BACBENCH: EVALUATING GENOMIC LANGUAGE MODELS FOR BACTERIA

Anonymous authors
Paper under double-blind review

ABSTRACT

Bacteria underpin key processes in health, ecology, and biotechnology, yet machine learning in bacterial genomics lacks systematic, large-scale evaluation resources. Current resources are typically limited to single-species datasets, where the small number of available genomes leaves species-specific models underpowered, underscoring the need for approaches that can generalize across the bacterial tree of life. To address this gap, we present BacBench, the first comprehensive benchmark for bacterial genomics. BacBench consists of 11 datasets across 6 tasks, including a newly generated dataset for operon identification derived from long-read RNA sequencing. BacBench covers gene-, system-, and genome-scale prediction tasks, spanning 67k genomes, 17.6k species and 255M proteins. We analyze the performance of state-of-the-art DNA LMs, protein LMs and bacterial LMs and find that while each approach excels at different scales—the existing models fail to accurately predict the bacterial phenotype at a whole-genome level, hampering the translation to high-impact applications such as antibiotic-resistance and bioproduction. Therefore, highlighting the need to develop methods that reason over the context of the entire genomes, exploiting genomic synteny and transfer across species. We outline the key requirements for such models and release a standardized library for preprocessing, embedding, and evaluation, fostering the development of methods that accurately represent bacterial genomes, and enabling reproducible comparison of diverse approaches under a unified framework. By providing the first comprehensive benchmark dedicated to bacterial genomics, BacBench lays the ground-work for developing machine learning models that truly exploit shared evolutionary patterns and generalize across the bacterial tree of life.

1 INTRODUCTION

Bacteria drive indispensable processes in medicine, ecology, and biotechnology (de Steenhuijsen Piters et al., 2015; Luo et al., 2024). They produce industrial enzymes and antibiotics (Ariaeenejad et al., 2024; Santos-Júnior et al., 2024), recycle nutrients, and are being engineered for carbon capture and waste remediation (Xu & Jiang, 2024). Unlocking this potential hinges on interpreting bacterial genomes at scale. A machine learning (ML) system that can embed and reason over entire bacterial genomes could predict clinically relevant traits, surface novel enzymes for biomanufacturing, and reveal how genetic variation translates into functional capabilities across species.

Traditionally, ML approaches in bacterial genomics have been species-specific. For example, genome-wide association studies (GWAS) have been successful in identifying genotype–phenotype associations within species, such as antimicrobial resistance or virulence traits (Lees et al., 2016; Power et al., 2017; San et al., 2020). However, species-specific datasets typically include only a few genomes, leaving models statistically underpowered and prone to overfitting given the vast mutation space of bacterial genomes.

Meaningful progress therefore requires models that can share information across species. Such transfer is biologically plausible: every bacterium carries a small core of universal single-copy proteins, and many additional gene families are conserved far beyond the species level (Wang et al., 2022; Lang et al., 2013; Coleman et al., 2021). Leveraging these shared signals allows models to capture the full extent of bacterial diversity, leading to predictions that are both robust and generalizable. Training on genomes from many species, laboratories, and environmental contexts

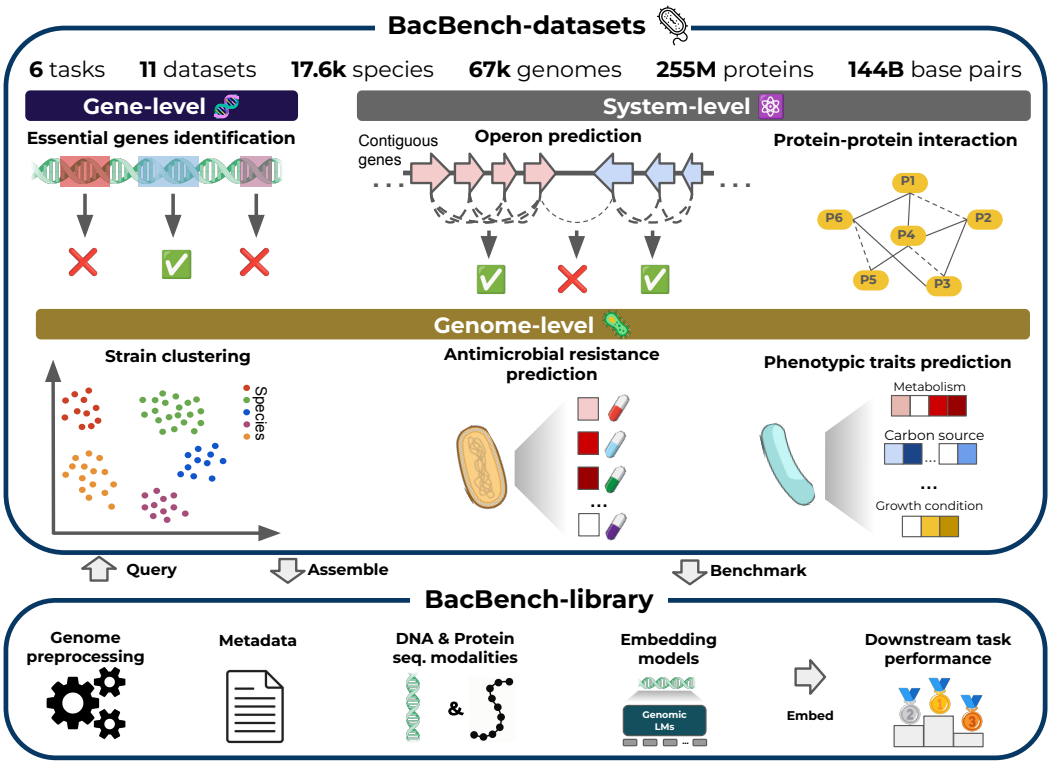


Figure 1: **BacBench overview.** We collected a diverse set of tasks at gene-, system- and genome-scales, spanning 17.6k bacterial species. `BacBench-library` provides support to preprocess and embed the datasets with various models. Finally, we performed systematic benchmarking for each task using diverse genomic LMs.

further guards against dataset-specific artifacts, making the resulting models less prone to sampling bias and more reliable when deployed on previously unseen strains or sequencing pipelines.

Recent breakthroughs in genomic sequence modeling show that the resulting representations can generalize across species and capture evolutionary signals (Nguyen et al., 2024; Dalla-Torre et al., 2024; Zhou et al., 2023; Lin et al., 2022; Elnaggar et al., 2021; Lin et al., 2023; Hayes et al., 2025). However, within the bacterial domain specifically, evaluation has been fragmented—either on narrow, single-task applications like antimicrobial resistance prediction (Wiatrak et al., 2024) or as part of cross-kingdom evaluations that fail to address bacteria-specific challenges (Nguyen et al., 2024). Consequently, the field lacks a dedicated, multi-scale benchmark for bacterial genomics, where genomic and metabolic mechanisms differ substantially from eukaryotes. Thus, leading to the development of models which can accurately model whole bacterial genomes.

Here, we introduce **BacBench**, the first multi-scale, multi-task and multi-species benchmark designed to evaluate ML for bacterial genomics (Fig. 1). BacBench has been collated from a diverse set of 11 datasets organized into 6 tasks spanning multiple biological scales: gene, system and genome. We consider essential gene prediction task at the gene-level, operon and protein–protein interaction prediction tasks at the system-level; and strain clustering, antibiotic resistance, and phenotypic traits prediction tasks at the genome-level. Collected from a diverse set of public resources and newly generated data, these datasets encompass 67k genomes spanning more than 17.6k bacterial species and 255M proteins. Using BacBench, we conduct comprehensive evaluation of distinct approaches to modeling bacterial genomes including DNA LMs, protein LMs (pLMs) and bacterial LMs (bLMs). We find that different modeling approaches excel at different tasks and scales, yet all models achieve low performance on phenotype prediction at a whole-genome level. Our results demonstrate *(i)* the need to develop ML approaches which can accurately model entire bacterial genomes, *(ii)* the benefits of pretraining on bacteria-specific corpora, rather than cross-kingdom ones, thus learning genomic mechanims that are unique to bacteria, and *(iii)* the importance of selecting the right model for the task at hand.

BacBench datasets and an accompanying toolkit for preprocessing and evaluation ensure straightforward reuse and extension[1]. By generating and integrating these datasets, we introduce the first benchmark suite for bacterial genomics, aiming to catalyze the development of ML methods that can transfer knowledge across species, unlocking new discoveries in bacterial genomics.

2 Related Work

Overall, existing resources for evaluating ML in bacterial genomics are limited in several key ways: they often lack functional labels, restrict evaluation to single species, or evaluate only one biological scale. BacBench addresses these limitations by providing a multi-scale and multi-task benchmark across the bacterial tree of life. It is specifically designed to reflect the core challenges in the field, including data sparsity within individual species, the need for cross-species generalization, and the importance of assessing model performance from the level of individual genes to entire genomes.

Bacterial genomics datasets. Large-scale sequence repositories such as MGnify (Mitchell et al., 2023), IMG/M (Markowitz et al., 2012), and AllTheBacteria (Blackwell et al., 2024) catalogue millions of bacterial assemblies. These resources are indispensable for comparative genomics, yet they provide limited metadata labels, making them ill-suited for training or benchmarking ML models. Simultaneously, task-specific collections provide information on essential genes and phenotypic traits (Zhang et al., 2004; Madin et al., 2020; Brbić et al., 2016; Weimann et al., 2016; Consortium, 2022)—supply richer annotations but usually cover only a handful of species and a single prediction setting. Diverse Genomic Embedding Benchmark (DGEB) (West-Roberts et al., 2024) offers tasks drawn from all domains of life, yet its bacterial coverage is mostly limited to single-species gene or short-segment datasets and lacks genome-scale evaluations such as broad phenotypic traits inference or strain-level clustering, leaving it unable to assess whether models generalize across thousands of species at multiple scales. BacBench complements these efforts by integrating six heterogeneous tasks, ranging from gene essentiality to genome-wide phenotype prediction - into a unified framework that explicitly tests generalization across 17.6 k bacterial species and three biological scales (Fig. 1).

Single-species bacterial genomics models. Despite the existence of large-scale bacterial genomics datasets, the ML applications in bacterial genomics have been mostly confined to single-species and single-task problems. For instance, genome-wide association studies (GWAS) have been successful in identifying genotype–phenotype associations within species, such as antimicrobial resistance or virulence traits (Lees et al., 2016; Power et al., 2017; San et al., 2020). More recent approaches such as unitig-based and deep learning models improved genotype–phenotype mapping by spanning the full pangenome (Lees et al., 2018; 2020) and predicting the effect of mutation based on the genomic context (Wiatrak et al., 2024). While the single-species models often perform well in their domains, they do not extend to other taxa and new genomic variants, and the huge genomic feature space relative to the number of labelled isolates leaves them prone to overfitting.

Genomic LMs. **DNA LMs** learn DNA sequence representations and have been shown to accurately represent long sequences (Zhou et al., 2023; Dalla-Torre et al., 2024; Jiang et al., 2023; Mourad, 2025; Nguyen et al., 2024; 2023), but are usually evaluated on human regulatory tasks, where transcriptional mechanisms and epigenomic regulation differ substantially from bacteria (Casadesús & Low, 2006). Moreover, even the DNA LMs with very large context window cannot span entire medium-sized bacterial genomes (Brixi et al., 2025). **Protein LMs** (pLMs) learn representations that correlate with structure, stability, and function (Elnaggar et al., 2021; Lin et al., 2022; 2023; Hayes et al., 2025). As bacterial genomes consist largely of coding sequence and possess simpler regulatory architectures than eukaryotes, pLMs can capture a substantial fraction of relevant biology. pLMs model proteins in isolation, however, characterizing bacterial genomes requires modeling the contextual interactions between the proteins present in the genome. Finally, we differentiate a third group of genomic LMs - **Bacterial LMs (bLMs)** which are recently proposed genomic LMs that are purposefully built to model bacterial genomes. These include gLM2 (Cornman et al., 2024), a mixed-modality genomic LM that represents coding regions as amino acids and intergenic regions as nucleotides to model contiguous genome context, and Bacformer, a genome-level contextual protein LM that treats each bacterial genome as an ordered sequence of proteins, refining each protein vector in the presence of

[1] https://anonymous.4open.science/r/BacBench-B6EF

Table 1: Summary of benchmarked models. "Max ctx." = maximum context length supported at inference; "dim" = dimensionality of the output of the last hidden layer. DNA LMs, pLMs and bLMs are separated by a horizontal line.

Model	Input	Objective	Tokenisation	Params	dim	Training corpus	Max ctx.
Mistral-DNA (Mourad, 2025)	DNA	Autoregressive	Byte-pair	138M	768	Bacteria	512
DNABERT-2 (Zhou et al., 2023)	DNA	Masked	Byte-pair	117M	768	Multi-kingdom	512
Nucleotide Transformer (Dalla-Torre et al., 2024)	DNA	Masked	*k*-mer	250M	768	Multi-kingdom	2,048
ProkBERT (Ligeti et al., 2024)	DNA	Masked	*k*-mer	27M	384	Bacteria	4,096
Evo (Nguyen et al., 2024)	DNA	Autoregressive	Single nucleotide	6.5B	4,096	Multi-kingdom	8,192
ESM-2 (Lin et al., 2022)	Single protein seq.	Masked	Single amino acid	35M	480	Multi-kingdom	1,024
ESM-C (ESM Team, 2024)	Single protein seq.	Masked	Single amino acid	300M	960	Multi-kingdom	1,024
ProtBERT (Elnaggar et al., 2021)	Single protein seq.	Masked	Single amino acid	420M	1,024	Multi-kingdom	1,024
gLM2 (Cornman et al., 2024)	Mixed modality (DNA & protein seq.)	Masked	Single nucleotide/amino acid	650M	1,280	Bacteria	4,096
Bacformer (Wiatrak et al., 2025)	Multiple protein seq.	Masked	Single protein	27M	480	Bacteria	6,000

all other proteins from the same genome, thus, encoding organism-level context. Both methods model the DNA or protein in the context of a bacterial genome and are pretrained on extensive bacterial corpora.

3 Bacterial Genome Representations & Baselines

We selected a diverse set of five DNA LMs, three pLMs and two bLMs to represent bacterial genomes and evaluate their performance across distinct tasks and scales. These genomic LMs take as input DNA, proteins, or both (gLM2) and can therefore generalize across the bacterial tree of life. Moreover, the suite spans modalities (DNA-only, single-protein, mixed DNA–protein, genome-level protein), objectives (masked vs. autoregressive), and training corpora (bacteria-specific vs. cross-kingdom), enabling a controlled comparison of how context length, modality, and pretraining data shape genome embeddings (Table 1).

Bacterial genomes representations. For the gene- and system-level tasks with **DNA LMs** we embed the coding sequence plus upstream promoter of the gene; we split sequences longer than the model limit L into overlapping windows of length L and average their embeddings across all windows. For genome-level tasks, we tile each genome into chunks with overlap, embed each chunk, and average the results to extract a genome embedding (Appendix B). For the **pLMs** we embed each protein present in the bacterial genome independently and average its residue embeddings. To generate genome-scale representations, we calculate the mean of all protein vectors (Appendix B). Finally, for **bLMs** we use contiguous, genome-aware inputs: for **gLM2**, we feed mixed-modality genomic segments that encode coding regions as amino acids and intergenic regions as nucleotides, preserving local genome context across adjacent genes; for **Bacformer**, we represent each genome as an ordered sequence of proteins by obtaining per-protein tokens from its base pLM (ESM-2 35M), and pass these through a transformer to learn contextualized, genome-level protein embeddings.

4 Prediction tasks and evaluation results

In BacBench, we consider six tasks across three scales: *(i)* gene, *(ii)* system, and *(iii)* genome. At gene-level, we consider the task of gene essentiality. At system-level, we assess operon identification and protein-protein interaction prediction tasks. Finally, at genome-level we evaluate methods on strain clustering, antibiotic resistance and phenotypic traits prediction tasks. We briefly describe each task and include further experimental details in the Appendices A & B.

4.1 Gene essentiality prediction

Identifying essential genes is crucial for *(i)* defining the minimal set of cellular functions, and *(ii)* prioritizing drug targets. To distinguish essential from non-essential genes, the methods need to generalize to phylogenetically diverse bacteria beyond the species in the training set. For each model we report the performance by *(i)* fitting a linear classifier on top of the frozen gene embeddings, *(ii)* finetuning the model to predict a binary label (i.e. essentiality of input gene; Appendix B).

Data. We compile the dataset from the Database of Essential Genes (DEG) (Zhang et al., 2004), a hand-curated resource which aggregates studies published for a broad range of bacteria. After quality-control filtering, the corpus comprises 51 distinct genomes spanning 37 species (Appendix

A). This amounts to $22,486$ essential and $146,922$ non-essential genes. To prevent train-test leakage, we split by genus—placing all genomes from a genus in one split—and evaluate on held-out genera, enforcing generalization to phylogenetically distant strains.

Metrics. Gene essentiality prediction is a binary classification task. We evaluate performance using AUROC and AUPRC and report macro-average metrics across test genomes.

Results. pLMs and bLMs substantially outperform DNA-based models in terms of both AUROC (Fig. 2) and AUPRC using both linear probing and finetuning (Appendix C), suggesting that essential genes share conserved protein motifs that the protein-based embeddings capture. bLMs perform the best, with gLM2 achieving the best results overall, and Bacformer significantly outperforming its backbone ESM-2 model, showing the benefits of incorporating genome context. When considering DNA LMs only, Evo achieves the best performance, demonstrating the benefits of scaling model and context size.

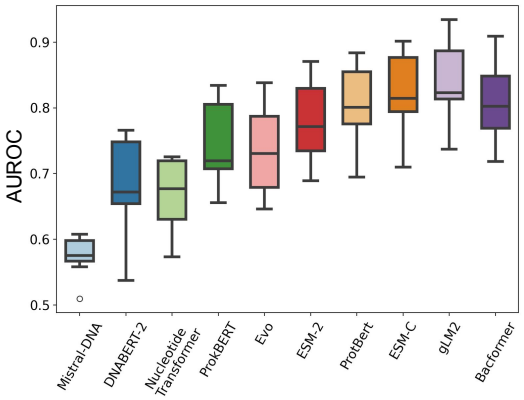


Figure 2: AUROC across genomes on essential gene prediction task using a linear model. The box spans the interquartile range with a line marking the median value.

4.2 Operon identification

Operons are multi-gene transcriptional units that underpin coordinated gene expression. Accurate operon maps are pivotal for *(i)* constructing gene regulatory networks (Fortino et al., 2014) and *(ii)* refining genome-scale metabolic models for strain engineering (Orth et al., 2010). In this task, the methods must predict whether the two neighbouring genes belong to the same operon, effectively predicting operon boundaries. Because available annotations are scarce we evaluate the task in a zero-shot setting.

Data. Due to the lack of experimentally validated operon annotations, we generated the operon labels by performing long-read RNA sequencing on a set of 5 diverse strains, amounting to $3,310$ unique operons (Appendix A).

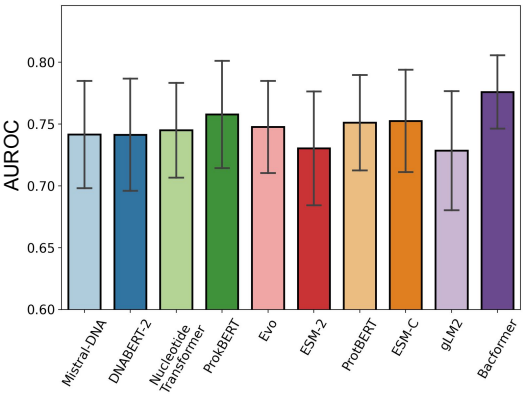


Figure 3: Zero-shot AUROC on operon identification task, the error bars represent standard error across strains.

Metrics. In BacBench, operon identification is a zero-shot binary classification task. We use the cosine similarity value between the two genes combined with the information on the genes' strand as a score indicating whether the genes belong to the same operon (Appendix A). We leverage AUROC and AUPRC for measuring performance. Finally, we report results across distinct strains.

Results. All methods except Bacformer attain similar performance (Fig. 3) in terms of both AUROC and AUPRC (Appendix C). We attribute the high performance of the Bacformer due to the *(i)* whole-genomic context of the model and *(ii)* extensive bacterial training corpus. ProkBERT performs 2nd best, showing that *(i)* scaling the model does not necessarily lead to improved results, *(ii)* the importance of a relevant pretraining corpus (ProkBERT was pretrained only on prokaryotes). Notably, both DNA and pLMs perform similarly on the task, indicating that the choice of modality is less important than incorporating genome-level context and domain-matched pretraining. Finally, the low overall performance across models highlights the need for task-specific methods and generating datasets which would allow for finetuning.

4.3 PROTEIN-PROTEIN INTERACTION PREDICTION

Mapping protein–protein interactions (PPI) is central to *(i)* reconstructing gene networks (Snider et al., 2015) and *(ii)* prioritizing drug-target combinations (Wilson et al., 2022). Compared to the operon identification task where interacting genes lie next to each other, in the PPI task proteins can be separated by millions of base pairs. In this task each input is a pair of proteins, and the goal is to predict whether two proteins interact. Because interaction data are available at the protein-level, we restrict the benchmark to models which only require protein sequence data, specifically, pLMs and Bacformer. We evaluate two regimes: *(i)* zero-shot: the cosine similarity between the frozen embeddings of the two proteins serves as the interaction score, *(ii)* finetuned: the averaged embeddings of the two proteins are passed through a linear classifier trained to output a binary class.

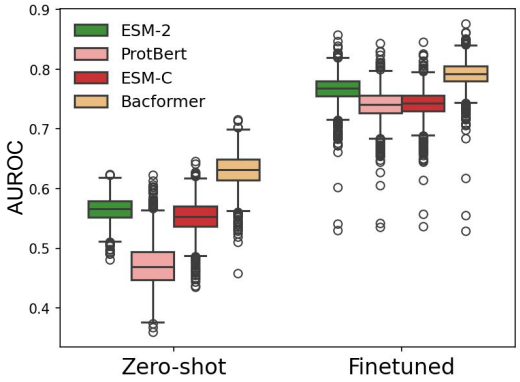


Figure 4: AUROC across genomes on PPI task in the zero-shot and finetuned setting. The box spans the interquartile range with a line marking the median value.

Data. We downloaded and processed all $10,533$ bacterial strains available in STRING DB (Szklarczyk et al., 2023) together with associated PPI scores. For every strain we extract the combined interaction score for each protein pair; the median strain contains almost $640,000$ scored pairs. We binarize the labels, resulting in roughly 10% of all interactions being positive (Appendix A).

Metrics. PPI is a binary classification task. We report performance with AUROC and AUPRC, macro-averaged over test genomes. Both the positive and negative labels are provided by STRING.

Results. In both zero-shot and finetuned setups, Bacformer achieves substantially higher scores than other methods. We attribute this to the rotary positional embeddings (Su et al., 2024), which increase the cosine similarity score between neighbouring genes. This is in line with experimental data, which shows that the neighbouring genes in the bacterial genomes tend to interact with each other (Dandekar et al., 1998). We also notice how the difference in performance between the methods decreases following finetuning, demonstrating how the models can adapt to the task during training. Finally, the performance of ESM-2 compared to ESM-C and ProtBERT shows that scaling up the training data and model size does not benefit PPI prediction in bacteria.

4.4 STRAIN CLUSTERING

Rapid clustering of whole genomes is valuable for placing newly sequenced metagenome assembled genomes (MAGs) into the bacterial tree of life and quality control. Therefore, we propose a metagenomic strain clustering task where the goal is to recover taxonomy using genome only. In this task, we feed every MAG to the model without using any species tokens or other metadata—so evaluation is fully zero-shot. We then evaluate whether models' genome embeddings preserve the taxonomy. A good embedding should cluster genomes from the same species close together and, at broader levels, conserve members of e.g. the same genus or family. We perform the clustering by computing k nearest neighbors across different k and running Leiden clustering (Traag et al., 2019) at various resolutions (Appendix B).

Data. In this BacBench task, we draw $6,071$ strains from MGnify (Mitchell et al., 2023), spanning 25 species distributed across 10 genera and 7 families chosen to give a balanced phylogenetic spread. We process DNA and protein inputs in the same way for every model; and average the vector from the final hidden layer over the whole genome to yield one fixed-length embedding per genome.

Metrics. We quantify clustering performance with the adjusted Rand index (ARI), normalized mutual information (NMI) and average silhouette width (ASW). Higher is better for all metrics. To obtain cluster assignments for each model we run Leiden clustering across resolutions, retaining the resolution that maximized the mean of the three metrics over the species, genus and family ranks. The ASW is unaffected by taxonomic rank, so we report it once in the *Combined* section.

Table 2: Clustering performance (higher is better) across taxonomic ranks and metrics. Best value for each metric is bolded. DNA LMs, pLMs and bLMs are separated by a horizontal line.

Method	Species		Genus		Family		Combined		
	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ASW
Mistral-DNA	89.57	94.85	69.04	86.60	48.52	78.56	69.04	86.67	39.88
DNABERT-2	97.10	98.33	64.98	85.81	43.85	76.35	68.64	86.83	63.10
Nucleotide Transformer	98.14	98.89	64.06	85.28	43.14	75.84	68.45	86.67	39.24
ProkBERT	**98.75**	**99.38**	63.55	84.88	42.76	75.46	68.35	86.58	**65.03**
Evo	55.56	76.98	49.28	72.42	35.53	66.58	46.79	72.00	25.33
ESM-2	50.94	71.67	56.82	72.20	46.84	69.06	51.53	70.97	16.75
ESM-C	72.65	87.70	79.76	89.87	60.39	83.43	70.93	87.00	29.55
ProtBERT	68.53	86.47	**85.98**	**92.40**	**65.78**	**86.42**	**73.43**	**88.43**	39.12
gLM2	72.33	84.03	66.60	82.27	49.03	75.93	62.65	80.74	40.16
Bacformer	79.39	90.62	77.92	90.20	54.74	80.71	70.68	87.17	30.12

Results. At the species level, DNA LMs attain the highest scores, indicating that species boundaries can be recovered from sequence alone and consistent with the extra signal carried in non-coding regions. Moving up to genus and family levels, the advantage shifts: pLMs and Bacformer overtake the DNA LMs, suggesting that protein embeddings retain deeper evolutionary relationships more faithfully. Aggregated across all ranks, ProtBERT shows the highest ARI and NMI values, whereas the DNA models have the overall highest ASW. Interestingly, gLM2 which is a mixed modality model combining DNA and protein sequences performs worse than other models. Upon investigation, we believe this is due to the large variance in its embeddings and suggest that a mixed modality method with improved regularization could capture both fine-grained strain identity and higher-order phylogeny in a single embedding space.

4.5 Antibiotic resistance prediction

Predicting antibiotic resistance is a task with *(i)* immediate clinical value for guiding antimicrobial drug therapy, *(ii)* monitoring the spread of resistant lineages and *(iii)* prioritizing compounds in drug-discovery pipelines. In this BacBench task, we define two subtasks: (1) given a bacterial genome, predict whether the strain is resistant or susceptible to a specific drug, and (2) given a bacterial genome, estimate its minimum inhibitory concentration (MIC). The first subtask is a binary classification problem (resistant vs susceptible), while the second subtask is a regression problem. Due to the computational complexity of computing genome-level embeddings on over $25k$ genomes (Appendix B), we *(i)* do not include Evo in the analysis due to its size (6.5B parameters), which makes it computationally infeasible to embed the entire corpus in most academic environments (see *Runtime analysis*; Appendix B), *(ii)* perform evaluation by stacking a linear layer on top of the frozen genome representation, and fine-tune a separate linear classifier for each model.

Data. We assemble a cross-species panel from the NIH Antimicrobial Susceptibility Test browser (National Center for Biotechnology Information, 2025), covering $25,032$ strains drawn from 38 bacterial species. After quality control and removal of sparsely sampled drugs, the dataset retains 36 antibiotics for binary label and 56 for regression prediction that span diverse classes of antibiotics (Appendix A). For the binary task we use the resistant/susceptible calls and discard any ambiguous entries (Appendix A). For the regression task we extract the raw MIC values, apply a $log1p$ transformation to dampen heavy tails and train a separate linear model for each drug–model pair.

Metrics. We report performance in the classification setting with AUROC and AUPRC across drugs. For the MIC regression we compute the *Pearson* correlation coefficient and the coefficient of determination (R^2) averaged across antibiotics. We report mean scores across antibiotics and include full per-antibiotic tables in the Appendix C.

Results. On both binary and regression setup, bLMs and pLMs tend to outperform DNA LMs. This may be explained by the fact that resistance is usually acquired through the mutation in the coding sequence, with studies showing that 90% of characterized resistance-conferring variants reside in coding regions (Sandgren et al., 2009; Farhat et al., 2019). Finally, the bLM-Bacformer achieves the best results implying the importance of epistatic effects on antibiotic resistance (Trindade et al., 2009) which the model considers by modeling the interactions between all of the proteins present in the genome. Notably, the methods record highly variable performance across antibiotics, underscoring the difficulty of building a single model that generalizes across the resistance mechanisms.

Table 3: Performance (%, higher is better) on the antibiotic-resistance prediction tasks across drugs. Values are mean ± standard deviation across 3 runs; best for each metric is bolded. DNA LMs, pLMs and bLMs are separated by a horizontal line.

Method	Binary		Regression	
	AUPRC	AUROC	R^2	Pearson
Mistral-DNA	49.86 ± 22.73	76.05 ± 9.14	19.71 ± 17.37	40.95 ± 21.51
DNABERT-2	52.35 ± 23.55	79.88 ± 7.26	23.61 ± 18.33	45.89 ± 19.27
Nucleotide Transformer	59.70 ± 22.90	84.22 ± 6.47	27.74 ± 17.80	51.19 ± 16.76
ProkBERT	58.90 ± 23.67	83.94 ± 6.39	28.00 ± 17.57	51.24 ± 16.91
ESM-2	62.04 ± 23.64	85.05 ± 6.63	31.18 ± 17.39	54.60 ± 16.15
ESM-C	63.41 ± 23.43	85.68 ± 6.81	33.15 ± 17.62	56.79 ± 15.46
ProtBERT	61.79 ± 23.19	84.51 ± 6.31	28.23 ± 18.46	51.56 ± 17.61
gLM2	55.80 ± 24.46	82.16 ± 6.75	26.15 ± 18.13	49.41 ± 17.14
Bacformer	**67.97** ± 20.73	**87.61** ± 5.68	**33.84** ± 18.60	**57.53** ± 15.43

4.6 PHENOTYPIC TRAITS PREDICTION

Accurately predicting phenotype from a genomic sequence enables *(i)* inference of the biological or ecological function of bacteria (Feldbauer et al., 2015) and *(ii)* engineering organisms with the exact metabolic traits needed for efficient waste remediation (Rafeeq et al., 2023), accelerating sustainable industrial bioproduction (Lawson et al., 2021). We evaluate whether the models' genome embeddings are predictive of diverse phenotypic traits. Here, given a genome embedding, the task is to predict a trait. Similarly as in the antibiotic resistance prediction task above, we perform linear probing evaluation, training a separate linear classifier per phenotype and exclude the Evo model due to the computational costs (Appendix B).

Data. To create the benchmark we collated large trait inventories (Madin et al., 2020; Brbić et al., 2016; Weimann et al., 2016), apply stringent quality filters and discard traits represented by only a handful of isolates (Appendix A). The final corpus covers 139 discrete phenotypes spanning $15,477$ bacterial species, making it the broadest dataset of its kind and challenging the models to generalize well beyond their training clades. We group traits into carbon utilisation, biochemical activity, growth conditions and cellular morphology (Appendix A), which we use later for stratified analysis. As similar genomes often share the same phenotype, we split the data for each phenotype by genus—placing all genomes from a genus in one split—and evaluate on held-out genera, enforcing generalization to phylogenetically distant strains.

Metrics. For BacBench, we restrict the phenotypic traits to categorical traits due to their sufficient number of available samples, and evaluate performance using the macro-averaged AUROC and AUPRC over phenotype categories. We report mean scores for every phenotype group and include full per-trait tables in the Appendix C.

Results. pLMs and bLMs outperform DNA LMs across phenotype groups and metrics (Fig. 5 & Appendix C). The relative difference is especially large for biochemical activity and carbon utilization traits, likely because these phenotypes hinge on enzyme active-site composition and pathway membership—information that is explicit in amino-acid space and can be better captured

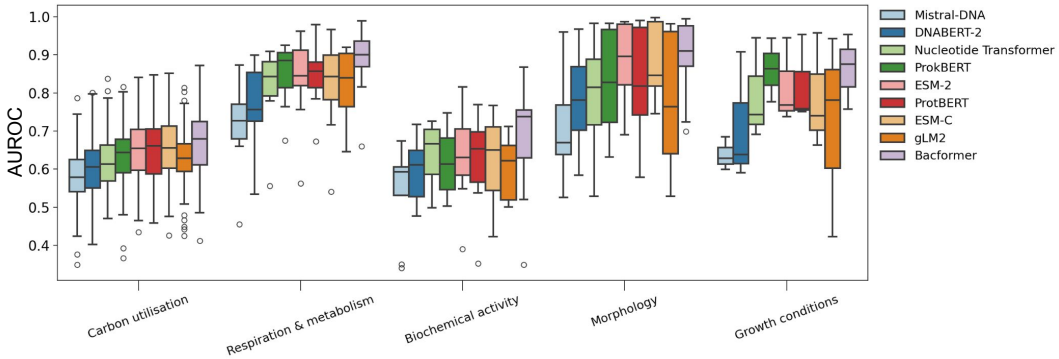


Figure 5: AUROC across diverse phenotypic traits groups and methods. The box spans the inter-quartile range with a line marking the median value.

by pLMs (Teukam et al., 2024; Lawson et al., 2021). The Bacformer bLM attains the highest scores, indicating that whole-genome context helps explain phenotypes arising from coordinated protein function and epistatic interactions (Trindade et al., 2009). In contrast, Mistral-DNA, the only autoregressive model in the task, lags well behind the other models. Overall performance remains moderate—especially for carbon utilization, biochemical activity and growth condition traits—highlighting considerable room for improvement. Progress will likely require more complex contextualized models that incorporate environmental metadata, together with larger, more balanced phenotype datasets.

