Group Brainstorm on Bacteriophage Engineering

Group Members: Sami Ur Rehman (2026a-sami-ur-rehman) and Edna Wanjiru Macharia (2026a-ednah-wanjiru)

Date: March 12th, 2026

Acknowledgments & Context

I want to begin by acknowledging that this assignment has been challenging for me. As someone with limited background in AI/ML and virology, digesting the dense research papers and understanding the computational tools required significant effort. I took help from **DeepSeek** to break down complex concepts, explain the tools, and guide me through the literature. This was essential because I was lagging in the course—I was busy with other commitments for last three weeks, and this Week 4 homework is now due when Week 7 is already halfway completed.

I also want to give credit to my batchmate **2026a-nourelden-rihan**, who is part of both the HTGAA course and the BioClub Tokyo node. I visited his published HTGAA page to understand the project expectations and see how he structured his approach. His group's proposal was particularly well-organized, and I built upon his framework while adding my own contributions. His work helped clarify concepts that I was struggling with, and I'm grateful for the collaborative spirit of our node.

With that context, here is my refined proposal, which combines the strongest elements from both his work and my own reading of the papers.

---

1. Selected Goals & Strategy

I will target the **MS2 L lysis protein** (75 amino acids), a prototypical amurin (single-gene lysis protein), with a dual engineering strategy:

- **Primary Goal (Medium difficulty): Increase lytic toxicity.** I aim to achieve faster, more complete bacterial lysis, leading to higher phage titers: a key

metric for therapeutic manufacturing and a direct path toward addressing the "higher toxicity" challenge from the project goals.

- **Secondary Goal (Enables primary goal): Modulate DnaJ interaction.** Rather than simple deletion of the N-terminal domain (which causes earlier lysis but removes all regulation), I will **redesign the N-terminal interface** to finely control or eliminate binding to the host chaperone DnaJ. This builds directly on **Chamakura, Tran, et al. (2017)** (*MS2 Lysis of Escherichia coli Depends on Host Chaperone DnaJ*), which shows that the N-terminus confers DnaJ dependence, and **Chamakura, Edwards, et al. (2017)** (*Mutational Analysis of the MS2 Lysis Protein L*), which identifies key residues through mutagenesis.

Design Constraints (Informed by Literature):

- **Must protect** the essential C-terminal transmembrane domain (residues 46–75) and the **LS motif (Leu48-Ser49)**, which **Chamakura, Edwards, et al. (2017)** shows is critical for function—most loss-of-function mutations cluster here, and even conservative substitutions (L44V, S49T) abolish lysis.
- **Must maintain** the ability to form high-order oligomers (>10 subunits) for pore formation, as demonstrated by cryo-EM and native mass spectrometry in **Mezhyrova et al. (2023)** (*In vitro characterization of the phage lysis protein MS2-L*).
- **Must avoid** disrupting overlapping reading frames: the *L* gene overlaps with the **Coat** and **Replicase** genes. Any engineered mutation must be cross-checked to ensure it does not introduce synonymous or non-synonymous changes that break these essential phage proteins.
- **Must preserve** membrane insertion competence: **Mezhyrova et al. (2023)** shows that membrane association is required for function, and mutations must not mislocalize the protein.

---

2. Computational Pipeline

I will employ a **sequence-to-structure-to-interaction** workflow that integrates evolutionary analysis, generative design, and structural validation. This pipeline was developed with guidance from DeepSeek to help me understand each tool's purpose and how they connect.

Stage 1: Evolutionary Profiling (BLAST + Clustal Omega)

- Identify remote homologs of single-gene lysis proteins from related leviviruses.
- Use Microviridae dataset as inspiration for finding related sequences. (King et al., 2025)
- Generate Multiple Sequence Alignment (MSA) to map:
    - Invariant residues (LS motif, pore-forming helix) that must remain untouched.
    - Variable regions (N-terminal domain candidates for modification)
    - Naturally occurring variations that might already alter DnaJ binding.
- This establishes the "allowed" sequence space and prevents mutating evolutionarily constrained positions.

