Phagos x AWS Hackdays 2025

A, Aragon, E. de Bézenac, F. Camaglia, M. Crilout
Train Tune Deploy (aka *Team Delta)*

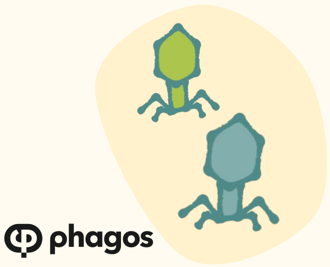
Introduction

- Imagine an unidentified bacterial strain is currently infecting Uncle Tom's farm.

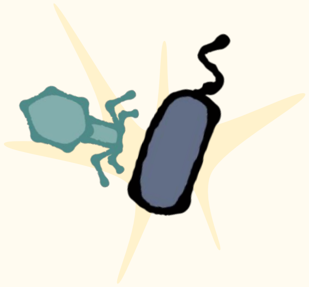
- Here at Phagos, we've isolated a collection of phages specific to this bacterium.

- From that collection, we want to select the phages exhibiting the strongest lytic activity.

→ Klebsiella



→ Boeckaerts et al. (2024)

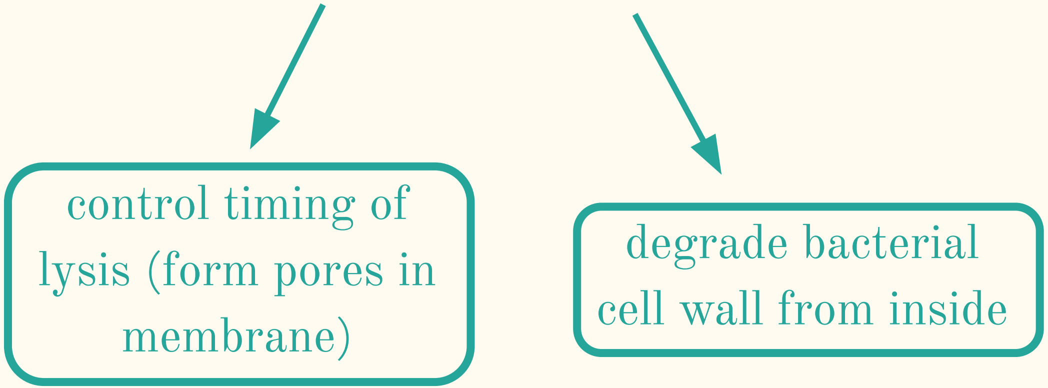


GOAL:

Transfer learning from other datasets!

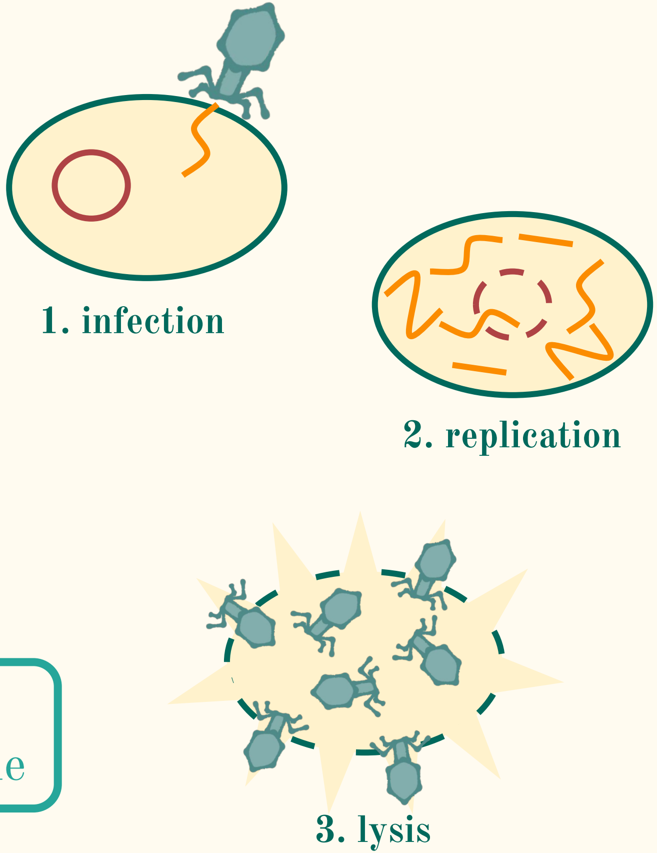
Lytic Functions in ESKAPE

- Genomes & proteins of phages targeting ESKAPE pathogens (staphylococcus, klebsiella, etc.)
- Select 14 infection protein classes including lysis proteins (Holins & Endolysins).



