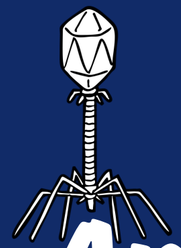
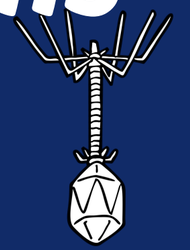
Universidade do Minho
Escola de Engenharia

CENTRE OF BIOLOGICAL ENGINEERING



Anti-Restriction: Phage Strategies to Evade Bacterial Defense Systems



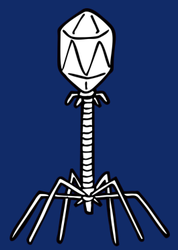
Bioinformatics Project

Master's degree in Bioinformatics

Academic Advisors: Fernanda Vieira, Ana Oliveira, Hugo Oliveira
Maria Carvalho | pg55130

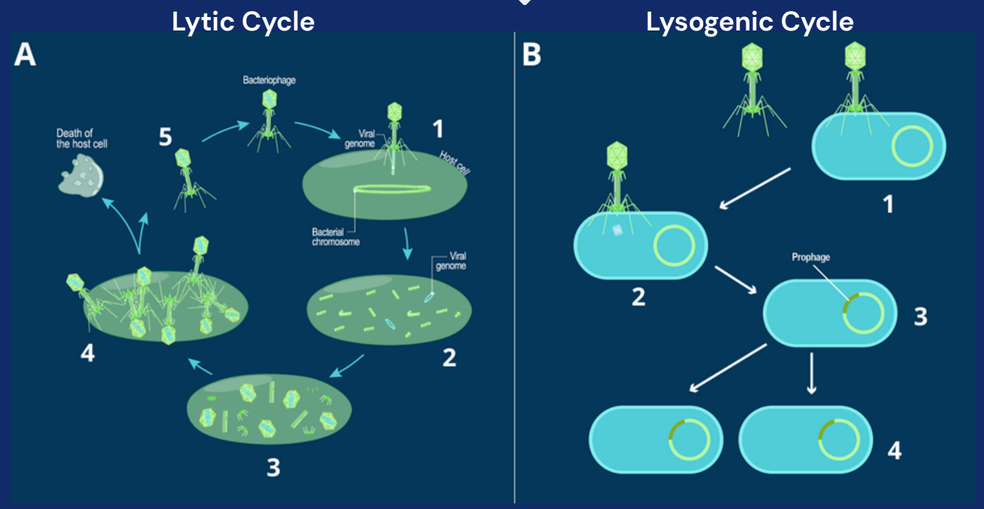
Introduction

Bacteriophages



Viruses that infect bacteria and can only replicate in bacterial cells.

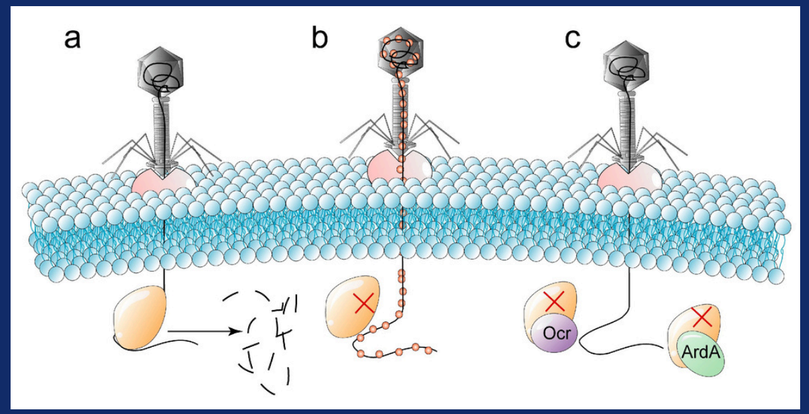
Phage - Bacteria



Phage takes control of the host's cellular machinery → replication → new viral particles.

Phage injects its DNA → integrates itself into the bacterial genome → prophage.

R-M Systems



Defends bacteria → methylate its own DNA → cleave unmethylated foreign DNA.