Stage 2: Generative Mutagenesis (Evo 2)

- Use protein language models to perform *in silico* saturation mutagenesis across the N-terminal domain (residues 1–36)
- **Specifically leverage Evo 2**, which **King et al. (2025)** shows was trained on >2 million phage genomes and can design functional phage sequences—making it uniquely suited for proposing mutations that "look phage-like".
- Generate mutation heatmaps predicting:
    - Evolutionary plausibility (fitness scores from masked language modeling)
    - Structural stability (via ESMFold pLDDT scores)

- Prioritize mutations that appear in multiple homologous sequences (from Stage 1) to ensure natural precedent.

Stage 3: Structure Prediction & Filtering (ESMFold)

- Model top one hundred candidates (single mutants and small combination libraries)
- Filter for:
    - High pLDDT (>70) in C-terminal domain (ensuring the essential region remains folded)
    - Preserved helical geometry in transmembrane region.
    - No predicted distortion of the LS motif or pore-forming helix
- This stage eliminates mutations that destabilize the core fold.

Stage 4: Aggregation Risk Assessment (TANGO / AGGRESCAN)

- Screen candidates for aggregation propensity.
- Filter out mutants with high aggregation scores, which could lead to non-functional inclusion bodies or improper membrane insertion.
- This step is critical because **Mezhyrova et al. (2023)** shows that L must remain membrane-competent, not aggregated.

Stage 5: Complex & Oligomer Prediction (AlphaFold-Multimer)

- Model mutant L proteins in complex with *E. coli* DnaJ (using DnaJ structure from AlphaFold DB or PDB: 1XBL)
- Select variants with:
    - Reduced interaction scores (low ipTM, high PAE at interface)—indicating successful DnaJ modulation
    - Maintained ability to form high-order oligomers (>10 subunits)—cross-validate with **Mezhyrova et al. (2023)'s** finding that oligomerization is driven by transmembrane domain

- Also, model L alone to ensure the monomeric structure is compatible with membrane insertion.

Stage 6: Overlapping Gene Constraint Check

- Back-translate mutant protein sequences to DNA
- Verify that codon changes do not introduce:
    - Premature stop codons in the Coat or Replicase reading frames
    - Deleterious missense mutations in those essential genes
- This is unique to MS2 biology and essential for any future incorporation into a full phage genome.

Stage 7: Final Selection

- Choose 10–15 top candidates balancing:
    - Reduced DnaJ binding (from Stage 5)
    - Maintained structural stability (from Stage 3)
    - Low aggregation risk (from Stage 4)
    - Preserved pore-forming capability (from Stage 5)
    - No disruption to overlapping genes (from Stage 6)

---

3. Rationale for Tool Selection

Tool	Why It Fits
BLAST + Clustal Omega	Helps me find related lysis proteins in nature and see which parts of L are "off-limits" for mutation. If a residue is identical across many species, it's probably important.
Evo 2	Trained on millions of phage genomes King et al. (2025), so it understands what "phage-like" sequences look like.

	It can suggest mutations that nature might approve of.
ESM-2 / ESMFold	Fast way to check if my mutant proteins will still fold correctly. High pLDDT scores mean the model is confident in the structure.
TANGO / AGGRESCAN	Checks if my mutants might clump together (aggregate). Aggregated proteins can't insert into membranes.
AlphaFold-Multimer	Predicts whether my mutant L still binds DnaJ (we want reduced binding) and whether it can still form pores (>10 subunits).
Codon optimization tools	Ensures I don't accidentally break the overlapping Coat and Replicase genes when I change L's DNA sequence.

---

4. Potential Pitfalls & Mitigations

Pitfall	**Mitigation (How** I will **address it)**
Limited structural data for L-DnaJ complex	No experimental structure exists. Use AF-Multimer predictions, but acknowledge this as a limitation. Plan experimental validation.
Model bias toward globular proteins	Use Evo 2 (phage-trained) instead of generic models; add membrane topology checks (TMHMM).
Poor annotation of amurins in databases	Supplement BLAST with sequences from King et al. (2025)'s dataset and recent NCBI phage collections.
Over-stabilization → loss of function	Apply dual filters: require good stability *but* avoid mutations that lock the protein rigidly. Screen for aggregation.

Context-dependent interactions	Cannot fully model membrane environment; plan experiments in lipid nanodiscs. **(Mezhyrova et al., 2023)**
Disruption of overlapping genes	Always back-translate and check Coat/Rep reading frames before final selection.
Host protease susceptibility	Use tools like PeptideCutter to screen for new cleavage sites; plan half-life experiments.
Lysis timing too early or too late	Our "redesign" approach aims for tunable control, not just elimination. Test multiple variants.

