Science
AAAS

Supplementary Materials for

Generative design of bacteriophages with genome language models

Samuel H. King *et al.*

Corresponding author: Brian L. Hie, brianhie@stanford.edu

Science **393**, eaec2657 (2026)
DOI: 10.1126/science.aec2657

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Materials and Methods

Generative modeling of bacteriophage genomes

Generating sequences with Evo

Sequence sampling was performed using a standard autoregressive sampling algorithm with the sampling code from https://github.com/evo-design/evo/ and https://github.com/arcinstitute/evo2, which leverages kv-caching of Transformer layers (*93*) and the recurrent formulation of Hyena layers for efficient, low-memory autoregressive generation (*26*, *28*, *94*). Given that Evo's vocabulary spans beyond the four nucleotides A, C, G, and T, unless otherwise noted, we sampled sequences with top-k = 4 and top-p = 1, to maximize next-token prediction of nucleotide characters only.

Generating sequences with bacteriophage realm prompts

For Evo 1 7B 131K, approximately 10,000 sequences of length 6,000 were sampled with temperature = 0.7, top-k = 4, and top-p = 1 for each prompt. Sequences were generated using the prompts `|r__Duplodnaviria;k__`, `|r__Monodnaviria;k__`, and `|r__Riboviria;k__`. For Evo 2 7B 1M, approximately 1,000 sequences of length 6,000 were sampled with temperature = 0.7, top-k = 4, and top-p = 1 for each prompt. Sequences were generated using the prompts `|r__Duplodnaviria;;;;;;;|`, `|r__Monodnaviria;;;;;;;|`, and `|r__Riboviria;;;;;;;|`.

Generated bacteriophage realm sequence analyses

For use as positive controls, natural sequences corresponding to the realms *Duplodnaviria*, *Monodnaviria*, and *Riboviria* were extracted from the OpenGenome stage 2 training dataset (*28*) by filtering entries whose taxonomy strings began with `|r__Duplodnaviria`, `|r__Monodnaviria`, or `|r__Riboviria`, respectively. From each realm, 10,000 sequences were randomly sampled. Sequences containing only canonical nucleotides (A, C, G, T) were retained for downstream use. For use as negative controls, scrambled natural sequences were generated by randomly permuting the nucleotide order of each sequence using a fixed random seed (42) to preserve overall nucleotide composition while ablating higher-order structure.

To analyze divergence of generated sequences from natural sequences, the generated sequences were searched against natural sequences in the core_nt database by nucleotide BLAST (blastn, E-value = 0.5, default settings) (*43*). Query coverage and percent identity of the top 100 target hits were collected for each generated sequence. To determine the proportion of generated sequences that classify as viral, the sequences were analyzed by geNomad version 1.8.0 (*42*). Coding density was determined by predicting ORFs using Prodigal version 2.6.3 (*95*) and dividing the total sum of ORF lengths per sequence with the total length of the sequence. Phage-like coding sequences were predicted with our gene annotation method (see *ORF calling and gene annotation*) and visualized by LoVis4u with flags `-hl`, `--set-category-colour`, `-c A4p2`, `-alip` (*96*). The pTM and mean pLDDT scores of predicted proteins were extracted from their structures folded by ESMFold (*44*), specifically the Hugging Face implementation of the `facebook/esmfold_v1` checkpoint accessed via the Transformers library. Self-consistency TM (scTM) scores were determined by inverse-folding ESMFold-predicted protein structures using ProteinMPNN (model v_48_020, sampling temperature 0.1, 1 sequence per target) (*17*) to produce a redesigned amino acid sequence. The redesigned sequence was re-predicted with ESMFold, and the TM-score between the original and re-predicted structures was computed using TM-align (tmtools),

normalized by the length of the original structure. Percent protein sequence identities of the generated proteins were determined by aligning them against Prodigal-predicted proteins in OpenGenome (*28*), and proteins in the PHROGs database (*45*) with MMseqs2 version 13.45111 (*97*) with sensitivity = 4.0. Functional annotations curated for PHROGs (*45*) were extracted for generated proteins with significant alignments against proteins in the PHROGs database.

Microviridae data

A total of 15,507 *Microviridae* genomes were collected from public databases as following: 4,498 genomes were downloaded with their associated metadata from NCBI Datasets on July 8, 2024 with the keyword 'Microviridae' and the filter 'complete' (*98*), 130 genomes were downloaded from the PhageScope database on July 31, 2024 with default filters, the taxonomy 'Microviridae', and sequence quality 'Complete' (*99*), and 10,879 genomes of the virus family 'Microviridae' were downloaded from the OpenGenome database (*26*, *28*). 261 genomes over 10 kb in length or containing characters other than 'A', 'C', 'G', or 'T' were removed. The filtered sequences were clustered at 99% identity by MMseqs2 version 13.45111 (*97*) and a total of 14,466 final sequences were extracted. The relative proportions of available sequencing data for bacteriophage families were analyzed from 'Complete' genomes collected from NCBI Virus on May 12, 2025 (*100*). For analyses involving ΦX174 variants, a total of 134 ΦX174 variant sequences were extracted from the raw *Microviridae* data by searching for the keyword 'phix174'.

Data preprocessing for supervised fine-tuning

Microviridae sequences were aligned to the ΦX174 NCBI reference sequence NC_001422.1 using `align.align_optimal` from the biotite Python package version 0.39.0 (*101*) and the sequences were prepended with tokens indicating their percent sequence identity to ΦX174. All sequences were prepended with the token "+" indicating *Microviridae*. After the "+" token, sequences with 95–100% identity to ΦX174 were prepended with "~", 80–95% with "^", 70–80% with "#", 50–70% with "$", and <50% with "!". Note that this soft-prompting strategy was not used for the final generation of generated phage candidates; however, the tokens "+~" were prepended to every prompt before generation. Before fine-tuning, phage genomes were randomly split into 14,266 training sequences, 100 validation sequences, and 100 test sequences.

Supervised fine-tuning of Evo 1

A full fine-tune of the Evo 1 7B 131K model was conducted for 5,000 iterations across 16 Nvidia H100 GPUs, with a batch size of 64 samples and a context length of 10,240 tokens, corresponding to a global batch size of 655,360 tokens. Each sample corresponded to a single phage genome, including special prompt tokens, where sequences shorter than 10,240 tokens were padded up to the context length and where the training loss was only defined on the sequence, excluding pad tokens. Most of the hyperparameters used during pretraining were retained (*28*) but the initial learning rate was set to 0.00009698 with linear warmup through 5% of the fine-tuning iterations followed by cosine decay through the remaining fine-tuning iterations to a minimum learning rate of 0.00003.

Supervised fine-tuning of Evo 2

A full fine-tune of the Evo 2 7B 8K model was conducted for 12,000 iterations across 32 Nvidia H100 GPUs, with a batch size of 32 samples and a context length of 10,240 tokens, corresponding to a global batch size of 327,680 tokens. Each sample corresponded to a single phage genome, including special prompt tokens, where sequences shorter than 10,240 tokens were padded up to the context length and where the training loss was only defined on the sequence, excluding pad tokens. Most of the hyperparameters used during pretraining were retained (*26*) but the initial learning rate was set to 0.00001 with linear warmup through 5% of the fine-tuning iterations followed by cosine decay through the remaining fine-tuning iterations to a minimum learning rate of 0.000001.

Positional entropy analysis

Per-position entropy was calculated as the entropy of the conditional probability $p(x_i \mid x_1, \ldots, x_{i-1})$ predicted by Evo at each position i for token x_i. The positional entropies of all ΦX174 variants extracted from the *Microviridae* dataset were calculated by Evo 1 7B 131K, Evo 2 7B 8K, and training checkpoints of the Evo 1 SFT and Evo 2 SFT models at 5K steps and 12K steps, respectively, using a custom script (**Data, code, and materials availability**). Per-position entropies were averaged across sequences of length 5386 and smoothed with a Savitzky–Golay filter, with a window length of 10 and polynomial order of 3.

Design constraints

Sequences were subject to a variety of nucleotide- and gene-level design constraints to determine their biological feasibility (**Data, code, and materials availability**). The design constraints were separated into three categories based on their purpose: quality control, host tropism specificity, and sequence diversification. Quality control constraints included filtering for sequences with only characters A, C, G, T, sequence length of 4–6 kb, 30–65% GC content, DNA homopolymer length of ≤ 10 nt, and predicted protein hit count ≥ 7 (**fig. S7A**). Protein sequences were predicted using our custom gene annotation method (see *ORF calling and gene annotation*). Sequences retained after applying quality control filters were passed to the tropism filter, which kept sequences containing a spike protein with ≥ 60% protein sequence identity to ΦX174 spike protein, determined by MMseqs2 version 13.45111 (*97*) with sensitivity = 4.0 (**fig. S7A**). The tropism-filtered sequences were then manually inspected and selected for DNA synthesis or passed to further diversification filters, each applied separately, including architectural similarity score ≤ 0.9, average amino acid identity (AAI) ≤ 95%, synteny break of one gene and total gene count of 10 or 12 genes (**fig. S7A**). The diversity-filtered sequences were then manually inspected and selected for DNA synthesis. A detailed breakdown of the DNA synthesis batches and steps taken throughout the design filtering process is described in **Figure S7C**. Generated sequences were additionally analyzed, but not filtered with, CheckV version 0.7.0 (*51*) with default settings, to determine viral sequence quality. Sequences were visualized by LoVis4u with flags `-hl`, `--set-category-colour`, `-c A4p2`, `-alip` (*96*) to aid manual inspection.

ORF calling and gene annotation

To enable accurate gene-level design constraints, various gene prediction methods were compared using the reference ΦX174 genome NC_001422.1. Prodigal version 2.6.3 was run with the flag `-p meta` (*95*). Pyrodigal-gv version 0.3.2 was run with default settings (*42*). Phanotate version 1.6.5 was run with default settings (*102*). pLannotate was run via the web server

(http://plannotate.barricklab.org/) using default parameters (*103*). Glimmer was run in Geneious Prime version 2025.1.2 with the following parameters: minimum gene length = 90, maximum overlap length = 1200, start codons = ATG, stop codons = TAG/TGA/TAA, and genetic code = 11 (*104*). GeneMark annotation was performed using MetaGeneMark version 3.42 via the web server (https://genemark.bme.gatech.edu/heuristic_gmhmmp.cgi) with default parameters (*105*). These methods failed to predict all 11 genes in the ΦX174 genome, which prompted us to create our own method tailored to ΦX174 (**fig. S4C**; **Data, code, and materials availability**). First, genomes were "pseudo-circularized", by searching for the first stop codon position in each reading frame, identifying the most downstream first stop codon position, extracting the sequence up to that position, and appending it to the end of the genome. Then, all possible ORFs were determined using orfipy (*106*) with the input start codon ATG, and stop codons TAA, TAG, and TGA, minimum and maximum ORF lengths of 90 and 1800 nucleotides, respectively, and strand 'f'. Finally, an all-by-all search was performed against the PHROGs database (*45*) using MMseqs2 version 13.45111 (*97*) with sensitivity = 4.0. For each hit against the PHROGs database, only the most significant hit (lowest E-value) was kept as the predicted annotation.