Introduction

R-M types

Type I	ATP-dependent, variable-distance cutter.
Type II	Mg^{2+}-dependent, fixed-site cutter.
Type III	ATP-dependent, needs two sites; hemimethylation protection.
Type IV	Cuts only modified (e.g., methylated) DNA.

Phage Evasion Strategies

Elimination of Recognition Sites	Phages modify or remove restriction enzyme recognition sequences from their genomes.
Production of Anti-Restriction Proteins	Phages encode proteins (Anti-REs) that directly inhibit host restriction enzymes.

Computational Approaches

DefenseFinder	Detect prokaryotic antiviral systems (including R-M systems) with high sensitivity (91.9% against the REBASE database).
PADLOC	Web-based platform that scans bacterial and archaeal genomes to identify antiviral defense loci.
DefensePredictor	Applies protein language model embeddings to classify defensive proteins in bacterial genomes. Although not explicitly tuned for R-M evasion, it shows promise for that use.

WorkPlan

Aim : investigate the interaction between phages and bacterial R-M systems, focusing on how phages evade these defense mechanisms

- Literature Review
- Construct a curated dataset of phage-host pairs, including associated anti-restriction proteins, to enable systematic analysis.
- Develop a data extraction script to collect enzyme attributes from the REBASE database for analyzing phage genome avoidance patterns.
- Train ML models on curated dbAPIS (a database of anti-prokaryotic immune system genes) protein sequences, phage–host metadata, and R-M system profiles to predict phage-encoded anti-restriction proteins.

Methods

Dataset Curation

Accession	Host	Lab_Host	Phage Description
PV287706.1	Escherichia coli 4s		phage Midge
PV252060.1	Escherichia coli	Escherichia coli M493	phage nithesis
PV204681.1	Escherichia coli O157		phage vB_EcoM_JQD51
PP453689.1	Escherichia coli K12 (RP4)		phage vB_EcoP_LHP
PV245935.1	Escherichia coli	Escherichia coli M491	phage kaset
PV245934.1	Escherichia coli	Escherichia coli M486	phage sutha
PV245933.1	Escherichia coli	Escherichia coli M485	phage nasanit

- Phage-host CSV
- Focused on *Escherichia coli*
- Added lab host info for enzyme matching

REBASE Data Extraction

ENZYME NAME	MICROORGANISM	RECOGNITION SEQUENCE
EciI	Escherichia coli	GGCGGA(11/9)
Eco9034II	Escherichia coli	GAAABCC
Eco13441IV	Escherichia coli	CANCATC
M.Eco9034Dam	Escherichia coli	GATC
M.Eco13441Dam	Escherichia coli	GATC
M.EcoEc67Dam	Escherichia coli	GATC

- From REBASE
- Output: Enzyme name, Microorganism, Recognition Site

Genome Search

Phage	Enzyme	Recognition Site	Start	End	Matched Sequence
PV287706.1	EciI	GGCGGA	632	638	GGCGGA
PV287706.1	EciI	GGCGGA	663	669	GGCGGA
PV287706.1	EciI	GGCGGA	1993	1999	GGCGGA
PV287706.1	EciI	GGCGGA	5595	5601	GGCGGA
PV287706.1	EciI	GGCGGA	6931	6937	GGCGGA
PV287706.1	EciI	GGCGGA	7416	7422	GGCGGA

- Translated IUPAC to regex
- Searched phage genomes for matches
- Output: phage, enzyme, recognition site, matched sequence

Methods

Enzymes Types

ENZYME NAME	MICROORGANISM	RECOGNITION SEQUENCE	ENZYME TYPE
EciI	Escherichia coli	GGCGGA(11/9)	I
Eco9034II	Escherichia coli	GAAABCC	II
Eco13441IV	Escherichia coli	CANCATC	IV
EcoHAI	Escherichia coli	YGGCCR	I
EcoNI	Escherichia coli	CCTNN^NNNAGG	I
Eco29kI	Escherichia coli	CCGC^GG	I
EcoprrI	Escherichia coli	CCANNNNNNNRTGC	I
EcoPI	Escherichia coli (PI)	AGACC(25/27)	I
EcoR124I	Escherichia coli (R124)	GAANNNNNNRTCG	I
EcoR124II	Escherichia coli (R124/3)	GAANNNNNNNRTCG	II

- Enzymes were categorized by type to facilitate subsequent prevalence analysis.