control timing of lysis (form pores in membrane)

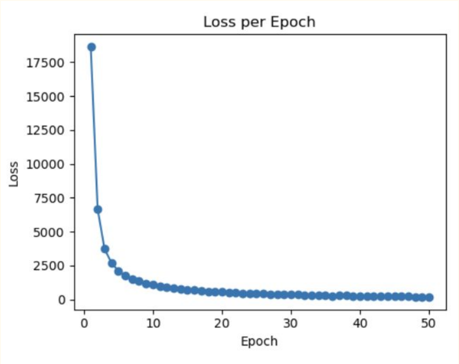
degrade bacterial cell wall from inside

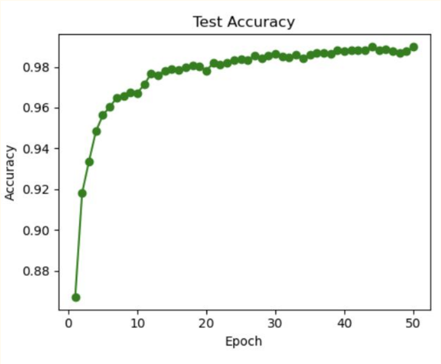


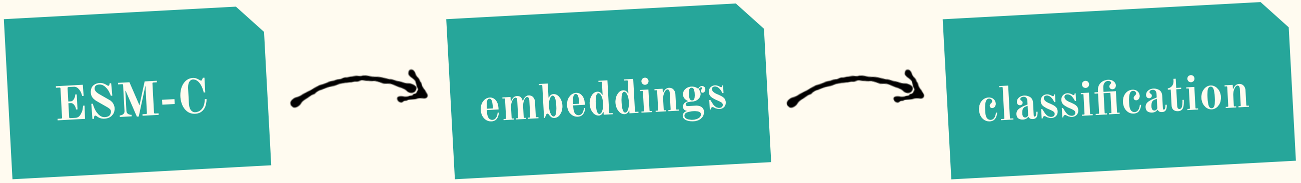
Classify Protein Functions

Phage proteins (AA) annotated as "lytic"

14k ESKAPE phage proteins
14 functional annotations
train/test 80/20%







6B parameters
ESM Team (2024)

960 features for AA
cached with sequence hash

2 Layers
Feed Forward

Prediction on test
validates model!

Accuracy ~0.99

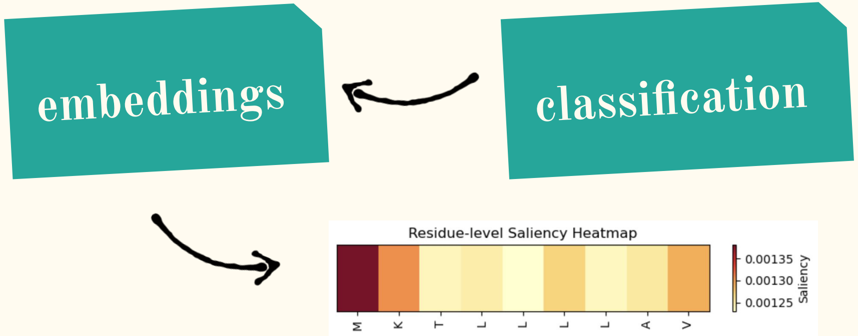
Peeking Inside the Model with Saliency Maps

⚙️ How It Works (Simplified)

1. **Backward Pass**: Compute gradients of the target class score **with respect to each input token (amino acid)**.