Eukaryotic viral CDS annotation analysis

As a biosecurity measure, the gene annotation method was applied to eukaryotic viruses and evaluated for prediction accuracy. Eukaryotic viral genomes were curated from Virus-Host DB (*107*) by downloading the entirety of the database on December 6, 2025, and retaining only genomes with host lineage starting with "cellular organisms; Eukaryota". CDSs were retrieved for each genome from the associated gbff file using their assigned RefSeq IDs. For genomes with multiple RefSeq IDs, the first ID listed was used. Genomes with CDS counts of 0 were removed from the data, resulting in a final set of 9806 eukaryotic viral genomes. As a test with the same settings tailored to ΦX174, all possible ORFs were determined using orfipy (*106*) with the input start codon ATG and stop codons TAA, TAG, and TGA, minimum and maximum ORF lengths of 90 and 1800 nucleotides, respectively, and strand 'f'. An all-by-all search was performed against the PHROGs database (*45*) using MMseqs2 version 13.45111 (*97*) with sensitivity = 4.0. To test the ability of the method to calibrate to eukaryotic viral genomes, ORFs were determined with input start codons ATG, TTG, CTG, GTG, and stop codons TAA, TAG, TGA, minimum and maximum ORF lengths of 90 and 6000 nucleotides, respectively, and strand 'b'. and all-by-all searches were performed against the PHROGs database or against viral proteins from the Swiss-Prot and TrEMBL databases downloaded on February 1, 2025 (*108*). Accuracy was measured as the percent difference in the number of predicted CDSs to the number of ground truth CDSs: 100 × ((predicted CDS count - true CDS count) / true CDS count)).

Genetic architecture similarity scoring

To evaluate architectural changes in genome sequences (i.e., the arrangement of open reading frames) compared to the reference ΦX174 genome NC_001422.1, we created a genetic architecture scoring algorithm (**fig. S5**; **Data, code, and materials availability**). Since local and global sequence alignment methods based on dynamic programming algorithms are compute-inefficient (*97*), and percent identity calculated from sequence alignment does not directly inform changes in genetic architecture, we reasoned that a simple dot product scoring method would be a fast way to broadly determine changes in the genetic architectures of sequences. The genes in ΦX174 begin with an ATG start codon and end with a stop codon (TAA, TAG, TGA). Thus, we

reasoned that we could quickly compare a generated architecture with the ground truth ΦX174 genome architecture by one-hot encoding start and stop codon positions, creating sparse vectors that represent the boundaries of genes. We created one-hot encoded 'truth' vectors of ΦX174: a vector *T* which is Gaussian blurred (σ = 5) to create fuzzy gene boundaries that tolerate slight positional shifts without over-penalizing near-matches, and a weight vector *W* that is multiplied by a weight factor equivalent to the total number of gene boundaries. A 'query' matrix *Q* of the one-hot encoded vectors of all circular permutations of a generated sequence is then scored with NumPy operations (*109*) `np.multiply(*W*, np.max(np.dot(*T*, *Q*)))/*N*`, where *N* is the genetic architecture score of ΦX174. Thus, a score of 1 indicates an exact match to the genetic architecture of ΦX174.

UMAP visualization of genetic architectures

One-hot encoded genetic architecture vectors with the highest genetic architecture similarity score to ΦX174 were computed for genomes in the *Microviridae* training data and stored in an AnnData object (*110*). Using Scanpy (*111*), the vectors were Z-score normalized and reduced to 50 principal components by PCA. A k-nearest neighbor graph (k = 50) was constructed on the PCA representation, UMAP was applied for non-linear dimensionality reduction, and the resulting UMAP embeddings were visualized.

ΦX174 genome prompt analysis

All ΦX174 variants extracted from the *Microviridae* training dataset were aligned with MAFFT (*112*) in Geneious Prime version 2025.1.2 (https://www.geneious.com/). The consensus start sequence was determined by analyzing the multiple sequence alignment by WebLogo (https://weblogo.berkeley.edu/logo.cgi) (*113*). Increasing amounts of nucleotides from the consensus start sequence were used to prompt the Evo 1 7B 131K, Evo 2 7B 1M, Evo 1 SFT, and Evo 2 SFT models with generation temperature = 1.1, top-k = 4, top-p = 1. Prompt lengths spanned 1 to 11 nucleotides, and all prompts were prepended with "+~" finetuning tokens except for when prompting the base models. Approximately 1,000 sequences of length 6,000 were generated with each prompt. Percent recall of ΦX174 per prompt was determined as the percent of generated sequences per prompt with a significant alignment against the reference genome ΦX174 NC_001422.1, using MMseqs2 version 13.45111 (*97*) with sensitivity = 7.5.

Model temperature and prompt sampling sweeps

To determine the optimal parameter combination for phage genome generation with the Evo 1 SFT and Evo 2 SFT models, sampling sweeps were conducted by systematically varying both model temperature and the nucleotide sequence prompt. Temperature sweeps were performed across five configurations: 0.3, 0.5, 0.7, 0.9, and 1.1. Prompt lengths spanned 1 to 11 nucleotides of the ΦX174 consensus start sequence. All prompts were prepended with "+~" fine-tuning tokens. Approximately 1,000 sequences were sampled per temperature–prompt parameter combination. All sampling runs were performed with top-k = 4 and top-p = 1. The parameter sweeps were evaluated by Shannon diversity (**Fig. 2J**), retention rate after filtering (**Fig. 2K**), and diversity of architectural similarity score vs. percent spike protein sequence identity (**fig. S6**). Additional sampling sweeps were performed across model checkpoints, fine-tuning tokens, top-k values, and top-p values, but were not used for final evaluation of generation conditions.

Shannon diversity analysis

Shannon entropy was used to quantify population-level sequence diversity (**Data, code, and materials availability**) in generated sequences, *Microviridae* genomes, scrambled *Microviridae* genomes, and the ΦX174 reference genome NC_001422.1. Generated sequences for each temperature and prompt length were filtered with quality control and tropism constraints before being analyzed. For each dataset, sequences were clustered using MMseqs2 version 13.45111 (*97*) at 99% nucleotide identity, then Shannon entropy was calculated on the resulting distribution of sequences per cluster.

Sequence filtering retention rate analysis

Generated sequences for each temperature and prompt length were filtered with quality control, tropism, and diversification constraints (**fig. S7A**), and the percentage of sequences passing each group of filters (retention rate) was calculated. The maximum retention rate across all temperature and prompt combinations was used to determine model performance. Design constraints were applied to *Microviridae*, scrambled *Microviridae*, and ΦX174 variant sequences as controls.

Functional assays of bacteriophage genomes

Biosafety and sterile technique

All experiments involving bacteriophage particles were conducted in a Class II Type A2 biosafety cabinet meeting NSF/ANSI 49, OEM specifications, and ISO 14644-1 standard. Pipettes and a 37 °C incubator were dedicated for use only with active bacteriophage cultures. Equipment was regularly sterilized with 70% ethanol, 10% bleach, and UV treatment. For all experiments, appropriate PPE was worn and sterile filter pipette tips were used. Any solid or liquid waste produced from experiments was discarded as biohazard waste.

Bacterial and bacteriophage growth media

Unless stated otherwise, bacteria were grown in liquid culture of "LB" consisting of LB Miller broth (Thermo Fisher Scientific #H26676). Bacteriophages were propagated with bacteria in liquid culture of "phage LB" media consisting of 10 g/L tryptone (RPI #T60060), 5 g/L yeast extract (IBI #IB49161), 10 g/L NaCl (Thermo Fisher Scientific #S271), and 2 mM $CaCl_2$ (J.T. Baker #1332). Bacterial colonies and bacteriophage plaques were grown in "phage LB agar" consisting of 7 g/L agar (Carolina #84-2133) in phage LB.

Bacterial strains

Lyophilized *E. coli* C derived from ATCC 13706 (Microbiologics #0747P) was rehydrated following the manufacturer's protocol in sterile conditions. The stock culture was inoculated in LB, incubated overnight at 37 °C with 170 rpm agitation, and mixed at a 1:1 volumetric ratio with 50% glycerol (Thermo Fisher Scientific #15514011) and stored at –80 °C. For preparation of *E. coli* C competent cells, a pipette tip was used to scrape glycerol stock and dipped in 5 mL of LB overnight at 37 °C with 170 rpm agitation, then 0.5 mL of the overnight culture was used with a Mix & Go *E. coli* Transformation Kit (Zymo Research #T3001) following the manufacturer's protocol. Competent cell stocks were stored at –80 °C and thawed only at the time of use. The

strains ATCC 25404 (*E. coli* K-12), Stbl3 (E. coli K-12/B HB101), JW5856 (*E. coli* K-12 W3110), Stellar (*E. coli* K-12 HST08), RosettaDE3 (*E. coli* B), Mach1 (*E. coli* W), and MDS69 (*E. coli* K-12 with reduced genome) (*114*) were a gift from Alex Gao's lab. Glycerol stocks and competent cell stocks for these strains were prepared the same way.