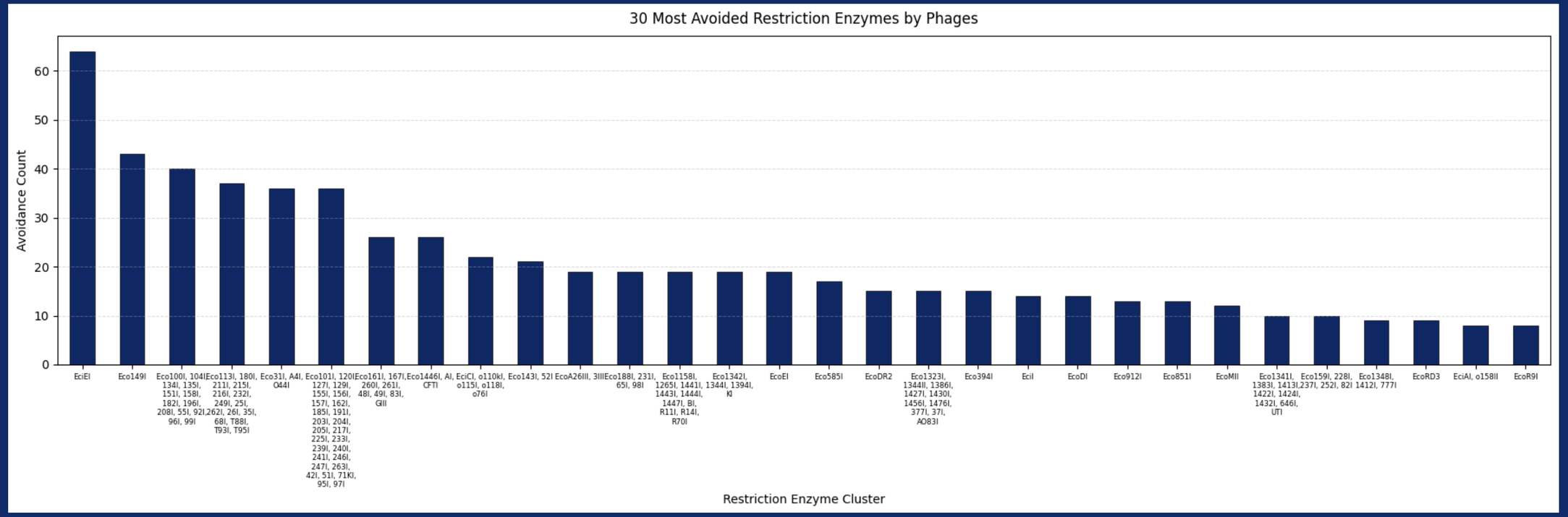
tBLASTn

```
[*] Running tblastn for each enzyme (tabular output)...

  • Enzyme: Eco9034II_GAAABCC  →  running tblastn (as 'Eco9034II_GAAABCC.tsv')... done.
  • Enzyme: Eco13441IV_CANCATC  →  running tblastn (as 'Eco13441IV_CANCATC.tsv')... done.
  • Enzyme: M.Eco9034Dam_GATC  →  running tblastn (as 'M_Eco9034Dam_GATC.tsv')... done.
  • Enzyme: M.Eco13441Dam_GATC  →  running tblastn (as 'M_Eco13441Dam_GATC.tsv')... done.
  • Enzyme: M.EcoEc67Dam_GATC  →  running tblastn (as 'M_EcoEc67Dam_GATC.tsv')... done.
  • Enzyme: Eco29kI_CCGC^GG  →  running tblastn (as 'Eco29kI_CCGC_GG.tsv')... done.
  • Enzyme: EcoprrI_CCANNNNNNNRTGC  →  running tblastn (as 'EcoprrI_CCANNNNNNNRTGC.tsv')... done.
  • Enzyme: EcoPI_AGACC(25/27)  →  running tblastn (as 'EcoPI_AGACC_25_27_.tsv')... done.
  • Enzyme: EcoR124II_GAANNNNNNNRTCG  →  running tblastn (as 'EcoR124II_GAANNNNNNNRTCG.tsv')... done.
  • Enzyme: M.Eco08Dcm_CCWGG  →  running tblastn (as 'M_Eco08Dcm_CCWGG.tsv')... done.
  • Enzyme: Eco1524I_AGG^CCT  →  running tblastn (as 'Eco1524I_AGG_CCT.tsv')... done.
  • Enzyme: EcoAI_GAGNNNNNNNGTCA  →  running tblastn (as 'EcoAI_GAGNNNNNNNGTCA.tsv')... done.
  • Enzyme: Eco4465II_GAAABCC  →  running tblastn (as 'Eco4465II_GAAABCC.tsv')... done.
  • Enzyme: M.Eco3936Dam_GATC  →  running tblastn (as 'M_Eco3936Dam_GATC.tsv')... done.
  • Enzyme: Eco9276II_CRARCAG  →  running tblastn (as 'Eco9276II_CRARCAG.tsv')... done.
  • Enzyme: Eco95NR1II_CRARCAG  →  running tblastn (as 'Eco95NR1II_CRARCAG.tsv')... done.
  • Enzyme: M.Eco3165Dam_GATC  →  running tblastn (as 'M_Eco3165Dam_GATC.tsv')... done.
  • Enzyme: M.EcoA13Dam_GATC  →  running tblastn (as 'M_EcoA13Dam_GATC.tsv')... done.
```

