



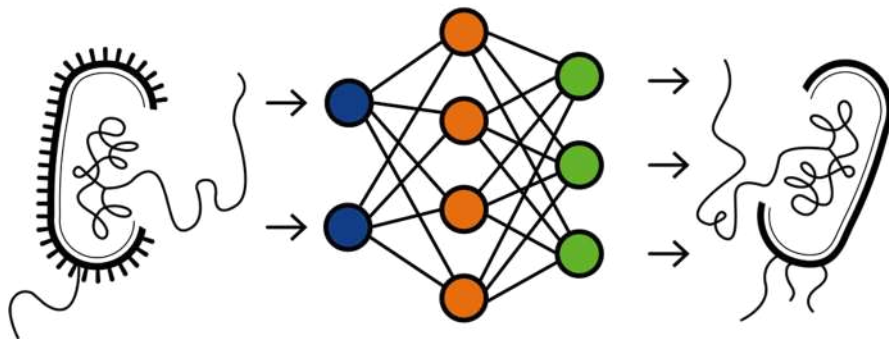
UNIVERSITY OF
CAMBRIDGE



Cambridge Centre
for AI in Medicine

Learning the language of bacterial genomes with genomic language models

Maciej Wiatrak | USCCN | 31.07.2026



The world runs on and suffers from bacteria



bacterial infections cause...

7.7M deaths

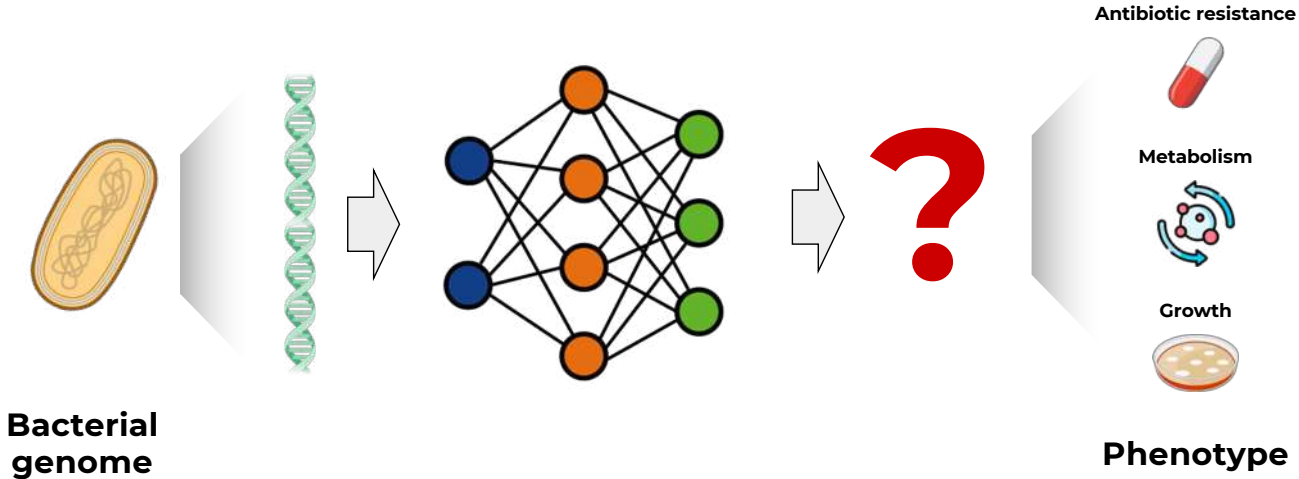
costing us...

\$693B

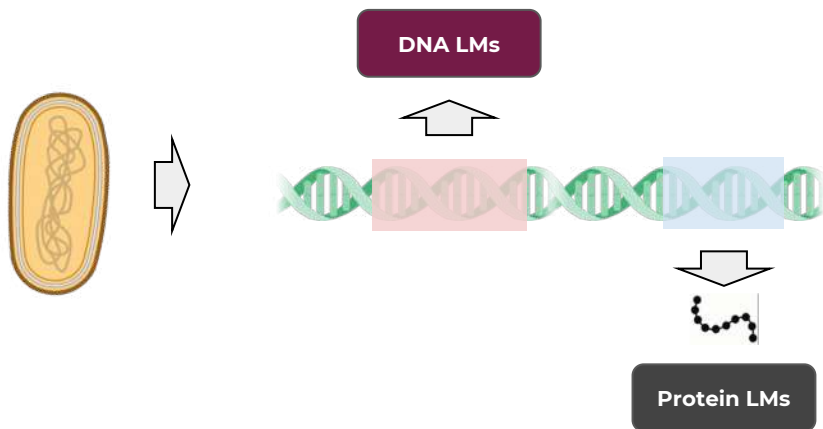
every year



The Problem: mapping genotype to phenotype



Existing methods fail to model bacteria



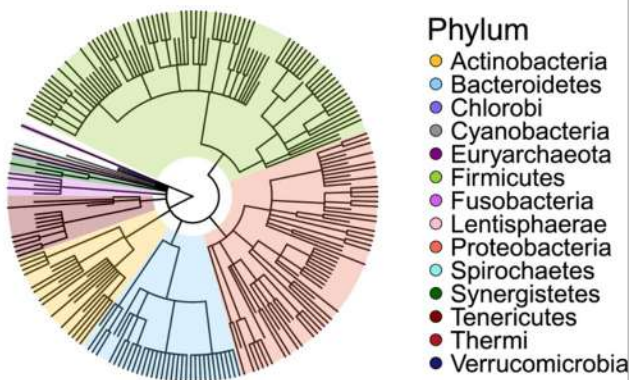
>8 TB GPU Mem

GPU memory required to model (forward+backward pass) an *E. Coli* genome using Evo* model.