5 DISCUSSION

Summary. BacBench addresses the lack of a comprehensive, cross-species evaluation resource for bacterial genomics by providing unified datasets and benchmarks, and by evaluating existing genomic language models across species, tasks, and biological scales. It introduces a newly generated dataset for operon identification—a key problem for refining genome-scale metabolic models—and curates five additional tasks (gene essentiality, protein–protein interaction, strain clustering, antibiotic resistance, and 139 phenotypic traits) into a single framework covering 67k genomes from 17.6k species. Our experiments show that *(i)* existing genomic LMs (DNA LMs, pLMs, and bLMs) capture core taxonomic structure and functional relatedness, providing a strong baseline representation of bacterial genomes, but fall short at accurate genome-to-phenotype prediction, as evidenced by antibiotic resistance and phenotype tasks; *(ii)* models purpose-built for bacteria (gLM2, Bacformer) or trained on bacteria-specific corpora (ProkBERT) tend to outperform broad, cross-kingdom counterparts across most tasks, underscoring the value of domain-matched pretraining and inductive biases; *(iii)* different modeling approaches excel at different problems—DNA LMs capture fine-grained taxonomic signals and do well on operon identification, pLMs better preserve deeper phylogeny and functional similarity for strain clustering, and bLMs like Bacformer perform best on tasks driven by multi-gene interactions (phenotypes, antibiotic resistance). All datasets are accompanied with extensive documentation; the embedding and evaluation library is provided at `https://anonymous.4open.science/r/BacBench-B6EF`.

Towards accurate genome-to-phenotype bacterial prediction. Our results suggest a practical path forward towards building model for genome-to-phenotype mapping in bacteria: *(i)* the model should represent entire genomes end-to-end to capture long-range, cross-protein dependencies (currently only feasible in Bacformer); *(ii)* pretrain on substantially larger, bacteria-focused corpora (we estimate >4M unique strains remain untapped (Mitchell et al., 2023; Markowitz et al., 2012; Blackwell et al., 2024)); *(iii)* integrate DNA and protein modalities—and where available RNA—to couple regulatory and coding signals in a single embedding while remaining computationally efficient; *(iv)* allow inclusion of structured priors (e.g., resistance gene catalogs, operon maps, HGT markers) to improve data efficiency on scarce phenotype labels; *(v)* expand high-quality phenotype supervision (including knock-outs and standardized trait panels) to close the supervision gap limiting genome-to-phenotype learning.

Limitations & Future Work. By releasing BacBench we provide a foundation for more expressive, cross-species models of bacterial genomics, yet the present benchmark covers only part of the functional landscape. Other tasksa re not yet included, and certain phenotypes and antibiotic classes remain sparse (Appendix A), underscoring the need for generating new data. We benchmarked a representative set of publicly available models capable of cross-species generalization, and expect the model suite to expand in future iterations. Finally, the current tasks do not include modalities such as transcriptomics and metabolomics due to data sparsity and inconsistent metadata. As community datasets mature, incorporating multi-omics will enable more faithful evaluation of causal, context-dependent genome-to-phenotype mappings.

We anticipate that subsequent iterations will broaden task and model coverage, ultimately enabling contextual, genome-scale representations for bacterial genomes. Community contributions of datasets, models, and evaluation routines are encouraged so that BacBench evolves into a continually updated standard for bacterial ML.

6 Reproducibility Statement

We release an *anonymous* codebase for preprocessing, embedding, and evaluation at `https://anonymous.4open.science/r/BacBench-B6EF`, together with helper utilities for bacterial genomics to lower the barrier for adding new models and tasks. All datasets in BacBench are fully documented and accompanied by anonymized MLCommons `Croissant` metadata files in the supplementary materials, enabling unambiguous data loading and lineage tracking. The Appendix details quality filtering, preprocessing pipelines, train/validation/test splits, model and training configurations, and hyperparameters; it also specifies random seeds, hardware, and optimization settings. Further details on the exact scripts to reproduce the results can be found in the anonymous repository linked above. These artifacts together provide the necessary pointers—code, data descriptors, and experimental specifications—to reproduce our results and to extend BacBench as an evolving benchmark for the ML community.

References

Shohreh Ariaeenejad, Javad Gharechahi, Mehdi Foroozandeh Shahraki, Fereshteh Fallah Atanaki, Jian-Lin Han, Xue-Zhi Ding, Falk Hildebrand, Mohammad Bahram, Kaveh Kavousi, and Ghasem Hosseini Salekdeh. Precision enzyme discovery through targeted mining of metagenomic data. *Natural Products and Bioprospecting*, 14(1):7, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Grace Blackwell, Karel M. Brinda, John A. Lees, Rebecca A. Gladstone, Philip Brownridge, et al. Allthebacteria: a uniformly assembled and searchable compendium of all public bacterial genomes. *bioRxiv*, pp. 2024.03.08.584059, 2024. doi: 10.1101/2024.03.08.584059.

Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. The landscape of microbial phenotypic traits and associated genes. *Nucleic acids research*, pp. gkw964, 2016.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pp. 2025–02, 2025.

Josep Casadesús and David Low. Epigenetic gene regulation in bacteria. *Molecular Microbiology*, 60(4):820–826, 2006. doi: 10.1111/j.1365-2958.2006.05136.x.

Gareth A Coleman, Adrián A Davín, Tara A Mahendrarajah, Lénárd L Szánthó, Anja Spang, Philip Hugenholtz, Gergely J Szöllősi, and Tom A Williams. A rooted phylogeny resolves early bacterial evolution. *Science*, 372(6542):eabe0511, 2021.

CRyPTIC Consortium. A data compendium associating the genomes of 12,289 mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics. *PLoS biology*, 20(8): e3001721, 2022.

Andre Cornman, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin Beracochea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. The omg dataset: An open metagenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, pp. 2024–08, 2024.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pp. 1–11, 2024.

Thomas Dandekar, Berend Snel, Martijn Huynen, and Peer Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*, 23(9):324–328, 1998.

Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 02 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008. URL https://doi.org/10.1093/gigascience/giab008.giab008.

Wouter AA de Steenhuijsen Piters, Elisabeth AM Sanders, and Debby Bogaert. The role of the local microbial ecosystem in respiratory health and disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1675):20140294, 2015.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning. https://evolutionaryscale.ai/blog/esm-cambrian, 2024. EvolutionaryScale website, published 4 December 2024, accessed 17 April 2025.

Maha R Farhat, Luca Freschi, Roger Calderon, Thomas Ioerger, Matthew Snyder, Conor J Meehan, Bouke de Jong, Leen Rigouts, Alex Sloutsky, Devinder Kaur, et al. Gwas for quantitative resistance phenotypes in mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nature communications*, 10(1):2128, 2019.

Roman Feldbauer, Frederik Schulz, Matthias Horn, and Thomas Rattei. Prediction of microbial phenotypes based on comparative genomics. *BMC bioinformatics*, 16:1–8, 2015.

Vittorio Fortino, Olli-Pekka Smolander, Petri Auvinen, Roberto Tagliaferri, and Dario Greco. Transcriptome dynamics-based operon prediction in prokaryotes. *BMC bioinformatics*, 15:1–14, 2014.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jenna Morgan Lang, Aaron E Darling, and Jonathan A Eisen. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PloS one*, 8(4):e62510, 2013.

Christopher E Lawson, Jose Manuel Martí, Tijana Radivojevic, Sai Vamshi R Jonnalagadda, Reinhard Gentz, Nathan J Hillson, Sean Peisert, Joonhoon Kim, Blake A Simmons, Christopher J Petzold, et al. Machine learning for metabolic engineering: A review. *Metabolic Engineering*, 63:34–60, 2021.

John A. Lees, Minna Vehkala, Niko Välimäki, Simon R. Harris, and Claire *et al.* Chewapreecha. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications*, 7:12797, 2016. doi: 10.1038/ncomms12797.

John A Lees, Marco Galardini, Stephen D Bentley, Jeffrey N Weiser, and Jukka Corander. Pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, 34(24):4310–4312, 2018.

John A Lees, T Tien Mai, Marco Galardini, Nicole E Wheeler, Samuel T Horsfield, Julian Parkhill, and Jukka Corander. Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *MBio*, 11(4):10–1128, 2020.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

Balázs Ligeti, István Szepesi-Nagy, Babett Bodnár, Noémi Ligeti-Nagy, and János Juhász. Prokbert family: genomic language models for microbiome applications. *Frontiers in Microbiology*, 14, 2024. ISSN 1664-302X. doi: 10.3389/fmicb.2023.1331233. URL https://www.frontiersin.org/articles/10.3389/fmicb.2023.1331233.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Zhaowei Luo, Zhanghua Qi, Jie Luo, and Tingtao Chen. Potential applications of engineered bacteria in disease diagnosis and treatment. *Microbiome Research Reports*, 4(1):N–A, 2024.

Joshua S Madin, Daniel A Nielsen, Maria Brbic, Ross Corkrey, David Danko, Kyle Edwards, Martin KM Engqvist, Noah Fierer, Jemma L Geoghegan, Michael Gillings, et al. A synthesis of bacterial and archaeal phenotypic trait data. *Scientific data*, 7(1):170, 2020.

Victor M Markowitz, I-Min A Chen, Krishna Palaniappan, Ken Chu, Ernest Szeto, Yuri Grechkin, Anna Ratner, Biju Jacob, Jinghua Huang, Peter Williams, et al. Img: the integrated microbial genomes database and comparative analysis system. *Nucleic acids research*, 40(D1):D115–D122, 2012.

Alex L. Mitchell, Antonio Almeida, Martin Beracochea, Matthew Boland, Jack Burgin, Guy Cochrane, et al. Mgnify: the microbiome analysis resource in 2023. *Nucleic Acids Research*, 51(D1):D723–D730, 2023. doi: 10.1093/nar/gkac1088.

Raphael Mourad. Mistral-DNA. https://github.com/raphaelmourad/Mistral-DNA, 2025. GitHub repository, accessed 17 April 2025.

National Center for Biotechnology Information. Antibiotic susceptibility test (ast) browser, 2025. URL https://www.ncbi.nlm.nih.gov/pathogens/ast. Accessed 24 Apr 2025.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

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. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Robert A. Power, Julian Parkhill, and Tulio de Oliveira. Microbial genome-wide association studies: lessons from human gwas. *Nature Reviews Genetics*, 18:41–50, 2017. doi: 10.1038/nrg.2016.132.

Hamza Rafeeq, Nadia Afsheen, Sadia Rafique, Arooj Arshad, Maham Intisar, Asim Hussain, Muhammad Bilal, and Hafiz MN Iqbal. Genetically engineered microorganisms for environmental remediation. *Chemosphere*, 310:136751, 2023.

James Emmanuel San, Shakuntala Baichoo, Aquillah Kanzi, Yumna Moosa, Richard Lessells, Vagner Fonseca, John Mogaka, Robert Power, and Tulio de Oliveira. Current affairs of microbial genome-wide association studies: approaches, bottlenecks and analytical pitfalls. *Frontiers in microbiology*, 10:3119, 2020.

Andreas Sandgren, Michael Strong, Preetika Muthukrishnan, Brian K Weiner, George M Church, and Megan B Murray. Tuberculosis drug resistance mutation database. *PLoS medicine*, 6(2):e1000002, 2009.

Célio Dias Santos-Júnior, Marcelo DT Torres, Yiqian Duan, Álvaro Rodríguez Del Río, Thomas SB Schmidt, Hui Chong, Anthony Fullam, Michael Kuhn, Chengkai Zhu, Amy Houseman, et al. Discovery of antimicrobial peptides in the global microbiome with machine learning. *Cell*, 187 (14):3761–3778, 2024.

Jamie Snider, Max Kotlyar, Punit Saraon, Zhong Yao, Igor Jurisica, and Igor Stagljar. Fundamentals of protein interaction network mapping. *Molecular systems biology*, 11(12):848, 2015.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

Yves Gaetan Nana Teukam, Loïc Kwate Dassi, Matteo Manica, Daniel Probst, Philippe Schwaller, and Teodoro Laino. Language models can identify enzymatic binding sites in protein sequences. *Computational and Structural Biotechnology Journal*, 23:1929–1937, 2024.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

Sandra Trindade, Ana Sousa, Karina Bivar Xavier, Francisco Dionisio, Miguel Godinho Ferreira, and Isabel Gordo. Positive epistasis drives the acquisition of multidrug resistance. *PLoS genetics*, 5(7): e1000578, 2009.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Saidi Wang, Minerva Ventolero, Haiyan Hu, and Xiaoman Li. A revisit to universal single-copy genes in bacterial genomes. *Scientific Reports*, 12(1):14550, 2022. doi: 10.1038/s41598-022-18762-z.

Aaron Weimann, Kyra Mooren, Jeremy Frank, Phillip B Pope, Andreas Bremges, and Alice C McHardy. From genomes to phenotypes: Traitar, the microbial trait analyzer. *MSystems*, 1(6): 10–1128, 2016.

Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse genomic embedding benchmark for functional evaluation across the tree of life. *bioRxiv*, pp. 2024–07, 2024.

Maciej Wiatrak, Aaron Weimann, Adam Dinan, Maria Brbić, and A. Floto, R.˙ Sequence-based modelling of bacterial genomes enables accurate antibiotic resistance prediction. *bioRxiv*, pp. 2024.01.03.574022, 2024. doi: 10.1101/2024.01.03.574022.

Maciej Wiatrak, Ramon Viñas Torné, Maria Ntemourtsidou, Adam Dinan, David C Abelson, Divya Arora, Maria Brbić, Aaron Weimann, and R Andres Floto. A contextualised protein language model reveals the functional syntax of bacterial evolution. *bioRxiv*, pp. 2025–07, 2025.

Jennifer L Wilson, Ethan Steinberg, Rebecca Racz, Russ B Altman, Nigam Shah, and Kevin Grimes. A network paradigm predicts drug synergistic effects using downstream protein–protein interactions. *CPT: Pharmacometrics & Systems Pharmacology*, 11(11):1527–1538, 2022.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Xihui Xu and Jiandong Jiang. Engineering microbiomes for enhanced bioremediation. *PLoS biology*, 22(12):e3002951, 2024.

Ren Zhang, Hong-Yu Ou, and Chun-Ting Zhang. Deg: a database of essential genes. *Nucleic acids research*, 32(suppl_1):D271–D272, 2004.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

A Additional Datasets & Tasks Details

In this section we outline the dataset details, including the overall and per dataset statistics as well as the preprocessing details. For each dataset we performed additional quality checks which we detail below. All of the genomes across datasets contain genome ID, which can be used to identify the genome source. Finally, we provide metadata on genes, proteins and genomes together with the datasets when available, together with extensive documentation.

Supplementary Table 1: Dataset summary for the six benchmark tasks used in this study.

Task	# Species	# Genomes	# Proteins	# Base pairs
Gene essentiality prediction	37	51	169 k	279 MB
Operon identification	5	5	22 k	25 MB
Protein–protein interaction prediction	6 956	10 533	36 M	N/A
Strain clustering	25	6 710	14 M	16 GB
Phenotypic traits prediction	15 477	24 462	100 M	111 GB
Antibiotic resistance prediction	38	26 302	105 M	112 GB

A.1 Gene essentiality prediction

We downloaded the gene essentiality annotations for bacteria across genomes from the Database of Essential Genes (DEG, `http://origin.tubic.org/deg`) (Zhang et al., 2004). Using the genome RefSeq ID provided in the database, we downloaded the associated genomes in both DNA and protein sequence modalities from the NCBI GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`). Across 66 genomes from DEG, there were multiple genomes with more than 98% overlap when it comes to annotations. We therefore removed these genomes, as including it could lead to inflated evaluation metrics, leaving us with 51 genomes across 37 distinct species. For each genome we provide start and end for each gene together with essentiality annotations (*Yes*=essential, *No*=non-essential), verifying the gene locations are correct. We also provide the strand of the gene to allow for the extraction of the region upstream of the gene.

Split. We performed a random data split into training, validation, and test sets in a 60 / 20 / 20 % ratio. Additionally, to prevent train-test leakage, we split by genus—placing all genomes from a genus in one split—and evaluate on held-out genera, enforcing generalization to phylogenetically distant strains.AUROC

A.2 Operon identification

Due to the lack or reliable whole-genome operon annotations, we performed long-read RNA sequencing to annotate operons across five distinct strains, processing three independent biological replicates for each strain.

Bacterial strains and culture conditions. Five bacterial strains were used in this experiment: *Staphylococcus aureus RN450* (*S. aureus* RN450), *Mycobacterium abscessus* ATCC 19977 (*M. ab*), ΔleuD ΔpanCD *Mycobacterium tuberculosis* H37Rv 102J23 (ΔleuD ΔpanCD *M. tb*), *Pseudomonas aeruginosa* PAO1, and *Escherichia coli DH5*α. Each strain was cultured in triplicate under nutrient-rich conditions until mid-exponential phase ($OD_{600} = 0.4$–0.6) was reached.

RNA isolation. Cells were harvested by centrifugation and processed for total RNA extraction using the MasterPure Complete DNA and RNA Purification Kit (Lucigen) with strain-specific modifications: For *S. aureus* RN450, cell pellets were pre-treated with lysostaphin (Tris buffer, pH 8.0) at 37 °C

for 30 min to aid lysis. For the mycobacterial strains (*M. ab* and ΔleuD ΔpanCD *M. tb*), cells were mechanically disrupted by bead beating in lysis buffer, followed by extraction with the standard kit protocol. Isolated RNA was treated twice with TURBO DNase (Invitrogen) to remove residual genomic DNA and purified using RNA Clean & Concentrator columns (Zymo Research). RNA integrity was assessed on an Agilent TapeStation RNA ScreenTape system, and concentrations were measured with a Qubit fluorometer (Invitrogen).

Library preparation and sequencing. For each replicate, 1,000 ng of total RNA underwent rRNA depletion with riboPOOLs (siTOOLs Biotech). The depleted RNA was polyadenylated with poly(A) polymerase (PAP) in the presence of a manganese catalyst, adding 50–90 adenosines per molecule. cDNA libraries were prepared with the Nanopore cDNA-PCR kit, pooled and sequenced on a PromethION device equipped with R10 flow cells (Oxford Nanopore Technologies).

Long-read RNA sequencing data preprocessing. For each sequenced strain, the following genome assemblies and gene annotations from NCBI RefSeq [ref] were used: `GCF_000013425.1` (***Staphylococcus aureus RN450***), `GCF_000069185.1` (***Mycobacterium abscessus ATCC 19977***), `GCF_000195955.2` (ΔleuD ΔpanCD ***Mycobacterium tuberculosis H37Rv 102J23***), `GCF_000006765.1` - (***Pseudomonas aeruginosa PAO1***), and `GCF_002899475.1` (***Escherichia coli DH5*α**).

ONT reads from each replicate were polished with Pychopper (v2.7.10), polyA tails longer than 10 bases and sequencing adapters were trimmed using cutadapt and mapped against the genome assemblies using Minimap2 (v2.29)(Li, 2018) and Samtools (v1.22) (Danecek et al., 2021). Candidate operons were identified from read alignments spanning at least two genes on the same strand and then extended by combining overlapping candidates at most 50 base pairs apart. Operons were then collated from the triplicates for each strain and used as our operon annotations.

Split. We evaluate the operon identification in a zero-shot manner, therefore, we do not split the data into train, validation and test splits and use the entire dataset as a test set.

A.3 Protein-protein interaction prediction

We downloaded all the data from the STRING DB download site (`https://string-db.org/cgi/download`). Using the species metadata file we selected only bacterial organisms and downloaded the protein sequences for them together with protein-protein interaction scores for protein pairs. After running the download scripts we ended up with $10,533$ unique strains across 6,956 species. We used the *combined* interaction score which combines information from various sources to get a final score. STRING DB provides only protein sequences and no DNA, and the interaction scores are computed mainly at the protein-level, therefore, for this dataset we only provide protein sequences and omit DNA. To binarize the interaction scores, we set the threshold at 0.6 ($\geq$ 0.6=interaction, ¡0.6=no interaction). This threshold was chosen through conducting small-scale experiments and looking at the average performance of AUROC and AUPRC on the validation set across genomes, choosing the threshold which attains the best average performance across the two.

Split. We performed a random data split into training, validation, and test sets in a 70 / 10 / 20 % ratio. The larger proportion of the train set compared to the gene essentiality task is motivated by the larger size of the overall dataset, with the 10% validation set still allowing for meaningful evaluation.

A.4 Strain clustering

To extract the metagenome assembled genomes (MAGs) for strain clustering, we use MGnify (Mitchell et al., 2023), which is a large-scale bacterial genomics database containing a diverse set of MAGs across numerous environments. The main reason for choosing MGnify over other potential resources is its large size combined with a uniform processing pipeline, providing comparable genomes. We wanted to evaluate whether various methods capture phylogenetic similarities across different taxonomic levels, therefore, we looked at strains which span different species, genera and families. These nested ranks provide three increasingly coarse resolutions to test whether an embedding preserves evolutionary signal. We use the taxonomic annotations provided by MGnify. The total number of unique genomes in the dataset is $6,071$.

To extract the genomes of interest, we extracted the most common species from MGnify from the corpus of 300k bacterial genomes and selected 25 species which are distributed across distinct genera and families. For a meaningful evaluation each genus (or family) must contain *at least two* species, otherwise the genus- or family-level clustering metrics become degenerate (every strain would be trivially assigned to a unique cluster at that rank). We download the chosen assemblies and use the accompanying annotations to translate gene DNA to protein sequences, while retaining the original DNA for the DNA-modality experiments.

Split. The strain clustering task is a fully unsupervised task, therefore, we do not split the data into train, validation and test splits and use the entire dataset as a test set.

A.5 Antibiotic resistance prediction

We leveraged the antibiotic sensitivity readings from the NCBI AST browser (`https://www.ncbi.nlm.nih.gov/pathogens/ast`), which contains hundreds of thousands of antibiotic-susceptibility test (AST) records for a diverse set of antibiotics. Using the genome assembly identifiers provided by the NCBI AST browser, we downloaded the DNA and protein sequences for each genome from the NCBI GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) and matched them with the antibiotic resistance readings. This left us with $26,052$ unique genomes with matched antibiotic resistance labels. We then processed the antibiotic resistance readings into *(i)* binary and *(ii)* regression labels decribed below.

Binary. For binary prediction, antibiotic sensitivity labels from NCBI AST browser of either *sensitive* (S) or *resistant* (R) were used for training and testing. If the antibiotic sensitivity test had no S/R label, they were not included. We remove antibiotics which 1) have less than 500 available unique genomes in total, and 2) have less than 50 unique genomes per class. This is motivated by the need to ensure that every classifier is trained on a sufficiently large and reasonably balanced data set; with fewer than 500 genomes overall, or fewer than 50 genomes in either class, the resulting model would suffer from poor statistical power and unreliable performance estimates. This resulted in 37 unique antibiotics. The Supp. Table 2 shows the number of available genomes per drug, including the number of susceptible and resistant genomes. The number of available readings varies strongly between antibiotics, which partly explains the high variance between the per drug performance.

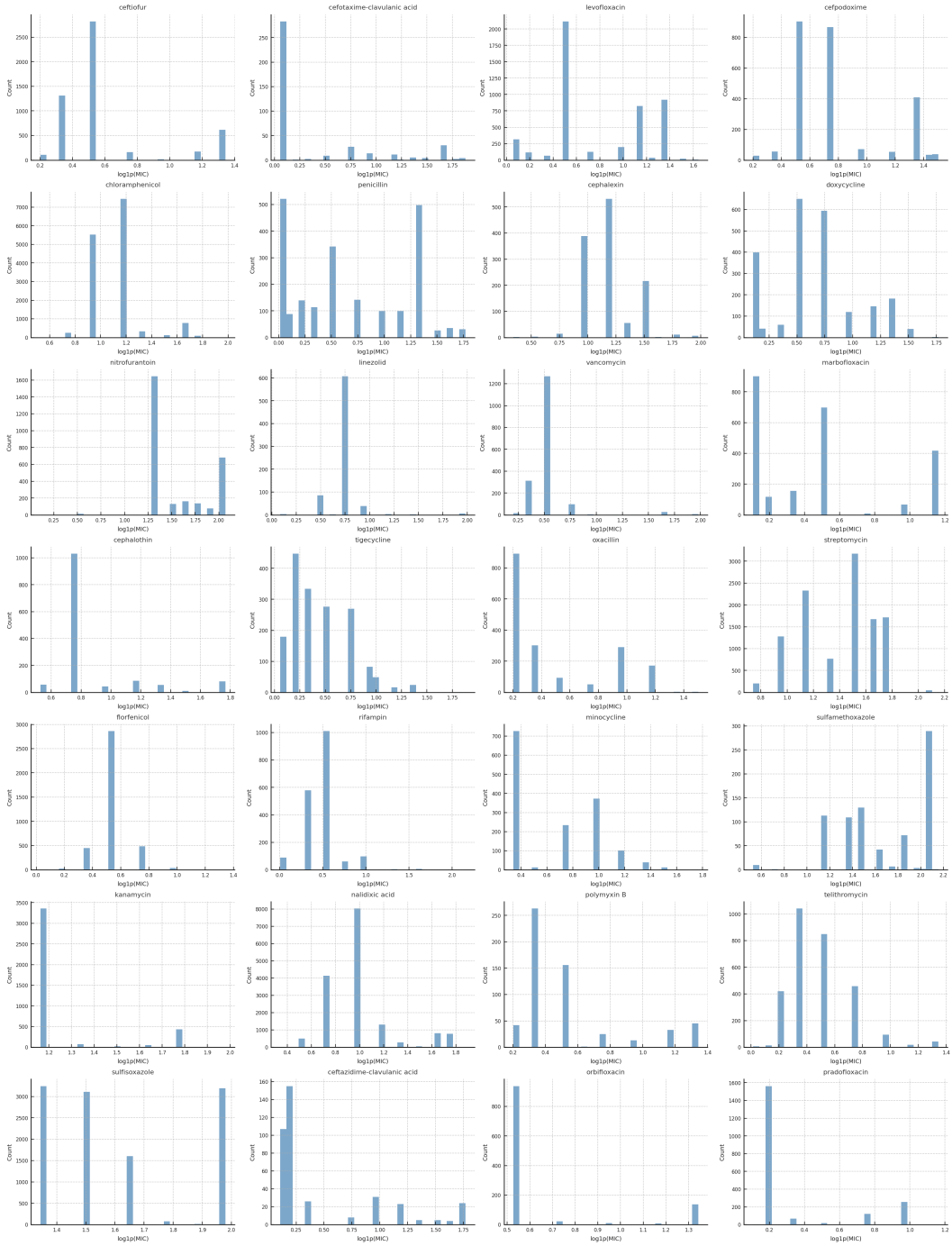
Regression. For regression MIC prediction, NCBI AST browser quotes minimum inhibitory concentrations (MICs) as $< x$, $\leq x$, $= x$, $\geq x$ and $> x$. These strings were translated to an actual number (y) by $y = x$ if NCBI quoted MIC=x, $\leq x$ or $\geq x$; $y = 2 \times x$ if MIC quoted as MIC $> x$, and $y = 0.5 \times x$ if MIC was $< x$. We filter out antibiotics with less than 500 available readings to ensure that the models are trained on a sufficiently large sample. This resulted in 56 antibiotics. To dampen the long tails, we normalized the MIC with a $log1p$ transformation. The final MIC distributions per antibiotic can be found in Supp. Fig 1 & 2.

Split. Due to low number of samples for many antibiotics and the variability between genomes, which may skew the results when using a single split, we trained and evaluated all models and antibiotics with k-fold split. Specifically, for each antibiotic we recommend: (1) splitting the available data into 5 equal splits using stratified split for the binary case and random split for regression. (2) In each split, further divide the larger train set into train and val, where validation makes up 20% of the train split. (3) Training the model on the train set and use the best performing model on the validation to evaluate the model on the test set.

Supplementary Table 2: Number of resistant, susceptible, and total labelled genomes for every antibiotic in the binary prediction setting.

Drug	# Resistant	# Susceptible	# Total
amikacin	7353	588	7941
ampicillin	10052	6211	16263
amoxicillin–clavulanic acid	12458	1683	14141
azithromycin	10497	2509	13006
cefazolin	1666	1897	3563
cefepime	3237	1188	4425
cefotaxime	566	446	1012
cefoxitin	12463	3902	16365
ceftazidime	2362	717	3079
ceftriaxone	13385	3469	16854
ciprofloxacin	17353	4280	21633
clindamycin	4327	337	4664
ertapenem	7708	373	8081
erythromycin	5605	1860	7465
fosfomycin	1186	396	1582
gentamicin	18242	2951	21193
imipenem	8485	409	8894
kanamycin	3768	410	4178
levofloxacin	3302	1106	4408
meropenem	11320	811	12131
nalidixic acid	1262	780	2042
nitrofurantoin	10535	5316	15851
oxacillin	6929	762	7691
piperacillin–tazobactam	3294	968	4262
rifampin	611	354	965
streptomycin	11965	2442	14407
sulfamethoxazole	3770	707	4477
sulfisoxazole	1569	323	1892
tetracycline	6936	3242	10178
tigecycline	662	357	1019
trimethoprim	2617	582	3199
ampicillin–sulbactam	1632	451	2083
aztreonam	2530	728	3258
ceftaroline	611	152	763
chloramphenicol	5345	1016	6361
colistin	572	290	862
daptomycin	492	131	623

Supplementary Figure 1: The $log1p$(MIC) distribution per antibiotic (1/2).



Supplementary Figure 2: The $log1p$(MIC) distribution per antibiotic (2/2).

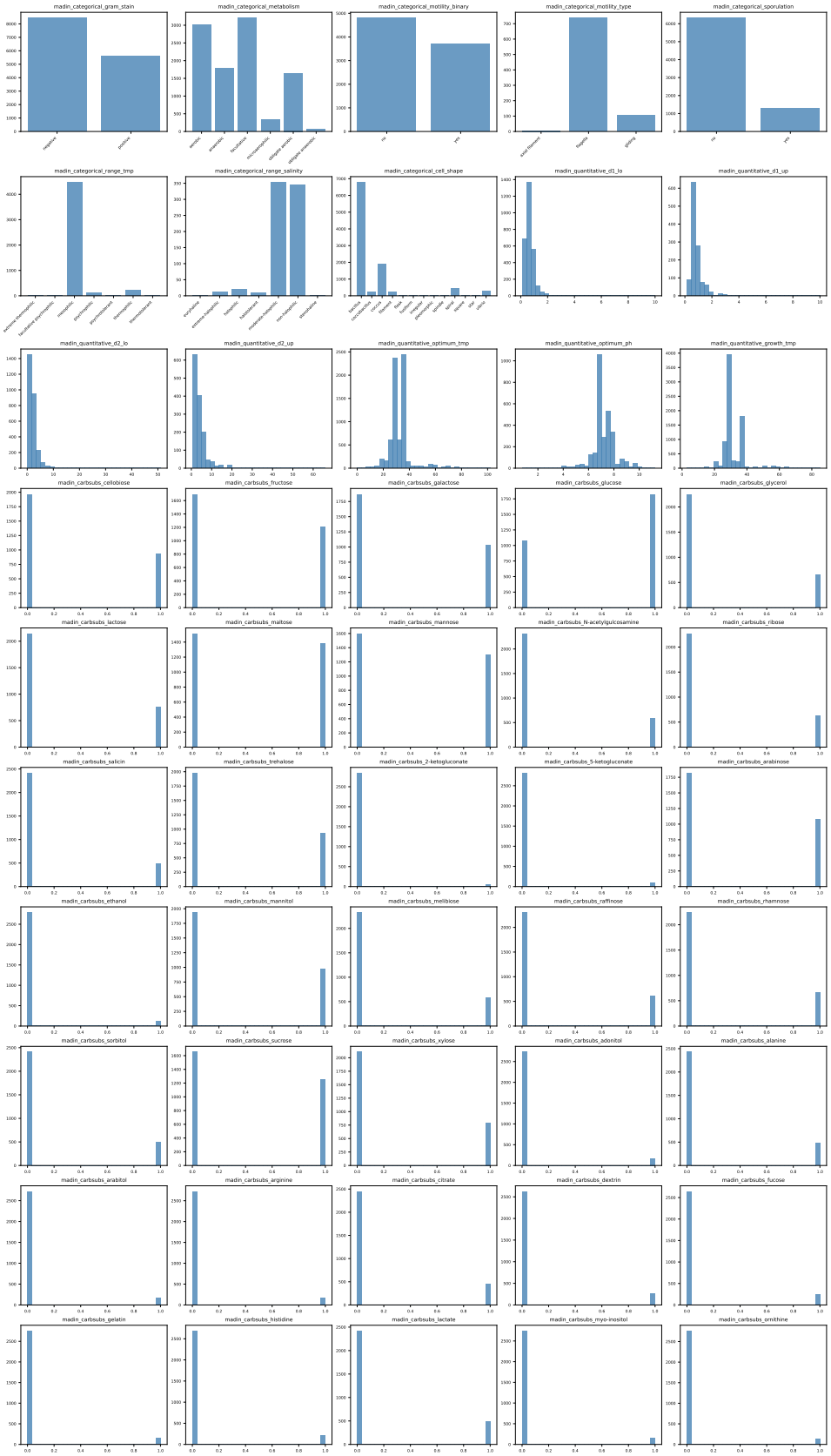
A.6 PHENOTYPIC TRAITS PREDICTION

We collected the phenotypic traits by collating two major sources (Madin et al., 2020; Weimann et al., 2016). To each phenotypic trait label we prepend its data source name so that the provenance of every label is explicit and any potential name collisions between the two catalogues are avoided. We keep duplicate traits that appear in both sources because they expand the number of labelled genomes without forcing us to merge measurements that were obtained with different experimental protocols. Using the taxonomy IDs and assembly accessions provided in the phenotypic traits datasets, we downloaded the associated genome DNA and protein sequences from the NCBI GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`). We limit ourselves to categorical phenotypes and filter out the phenotypic traits with less than 500 genomes to ensure that the models are trained and evaluated on a sufficiently large sample. Additionally, we removed the classes with less than 50 samples. This resulted in 139 unique phenotypes across $24,462$ genomes. We group the phenotypic traits into 5 distinct groups according to the type of biological information they capture (Supp. Fig. 3). We include the distributions of each phenotypic trait label in the Supp. Fig. 3-5 which shows large variation in the number of available labels per phenotypic trait which affects the final results.