Bacteriophage genome assembly from ΦX174 RFI DNA

ΦX174 *am*3 *cs*70 RFI DNA (NEB #N3021) was PCR-amplified into separate sets of two or three fragments (**data S1**). The following pairs of primers were used to PCR-amplify fragments for two-fragment Gibson assembly of lysis mutant ΦX174 genomes with the *amber* (*am3*) mutation in the lysis gene: SK#327 + SK#324 and SK#323 + SK#328. The following pairs of primers were used to PCR-amplify fragments for two-fragment Gibson assembly of wild-type ΦX174 genomes by reverting the *am3* mutation to the wild-type codon: SK#333 + SK#326 and SK#334 + SK#325. The following pairs of primers were used to PCR-amplify fragments for three-fragment Gibson assembly of lysis mutant ΦX174 genomes with the *am3* mutation: SK#327 + SK#332 and SK#331 + SK#326 and SK#325 + SK#328. The following pairs of primers were used to PCR-amplify fragments for three-fragment Gibson assembly of wild-type ΦX174 genomes by reverting the *am3* mutation to the wild-type codon: SK#333 + SK#322 and SK#321 + SK#318 and SK#317 + SK#334. The fragments were amplified using 1 uL of 1 ng/uL ΦX174 *am*3 *cs*70 RFI DNA with 0.5 uL each of 10 uM primers, 10.5 uL of dH_2O, and 12.5 uL of 2× CloneAmp HiFi PCR Premix (Takara Bio #639298) with the following settings: 98 °C for 30 sec, 98 °C for 10 sec and 62 °C for 10 sec and 72 °C for 20 sec repeated 30 times, and 72 °C for 5 min. The PCR amplicons were separated in 1% (w/v) UltraPure Agarose (Thermo Fisher Scientific #16500500) in 1× TAE gels stained with 1× SYBR Safe (Thermo Fisher Scientific #S33102) run at 120 V for 30 min, with GeneRuler 1 kb Plus DNA Ladder (Thermo Fisher Scientific #SM1331). PCR products of the correct size were purified by gel extraction using a QIAquick Gel Extraction Kit (Qiagen #28704) following the manufacturer's protocol. Assembly fragments were mixed with NEBuilder HiFi DNA Assembly Master Mix (NEB #E2621) with the volumes of each fragment calculated on https://nebuildercalculator.neb.com/, and incubated at 50 °C for 1 hour. Assemblies were stored at –20 °C until use.

Bacteriophage genome assembly from synthesized DNA

Sequences were split into two Gibson assembly-compatible fragments and synthesized as Gene Fragments by Twist Biosciences, without adapters, dried down and normalized to 250 ng in 96-well plate wells. Optimal split points and overhangs for each Gibson assembly were designed with a custom Python script (**Data, code, and materials availability**) that outputs each fragment to be synthesized. The synthesized fragments were resuspended in dH_2O to 10, 25, or 50 ng/uL and assembled with the following conditions: each of two fragments per genome were mixed at a 1:1 (v/v) ratio to a total volume of 2 uL. 2 uL of NEBuilder HiFi DNA Assembly Master Mix (NEB #E2621) was added to the mixture on ice to a total reaction volume of 4 uL and the reaction was incubated at 50 °C for 1 hour. Assemblies were stored at –20 °C until use. The gene fragments were sealed in their original 96-well plates with foil seals (Bio-Rad #MSF1001) and stored at –20 °C.

Bacteriophage genome assembly for mutant viability assays

To test the viability of Evo-Φ5 and Evo-Φ3392 with their acquired mutations reverted back to their original generated sequences, the original generated sequences were synthesized again but by Integrated DNA Technologies through their gBlocks Gene Fragments service and resuspended in dH_2O to 25 ng/uL. The Gibson assembly was performed as described above. To test the viability of Evo-Φ36 gene J in the genomic context of ΦX174 and NC51, the coding sequences of gene J in the genomes of ΦX174 and NC51 were directly replaced with the coding sequence of Evo-Φ36 gene J. The Gibson assembly-compatible fragments were designed, synthesized, and assembled as described above. To test the viability of Evo-Φ63 without the second downstream gene J, the coding sequence of the downstream gene J at nucleotide positions 1116-1232 was deleted from the Gibson assembly-compatible fragment designed as described above. The fragment containing the deletion was synthesized by Integrated DNA Technologies through their gBlocks Gene Fragments service and resuspended in dH_2O to 25 ng/uL. The Gibson assembly was performed as described above. To test the viability of the noncoding deletion mutants of Evo-Φ2147, Evo-Φ2483, and Evo-Φ2498, the noncoding regions were deleted at nucleotide positions 978-1246, 2418-2611, and 2418-2572, respectively. The Gibson assembly-compatible fragments were designed, synthesized, and assembled as described above.

Bacteriophage genome assembly plaque assay

Phage genomes were assembled by Gibson assembly and 4 uL of the assembly was mixed with 100 uL of *E. coli* C competent cells and left on ice for 10 min. The transformation was gently mixed with 0.5 mL of phage LB and added to 7 mL of phage LB agar at a temperature of 42–46 °C monitored using an Infrared Thermometer (Ketokek #KT600B), then plated immediately on 10 cm Petri dishes. The plates were incubated at 37 °C for 3 hours, wrapped with a thin strip of Parafilm (Millipore Sigma #HS234526B) to retain moisture, and stored at 4 °C.

Bacteriophage genome assembly growth assay

Phage genomes were assembled by Gibson assembly and 1 uL of the assembly was mixed with 15 uL of *E. coli* C or *E. coli* K-12 competent cells and left on ice for 10 min. The transformation was gently mixed with 735 uL of phage LB and split into three wells of a flat-bottom 96-well plate (Thermo Fisher Scientific #167008) with 250 uL of per well. Three wells with 250 uL phage LB only were added to the plate to normalize the OD_{600} measurements. The 96-well plate was incubated at 37 °C with orbital shaking at 1.5 mm amplitude and 360 rpm in a Tecan Spark Multimode Microplate Reader, with OD_{600} measured every 15 min. After 6 hours, the plate was removed from the microplate reader and stored at 4 °C.

Bacteriophage glycerol stocks

Bacterial debris in phage-clarified cultures were pelleted by centrifugation at 3,000 × *g* for 10 min at 4 °C. The supernatant was sterile-filtered through a 0.22 um cellulose acetate Spin-X Centrifuge Tube Filter (Thermo Fisher Scientific #07-200-385) by centrifugation at 10,000 × *g* for 3 min at 4 °C. Glycerol stocks were prepared by mixing the flow-through with 0.22 um-filtered 50% glycerol (Thermo Fisher Scientific #15514011) at a 1:1 (v/v) ratio and stored at –80 °C in cryogenic tubes (Thermo Fisher Scientific #11-676-48).

Long-read sequencing of bacteriophage genomes from glycerol stocks

Glycerol stocks prepared from the phage growth assays were individually scraped using sterile pipette tips and dipped into 6 uL of 1× phosphate buffered saline (PBS). The picked phage genomes were amplified using 1 uL of the solution in a PCR reaction with 0.5 uL each of 10 uM primers (**data S1**), 10.5 uL of dH_2O, and 12.5 uL of 2× CloneAmp HiFi PCR Premix (Takara Bio #639298) with the following settings: 98 °C for 30 sec, 98 °C for 10 sec and 62 °C for 10 sec and 72 °C for 20 sec repeated 30 times, and 72 °C for 5 min. The PCR amplicons were separated in 1% (w/v) UltraPure Agarose (Thermo Fisher Scientific #16500500) in 1× TAE gels stained with 1× SYBR Safe (Thermo Fisher Scientific #S33102) run at 120 V for 30 min, with GeneRuler 1 kb Plus DNA Ladder (Thermo Fisher Scientific #SM1331). PCR products of the correct size were purified by gel extraction using a QIAquick Gel Extraction Kit (Qiagen #28704) following the manufacturer's protocol and sequenced using Plasmidsaurus' Standard Purified Linear/PCR sequencing service. Sequences were aligned by MAFFT v7 (*112*) on Benchling (https://www.benchling.com/).

Bacteriophage propagation and harvesting

A glycerol scrape of *E. coli* C was inoculated in 5 mL of LB and incubated at 37 °C overnight with 200 rpm agitation. 3 mL of the overnight culture was mixed with 247 mL of phage LB and grown to an OD_{600} of ~0.4. A scrape of phage glycerol stock was added to the culture and clarified for up to 9 hours. The lysed debris was pelleted by centrifugation at 4,000 × *g* for 10 min at 4 °C and sterile-filtered through a 0.22 um pore size PES membrane filter (Millipore Sigma #S2GPU05RE). The phages Evo-Φ46, Evo-Φ111, Evo-Φ114, ID22, NC5, NC16, and NC37 were double propagated from their first propagation due to low titer. The phages Evo-Φ36, Evo-Φ108, NC1, and NC7 were triple propagated due to low titer.

Bacteriophage titering

A glycerol scrape of *E. coli* C was inoculated in 5 mL of LB and incubated at 37 °C overnight with 200 rpm agitation. 50 uL of the overnight culture was mixed with 5 mL of phage LB and incubated at 37 °C with 200 rpm agitation. The culture was grown to an OD_{600} of 0.3–0.4 and 675 uL of the culture was added to 22.5 mL of phage LB agar at a temperature of 42–46 °C. The temperature was monitored using an Infrared Thermometer (Ketokek #KT600B). The mixture was poured in a 15 cm Petri dish and allowed to solidify for up to 5 min. 2 uL of 10-fold serial dilutions of phage in phage LB from 10^0 to 10^{-10} was spotted on the agar and fully dried for up to 15 min. An additional 2 uL spot of phage LB only was also spotted as a negative control. The plates were inverted and incubated at 37 °C for 3 hours, imaged, wrapped with a thin strip of Parafilm (Millipore Sigma #HS234526B) to retain moisture, and stored at 4 °C. Images were taken on an iPhone with a custom backlight setup and converted to black and white by decreasing the saturation to 0. Individual visible plaques were counted at the highest dilution where they were present across all three replicates, and the titer (plaque forming units (PFU) / mL) per phage was calculated as ((average plaque count / spot volume (mL)) × dilution factor).

Host tropism assay

For each *E. coli* strain, a glycerol scrape was inoculated in 5 mL of LB and incubated at 37 °C overnight with 200 rpm agitation. 250 uL of the overnight culture was mixed with 25 mL of phage LB and incubated at 37 °C with 200 rpm agitation. The culture was grown to an OD_{600} of 0.4–0.6 and 200 uL per well was plated in a flat-bottom 96-well plate (Thermo Fisher Scientific #167008).

50 uL of phage stock at a concentration of ~10^5 PFU/mL was then added to each well. As negative controls, 50 uL of phage LB was added to each well instead of phage. Three wells with 250 uL phage LB only were added to the plate to normalize the OD_{600} measurements. The 96-well plate was incubated at 37 °C with orbital shaking at 1.5 mm amplitude and 360 rpm in a Tecan Spark Multimode Microplate Reader, with OD_{600} measured every 15 min. After ~12 hours, the plate was removed from the microplate reader and stored at 4 °C.

Phylogenetic analysis of bacteriophages

Natural genome datasets

Unless otherwise noted, the wild-type ΦX174 reference genome used for all phylogenetic analyses was sourced from NCBI accession NC_001422.1, and the wild-type G4 reference genome used for all phylogenetic analyses was sourced from NCBI accession NC_001420.2.