Existing genomic LMs model only a small subset of bacterial genomes

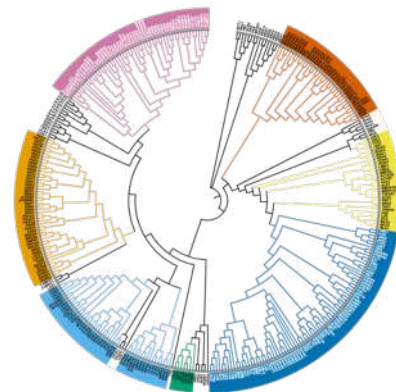
Transfer learning across bacteria

**Bacteria share a lot of similarities
across species**

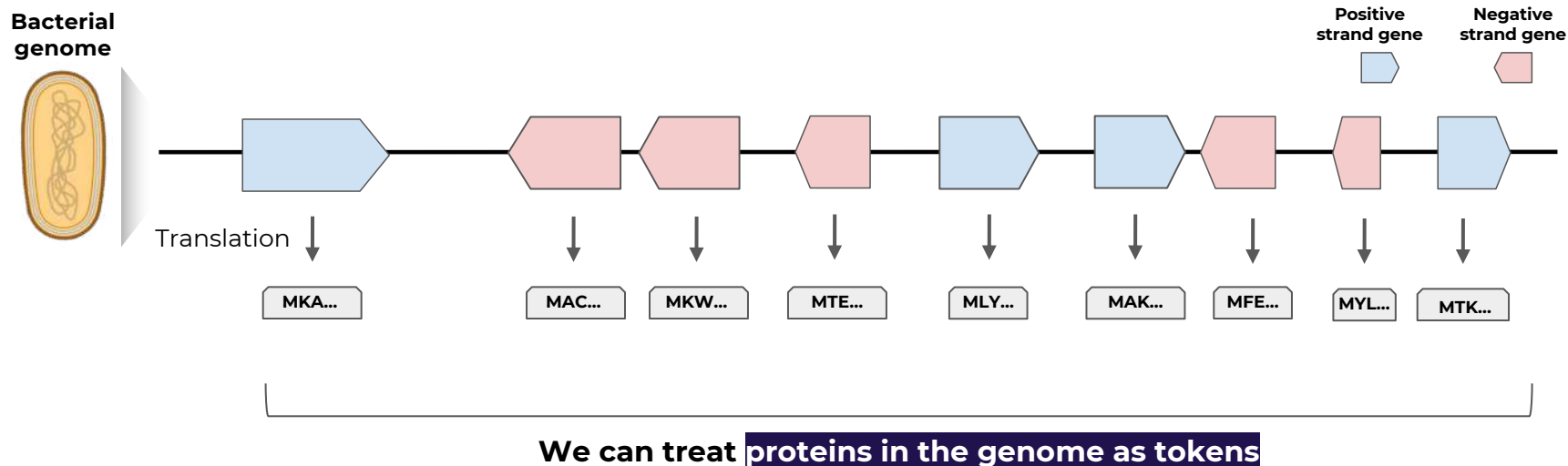


85-90%
of the genome is
protein-coding
across diverse bacteria

**Proteins can be
clustered into groups**

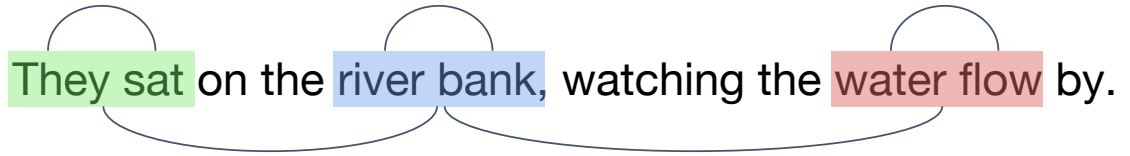


Unified representation of bacteria



Bacteria-specific architecture

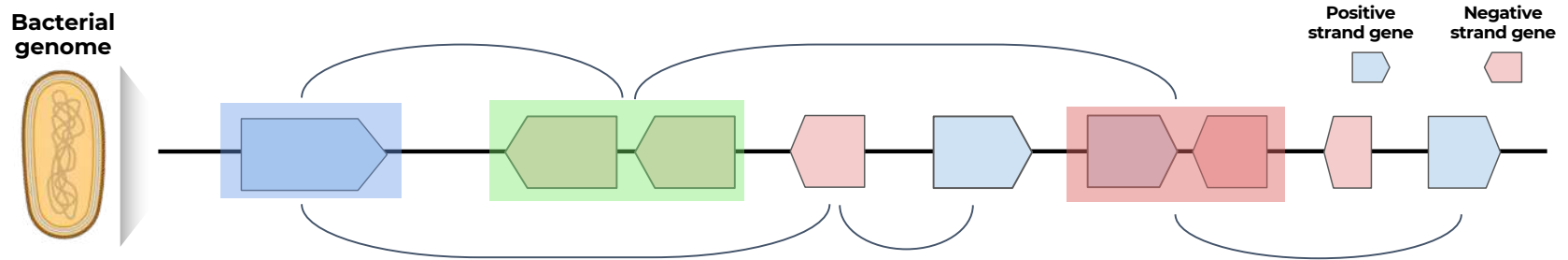
In text - word syntax defines the sentence meaning.



In proteins - amino acid syntax defines protein function.