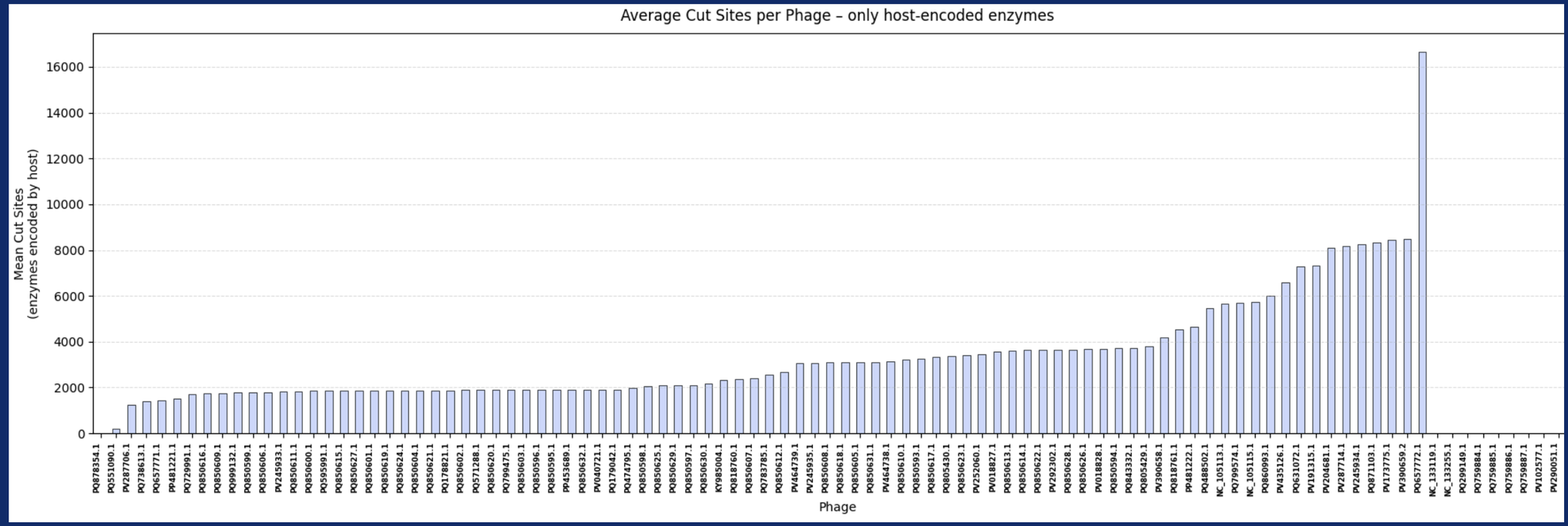
- Identifying which enzymes are most widespread in bacteria—and which bacteria share those enzymes