To interpret generated sequence novelty in the context of ΦX174-like phages, we collected three datasets of natural phages to compare generated genomes against (**fig. S15H**): 134 ΦX174 variants from our unprocessed *Microviridae* training data (see *Microviridae data*), 39 ΦX174 variants evolved over 180 days of laboratory evolution, equivalent to ~13,000 phage generations (*60*), and 16 ΦX174-like phages isolated by screening 734 phage isolates collected from environmental samples from three locations across the United States (North Carolina, Idaho, Washington) (*61*). For the lab-evolved ΦX174 variants, the ancestral genome ΦX174 Anc AF176034.1 was included in the dataset. The aligned lab-evolved ΦX174 variant sequencing data in plain text format was downloaded (*60*) and reformatted into a fasta file, ambiguous bases were replaced with 'N', and gap characters consisting of '.' were removed. The ΦX174-like isolates (*61*) were each sourced from their NCBI accessions: NC51 (DQ079891.1), WA10 (DQ079894.1), ID45 (DQ079883.1), NC5 (DQ079885.1), NC41 (DQ079890.1), NC16 (DQ079888.1), NC37 (DQ079889.1), NC56 (DQ079892.1), NC11 (DQ079887.1), NC1 (DQ079884.1), NC7 (DQ079886.1), ID22 (DQ079881.1), ID34 (DQ079882.1), ID1 (DQ079880.1), WA4 (DQ079893.1), and WA11 (DQ079895.1).

Synteny analysis

Synteny was visualized by LoVis4u with flags `-hl`, `--set-category-colour`, `-c A4p2`, `-alip`, (*96*) using GFF3 files for each phage genome created with a custom script (**Data, code, and materials availability**). Pairwise synteny between each phage genome was determined in the order presented (**Fig. 4A**) and consolidated into a single synteny plot. Since our gene annotation method only partially predicted gene A*, it was omitted from synteny visualization. Synonymous substitutions, nonsynonymous substitutions, indels, and intergenic substitutions were determined relative to each genome's nearest natural genome using a custom script (**Data, code, and materials availability**; see *Mutational comparison to nearest natural genomes*) Genes sharing synteny with ΦX174 were determined using a custom script (**Data, code, and materials availability**) analyzing the pairwise protein identity matrix calculated by LoVis4u. For simply visualizing SNVs relative to ΦX174, ΦX174 was set as the reference sequence and the default mutation highlighting in Geneious Prime was exported and overlaid on the synteny visualization.

Whole-genome alignment

Unless otherwise noted, whole-genome alignments were performed with MAFFT (*112*) in Geneious Prime version 2025.1.2 (https://www.geneious.com/). To determine percent genome identity of generated phage genome candidates compared to sequences in the *Microviridae* training data, genomes were aligned by nucleotide BLAST (blastn, E-value = 0.5, default settings) (*43*) in Geneious Prime, and percent genome identity was calculated as (percent identity × percent query coverage). Percent genome identity of sequences in the *Microviridae* training data compared to ΦX174 were determined by MMseqs2 version 13.45111 (*97*) with default settings.

Mutational comparison to nearest natural genomes

To characterize the amount and types of mutations distinguishing generated and natural phages from their nearest natural genomes, we performed a systematic pairwise mutation classification analysis (**Data, code, and materials availability**). Each genome was queried against the NCBI nucleotide (core_nt) database using BLASTn (megablast, E-value = 0.5, default parameters) via the NCBI remote API (Bio.Blast.NCBIWWW.qblast; Biopython v1.82, last updated March 6, 2026). For each query, the top 10 hits were retained (excluding self-hits where the query accession matched a subject accession). Multiple hits per query were used rather than a single top hit to reduce sensitivity to minor ranking differences among highly similar neighbors. Percent sequence similarity of each query sequence to its nearest natural sequence was calculated as (percent identity × percent query coverage). GenBank records for all unique hit accessions were retrieved via NCBI Entrez and were required to contain at least 10 annotated CDSs and at least 8 of the 9 core ΦX174 genes (A, B, C, D, E, F, G, H, J) to be included in downstream analysis. For each query-hit pair, a pairwise nucleotide alignment was generated using MAFFT v7.505 (--auto mode) (*112*). Mutations were classified by iterating over each alignment column. Substitutions (columns where both sequences contained a non-gap base that differed) were mapped to CDS annotations in the hit GenBank record. For each CDS overlapping the substitution position, the reading frame was determined directly from the annotated CDS boundaries in the GenBank record. Each CDS defines an ordered sequence of genome positions; the substitution's index within this sequence determines which codon it belongs to (codon number = floor(index/3)), and the three genome positions comprising that codon were identified accordingly. For genes spanning the genome origin via join() locations, the position sequences from all segments were concatenated in annotated order, preserving correct codon boundaries across the origin. The three codon positions were then located in the pairwise alignment, and the corresponding bases from both the query and hit sequences were extracted, translated using the standard genetic code, and compared: if the encoded amino acids were identical, the substitution was classified as synonymous; otherwise, as nonsynonymous. Positions falling in overlapping reading frames were classified independently in each frame, so a single mutation in an overlap region can contribute one mutation count per overlapping gene. Substitutions at positions outside any annotated CDS were classified as intergenic. Indels were identified as contiguous gap runs in either sequence and counted as single events regardless of length. Substitutions were counted per-position: adjacent substitutions (e.g., two consecutive mismatches) were treated as independent events, each classified by its own codon context. The total number of SNVs in each genome was estimated as ((1 − (percent sequence similarity / 100)) × nearest natural genome length). The total number of mutation events in each genome was determined as the sum of synonymous substitution counts, nonsynonymous substitution counts, indel events, and intergenic substitution counts.

Phylogenetic tree construction

A multiple sequence alignment of representative *Microviridae* and generated phage genomes was constructed with MAFFT (*112*) in Geneious Prime version 2025.1.2 (https://www.geneious.com/). The resulting MSA was used to build a Neighbor-Joining tree with the Geneious Tree Builder tool with a Jukes-Cantor genetic distance model and no outgroup. The following natural reference genomes collected from NCBI were used for phylogenetic trees: NC_001422.1 (ΦX174), KY653237.1 (alpha-α), DQ079890.1 (NC41), DQ079885.1 (NC5), DQ079891.1 (NC51), and AF274751.1 (S13), NC_001420.2 (G4), NC_012868.1 (St-1), NC_001730.1 (ΦK), NC_007821.1 (WA13), NC_001330.1 (α3).

ΦX174-like clade MSA construction

Sequences from the raw *Microviridae* dataset were searched against the ΦX174 reference genome using BLASTn (E-value ≤ 1×10^{-5}, max_target_seqs = 1), and 532 sequences with ≥ 80% nucleotide identity were retained to minimize gaps in the MSA. These were aligned using MAFFT v7.505 L-INS-i (--localpair --maxiterate 1000) (*112*). Because *Microviridae* genomes are circular and have been linearized at variable origins, each sequence was rotated to match the ΦX174 linearization origin: a concatenated duplicate of each genome was searched against the reference using BLASTn to identify the position corresponding to the reference start, and the sequence was circularly permuted to that position. The rotated sequences were then re-aligned with MAFFT --auto.

Site-independent MSA sampling analysis

To assess whether viable AI-generated phage genomes could plausibly arise from simple position-wise sampling of known sequence diversity, a site-independent null model was constructed from an MSA of 532 ΦX174-like *Microviridae* genomes (see *ΦX174-like clade MSA construction*). For each of the 5,386 ΦX174 reference-aligned positions, nucleotide frequencies (A, C, G, T) were computed across all aligned sequences, a pseudocount of 10^{-6} was added to avoid zero probabilities, and each position was normalized to sum to one. From this model, 10,000 null genomes were sampled by independently drawing a nucleotide at each position according to its estimated probability distribution. The total $\log_2$-likelihood of each sampled genome was computed as the sum of $\log_2$-transformed positional probabilities for the drawn nucleotides. To score the generated genomes under the same model, each sequence was aligned to the ΦX174 reference by global pairwise alignment using Biopython's PairwiseAligner (match score = 2, mismatch = −1, gap open = −3, gap extend = −0.5). To account for any minor differences in the number of aligned positions, the $\log_2$-likelihood sum was normalized to a per-position average and then multiplied by the full reference length, making scores directly comparable to the null samples.

Positional mutation frequency and enrichment analysis

To characterize the nucleotide-level positional distribution of mutations across the generated and natural phage genomes, we compared viable generated genomes, nonviable generated genomes, and natural ΦX174-clade sequences obtained from an MSA (see *ΦX174-like clade MSA construction*). For each genome in a given dataset, we identified its nearest neighbor by BLASTn search (E-value ≤ 1e-10) against a database of rotated MSA sequences, selecting the top hit by sequence similarity, with self-hits excluded to prevent a natural sequence from matching itself. Each query genome was then globally pairwise-aligned to its nearest neighbor using Biopython's pairwise2.align.globalms (match = 2, mismatch = −1, gap open = −3, gap extend = −0.5), and

mutations were classified as substitutions (single-base mismatches), deletions, or insertion events (contiguous inserted blocks). Genomes whose length differed from their nearest neighbor by more than 10% were excluded. To put mutations on a common coordinate system, mutation positions were mapped to the ΦX174 reference coordinate system (5,386 positions). Because each genome was aligned to its nearest neighbor rather than directly to the reference, mutation positions were projected onto the reference coordinate system using the MSA as an intermediary. Each neighbor sequence appears in the MSA alongside the reference, so any position in the neighbor can be mapped to the reference via their shared MSA column. Substitutions and deletions were mapped directly through the neighbor position at which they occurred. Insertions, which correspond to bases in the query that have no counterpart in the neighbor, were assigned to the midpoint between the flanking neighbor positions that could be mapped. Per-position mutation counts were normalized by the number of genomes in each dataset to yield mutation events per genome.

For enrichment analysis, the per-genome density profiles were further normalized to sum to 1.0 across the genome, yielding relative mutation frequency distributions. Log_2 fold-change enrichment was then computed for each pairwise comparison (viable/nonviable, viable/natural, natural/nonviable) with a pseudocount of 1×10^{-8} and clipped to the range [−2, 2]. All profiles were smoothed using a 25-bp uniform-kernel rolling average, with correction for edge effects at positions with missing values.

To assess how mutations in viable generated genomes were distributed across genetic elements, we computed the mean mutation frequency for each annotated ΦX174 feature. For each feature, we counted the number of genomes carrying a mutation (substitution, deletion, or insertion) at each reference position within that feature's boundaries, then averaged across positions to obtain the mean mutation frequency.