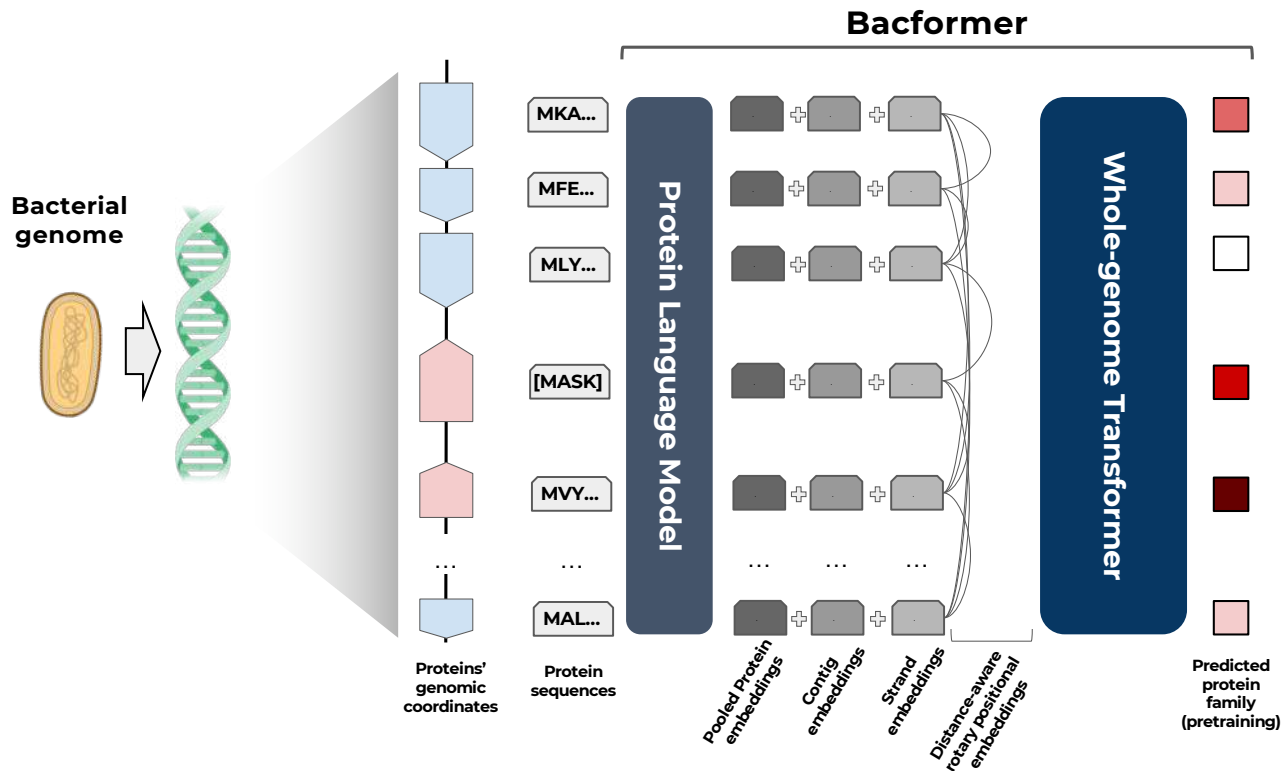


Bacteria-specific architecture



In bacteria - protein syntax defines the phenotype.

Modeling whole bacterial genomes with Bacformer

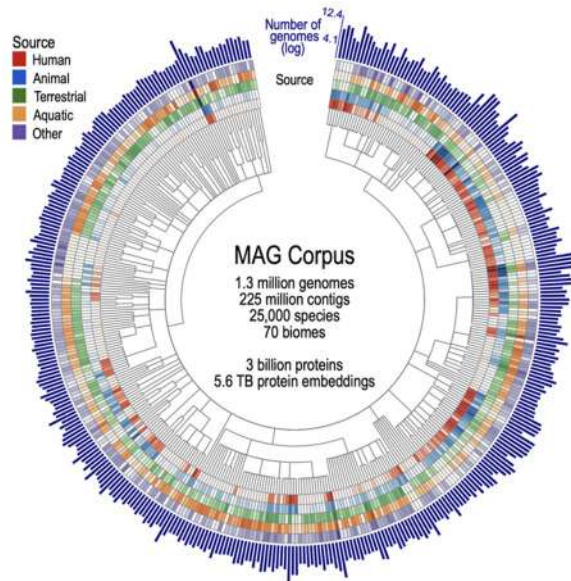
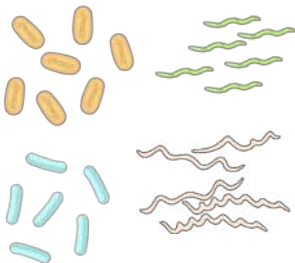


Pretraining corpus

Metagenomic samples and
bacterial isolates from
diverse environments

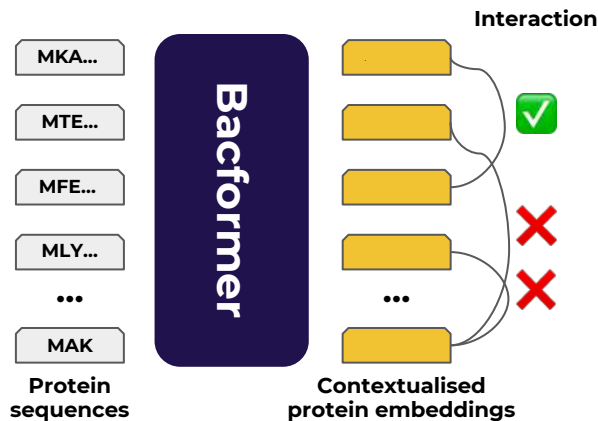


Diverse bacterial
genomes

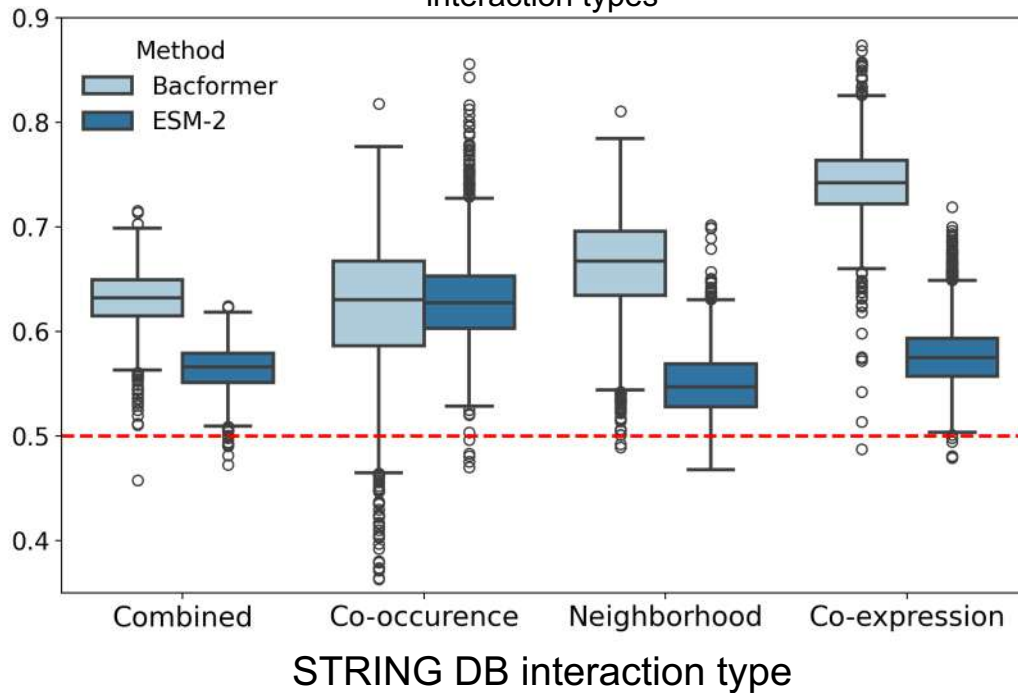


Bacfomer
pretraining

Bacformer predicts protein-protein interactions (1)