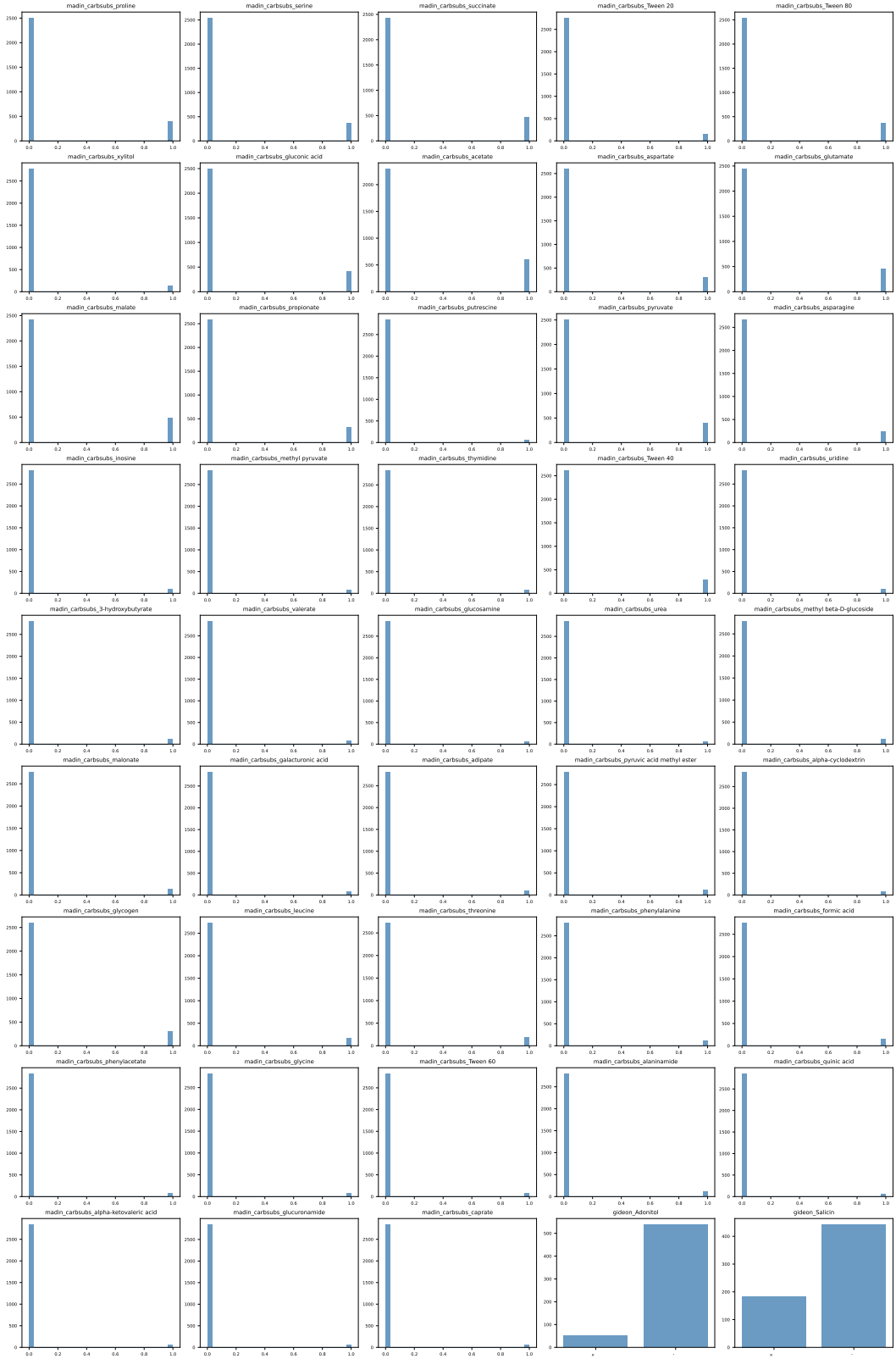
Split. For each phenotype, we split the data into $60/20/20$ train, validation and test partitions respectively. As similar genomes often share the same phenotype, we split the data for each phenotype by genus—placing all genomes from a genus in one split—and evaluate on held-out genera, enforcing generalization to phylogenetically distant strains. To obtain stable estimates, as many traits are rare, we aggregate the per-phenotype results across 5 independent runs with different data splits.

Supplementary Table 3: Phenotype groups and their associated phenotypes. The name before the first "_" symbolizes the dataset source. *madin* phenotypes were extracted from Madin et al. (2020) and *gideon* from Weimann et al. (2016).

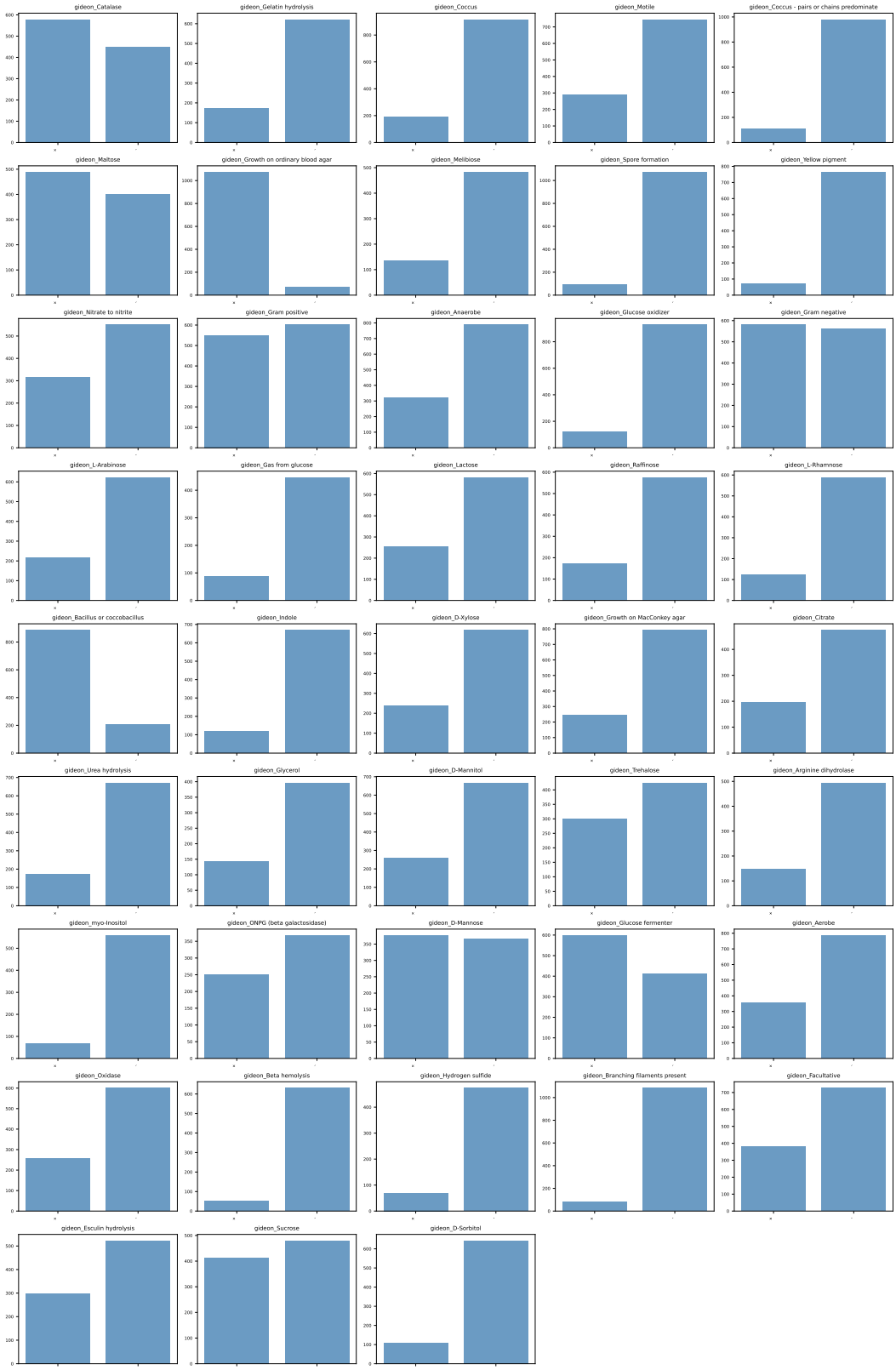
Phenotype group	Phenotype
Biochemical Activity	gideon_Gelatin hydrolysis, gideon_Indole, gideon_Urea hydrolysis, gideon_Methyl red, gideon_VP (Voges Proskauer), gideon_-Gal (beta-galactosidase), gideon_Beta hemolysis, gideon_Hydrogen sulfide, gideon_Esculin hydrolysis
Carbon Utilization	madin_carbsubs_cellobiose, madin_carbsubs_glucose, madin_carbsubs_glycerol, madin_carbsubs_lactose, madin_carbsubs_maltose, madin_carbsubs_mannitol, madin_carbsubs_sucrose, madin_carbsubs_xylose, gideon_D-Arabitol, gideon_D-Mannose, gideon_Sucrose, gideon_D-Sorbitol
Growth Conditions	madin_categorical_range_tmp, madin_categorical_range_pH, madin_quantitative_optimum_tmp, madin_quantitative_optimum_pH, madin_quantitative_optimum_O2, madin_quantitative_growth_rate, gideon_Growth on ordinary blood agar, gideon_Growth on MacConkey agar
Morphology	madin_categorical_gram_stain, madin_categorical_sporulation, madin_categorical_cell_shape, madin_quantitative_average_cell_size, gideon_Motility, gideon_Spores, gideon_Shape: bacillus or coccobacillus, gideon_Branching filaments present
Respiration Metabolism	madin_categorical_metabolism, gideon_Nitrate reduction, gideon_Alanine aminopeptidase, gideon_O/F glucose oxidizer, gideon_Gas from glucose, gideon_Glucose fermenter, gideon_Aerobe, gideon_Oxidase, gideon_Facultative



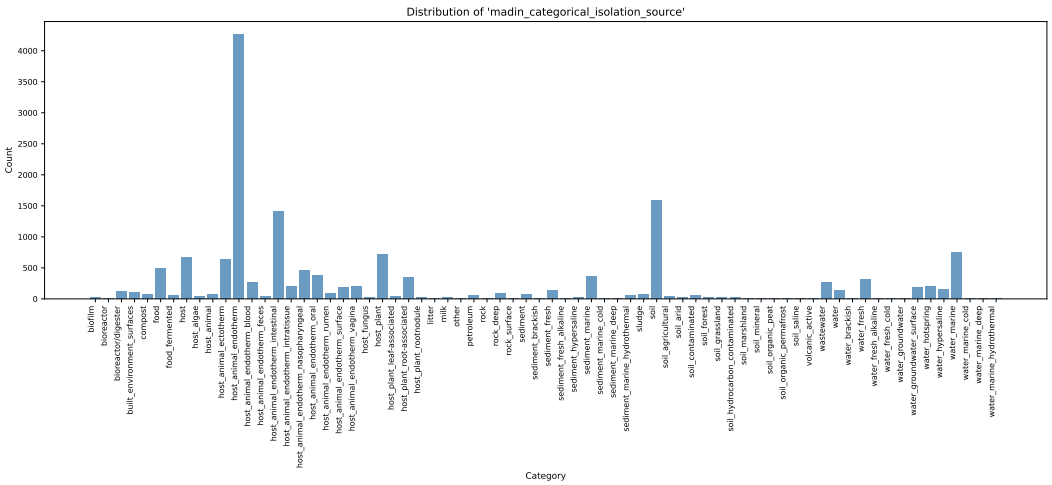
Supplementary Figure 3: The distribution of phenotypic traits labels (1/4).



Supplementary Figure 4: The distribution of phenotypic traits labels (2/4).



Supplementary Figure 5: The distribution of phenotypic traits labels (3/4).



Supplementary Figure 6: The distribution of phenotypic traits labels (4/4).

B Additional Modeling & Evaluation Details

We outline the experimental details used for modeling & evaluation for all model types; DNA Language Models (DNA LMs), protein Language Models (pLMs) and bacterial Language Models (bLMs). Implementation and further details can be found at `https://anonymous.4open.science/r/BacBench-B6EF`. All of the modeling code has been implemented in `PyTorch` (Paszke, 2019).

DNA LMs. For every model we load the public checkpoint (Supp. Table 4) and keep the hidden states of the *last* encoder layer. If a DNA string has length of G tokens, the encoder produces $\mathbf{X} \in \mathbb{R}^{G \times D}$; we collapse it with a simple mean-pool to obtain a vector

$$\mathbf{z} = \frac{1}{G} \sum_{g=1}^{G} \mathbf{X}_{g,\cdot} \in \mathbb{R}^D.$$

Gene- & system-level. When the input gene (plus upstream promoter) is longer than the model context C, we slide a window of length C with stride s and embed each window. Averaging the M window vectors $\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}\}$ gives the final gene representation $\bar{\mathbf{z}}_{\text{gene}} = \frac{1}{M} \sum_m \mathbf{z}^{(m)}$.

Genome-level. We treat the whole genome in the same way: split into C-bp windows, embed each, and average; if several contigs are present we first average per-contig and then across contigs.

pLMs. A protein of K amino acids yields $\mathbf{X} \in \mathbb{R}^{K \times D}$ and $\mathbf{z} = \frac{1}{K} \sum_k \mathbf{X}_{k,\cdot}$. For genome-level tasks we aggregate the M protein vectors in that genome via $\bar{\mathbf{z}}_{\text{prot}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{z}_i$. Gene- and operon-level tasks use only the proteins involved.

bLMs. **gLM2** ingests *mixed-modality* genomic scaffolds in which protein-coding regions are translated to amino-acid tokens and intergenic regions remain as nucleotide tokens. We tokenize each scaffold into a single sequence up to the model context C (4,096 tokens); longer scaffolds are processed with a sliding window of length C and stride s, mirroring the DNA LM setup. For each window, we retain the last-layer hidden states and mean-pool to obtain a window vector $\mathbf{z}^{(m)} \in \mathbb{R}^D$; genome-level embeddings average across windows (and across contigs when present):

$$\bar{\mathbf{z}}_{\text{genome}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{z}^{(m)}.$$

For gene essentiality and operon tasks, inputs include contextual flanking sequence to supply regulatory/positional cues, but the gene embedding is computed by averaging only the hidden states of tokens that fall within the gene's coordinates; empirically this has shown to outperform averaging the entire slice.

Bacformer model takes as input *local* protein vectors $\mathbf{z}_1, \ldots, \mathbf{z}_M$ obtained exactly as in the pLM setting. They are then **ordered by their genomic coordinates** (chromosome followed by plasmids) so that the model can "see" genome organisation. Rotary positional embeddings (Su et al., 2024) are added to these vectors and the ordered sequence is fed to a *genome-level* Transformer encoder (Vaswani et al., 2017) with L layers,

$$\mathbf{H}^{(0)} = [\mathbf{z}_1; \ldots; \mathbf{z}_M], \qquad \mathbf{H}^{(\ell+1)} = \text{Transformer}^{(\ell)}\big(\mathbf{H}^{(\ell)}\big),\ \ell = 0, \ldots, L-1.$$

The output $\mathbf{H}^{(L)} = [\tilde{\mathbf{z}}_1; \ldots; \tilde{\mathbf{z}}_M]$ contains **contextualised protein embeddings** $\tilde{\mathbf{z}}_i \in \mathbb{R}^D$ that encode both the protein sequence and its genomic neighbourhood. During pre-training Bacformer learns to predict which proteins co-evolved, so these embeddings capture functional coupling across the genome.

Empirically, averaging token embeddings performed slightly better than using the special `[CLS]` token, so mean pooling was used for every model throughout the paper.

B.1 General training details.

For all tasks and settings which required finetuning except gene essentiality prediction, we kept the frozen backbone encoder model frozen, and only finetuned the neural network layer(s) stacked on

Supplementary Table 4: Summary of benchmarked models. "Max ctx." = maximum context length supported at inference; for DNA models measured in base pairs, for pLMs in amino acids, and for Bacformer in number of proteins present in the genome. "dim" = dimensionsionality of the output of the last hidden layer. *We use either `bacformer-masked-complete-genomes` or `bacformer-masked-MAG` depending on the genome type used as input. The DNA LMs, pLMs and bLMs are separated by a horizontal line.

Model	Input	Variant / Checkpoint	Objective	Tokenisation	Params	dim	Training corpus	Max ctx.
Mistral-DNA (Mourad, 2025)	DNA	`Mistral-DNA-v1-138M-bacteria`	Autoregressive	Byte-pair	138M	768	Bacteria	512
DNABERT-2 (Zhou et al., 2023)	DNA	`DNABERT-2-117M`	Masked	Byte-pair	117M	768	Multi-kingdom	512
Nucleotide Transformer (Dalla-Torre et al., 2024)	DNA	`nucleotide-transformer-v2-250m-multi-species`	Masked	*k*-mer	250M	768	Multi-kingdom	2,048
ProkBERT (Ligeti et al., 2024)	DNA	`neuralbioinfo/prokbert-mini-long`	Masked	*k*-mer	27M	384	Bacteria	4,096
Evo (Nguyen et al., 2024)	DNA	`evo-1-8k-base (1.1_fix)`	Autoregressive	Single nucleotide	6.5B	4,096	Multi-kingdom	8,192
ESM-2 (Lin et al., 2022)	Single protein seq.	`esm2_t12_35M_UR50D`	Masked	Single amino acid	35M	480	Multi-kingdom	1,024
ESM-C (ESM Team, 2024)	Single protein seq.	`esmc_300m`	Masked	Single amino acid	300M	960	Multi-kingdom	1,024
ProtBERT (Elnaggar et al., 2021)	Single protein seq.	`prot_bert`	Masked	Single amino acid	420M	1,024	Multi-kingdom	1,024
gLM2 (Cornman et al., 2024)	Mixed modality (DNA & protein seq.)	`tattabio/gLM2_650M`	Masked	Single nucleotide/amino acid	650M	1,280	Bacteria	4,096
Bacformer (Wiatrak et al., 2025)	Multiple protein seq.	`macwiatrak/bacformer-masked-complete-genomes`*	Masked	Single protein	27M	480	Bacteria	6,000

top of the model encoder. This was motivated by the computational cost required to embed all 67k genomes with all models. Further details on runtime and computational cost can be found below. We used Adam optimizer (Kingma, 2014) in all finetuning setups. All of the checkpoints used have been downloaded directly from HuggingFace and are specified in Supp. Table 4. For each task and model combination, we tuned the learning rate keeping other parameters unchanged. Further details on hyperparameters used for each task and setup can be found in task-specific sections below.

Supplementary Table 5: Model-specific context parameters used in this study. "Max ctx." is the maximum input length; "DNA-seq overlap" is the stride between consecutive windows when sliding across a genome; "Promoter length" is the upstream sequence length concatenated for promoter prediction.* For Bacformer the maximum input size of a protein is $1,024$ amino acids and maximum number of proteins in the genome is set to $6,000$.

Model	Max ctx.	DNA-seq overlap	Promoter length
Mistral-DNA	512	16	128
DNABERT-2	512	16	128
Nucleotide Transformer	2,048	32	128
ProkBERT	4,096	64	128
Evo	8,192	32	128
ProtBERT	1,024	N/A	N/A
ESM-2	1,024	N/A	N/A
ESM-C	1,024	N/A	N/A
gLM2	4,096	64	128
Bacformer	1,024 / 6,000*	N/A	N/A

B.2 TASK DETAILS

We outline the experimental details for each task, outlining the hyperparameters used and evaluation setup. All of the experiments have been performed on a single NVIDIA A100 with 32 CPU cores.

Gene essentiality prediction. We stacked a single linear layer preceeded by a dropout of 0.2 and layer normalization (Ba et al., 2016) on top of the gene embeddings and trained it with binary cross-entropy loss to predict gene essentiality (1=essential, 0=non-essential). We trained the model for maximum of 100 epochs with early stopping patience of 10, monitoring the macro AUROC across genomes on the validation set. For each model we tuned the learning rate specified in Supp. Table 6. The weight decay for the Adam optimizer has been set to 0.02 for all models.

Evo. Evo natively does not provide straight-forward access to the output of the last hidden layer of the model. Therefore, we experimented with two ways of extracting the gene embeddings from Evo. Given a gene sequence G of size N, 1) we used the script provided as part of the Evo implementation (`https://github.com/evo-design/evo`) to score the log-likelihoods of the nucleotides in a sequence, resulting in a vector $z \in \mathbb{R}^N$, and 2) modified the Evo model to return the output of the last hidden state, resulting in a matrix $X \in \mathbb{R}^{N \times D}$, where D is model dimensionality which here equals $4,096$. We then similarly as with other models took the average of all sequence tokens resulting in a vector $x \in \mathbb{R}^{\mathbb{D}}$. We include this Evo implementation in the BacBench code repository (`https://anonymous.4open.science/r/BacBench-B6EF`). The option 1) yielded much better results on the validation set, therefore, we used it for final benchmarking.

Supplementary Table 6: Learning rates used for essential genes linear layer.

Method	Learning rate
Mistral-DNA	0.005
DNABERT-2	0.005
Nucleotide Transformer	0.005
ProkBERT	0.01
Evo	0.001
ProtBERT	0.005
ESM-2	0.005
ESM-C	0.005
gLM2	0.001
Bacformer	0.005

Operon identification. The operon–identification task is evaluated *zero-shot*; no fine-tuning is performed. We formulated operon prediction as a binary boundary classification problem, evaluating whether two contiguous genes form an operon. To address this, we developed a method that incorporates three features: (1) gene embeddings and (2) the strand of each gene.

First, we compute cosine similarity between adjacent genes using embeddings from the last hidden layer of pretrained models. We also record each gene's transcriptional strand from the genome assembly.

We then define a simple score that combines similarity and strand co-orientation:

$$s_i \;=\; c_i \, I_{\text{strand}}, \qquad c_i \;=\; \tfrac{1}{2}\big(1 + \cos(\hat{h}_i, \hat{h}_{i+1})\big),$$

where $I_{\text{strand}} = 1$ if both genes are on the same strand and 0 otherwise, and $\hat{h}_i, \hat{h}_{i+1}$ are the ℓ_2-normalised embeddings of genes i and $i{+}1$. The score $s_i \in [0, 1]$ serves as the operon-membership score; a pair is classified as belonging to the same operon when s_i exceeds a threshold. Mapping cosine similarity to $[0, 1]$ stabilizes the scale across models, while the strand indicator provides a hard veto since genes on opposite strands do not belong to the same operon.

Performance is computed per strain.

Evo. Evo natively does not provide straight-forward access to the output of the last hidden layer of the model. Therefore, we experimented with two ways of extracting the gene embeddings from Evo. Given a gene sequence G of size N, 1) we used the script provided as part of the Evo implementation (`https://github.com/evo-design/evo`) to score the log-likelihoods of the nucleotides in a sequence, resulting in a vector $z \in \mathbb{R}^N$, and 2) modified the Evo model to return the output of the last hidden state, resulting in a matrix $X \in \mathbb{R}^{N \times D}$, where D is model dimensionality which here equals $4,096$. We then similarly as with other models took the average of all sequence tokens resulting in a vector $x \in \mathbb{R}^{\mathbb{D}}$. We include this Evo implementation in the BacBench code repository (`https://anonymous.4open.science/r/BacBench-B6EF`). The option 2) yielded significantly better results on operon identification, therefore, we used it for final benchmarking.

Protein-protein interaction prediction. To predict whether the two proteins interact, we fed the two protein embeddings into a linear model, which is trained to predict a binary label (1=interaction, 0=no interaction). The linear model is a single-layer neural network preceeded by a dropout of 0.2 and layer normalization (Ba et al., 2016). The protein embeddings are fed into the linear classifier, after which the pairs of interacting proteins are averaged and passed through a final binary classification layer preceeded by a dropout of 0.2. The model is trained to minimize the binary cross-entropy loss. We experimented with different learning rates for all the models and set the final learning rate to 0.001, which has shown to perform the best for all the models. We trained the model for the maximum epochs of 10 and no early stopping patience. We set the maximum gradient norm to 2.0 and monitor the validation loss across genomes. The weight decay for the Adam optimizer is set to 0.01.

Strain clustering. To compute the strain-clustering metrics we run Leiden clustering (Traag et al., 2019) over a grid of parameters. We vary the *resolution* in `[0.1, 0.25, 0.5, 1.0]` and the

number of neighbours in `[5, 10, 15]`, evaluating every pairwise combination. Lower resolutions are omitted to avoid collapsing many genomes into a single (or just a few) giant clusters. After computing the clustering metrics for every parameter pair, we keep for each method the combination that maximises the mean performance across species-, genus- and family-level labels. The Leiden clustering is performed using the `scanpy` package (Wolf et al., 2018).

Antibiotic resistance prediction. To predict the antibiotic resistance, we train a linear model for each drug and method combination. The linear model is a single-layer neural network preceeded by a dropout of 0.2 and layer normalization (Ba et al., 2016). We train the models separately for the *(i)* binary and *(ii)* regression MIC prediction case. We optimize the former for the binary cross entropy loss and the latter for the mean squared error loss. We train all models for the maximum of 100 epochs with early stopping patience of 5, monitoring the validation AUPRC in the binary setup and validation R^2 in the regresssion setup. We experimented with various learning rates for each model, setting the final learning rate to 0.005 which attained the best results across folds and seeds for all models. We set the weight decay in the Adam optimizer to 0.01.

We have also experimented with training a multi-task linear model, which simultaneously predicts antibiotic resistance of a genome to multiple drugs, however, it performed worse then a separate linear model for each drug.

Phenotypic traits prediction. To predict a phenotypic trait from a genome-level embedding, we train an linear model for each phenotype and method combination. The linear model is a single-layer neural network preceeded by a dropout of 0.2 and layer normalization (Ba et al., 2016). As all labels are categorical, we optimize it to minimize the cross-entropy loss. We set the maximum number of epochs to $2,000$ and early stopping patience to 50, monitoring the validation loss. The learning rate for all models was set to 0.01. We use the cross-entropy loss. To account for the class imbalance, we weigh each class according to:

$$w_c = \frac{N}{K\, n_c}, \tag{1}$$

where n_c is the number of training samples in class c, K is the total number of classes, and N is the total number of training samples.

Runtime analysis. A typical bacterial genome contains on the order of $3{,}000{-}5{,}000$ genes and $\sim 4{-}6$ Mbp of DNA, with average protein lengths of ~ 300 amino acids; embedding an entire genome therefore entails thousands of forward passes for protein LMs and millions of tokens for DNA LMs, making raw throughput a practical bottleneck for population-scale studies. We measured the wall-clock time required to embed the *11* genomes used in the operon identification task on a single NVIDIA A100 and extrapolated to BacBench's full collection of $67{,}000$ genomes (Table 7). The fastest models are **ESM-2** and **Bacformer** (64–67 s for 11 genomes; ≈1.2–1.25k GPU-hours for 67k genomes). DNA LMs add a modest cost (e.g., 93 s for Mistral-DNA; 168 s for DNABERT-2) yet remain tractable on a single GPU. In stark contrast, the 6.5B-parameter **Evo** is $\sim 10^4\times$ slower (5.76×10^5 s for 11 genomes, i.e., ∼14 h per genome), yielding an impractical $\sim 1.07\times10^7$ GPU-hours for 67k genomes on one GPU. These measurements underline that—even when accuracy is the primary goal—runtime quickly becomes the limiting factor for population-scale analyses.

Beyond embedding costs, fine-tuning on genome-level tasks requires backpropagating through *all* proteins per genome (median ∼2,506 proteins across our datasets) or through multi-megabase DNA contexts, which dramatically amplifies memory and compute, requiring often ¿200 NVIDIA A100 GPUs for a single genome. Practically, this makes genome-scale fine-tuning out of reach for most academic labs for DNA LMs and pLMs, while remaining feasible for Bacformer-style bLMs; thus, linear-probe evaluations are a necessary, controlled proxy for model selection at scale.

B.3 Datasets & models licenses

To ensure that our datasets and benchmarks are reusable in the academic setting, we checked the license for each resource used in the manuscript. The Supp. Table 8 details a license for all models and datasets used in the manuscript.

Supplementary Table 7: Embedding runtime (s=seconds) on the operon identification task (11 genomes) on one NVIDIA A100 GPU and the extrapolated wall-clock time (h=hours) to process the entire collection of 67k bacterial genomes on the same hardware.

Model	Runtime (s)	Estimated time for 67 k genomes (h)
Mistral-DNA	93	1,731
DNABERT-2	168	3,127
Nucleotide Transformer	100	1,861
ProkBERT	106	1,973
Evo	575,629	10,712,540
ProtBERT	95	1,768
ESM-2	64	1,191
ESM-C	137	2,550
gLM2	182	3,387
Bacformer	67	1,247

Supplementary Table 8: External resources used in this study and their licences.

Resource	Type	Licence
ESM-2	Model	MIT
ESM-C	Model	Cambrian Open License Agreement
ProtBERT	Model	Academic Free License v. 3.0
gLM2	Model	Apache 2.0
Bacformer	Model	Apache 2.0
DNABERT-2	Model	Apache 2.0
ProkBERT	Model	MIT
Evo	Model	Apache 2.0
Nucleotide Transformer	Model	CC BY-NC-SA 4.0
Mistral-DNA	Model	Apache 2.0
MGnify	Data	CC0 1.0 Universal
NCBI AST Browser	Data	Public domain (U.S. Gov data)
NCBI GenBank/RefSeq	Data	Public domain (U.S. Gov data)
Database of Essential Genes	Data	CC BY 4.0
Operon DB	Data	CC BY 4.0
STRING DB	Data	CC BY 4.0

C Additional Results

We show and discuss further results across all tasks. To allow for comparability in future work, we include tables with all numerical results as well as per antibiotic and phenotypic traits scores.

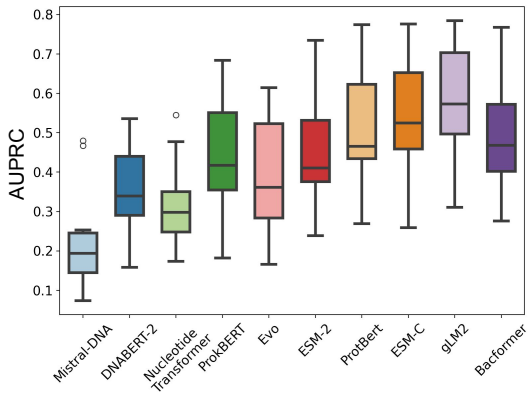
C.1 Gene essentiality prediction

The results on AUPRC show the bLMs and pLMs outperforming DNA LMs (Supp. Fig 7). gLM2 performs the best, showing the benefits of taking as input both DNA as well as protein sequence information. Moreover, increasing the pLM size appears to further boost performance, as demonstrated by ESM-C (300M) and ProtBERT (420M) outperforming ESM-2 (35M). Bacformer outperforms its protein representation backbone ESM-2, showing the benefits of incorporating whole-genome context. Evo outperforms other DNA LMs, except ProkBERT, demonstrating the performance gain by conducting pretraining on a relevant corpus. Finally the results on AUPRC show that there is large room for improvement. We believe that increasing the number and diversity of annotated genomes would significantly boost model performance.

The Supp. Table 9 shows the exact results across AUROC and AUPRC measured across disinct genomes.

Finetuning performance. To further analyse model performance, we have conducted finetuning on the gene essentiality task. The results show that finetuning boosts performance (Table 10); however, the model ranking remains largely unchanged, with the pLMs and bLMs outperforming DNA LMs. We have excluded Evo from finetuning due to the computational complexity required to finetune

Evo (Table 7). Gene essentiality is the only task for which we finetuned all models; genome-level tasks such as antibiotic-resistance prediction require end-to-end, whole-genome context and are prohibitively expensive for all models except Bacformer—underscoring the need for models that can be finetuned efficiently for whole-genome tasks.



Supplementary Figure 7: AUPRC across test genomes on the gene–essentiality prediction task. The box spans the inter-quartile range with a line marking the median value.

Supplementary Table 9: Overall performance on gene–essentiality prediction. Values are *mean* ± *standard deviation* over 3 random seeds; the best score for each metric is highlighted in bold.

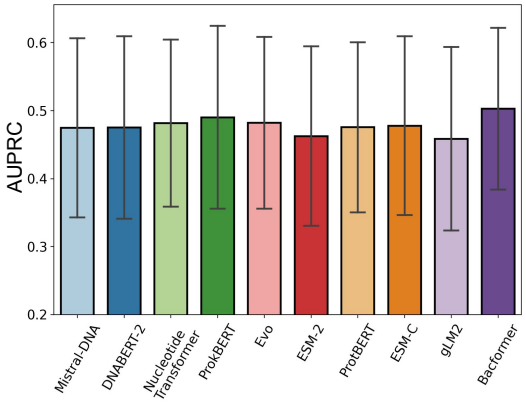
Method	AUROC	AUPRC
Mistral-DNA	57.52 ± 2.86	22.69 ± 14.21
DNABERT-2	68.21 ± 7.17	35.76 ± 12.46
Nucleotide Transformer	67.03 ± 5.56	31.78 ± 11.88
ProkBERT	74.79 ± 6.22	44.79 ± 15.25
Evo	73.71 ± 7.10	39.08 ± 15.21
ESM-2	77.99 ± 5.82	46.23 ± 15.87
ESM-C	82.25 ± 6.04	55.08 ± 15.83
ProtBERT	80.72 ± 5.85	52.00 ± 15.90
gLM2	**83.77 ± 6.17**	**58.61 ± 14.96**
Bacformer	80.72 ± 5.87	50.33 ± 15.43

Supplementary Table 10: Overall finetuning performance on gene–essentiality prediction. Values are *mean* ± *standard deviation* over 3 random seeds; the best score for each metric is highlighted in bold.

Method	AUROC	AUPRC
Mistral-DNA	62.18 ± 7.87	28.81 ± 13.51
DNABERT-2	74.88 ± 16.75	56.12 ± 29.71
Nucleotide Transformer	73.06 ± 14.99	48.56 ± 28.65
ProkBERT	69.88 ± 8.60	44.98 ± 14.36
ESM-2	85.36 ± 9.12	64.42 ± 17.79
ESM-C	89.96 ± 7.53	71.85 ± 14.49
ProtBERT	**91.31 ± 73.02**	73.02 ± 16.56
gLM2	90.12 ± 7.87	**74.03 ± 21.70**
Bacformer	90.31 ± 7.40	72.13 ± 15.16

C.2 Operon identification

The AUPRC follows the same trend as the AUROC described in the main manuscript: Bacformer attains the best scores, ProkBERT ranks second, and the remaining DNA LMs and pLMs form a tight cluster with broadly comparable performance—reinforcing that genome-level context and bacteria-specific pretraining matter more than modality alone; absolute AUPRC values are lower (as expected under class imbalance) but track the same model ordering.



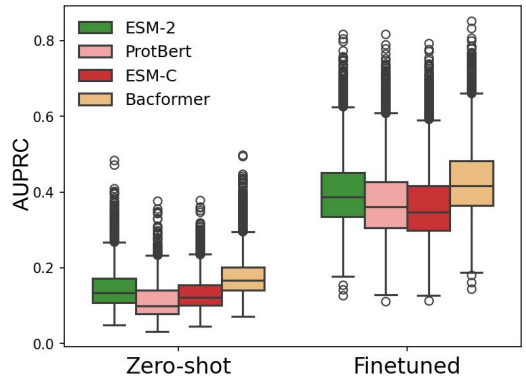
Supplementary Figure 8: AUPRC across test genomes on gene essentiality prediction task. The box spans the inter-quartile range with a line marking the median value.

Supplementary Table 11: Overall performance on operon identification. Values are *mean* ± *standard deviation*; the best score for each metric is highlighted in bold.