2. **Magnitude of Gradient**: The bigger the gradient, the more that input position affected the output.

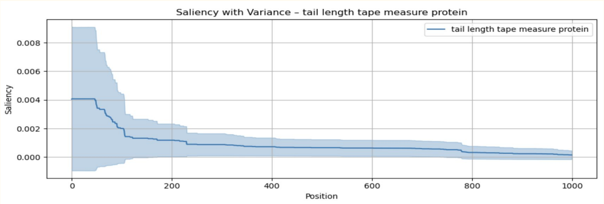
3. **Visualization**: You can plot these magnitudes (optionally multiplied by the embeddings) as a **heatmap or barplot** to see the "important" residues.



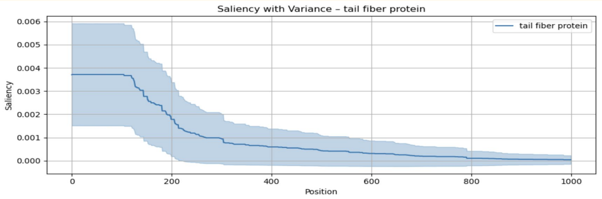
embeddings

classification

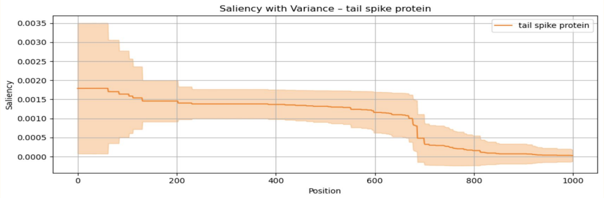
We aggregate saliency maps per class to look at patterns at amino acid level!



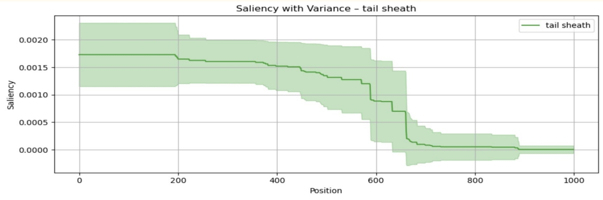
Saliency with Variance – tail length tape measure protein