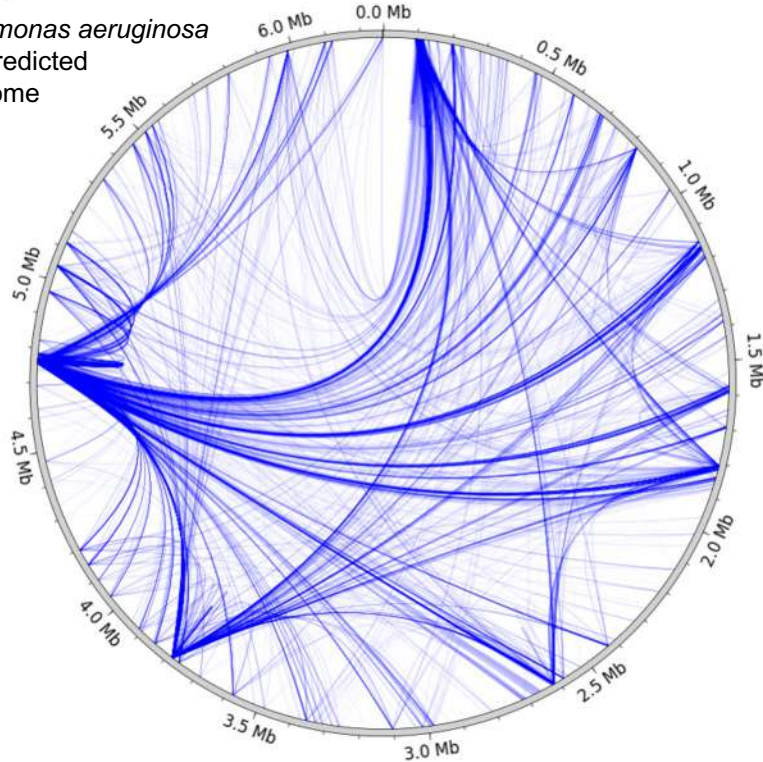


Zero-shot protein-protein interaction prediction performance across interaction types

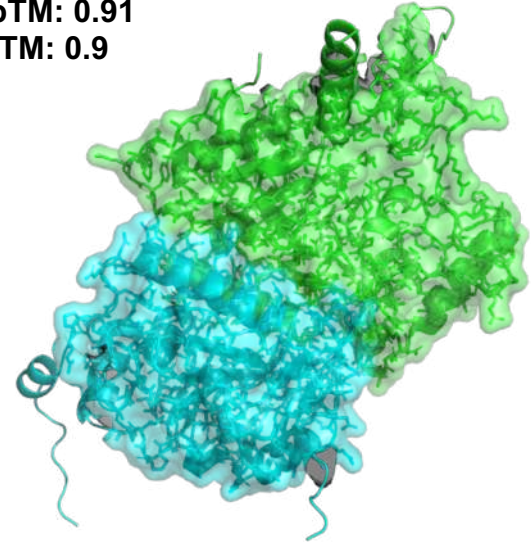


Bacformer predicts protein-protein interactions (2)

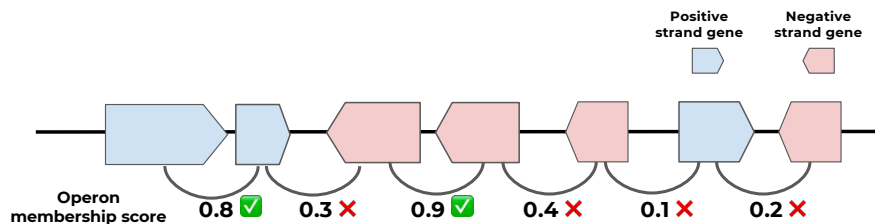
Pseudomonas aeruginosa
PAO1 predicted
interactome



phnJ & phnH
ipTM: 0.91
pTM: 0.9



Validating Bacformer with operon prediction



Operon-Membership Score (s_i) Model

$$s_i = c_i \cdot e^{-d_i/\lambda} \cdot I_{\text{strand}} \cdot I_{\text{gap}}$$

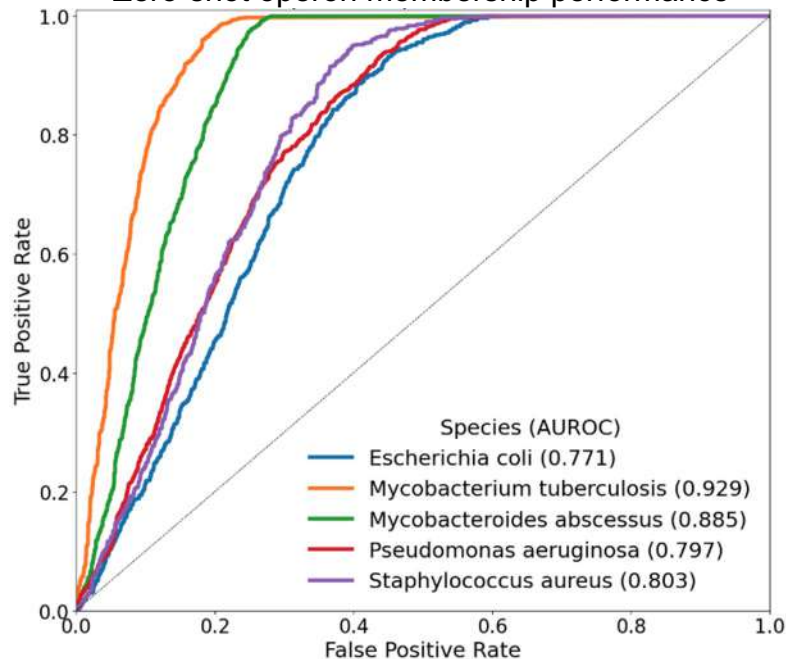
Continuous Factors

- $s_i \in [0, 1]$: Binary score indicating operon membership threshold.
- $c_i = \frac{1}{2}(1 + \cos(\hat{h}_i, \hat{h}_{i+1}))$: Similarity term based on ℓ_2 -normalised Bacformer embeddings ($\hat{h}$) of genes $i, i+1$.
- d_i : Intergenic distance between genes.
- $e^{-d_i/\lambda}$: Exponential decay factor reducing score with distance.
- $\lambda = 500$ bp: Length scale constant.