Method	**AUROC**	**AUPRC**
Mistral-DNA	74.16 ± 9.70	47.48 ± 29.47
DNABERT-2	74.13 ± 10.14	47.53 ± 30.07
Nucleotide Transformer	74.50 ± 8.57	48.19 ± 27.47
ProkBERT	75.77 ± 9.69	49.03 ± 30.07
Evo	74.77 ± 8.33	48.21 ± 28.27
ESM-2	73.02 ± 10.30	46.27 ± 29.55
ESM-C	75.25 ± 9.26	47.79 ± 29.39
ProtBERT	75.11 ± 8.65	47.58 ± 27.98
gLM2	72.85 ± 10.77	45.86 ± 30.18
Bacformer	**77.59 ± 6.64**	**50.31 ± 26.61**

C.3 Protein-protein interaction prediction

The protein-protein interaction (PPI) prediction results on AUPRC (Supp. Fig. 9) show that contextual pLM, Bacformer, consistently outperforms other methods. We credit it to its usage of the genomic context. Moreover, the performance does not increase by scaling the model size, as shown by ESM-C and ProtBERT underperforming ESM-2. The overall results (Supp. Table 12) show relatively low performance considering the trainins set size (¿7k genomes) even in the finetuned setting. We believe this is due to the 1) complexity of the task, 2) noisy source data, as STRING DB where the interactions have been extracted from collates information from a variety of sources and is not limited to experimentally validated interactions, highlighting the importance of building high quality PPI datasets.



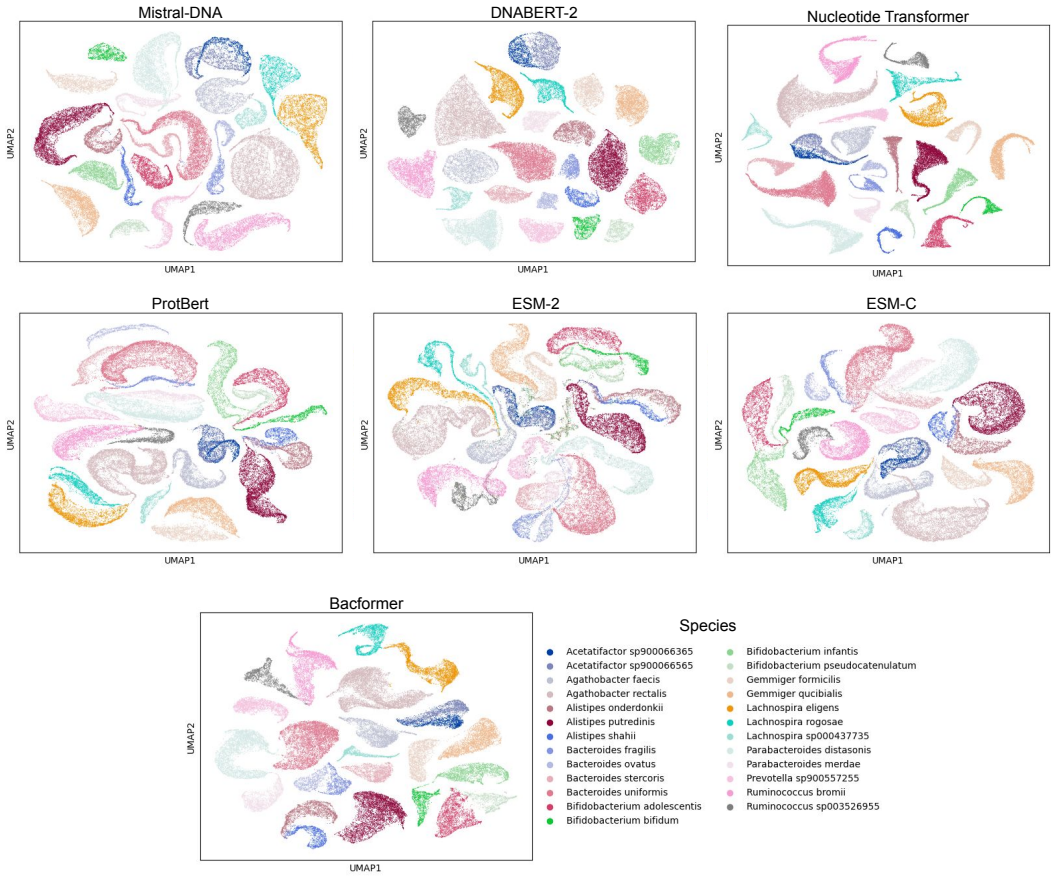
Supplementary Figure 9: AUPRC across test genomes on the protein–protein-interaction task in both zero-shot and fine-tuned settings. The box spans the inter-quartile range with a line marking the median.

Supplementary Table 12: Protein–protein-interaction prediction on the held-out genomes. Values are *mean* ± *standard deviation* over five seeds; the best score for each metric is shown in bold.

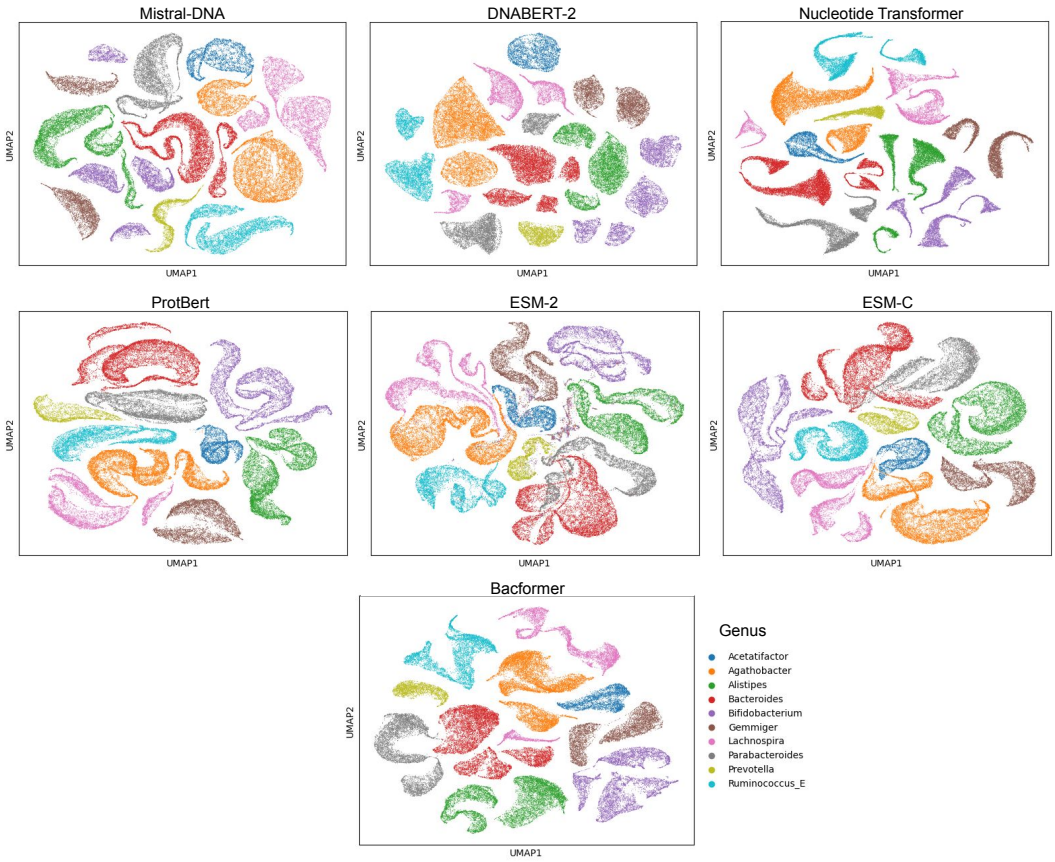
Method	**Zero-shot**		**Finetuned**	
	AUROC	AUPRC	AUROC	AUPRC
ESM-2	56.46 ± 2.10	14.94 ± 6.04	76.62 ± 2.35	40.68 ± 10.71
ProtBERT	47.20 ± 3.88	11.45 ± 4.89	74.05 ± 2.53	38.15 ± 11.16
ESM-C	55.17 ± 2.66	13.33 ± 4.61	74.21 ± 2.37	37.30 ± 10.98
Bacformer	**63.09 ± 2.73**	**18.20 ± 6.28**	**79.09 ± 2.25**	**43.47 ± 10.61**

C.4 STRAIN CLUSTERING

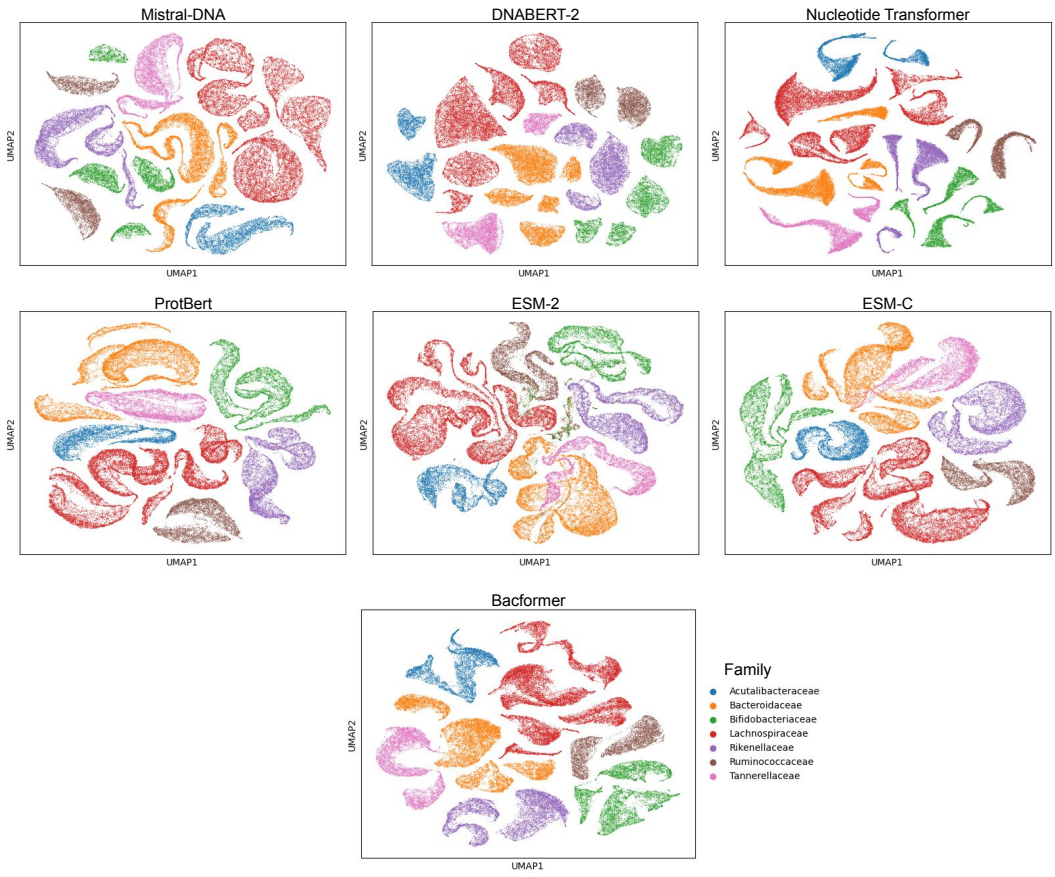
In addition to the strain clustering metrics included in the main manuscript, we plotted 2-dimensional UMAP results colored by species (Supp. Fig. 10), genus (Supp. Fig. 11) and family (Supp. Fig. 12) for a subset of models for further investigation. The UMAPs show that the strain representations differ between the models. All models cluster strains into separate species clusters, however, not all of them retain the phylogenetic similarities between species in the same genus or family. We also notice that the DNA LMs tend to output less overlapping species clusters, which boosts its performance at species level, but leads to lower results at higher taxonomic levels (genus and family).



Supplementary Figure 10: UMAP plots of strains (i.e. genome-level embeddings) across diverse methods colored by species.



Supplementary Figure 11: UMAP plots of strains (i.e. genome-level embeddings) across diverse colored by genus.



Supplementary Figure 12: UMAP plots of strains (i.e. genome-level embeddings) across diverse methods colored by family.

C.5 ANTIBIOTIC RESISTANCE PREDICTION

On the following pages we present the results per antibiotic across metrics. This includes binary prediction results (susceptible/resistant, Supp. Table 13) and MIC regression prediction results (Supp. Table 14). Each antibiotic was run across 5-folds with 3 random seeds to avoid variance stemming from random initialization and data-split bias. In the binary prediction setting the bLM-Bacformer outperforms other methods, achieving the best AUROC on 26 drugs, AUPRC on 27 drugs and F1 on 24 drugs out of 37 in total. Thus, showcasing the benefits of considering the interactions between proteins present in the genome. In the regression setting, Bacformer attains the best result on 44 drugs on R^2 and 45 on *Pearson* correlation coefficient out of total of 56 antibiotics. Finally, we see large variance across as well as within drugs. The former can be partly explained by the variable number of labels available per antibiotic, while the latter shows that, even within a single antibiotic, model performance can fluctuate markedly across folds and random seeds—highlighting sensitivity to sample composition and pointing to the need for larger, more balanced datasets or stronger regularisation to obtain stabler estimates.

Supplementary Table 13: Per-antibiotic binary antibiotic resistance prediction. Values are mean ± standard deviation across 3 seeds. Bold indicates the highest mean for each metric.

Antibiotic	Mistral-DNA		DNABERT-2		Nucleotide Transformer		ProkBERT		ESM-2		ESM-C		ProtBERT		gLM2		Bacformer	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
amikacin	90.24 ± 1.09	33.09 ± 1.79	88.46 ± 6.35	33.23 ± 9.22	90.77 ± 1.07	34.84 ± 3.83	90.47 ± 0.46	30.44 ± 0.30	91.62 ± 0.47	41.33 ± 0.75	91.94 ± 0.94	48.04 ± 3.46	91.70 ± 2.55	49.59 ± 6.31	88.94 ± 1.87	40.32 ± 2.48	**94.43 ± 1.12**	**59.00 ± 4.79**
amoxicillin-clavulanic acid	69.69 ± 1.80	28.06 ± 1.91	76.07 ± 1.53	37.21 ± 3.28	76.99 ± 2.30	41.06 ± 6.41	78.61 ± 0.50	42.47 ± 1.64	82.51 ± 1.63	48.70 ± 3.51	85.17 ± 1.03	54.77 ± 2.83	80.40 ± 1.07	48.71 ± 4.38	78.33 ± 1.48	42.78 ± 3.39	**85.86 ± 0.73**	**59.63 ± 1.54**
ampicillin	74.49 ± 1.42	68.27 ± 1.68	76.99 ± 1.32	69.81 ± 2.23	78.74 ± 0.94	72.00 ± 1.70	78.11 ± 0.47	71.77 ± 1.00	79.42 ± 0.24	73.05 ± 0.70	82.67 ± 0.59	77.14 ± 0.88	80.23 ± 0.47	74.55 ± 0.35	78.97 ± 0.67	71.77 ± 1.40	**82.72 ± 0.84**	**77.85 ± 0.98**
ampicillin-sulbactam	75.17 ± 4.69	77.76 ± 5.53	78.75 ± 4.65	79.29 ± 5.24	82.87 ± 3.82	82.53 ± 6.19	83.21 ± 4.42	81.04 ± 5.30	78.52 ± 4.68	81.63 ± 4.88	78.37 ± 3.48	81.83 ± 4.31	77.69 ± 6.97	81.15 ± 6.74	78.65 ± 3.30	78.09 ± 4.95	**86.72 ± 5.01**	**88.06 ± 5.39**
azithromycin	73.61 ± 7.84	3.57 ± 2.52	80.34 ± 3.02	7.31 ± 5.64	**86.53 ± 3.95**	9.15 ± 5.02	79.96 ± 6.85	4.62 ± 2.77	82.37 ± 6.40	8.08 ± 4.63	83.74 ± 5.03	7.91 ± 2.87	83.64 ± 2.58	7.30 ± 2.63	68.52 ± 12.37	6.76 ± 8.30	83.00 ± 3.73	**9.86 ± 5.80**
aztreonam	90.11 ± 0.68	90.05 ± 0.97	88.46 ± 1.67	89.37 ± 1.53	93.44 ± 0.90	93.24 ± 0.84	93.89 ± 1.14	93.53 ± 1.09	95.30 ± 0.39	94.25 ± 0.74	**95.76 ± 0.37**	**94.74 ± 0.56**	93.73 ± 0.89	93.40 ± 0.99	92.54 ± 0.65	92.28 ± 0.65	93.91 ± 1.21	93.09 ± 0.90
cefazolin	78.99 ± 1.58	83.03 ± 0.82	81.08 ± 1.32	84.62 ± 0.78	85.30 ± 2.61	89.63 ± 2.47	85.18 ± 2.72	90.00 ± 2.15	88.81 ± 3.86	91.90 ± 2.71	**92.23 ± 1.25**	**94.13 ± 1.11**	86.01 ± 0.51	90.29 ± 0.50	86.99 ± 0.26	90.97 ± 0.34	90.68 ± 1.24	93.61 ± 0.86
cefepime	71.17 ± 3.51	42.30 ± 3.73	71.83 ± 4.55	44.05 ± 4.25	75.30 ± 2.94	51.49 ± 6.06	76.21 ± 2.79	52.50 ± 2.25	79.50 ± 5.92	57.81 ± 12.81	82.08 ± 2.32	65.81 ± 4.89	78.94 ± 3.77	63.77 ± 5.14	76.11 ± 1.09	52.36 ± 2.19	**84.75 ± 2.00**	**68.73 ± 5.33**
cefotaxime	56.87 ± 1.73	79.94 ± 0.25	66.74 ± 3.36	89.85 ± 2.20	84.73 ± 3.33	95.39 ± 1.35	84.76 ± 4.79	94.30 ± 2.69	91.83 ± 2.39	97.54 ± 1.26	97.78 ± 1.05	99.38 ± 2.98	90.98 ± 2.98	97.36 ± 1.35	84.63 ± 2.64	95.96 ± 0.69	91.87 ± 2.58	**97.82 ± 0.89**
cefoxitin	76.41 ± 1.43	53.21 ± 0.34	80.23 ± 3.29	56.91 ± 4.44	82.94 ± 2.25	64.42 ± 2.09	82.10 ± 1.94	60.50 ± 1.73	85.36 ± 1.20	65.00 ± 1.26	87.56 ± 0.70	68.87 ± 1.68	84.79 ± 2.05	63.05 ± 5.57	81.91 ± 1.59	61.63 ± 1.07	**89.03 ± 0.97**	**72.49 ± 2.16**
ceftazidime	72.74 ± 2.12	60.18 ± 1.92	72.75 ± 1.58	60.63 ± 1.92	79.65 ± 0.63	69.71 ± 1.64	80.47 ± 0.61	71.54 ± 1.28	78.73 ± 5.68	68.29 ± 8.95	83.93 ± 0.26	75.92 ± 0.87	79.18 ± 1.04	73.29 ± 1.29	78.07 ± 2.83	66.86 ± 4.36	**86.26 ± 0.30**	**80.74 ± 0.92**
ceftiofur	74.81 ± 2.82	27.04 ± 3.63	81.13 ± 1.89	35.06 ± 5.20	86.11 ± 1.15	50.21 ± 4.64	84.02 ± 1.50	42.97 ± 3.83	81.77 ± 5.89	46.11 ± 10.96	86.39 ± 2.72	52.22 ± 6.91	81.33 ± 3.42	41.22 ± 9.21	77.05 ± 0.85	30.72 ± 1.76	**88.73 ± 1.10**	**58.91 ± 3.37**
ceftriaxone	74.60 ± 4.16	50.22 ± 3.72	74.00 ± 1.38	50.67 ± 2.37	85.10 ± 0.46	67.21 ± 0.77	82.75 ± 0.22	63.59 ± 0.74	87.59 ± 0.40	69.16 ± 1.52	**90.51 ± 0.26**	**75.08 ± 0.75**	81.72 ± 1.18	58.92 ± 1.22	85.49 ± 0.78	66.11 ± 0.97	88.98 ± 0.25	72.57 ± 0.65
chloramphenicol	72.18 ± 5.04	28.68 ± 4.05	75.99 ± 5.16	17.39 ± 1.97	83.29 ± 2.96	30.40 ± 9.12	83.25 ± 2.36	23.86 ± 5.20	84.62 ± 1.46	30.79 ± 3.68	86.07 ± 0.72	42.85 ± 1.50	81.35 ± 3.31	23.17 ± 6.42	77.03 ± 2.20	19.77 ± 4.45	**89.13 ± 1.81**	**49.39 ± 1.54**
ciprofloxacin	80.24 ± 0.39	54.97 ± 1.18	79.77 ± 1.15	54.18 ± 3.12	82.21 ± 0.33	59.35 ± 0.62	80.76 ± 0.42	58.02 ± 0.79	83.50 ± 0.33	60.83 ± 0.59	85.44 ± 0.24	64.16 ± 0.46	82.67 ± 0.24	59.71 ± 1.52	83.26 ± 0.44	58.98 ± 0.86	**86.09 ± 0.29**	**66.14 ± 1.16**
clindamycin	81.65 ± 2.09	22.58 ± 1.27	84.21 ± 1.56	28.39 ± 7.35	86.49 ± 0.60	38.99 ± 5.93	85.95 ± 1.57	35.94 ± 6.67	84.51 ± 1.80	33.87 ± 2.48	83.57 ± 1.45	29.92 ± 4.37	86.92 ± 1.80	35.10 ± 7.06	84.46 ± 2.21	28.21 ± 3.30	**89.88 ± 0.67**	**50.23 ± 7.67**
colistin	64.04 ± 8.15	17.71 ± 3.46	72.32 ± 6.37	19.23 ± 3.82	69.83 ± 15.10	21.46 ± 9.81	72.93 ± 5.68	28.63 ± 10.49	66.23 ± 8.95	19.00 ± 3.41	61.74 ± 6.24	14.10 ± 1.66	**75.62 ± 2.60**	**28.65 ± 2.15**	71.59 ± 4.39	19.46 ± 2.25	68.04 ± 7.94	21.82 ± 11.69
doripenem	71.13 ± 5.41	72.67 ± 4.89	75.90 ± 7.91	75.69 ± 9.47	78.05 ± 6.21	79.04 ± 7.76	77.53 ± 8.36	75.94 ± 12.42	82.87 ± 5.17	83.73 ± 6.94	83.77 ± 5.16	83.79 ± 5.69	**84.97 ± 7.22**	**85.05 ± 8.64**	78.02 ± 7.76	78.97 ± 7.51	82.77 ± 10.78	83.83 ± 10.40
ertapenem	82.31 ± 0.82	32.40 ± 1.20	84.36 ± 1.08	34.89 ± 1.16	90.39 ± 1.18	45.48 ± 1.12	89.47 ± 4.60	52.84 ± 16.34	85.85 ± 1.62	53.35 ± 8.86	86.21 ± 1.13	53.82 ± 3.27	**93.60 ± 2.55**	**75.01 ± 4.84**	90.10 ± 2.37	42.57 ± 6.62	93.02 ± 0.44	64.03 ± 11.14
erythromycin	87.71 ± 1.42	40.77 ± 4.66	90.44 ± 0.92	46.54 ± 3.00	90.88 ± 1.27	58.28 ± 4.82	91.30 ± 1.40	62.14 ± 7.46	89.90 ± 0.66	50.89 ± 6.47	89.42 ± 0.88	50.56 ± 3.86	91.50 ± 0.46	54.53 ± 7.56	89.92 ± 2.10	53.04 ± 6.21	**93.00 ± 0.46**	**63.23 ± 4.87**
gentamicin	82.55 ± 1.06	47.47 ± 2.30	81.20 ± 2.05	46.15 ± 4.86	86.73 ± 0.97	59.67 ± 3.95	85.32 ± 0.94	58.64 ± 3.00	86.51 ± 0.88	59.49 ± 1.97	87.52 ± 0.70	61.93 ± 1.60	85.36 ± 1.28	57.71 ± 1.84	85.81 ± 1.06	59.44 ± 2.87	**87.69 ± 0.94**	**65.02 ± 1.94**
imipenem	90.92 ± 0.62	64.56 ± 2.70	89.20 ± 2.02	61.91 ± 2.71	91.22 ± 0.78	68.93 ± 1.99	91.30 ± 0.46	70.77 ± 1.43	93.70 ± 0.82	79.74 ± 1.28	93.12 ± 0.74	78.37 ± 1.65	94.18 ± 0.13	79.33 ± 1.22	89.75 ± 3.03	64.90 ± 6.61	**94.53 ± 0.59**	**80.69 ± 2.68**
kanamycin	64.70 ± 1.04	18.18 ± 0.26	71.85 ± 2.21	23.02 ± 1.92	77.19 ± 4.45	29.35 ± 6.66	77.35 ± 4.34	27.77 ± 4.10	78.18 ± 0.75	28.80 ± 3.49	77.84 ± 0.70	29.09 ± 2.92	78.24 ± 2.41	28.41 ± 1.43	72.24 ± 1.98	22.27 ± 0.72	**81.34 ± 2.08**	**37.84 ± 2.78**
levofloxacin	68.37 ± 1.84	56.74 ± 1.30	73.88 ± 1.64	66.24 ± 1.12	79.21 ± 0.48	69.32 ± 1.70	78.34 ± 0.69	69.04 ± 1.25	82.73 ± 0.99	76.67 ± 1.34	**86.34 ± 0.45**	79.94 ± 0.90	77.27 ± 0.94	71.42 ± 1.55	79.97 ± 0.78	68.14 ± 1.05	86.08 ± 0.52	**81.91 ± 0.79**
meropenem	94.12 ± 0.88	47.47 ± 5.07	93.73 ± 0.70	46.05 ± 2.09	95.97 ± 0.26	57.97 ± 1.29	94.13 ± 1.14	47.88 ± 3.36	96.43 ± 0.82	69.00 ± 2.92	94.37 ± 1.31	57.75 ± 7.88	96.05 ± 1.20	66.99 ± 5.32	92.58 ± 1.49	48.79 ± 3.97	**96.92 ± 0.26**	**69.43 ± 0.88**
nalidixic acid	73.70 ± 0.67	22.59 ± 0.66	72.87 ± 0.91	21.57 ± 0.83	77.79 ± 1.16	27.59 ± 2.45	75.44 ± 0.59	23.98 ± 0.69	78.55 ± 2.27	30.73 ± 3.25	79.03 ± 0.71	39.48 ± 2.36	75.83 ± 1.91	24.77 ± 3.19	75.88 ± 1.51	21.49 ± 0.62	**84.75 ± 2.03**	**42.70 ± 4.50**
nitrofurantoin	90.48 ± 8.35	95.04 ± 3.95	95.00 ± 4.59	96.21 ± 3.82	97.65 ± 0.84	98.47 ± 0.42	**98.15 ± 0.43**	**98.69 ± 0.25**	97.31 ± 1.16	98.24 ± 0.74	97.22 ± 1.35	98.24 ± 0.83	96.54 ± 1.30	97.97 ± 0.77	96.97 ± 0.60	97.86 ± 0.46	97.24 ± 0.40	98.19 ± 0.25
oxacillin	76.85 ± 12.08	51.20 ± 10.97	85.42 ± 1.35	62.42 ± 3.28	85.46 ± 6.33	61.36 ± 9.53	**90.01 ± 5.05**	**71.74 ± 7.06**	84.98 ± 9.48	68.11 ± 16.77	80.37 ± 6.06	60.87 ± 9.43	81.64 ± 6.87	64.75 ± 9.49	78.32 ± 3.87	52.79 ± 5.67	85.68 ± 3.99	69.21 ± 5.75
penicillin	64.09 ± 25.33	83.31 ± 10.29	91.37 ± 2.23	94.66 ± 0.90	90.12 ± 0.54	93.33 ± 1.10	**94.31 ± 2.83**	**96.37 ± 2.02**	93.41 ± 0.64	95.93 ± 0.52	93.81 ± 0.63	96.31 ± 0.61	91.11 ± 0.36	93.57 ± 0.36	92.72 ± 0.91	95.26 ± 0.43	92.30 ± 0.88	94.90 ± 1.00
piperacillin-tazobactam	81.84 ± 0.34	47.15 ± 2.94	80.82 ± 0.65	46.56 ± 1.45	89.49 ± 1.28	68.76 ± 6.35	87.84 ± 1.12	64.93 ± 5.61	89.66 ± 0.46	71.28 ± 1.81	85.68 ± 3.92	59.66 ± 9.99	89.44 ± 1.14	71.98 ± 5.26	84.28 ± 0.72	52.25 ± 2.42	**90.44 ± 0.84**	**74.19 ± 4.78**
streptomycin	69.88 ± 0.98	60.00 ± 0.96	72.75 ± 2.39	57.81 ± 2.31	79.39 ± 1.32	65.04 ± 1.75	80.19 ± 0.57	68.35 ± 1.42	84.52 ± 1.05	75.78 ± 0.91	83.54 ± 0.79	73.32 ± 1.37	78.41 ± 0.44	67.28 ± 1.40	76.70 ± 1.30	63.86 ± 1.89	**84.77 ± 0.91**	**76.28 ± 1.28**
sulfamethoxazole	67.20 ± 1.91	57.04 ± 0.04	75.68 ± 3.25	**63.45 ± 2.46**	77.28 ± 3.16	62.64 ± 3.89	**77.65 ± 2.04**	63.08 ± 2.93	76.11 ± 1.62	59.78 ± 1.51	76.43 ± 2.27	61.79 ± 2.00	75.77 ± 2.28	61.74 ± 0.96	75.02 ± 1.85	59.01 ± 2.08	76.15 ± 0.63	61.17 ± 2.21
sulfisoxazole	70.34 ± 0.46	56.74 ± 0.94	80.33 ± 1.62	62.12 ± 4.19	88.55 ± 0.77	79.66 ± 0.63	87.00 ± 1.67	73.56 ± 3.08	90.68 ± 0.73	83.47 ± 1.31	**90.93 ± 0.50**	**84.73 ± 0.46**	87.77 ± 0.42	78.33 ± 1.46	81.49 ± 0.76	62.20 ± 1.61	88.19 ± 0.16	81.23 ± 0.66
telithromycin	73.42 ± 2.11	9.77 ± 4.96	82.10 ± 3.27	9.29 ± 5.93	82.09 ± 2.66	12.04 ± 3.07	81.87 ± 3.54	10.31 ± 3.10	79.66 ± 5.11	15.38 ± 7.77	76.61 ± 11.40	16.08 ± 3.25	82.80 ± 4.72	17.26 ± 5.73	79.80 ± 5.65	6.55 ± 2.37	**84.43 ± 5.78**	**25.38 ± 12.19**
tetracycline	59.30 ± 0.58	58.20 ± 0.67	64.08 ± 0.15	60.32 ± 0.87	71.53 ± 0.56	67.32 ± 0.63	71.60 ± 0.86	67.20 ± 0.95	75.71 ± 0.57	71.43 ± 0.51	80.07 ± 0.30	76.28 ± 0.25	74.41 ± 1.41	70.06 ± 1.61	74.96 ± 0.50	70.75 ± 0.47	**81.71 ± 0.34**	**79.07 ± 0.23**
tobramycin	82.21 ± 2.01	55.51 ± 4.28	83.01 ± 0.66	57.92 ± 0.35	87.93 ± 1.51	69.18 ± 5.10	86.83 ± 1.24	67.77 ± 3.40	86.69 ± 3.28	67.15 ± 8.62	86.84 ± 2.12	67.68 ± 6.40	87.73 ± 0.70	71.30 ± 3.07	84.06 ± 1.39	57.40 ± 3.22	**89.92 ± 0.60**	**76.53 ± 1.78**
trimethoprim-sulfamethoxazole	85.78 ± 0.58	47.47 ± 0.92	82.55 ± 0.28	46.81 ± 2.77	88.88 ± 0.33	64.36 ± 1.46	87.65 ± 0.51	62.72 ± 1.04	91.08 ± 0.53	69.30 ± 1.01	**91.61 ± 0.86**	**71.11 ± 2.02**	87.18 ± 0.07	59.38 ± 1.54	88.95 ± 1.00	63.99 ± 0.44	90.52 ± 0.57	70.13 ± 1.72

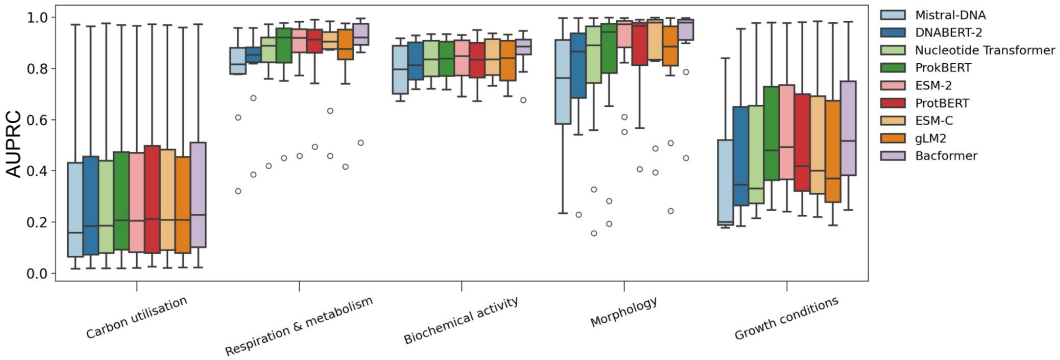
Supplementary Table 14: Per-antibiotic MIC regression prediction results across 3 seeds. Values are mean ± std. Bold indicates the best mean for each metric.