---

5. Expected Outcomes & Applications

I expect to generate MS2 L variants with:

- **Faster lysis kinetics** compared to wild-type (measured by OD600 drop)
- **Reduced DnaJ binding** (validated by pulldown assays)
- **Maintained pore formation** (confirmed by cryo-EM in nanodiscs, following Mezhyrova et al. (2023)
- **No disruption to overlapping genes** (verified by sequencing)

Potential Applications:

- **Synthetic phage cocktails** with tunable lysis timing for therapeutic use (addressing needs from **Strathdee et al. (2023) & (*Phage Therapy: Past, Present and Future | ASM.org*, 2026)** on phage therapy)
- **Bacterial ghost production** for vaccines (faster lysis = more efficient production)
- **Study of lysis regulation** in phage infection cycles

---

6. Validation Roadmap (Future Work)

1. Clone top fifteen mutants into expression system (as in Chamakura, Tran, et al. (2017)
2. Express in *E. coli* (wild-type and *DnaJ* knockout strains)
3. Measure lysis kinetics (OD600 over time)
4. Assess protein stability (Western blot half-life)
5. Test DnaJ interaction (pulldown assays)
6. Visualize pore formation (cryo-EM in nanodiscs)
7. Check membrane localization (subcellular fractionation)
8. Test in full-phage context (if time permits)
9. Iterate design based on results

---

7. Pipeline Schematic

[WT MS2 L Sequence (75 aa)]

↓

Stage 1: BLAST + Clustal Omega → Find conserved vs. mutable regions

↓

Stage 2: Evo 2 Mutagenesis → Generate mutation heatmaps (N-terminal domain)

↓

Stage 3: ESMFold Filtering → Keep mutants with good structure (pLDDT >70)

↓

Stage 4: TANGO/AGGRESCAN → Remove aggregation-prone mutants

↓

Stage 5: AlphaFold-Multimer → Check DnaJ binding + pore formation

↓

Stage 6: Overlapping Gene Check → Ensure Coat/Rep genes are safe

↓

Stage 7: Final Selection → 10–15 top candidates for experiments

---

References:

Chamakura, K. R., Tran, J. S., & Young, R. (2017c). MS2 Lysis of Escherichia coli Depends on Host Chaperone DnaJ. *Journal of Bacteriology*, *199*(12). https://doi.org/10.1128/jb.00058-17

Chamakura, K. R., Edwards, G. B., & Young, R. (2017b). Mutational analysis of the MS2 lysis protein L. *Microbiology*, *163*(7), 961–969. https://doi.org/10.1099/mic.0.000485

Mezhyrova, J., Martin, J., Börnsen, C., Dötsch, V., Frangakis, A. S., Morgner, N., & Bernhard, F. (2023b). In vitro characterization of the phage lysis protein MS2-L. *Microbiome Research Reports*, *2*(4), 28. https://doi.org/10.20517/mrr.2023.28

Strathdee, S. A., Hatfull, G. F., Mutalik, V. K., & Schooley, R. T. (2023b). Phage therapy: From biological mechanisms to future directions. *Cell*, *186*(1), 17–31. https://doi.org/10.1016/j.cell.2022.11.017

King, S. H., Driscoll, C. L., Li, D. B., Guo, D., Merchant, A. T., Brixi, G., Wilkinson, M. E., & Hie, B. L. (2025). *Generative design of novel bacteriophages with genome language models*. bioRxiv. https://doi.org/10.1101/2025.09.12.675911

Barron, M. (2022, August 31). *Phage therapy: Past, present and future*. American Society for Microbiology. https://asm.org/articles/2022/august/phage-therapy-past,-present-and-future

Personal Note

I want to be honest: this assignment pushed me far outside my comfort zone. The combination of virology, protein engineering, and AI/ML tools was overwhelming at first. I relied on DeepSeek to help me understand each paper, explain what tools like ESMFold and AlphaFold do, and connect the concepts to the project goals. I

also learned a lot by looking at how my batchmate 2026a-nourelden-rihan structured his proposal—his clarity helped me see what a strong submission looks like.

I am a committed listener from Pakistan, part of the BioClub Tokyo node, and I am determined to catch up. This Week 4 homework is my stepping stone. I am proud of what I have produced, even if the journey was difficult, and I am grateful for the supportive community that makes HTGAA possible.