Saliency with Variance – tail fiber protein

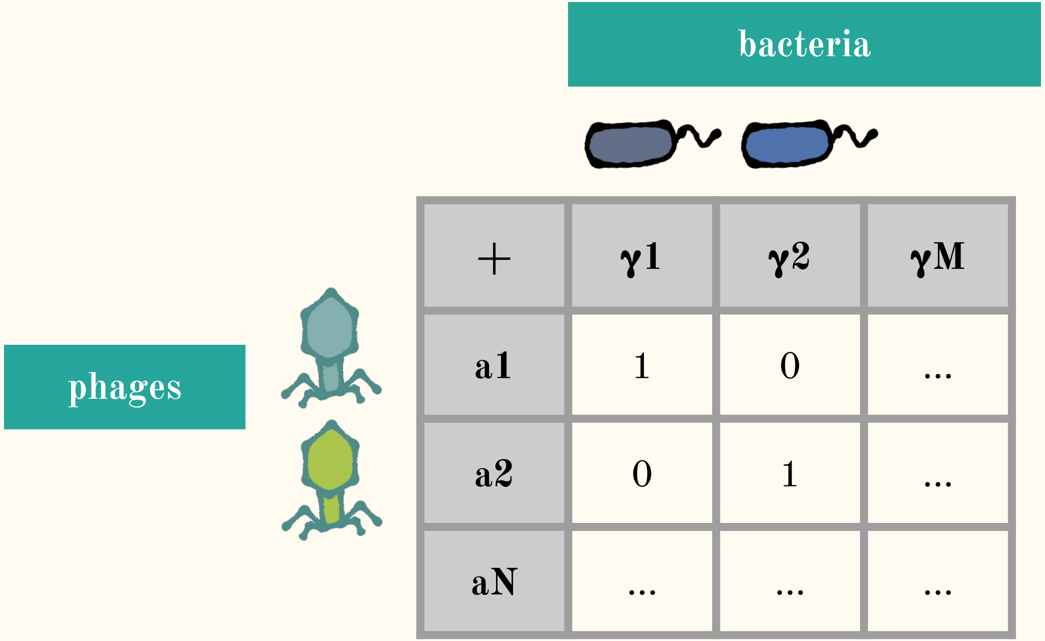


Saliency with Variance – tail spike protein



Saliency with Variance – tail sheath

Phage-Klebsiella interactions from PHL-Klebsiella



bacteria

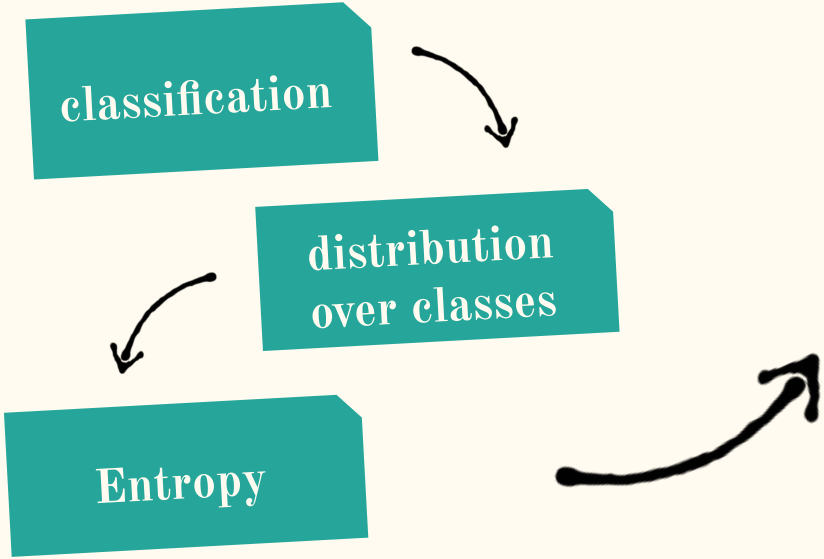
phages

+	γ1	γ2	γM
a1	1	0	...
a2	0	1	...
aN	...	...	...

- PHL-Klebsiella dataset provides an **interaction matrix** between phages and klebsiella strains (Boeckaerts et al., 2024).

- Use matrix to calculate interaction score for each phage.