Antibiotic	Mistral-DNA		DNABERT-2		Nucleotide Transformer		ProkBERT		ESM-2		ESM-C		ProtBERT		gLM2		Bacformer	
	Pearson	R^2	Pearson	R^2	Pearson	R^2	Pearson	R^2	Pearson	R^2	Pearson	R^2	Pearson	R^2	Pearson	R^2	Pearson	R^2
amikacin	67.35 ± 0.91	45.08 ± 1.48	68.37 ± 1.03	46.63 ± 1.44	70.01 ± 1.02	48.82 ± 1.47	69.78 ± 0.74	48.66 ± 1.04	69.95 ± 1.08	48.63 ± 1.71	71.13 ± 0.99	**50.50 ± 1.47**	69.36 ± 0.85	48.01 ± 1.21	69.42 ± 1.67	48.00 ± 2.42	**71.32 ± 1.59**	50.46 ± 2.08
amoxicillin-clavulanic acid	46.27 ± 1.45	21.06 ± 1.03	48.59 ± 2.53	23.24 ± 2.06	51.50 ± 4.26	26.41 ± 4.21	52.04 ± 1.26	26.89 ± 1.14	53.13 ± 0.81	27.55 ± 0.93	56.14 ± 0.48	30.95 ± 0.30	54.47 ± 0.93	29.13 ± 1.12	51.14 ± 1.11	25.24 ± 1.01	**57.28 ± 0.47**	**32.09 ± 0.70**
ampicillin	46.43 ± 0.78	20.94 ± 0.81	51.01 ± 0.80	25.35 ± 1.03	52.72 ± 0.51	27.40 ± 0.87	51.78 ± 0.63	26.68 ± 0.69	54.19 ± 0.75	28.81 ± 0.92	57.00 ± 0.24	**31.94 ± 0.41**	52.20 ± 1.74	26.87 ± 1.74	51.62 ± 0.07	25.74 ± 0.48	**57.06 ± 1.09**	31.79 ± 1.36
ampicillin-sulbactam	40.77 ± 3.28	16.57 ± 2.74	41.23 ± 3.29	16.77 ± 2.37	45.36 ± 5.54	20.43 ± 4.64	42.14 ± 3.78	17.23 ± 2.85	45.49 ± 2.52	20.52 ± 2.23	46.60 ± 3.46	21.33 ± 2.96	42.40 ± 1.66	17.46 ± 1.20	47.21 ± 3.75	22.04 ± 3.45	**47.78 ± 7.02**	**22.09 ± 7.02**
azithromycin	79.15 ± 0.38	62.59 ± 0.66	79.57 ± 0.06	63.18 ± 0.10	80.01 ± 0.16	63.94 ± 0.21	79.74 ± 0.05	63.42 ± 0.09	79.63 ± 0.07	63.31 ± 0.15	**80.43 ± 0.58**	**64.50 ± 1.02**	79.97 ± 0.14	63.82 ± 0.24	79.73 ± 0.14	63.40 ± 0.28	80.27 ± 0.33	64.33 ± 0.70
aztreonam	72.00 ± 1.08	50.51 ± 2.22	72.36 ± 1.57	51.22 ± 2.83	74.70 ± 2.30	55.04 ± 3.92	74.48 ± 2.44	54.14 ± 4.96	77.48 ± 1.57	59.54 ± 2.81	78.07 ± 2.12	60.23 ± 3.90	77.07 ± 2.08	58.61 ± 3.02	74.41 ± 0.82	54.77 ± 1.17	**79.38 ± 1.96**	**62.61 ± 3.12**
cefazolin	65.50 ± 0.57	41.05 ± 1.29	66.18 ± 1.10	43.27 ± 1.77	69.40 ± 1.10	47.84 ± 1.98	68.16 ± 1.44	46.19 ± 2.16	71.38 ± 1.29	50.22 ± 2.12	**75.37 ± 1.85**	**56.05 ± 2.89**	67.57 ± 1.06	45.26 ± 1.74	70.19 ± 0.19	48.77 ± 0.96	73.18 ± 0.83	53.21 ± 1.30
cefepime	48.40 ± 2.99	23.05 ± 2.62	49.95 ± 2.10	24.28 ± 1.43	53.21 ± 2.92	28.21 ± 3.25	53.28 ± 1.58	28.14 ± 1.85	58.39 ± 3.10	33.82 ± 3.44	59.39 ± 1.85	34.51 ± 2.59	56.44 ± 3.72	31.56 ± 4.15	52.34 ± 0.59	27.11 ± 0.72	**60.53 ± 1.95**	**36.34 ± 2.49**
cefotaxime	15.28 ± 5.95	1.04 ± 0.84	33.60 ± 12.14	10.01 ± 6.23	51.74 ± 9.12	25.09 ± 9.05	49.80 ± 7.62	23.72 ± 8.13	57.33 ± 1.90	31.67 ± 2.40	58.41 ± 2.50	33.13 ± 3.77	54.12 ± 3.05	27.74 ± 5.29	46.18 ± 4.42	20.12 ± 5.28	**63.09 ± 1.48**	**38.04 ± 3.43**
cefovecin	9.42 ± 6.07	-2.14 ± 5.56	16.40 ± 8.74	0.97 ± 4.55	41.05 ± 2.51	12.55 ± 2.54	35.63 ± 19.51	9.99 ± 12.35	49.22 ± 0.72	21.17 ± 2.60	51.54 ± 1.83	24.29 ± 0.56	39.68 ± 6.29	13.92 ± 6.22	31.62 ± 9.08	6.77 ± 4.58	**52.60 ± 1.88**	**24.89 ± 3.18**
cefoxitin	54.85 ± 1.67	29.82 ± 1.66	60.37 ± 0.97	36.04 ± 0.86	63.07 ± 0.88	39.67 ± 1.15	62.02 ± 1.02	38.31 ± 1.18	62.75 ± 1.06	39.14 ± 1.13	65.50 ± 0.88	42.69 ± 1.05	63.16 ± 0.61	39.75 ± 0.69	61.36 ± 0.39	37.35 ± 0.50	**66.00 ± 1.27**	**43.21 ± 1.46**
cefpodoxime	32.75 ± 4.16	10.27 ± 2.16	32.72 ± 3.35	10.00 ± 1.46	40.53 ± 4.54	15.11 ± 2.73	42.37 ± 0.25	16.51 ± 0.90	47.28 ± 0.40	21.11 ± 0.21	48.70 ± 2.08	22.66 ± 1.90	45.07 ± 1.70	18.94 ± 0.33	30.86 ± 5.52	8.71 ± 3.37	**52.81 ± 1.99**	**26.55 ± 1.40**
ceftazidime	44.76 ± 0.70	19.70 ± 0.71	46.81 ± 0.74	21.63 ± 0.69	54.96 ± 1.97	29.66 ± 2.24	54.68 ± 1.50	29.38 ± 1.29	59.62 ± 1.22	33.83 ± 2.25	60.87 ± 0.62	36.49 ± 0.84	50.81 ± 2.20	25.48 ± 1.95	49.58 ± 1.50	24.17 ± 1.82	**61.88 ± 1.62**	**36.95 ± 2.55**
ceftiofur	4.56 ± 3.02	-0.02 ± 0.30	38.41 ± 1.93	9.77 ± 0.87	37.17 ± 9.23	11.15 ± 6.00	38.19 ± 2.72	12.09 ± 2.73	47.37 ± 0.13	20.69 ± 0.57	**48.21 ± 0.04**	**21.82 ± 0.64**	41.84 ± 1.81	16.04 ± 2.22	32.45 ± 3.89	9.84 ± 2.07	45.78 ± 2.60	19.01 ± 2.91
ceftriaxone	56.54 ± 0.78	31.78 ± 1.01	57.93 ± 1.22	33.37 ± 1.43	63.17 ± 1.83	39.66 ± 2.12	62.08 ± 1.43	38.37 ± 1.64	64.81 ± 1.96	41.65 ± 2.45	**69.00 ± 1.59**	**47.11 ± 2.05**	63.53 ± 1.58	39.97 ± 1.81	62.63 ± 1.41	38.71 ± 1.61	67.66 ± 1.71	45.43 ± 2.16
cephalexin	6.48 ± 2.58	0.05 ± 0.19	7.01 ± 4.62	0.24 ± 0.54	19.53 ± 6.34	3.33 ± 1.91	16.96 ± 4.27	2.25 ± 1.92	20.39 ± 6.07	3.91 ± 2.25	35.04 ± 5.97	9.38 ± 2.38	18.65 ± 1.38	2.94 ± 0.97	15.51 ± 13.44	2.38 ± 3.24	**35.36 ± 12.54**	**10.53 ± 7.83**
cephalothin	58.16 ± 8.99	30.88 ± 10.42	58.18 ± 8.67	30.84 ± 9.15	59.02 ± 8.53	31.58 ± 11.04	63.20 ± 8.52	37.95 ± 11.09	61.98 ± 8.07	37.00 ± 10.81	60.87 ± 8.15	35.29 ± 10.44	63.70 ± 10.05	36.95 ± 14.57	60.98 ± 7.69	35.71 ± 10.05	**66.14 ± 10.27**	**41.47 ± 13.32**
chloramphenicol	11.59 ± 0.68	1.27 ± 0.12	16.99 ± 1.50	2.82 ± 0.44	22.13 ± 2.63	4.62 ± 0.89	18.82 ± 2.22	3.34 ± 0.74	29.32 ± 2.20	7.60 ± 0.94	26.42 ± 5.41	6.74 ± 2.38	20.58 ± 4.21	4.07 ± 1.53	19.72 ± 4.39	3.64 ± 1.38	**34.47 ± 1.39**	**10.99 ± 0.56**
ciprofloxacin	57.09 ± 0.62	31.93 ± 0.61	57.20 ± 1.51	32.40 ± 1.58	58.82 ± 0.82	34.12 ± 1.36	58.32 ± 0.40	33.72 ± 0.76	59.10 ± 0.89	34.43 ± 0.89	59.13 ± 1.11	34.60 ± 1.06	58.53 ± 1.48	34.02 ± 1.55	56.44 ± 0.48	31.26 ± 0.33	**61.75 ± 0.67**	**37.71 ± 0.48**
clindamycin	55.41 ± 2.48	30.40 ± 2.62	60.77 ± 3.63	36.30 ± 4.38	63.33 ± 4.59	39.64 ± 6.13	64.95 ± 4.05	41.74 ± 5.79	66.84 ± 3.26	44.16 ± 4.41	67.99 ± 2.31	45.62 ± 3.18	62.26 ± 5.11	38.42 ± 6.50	62.99 ± 3.05	39.25 ± 3.91	**70.37 ± 2.85**	**49.06 ± 4.04**
colistin	22.28 ± 6.30	4.73 ± 2.59	26.95 ± 3.60	7.03 ± 1.72	41.61 ± 5.06	16.64 ± 3.73	40.77 ± 3.28	15.81 ± 1.90	32.42 ± 3.91	9.81 ± 1.90	43.88 ± 6.65	17.47 ± 4.24	45.48 ± 5.02	19.00 ± 4.77	32.54 ± 4.17	8.86 ± 1.53	**50.27 ± 7.21**	**22.54 ± 5.30**
daptomycin	39.30 ± 20.02	13.52 ± 13.63	40.87 ± 18.10	13.90 ± 11.44	46.90 ± 9.64	21.12 ± 10.00	52.22 ± 4.53	25.08 ± 4.64	43.66 ± 15.77	15.37 ± 18.13	46.33 ± 14.40	11.59 ± 30.18	43.99 ± 15.22	13.23 ± 22.82	**54.45 ± 4.38**	26.53 ± 5.14	53.99 ± 9.93	**28.48 ± 9.28**
doripenem	44.81 ± 7.03	18.72 ± 6.94	44.48 ± 7.11	19.01 ± 6.14	50.00 ± 4.35	23.16 ± 4.30	45.99 ± 4.07	20.23 ± 2.64	47.62 ± 13.69	22.04 ± 14.10	**60.01 ± 6.35**	**33.93 ± 6.73**	47.19 ± 7.18	18.55 ± 11.45	48.11 ± 5.74	22.84 ± 5.62	51.85 ± 3.34	25.74 ± 2.71
doxycycline	48.62 ± 2.95	23.17 ± 3.47	50.81 ± 2.43	25.53 ± 2.70	52.66 ± 1.16	27.19 ± 1.40	53.69 ± 1.54	28.20 ± 2.02	57.48 ± 1.18	32.22 ± 1.39	57.84 ± 1.50	32.58 ± 2.43	53.76 ± 1.33	28.59 ± 1.63	51.09 ± 2.47	26.02 ± 2.57	**58.22 ± 1.21**	**33.35 ± 1.46**
enrofloxacin	23.72 ± 2.70	5.16 ± 0.99	26.65 ± 1.42	6.82 ± 0.52	47.42 ± 2.82	20.99 ± 2.01	51.24 ± 2.65	23.83 ± 2.21	58.55 ± 1.40	32.80 ± 1.69	**62.78 ± 1.30**	**38.78 ± 1.71**	47.04 ± 3.52	20.45 ± 3.40	39.53 ± 1.28	14.29 ± 1.70	60.94 ± 3.94	35.63 ± 4.48
ertapenem	64.83 ± 1.13	40.73 ± 2.02	66.07 ± 0.89	42.60 ± 1.09	67.35 ± 0.75	44.64 ± 1.28	66.82 ± 0.48	43.91 ± 0.73	70.87 ± 0.87	49.68 ± 1.34	69.49 ± 1.48	47.94 ± 1.93	69.96 ± 1.36	48.70 ± 1.84	63.31 ± 1.49	39.23 ± 2.09	**71.99 ± 1.53**	**51.63 ± 2.18**
erythromycin	42.05 ± 1.24	16.50 ± 0.80	50.18 ± 1.28	23.97 ± 1.66	53.06 ± 0.29	26.90 ± 0.36	54.88 ± 2.60	28.60 ± 1.17	56.35 ± 2.57	29.53 ± 1.77	58.52 ± 2.39	32.57 ± 1.44	54.07 ± 2.06	27.51 ± 1.54	52.03 ± 2.22	25.13 ± 0.98	**59.05 ± 1.36**	**33.53 ± 0.57**
florfenicol	42.43 ± 1.86	13.63 ± 3.73	**46.00 ± 2.40**	19.00 ± 1.91	42.77 ± 8.03	17.75 ± 6.00	43.86 ± 9.39	18.89 ± 7.40	45.44 ± 9.07	20.27 ± 7.19	44.23 ± 10.07	19.38 ± 8.23	44.14 ± 7.30	19.10 ± 5.92	39.97 ± 8.24	16.05 ± 5.95	45.56 ± 9.11	**20.33 ± 8.26**
gentamicin	55.79 ± 0.61	30.47 ± 0.75	58.11 ± 0.79	33.46 ± 0.94	59.19 ± 1.02	34.76 ± 1.43	58.66 ± 0.67	34.25 ± 0.83	59.58 ± 0.22	35.19 ± 0.31	59.57 ± 1.23	35.36 ± 1.56	58.61 ± 0.60	34.24 ± 0.68	58.71 ± 1.17	34.30 ± 1.36	**60.23 ± 0.83**	**36.01 ± 0.92**
imipenem	70.39 ± 1.64	49.28 ± 2.42	70.87 ± 1.50	49.89 ± 2.37	71.28 ± 1.70	50.45 ± 2.60	71.27 ± 1.81	50.52 ± 2.70	71.83 ± 1.81	51.33 ± 2.78	71.61 ± 1.39	50.89 ± 1.85	72.22 ± 1.30	51.62 ± 1.96	71.16 ± 0.94	50.42 ± 1.38	**73.39 ± 1.88**	**53.62 ± 2.66**
kanamycin	0.75 ± 5.27	-0.19 ± 0.24	23.16 ± 2.69	4.62 ± 0.85	36.53 ± 2.17	12.96 ± 1.32	**36.88 ± 2.80**	**12.97 ± 1.59**	34.85 ± 4.49	11.67 ± 2.87	34.75 ± 4.24	11.61 ± 2.80	33.68 ± 2.35	10.99 ± 1.38	30.12 ± 2.79	8.57 ± 1.94	36.26 ± 3.37	12.59 ± 2.59
levofloxacin	28.22 ± 10.03	7.61 ± 4.71	42.69 ± 1.33	17.54 ± 0.91	48.47 ± 0.58	22.66 ± 0.47	44.63 ± 1.53	18.81 ± 0.32	53.69 ± 2.59	27.87 ± 2.97	**58.67 ± 2.39**	32.76 ± 2.13	49.48 ± 2.47	23.69 ± 2.09	50.55 ± 1.89	24.93 ± 1.49	58.59 ± 1.18	**33.35 ± 1.64**
linezolid	8.05 ± 15.49	**0.22 ± 2.85**	7.73 ± 14.70	-5.96 ± 10.47	7.37 ± 1.10	-16.82 ± 17.24	**16.07 ± 17.49**	-7.48 ± 20.22	10.06 ± 18.14	-5.13 ± 12.04	5.68 ± 10.41	-14.01 ± 26.52	3.85 ± 25.12	-21.28 ± 35.29	10.49 ± 15.02	-15.08 ± 25.90	2.86 ± 9.77	-34.65 ± 30.70
marbofloxacin	41.18 ± 6.12	15.70 ± 6.36	42.54 ± 4.39	16.75 ± 5.53	50.89 ± 2.76	24.93 ± 3.20	49.17 ± 10.47	23.46 ± 11.50	61.45 ± 4.60	35.53 ± 5.81	**65.92 ± 2.60**	**42.74 ± 4.01**	53.88 ± 4.92	28.06 ± 5.22	47.65 ± 9.41	21.83 ± 9.38	62.21 ± 5.45	37.09 ± 8.55
meropenem	73.25 ± 2.30	52.37 ± 2.49	74.71 ± 0.93	55.22 ± 0.88	76.01 ± 1.55	57.41 ± 2.41	75.94 ± 1.37	57.55 ± 2.04	75.50 ± 0.98	56.73 ± 1.26	75.93 ± 1.05	57.42 ± 1.49	76.12 ± 1.23	57.58 ± 1.75	73.45 ± 0.80	53.49 ± 1.07	**76.95 ± 1.48**	**58.78 ± 1.98**
minocycline	51.40 ± 1.91	25.51 ± 1.27	59.32 ± 0.72	33.72 ± 0.48	61.16 ± 2.64	36.28 ± 2.54	61.22 ± 3.01	37.25 ± 3.48	69.82 ± 1.51	47.41 ± 1.35	**70.19 ± 1.09**	**48.27 ± 2.01**	65.91 ± 1.33	42.53 ± 1.68	60.95 ± 1.28	36.28 ± 1.27	66.34 ± 4.26	43.54 ± 5.42
nalidixic acid	42.25 ± 2.47	17.63 ± 1.76	43.20 ± 2.09	18.45 ± 1.44	45.64 ± 2.06	20.62 ± 1.96	43.62 ± 2.13	19.01 ± 1.85	47.84 ± 1.24	22.59 ± 1.14	49.79 ± 1.83	24.44 ± 1.54	45.39 ± 1.63	20.08 ± 1.74	44.65 ± 2.80	19.76 ± 2.41	**54.10 ± 1.14**	**28.32 ± 1.20**
nitrofurantoin	91.48 ± 1.91	62.92 ± 2.32	94.36 ± 0.93	88.58 ± 1.68	94.72 ± 1.12	89.60 ± 2.22	95.05 ± 1.36	90.30 ± 2.51	94.93 ± 0.94	90.05 ± 1.78	94.87 ± 0.81	89.97 ± 1.50	95.10 ± 1.00	90.42 ± 1.88	94.76 ± 1.08	89.78 ± 2.02	**95.18 ± 1.05**	**90.55 ± 1.99**
orbifloxacin	6.61 ± 4.56	-4.38 ± 4.73	10.49 ± 5.31	-2.52 ± 3.35	10.79 ± 1.15	-5.91 ± 6.96	3.74 ± 1.34	-5.61 ± 5.73	14.22 ± 8.99	-4.05 ± 4.23	24.15 ± 9.09	-2.85 ± 13.33	8.81 ± 7.84	-5.34 ± 5.05	17.65 ± 9.42	-0.58 ± 3.66	**26.94 ± 13.87**	**4.06 ± 3.81**
oxacillin	39.98 ± 5.97	13.98 ± 6.21	54.68 ± 3.43	29.23 ± 3.52	56.74 ± 2.49	30.98 ± 2.18	63.44 ± 0.42	38.44 ± 0.91	64.06 ± 1.88	40.62 ± 2.06	66.14 ± 3.02	43.00 ± 3.14	62.72 ± 3.66	38.87 ± 4.20	58.80 ± 4.05	33.19 ± 3.74	**70.36 ± 2.44**	**48.92 ± 2.99**
penicillin	39.90 ± 1.98	15.08 ± 1.72	43.53 ± 2.53	18.16 ± 3.47	48.33 ± 2.03	22.67 ± 2.41	49.51 ± 3.17	23.54 ± 3.36	55.46 ± 3.46	29.58 ± 4.51	55.16 ± 5.06	29.68 ± 5.60	52.99 ± 2.74	27.62 ± 3.34	49.39 ± 2.46	23.52 ± 3.01	**60.02 ± 4.51**	**35.04 ± 5.81**
piperacillin-tazobactam	56.97 ± 0.92	32.01 ± 1.01	58.82 ± 0.85	34.18 ± 0.79	67.38 ± 2.28	44.81 ± 2.91	66.31 ± 1.20	43.06 ± 0.90	68.88 ± 2.18	46.45 ± 2.72	71.07 ± 1.08	49.57 ± 1.36	67.21 ± 0.99	44.58 ± 0.79	64.52 ± 0.58	41.01 ± 0.57	**72.30 ± 2.04**	**51.81 ± 2.56**
polymyxin B	23.37 ± 1.83	0.06 ± 8.91	29.67 ± 9.17	6.46 ± 8.12	43.30 ± 5.78	18.56 ± 4.72	46.37 ± 3.09	20.41 ± 2.04	46.17 ± 6.72	18.82 ± 8.05	44.44 ± 11.11	19.31 ± 10.42	**46.82 ± 5.68**	**21.02 ± 6.42**	31.96 ± 13.30	8.46 ± 9.67	41.83 ± 12.62	17.56 ± 9.45
pradofloxacin	13.71 ± 4.66	1.44 ± 0.97	14.28 ± 0.85	1.47 ± 0.25	28.36 ± 4.31	7.07 ± 3.19	35.05 ± 11.51	9.98 ± 5.43	48.88 ± 4.54	21.93 ± 3.73	**57.13 ± 3.18**	**30.88 ± 3.04**	35.33 ± 2.14	12.10 ± 1.33	33.54 ± 11.82	8.98 ± 4.66	51.13 ± 3.26	25.49 ± 2.82
rifampin	65.70 ± 10.55	38.90 ± 12.54	65.43 ± 11.25	42.08 ± 13.97	64.59 ± 7.60	39.63 ± 7.41	67.21 ± 8.07	44.45 ± 9.87	**68.89 ± 8.92**	46.59 ± 11.82	66.77 ± 7.60	43.27 ± 10.46	62.28 ± 5.46	37.65 ± 6.19	67.05 ± 8.26	44.57 ± 11.06	68.45 ± 6.02	**46.60 ± 8.24**
streptomycin	34.52 ± 2.22	11.36 ± 1.24	41.54 ± 2.45	16.58 ± 1.95	46.45 ± 0.99	21.34 ± 0.77	45.58 ± 1.35	20.27 ± 0.78	55.58 ± 1.49	29.25 ± 1.21	54.86 ± 2.49	28.39 ± 1.80	41.41 ± 5.15	16.96 ± 3.95	43.94 ± 1.97	19.09 ± 1.57	**57.69 ± 2.42**	**32.43 ± 2.43**
sulfamethoxazole	19.04 ± 3.72	-0.74 ± 1.05	11.66 ± 15.25	-2.67 ± 2.93	18.94 ± 2.23	2.40 ± 1.96	20.07 ± 2.99	3.08 ± 2.02	**28.34 ± 9.00**	5.68 ± 5.18	27.12 ± 5.70	**5.91 ± 2.39**	10.19 ± 2.35	-0.73 ± 1.21	24.86 ± 5.28	4.20 ± 3.41	23.63 ± 3.14	2.55 ± 2.83
sulfisoxazole	19.83 ± 0.36	3.56 ± 0.64	45.59 ± 2.22	19.43 ± 1.51	55.93 ± 3.70	30.50 ± 4.43	53.53 ± 1.14	27.34 ± 1.24	62.66 ± 0.37	38.26 ± 0.66	**65.77 ± 0.46**	**42.18 ± 0.92**	57.25 ± 1.24	31.15 ± 1.75	46.49 ± 1.57	20.29 ± 1.30	60.71 ± 1.40	35.84 ± 1.63
telithromycin	41.16 ± 2.74	12.00 ± 8.06	43.11 ± 2.49	16.37 ± 2.49	42.63 ± 3.04	17.63 ± 2.78	41.81 ± 2.89	16.78 ± 3.28	45.30 ± 3.50	19.37 ± 3.01	45.50 ± 2.78	19.87 ± 2.55	42.61 ± 3.40	17.77 ± 3.03	33.87 ± 1.34	11.00 ± 1.42	**45.80 ± 0.55**	**20.27 ± 0.62**
tetracycline	25.72 ± 0.94	6.50 ± 0.38	33.08 ± 1.60	10.85 ± 1.00	36.88 ± 2.22	13.39 ± 1.55	39.37 ± 1.38	15.15 ± 0.91	44.68 ± 1.24	19.60 ± 0.79	49.46 ± 1.65	23.74 ± 1.49	40.02 ± 7.56	16.03 ± 5.47	45.46 ± 1.50	20.04 ± 1.46	**52.35 ± 0.84**	**26.20 ± 1.24**
tigecycline	37.97 ± 3.94	13.42 ± 2.80	48.96 ± 4.84	23.52 ± 4.10	51.95 ± 4.02	26.61 ± 4.52	51.92 ± 3.85	26.81 ± 3.90	55.00 ± 3.40	29.33 ± 3.28	54.76 ± 4.72	29.39 ± 4.57	53.01 ± 4.70	27.83 ± 5.41	52.17 ± 2.08	26.74 ± 2.32	**59.21 ± 3.78**	**34.44 ± 4.14**
tobramycin	49.90 ± 1.32	24.02 ± 1.46	52.86 ± 2.62	27.33 ± 2.02	54.22 ± 3.28	28.83 ± 3.49	55.15 ± 4.08	29.88 ± 4.33	56.79 ± 3.17	31.09 ± 2.40	57.85 ± 3.87	32.07 ± 3.30	57.10 ± 4.72	32.12 ± 5.04	51.21 ± 1.17	25.42 ± 0.54	**59.92 ± 2.89**	**35.16 ± 2.48**
trimethoprim-sulfamethoxazole	58.87 ± 0.77	34.10 ± 0.77	61.77 ± 1.48	37.84 ± 1.71	63.94 ± 1.22	40.61 ± 1.38	64.00 ± 1.32	40.80 ± 1.66	66.66 ± 0.86	44.05 ± 1.08	**68.82 ± 0.78**	**46.89 ± 0.98**	65.03 ± 1.25	41.87 ± 1.53	63.69 ± 0.45	39.83 ± 0.44	68.02 ± 0.62	45.93 ± 0.96
vancomycin	14.32 ± 5.27	-0.57 ± 4.11	30.27 ± 7.49	8.39 ± 4.89	49.51 ± 4.69	23.33 ± 5.10	43.46 ± 10.95	17.68 ± 10.86	39.25 ± 8.09	13.18 ± 6.24	**51.69 ± 2.31**	**25.36 ± 2.71**	40.21 ± 10.64	12.70 ± 10.03	33.83 ± 9.94	10.08 ± 8.51	45.52 ± 8.99	18.05 ± 10.09

C.6 PHENOTYPIC TRAITS PREDICTION

We measure the AUPRC across phenotype groups (Supp. Fig. 13) as well as overall performance across metrics and all phenotypic traits (Supp. Table 15), and groups (Supp. Table 16). Finally, we include the results for each individual phenotype (Supp. Table 17 & 18). Analyzing performance across phenotype groups and metrics (Supp. Fig. 13 & Supp. Table 16), bLM-Bacformer achieves the best or combined best result across all groups and metrics, showing the benefits of incorporating genomic interactions into phenotype prediction models. Across metrics and phenotypes, Bacformer achieves the best result on 80 phenotypes on AUROC, 78 on AUPRC and 86 on F1 out of a total of 139 phenotypes. Therefore, showing that there is a considerable variation between phenotypes and one should choose a model specific for a phenotype. We believe this is due to the variable number of labels available for a phenotype as well as the inherent differences in the phenotypes themselves, which make some phenotypes easier to predict using a computational approach than others.

Predicting phenotypes is often a very challenging task which includes understanding the effect of mutations and multi-level genomic interactions. However, accurately predicting phenotypes could allow us to engineer genomes for a desired purpose, such as sustainable bioproduction, thus having potentially massive positive impact. We believe that next-generation models should consider the genomic context and incorporate the prior knowledge, such as genome-scale metabolic models to make the most out of available data for a given phenotype.



Supplementary Figure 13: AUPRC across diverse phenotypic traits groups and methods. The box spans the inter-quartile range with a line marking the median value. The results were macro-averaged across classes for each phenotype.

Supplementary Table 15: Overall phenotypic traits prediction performance across all phenotypes. Values are mean ± standard deviation across 5 seeds. The results were macro-averaged across classes for each phenotype.

Method	AUROC	AUPRC	F1
Mistral-DNA	60.34 ± 9.81	39.72 ± 33.09	47.25 ± 6.67
DNABERT-2	63.83 ± 11.06	42.10 ± 34.24	49.42 ± 8.99
Nucleotide Transformer	65.98 ± 11.45	43.22 ± 34.85	52.04 ± 11.07
ProkBERT	67.64 ± 11.87	44.80 ± 34.72	52.03 ± 11.56
ESM-2	68.71 ± 12.36	45.61 ± 35.45	53.10 ± 11.90
ESM-C	69.07 ± 11.98	45.54 ± 35.38	52.28 ± 11.39
ProtBERT	68.47 ± 11.97	45.47 ± 34.89	53.71 ± 11.65
gLM2	65.82 ± 11.41	44.00 ± 34.40	51.11 ± 10.86
Bacformer	**71.34 ± 12.91**	**47.80 ± 35.99**	**56.14 ± 13.19**

Supplementary Table 16: Phenotype-group prediction results; *mean* ± *standard deviation* over five random seeds. The highest score in each column is shown in bold. Results are macro-averaged across classes for each phenotype.

Method	Biochemical activity			Carbon utilisation			Growth conditions			Morphology			Respiration & metabolism		
	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1
Mistral-DNA	54.67 ± 12.73	79.60 ± 10.06	44.48 ± 2.40	58.17 ± 7.25	27.95 ± 27.92	46.67 ± 3.25	63.73 ± 4.34	40.58 ± 37.70	36.80 ± 17.74	70.07 ± 11.96	72.05 ± 23.63	52.81 ± 9.92	71.79 ± 11.54	77.30 ± 18.47	50.88 ± 14.54
DNABERT-2	59.88 ± 9.16	82.39 ± 8.55	46.19 ± 6.05	60.46 ± 7.60	29.24 ± 28.29	47.63 ± 3.90	71.22 ± 17.07	49.41 ± 40.69	45.89 ± 27.95	79.15 ± 12.39	79.75 ± 22.57	56.83 ± 14.23	76.88 ± 11.06	80.97 ± 16.72	60.52 ± 15.97
Nucleotide Transformer	63.74 ± 8.59	83.40 ± 8.47	50.62 ± 4.88	62.07 ± 7.58	30.06 ± 28.72	49.20 ± 4.03	79.33 ± 13.37	50.82 ± 41.12	51.73 ± 33.88	79.63 ± 14.10	78.94 ± 26.27	61.55 ± 20.66	81.74 ± 10.26	83.90 ± 16.08	66.91 ± 14.99
ProkBERT	61.84 ± 8.89	83.41 ± 8.29	49.52 ± 5.28	63.50 ± 7.72	31.50 ± 28.50	49.34 ± 4.49	86.10 ± 8.33	56.82 ± 37.42	49.07 ± 29.41	82.76 ± 12.95	81.29 ± 26.40	61.87 ± 23.45	84.91 ± 8.16	85.55 ± 16.14	66.74 ± 14.85
ESM-2	62.86 ± 12.75	83.30 ± 9.29	51.30 ± 9.54	64.60 ± 8.18	31.77 ± 29.34	49.38 ± 4.58	81.70 ± 11.20	56.97 ± 37.48	52.70 ± 29.63	87.97 ± 10.47	89.36 ± 14.86	**71.41 ± 18.24**	83.46 ± 11.48	86.63 ± 15.62	66.11 ± 16.31
ESM-C	62.71 ± 11.90	83.90 ± 8.16	50.82 ± 9.24	65.37 ± 8.16	32.20 ± 29.54	49.27 ± 4.24	78.71 ± 15.27	53.44 ± 40.05	55.72 ± 35.81	88.05 ± 9.67	87.26 ± 20.11	67.68 ± 20.43	81.79 ± 12.13	85.05 ± 16.95	61.05 ± 14.90
ProtBERT	61.93 ± 13.25	82.30 ± 10.18	50.78 ± 7.31	64.86 ± 8.29	32.22 ± 29.14	50.30 ± 5.05	82.07 ± 11.45	54.06 ± 39.32	54.87 ± 33.97	83.27 ± 13.36	86.10 ± 18.70	68.60 ± 18.70	84.71 ± 8.40	86.48 ± 14.80	68.69 ± 14.17
gLM2	60.39 ± 8.39	82.71 ± 9.45	47.76 ± 4.82	62.78 ± 7.22	30.97 ± 28.27	48.65 ± 4.06	71.55 ± 26.57	51.10 ± 41.36	45.49 ± 27.39	77.44 ± 16.90	82.79 ± 21.88	62.18 ± 20.77	82.26 ± 9.24	84.22 ± 16.68	64.43 ± 16.91
Bacformer	**67.42 ± 16.45**	**86.32 ± 9.03**	**55.09 ± 11.26**	**66.94 ± 8.94**	**33.94 ± 30.24**	**51.99 ± 6.44**	**86.18 ± 9.85**	**58.25 ± 37.22**	**59.53 ± 33.08**	**89.53 ± 10.01**	**91.13 ± 15.07**	70.20 ± 19.26	**88.18 ± 9.31**	**89.19 ± 14.12**	**77.15 ± 15.49**

Supplementary Table 17: Per-phenotype performance (1/2). Values are mean ± standard deviation across 5 random seeds. Bold highlights the best mean per metric. The results were macro-averaged across classes for each phenotype.