Hard Veto Indicators ($I \in \{0, 1\}$)

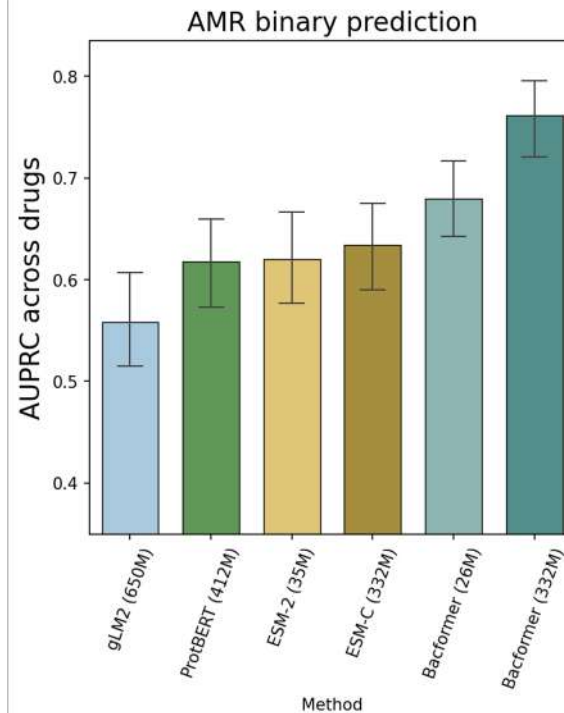
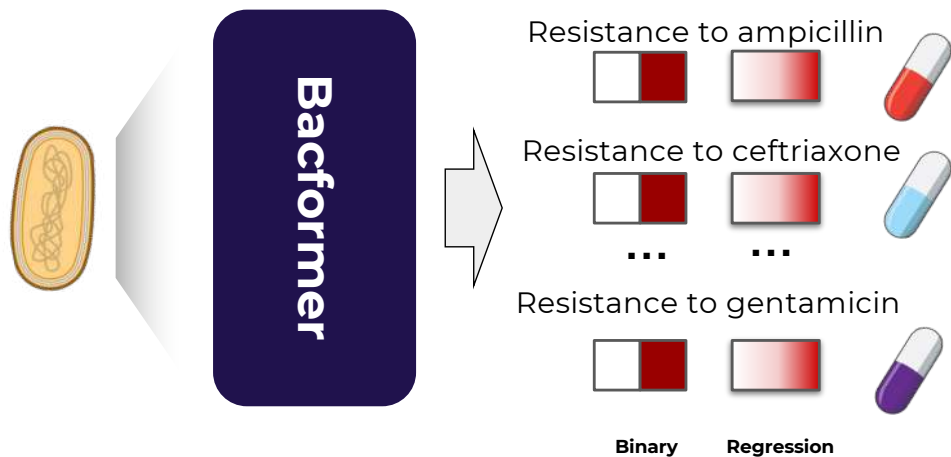
- I_{strand} : 1 if genes are on the SAME strand, 0 otherwise.
- I_{gap} : 1 if intergenic distance $d_i \leq 500$ bp, 0 otherwise.

Zero-shot operon membership performance

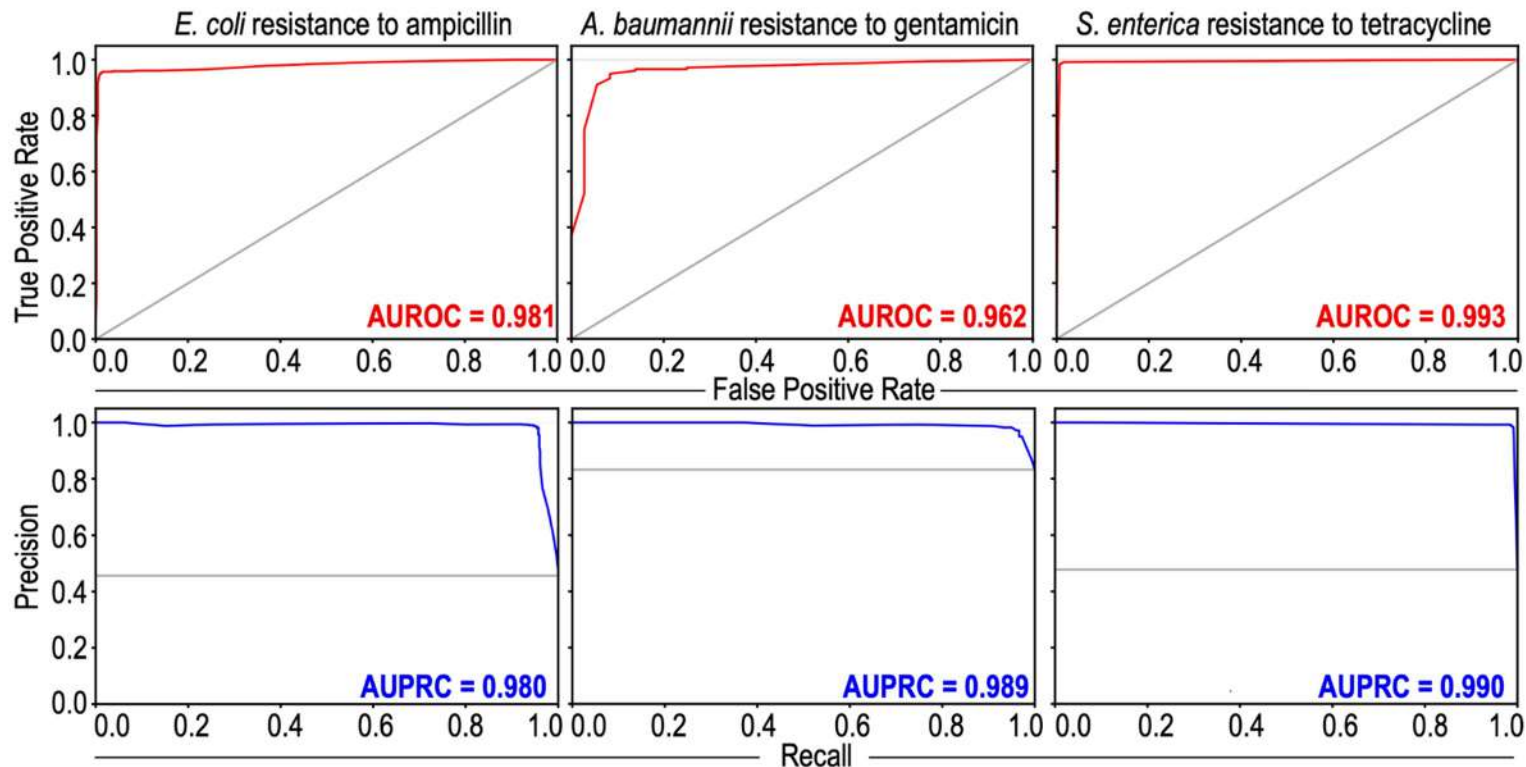


Bacformer accurately predicts antimicrobial resistance (AMR)