Transfer Learning



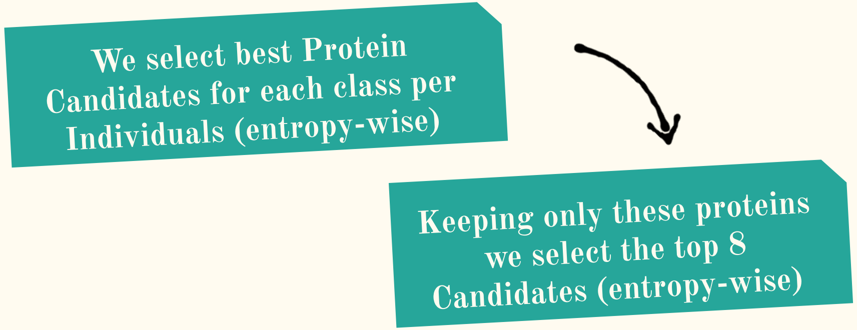
classification

distribution over classes

Entropy

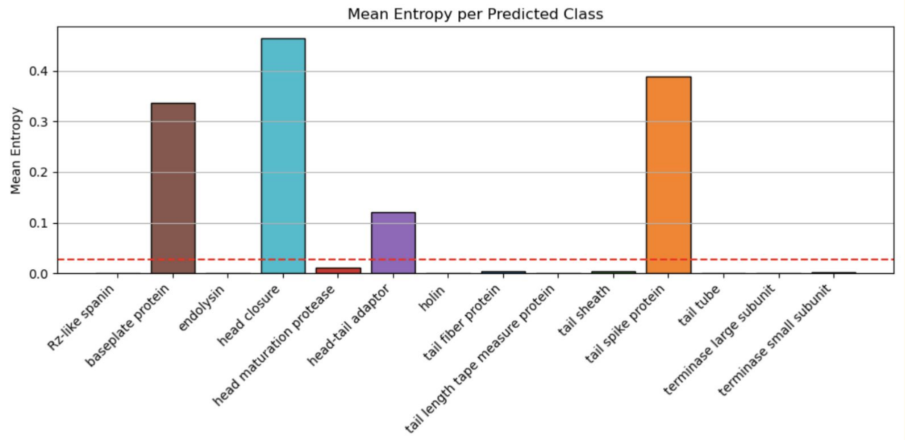
Mean Entropy over Train set is 0.027 (uniformity is 2.63 (very good))

Best individuals selection procedure for PhL-Klebsiella Dataset



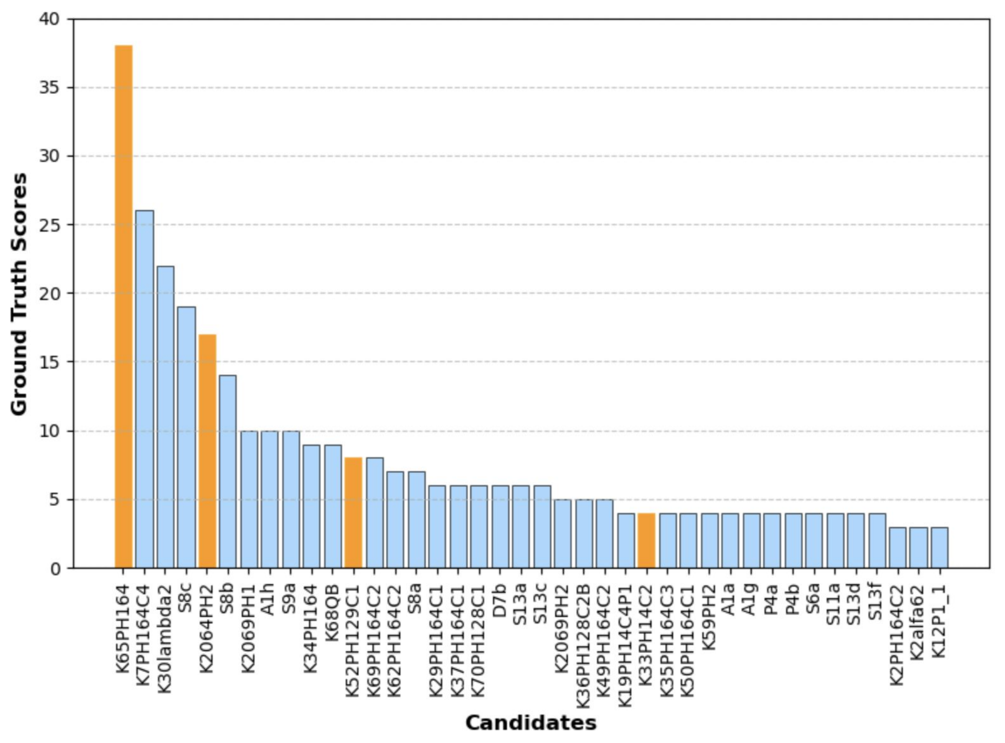
We select best Protein Candidates for each class per Individuals (entropy-wise)

Keeping only these proteins we select the top 8 Candidates (entropy-wise)



```
Ratio of entropy < 0.027: 0.8500
```