Null survival model

To assess whether generated phage genomes exhibit viability rates above what would be expected from random mutagenesis, we compared the observed fraction of viable designs to an analytic null expectation derived from empirical nonsynonymous mutation fitness data (*16*). Genomes were binned by their nonsynonymous mutation counts in intervals of 25 mutations, with the upper bound not included in each interval except for the last interval. The null expectation models the probability that a genome with k nonsynonymous mutations relative to a natural genome remains viable, assuming each mutation independently confers lethality with probability 0.2, based on the finding of Domingo-Calap et al. (*16*) that approximately 20% of random single nonsynonymous substitutions in ΦX174 are lethal. Under this framework, the survival probability for a genome is $P(survive) = (1 - 0.2)^k$. We did not consider the effects of synonymous substitutions, indels, and intergenic substitutions in this model, and genomes that failed DNA synthesis were excluded. For each nonsynonymous mutation bin, the survival probabilities of all genomes in that bin were averaged, giving the null expectation for that bin.

Cumulative genome attribution analysis

To determine if mutations in the generated genomes could be attributed to existing mutations in natural genomes, the top 1,000 alignments (E value < 1.0) for each generated genome were determined using nucleotide BLAST (blastn, E-value = 0.05, default settings) (*43*). Base pair-level attributions followed a greedy approach, where exact nucleotide matches were first assigned to the highest-scoring BLAST hit, then remaining unassigned positions were iteratively assigned to

lower-scoring hits in descending order until complete assignment or no additional matches were possible.

Structural analysis of bacteriophages

Protein structure prediction

F (capsid), G (spike) and J (DNA packaging) proteins were co-folded by AlphaFold 3 (*65*) via the online web server (https://alphafoldserver.com/). PAE was visualized with https://thecodingbiologist.com/tools/pae.html and predicted structures were visualized with UCSF ChimeraX version 1.7.1 (*115*).

Bacteriophage purification for cryo-EM

E. coli C was grown in 300 mL of phage LB at 37 °C with 200 rpm agitation until reaching OD_{600} 0.4–0.6, at which they were inoculated with 3 mL of phage for 3.5–8 hours. Chloroform was added to cultures to a final dilution of 1% and incubated for 15 min at 25 °C with 200 rpm agitation for complete lysis. Lysed cultures were pelleted by centrifugation (3,000 × *g*, 10 min, 4 °C), and sterile-filtered through a 0.22 um pore size PES membrane filter (Millipore Sigma #S2GPU05RE). Filtered lysate was supplemented with polyethylene glycol 8,000, pH 7.4 (8%), 100 mM NaCl and incubated with end-over-end rotation at 4 °C for 1 hour, before incubation at 4 °C for 2–16 hours without rotation. The phage pellet was obtained by centrifugation (11,000 × *g*, 20 min, 4 °C) and resuspended in 4 mL PBS pH 7.4, 10 mM $MgCl_2$, to which 50 U/mL Benzonase Nuclease (Teknova #E1014) was subsequently added. The mixture was incubated for 45 min at 37 °C, with gentle mixing every 15 min. The phage suspension was carefully placed on top of a manually constructed discontinuous iodixanol (Teknova #21449, #21443, #21431, #21425) density gradient (60%, 40%, 25%, 15%), and subjected to ultracentrifugation (200,000 × *g*, 2–4 hours, 4 °C). Visible phage bands (at the 60% / 40% fraction boundary) were carefully extracted and buffer exchanged by four rounds of centrifugation (15,000 × *g*, 10 min, 4 °C) in 100 kDa molecular weight cut-off Amicon centrifugal filters (Millipore #UFC510024) followed by resuspension in 20 mM Tris pH 7.4, 100 mM NaCl.

Polyacrylamide gel electrophoresis (PAGE)

SDS-PAGE was performed using 4–20% SurePAGE precast gels (GenScript #M00657) in XCell SureLock Mini-cell gel chambers (Thermo Fisher Scientific #EI0001). Before loading onto gels, samples were mixed with 4× Laemmli buffer (62.5 mM Tris HCl, pH 6.8, containing 2% (w/v) SDS, 10% (v/v) glycerol, and 0.002% (w/v) bromophenol blue) and boiled at 99 °C for 3 min. Gels were run at 200 V in 1× MES SDS Running Buffer (Genscript #M00677) alongside SeeBlue Plus2 Pre-stained Protein Standard ladder (Invitrogen #LC5925) or PageRuler Plus Prestained Protein ladder (Thermo Fisher Scientific #26619). Gels were stained with InstantBlue Coomassie (AbCam #ab119211) for up to 1 hour with gentle shaking and de-stained for at least 1 hour with Milli-Q water prior to imaging on a GelDoc Go imager (Bio-Rad). ImageLab version 6.1.0 (Bio-Rad) was used for analysis.

Grid preparation for cryo-EM

Phage samples were purified as described above and stored at 4 °C prior to cryo-EM grid preparation. 3.2 uL of the purified phage was applied to R1.2/1.3, 300 mesh carbon Cu grids (QUANTIFOIL #Q3100CR1.3) which were glow-discharged using a PELCO easiGlow system with 10 mA negative current for a total period of 90 s at 0.26 mBar with a 10 s hold period. The grids were plunge-frozen in liquid ethane using a Vitrobot Mark IV (Thermo Fisher Scientific) maintained at 100% humidity and 8 °C with a blot time of 3 s and a blot force of 3.5. The grids were stored in liquid nitrogen.

Cryo-EM data acquisition

Grids of ΦX174 and Evo-Φ36 were initially screened on a Glacios electron microscope (Thermo Fisher Scientific). One high-quality grid for each phage was selected for data collection and imaged on a Glacios electron microscope (Thermo Fisher Scientific) operated at 200 kV and equipped with a Falcon 4i direct electron detector. Movies were automatically collected with EPU software (Thermo Fisher Scientific) at 150,000× magnification, corresponding to a real pixel size of 0.923 Å at the specimen level. For ΦX174, 3,372 movies were collected with a defocus range of -1.5 to -3.0 microns. Each movie was recorded over 3.52 s with a total accumulated dose of 41.15 $e^{-}/Å^{2}$, fractionated into 40 frames (~1.03 $e^{-}/Å^{2}$ per frame). For Evo-Φ36, 4,796 movies were collected with a defocus range of -1.5 to -2.5 microns. Each movie was recorded over 3.79 s with a total accumulated dose of 50.00 $e^{-}/Å^{2}$, fractionated into 43 frames (~1.16 $e^{-}/Å^{2}$ per frame).

Cryo-EM data processing

Cryo-EM data were processed using cryoSPARC version 4.6.2 (*116*). Movies were motion-corrected and CTF parameters estimated using the patch motion correction and patch-based CTF estimation jobs. Initial particles were picked on denoised micrographs by blob picking using a circular blob with particle diameter between 250 Å and 325 Å. Picked particles were extracted with a box size of 512 pixels and classified in 2D to select representative 2D classes for use as templates. Resulting particles from template-guided particle picking were extracted with a box size of 512 pixels and subjected to 2–3 rounds of 2D classification, yielding 24,999 particles for ΦX174 and 30,228 particles for Evo-Φ36. Multiple rounds of three-dimensional (3D) reconstruction and refinement (uniform and non-uniform with icosahedral symmetry imposed, as well as global and local CTF refinement) were performed without further particle sorting, with one round of reference-based motion correction, with 24,547 particles and 29,708 particles used in the final refinements for ΦX174 and Evo-Φ36 respectively. This resulted in final maps for ΦX174 and Evo-Φ36 with resolutions of 2.76 Å and 2.90 Å, respectively, as determined by their gold standard Fourier shell correlation with a threshold of 0.143 (*117*). Further processing details are provided in **Figure S23**. Final maps for model building were sharpened using a postprocessing job in RELION 5.0 with masks constructed in RELION using a low-pass filter of 20 Å, 1 pixel extension, and 8 pixel soft-edge with a binarization threshold of 0.015 and 0.01 for ΦX174 and Evo-Φ36, respectively (*118*).

Cryo-EM model building

AlphaFold 3 models of the F, G, and J proteins co-folded for either ΦX174 or Evo-Φ36 were docked into their respective sharpened cryo-EM densities in UCSF ChimeraX before manual adjustment in Coot version 0.9.8.96 (*119*). Interactive molecular dynamics-based refinement was performed in ISOLDE version 1.9 (*120*) to correct local geometry and improve fit to density. Final

refinement was carried out in Phenix version 1.21.2 (*121*) against the cryo-EM density maps, using parameters generated with the 'isolde write phenixRsrInput' command with restrainPositions set to true. Structures were visualized with UCSF ChimeraX version 1.7.1 or 1.9.1 (*115*).

Fitness assays of bacteriophage genomes

Bacteriophage co-infection assay

A glycerol scrape of *E. coli* C was inoculated in 100 mL of LB and incubated at 37 °C overnight with 200 rpm agitation. 2 uL of the overnight culture was mixed with 198 uL of phage LB in a flat-bottom 96-well plate (Thermo Fisher Scientific #167008) wells and incubated at 37 °C with orbital shaking at 1.5 mm amplitude and 360 rpm in a Tecan Spark Multimode Microplate Reader. Three wells with 250 uL phage LB only were added to the plate to normalize the OD_{600} measurements. Once the cultures reached OD_{600} ~0.4, 50 uL of a phage mixture consisting of ~1×10^5 PFU/mL per phage of all 16 generated phages and ΦX174 was added to each well. The plate was again incubated in the microplate reader with the same conditions, with OD_{600} measured every 10 min. Upon the initial infection, 10 uL was extracted from each culture, then 10 uL was extracted every 30 min for 3 hours, then every 60 min for 3 hours, for a total of 6 hours. Each extraction was immediately boiled in a thermal cycler at 98 °C for 45 s and stored in –20 °C. 2 uL of the lysate at each time point was used in a PCR reaction with 1 uL each of 10 uM primers SK#324 and SK#359, 21 uL of dH_2O, and 25 uL of 2× CloneAmp HiFi PCR Premix (Takara Bio #639298) with the following settings: 98 °C for 30 sec, 98 °C for 10 sec and 62 °C for 10 sec and 72 °C for 40 sec repeated 25 times, and 72 °C for 5 min. The PCR amplicons were separated in 1.5% (w/v) UltraPure Agarose (Thermo Fisher Scientific #16500500) in 1× TAE gels stained with 1× SYBR Safe (Thermo Fisher Scientific #S33102) run at 120 V for 30 min, with GeneRuler 1 kb Plus DNA Ladder (Thermo Fisher Scientific #SM1331). PCR products of the correct size were purified by gel extraction using a QIAquick Gel Extraction Kit (Qiagen #28704) following the manufacturer's protocol and sequenced using Plasmidsaurus' Standard PCR Premium sequencing service.