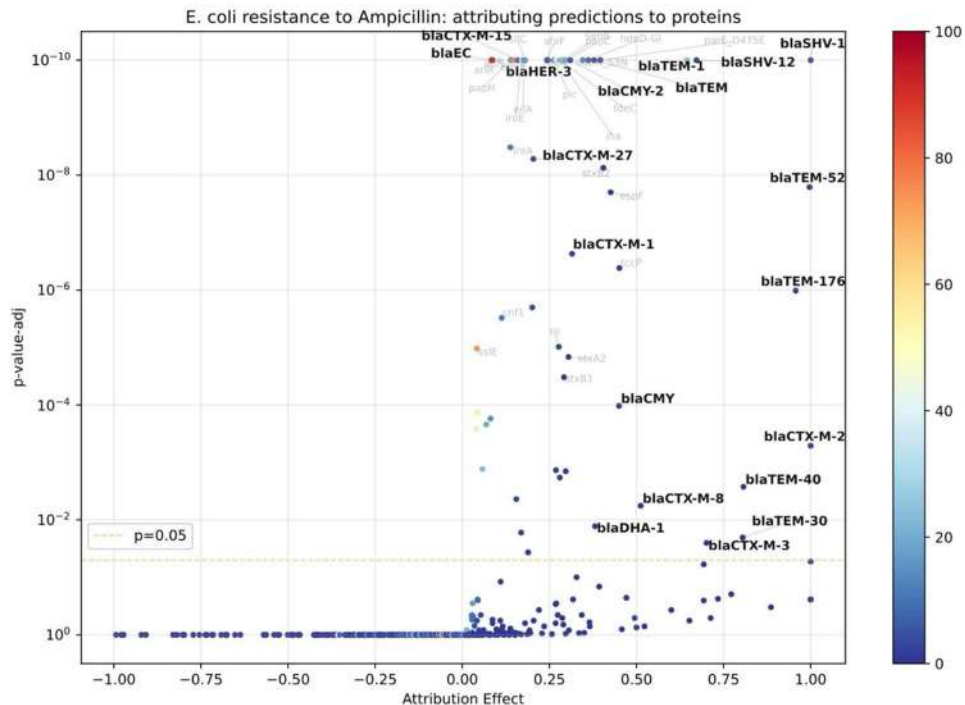
Antimicrobial Resistance (AMR) Prediction



Bacformer accurately predicts antimicrobial resistance (AMR)



Bacformer identifies AMR resistance genes



Bacformer identifies known AMR genes with

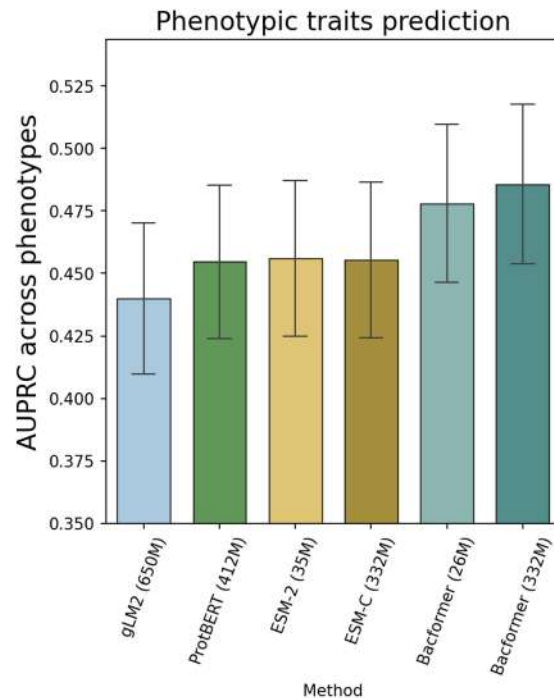
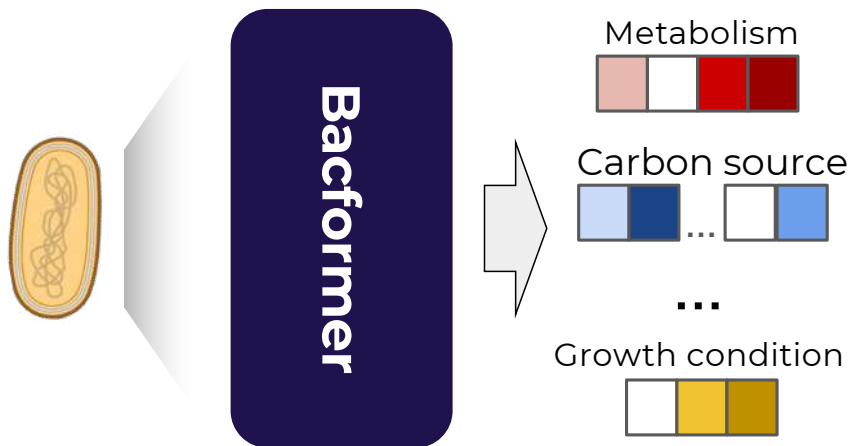
>95%
specificity

In *E. coli* / ampicillin (drug)