Comparing to Biological Assays



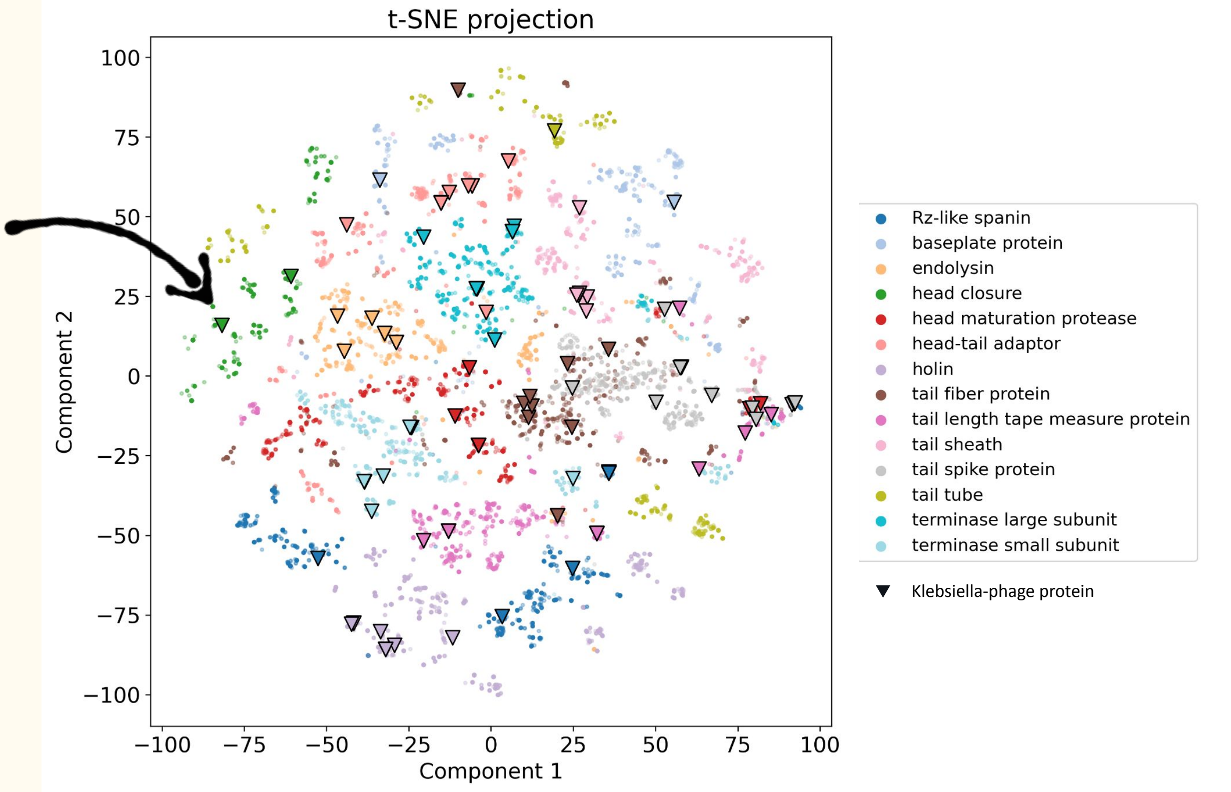
- PHL-Klebsiella phages ranked by interaction scores (top 40).

- In orange: 4 of the top 8 scorers from pipeline.

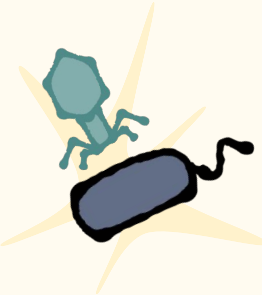
Predicted functions of Klebsiella phages

Klebsiella-Phages Function Prediction

...consistent with ESKAPE dataset!



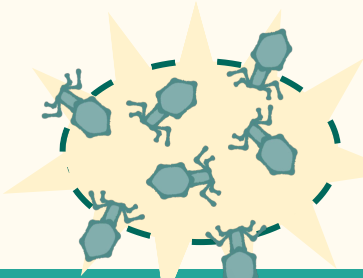
Discussion



- Even if we didn’t aim at phage-host affinity, there is some signal in the strength of **lytic activity** (to be tested against null model).

- Classification can allow for **more functional classes**, to be **better annotated** through biologically informed language models.

- Embeddings obtained with **LLM models** like ESM-C are really good at capturing motives across different proteins.



Thank you!

A, Aragon, E. de Bézenac, F. Camaglia, M. Crilout
Train Tune Deploy (aka *Team Delta)*