Bacteriophage co-infection sequencing analysis

The sequencing reads from each competition time point were analyzed using a custom script (**Data, code, and materials availability**). Raw nanopore sequencing reads were quality-filtered using fastp version 0.24.0 (minimum length 250 bp, minimum quality score 20) and aligned to reference amplicon sequences using minimap2 version 2.28 (map-ont preset, --secondary=no) to retain only primary alignments. Rather than relying on the aligner's best-hit assignment, reads were assigned to individual phage genotypes using a SNV-based scoring approach. For each read, the base at every variable position across all reference sequences was compared to the expected base for each phage, and the read was assigned to the phage with the highest number of matching positions. Reads were classified as ambiguous and discarded if the best and second-best scores were tied. An MSA was constructed via pairwise alignment of each reference to an anchor sequence (ΦX174, NC_001422.1) using Biopython's PairwiseAligner, variable columns were identified in MSA coordinates, with a minimum percent identity of 90% and alignment fraction of 90%. To detect chimeric reads that were likely artifacts of PCR template switching, variable positions in amplicons were divided into five windows each containing an approximately equal number of variable positions, and the best-matching phage was determined independently within

each window. A read was flagged as chimeric if any window assigned the read to a different phage than the overall best match, provided that window contained at least 5 variable positions and the alternative phage's score exceeded the overall best match's score by at least three positions. Any reads flagged as chimeric this way were discarded from the final analysis. Relative abundance was quantified as the cumulative $\log_2$ fold change in read proportion over time, calculated from the earliest time point at which each phage was detected. Overall fitness was summarized as the signed area under the curve (AUC) of the cumulative $\log_2$ fold change trajectory, computed via the trapezoidal rule with time points converted to hours.

Bacterial growth rate assay

A glycerol scrape of *E. coli* C was inoculated in 5 mL of LB and incubated at 37 °C overnight with 200 rpm agitation. 250 uL of the overnight culture was mixed with 25 mL of phage LB and incubated at 37 °C with 200 rpm agitation. The culture was grown to an OD_{600} of ~0.4 and 200 uL per well was plated in a flat-bottom 96-well plate (Thermo Fisher Scientific #167008). 50 uL of each phage stock at a concentration of ~10^5 PFU/mL was then added to each well. As negative controls, 50 uL of phage LB was added to each well instead of phage. Three wells with 250 uL phage LB only were added to the plate to normalize the OD_{600} measurements. The 96-well plate was incubated at 37 °C with orbital shaking at 1.5 mm amplitude and 360 rpm in a Tecan Spark Multimode Microplate Reader, with OD_{600} measured every 15 min. After ~12 hours, the plate was removed from the microplate reader and stored at 4 °C.

Bacterial growth rate determination

Bacterial growth rates were derived from OD_{600} measurements collected at fixed time intervals. For each replicate, growth rate was computed as the numerical derivative of OD_{600} with respect to time using NumPy's operation `np.gradient` in Python (*109*), yielding instantaneous rates expressed as ΔOD_{600} per minute. Replicate trajectories were then aggregated to calculate summary statistics (mean, standard deviation, and minimum growth rates), which were used for downstream comparative analyses across phage treatments.

Statistical analysis

Statistical analyses were performed in Python using the statsmodels package and SciPy library (*122*, *123*). For each assay metric (signed area under the curve, minimum OD_{600} after peak, time to post-peak minimum, and minimum growth rate), replicate-level values were first cleaned by removing missing values. A type II one-way analysis of variance (ANOVA) was used to test for overall differences across phage groups, with statistical significance set at $\alpha = 0.05$. Where the ANOVA indicated significance, post-hoc pairwise comparisons between phage groups were performed using Tukey's honestly significant difference (HSD) test, which controls the family-wise error rate for multiple comparisons. Replicate distributions were visualized alongside group means and standard deviations. Spearman rank correlation was used to assess the relationship between mean signed AUC and mutational distance from the nearest natural genome.

Bacteriophage resistance assay

ΦX174 was assembled by Gibson assembly and 1 uL of the assembly was mixed with 15 uL of *E. coli* C competent cells and left on ice for 10 min. The transformation was gently mixed with 735

uL of phage LB and split into three wells of a flat-bottom 96-well plate (Thermo Fisher Scientific #167008) with 250 uL per well. 15 uL of non-transformed competent cells were mixed with 735 uL of phage LB and plated in another three wells. Three wells with 250 uL phage LB only were added to the plate to normalize the OD_{600} measurements. The 96-well plate was incubated at 37 °C with orbital shaking at 1.5 mm amplitude and 360 rpm in a Tecan Spark Multimode Microplate Reader, with OD_{600} measured every 15 min. After 24 hours, the transformed cultures reached stable resistance against ΦX174, and the non-transformed, ΦX174-susceptible cultures also reached equilibrium growth. 250 uL of each culture was mixed with 50% glycerol (Thermo Fisher Scientific #15514011) at a 1:1 (v/v) ratio and stored at –80 °C. To isolate single colonies from each strain, the glycerol stocks were individually streaked out on phage LB agar plates and incubated at 37 °C overnight. Colonies were picked, dipped in 250 uL of phage LB in a 96-well plate, and incubated at 37 °C overnight with 200 rpm agitation. 50 uL of each culture was mixed with 50% glycerol at a 1:1 (v/v) ratio and stored at –80 °C for final stocks of the resistant and susceptible *E. coli* C strains. Each strain's genome was sequenced using Plasmidsaurus' Standard Bacteria Genome with Extraction service.

To test the ability of ΦX174, natural ΦX174-like phages, and generated phages to inhibit growth of the resistant and susceptible strains, we devised a bacteriophage counter-resistance evolution assay similar to a previously described protocol (*69*). The following mixtures of phages were prepared: ΦX174 diluted to a concentration of ~17×10^5 PFU/mL in phage LB ("ΦX174-only cocktail"), all 16 ΦX174-like phages (see *Natural genome datasets*) and ΦX174 diluted together in phage LB each at a concentration of ~1×10^5 PFU/mL ("natural phage cocktail"), and all 16 generated phages and ΦX174 diluted together in phage LB each at a concentration of ~1×10^5 PFU/mL ("generated phage cocktail"). Note that due to failed rebooting, the ΦX174-like phages ID1 and NC56 were excluded, and ΦX174-like phages S13 (AF274751.1) and MED1 (KJ997912.1) were rebooted and included instead, given their frequent appearance amongst the nearest natural genomes (**fig. S26**; **data S1**). A scrape of each glycerol stock of the resistant and susceptible strains was inoculated in 250 uL phage LB and incubated at 37 °C with 200 rpm agitation overnight. For each resistant strain tested, three different conditions were prepared in a 96-well "seed" plate: a "mixed well" consisting of 1 uL of overnight resistant culture with 1 uL of overnight susceptible culture in 198 uL of phage LB (to facilitate evolution of the phages in susceptible cells in the presence of resistant cells), a resistant-only well consisting of 2 uL of overnight resistant culture in 198 uL of phage LB, and a susceptible-only well consisting of 2 uL of overnight susceptible culture in 198 uL of phage LB. The plate was incubated at 37 °C with orbital shaking at 1.5 mm amplitude and 360 rpm in a Tecan Spark Multimode Microplate Reader, with OD_{600} measured every 15 min until the cultures reached OD_{600} ~0.4. 50 uL of the ΦX174-only cocktail and the generated phage cocktail were added to the wells as initial infections, and three wells with 250 uL phage LB only were added to the plate to normalize the OD_{600} measurements. The plate was again incubated in the microplate reader the same way, for 3 hours. Resistant-only wells were observed for growth inhibition, and in the meantime, another seed plate was prepared. Each 250 uL culture in the phage-infected plate was transferred into a 96-well plate (Thermo Fisher Scientific #N8010560) and the cells were pelleted at 3,000 × *g* for 10 min at 4 °C. Approximately 200 uL of the supernatants were collected and 50 uL of the supernatant from each mixed well was passaged into their corresponding resistant conditions in the new seed plate. Passages were repeated five times. At each passage, any resistant-only culture with inhibited growth due to a successfully counter-evolved phage cocktail was harvested: the cells were pelleted at 3,000 × *g* for 10 min at 4 °C, and the supernatant was sterile-filtered through a 0.22 um cellulose

acetate Spin-X Centrifuge Tube Filter (Thermo Fisher Scientific #07-200-385) by centrifugation at 10,000 × *g* for 3 min at 4 °C. The harvested phage cocktails were stored at 4 °C.

Bacteriophage cocktail isolate sequencing

A glycerol scrape of resistant *E. coli* C corresponding to each evolved phage cocktail was inoculated in 5 mL of LB and incubated at 37 °C overnight with 200 rpm agitation. 50 uL of the overnight culture was mixed with 5 mL of phage LB and incubated at 37 °C with 200 rpm agitation. The culture was grown to an OD_{600} of ~0.4 and 300 uL of the culture was added to 7 mL phage LB agar at a temperature of 42–46 °C monitored using an Infrared Thermometer (Ketokek #KT600B). To sequence individual phages in the evolved phage cocktails, a pipette tip was dipped in the cocktails and scraped on the surface of the agar plate and incubated at 37 °C for 3 hours. Eight individual plaques per plate were randomly picked and dipped in 6 uL of phage LB. 2 uL of the phage LB was used per PCR reaction with 1 uL each of 10 uM primers (**data S1**), 21 uL of dH_2O, and 25 uL of 2× CloneAmp HiFi PCR Premix (Takara Bio #639298) with the following settings: 98 °C for 30 sec, 98 °C for 10 sec and 62 °C for 10 sec and 72 °C for 40 sec repeated 25 times, and 72 °C for 5 min. The PCR amplicons were separated in 1.5% (w/v) UltraPure Agarose (Thermo Fisher Scientific #16500500) in 1× TAE gels stained with 1× SYBR Safe (Thermo Fisher Scientific #S33102) run at 120 V for 30 min, with GeneRuler 1 kb Plus DNA Ladder (Thermo Fisher Scientific #SM1331). PCR products of the correct size were purified by gel extraction using a QIAquick Gel Extraction Kit (Qiagen #28704) following the manufacturer's protocol and sequenced using Plasmidsaurus' Standard Purified Linear/PCR sequencing service. The sequencing fragments were aligned against all generated phage genomes and ΦX174 by MAFFT (*112*) with default settings in Geneious Prime version 2025.1.2 (https://www.geneious.com/) and analyzed for regions of homology. Evo-ΦR1 was chosen as the sequence in which eight out of eight plaques were identical from the cocktail evolved against resistant strain CR1, Evo-ΦR2 was chosen as the sequence in which six out of eight plaques were identical from the cocktail against resistant strain CR2.