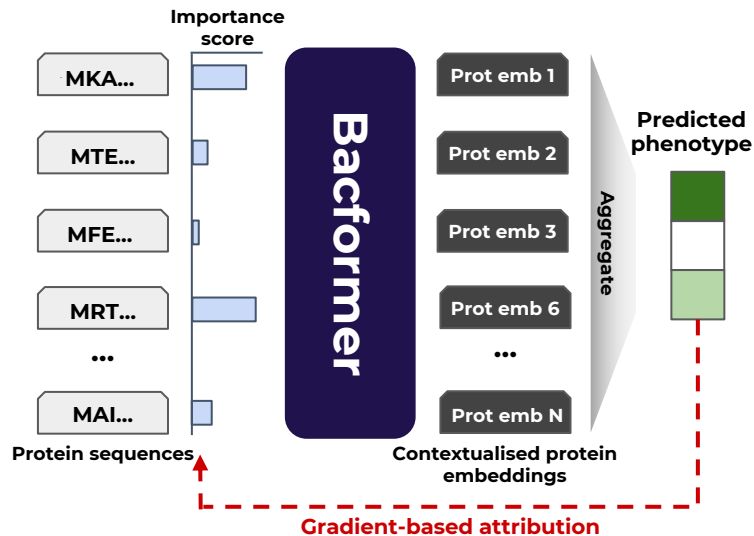
using *in silico* feature ablation

Bacformer predicts phenotypic traits

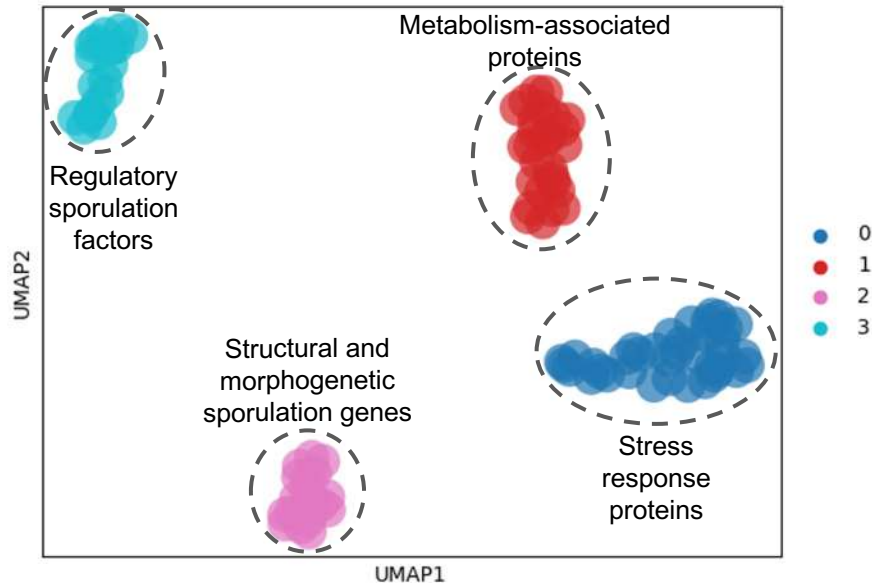
Phenotypic traits prediction



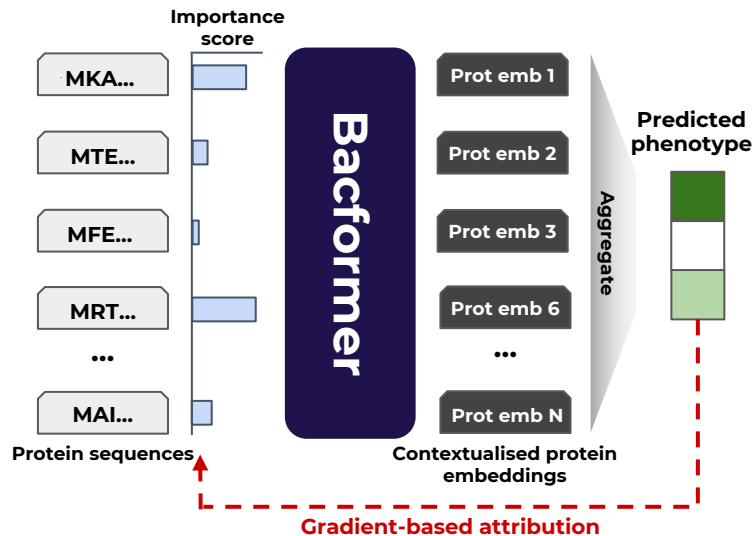
Understanding the genes driving the phenotypic traits



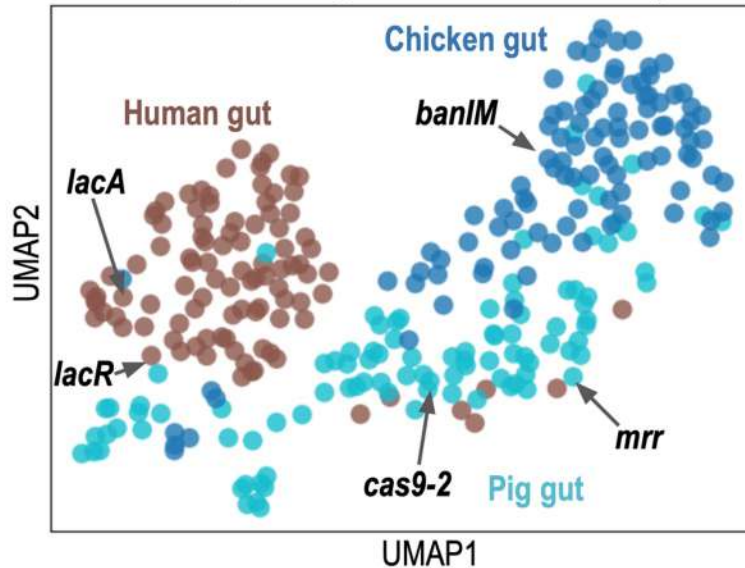
Leiden clustering of sporulation-specific genes



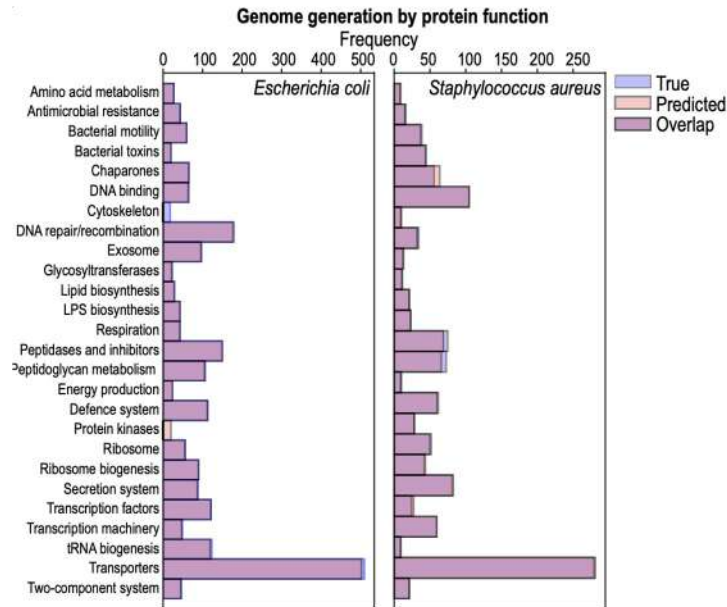
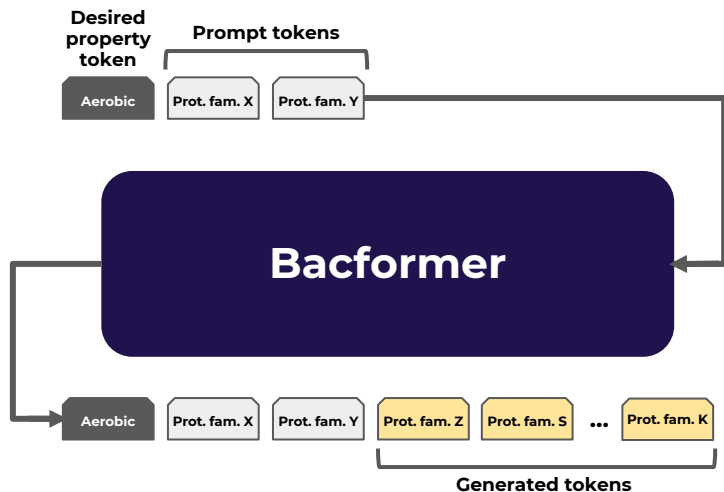
Understanding the genes driving the phenotypic traits