Phenotype	Mistral-DNA			DNABERT-2			Nucleotide Transformer			ProkBERT			ESM-2			ESM-C			ProtBERT			gLM2			Bacformer		
	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1
gideon_Adonitol	78.65 ± 8.07	97.09 ± 3.12	47.84 ± 1.49	79.77 ± 9.66	96.42 ± 4.56	47.77 ± 1.40	**83.76 ± 9.10**	**97.56 ± 2.88**	51.74 ± 8.04	81.54 ± 8.40	96.96 ± 3.64	51.76 ± 7.06	77.70 ± 8.09	96.64 ± 3.56	47.84 ± 1.49	78.03 ± 9.47	97.03 ± 3.26	47.84 ± 1.49	79.61 ± 6.84	97.24 ± 3.01	47.77 ± 1.40	77.22 ± 5.27	95.94 ± 4.87	47.84 ± 1.49	80.67 ± 9.67	97.22 ± 3.10	**61.31 ± 6.11**
gideon_Aerobe	76.51 ± 6.25	78.58 ± 8.18	62.95 ± 25.02	85.67 ± 3.39	88.52 ± 4.93	62.40 ± 19.50	88.49 ± 1.52	90.68 ± 3.07	74.74 ± 4.43	92.43 ± 2.46	94.96 ± 1.72	82.49 ± 4.69	93.05 ± 1.45	95.67 ± 0.59	87.14 ± 2.79	90.71 ± 2.42	94.24 ± 1.75	65.03 ± 17.82	93.49 ± 2.35	96.07 ± 1.14	72.43 ± 15.88	91.92 ± 2.58	94.70 ± 1.97	79.99 ± 4.48	**96.51 ± 0.82**	**98.23 ± 0.31**	**92.05 ± 1.21**
gideon_Anaerobe	77.21 ± 2.99	88.91 ± 1.84	46.30 ± 4.48	88.84 ± 4.07	95.08 ± 2.20	83.13 ± 2.79	90.90 ± 1.02	95.39 ± 1.61	81.25 ± 7.24	92.50 ± 1.52	96.59 ± 0.90	81.41 ± 1.93	96.14 ± 1.34	98.28 ± 0.68	86.50 ± 2.68	96.57 ± 2.93	98.40 ± 1.16	76.33 ± 30.20	97.92 ± 1.07	99.08 ± 0.54	89.29 ± 2.14	91.08 ± 1.84	96.37 ± 0.67	78.24 ± 3.99	**98.92 ± 0.37**	**99.53 ± 0.19**	**95.64 ± 1.76**
gideon_Arginine dihydrolase	35.07 ± 9.42	69.15 ± 4.22	42.87 ± 1.67	49.40 ± 22.04	77.45 ± 7.76	42.87 ± 1.67	49.88 ± 5.19	77.20 ± 10.93	**43.84 ± 2.86**	50.30 ± 19.51	77.40 ± 15.50	42.16 ± 1.56	39.06 ± 8.47	72.64 ± 9.78	42.87 ± 1.67	42.30 ± 17.13	73.25 ± 8.22	42.81 ± 1.62	35.20 ± 13.64	67.28 ± 5.88	42.87 ± 1.67	**52.26 ± 19.08**	**78.49 ± 14.18**	42.87 ± 1.67	34.86 ± 7.48	67.62 ± 4.60	41.99 ± 1.53
gideon_Bacillus or coccobacillus	63.77 ± 12.86	23.31 ± 12.78	46.05 ± 1.44	66.65 ± 10.66	22.91 ± 11.29	47.58 ± 2.77	67.89 ± 16.70	32.81 ± 22.25	56.35 ± 11.94	70.48 ± 11.47	28.20 ± 15.40	51.87 ± 4.43	75.66 ± 4.68	**55.22 ± 7.09**	65.56 ± 5.11	74.54 ± 8.99	39.42 ± 16.19	66.83 ± 7.36	74.26 ± 14.65	40.67 ± 26.37	62.78 ± 16.37	65.71 ± 8.67	24.28 ± 12.30	48.21 ± 3.65	**78.62 ± 6.11**	45.10 ± 21.01	**70.62 ± 8.52**
gideon_Beta hemolysis	**63.40 ± 15.20**	91.72 ± 5.40	47.08 ± 0.64	62.26 ± 16.78	**92.21 ± 5.97**	47.08 ± 0.64	60.54 ± 16.54	90.78 ± 6.14	47.08 ± 0.64	53.06 ± 5.76	88.96 ± 3.05	47.08 ± 0.64	54.80 ± 10.50	88.61 ± 5.10	47.08 ± 0.64	54.08 ± 6.86	90.87 ± 3.83	47.08 ± 0.64	53.71 ± 12.93	89.32 ± 5.24	**49.23 ± 3.82**	50.10 ± 5.57	89.60 ± 1.81	47.08 ± 0.64	52.05 ± 8.24	88.81 ± 5.26	47.08 ± 0.64
gideon_Branching filaments present	95.93 ± 2.98	99.76 ± 0.31	57.91 ± 17.80	96.53 ± 2.17	99.78 ± 0.30	56.70 ± 15.71	96.34 ± 2.17	99.65 ± 0.55	54.28 ± 11.56	96.88 ± 1.63	99.80 ± 0.27	53.06 ± 9.48	98.23 ± 1.75	99.70 ± 0.50	**63.26 ± 27.02**	**98.66 ± 1.33**	**99.82 ± 0.31**	56.73 ± 15.76	87.65 ± 17.41	97.92 ± 3.55	56.73 ± 15.76	96.99 ± 0.65	99.78 ± 0.32	48.37 ± 2.30	96.69 ± 0.82	99.75 ± 0.36	48.37 ± 2.30
gideon_Catalase	87.35 ± 9.15	77.79 ± 13.31	81.13 ± 9.39	89.94 ± 6.15	81.92 ± 9.71	85.77 ± 5.39	89.56 ± 8.99	81.57 ± 15.47	81.70 ± 4.56	89.57 ± 6.13	81.72 ± 9.09	84.23 ± 3.91	94.35 ± 3.43	90.67 ± 6.24	87.54 ± 3.48	90.54 ± 5.98	87.23 ± 9.12	88.58 ± 5.83	80.88 ± 5.40	74.20 ± 7.17	73.54 ± 11.69	91.19 ± 5.12	83.68 ± 9.00	86.00 ± 4.23	**94.73 ± 2.93**	**91.19 ± 5.81**	**90.22 ± 2.33**
gideon_Citrate	57.87 ± 6.09	76.67 ± 9.44	43.66 ± 6.78	63.63 ± 1.99	80.56 ± 5.40	43.50 ± 6.26	66.23 ± 8.61	83.51 ± 8.23	56.97 ± 10.62	**75.39 ± 12.62**	**85.61 ± 9.63**	55.97 ± 18.23	72.41 ± 15.05	84.18 ± 13.50	61.14 ± 19.50	72.19 ± 10.49	84.31 ± 10.80	55.80 ± 14.02	74.15 ± 10.85	84.47 ± 11.73	**64.05 ± 8.14**	74.87 ± 8.46	85.35 ± 6.48	53.78 ± 14.29	69.30 ± 20.62	82.79 ± 12.39	62.19 ± 18.43
gideon_Coccus	68.59 ± 4.91	91.21 ± 2.80	47.20 ± 6.80	70.21 ± 13.51	93.78 ± 2.55	46.63 ± 0.68	81.38 ± 9.72	95.67 ± 2.41	64.43 ± 17.13	77.49 ± 10.39	94.38 ± 2.94	62.52 ± 14.51	91.16 ± 5.58	97.30 ± 1.73	75.01 ± 17.61	77.96 ± 26.27	96.02 ± 3.79	69.91 ± 19.30	81.81 ± 15.74	96.59 ± 1.94	65.29 ± 18.71	53.07 ± 34.14	87.81 ± 8.74	45.85 ± 1.30	**93.94 ± 2.10**	**98.66 ± 0.69**	**81.98 ± 4.34**
gideon_Coccus - pairs or chains predominate	79.84 ± 3.05	98.98 ± 0.85	48.77 ± 1.18	83.59 ± 0.98	99.11 ± 0.84	48.77 ± 1.18	83.63 ± 5.81	99.26 ± 0.52	51.39 ± 5.03	83.22 ± 11.78	99.44 ± 0.24	52.22 ± 4.80	86.92 ± 16.60	99.57 ± 0.50	**67.14 ± 18.28**	84.58 ± 15.95	99.47 ± 0.47	48.77 ± 1.18	65.23 ± 6.30	97.60 ± 2.71	48.35 ± 0.91	52.85 ± 38.25	97.67 ± 1.59	48.77 ± 1.18	**90.94 ± 7.58**	**99.71 ± 0.19**	52.42 ± 3.23
gideon_D-Mannitol	67.26 ± 1.13	81.39 ± 1.66	43.97 ± 6.96	76.43 ± 6.87	86.62 ± 3.52	48.72 ± 11.13	75.88 ± 4.69	86.52 ± 2.79	51.38 ± 9.71	80.31 ± 3.43	88.86 ± 2.37	49.10 ± 21.79	77.25 ± 8.27	87.75 ± 3.59	56.58 ± 15.49	74.89 ± 7.70	86.46 ± 2.48	49.33 ± 16.45	80.28 ± 5.57	87.06 ± 2.87	**63.15 ± 2.26**	**81.35 ± 5.08**	**88.96 ± 2.73**	62.90 ± 3.58	77.72 ± 1.44	87.70 ± 1.21	55.76 ± 14.06
gideon_D-Mannose	47.87 ± 9.11	48.12 ± 6.55	50.56 ± 9.26	48.67 ± 10.77	49.48 ± 11.58	40.82 ± 10.12	51.65 ± 8.16	53.79 ± 7.16	42.33 ± 12.64	60.94 ± 12.28	61.51 ± 1.71	46.49 ± 11.84	49.75 ± 5.47	51.26 ± 9.23	41.65 ± 7.39	66.94 ± 7.88	65.34 ± 3.77	56.30 ± 3.31	66.58 ± 12.40	64.64 ± 8.35	58.98 ± 9.42	62.13 ± 3.61	61.54 ± 8.03	47.70 ± 14.28	**73.86 ± 16.47**	**76.60 ± 12.59**	**63.37 ± 12.23**
gideon_D-Sorbitol	65.14 ± 12.52	92.68 ± 3.01	46.00 ± 1.89	53.10 ± 19.68	88.06 ± 7.50	46.00 ± 1.89	**72.23 ± 5.02**	93.97 ± 1.84	**58.71 ± 0.54**	64.39 ± 16.30	89.39 ± 5.49	49.00 ± 6.83	62.34 ± 18.02	91.75 ± 2.77	47.28 ± 3.03	70.87 ± 12.91	**94.01 ± 2.64**	48.22 ± 5.50	66.15 ± 15.81	92.51 ± 0.61	49.75 ± 4.96	59.48 ± 18.72	88.40 ± 5.59	46.00 ± 1.89	54.91 ± 23.09	89.10 ± 5.27	47.67 ± 4.58
gideon_D-Xylose	62.12 ± 2.18	83.82 ± 1.21	42.71 ± 1.19	70.19 ± 2.64	87.07 ± 2.03	56.95 ± 12.84	72.62 ± 7.43	88.91 ± 3.37	55.44 ± 0.83	76.35 ± 6.16	89.51 ± 2.89	**67.75 ± 4.90**	66.86 ± 5.41	87.35 ± 2.10	52.65 ± 7.77	71.17 ± 10.79	89.49 ± 3.55	56.29 ± 11.66	75.51 ± 5.16	89.25 ± 2.82	66.06 ± 3.50	69.35 ± 1.00	85.93 ± 1.47	61.46 ± 6.87	**76.69 ± 7.51**	**89.85 ± 3.36**	65.33 ± 1.52
gideon_Esculin hydrolysis	59.34 ± 12.27	70.35 ± 14.88	40.54 ± 1.29	62.60 ± 8.34	71.86 ± 16.29	40.52 ± 5.67	70.34 ± 7.42	77.76 ± 10.11	60.38 ± 3.69	63.40 ± 3.62	71.78 ± 4.90	47.22 ± 17.60	59.67 ± 17.17	69.05 ± 15.05	55.48 ± 16.58	69.51 ± 19.30	75.15 ± 17.42	60.51 ± 19.25	72.27 ± 13.64	**80.03 ± 8.55**	53.54 ± 17.86	63.50 ± 8.47	73.48 ± 8.04	48.88 ± 14.14	**76.08 ± 18.25**	78.69 ± 18.41	**71.95 ± 10.53**
gideon_Facultative	71.07 ± 12.49	88.07 ± 7.42	43.12 ± 2.72	68.88 ± 30.25	83.53 ± 14.34	66.85 ± 21.48	86.07 ± 7.98	92.05 ± 6.43	79.25 ± 10.60	**90.83 ± 4.94**	**95.80 ± 4.08**	67.98 ± 24.39	83.51 ± 8.39	93.07 ± 4.47	67.21 ± 23.75	77.57 ± 13.56	91.69 ± 5.19	62.21 ± 22.18	86.00 ± 9.58	92.57 ± 6.05	81.17 ± 10.33	73.03 ± 27.55	86.66 ± 12.70	52.26 ± 18.35	89.95 ± 4.08	93.09 ± 4.96	**84.52 ± 4.70**
gideon_Gas from glucose	45.55 ± 12.51	84.64 ± 6.25	45.71 ± 1.83	53.39 ± 14.32	85.86 ± 2.44	50.12 ± 8.70	55.63 ± 15.62	87.21 ± 3.30	54.09 ± 9.12	**67.48 ± 6.90**	89.21 ± 5.13	51.10 ± 10.39	56.26 ± 15.56	88.14 ± 2.96	51.55 ± 11.14	54.02 ± 11.47	89.13 ± 1.71	45.71 ± 1.83	67.28 ± 17.75	**90.02 ± 6.42**	59.33 ± 1.92	64.60 ± 10.16	88.42 ± 4.01	51.10 ± 10.39	65.93 ± 15.42	89.96 ± 4.14	**61.81 ± 15.65**
gideon_Gelatin hydrolysis	59.28 ± 7.85	77.91 ± 5.26	41.80 ± 1.53	53.84 ± 9.03	76.09 ± 5.50	41.80 ± 1.53	52.84 ± 9.80	75.81 ± 7.65	51.59 ± 10.25	55.20 ± 13.82	78.84 ± 8.14	50.24 ± 9.44	60.06 ± 8.00	78.78 ± 6.15	41.80 ± 1.53	60.40 ± 11.24	80.62 ± 9.65	41.80 ± 1.53	57.56 ± 11.62	78.79 ± 6.07	47.76 ± 2.86	50.75 ± 18.39	75.99 ± 11.95	42.95 ± 1.31	**72.97 ± 5.20**	**87.75 ± 3.81**	**56.68 ± 3.98**
gideon_Glucose fermenter	67.90 ± 3.09	60.96 ± 6.11	54.16 ± 18.46	71.96 ± 12.66	68.45 ± 12.45	54.86 ± 14.11	80.33 ± 10.15	75.86 ± 16.46	68.71 ± 11.37	81.54 ± 10.00	75.05 ± 19.23	68.90 ± 11.12	81.55 ± 5.23	77.21 ± 10.89	68.60 ± 18.54	71.61 ± 14.38	63.45 ± 20.43	50.94 ± 23.49	88.68 ± 6.59	84.73 ± 12.56	76.20 ± 8.41	80.43 ± 7.18	74.02 ± 15.47	69.44 ± 7.79	**89.98 ± 3.90**	**86.31 ± 7.88**	**83.65 ± 4.72**
gideon_Glucose oxidizer	83.97 ± 4.94	95.84 ± 3.32	46.11 ± 2.76	84.60 ± 5.25	95.79 ± 3.33	46.11 ± 2.76	87.25 ± 5.58	97.20 ± 2.54	56.47 ± 20.27	89.34 ± 3.81	97.83 ± 2.10	54.38 ± 8.89	85.61 ± 3.00	97.25 ± 2.07	52.26 ± 3.66	88.09 ± 7.18	97.50 ± 2.86	49.93 ± 3.85	78.39 ± 13.91	96.31 ± 2.00	46.99 ± 3.36	87.37 ± 2.03	97.65 ± 2.04	46.06 ± 2.83	**90.01 ± 3.45**	**98.10 ± 1.80**	**64.55 ± 14.18**
gideon_Glycerol	69.19 ± 17.58	84.50 ± 15.97	49.11 ± 8.80	63.20 ± 14.42	82.40 ± 13.28	49.78 ± 8.75	71.17 ± 18.47	85.34 ± 15.36	56.02 ± 15.29	68.21 ± 13.95	85.98 ± 12.32	43.72 ± 0.89	71.71 ± 14.45	87.31 ± 10.38	50.81 ± 5.52	**74.94 ± 15.00**	**88.99 ± 9.87**	**61.74 ± 15.38**	66.91 ± 16.02	85.61 ± 9.57	46.73 ± 8.52	69.68 ± 15.55	85.01 ± 14.87	48.42 ± 8.64	74.23 ± 13.67	88.93 ± 9.42	56.39 ± 9.20
gideon_Gram negative	67.08 ± 13.42	67.37 ± 6.34	50.41 ± 12.62	85.33 ± 0.50	85.84 ± 4.62	62.38 ± 23.73	89.11 ± 8.97	88.40 ± 9.71	81.49 ± 8.53	97.09 ± 2.46	97.37 ± 1.93	86.66 ± 9.93	98.03 ± 2.44	98.19 ± 2.11	88.97 ± 10.79	98.70 ± 1.99	98.44 ± 2.39	91.74 ± 5.48	99.02 ± 0.42	99.05 ± 0.58	94.01 ± 2.96	88.52 ± 12.16	85.39 ± 18.57	80.78 ± 18.81	**99.47 ± 0.22**	**99.16 ± 0.61**	**97.75 ± 1.05**
gideon_Gram positive	65.58 ± 6.77	65.98 ± 13.59	54.27 ± 1.52	86.88 ± 5.47	86.63 ± 7.71	73.17 ± 7.60	88.12 ± 8.21	89.65 ± 6.03	75.83 ± 14.34	97.07 ± 3.41	97.53 ± 2.96	93.17 ± 3.71	98.48 ± 1.69	98.58 ± 1.62	93.26 ± 6.92	**99.73 ± 0.29**	**99.71 ± 0.34**	**94.29 ± 3.85**	97.50 ± 2.39	97.82 ± 1.88	92.68 ± 2.81	96.06 ± 4.29	96.69 ± 3.73	92.29 ± 3.52	97.61 ± 3.16	98.27 ± 2.33	76.39 ± 37.69
gideon_Growth on MacConkey agar	62.83 ± 6.12	84.09 ± 6.04	44.84 ± 3.76	90.74 ± 3.82	95.45 ± 2.59	72.09 ± 8.48	94.48 ± 3.70	97.82 ± 1.48	84.02 ± 3.73	94.31 ± 3.77	97.88 ± 1.06	74.04 ± 14.03	94.51 ± 4.02	97.73 ± 2.21	76.76 ± 25.21	**95.78 ± 2.05**	**98.49 ± 0.92**	**86.61 ± 6.47**	95.29 ± 3.19	98.06 ± 1.52	85.86 ± 7.89	94.25 ± 2.77	97.69 ± 0.88	70.89 ± 10.37	95.31 ± 3.57	98.31 ± 1.51	82.14 ± 12.99
gideon_Growth on ordinary blood agar	68.45 ± 12.66	19.93 ± 24.85	49.10 ± 0.77	59.06 ± 33.91	34.49 ± 26.06	49.10 ± 0.77	74.35 ± 19.44	33.12 ± 29.25	54.70 ± 9.92	**86.35 ± 16.24**	47.97 ± 32.86	56.51 ± 13.04	76.84 ± 19.17	49.20 ± 33.21	61.73 ± 22.09	66.33 ± 24.10	39.96 ± 35.90	64.09 ± 26.18	75.11 ± 32.78	41.78 ± 35.31	60.21 ± 19.56	78.07 ± 9.74	36.90 ± 30.15	49.10 ± 0.77	75.73 ± 32.26	**51.72 ± 31.62**	**74.89 ± 16.84**
gideon_Hydrogen sulfide	34.06 ± 9.54	81.32 ± 4.20	46.26 ± 0.73	47.63 ± 16.33	85.18 ± 4.07	46.26 ± 0.73	64.32 ± 6.33	89.30 ± 5.73	52.96 ± 5.56	59.32 ± 15.68	90.54 ± 5.10	50.07 ± 6.72	71.30 ± 13.29	93.01 ± 3.52	51.20 ± 8.92	69.96 ± 17.58	92.92 ± 3.09	46.26 ± 0.73	**76.96 ± 6.94**	**95.06 ± 1.41**	49.34 ± 5.71	60.86 ± 7.42	91.17 ± 2.81	47.72 ± 3.44	75.28 ± 10.50	94.75 ± 1.79	**58.89 ± 8.87**
gideon_Indole	59.14 ± 6.19	87.96 ± 0.99	45.46 ± 0.60	60.00 ± 12.03	89.55 ± 4.64	45.46 ± 0.60	**68.84 ± 7.46**	**90.93 ± 3.07**	**50.43 ± 9.13**	66.92 ± 17.26	90.55 ± 4.93	45.39 ± 0.62	66.13 ± 11.45	90.31 ± 4.24	46.74 ± 2.91	54.46 ± 9.52	86.49 ± 5.14	45.46 ± 0.60	62.03 ± 20.09	86.59 ± 10.60	45.46 ± 0.60	65.16 ± 13.40	90.61 ± 2.82	46.54 ± 1.92	66.65 ± 8.35	90.38 ± 3.32	46.42 ± 1.83
gideon_L-Arabinose	72.55 ± 10.19	84.21 ± 11.77	55.04 ± 19.59	74.55 ± 8.25	85.83 ± 9.27	60.75 ± 11.80	75.67 ± 8.09	85.95 ± 9.47	58.74 ± 9.35	79.79 ± 4.59	88.26 ± 8.30	65.82 ± 6.95	79.00 ± 4.77	87.32 ± 10.69	56.32 ± 12.54	77.02 ± 9.79	87.58 ± 10.66	59.91 ± 13.48	82.00 ± 2.66	89.75 ± 6.55	68.19 ± 11.38	80.32 ± 2.27	89.52 ± 5.30	62.72 ± 15.43	**82.80 ± 0.70**	**90.95 ± 4.21**	**73.49 ± 5.50**
gideon_L-Rhamnose	69.02 ± 7.46	91.83 ± 2.26	44.77 ± 1.99	75.02 ± 12.50	93.84 ± 1.81	58.41 ± 9.93	80.50 ± 6.12	95.04 ± 0.42	**61.14 ± 11.89**	67.76 ± 10.33	91.08 ± 3.66	46.24 ± 2.20	76.16 ± 9.27	94.58 ± 1.84	44.77 ± 1.99	79.36 ± 1.75	95.02 ± 1.47	44.71 ± 1.87	**83.39 ± 5.73**	**95.93 ± 1.70**	55.96 ± 7.14	68.98 ± 17.61	92.67 ± 1.59	59.57 ± 12.06	79.61 ± 11.42	93.93 ± 4.32	59.57 ± 14.25
gideon_Lactose	49.82 ± 11.14	72.76 ± 7.55	41.21 ± 2.25	49.42 ± 13.86	70.40 ± 14.96	48.09 ± 8.45	64.61 ± 4.11	81.30 ± 3.66	59.42 ± 4.82	73.25 ± 7.74	82.04 ± 9.18	68.67 ± 7.34	70.23 ± 7.16	82.13 ± 6.56	62.63 ± 6.81	74.99 ± 5.59	86.87 ± 6.51	63.08 ± 3.96	60.69 ± 24.89	78.94 ± 17.69	56.72 ± 16.16	72.80 ± 6.27	83.68 ± 7.47	52.75 ± 20.94	**82.05 ± 4.36**	**91.00 ± 0.71**	**74.48 ± 5.62**
gideon_Maltose	57.91 ± 1.00	55.30 ± 8.84	47.22 ± 6.44	58.18 ± 22.63	58.08 ± 24.71	54.22 ± 14.94	66.58 ± 12.38	62.00 ± 19.12	54.05 ± 15.06	67.24 ± 7.49	59.26 ± 17.49	56.15 ± 2.74	74.67 ± 1.66	72.94 ± 7.65	65.75 ± 3.22	71.11 ± 11.70	69.61 ± 1.98	48.26 ± 17.97	71.89 ± 7.93	71.22 ± 9.94	61.95 ± 7.69	61.95 ± 5.23	56.15 ± 9.43	37.46 ± 7.87	**83.91 ± 3.76**	**82.49 ± 8.20**	**73.60 ± 3.05**
gideon_Melibiose	53.02 ± 10.99	81.48 ± 1.69	40.56 ± 4.24	53.22 ± 6.67	80.43 ± 1.65	43.27 ± 1.25	48.54 ± 12.83	78.12 ± 3.54	43.08 ± 3.55	49.02 ± 6.05	75.76 ± 2.23	44.91 ± 1.23	52.09 ± 16.81	79.72 ± 5.27	40.81 ± 3.83	55.62 ± 14.34	82.83 ± 2.66	41.10 ± 3.15	55.17 ± 9.81	81.22 ± 4.75	47.70 ± 5.65	47.86 ± 4.96	75.94 ± 3.27	50.47 ± 6.12	**64.98 ± 13.08**	**86.77 ± 5.30**	**50.99 ± 2.97**
gideon_Motile	66.98 ± 9.54	77.79 ± 6.46	43.13 ± 5.84	76.09 ± 2.49	84.66 ± 0.67	47.19 ± 6.11	81.63 ± 3.84	87.49 ± 2.55	66.00 ± 14.77	64.59 ± 5.22	77.20 ± 1.69	46.32 ± 12.48	82.07 ± 4.29	88.24 ± 1.71	70.20 ± 11.65	74.70 ± 8.69	82.96 ± 6.19	52.81 ± 23.70	79.50 ± 6.52	81.25 ± 5.43	**70.29 ± 11.87**	63.98 ± 5.82	77.26 ± 4.27	47.11 ± 9.46	**87.01 ± 9.71**	**91.10 ± 7.41**	69.68 ± 27.96
gideon_Nitrate to nitrite	65.98 ± 7.21	78.20 ± 6.12	48.95 ± 11.46	75.18 ± 4.18	84.73 ± 4.76	61.95 ± 8.20	77.91 ± 7.68	84.88 ± 6.40	60.71 ± 22.35	76.39 ± 15.52	83.96 ± 13.73	58.99 ± 29.05	75.60 ± 11.61	85.66 ± 6.94	53.70 ± 20.78	80.16 ± 11.98	88.77 ± 6.94	66.94 ± 11.92	**85.33 ± 9.63**	**91.13 ± 6.07**	**76.12 ± 10.70**	75.32 ± 19.42	83.56 ± 15.71	66.89 ± 17.98	81.54 ± 13.43	89.02 ± 6.87	70.42 ± 13.71
gideon_ONPG (beta galactosidase)	59.67 ± 8.24	67.14 ± 8.62	46.07 ± 6.43	71.63 ± 10.84	73.93 ± 9.74	60.04 ± 17.28	70.53 ± 6.45	72.03 ± 10.86	50.69 ± 10.83	74.80 ± 8.38	75.81 ± 5.94	57.43 ± 26.22	81.60 ± 6.41	81.15 ± 11.36	**71.65 ± 14.25**	76.65 ± 9.60	78.21 ± 4.09	67.41 ± 3.54	68.94 ± 12.20	69.56 ± 18.47	66.97 ± 12.46	69.27 ± 6.67	69.11 ± 21.16	58.34 ± 28.04	**86.81 ± 6.65**	**88.21 ± 2.32**	69.95 ± 14.30
gideon_Oxidase	74.24 ± 7.35	87.86 ± 7.56	55.48 ± 14.06	74.34 ± 14.10	87.28 ± 10.78	60.90 ± 18.80	82.42 ± 7.15	92.15 ± 4.79	76.51 ± 9.75	87.62 ± 7.74	95.30 ± 3.46	76.77 ± 11.17	85.52 ± 5.18	94.40 ± 3.57	62.87 ± 3.98	85.35 ± 13.50	94.23 ± 6.98	65.49 ± 10.93	82.68 ± 10.61	91.31 ± 6.93	65.73 ± 21.38	**88.15 ± 7.77**	95.44 ± 3.38	78.07 ± 11.32	87.49 ± 11.49	**95.50 ± 4.90**	**81.04 ± 7.89**
gideon_Raffinose	58.43 ± 13.93	87.67 ± 2.53	44.29 ± 1.90	66.08 ± 9.36	87.88 ± 0.82	47.67 ± 6.87	63.31 ± 18.28	89.25 ± 4.49	50.64 ± 7.38	74.47 ± 14.71	90.34 ± 7.60	56.60 ± 12.10	73.48 ± 10.27	91.95 ± 2.13	57.34 ± 10.13	73.86 ± 11.52	92.10 ± 2.56	61.25 ± 13.90	68.52 ± 14.54	89.99 ± 3.56	59.93 ± 12.11	70.24 ± 15.00	90.09 ± 5.42	52.83 ± 6.88	**81.69 ± 9.85**	**94.83 ± 2.68**	**63.74 ± 8.19**
gideon_Salicin	52.52 ± 4.92	72.80 ± 6.01	39.85 ± 1.06	46.99 ± 24.89	68.07 ± 18.52	40.60 ± 1.78	47.05 ± 8.77	67.42 ± 11.22	39.95 ± 1.07	56.69 ± 25.15	72.59 ± 19.77	44.33 ± 7.78	63.14 ± 13.79	80.66 ± 6.83	48.12 ± 14.31	59.98 ± 18.03	76.10 ± 10.66	44.21 ± 7.59	55.25 ± 28.23	73.10 ± 20.44	48.83 ± 15.54	56.35 ± 15.29	71.32 ± 9.21	39.95 ± 1.07	**72.08 ± 19.71**	**84.76 ± 12.21**	**55.61 ± 14.66**
gideon_Spore formation	52.59 ± 21.62	89.77 ± 5.09	46.72 ± 1.71	72.71 ± 16.94	96.27 ± 1.53	46.72 ± 1.71	79.25 ± 27.78	97.14 ± 3.40	51.50 ± 8.72	83.66 ± 22.84	98.19 ± 2.03	**55.87 ± 14.28**	89.55 ± 8.27	98.89 ± 0.71	48.50 ± 1.64	89.91 ± 12.13	99.01 ± 0.89	46.72 ± 1.71	90.68 ± 7.52	98.87 ± 0.30	55.64 ± 13.89	74.53 ± 30.76	96.22 ± 3.20	46.72 ± 1.71	**90.85 ± 10.38**	**99.09 ± 0.67**	46.29 ± 1.22
gideon_Sucrose	55.37 ± 6.42	58.50 ± 6.96	38.05 ± 3.52	68.64 ± 10.82	74.75 ± 10.50	62.72 ± 7.09	70.79 ± 7.20	75.94 ± 7.02	48.38 ± 16.02	71.19 ± 4.74	75.27 ± 7.30	64.27 ± 3.91	72.30 ± 4.40	76.48 ± 6.15	49.67 ± 13.88	76.03 ± 4.41	78.95 ± 3.24	64.30 ± 3.80	67.99 ± 8.80	70.46 ± 3.17	55.37 ± 8.00	64.82 ± 5.59	68.90 ± 8.49	60.44 ± 3.86	**81.50 ± 5.20**	**85.10 ± 5.54**	**69.96 ± 4.71**
gideon_Trehalose	64.66 ± 8.83	66.39 ± 10.75	41.12 ± 7.45	73.73 ± 9.10	75.09 ± 8.03	57.29 ± 10.60	74.83 ± 9.37	76.43 ± 8.99	63.59 ± 10.52	71.43 ± 11.03	72.99 ± 10.39	56.44 ± 6.11	74.71 ± 7.56	73.69 ± 9.15	68.53 ± 7.30	75.98 ± 5.39	75.93 ± 8.81	55.45 ± 19.82	**78.99 ± 5.11**	**78.57 ± 6.23**	**68.69 ± 4.12**	69.37 ± 13.20	70.46 ± 12.35	55.06 ± 13.16	78.30 ± 5.12	76.44 ± 8.00	59.82 ± 16.01
gideon_Urea hydrolysis	67.39 ± 9.31	91.28 ± 0.89	45.78 ± 0.63	71.71 ± 11.97	92.87 ± 2.92	45.49 ± 0.48	72.61 ± 11.93	93.36 ± 2.24	48.03 ± 2.57	71.68 ± 11.90	93.36 ± 1.46	**56.57 ± 8.14**	70.27 ± 3.80	92.84 ± 1.64	53.59 ± 5.20	74.29 ± 9.15	93.69 ± 1.07	55.26 ± 5.21	68.73 ± 8.30	91.78 ± 3.86	51.03 ± 6.06	71.19 ± 11.70	93.25 ± 1.22	47.67 ± 3.43	**74.68 ± 13.08**	**94.34 ± 1.73**	47.74 ± 3.55
gideon_Yellow pigment	61.93 ± 23.77	93.80 ± 5.14	52.90 ± 9.29	58.44 ± 23.82	92.37 ± 5.88	47.78 ± 0.42	56.24 ± 22.96	92.39 ± 6.71	50.80 ± 5.66	63.14 ± 19.58	94.14 ± 5.34	50.80 ± 5.66	69.04 ± 25.74	95.54 ± 4.67	47.78 ± 0.42	**83.95 ± 6.66**	**98.18 ± 1.06**	47.78 ± 0.42	57.86 ± 22.22	93.89 ± 4.31	50.82 ± 5.26	61.95 ± 16.66	94.32 ± 3.53	47.78 ± 0.42	72.38 ± 24.46	95.55 ± 4.75	**56.44 ± 8.49**
gideon_myo-Inositol	74.39 ± 6.10	95.02 ± 2.06	46.94 ± 1.51	66.91 ± 8.64	93.24 ± 4.83	46.94 ± 1.51	**76.08 ± 1.22**	95.30 ± 2.49	46.51 ± 1.43	75.71 ± 2.18	**95.94 ± 2.22**	**54.50 ± 7.70**	69.13 ± 8.02	92.67 ± 5.39	46.77 ± 1.44	72.10 ± 8.69	93.41 ± 3.65	46.75 ± 1.33	69.17 ± 8.41	93.68 ± 3.98	46.94 ± 1.51	75.62 ± 2.83	95.79 ± 2.26	46.94 ± 1.51	69.43 ± 6.30	94.34 ± 2.53	46.94 ± 1.51
madin_carbsubs_2-ketogluconate	51.01 ± 6.03	2.92 ± 1.17	**49.41 ± 0.06**	49.91 ± 11.97	2.56 ± 0.88	**49.41 ± 0.06**	**64.84 ± 6.91**	**4.96 ± 2.88**	**49.41 ± 0.06**	60.45 ± 3.39	4.19 ± 1.03	**49.41 ± 0.06**	56.51 ± 12.69	4.74 ± 3.08	**49.41 ± 0.06**	52.26 ± 10.65	2.96 ± 1.38	**49.41 ± 0.06**	57.83 ± 10.23	3.45 ± 1.30	**49.41 ± 0.06**	52.71 ± 7.89	4.75 ± 2.11	**49.41 ± 0.06**	60.10 ± 6.40	4.09 ± 1.37	**49.41 ± 0.06**
madin_carbsubs_3-hydroxybutyrate	65.81 ± 5.31	7.43 ± 1.94	**49.00 ± 0.05**	62.07 ± 4.94	7.91 ± 2.38	**49.00 ± 0.05**	72.36 ± 6.51	10.62 ± 1.65	**49.00 ± 0.05**	66.78 ± 7.44	8.66 ± 2.14	**49.00 ± 0.05**	67.14 ± 6.77	10.21 ± 4.27	48.96 ± 0.07	**73.28 ± 4.30**	**12.91 ± 2.86**	**49.00 ± 0.05**	69.81 ± 7.30	11.96 ± 3.03	**49.00 ± 0.05**	59.04 ± 4.15	6.65 ± 2.21	**49.00 ± 0.05**	68.38 ± 9.52	11.35 ± 4.44	**49.00 ± 0.05**
madin_carbsubs_5-ketogluconate	46.85 ± 0.59	**4.59 ± 2.76**	**49.14 ± 0.17**	**51.24 ± 3.66**	4.25 ± 1.67	**49.14 ± 0.17**	47.15 ± 7.23	3.47 ± 0.56	**49.14 ± 0.17**	47.97 ± 1.50	3.75 ± 0.17	**49.14 ± 0.17**	43.41 ± 1.70	3.26 ± 0.78	**49.14 ± 0.17**	50.00 ± 3.48	4.24 ± 0.95	**49.14 ± 0.17**	49.51 ± 0.25	3.99 ± 0.67	49.13 ± 0.15	46.59 ± 5.00	4.44 ± 1.42	**49.14 ± 0.17**	41.16 ± 4.52	3.46 ± 0.89	**49.14 ± 0.17**
madin_carbsubs_N-acetylgulcosamine	62.78 ± 4.56	28.16 ± 7.31	44.29 ± 0.54	64.19 ± 4.40	28.06 ± 4.85	44.29 ± 0.54	64.80 ± 3.99	30.16 ± 3.38	47.26 ± 1.45	65.31 ± 3.91	32.74 ± 5.58	45.01 ± 1.39	70.79 ± 2.87	38.08 ± 2.64	49.16 ± 3.49	70.74 ± 2.00	38.01 ± 1.23	47.37 ± 3.00	70.57 ± 0.93	38.41 ± 2.73	47.67 ± 4.57	64.80 ± 4.89	32.64 ± 6.64	45.33 ± 1.45	**71.20 ± 2.68**	**40.44 ± 3.01**	**51.65 ± 4.06**
madin_carbsubs_Tween 20	57.51 ± 3.75	7.59 ± 1.55	**48.51 ± 0.06**	61.36 ± 1.49	9.77 ± 3.22	**48.51 ± 0.06**	61.13 ± 5.07	11.43 ± 2.86	**48.51 ± 0.06**	64.47 ± 1.47	14.24 ± 3.75	**48.51 ± 0.06**	75.58 ± 3.17	16.70 ± 2.76	**48.51 ± 0.06**	77.16 ± 1.79	19.07 ± 4.10	**48.51 ± 0.06**	75.57 ± 2.56	17.73 ± 3.36	**48.51 ± 0.06**	65.13 ± 2.50	14.14 ± 4.10	**48.51 ± 0.06**	**77.99 ± 4.64**	**20.20 ± 2.21**	**48.51 ± 0.06**
madin_carbsubs_Tween 40	54.09 ± 5.48	12.28 ± 5.13	47.42 ± 0.64	59.16 ± 1.35	12.76 ± 1.74	47.42 ± 0.64	60.10 ± 3.77	13.60 ± 3.82	47.42 ± 0.64	61.43 ± 4.91	15.39 ± 5.45	47.42 ± 0.64	69.08 ± 4.43	20.51 ± 6.97	47.78 ± 0.89	69.35 ± 4.83	**20.86 ± 6.27**	47.42 ± 0.64	**70.42 ± 4.17**	19.81 ± 8.77	47.42 ± 0.64	62.84 ± 4.40	16.60 ± 4.92	47.42 ± 0.64	69.14 ± 3.90	18.56 ± 8.50	**48.60 ± 1.70**
madin_carbsubs_Tween 60	49.02 ± 11.61	3.35 ± 1.52	**49.29 ± 0.20**	59.23 ± 6.26	3.44 ± 0.84	**49.29 ± 0.20**	54.38 ± 9.27	4.58 ± 1.71	**49.29 ± 0.20**	61.66 ± 5.61	4.85 ± 0.95	**49.29 ± 0.20**	72.06 ± 1.83	6.90 ± 1.99	**49.29 ± 0.20**	73.99 ± 3.29	6.87 ± 1.71	**49.29 ± 0.20**	**74.70 ± 2.86**	6.80 ± 1.42	**49.29 ± 0.20**	61.30 ± 2.76	4.14 ± 0.93	**49.29 ± 0.20**	72.46 ± 5.93	**11.96 ± 6.51**	**49.29 ± 0.20**
madin_carbsubs_Tween 80	53.72 ± 2.95	14.20 ± 1.97	47.03 ± 0.44	64.30 ± 3.97	19.87 ± 3.33	47.03 ± 0.44	62.02 ± 1.27	16.41 ± 2.11	47.03 ± 0.44	67.77 ± 1.67	20.81 ± 3.54	48.01 ± 1.47	68.46 ± 3.62	23.61 ± 4.55	48.54 ± 1.27	**69.32 ± 2.87**	**25.76 ± 3.99**	48.16 ± 1.73	68.70 ± 6.17	23.03 ± 6.36	47.03 ± 0.44	66.83 ± 2.57	20.15 ± 2.78	47.03 ± 0.44	68.42 ± 3.70	24.59 ± 2.26	**49.73 ± 5.12**
madin_carbsubs_acetate	55.25 ± 6.25	21.92 ± 3.07	44.39 ± 0.15	66.16 ± 1.33	31.86 ± 3.63	44.37 ± 0.13	65.96 ± 2.82	31.72 ± 3.58	47.98 ± 4.26	62.85 ± 2.95	29.90 ± 3.54	54.01 ± 8.53	62.31 ± 8.61	26.54 ± 6.24	46.52 ± 1.90	67.63 ± 1.67	32.84 ± 4.64	47.52 ± 3.84	**70.29 ± 1.00**	**34.50 ± 2.81**	48.63 ± 4.45	63.68 ± 2.75	31.29 ± 3.92	44.39 ± 0.15	68.24 ± 2.21	34.02 ± 6.25	**58.86 ± 1.50**
madin_carbsubs_adipate	68.72 ± 3.56	7.84 ± 3.19	**49.09 ± 0.17**	74.00 ± 5.22	11.33 ± 7.18	**49.09 ± 0.17**	73.19 ± 4.60	**15.61 ± 8.23**	**49.09 ± 0.17**	68.10 ± 9.48	13.02 ± 8.89	**49.09 ± 0.17**	74.22 ± 6.11	14.17 ± 7.92	**49.09 ± 0.17**	75.18 ± 4.65	15.05 ± 3.78	**49.09 ± 0.17**	72.33 ± 4.87	9.90 ± 3.79	49.08 ± 0.17	70.18 ± 7.34	13.89 ± 10.58	**49.09 ± 0.17**	**75.44 ± 4.47**	12.46 ± 2.69	**49.09 ± 0.17**
madin_carbsubs_adonitol	59.96 ± 6.96	8.05 ± 1.58	**48.27 ± 0.54**	62.22 ± 9.20	8.69 ± 2.15	**48.27 ± 0.54**	60.45 ± 6.85	8.52 ± 2.35	**48.27 ± 0.54**	64.59 ± 7.49	12.49 ± 3.96	**48.27 ± 0.54**	57.42 ± 6.36	8.31 ± 3.44	**48.27 ± 0.54**	64.20 ± 5.57	9.48 ± 3.15	**48.27 ± 0.54**	65.38 ± 12.66	**13.70 ± 2.96**	**48.27 ± 0.54**	62.98 ± 6.96	11.94 ± 3.82	**48.27 ± 0.54**	**65.49 ± 9.35**	12.68 ± 3.69	**48.27 ± 0.54**
madin_carbsubs_alaninamide	58.81 ± 9.93	5.06 ± 1.69	**49.20 ± 0.05**	58.61 ± 2.47	6.35 ± 3.56	**49.20 ± 0.05**	61.32 ± 1.22	**7.80 ± 2.94**	**49.20 ± 0.05**	58.32 ± 5.50	4.08 ± 0.58	**49.20 ± 0.05**	61.32 ± 3.94	6.11 ± 1.37	**49.20 ± 0.05**	60.65 ± 5.41	5.16 ± 1.55	**49.20 ± 0.05**	59.12 ± 3.81	6.65 ± 3.17	**49.20 ± 0.05**	60.49 ± 3.78	4.48 ± 0.47	**49.20 ± 0.05**	**64.38 ± 6.35**	6.35 ± 2.11	**49.20 ± 0.05**
madin_carbsubs_alanine	57.50 ± 4.85	20.63 ± 6.77	45.72 ± 0.43	60.59 ± 0.86	22.19 ± 1.17	45.72 ± 0.43	60.46 ± 4.37	22.11 ± 1.37	45.72 ± 0.43	63.21 ± 4.16	26.30 ± 4.01	46.42 ± 1.59	64.35 ± 2.22	24.66 ± 5.33	45.72 ± 0.43	58.29 ± 12.37	22.66 ± 6.91	45.72 ± 0.43	**66.49 ± 1.11**	**28.01 ± 2.66**	46.47 ± 1.67	63.58 ± 3.08	27.24 ± 2.91	46.09 ± 1.02	63.91 ± 2.95	24.55 ± 5.69	**50.76 ± 4.77**
madin_carbsubs_alpha-cyclodextrin	63.29 ± 7.64	3.99 ± 1.08	**49.49 ± 0.04**	65.78 ± 2.57	3.98 ± 0.47	**49.49 ± 0.04**	73.98 ± 5.79	4.68 ± 0.95	**49.49 ± 0.04**	73.29 ± 4.87	6.64 ± 1.85	**49.49 ± 0.04**	77.25 ± 4.43	9.32 ± 5.65	**49.49 ± 0.04**	79.95 ± 4.55	9.13 ± 3.42	**49.49 ± 0.04**	76.21 ± 6.80	8.02 ± 1.12	**49.49 ± 0.04**	72.97 ± 4.28	7.39 ± 3.14	**49.49 ± 0.04**	**80.65 ± 7.78**	**9.91 ± 1.24**	**49.49 ± 0.04**
madin_carbsubs_alpha-ketovaleric acid	61.77 ± 1.55	3.80 ± 0.89	**49.60 ± 0.20**	49.32 ± 8.16	1.89 ± 0.29	**49.60 ± 0.20**	56.45 ± 10.22	3.50 ± 1.87	**49.60 ± 0.20**	52.65 ± 5.11	2.24 ± 0.72	**49.60 ± 0.20**	**62.88 ± 2.89**	4.12 ± 2.51	**49.60 ± 0.20**	52.35 ± 7.85	2.35 ± 0.55	**49.60 ± 0.20**	60.26 ± 14.02	2.73 ± 0.39	**49.60 ± 0.20**	56.47 ± 6.49	2.47 ± 0.96	**49.60 ± 0.20**	62.66 ± 5.97	**4.27 ± 3.41**	**49.60 ± 0.20**
madin_carbsubs_arabinose	58.52 ± 0.82	47.12 ± 5.31	42.38 ± 8.83	61.08 ± 0.53	48.02 ± 2.50	45.95 ± 10.78	58.07 ± 1.71	46.75 ± 1.15	43.96 ± 11.56	62.46 ± 2.52	50.08 ± 3.21	44.25 ± 7.53	68.37 ± 0.81	56.23 ± 2.70	47.69 ± 8.46	64.99 ± 3.56	52.22 ± 4.12	50.55 ± 4.91	68.18 ± 2.23	53.78 ± 6.04	56.97 ± 6.44	60.98 ± 4.35	48.66 ± 5.38	49.49 ± 9.36	**73.46 ± 1.22**	**61.05 ± 1.38**	**63.85 ± 3.64**
madin_carbsubs_arabitol	55.85 ± 1.26	8.93 ± 1.40	**48.21 ± 0.24**	54.97 ± 2.44	8.64 ± 1.10	**48.21 ± 0.24**	56.36 ± 3.77	8.02 ± 1.41	**48.21 ± 0.24**	57.76 ± 3.59	**10.39 ± 1.08**	**48.21 ± 0.24**	54.68 ± 3.58	7.80 ± 0.89	**48.21 ± 0.24**	53.40 ± 2.11	7.94 ± 0.52	**48.21 ± 0.24**	57.57 ± 4.39	9.48 ± 2.65	48.19 ± 0.25	**58.27 ± 3.36**	10.31 ± 0.79	**48.21 ± 0.24**	57.63 ± 2.32	8.29 ± 1.25	**48.21 ± 0.24**
madin_carbsubs_arginine	58.62 ± 5.68	7.82 ± 1.48	**48.49 ± 0.26**	57.69 ± 2.12	7.01 ± 1.05	**48.49 ± 0.26**	57.62 ± 0.60	8.41 ± 1.07	**48.49 ± 0.26**	61.55 ± 4.06	9.68 ± 3.62	**48.49 ± 0.26**	**64.14 ± 4.29**	9.66 ± 1.36	48.45 ± 0.26	56.46 ± 10.85	8.51 ± 3.50	48.48 ± 0.29	51.68 ± 15.68	7.26 ± 1.42	48.48 ± 0.25	59.87 ± 2.58	**9.78 ± 3.50**	**48.49 ± 0.26**	57.04 ± 5.57	8.14 ± 2.28	**48.49 ± 0.26**
madin_carbsubs_asparagine	53.33 ± 8.71	8.75 ± 1.83	**48.11 ± 0.19**	62.07 ± 0.74	13.22 ± 1.14	**48.11 ± 0.19**	60.38 ± 6.61	13.10 ± 1.82	**48.11 ± 0.19**	**65.08 ± 1.51**	**13.73 ± 1.05**	**48.11 ± 0.19**	63.30 ± 2.38	12.50 ± 1.61	**48.11 ± 0.19**	62.38 ± 2.00	10.97 ± 0.94	**48.11 ± 0.19**	58.17 ± 9.08	11.86 ± 2.69	**48.11 ± 0.19**	62.79 ± 2.42	12.40 ± 1.56	**48.11 ± 0.19**	61.18 ± 3.44	10.27 ± 1.75	48.05 ± 0.28
madin_carbsubs_aspartate	57.58 ± 7.37	13.55 ± 2.68	47.27 ± 0.17	64.32 ± 3.18	16.83 ± 2.31	47.27 ± 0.17	61.26 ± 8.81	15.53 ± 4.12	47.27 ± 0.17	65.94 ± 2.15	17.02 ± 0.95	**49.52 ± 2.13**	67.16 ± 1.72	15.90 ± 2.58	47.27 ± 0.17	**67.48 ± 1.19**	15.18 ± 1.24	47.27 ± 0.17	66.60 ± 6.58	**19.28 ± 5.94**	47.27 ± 0.17	67.33 ± 0.94	18.82 ± 1.80	47.27 ± 0.17	66.72 ± 3.47	15.91 ± 3.56	47.27 ± 0.17