Results



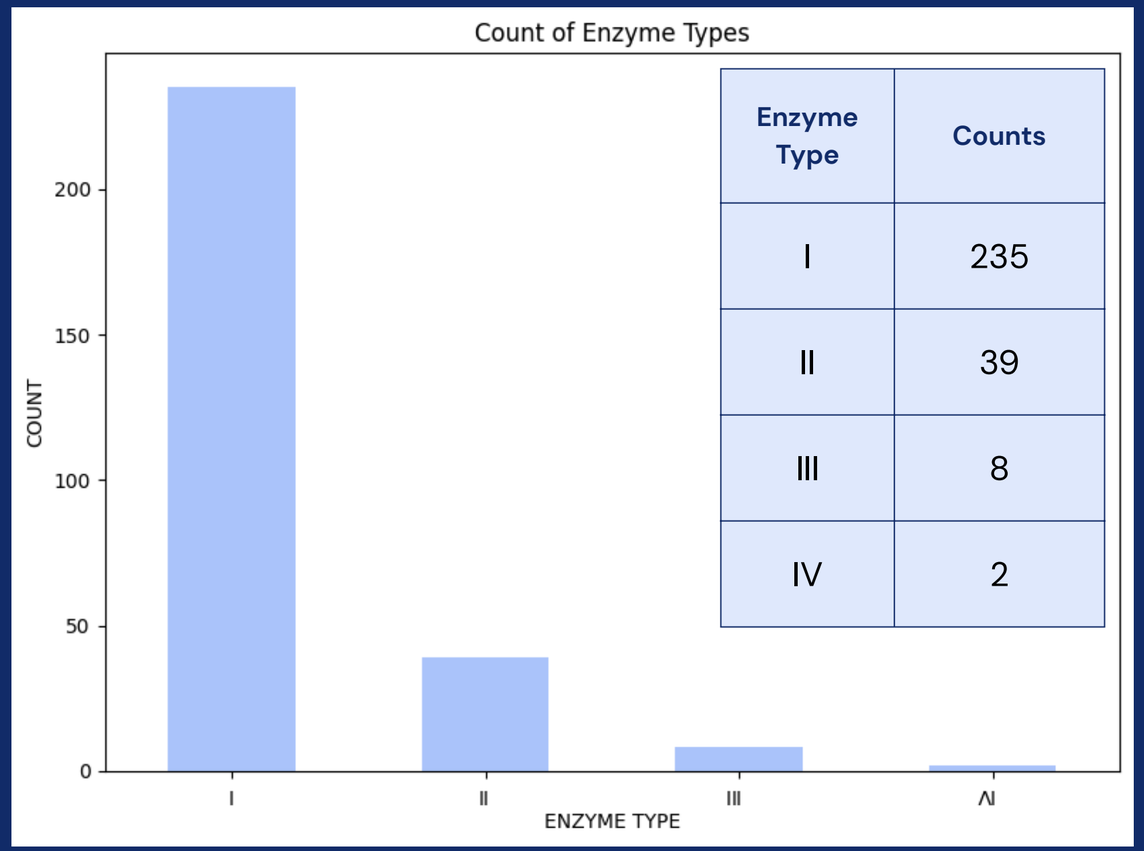
- Enzymes grouped by common recognition site.
- Highest-avoided is EciEI ~65.
- There is a steep drop from ~65 avoidance down to the low-40s (Eco149I)
- Several multi-enzyme clusters appear around 37.
- Avoidance counts continue to taper off, with many clusters around 15–20 avoidances, and the least-avoided clusters each around 8–10.

Results



- Total phages analyzed: 88 (out of 98).
- One phage sits around 16000 mean cut sites, towering above all others.
- A small group of phages falls in the 4,000–5,000 mean cut-site window.
- Most phages have between 1,000 and 3,000 mean cut sites.
- The bulk of phages cluster in the 1,000–2,500 window.