Whole-genome sequencing analysis of ΦX174-resistant E. coli strains

To identify mutational differences potentially conferring resistance amongst ΦX174-resistant *E. coli* C strains, protein FASTA files generated by Bakta (*124*), provided by Plasmidsaurus' Standard Bacteria Genome with Extraction service, for resistant and susceptible *E. coli* C populations were compared. Briefly, each FASTA file was parsed to extract standardized protein descriptors and sequences. For all proteins with identical descriptors, corresponding amino acid sequences were initially compared using simple string equality, with identical sequences being classified as exact matches between resistant and susceptible strains. Non-identical protein sequences were subsequently aligned using MAFFT version 7.525 (*112*) to identify specific amino acid-level differences. Any remaining proteins with non-matching descriptors between the resistant and susceptible strains were then aligned against each other using MAFFT to account for potential large truncations, insertions, or deletions. Differences were then compiled and visualized using Matplotlib.

Supplementary Text: Discussion on biosafety, biocontainment, and biosecurity

The generative design of complete bacteriophage genomes represents a milestone in our ability to engineer biological systems, offering paths to new biotechnologies and therapeutics. To ensure responsible development, this powerful capability advance necessitates a careful and unique consideration of biosafety, biocontainment, and biosecurity, and current institutional and policymaker guidance will need to evolve accordingly (*125*). The ability to understand and manipulate complex biological systems to cure disease and ease human suffering has been a longstanding goal of scientific research. If done safely and responsibly, generative AI systems have immense promise for advancing humanity toward this goal.

The goal of this work was to design bacteriophage genomes using generative models, but hypothetical applications of this framework to other viruses, such as eukaryotic viruses, might be feasible at a similar scale. While applying genome language models to eukaryotic viral sequences may offer scientific and clinical value—for example, in forecasting viral evolution or optimizing gene therapy vectors—such work should proceed only after careful review and deliberation with other researchers and experts in biosafety, biocontainment, and biosecurity, as well as adhering to all current and future governance structures and best practices (*126–128*).

Biosafety

To devise appropriate safety strategies, researchers seeking to engage in generative phage design should consult with experts to plan their experiments at key steps before and during their investigations. Our study was conducted with appropriate institutional biosafety review and was designed to be intrinsically safe and controllable. We selected the lytic phage ΦX174, as well as its non-pathogenic host, *E. coli* C, as our design template. This system is well-studied, tractable, and has a long history of safe use and engineering in molecular biology research (*13*, *36*, *39–41*). Our method allows for precise, user-specified constraints, including a strict tropism filter to ensure that generated phages specifically target our laboratory host strain, which offers more control than traditional "phage hunting" methods that isolate phages from the environment with unknown host ranges or phenotypes. The success of our tropism constraint was validated experimentally, as none of the 285 tested assemblies showed activity against off-target *E. coli* K-12 strains.

Moving forward, understanding the limits of current model capabilities, guidelines around training data curation, and development of responsive policies are important steps in understanding and mitigating the potential for unintended harm (*78*, *79*, *129*). Future biosafety frameworks should also carefully balance risk mitigation alongside the ability to conduct beneficial technological development (*130*).

Biocontainment

As with most biological experiments involving synthetic DNA sequences, generative phage design creates the potential for such sequences to move beyond controlled laboratory environments into natural settings and ecosystems. To ensure proper containment, experiments should be conducted in adequately isolated environments. In planning and conducting our experiments, we leveraged decades of precedent and established biocontainment protocols for working with phage and non-pathogenic bacterial hosts (*77*). As an additional precaution, we performed all work involving phages within a biosafety cabinet with appropriate PPE and dedicated equipment, which was regularly sterilized with 70% ethanol, 10% bleach, and UV treatment, and we disposed of any

waste as biohazardous waste. Our previously described tropism constraint also provided an additional containment measure. Furthermore, the risk of significant ecological disruption by any designed ΦX174-variant phage is minimal, given both our containment strategies and the powerful systems that bacteria have evolved for phage defense.

Importantly, design of modified viruses that infect humans (especially of select agents or pathogens with enhanced pandemic potential) require much stricter biosafety review and biocontainment protocols than those described in this study. Such a hypothetical study would require compliance with government regulations on dual use research of concern (DURC) (*126–128*) and would most likely require experimental facilities that adhere to biosafety level (BSL) 3 or 4 containment (*131*). We note that bacteriophages and laboratory strains of *E. coli* do not fall under these categories.

Biosecurity

Generative biological models, including those for genome design, might be repurposed by individual-, institutional-, or state-level actors for intended harm (*75*). Hence, it is important that researchers developing or leveraging these tools for viral genome design seek out biosecurity expertise. Prior to initiating this work, and throughout its execution, we consulted with biosecurity experts to identify appropriate safeguards, which informed the implementation of model-intrinsic guardrails, constrained computational design criteria, open-sourcing of our data, methods, and models, and reliance on commercial DNA synthesis screening.

A primary layer of security is built directly into the genome language models themselves, as the generative capabilities of these models are fundamentally shaped by their training data (*26*, *78*, *129*). In this and previous work, we have deliberately withheld all viruses with eukaryotic hosts, including those pathogenic to humans, from the models' training data. As we have demonstrated previously, this data exclusion strategy successfully prevents Evo 2 from prediction and design tasks related to human viruses (*26*). Enabling the generation of harmful viruses currently requires significant investment in data, computational and experimental resources, methodological development, and domain-specific knowledge, but the growing accessibility to such infrastructure makes malicious use cases increasingly possible (*75*). Securing dual-use pathogenic data could help mitigate these risks (*78*), as well as restricting and monitoring usage capabilities for publicly available large language models that currently make high-risk knowledge increasingly accessible (*132*).

We have shown that both pretraining and fine-tuning were necessary for coherent generation of the desired phages and of other complex systems such as CRISPR-Cas operons (*28*). For this study, we specialized the models by fine-tuning them exclusively on a curated dataset of publicly available *Microviridae* genomes, a viral family composed entirely of bacteriophages. This targeted training means the model retains its poor performance on eukaryotic viruses while enhancing its ability to generate sequences within a specific phage family. Recent work further showed that Evo 2 fine-tuned on bacteriophages does not have improved predictive capabilities on pathogenic human viruses (*129*).

We have made the *Microviridae* data and fine-tuned models publicly available for the benefit of the research community. Open models can be rigorously examined by researchers, including for their safety profiles, and can be repurposed for pathogen surveillance or for DNA synthesis screening (*80*). While it may be possible to fine-tune Evo 2 on harmful human viruses, this involves non-trivial technical expertise and computational resources, and may not necessarily lead to improved pathogenicity prediction over task-specific models (*129*). While many protein

and genome language models with explicit training on human viral sequences already exist in the public domain (*44*, *133–139*), future model development efforts could implement data exclusions, as we have done, or even consider withholding certain model checkpoints from public use.

Our design criteria for generative phage design implemented commonly used computational biology tools that are all publicly available such as MMseqs2, ESMFold, geNomad, and BLAST (*42–44*, *97*). We also developed two new bioinformatic methods that may be useful to bacteriophage researchers—a bespoke ΦX174 gene annotation tool and an algorithm for scoring genetic architecture similarity between genomes—both of which are highly specialized to ΦX174 and similar bacteriophages (**Materials and Methods**). Our phage genome annotation technique does not perform well on eukaryotic viruses, which have distinct genetic architectures (for example, long genomic polyproteins), mitigating the potential for misuse (**fig. S4D**).

Generative biology can also be secured by improved DNA synthesis screening and sequence tracing (*127*, *140*), including the use of generative models as part of the screening process (*80*). Future work could also leverage the Evo models to enable intelligent synthesis screening at the genome scale, beyond individual proteins or genes. For this study, we leveraged commercial synthesis vendors with compliant screening protocols.

Biosecurity remains an open and evolving field, and evaluating broader societal risks requires dedicated expertise, including through continued investment in biosecurity organizations. Matters involving security should be primarily led by security professionals, but clear pathways for engagement with researchers will be essential as generative biology matures. Close collaboration involving researchers, ethicists, policymakers, patients, members of the public, and other stakeholders will help to safely and responsibly realize the immense potential of both generative phage design and generative genomics.

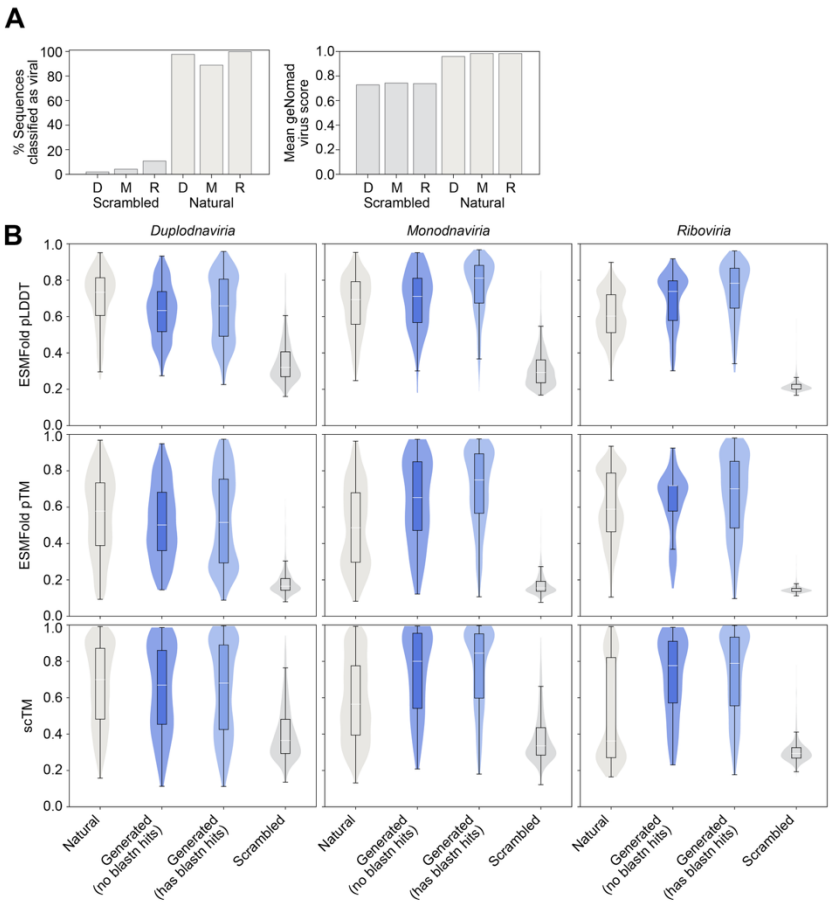

Fig. S1. Supplementary data for bacteriophage realm sequence generation. (**A**) Percent of natural and scrambled natural sequences classified as viral by geNomad (left), and their mean geNomad virus scores (right). D, *Duplodnaviria*; M, *Monodnaviria*; R, *Riboviria*. geNomad effectively discriminates real natural sequences from scrambled ones. (**B**) Generated sequences stratified by whether they have a nucleotide blast (blastn) hit or not both have ESMFold-predicted protein structures with mean predicted and self-consistency template modeling (pTM and scTM) scores similar to natural proteins, higher than scrambled natural sequence controls.