Supplementary Table 18: Per-phenotype performance (2/2). Values are mean ± standard deviation across 5 random seeds. Bold highlights the best mean per metric. The results were macro-averaged across classes for each phenotype.

Phenotype	Mistral-DNA AUROC	Mistral-DNA AUPRC	Mistral-DNA F1	DNABERT-2 AUROC	DNABERT-2 AUPRC	DNABERT-2 F1	Nucleotide Transformer AUROC	Nucleotide Transformer AUPRC	Nucleotide Transformer F1
madin_carbsubs_caprate	68.67 ± 18.55	4.91 ± 4.03	**49.53 ± 0.18**	80.05 ± 4.66	7.98 ± 4.57	**49.53 ± 0.18**	78.69 ± 11.25	10.79 ± 9.42	**49.53 ± 0.18**
madin_carbsubs_cellobiose	61.76 ± 2.48	42.87 ± 2.60	41.28 ± 2.38	61.98 ± 3.47	42.19 ± 3.58	40.40 ± 0.89	61.67 ± 2.17	40.63 ± 4.66	**53.39 ± 8.56**
madin_carbsubs_citrate	64.02 ± 4.53	22.50 ± 1.01	46.02 ± 0.38	66.84 ± 2.76	25.25 ± 2.20	46.63 ± 0.95	64.85 ± 4.41	24.93 ± 3.17	48.21 ± 1.87
madin_carbsubs_dextrin	49.78 ± 3.75	8.80 ± 1.28	**47.81 ± 0.31**	53.60 ± 2.15	11.14 ± 0.23	**47.81 ± 0.31**	56.51 ± 1.77	10.46 ± 1.53	47.75 ± 0.34
madin_carbsubs_ethanol	61.06 ± 3.96	5.24 ± 0.73	49.06 ± 0.17	55.11 ± 15.45	4.59 ± 1.13	49.06 ± 0.17	66.31 ± 5.40	9.76 ± 2.65	49.06 ± 0.17
madin_carbsubs_formic acid	58.21 ± 13.07	6.35 ± 2.00	48.78 ± 0.15	70.35 ± 6.84	11.23 ± 3.63	48.78 ± 0.15	69.32 ± 5.80	13.75 ± 3.60	49.76 ± 1.58
madin_carbsubs_fructose	53.77 ± 1.57	43.72 ± 2.98	40.88 ± 7.24	59.90 ± 1.71	49.39 ± 2.50	46.99 ± 8.48	57.79 ± 1.71	48.49 ± 2.14	52.22 ± 1.47
madin_carbsubs_fucose	50.66 ± 8.81	10.81 ± 5.73	**47.71 ± 0.29**	53.74 ± 7.43	11.77 ± 3.38	**47.71 ± 0.29**	52.98 ± 14.69	10.47 ± 2.59	**47.71 ± 0.29**
madin_carbsubs_galactose	55.32 ± 4.32	39.71 ± 4.34	39.39 ± 0.47	58.12 ± 3.00	42.87 ± 2.01	39.39 ± 0.47	55.09 ± 3.05	39.40 ± 4.05	42.09 ± 2.41
madin_carbsubs_galacturonic acid	62.04 ± 11.24	**6.60 ± 5.23**	**49.41 ± 0.08**	**65.17 ± 0.37**	4.57 ± 0.81	**49.41 ± 0.08**	50.76 ± 11.37	2.62 ± 1.00	**49.41 ± 0.08**
madin_carbsubs_gelatin	51.62 ± 6.68	5.74 ± 1.55	**48.69 ± 0.16**	56.85 ± 0.98	6.62 ± 1.78	**48.69 ± 0.16**	56.25 ± 0.60	7.24 ± 0.93	**48.69 ± 0.16**
madin_carbsubs_gluconic acid	57.20 ± 1.35	17.04 ± 3.52	**46.32 ± 0.97**	60.71 ± 2.60	18.41 ± 4.89	**46.32 ± 0.97**	63.29 ± 3.16	20.64 ± 5.38	**46.32 ± 0.97**
madin_carbsubs_glucosamine	57.91 ± 4.33	3.35 ± 1.35	**49.40 ± 0.16**	51.11 ± 7.38	2.65 ± 0.04	**49.40 ± 0.16**	57.14 ± 6.10	3.22 ± 1.31	**49.40 ± 0.16**
madin_carbsubs_glucose	49.81 ± 5.94	63.97 ± 3.80	43.30 ± 6.99	53.76 ± 7.58	67.77 ± 6.49	40.45 ± 1.72	54.04 ± 5.73	67.56 ± 5.49	38.94 ± 10.45
madin_carbsubs_glucuronamide	42.39 ± 7.20	3.17 ± 3.12	**49.56 ± 0.04**	56.89 ± 4.00	4.05 ± 2.14	**49.56 ± 0.04**	58.52 ± 10.28	2.70 ± 1.24	**49.56 ± 0.04**
madin_carbsubs_glutamate	57.83 ± 2.42	18.00 ± 1.24	46.27 ± 0.31	62.24 ± 3.57	19.59 ± 1.55	46.27 ± 0.31	64.48 ± 8.41	22.38 ± 3.75	**48.76 ± 4.52**
madin_carbsubs_glycerol	53.94 ± 7.11	26.43 ± 3.33	43.53 ± 0.79	53.09 ± 5.79	25.09 ± 1.96	43.53 ± 0.79	56.65 ± 6.17	26.95 ± 3.04	45.21 ± 3.71
madin_carbsubs_glycine	34.86 ± 3.88	2.20 ± 0.62	**49.26 ± 0.19**	40.22 ± 8.03	2.56 ± 0.96	**49.26 ± 0.19**	**53.94 ± 4.39**	**3.39 ± 0.81**	**49.26 ± 0.19**
madin_carbsubs_glycogen	51.01 ± 4.95	12.66 ± 2.23	**47.05 ± 0.18**	51.58 ± 1.47	12.29 ± 1.19	**47.05 ± 0.18**	53.26 ± 5.00	12.29 ± 1.12	47.02 ± 0.15
madin_carbsubs_histidine	59.81 ± 6.43	11.36 ± 1.34	**48.01 ± 0.09**	58.96 ± 4.69	10.82 ± 1.58	**48.01 ± 0.09**	59.73 ± 5.35	11.31 ± 0.77	**48.01 ± 0.09**
madin_carbsubs_inosine	**63.03 ± 3.43**	5.17 ± 2.32	**49.24 ± 0.12**	57.88 ± 15.05	5.01 ± 2.55	**49.24 ± 0.12**	59.65 ± 2.11	6.33 ± 1.26	**49.24 ± 0.12**
madin_carbsubs_lactate	67.30 ± 3.44	29.54 ± 3.48	45.60 ± 0.37	67.55 ± 4.35	29.39 ± 1.87	45.60 ± 0.37	67.88 ± 4.79	29.07 ± 2.83	49.80 ± 7.70
madin_carbsubs_lactose	61.68 ± 4.23	33.46 ± 8.41	42.68 ± 1.15	64.84 ± 2.44	38.00 ± 5.41	43.45 ± 2.43	64.72 ± 3.63	37.65 ± 9.32	54.06 ± 5.30
madin_carbsubs_leucine	47.12 ± 3.82	7.08 ± 0.82	**48.49 ± 0.03**	56.98 ± 4.33	9.22 ± 1.36	**48.49 ± 0.03**	57.20 ± 4.35	9.73 ± 1.21	**48.49 ± 0.03**
madin_carbsubs_malate	64.71 ± 8.27	25.38 ± 4.53	45.30 ± 0.36	67.92 ± 5.16	26.94 ± 1.88	45.30 ± 0.36	68.46 ± 4.12	28.90 ± 4.45	**50.52 ± 4.62**
madin_carbsubs_malonate	67.05 ± 3.36	10.85 ± 1.42	**48.63 ± 0.35**	**70.79 ± 0.34**	10.56 ± 2.50	**48.63 ± 0.35**	62.29 ± 5.83	7.51 ± 1.31	**48.63 ± 0.35**
madin_carbsubs_maltose	55.89 ± 2.91	52.85 ± 0.72	55.50 ± 1.86	58.77 ± 1.45	54.23 ± 1.76	48.40 ± 5.67	57.46 ± 1.11	54.64 ± 4.70	46.99 ± 10.75
madin_carbsubs_mannitol	57.55 ± 1.94	41.59 ± 1.71	39.44 ± 0.64	61.56 ± 1.03	46.11 ± 2.33	40.56 ± 1.71	61.32 ± 1.17	45.32 ± 1.01	48.86 ± 9.11
madin_carbsubs_mannose	55.68 ± 4.78	48.29 ± 4.04	46.25 ± 6.97	59.60 ± 3.95	52.59 ± 4.45	45.86 ± 10.12	59.11 ± 1.80	53.56 ± 5.39	49.90 ± 12.19
madin_carbsubs_melibiose	58.85 ± 4.97	27.03 ± 6.20	44.38 ± 0.68	60.85 ± 3.77	30.28 ± 4.07	44.38 ± 0.68	62.47 ± 3.08	32.17 ± 3.86	48.75 ± 4.58
madin_carbsubs_methyl beta-D-glucoside	59.86 ± 5.82	5.36 ± 2.26	**49.04 ± 0.40**	60.17 ± 7.39	7.49 ± 1.67	**49.04 ± 0.40**	58.33 ± 6.62	6.12 ± 1.69	**49.04 ± 0.40**
madin_carbsubs_methyl pyruvate	56.03 ± 4.29	3.02 ± 1.50	**49.39 ± 0.32**	64.84 ± 5.97	3.83 ± 1.76	**49.39 ± 0.32**	63.99 ± 3.53	8.76 ± 2.08	**49.39 ± 0.32**
madin_carbsubs_myo-inositol	60.51 ± 3.95	9.03 ± 3.21	48.20 ± 0.56	51.61 ± 9.17	7.35 ± 3.18	48.20 ± 0.56	55.33 ± 5.32	10.01 ± 3.32	48.20 ± 0.56
madin_carbsubs_ornithine	57.04 ± 11.29	6.13 ± 3.42	**48.88 ± 0.29**	59.18 ± 12.41	6.80 ± 3.65	**48.88 ± 0.29**	68.87 ± 5.95	8.96 ± 3.74	**48.88 ± 0.29**
madin_carbsubs_phenylacetate	62.53 ± 16.32	7.13 ± 7.48	**49.24 ± 0.42**	56.44 ± 11.60	3.70 ± 1.83	**49.24 ± 0.42**	65.48 ± 15.58	5.90 ± 4.90	**49.24 ± 0.42**
madin_carbsubs_phenylalanine	66.43 ± 2.81	5.14 ± 0.28	**49.18 ± 0.13**	65.15 ± 2.58	5.31 ± 0.26	**49.18 ± 0.13**	64.65 ± 4.27	6.46 ± 1.41	**49.18 ± 0.13**
madin_carbsubs_proline	55.28 ± 8.35	16.10 ± 4.93	46.36 ± 0.21	61.20 ± 3.10	20.07 ± 2.36	46.36 ± 0.21	59.26 ± 7.02	18.85 ± 5.11	46.36 ± 0.21
madin_carbsubs_propionate	57.31 ± 9.24	13.67 ± 4.50	47.11 ± 0.39	59.01 ± 8.02	14.58 ± 5.38	47.11 ± 0.39	59.93 ± 3.79	14.74 ± 1.48	47.27 ± 0.53
madin_carbsubs_putrescine	44.89 ± 6.05	1.78 ± 0.40	**49.52 ± 0.17**	52.00 ± 7.25	2.35 ± 1.41	**49.52 ± 0.17**	50.27 ± 8.03	2.20 ± 0.89	**49.52 ± 0.17**
madin_carbsubs_pyruvate	59.60 ± 4.16	21.41 ± 1.98	46.43 ± 0.18	64.22 ± 3.53	23.27 ± 3.06	46.81 ± 0.75	62.52 ± 2.27	22.62 ± 3.76	50.56 ± 3.73
madin_carbsubs_pyruvic acid methyl ester	61.46 ± 10.82	9.07 ± 6.67	**49.02 ± 0.16**	63.90 ± 10.25	8.73 ± 4.78	**49.02 ± 0.16**	62.50 ± 6.96	6.35 ± 1.82	48.92 ± 0.02
madin_carbsubs_quinic acid	53.15 ± 15.96	7.35 ± 9.10	**49.59 ± 0.08**	53.57 ± 14.93	**7.84 ± 10.21**	**49.59 ± 0.08**	50.59 ± 9.66	1.90 ± 0.22	**49.59 ± 0.08**
madin_carbsubs_raffinose	63.10 ± 6.30	31.16 ± 6.18	43.87 ± 0.58	65.04 ± 7.99	34.35 ± 10.88	43.84 ± 0.61	66.74 ± 3.10	32.96 ± 4.62	43.85 ± 0.55
madin_carbsubs_rhamnose	67.25 ± 3.98	42.01 ± 1.83	43.18 ± 1.61	67.97 ± 5.24	42.40 ± 2.49	43.09 ± 1.19	69.65 ± 4.48	42.43 ± 6.55	43.10 ± 0.91
madin_carbsubs_ribose	57.83 ± 1.42	26.99 ± 2.55	43.97 ± 0.66	62.24 ± 5.97	29.13 ± 2.06	43.97 ± 0.66	65.08 ± 1.40	32.17 ± 1.43	45.91 ± 2.76
madin_carbsubs_salicin	60.65 ± 4.60	25.00 ± 2.50	45.78 ± 1.01	62.80 ± 0.66	23.64 ± 2.97	45.21 ± 0.68	67.01 ± 8.32	29.18 ± 11.28	**51.60 ± 5.79**
madin_carbsubs_serine	58.47 ± 4.23	15.79 ± 2.42	46.98 ± 0.44	61.73 ± 2.38	17.69 ± 3.27	46.98 ± 0.44	62.35 ± 2.00	18.54 ± 2.03	**47.83 ± 1.58**
madin_carbsubs_sorbitol	54.89 ± 4.92	18.23 ± 0.73	**45.29 ± 0.43**	55.93 ± 3.58	20.44 ± 1.29	**45.29 ± 0.43**	55.67 ± 5.39	19.70 ± 2.71	**45.29 ± 0.43**
madin_carbsubs_succinate	66.04 ± 4.27	27.28 ± 0.34	45.45 ± 0.15	68.45 ± 1.64	27.33 ± 1.97	45.45 ± 0.15	66.31 ± 3.47	29.69 ± 2.40	**49.39 ± 3.08**
madin_carbsubs_sucrose	58.76 ± 5.88	52.39 ± 7.55	45.14 ± 6.73	60.52 ± 6.59	54.04 ± 11.22	45.71 ± 11.83	62.82 ± 2.84	56.34 ± 7.85	**58.24 ± 3.45**
madin_carbsubs_threonine	54.40 ± 2.07	9.27 ± 1.53	48.29 ± 0.23	54.85 ± 3.16	8.50 ± 2.46	48.29 ± 0.23	55.45 ± 1.92	9.58 ± 2.90	48.29 ± 0.23
madin_carbsubs_thymidine	51.90 ± 5.12	2.28 ± 0.58	**49.46 ± 0.09**	44.26 ± 4.51	2.03 ± 0.46	**49.46 ± 0.09**	**56.60 ± 5.56**	**3.97 ± 1.70**	**49.46 ± 0.09**
madin_carbsubs_trehalose	57.41 ± 2.00	36.34 ± 3.38	41.70 ± 1.68	54.61 ± 0.18	33.72 ± 2.62	40.94 ± 1.09	58.17 ± 4.23	38.10 ± 6.12	44.55 ± 2.40
madin_carbsubs_urea	37.59 ± 6.62	1.74 ± 0.39	**49.46 ± 0.12**	48.06 ± 5.35	2.45 ± 1.00	**49.46 ± 0.12**	49.89 ± 9.86	2.39 ± 0.57	**49.46 ± 0.12**
madin_carbsubs_uridine	62.60 ± 3.69	**5.14 ± 0.51**	**49.21 ± 0.13**	60.27 ± 4.46	4.71 ± 1.48	**49.21 ± 0.13**	57.36 ± 1.67	5.02 ± 1.05	**49.21 ± 0.13**
madin_carbsubs_valerate	59.41 ± 2.81	3.73 ± 1.73	**49.35 ± 0.31**	55.52 ± 5.15	5.28 ± 2.58	**49.35 ± 0.31**	60.77 ± 2.94	5.11 ± 2.36	**49.35 ± 0.31**
madin_carbsubs_xylitol	50.19 ± 4.95	4.76 ± 0.27	**48.81 ± 0.12**	50.24 ± 7.82	5.50 ± 0.56	**48.81 ± 0.12**	50.26 ± 6.86	5.02 ± 0.73	**48.81 ± 0.12**
madin_carbsubs_xylose	65.29 ± 1.53	43.41 ± 1.49	41.67 ± 0.81	66.69 ± 3.10	45.09 ± 2.36	41.67 ± 0.81	65.36 ± 2.26	41.15 ± 2.52	47.80 ± 11.42
madin_categorical_gram_stain	87.51 ± 6.69	76.25 ± 14.86	78.98 ± 8.20	96.71 ± 1.37	93.15 ± 3.14	91.68 ± 2.64	98.24 ± 0.20	96.72 ± 0.63	95.25 ± 1.28
madin_categorical_metabolism	68.09 ± 0.56	32.14 ± 2.82	24.91 ± 6.71	75.99 ± 2.04	38.55 ± 3.42	33.08 ± 1.15	78.81 ± 2.89	41.98 ± 3.42	35.66 ± 6.54
madin_categorical_motility_binary	65.66 ± 3.38	58.37 ± 15.70	63.57 ± 4.18	64.77 ± 3.61	59.78 ± 12.23	58.77 ± 5.81	73.12 ± 9.84	71.49 ± 3.32	64.51 ± 4.45
madin_categorical_sporulation	76.80 ± 2.54	38.90 ± 11.24	54.69 ± 7.16	92.95 ± 3.15	68.42 ± 19.89	69.52 ± 15.64	95.78 ± 1.51	83.02 ± 8.17	89.45 ± 2.93