Results



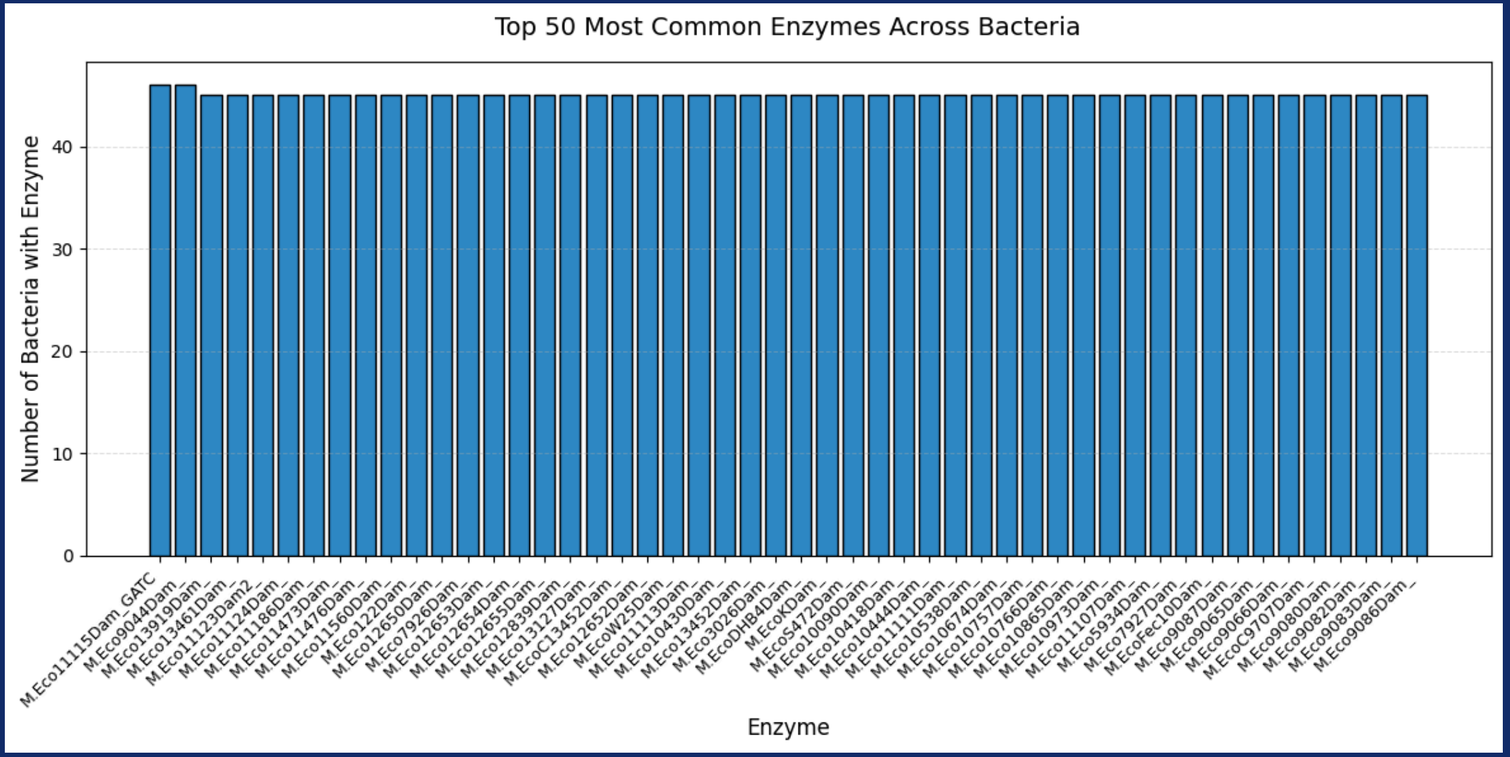
Enzyme Type	Counts
I	235
II	39
III	8
IV	2

- Type I → vast majority of enzymes in this dataset (235)
- Type II → less common (39)
- Type III → rare (8)
- Type IV → just appears 2 times

Type I enzymes dominate the sample, while the other three classes are present only in very small numbers.