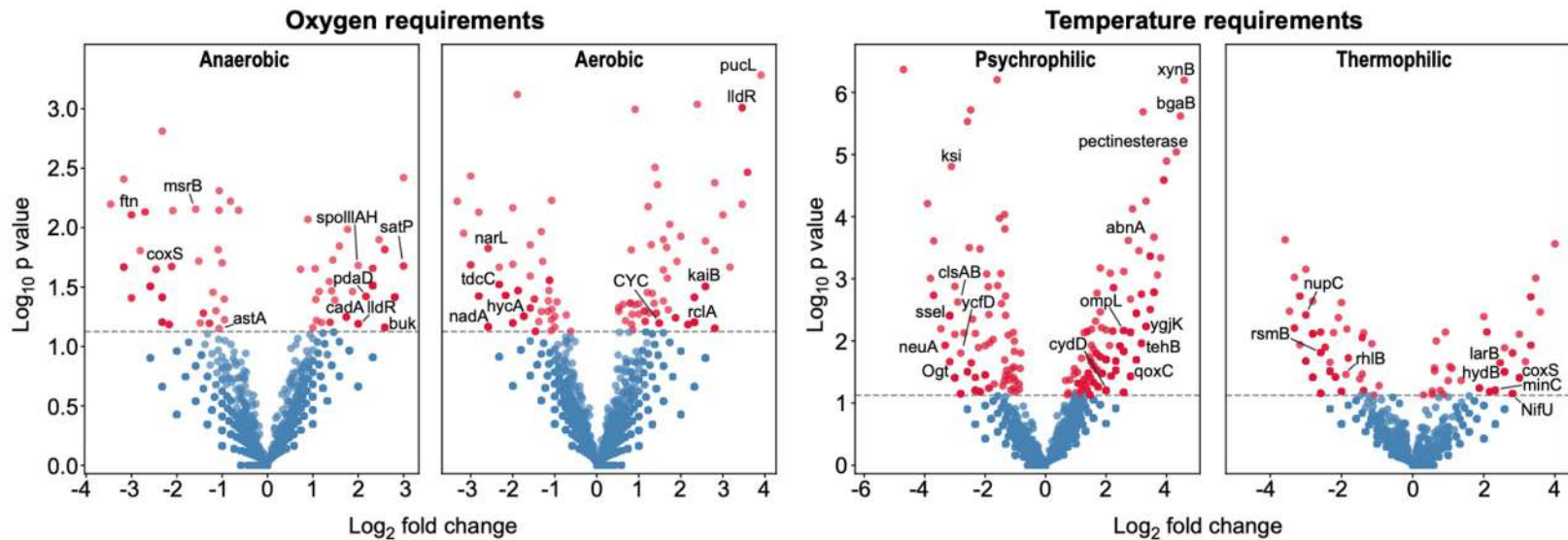
Environment-specific genes in *Lactobacillus johnsonii*



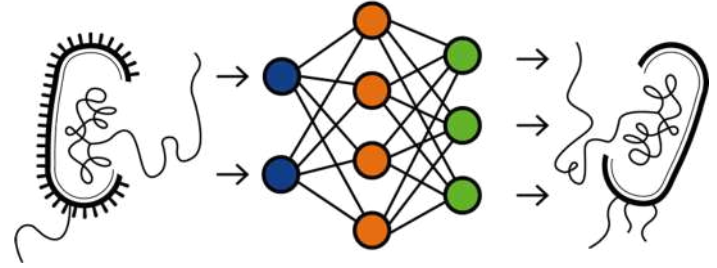
Bacformer generates genomes with diverse protein functions



Bacformer generates genomes with desired properties



Conclusions



Conclusions

- Bacformer produces **accurate bacterial genome representations** across species
- Bacformer exhibits **scaling laws** with gains as model size increases
- Genomic organisation in bacteria provides a strong inductive bias
- Bacformer achieves state-of-the-art results on **genome to phenotype** prediction
- Bacformer allows for genome completion

Acknowledgements

Maciej Wiatrak
(Uni of Cam)



David Abelson
(Uni of Cam)



Maria Ntemortsidou
(Uni of Cam)



Sam Horsfield
(EBI)



Divya Arora
(Uni of Cam)



Adam Dinan
(Uni of Cam)



Ramon Viñas Torné
(EPFL)



Maria Brbić
(EPFL)



John Lees
(EBI)



Aaron Weimann
(Uni of Cam)



Andres Floto
(Uni of Cam)



Anja Hagting (Uni of Cam)

Chris Ruis (Uni of Cam)

Sundeeep Vedithi (Uni of Cam)

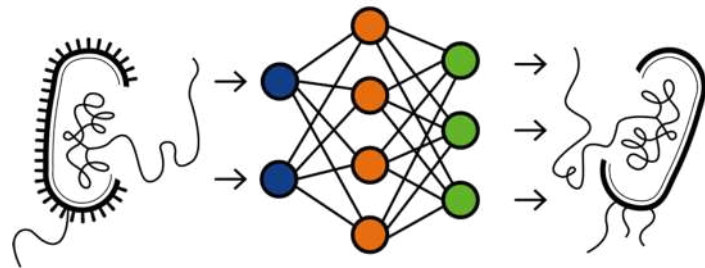
Oliver Hernandez (Imperial College London)

Chris Barnes (UCL)

Funders



Thank you!



HuggingFace 🤗



GitHub 🖨



Preprint 📖