Phenotype	ProkBERT AUROC	ProkBERT AUPRC	ProkBERT F1	ESM-2 AUROC	ESM-2 AUPRC	ESM-2 F1	ESM-C AUROC	ESM-C AUPRC	ESM-C F1
madin_carbsubs_caprate	76.79 ± 13.71	10.22 ± 8.00	**49.53 ± 0.18**	84.02 ± 6.90	14.20 ± 8.97	49.47 ± 0.16	85.20 ± 5.81	11.21 ± 7.49	**49.53 ± 0.18**
madin_carbsubs_cellobiose	61.85 ± 7.56	42.32 ± 8.10	46.20 ± 5.28	64.32 ± 5.24	**44.50 ± 6.77**	48.65 ± 4.62	62.06 ± 2.53	40.64 ± 3.97	43.35 ± 3.52
madin_carbsubs_citrate	67.67 ± 5.67	32.09 ± 3.96	46.39 ± 0.53	68.26 ± 4.12	30.04 ± 10.84	48.16 ± 3.55	68.19 ± 5.36	30.17 ± 12.13	46.02 ± 0.38
madin_carbsubs_dextrin	**60.70 ± 2.07**	12.83 ± 1.63	**47.81 ± 0.31**	58.12 ± 1.57	11.46 ± 3.64	**47.81 ± 0.31**	59.64 ± 0.68	12.11 ± 2.12	**47.81 ± 0.31**
madin_carbsubs_ethanol	65.93 ± 8.80	13.89 ± 14.99	49.06 ± 0.17	73.30 ± 12.74	10.84 ± 3.16	49.06 ± 0.17	73.45 ± 4.49	9.84 ± 1.92	49.06 ± 0.17
madin_carbsubs_formic acid	63.06 ± 15.08	8.25 ± 2.65	48.78 ± 0.15	74.65 ± 3.67	11.81 ± 1.53	48.77 ± 0.16	75.28 ± 2.98	11.84 ± 1.78	48.78 ± 0.15
madin_carbsubs_fructose	60.24 ± 3.13	51.70 ± 1.46	47.86 ± 9.29	63.47 ± 1.88	53.61 ± 0.69	**55.70 ± 4.69**	63.77 ± 2.40	54.31 ± 1.77	53.28 ± 4.31
madin_carbsubs_fucose	57.24 ± 6.22	12.90 ± 3.37	**47.71 ± 0.29**	**66.42 ± 3.95**	**18.43 ± 6.14**	**47.71 ± 0.29**	63.55 ± 2.79	13.40 ± 2.06	**47.71 ± 0.29**
madin_carbsubs_galactose	61.21 ± 2.21	44.61 ± 3.10	46.10 ± 3.71	62.94 ± 3.26	44.54 ± 3.13	45.76 ± 3.58	61.76 ± 1.92	45.08 ± 5.33	41.75 ± 3.62
madin_carbsubs_galacturonic acid	53.91 ± 7.85	3.18 ± 0.95	**49.41 ± 0.08**	48.75 ± 12.17	3.26 ± 1.37	**49.41 ± 0.08**	60.27 ± 19.02	5.23 ± 3.00	**49.41 ± 0.08**
madin_carbsubs_gelatin	58.54 ± 5.59	7.00 ± 0.59	**48.69 ± 0.16**	51.45 ± 6.22	5.60 ± 1.49	48.63 ± 0.07	**61.72 ± 6.27**	7.24 ± 0.70	**48.69 ± 0.16**
madin_carbsubs_gluconic acid	62.53 ± 6.26	22.93 ± 6.92	**46.32 ± 0.97**	61.08 ± 3.84	18.47 ± 8.26	**46.32 ± 0.97**	63.60 ± 6.17	22.76 ± 7.39	**46.32 ± 0.97**
madin_carbsubs_glucosamine	58.52 ± 8.40	3.66 ± 0.97	**49.40 ± 0.16**	61.62 ± 1.08	4.87 ± 0.88	**49.40 ± 0.16**	57.94 ± 8.00	3.88 ± 1.48	**49.40 ± 0.16**
madin_carbsubs_glucose	51.91 ± 8.63	65.74 ± 6.61	43.27 ± 7.69	59.15 ± 3.74	72.53 ± 3.03	54.78 ± 2.12	58.44 ± 3.22	71.42 ± 1.50	49.16 ± 6.42
madin_carbsubs_glucuronamide	53.00 ± 7.24	3.28 ± 2.26	**49.56 ± 0.04**	49.67 ± 17.08	2.13 ± 1.24	**49.56 ± 0.04**	**70.96 ± 6.30**	**4.72 ± 2.17**	**49.56 ± 0.04**
madin_carbsubs_glutamate	69.27 ± 3.06	**24.58 ± 0.90**	46.75 ± 0.51	66.30 ± 1.26	20.78 ± 1.79	46.65 ± 0.92	65.23 ± 7.23	21.52 ± 5.53	46.27 ± 0.31
madin_carbsubs_glycerol	59.31 ± 7.53	30.07 ± 4.87	**45.45 ± 3.14**	58.99 ± 7.07	30.05 ± 2.30	43.53 ± 0.79	**62.80 ± 4.85**	**32.42 ± 3.85**	45.25 ± 1.26
madin_carbsubs_glycine	39.24 ± 3.22	2.36 ± 0.67	**49.26 ± 0.19**	46.49 ± 12.05	2.90 ± 1.22	49.24 ± 0.17	52.86 ± 9.43	3.25 ± 0.89	**49.26 ± 0.19**
madin_carbsubs_glycogen	59.18 ± 2.54	14.87 ± 3.05	**47.05 ± 0.18**	**62.89 ± 4.27**	**16.46 ± 1.69**	**47.05 ± 0.18**	61.90 ± 7.17	16.23 ± 1.33	**47.05 ± 0.18**
madin_carbsubs_histidine	59.85 ± 5.75	**13.30 ± 3.28**	**48.01 ± 0.09**	59.30 ± 3.90	11.47 ± 1.68	**48.01 ± 0.09**	57.65 ± 7.72	10.57 ± 1.94	**48.01 ± 0.09**
madin_carbsubs_inosine	57.16 ± 1.06	4.54 ± 1.35	**49.24 ± 0.12**	62.84 ± 5.40	6.06 ± 2.52	**49.24 ± 0.12**	57.38 ± 9.07	6.20 ± 4.21	**49.24 ± 0.12**
madin_carbsubs_lactate	**68.68 ± 4.59**	**32.63 ± 4.21**	45.60 ± 0.37	67.53 ± 5.23	28.30 ± 4.15	45.60 ± 0.37	67.89 ± 6.34	29.96 ± 5.05	47.88 ± 4.30
madin_carbsubs_lactose	66.33 ± 2.61	42.95 ± 4.25	48.11 ± 3.78	68.90 ± 2.70	43.82 ± 6.61	**60.90 ± 1.60**	68.01 ± 3.56	44.30 ± 5.74	50.62 ± 6.45
madin_carbsubs_leucine	57.81 ± 4.76	9.05 ± 3.10	**48.49 ± 0.03**	59.81 ± 2.11	8.69 ± 0.33	**48.49 ± 0.03**	60.94 ± 2.42	8.79 ± 0.70	**48.49 ± 0.03**
madin_carbsubs_malate	69.27 ± 4.55	31.78 ± 8.91	48.94 ± 2.81	72.60 ± 5.35	33.97 ± 5.47	45.30 ± 0.36	72.99 ± 3.65	33.40 ± 5.06	45.62 ± 0.91
madin_carbsubs_malonate	64.71 ± 5.91	11.71 ± 4.85	**48.63 ± 0.35**	69.07 ± 4.32	10.57 ± 1.74	**48.63 ± 0.35**	68.28 ± 4.67	10.71 ± 1.15	**48.63 ± 0.35**
madin_carbsubs_maltose	58.90 ± 2.78	56.34 ± 4.55	48.71 ± 3.11	61.21 ± 1.34	57.87 ± 4.00	51.02 ± 2.22	61.89 ± 5.68	60.04 ± 6.63	**58.75 ± 6.37**
madin_carbsubs_mannitol	66.28 ± 2.79	50.09 ± 3.84	51.89 ± 4.48	67.61 ± 3.56	49.57 ± 2.19	49.42 ± 11.38	68.47 ± 2.53	51.43 ± 1.97	48.50 ± 4.79
madin_carbsubs_mannose	61.85 ± 1.76	56.77 ± 4.46	41.17 ± 12.30	65.32 ± 1.61	60.17 ± 1.14	58.83 ± 3.86	63.98 ± 2.02	59.88 ± 2.50	58.25 ± 3.72
madin_carbsubs_melibiose	68.23 ± 3.60	42.34 ± 6.67	50.69 ± 6.07	70.32 ± 2.31	43.88 ± 5.32	47.55 ± 6.16	70.87 ± 3.45	44.35 ± 5.45	48.49 ± 4.69
madin_carbsubs_methyl beta-D-glucoside	64.91 ± 8.97	10.21 ± 5.55	**49.04 ± 0.40**	**65.51 ± 0.78**	6.79 ± 3.06	**49.04 ± 0.40**	59.40 ± 10.40	6.86 ± 2.70	**49.04 ± 0.40**
madin_carbsubs_methyl pyruvate	64.92 ± 16.08	**9.28 ± 7.45**	**49.39 ± 0.32**	**67.61 ± 9.27**	4.29 ± 1.52	**49.39 ± 0.32**	61.38 ± 12.90	3.34 ± 1.42	**49.39 ± 0.32**
madin_carbsubs_myo-inositol	66.59 ± 6.33	16.10 ± 3.79	48.20 ± 0.56	54.32 ± 7.51	9.30 ± 5.15	48.20 ± 0.56	53.07 ± 13.71	9.64 ± 4.30	48.20 ± 0.56
madin_carbsubs_ornithine	71.02 ± 5.71	8.08 ± 3.12	**48.88 ± 0.29**	71.28 ± 5.19	10.43 ± 3.20	**48.88 ± 0.29**	73.00 ± 4.99	9.83 ± 2.43	**48.88 ± 0.29**
madin_carbsubs_phenylacetate	**72.21 ± 7.69**	**12.09 ± 14.75**	49.23 ± 0.40	58.21 ± 13.94	4.09 ± 2.07	**49.24 ± 0.42**	66.97 ± 16.96	9.20 ± 9.61	**49.24 ± 0.42**
madin_carbsubs_phenylalanine	**67.48 ± 3.25**	7.54 ± 1.78	**49.18 ± 0.13**	66.32 ± 1.59	5.66 ± 1.19	**49.18 ± 0.13**	64.49 ± 5.95	**8.09 ± 5.51**	**49.18 ± 0.13**
madin_carbsubs_proline	64.12 ± 1.83	20.64 ± 2.95	46.36 ± 0.21	65.44 ± 2.46	20.86 ± 1.11	46.34 ± 0.20	66.83 ± 2.17	21.27 ± 1.63	46.36 ± 0.21
madin_carbsubs_propionate	60.34 ± 4.22	17.29 ± 5.81	47.11 ± 0.39	56.56 ± 9.21	13.08 ± 4.04	47.11 ± 0.39	64.32 ± 2.41	15.90 ± 0.33	47.11 ± 0.39
madin_carbsubs_putrescine	**52.64 ± 14.29**	2.75 ± 0.83	**49.52 ± 0.17**	46.98 ± 6.62	2.06 ± 0.79	**49.52 ± 0.17**	49.74 ± 12.46	2.83 ± 2.08	**49.52 ± 0.17**
madin_carbsubs_pyruvate	67.96 ± 1.61	26.52 ± 1.77	51.23 ± 0.80	71.55 ± 3.43	29.67 ± 1.20	49.12 ± 1.28	71.36 ± 2.73	29.00 ± 3.74	48.86 ± 2.15
madin_carbsubs_pyruvic acid methyl ester	65.28 ± 4.64	**12.34 ± 8.18**	**49.02 ± 0.16**	62.44 ± 10.95	8.04 ± 4.48	**49.02 ± 0.16**	68.81 ± 7.34	10.26 ± 4.21	**49.02 ± 0.16**
madin_carbsubs_quinic acid	59.21 ± 3.90	4.61 ± 3.04	**49.59 ± 0.08**	53.52 ± 16.07	2.56 ± 1.36	**49.59 ± 0.08**	60.15 ± 18.80	2.46 ± 0.82	**49.59 ± 0.08**
madin_carbsubs_raffinose	65.42 ± 7.31	35.51 ± 8.34	**50.23 ± 5.37**	71.19 ± 3.71	41.14 ± 7.03	47.55 ± 3.10	65.73 ± 7.62	33.58 ± 11.30	46.14 ± 3.72
madin_carbsubs_rhamnose	71.36 ± 0.93	41.60 ± 4.79	49.44 ± 9.04	72.31 ± 3.91	44.06 ± 10.12	43.28 ± 0.69	72.70 ± 0.57	**44.16 ± 8.29**	43.30 ± 0.72
madin_carbsubs_ribose	64.61 ± 4.65	32.49 ± 4.46	45.42 ± 1.89	66.75 ± 3.18	31.76 ± 0.84	44.41 ± 1.34	65.52 ± 3.56	34.01 ± 4.08	44.24 ± 1.11
madin_carbsubs_salicin	69.68 ± 2.62	29.99 ± 7.38	45.76 ± 1.12	70.42 ± 2.82	31.13 ± 5.47	45.51 ± 0.71	**73.07 ± 4.39**	**34.02 ± 9.67**	45.74 ± 1.02
madin_carbsubs_serine	66.21 ± 2.63	20.60 ± 3.51	46.98 ± 0.44	65.48 ± 1.62	18.29 ± 3.20	46.98 ± 0.44	67.21 ± 2.15	20.79 ± 5.36	46.98 ± 0.44
madin_carbsubs_sorbitol	57.83 ± 4.01	**23.80 ± 1.27**	45.25 ± 0.39	59.62 ± 3.63	21.95 ± 2.03	**45.29 ± 0.43**	55.05 ± 7.73	19.64 ± 2.02	45.25 ± 0.39
madin_carbsubs_succinate	69.03 ± 2.12	**34.26 ± 7.19**	48.50 ± 3.70	69.38 ± 4.09	28.73 ± 5.02	45.45 ± 0.15	70.03 ± 3.56	30.38 ± 3.10	45.45 ± 0.15
madin_carbsubs_sucrose	60.94 ± 3.59	54.53 ± 5.34	48.51 ± 12.23	66.26 ± 4.07	**60.73 ± 7.05**	57.08 ± 7.78	63.46 ± 7.06	55.71 ± 11.34	48.71 ± 9.70
madin_carbsubs_threonine	58.47 ± 1.10	11.10 ± 2.13	48.29 ± 0.23	56.76 ± 2.34	9.00 ± 1.52	48.29 ± 0.23	57.09 ± 3.63	9.09 ± 0.72	48.29 ± 0.23
madin_carbsubs_thymidine	53.62 ± 12.10	2.78 ± 1.18	**49.46 ± 0.09**	54.95 ± 2.55	2.53 ± 0.34	**49.46 ± 0.09**	42.61 ± 4.86	1.98 ± 0.24	**49.46 ± 0.09**
madin_carbsubs_trehalose	61.06 ± 2.76	41.33 ± 5.35	49.75 ± 5.24	**63.65 ± 2.27**	**43.39 ± 4.23**	45.23 ± 1.94	63.21 ± 3.34	43.07 ± 6.30	47.00 ± 6.62
madin_carbsubs_urea	36.66 ± 9.77	1.81 ± 0.13	**49.46 ± 0.12**	47.46 ± 11.56	2.15 ± 0.17	**49.46 ± 0.12**	47.58 ± 9.26	2.17 ± 0.19	**49.46 ± 0.12**
madin_carbsubs_uridine	58.68 ± 2.98	4.24 ± 0.98	**49.21 ± 0.13**	**63.29 ± 2.21**	4.98 ± 0.70	**49.21 ± 0.13**	56.94 ± 4.66	4.52 ± 1.57	**49.21 ± 0.13**
madin_carbsubs_valerate	58.51 ± 3.52	3.97 ± 1.41	**49.35 ± 0.31**	60.27 ± 5.90	4.43 ± 2.11	**49.35 ± 0.31**	62.90 ± 7.86	4.89 ± 2.37	**49.35 ± 0.31**
madin_carbsubs_xylitol	**54.22 ± 3.70**	6.38 ± 1.16	**48.81 ± 0.12**	51.09 ± 8.30	5.47 ± 0.45	**48.81 ± 0.12**	49.04 ± 4.20	4.77 ± 0.07	**48.81 ± 0.12**
madin_carbsubs_xylose	64.99 ± 2.31	41.14 ± 1.05	50.16 ± 10.52	68.52 ± 2.74	42.71 ± 5.41	45.80 ± 3.85	67.54 ± 1.92	44.14 ± 0.77	**54.36 ± 10.80**
madin_categorical_gram_stain	98.29 ± 0.04	96.97 ± 1.29	93.68 ± 0.50	98.70 ± 0.22	97.62 ± 0.85	96.91 ± 0.71	98.83 ± 0.09	97.91 ± 0.87	97.12 ± 0.57
madin_categorical_metabolism	81.33 ± 1.64	45.04 ± 5.20	41.21 ± 3.47	83.01 ± 1.89	45.93 ± 3.91	43.75 ± 3.69	83.28 ± 3.42	45.85 ± 4.95	39.30 ± 3.27
madin_categorical_motility_binary	82.39 ± 5.25	81.20 ± 6.01	76.59 ± 3.37	85.73 ± 6.11	82.24 ± 7.14	77.58 ± 2.21	83.77 ± 4.53	83.31 ± 1.64	72.39 ± 3.69
madin_categorical_sporulation	96.03 ± 2.38	89.07 ± 6.56	89.48 ± 1.89	97.31 ± 1.06	89.51 ± 4.58	92.26 ± 2.94	97.52 ± 1.67	91.30 ± 6.21	**92.76 ± 2.90**

Phenotype	ProtBERT AUROC	ProtBERT AUPRC	ProtBERT F1	gLM2 AUROC	gLM2 AUPRC	gLM2 F1	Bacformer AUROC	Bacformer AUPRC	Bacformer F1
madin_carbsubs_caprate	84.68 ± 7.18	13.38 ± 9.29	49.51 ± 0.16	78.00 ± 13.09	11.39 ± 8.08	**49.53 ± 0.18**	**87.24 ± 4.48**	**16.13 ± 7.34**	**49.53 ± 0.18**
madin_carbsubs_cellobiose	63.12 ± 5.60	43.44 ± 7.88	47.19 ± 4.45	61.50 ± 7.73	42.21 ± 7.75	42.19 ± 1.65	**65.19 ± 3.25**	43.23 ± 2.36	51.73 ± 10.61
madin_carbsubs_citrate	72.12 ± 3.83	35.53 ± 3.27	**52.01 ± 5.01**	67.27 ± 5.70	30.91 ± 4.58	46.39 ± 0.53	**72.54 ± 5.10**	**38.39 ± 6.16**	51.01 ± 6.63
madin_carbsubs_dextrin	54.12 ± 4.24	9.91 ± 2.62	**47.81 ± 0.31**	60.39 ± 1.92	**14.00 ± 1.45**	**47.81 ± 0.31**	58.35 ± 3.78	11.25 ± 1.44	47.78 ± 0.30
madin_carbsubs_ethanol	**75.78 ± 2.70**	12.34 ± 5.08	49.06 ± 0.17	63.86 ± 5.08	6.11 ± 0.64	49.06 ± 0.17	71.87 ± 6.50	**13.97 ± 5.57**	**50.47 ± 2.59**
madin_carbsubs_formic acid	74.87 ± 4.43	10.75 ± 2.31	48.78 ± 0.15	62.89 ± 14.67	7.61 ± 2.82	48.78 ± 0.15	**78.97 ± 2.43**	**16.06 ± 3.40**	**50.80 ± 1.61**
madin_carbsubs_fructose	61.10 ± 3.55	53.57 ± 1.07	47.01 ± 6.47	60.52 ± 3.59	50.73 ± 2.08	52.68 ± 2.83	**65.87 ± 2.11**	**55.64 ± 0.89**	51.63 ± 9.04
madin_carbsubs_fucose	58.62 ± 15.47	14.94 ± 8.71	**47.71 ± 0.29**	54.13 ± 12.04	11.75 ± 4.03	**47.71 ± 0.29**	65.08 ± 4.55	18.13 ± 4.72	**47.71 ± 0.29**
madin_carbsubs_galactose	65.22 ± 2.88	47.04 ± 3.38	55.43 ± 10.43	60.16 ± 2.15	42.91 ± 3.89	42.03 ± 2.35	**68.15 ± 1.36**	**48.03 ± 2.50**	**60.34 ± 4.76**
madin_carbsubs_galacturonic acid	49.58 ± 13.40	2.86 ± 0.79	**49.41 ± 0.08**	55.79 ± 8.71	3.18 ± 1.09	**49.41 ± 0.08**	56.89 ± 18.86	6.46 ± 3.92	**49.41 ± 0.08**
madin_carbsubs_gelatin	58.53 ± 2.57	7.75 ± 1.77	**48.69 ± 0.16**	59.39 ± 4.38	7.07 ± 0.82	**48.69 ± 0.16**	59.42 ± 8.79	**9.21 ± 3.67**	**48.69 ± 0.16**
madin_carbsubs_gluconic acid	**66.41 ± 3.13**	23.41 ± 5.70	**46.32 ± 0.97**	63.23 ± 5.65	23.15 ± 7.43	**46.32 ± 0.97**	62.14 ± 1.01	**24.55 ± 6.87**	46.31 ± 1.00
madin_carbsubs_glucosamine	57.92 ± 5.69	**4.91 ± 1.64**	**49.40 ± 0.16**	**61.87 ± 6.54**	3.91 ± 0.87	**49.40 ± 0.16**	54.93 ± 5.82	3.35 ± 0.75	**49.40 ± 0.16**
madin_carbsubs_glucose	55.78 ± 6.18	69.74 ± 4.57	46.88 ± 5.93	52.07 ± 7.54	65.61 ± 5.02	44.09 ± 9.35	**62.28 ± 1.50**	**73.86 ± 1.05**	**55.38 ± 4.42**
madin_carbsubs_glucuronamide	50.92 ± 3.94	3.97 ± 2.77	**49.56 ± 0.04**	57.63 ± 9.18	3.46 ± 1.22	**49.56 ± 0.04**	69.06 ± 1.44	3.49 ± 0.69	**49.56 ± 0.04**
madin_carbsubs_glutamate	66.74 ± 3.42	20.84 ± 0.92	46.21 ± 0.26	**69.54 ± 3.41**	24.44 ± 1.00	46.64 ± 0.79	66.39 ± 4.20	21.90 ± 4.11	47.06 ± 1.07
madin_carbsubs_glycerol	60.06 ± 6.91	31.27 ± 3.14	44.22 ± 1.13	59.72 ± 6.74	30.78 ± 3.85	44.42 ± 1.26	60.47 ± 7.18	29.30 ± 2.12	43.72 ± 0.71
madin_carbsubs_glycine	53.83 ± 6.53	3.25 ± 0.51	**49.26 ± 0.19**	44.89 ± 2.18	2.66 ± 0.61	**49.26 ± 0.19**	49.84 ± 6.47	3.35 ± 0.92	**49.26 ± 0.19**
madin_carbsubs_glycogen	56.31 ± 2.78	14.31 ± 1.85	**47.05 ± 0.18**	54.84 ± 7.37	13.41 ± 3.21	**47.05 ± 0.18**	59.65 ± 4.75	14.59 ± 0.94	**47.05 ± 0.18**
madin_carbsubs_histidine	58.88 ± 4.59	12.57 ± 2.25	**48.01 ± 0.09**	58.83 ± 5.01	12.07 ± 3.14	**48.01 ± 0.09**	**62.25 ± 2.11**	13.05 ± 3.49	47.99 ± 0.12
madin_carbsubs_inosine	59.67 ± 7.03	**6.60 ± 3.47**	**49.24 ± 0.12**	61.72 ± 1.72	5.33 ± 0.80	**49.24 ± 0.12**	55.59 ± 8.76	4.42 ± 2.31	**49.24 ± 0.12**
madin_carbsubs_lactate	66.86 ± 1.94	25.18 ± 3.56	45.78 ± 0.67	64.94 ± 2.26	29.40 ± 1.21	45.60 ± 0.37	68.41 ± 6.79	31.47 ± 6.77	**51.96 ± 5.91**
madin_carbsubs_lactose	68.30 ± 2.01	44.74 ± 6.73	47.41 ± 9.26	64.78 ± 3.09	42.28 ± 4.59	50.92 ± 6.92	**71.37 ± 2.07**	**47.52 ± 5.56**	50.79 ± 6.66
madin_carbsubs_leucine	59.67 ± 1.90	9.54 ± 1.29	**48.49 ± 0.03**	57.76 ± 5.74	**10.91 ± 6.77**	**48.49 ± 0.03**	**63.18 ± 2.49**	10.10 ± 1.95	**48.49 ± 0.03**
madin_carbsubs_malate	72.07 ± 3.99	31.34 ± 6.49	45.55 ± 0.31	69.38 ± 3.71	31.48 ± 6.88	45.30 ± 0.36	**73.26 ± 5.44**	**34.83 ± 7.36**	45.73 ± 0.75
madin_carbsubs_malonate	66.98 ± 4.55	8.62 ± 1.67	48.61 ± 0.34	67.06 ± 2.24	**12.43 ± 1.95**	**48.63 ± 0.35**	67.62 ± 5.16	10.55 ± 2.68	**48.63 ± 0.35**
madin_carbsubs_maltose	61.51 ± 2.22	58.73 ± 1.76	58.51 ± 2.15	58.00 ± 2.28	56.71 ± 5.76	51.95 ± 5.98	**63.76 ± 6.03**	**60.73 ± 5.22**	57.82 ± 3.25
madin_carbsubs_mannitol	68.05 ± 2.61	52.54 ± 2.18	49.09 ± 8.48	63.77 ± 6.42	47.95 ± 5.88	48.74 ± 9.86	**71.31 ± 2.46**	**54.08 ± 2.87**	**59.78 ± 1.12**
madin_carbsubs_mannose	64.34 ± 1.90	59.17 ± 4.38	57.90 ± 1.83	61.87 ± 1.69	56.83 ± 3.89	49.85 ± 6.51	**67.24 ± 2.04**	**61.29 ± 3.23**	**60.06 ± 5.68**
madin_carbsubs_melibiose	71.94 ± 3.36	46.10 ± 6.03	51.23 ± 6.17	66.54 ± 6.11	40.04 ± 9.18	47.42 ± 3.13	**74.12 ± 1.81**	**47.13 ± 2.35**	**52.31 ± 2.34**
madin_carbsubs_methyl beta-D-glucoside	59.61 ± 7.74	6.04 ± 1.70	**49.04 ± 0.40**	65.51 ± 8.61	**10.66 ± 6.08**	**49.04 ± 0.40**	58.02 ± 13.24	6.57 ± 4.02	**49.04 ± 0.40**
madin_carbsubs_methyl pyruvate	59.49 ± 12.57	3.48 ± 2.07	49.37 ± 0.35	64.84 ± 9.33	6.60 ± 3.61	**49.39 ± 0.32**	58.37 ± 14.34	3.48 ± 0.25	**49.39 ± 0.32**
madin_carbsubs_myo-inositol	**66.83 ± 10.94**	**17.62 ± 3.53**	**49.03 ± 1.68**	65.83 ± 8.00	14.23 ± 4.46	48.20 ± 0.56	59.61 ± 12.50	10.69 ± 6.34	48.20 ± 0.56
madin_carbsubs_ornithine	74.62 ± 4.74	**11.58 ± 3.83**	**48.88 ± 0.29**	70.58 ± 6.25	8.17 ± 3.08	**48.88 ± 0.29**	**75.43 ± 3.46**	11.03 ± 3.34	**48.88 ± 0.29**
madin_carbsubs_phenylacetate	64.02 ± 14.32	7.82 ± 8.07	49.23 ± 0.40	70.40 ± 11.60	11.52 ± 13.94	**49.24 ± 0.42**	62.94 ± 20.39	10.80 ± 13.61	**49.24 ± 0.42**
madin_carbsubs_phenylalanine	59.87 ± 10.65	5.32 ± 1.11	**49.18 ± 0.13**	63.76 ± 5.87	7.18 ± 0.69	**49.18 ± 0.13**	60.96 ± 1.31	4.37 ± 0.91	**49.18 ± 0.13**
madin_carbsubs_proline	67.73 ± 2.06	23.65 ± 3.04	46.34 ± 0.20	64.38 ± 1.83	21.42 ± 2.83	46.36 ± 0.21	**68.55 ± 2.39**	**24.17 ± 3.77**	**46.77 ± 0.89**
madin_carbsubs_propionate	63.47 ± 6.15	17.32 ± 6.38	**47.66 ± 1.25**	60.48 ± 2.75	16.55 ± 3.74	47.11 ± 0.39	**70.03 ± 1.48**	**23.29 ± 4.77**	47.10 ± 0.38
madin_carbsubs_putrescine	45.88 ± 7.99	**3.18 ± 2.82**	49.49 ± 0.15	42.52 ± 12.73	2.13 ± 1.08	**49.52 ± 0.17**	48.50 ± 13.51	2.65 ± 1.93	**49.52 ± 0.17**
madin_carbsubs_pyruvate	72.22 ± 4.68	30.53 ± 2.34	51.25 ± 3.26	66.76 ± 1.26	24.67 ± 1.16	47.92 ± 2.44	**72.41 ± 3.77**	**32.37 ± 2.56**	**52.11 ± 5.15**
madin_carbsubs_pyruvic acid methyl ester	57.51 ± 11.72	5.84 ± 2.70	**49.02 ± 0.16**	65.60 ± 6.40	11.23 ± 9.37	**49.02 ± 0.16**	**71.49 ± 3.40**	10.39 ± 2.18	**49.02 ± 0.16**
madin_carbsubs_quinic acid	56.91 ± 10.86	2.47 ± 0.96	**49.59 ± 0.08**	**61.87 ± 2.41**	4.45 ± 3.27	**49.59 ± 0.08**	51.96 ± 17.78	2.24 ± 1.00	**49.59 ± 0.08**
madin_carbsubs_raffinose	67.47 ± 3.91	36.32 ± 6.14	48.14 ± 4.65	63.66 ± 5.43	35.09 ± 7.33	44.83 ± 1.50	**73.50 ± 1.06**	**44.30 ± 3.95**	46.83 ± 2.94
madin_carbsubs_rhamnose	71.57 ± 1.39	42.18 ± 7.37	48.00 ± 7.90	65.52 ± 11.81	37.62 ± 13.79	43.50 ± 0.67	**73.80 ± 1.13**	43.89 ± 8.22	**53.73 ± 6.27**
madin_carbsubs_ribose	65.06 ± 4.30	30.55 ± 2.01	43.97 ± 0.66	64.95 ± 5.84	33.17 ± 6.38	44.92 ± 1.52	**67.96 ± 4.50**	**34.45 ± 2.88**	**48.12 ± 6.63**
madin_carbsubs_salicin	71.77 ± 1.33	32.54 ± 6.43	46.53 ± 1.72	63.53 ± 8.25	25.99 ± 9.00	45.19 ± 0.69	71.53 ± 3.15	33.12 ± 6.49	45.95 ± 0.98
madin_carbsubs_serine	66.07 ± 1.47	21.05 ± 5.51	47.52 ± 1.31	66.02 ± 2.53	20.73 ± 2.18	46.98 ± 0.44	**68.08 ± 2.62**	**21.83 ± 5.21**	46.98 ± 0.44
madin_carbsubs_sorbitol	58.49 ± 3.76	22.08 ± 0.38	**45.29 ± 0.43**	58.01 ± 3.25	23.77 ± 0.90	**45.29 ± 0.43**	**61.25 ± 2.10**	22.70 ± 1.20	45.23 ± 0.45
madin_carbsubs_succinate	**72.44 ± 2.58**	31.98 ± 3.66	45.75 ± 0.49	69.46 ± 2.03	32.80 ± 4.06	45.98 ± 0.82	72.43 ± 1.33	33.88 ± 2.53	47.41 ± 3.34
madin_carbsubs_sucrose	65.90 ± 1.99	59.91 ± 7.17	56.92 ± 4.83	59.27 ± 5.32	53.73 ± 8.54	51.59 ± 6.95	**66.62 ± 7.70**	58.79 ± 12.54	55.71 ± 15.99
madin_carbsubs_threonine	55.02 ± 6.81	10.18 ± 3.64	48.29 ± 0.23	**58.54 ± 2.42**	11.28 ± 3.06	48.29 ± 0.23	57.79 ± 1.92	**11.67 ± 3.77**	**49.44 ± 2.13**
madin_carbsubs_thymidine	50.07 ± 3.23	2.50 ± 0.21	**49.46 ± 0.09**	53.82 ± 11.94	2.88 ± 1.08	**49.46 ± 0.09**	49.41 ± 7.21	2.35 ± 0.57	**49.46 ± 0.09**
madin_carbsubs_trehalose	62.10 ± 2.90	41.12 ± 5.41	48.92 ± 5.56	60.50 ± 2.87	41.08 ± 5.77	49.49 ± 0.44	62.89 ± 3.71	43.30 ± 4.70	**51.63 ± 1.37**
madin_carbsubs_urea	**52.96 ± 8.87**	**2.98 ± 1.32**	**49.46 ± 0.12**	44.22 ± 1.38	2.12 ± 0.33	**49.46 ± 0.12**	51.99 ± 9.37	2.26 ± 0.32	**49.46 ± 0.12**
madin_carbsubs_uridine	61.27 ± 7.91	5.06 ± 1.63	**49.21 ± 0.13**	61.19 ± 2.35	4.88 ± 0.33	**49.21 ± 0.13**	56.70 ± 6.48	4.37 ± 1.23	**49.21 ± 0.13**
madin_carbsubs_valerate	60.39 ± 6.62	5.10 ± 2.50	**49.35 ± 0.31**	56.59 ± 3.11	4.27 ± 1.90	**49.35 ± 0.31**	**63.68 ± 9.04**	**9.55 ± 5.59**	**49.35 ± 0.31**
madin_carbsubs_xylitol	50.83 ± 4.28	6.20 ± 1.14	**48.81 ± 0.12**	50.79 ± 2.57	5.93 ± 1.05	**48.81 ± 0.12**	51.02 ± 4.85	**7.77 ± 1.88**	**48.81 ± 0.12**
madin_carbsubs_xylose	66.75 ± 4.12	44.08 ± 2.80	49.77 ± 7.72	62.15 ± 5.21	39.38 ± 5.84	41.65 ± 0.82	**70.98 ± 1.66**	**46.54 ± 2.86**	47.29 ± 9.36
madin_categorical_gram_stain	**99.02 ± 0.19**	**98.12 ± 0.61**	**97.36 ± 0.61**	98.12 ± 0.01	96.47 ± 1.51	93.46 ± 0.40	98.92 ± 0.27	97.86 ± 0.87	97.25 ± 0.48
madin_categorical_metabolism	86.41 ± 1.13	49.38 ± 2.11	46.07 ± 2.73	79.50 ± 4.32	41.65 ± 6.47	36.22 ± 6.56	**86.68 ± 1.62**	**50.99 ± 1.73**	**47.58 ± 1.38**
madin_categorical_motility_binary	78.93 ± 3.06	72.34 ± 18.21	68.58 ± 6.70	82.41 ± 5.34	81.05 ± 3.71	75.79 ± 2.25	**89.80 ± 3.04**	**89.87 ± 3.73**	**81.67 ± 2.29**
madin_categorical_sporulation	97.19 ± 1.06	88.42 ± 8.07	87.37 ± 10.88	96.27 ± 2.32	88.49 ± 8.00	91.26 ± 2.73	**97.83 ± 1.17**	**91.95 ± 3.60**	90.11 ± 4.98

D BROADER IMPACT & LIMITATIONS

Broader impact BacBench provides the first public, multi-task testbed that spans gene-, system- and genome-scale prediction problems over 67k genomes from 17.6k species. By standardising data splits, evaluation code and baseline embeddings, it lowers the entry barrier for machine-learning researchers who lack domain-specific pipelines yet want to work on microbial genomics. In the near term, more reliable essential-gene or antibiotic-resistance predictors could shorten drug-development cycles and inform stewardship policies, while better phenotype-from-genome models will accelerate the search for chassis strains that sequester carbon, degrade waste or synthesise valuable biochemicals. Because the benchmark emphasises cross-species generalization, methods that succeed on BacBench are naturally suited to poorly studied or newly sequenced taxa, helping global health laboratories track emerging pathogens even when only draft assemblies are available. Finally, releasing all data under permissive licences and exposing a HuggingFace hub invites continual community contributions, which should foster an open, comparative culture similar to computer-vision or NLP benchmarks and, in turn, drive rapid, reproducible advances in microbial bio-AI research.

Limitations Despite its breadth, BacBench still samples an uneven slice of bacterial diversity: phenotypic-trait labels cluster heavily around medically important genera, and some antibiotic classes remain sparsely represented, which could bias models towards well-studied lineages and mechanisms. Tasks that matter for ecology and biotechnology—horizontal-gene-transfer detection, host–phage interaction, metabolic-flux prediction or transcriptome conditioning—are absent, so performance on BacBench should not be interpreted as general mastery of bacterial genomics. Moreover, the benchmark inherits experimental noise from upstream databases: STRING DB interaction scores mix heterogeneous evidence; operon annotations are incomplete; and phenotype labels amalgamate disparate growth protocols, introducing label uncertainty that caps achievable accuracy. Finally, computing embeddings for every update is resource-intensive, which may hinder participation from groups without access to multi-GPU servers, although smaller surrogate splits are planned for future releases.