Results

tBLASTn Survey of Enzyme Prevalence Across Bacterial Genomes



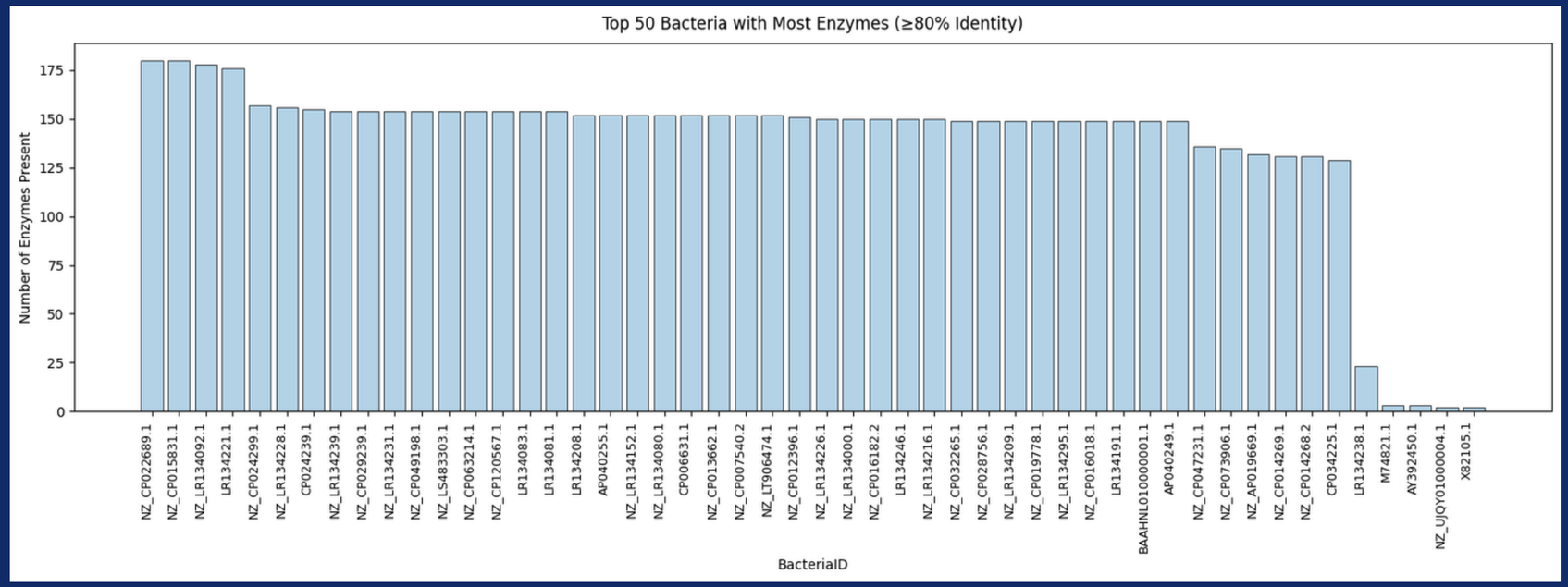
- Every enzyme is found in ~ 45 different bacterial genomes. → Dam-type enzymes predominate.

This uniformity underscores how highly conserved and widespread the Dam→GATC methylation system is across bacteria.

Results

tBLASTn Survey of Enzyme Prevalence Across Bacterial Genomes

- Top four genomes have ~176–178 enzymes, dropping to ~154–157 and tapering to ~130–135. → Counts drop to ~23 , showing a few genomes hold most hits while the rest have far fewer.



Only a handful of bacterial genomes encode the vast majority of these enzyme hit.

Conclusions

Within our curated REBASE subset 235 Type I enzymes were identified, 39 Type II, 8 Type III, and 2 Type IV, underscoring that most anti-restriction effort by phages is directed at Type I systems.

Because Type IV enzymes are underrepresented, we lack sufficient training examples—and will therefore exclude them to prevent overfitting and bias.

Our *E. coli* phage data show that avoiding conserved Dam-methylation sites is a common evasion tactic.

Although many bacteria carry R–M enzymes, a small number of strains account for most variants (> 170 hits)—phages must optimize evasion for these key hosts.

Future Perspectives

Compute sequence-depletion scores and other genomic features from our phage dataset → train a machine-learning model that predicts anti-restriction fuctions in novel phage genomes

Annotate phage proteomes for known Anti-RE families (e.g., ArdA, Ocr, Stp) → reveal which phages rely on active inhibition versus passive site depletion

Curate additional phage–host pairs across multiple genera → avoidance patterns and the prevalence of certain Anti-REs are conserved across diverse bacterial clades