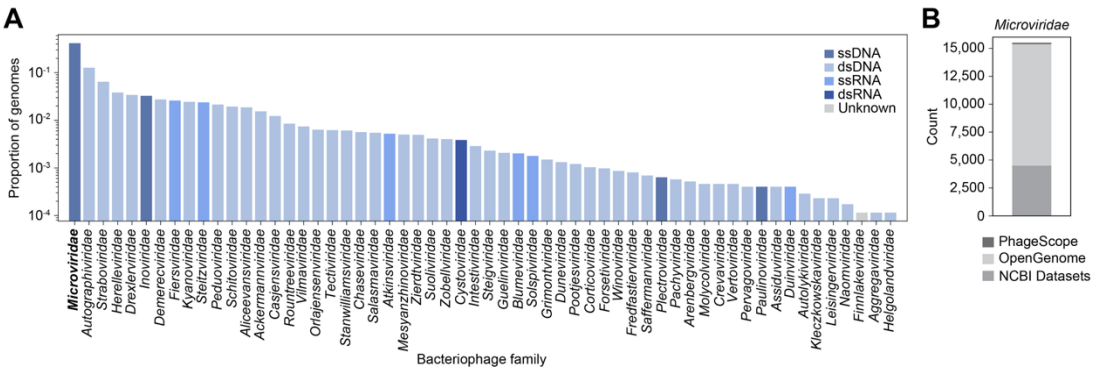

Fig. S2. *Microviridae* sequence data. (A) Relative proportions of complete phage genomes on NCBI Virus as of May 12, 2025 show that *Microviridae* sequence data are the most abundantly available amongst phage families. (**B**) *Microviridae* data for fine-tuning collected from PhageScope, OpenGenome, and NCBI Datasets.

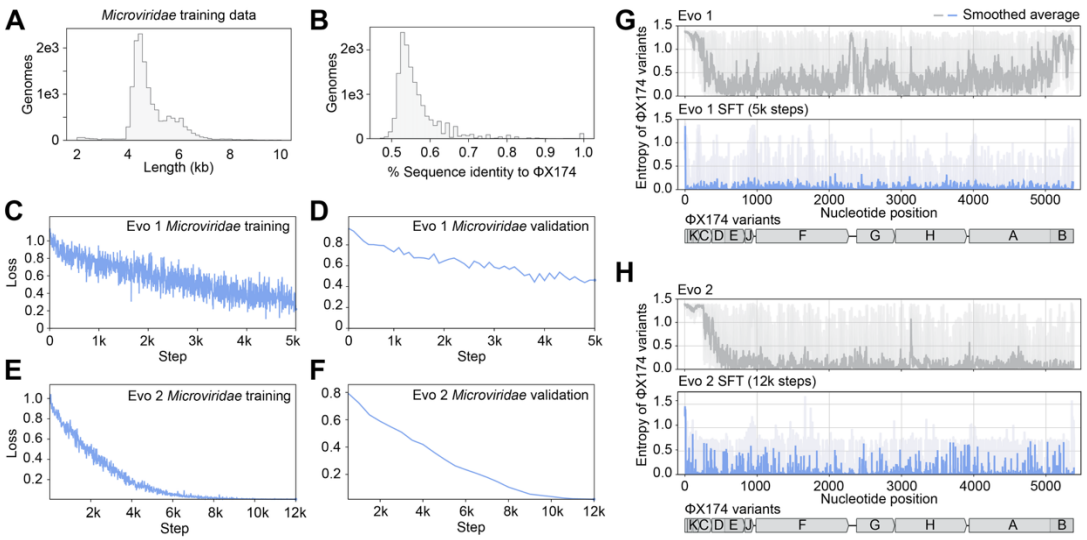

Fig. S3. Fine-tuning Evo 1 and Evo 2 on *Microviridae* genomes. (**A**) Lengths of genomic sequences in the *Microviridae* training data. kb, kilobase. (**B**) Percent sequence identity of genomes in the *Microviridae* data aligned to ΦX174. (**C–F**) Training and validation loss curves for Evo 1 fine-tuning (**C–D**) and Evo 2 fine-tuning (**E–F**). (**G–H**) Per-position entropy of ΦX174 variant sequences calculated by Evo 1 7B 131K and Evo 1 SFT at 5k training steps (**G**), and Evo 2 7B 8K and Evo 2 SFT at 12k training steps (**H**). Dark gray and dark blue, smoothed average of positional entropies; light gray and light blue, range of position entropies. Schematic of the ΦX174 genome is shown below.

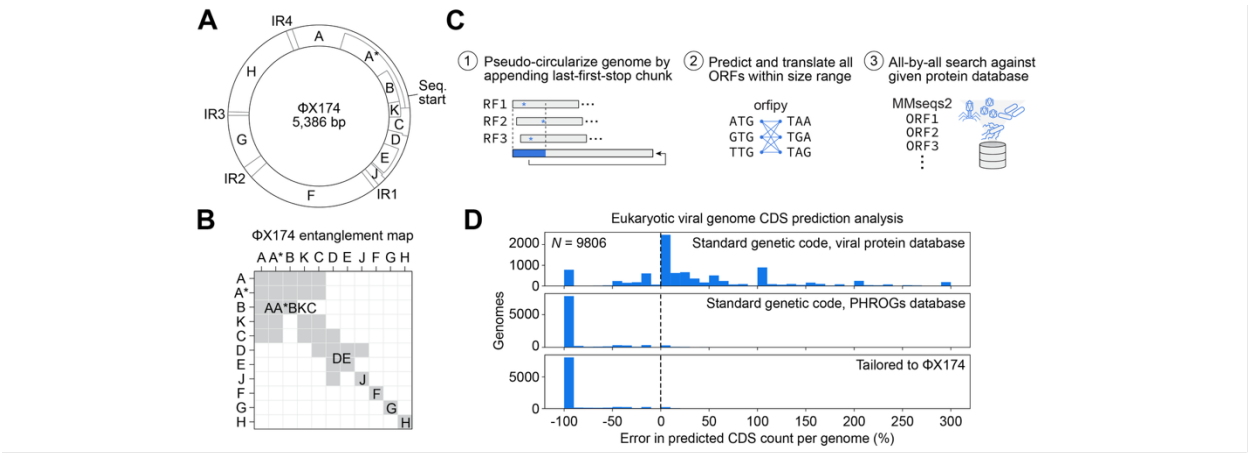

Fig. S4. Method for predicting genes in ΦX174-like sequences. (**A**) Circular schematic of the ΦX174 genome. Seq., sequence; IR, intergenic region. (**B**) Pairwise matrix indicating overlapping genes (gray) in the ΦX174 genome. (**C**) Steps of our coding sequence (CDS) prediction method. First, sequences are "pseudo-circularized" by searching for the first stop codon in each reading frame (RF), identifying the most downstream first stop codon position, extracting the sequence up to the position, and appending it to the end of the genome. Then, all possible open reading frames (ORFs) are determined with input start and stop codons using orfipy (*106*). Finally, an all-by-all search of the ORFs against an input protein database using MMseqs2 (*97*) is performed. (**D**) The method performs poorly at predicting CDS counts of eukaryotic viral genomes.

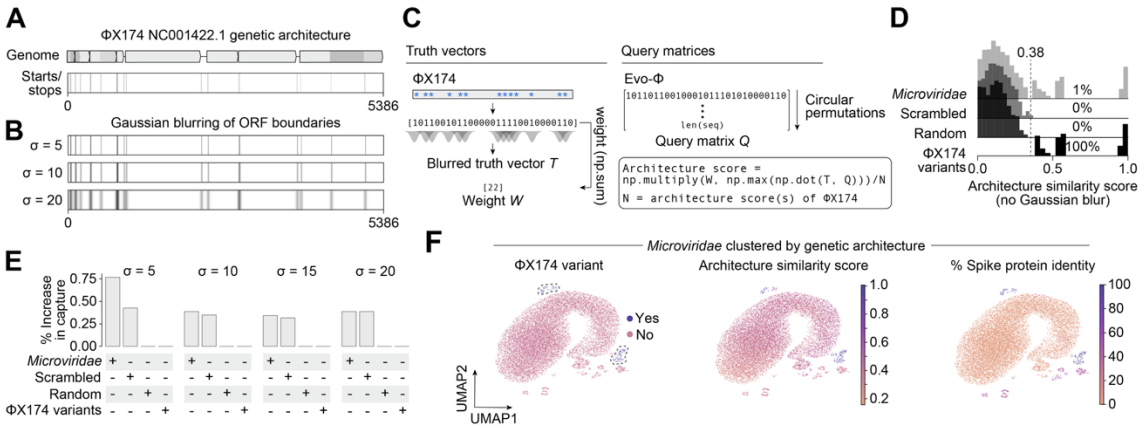

Fig. S5. Scoring architectural similarity of genomes. (**A**) Top: Schematic of ΦX174 NC_001422.1 genome architecture. Bottom: Visualization of one-hot encoding of the known start and stop codon boundaries in the ΦX174 genome. (**B**) Applying Gaussian blurs to the one-hot encoding produces smoothed open reading frame (ORF) boundary profiles, making the similarity metric less sensitive to exact start/stop positions. (**C**) Architectural similarity scoring process (**Materials and Methods**). (**D**) Distribution of architecture similarity scores normalized to ΦX174 NC_001422.1 without Gaussian blurring show that ΦX174 variants score highly (close to 1), while natural *Microviridae*, scrambled *Microviridae*, and random sequences have low scores. A score greater than 0.38 reliably delineates ΦX174-like sequences. (**E**) Increasing the Gaussian blur parameter σ expands the capture of sequences above the 0.38 threshold. At σ = 5, the method balances sensitivity (capturing more ΦX174-like variants) with specificity (limiting scrambled or random sequences scoring as high). (**F**) *Microviridae* sequences clustered by their nearest one-hot encoded start/stop codon vectors relative to the ΦX174 reference. ΦX174 variants form distinct clusters with both high architecture similarity scores and high spike protein sequence identity to the ΦX174 spike protein.

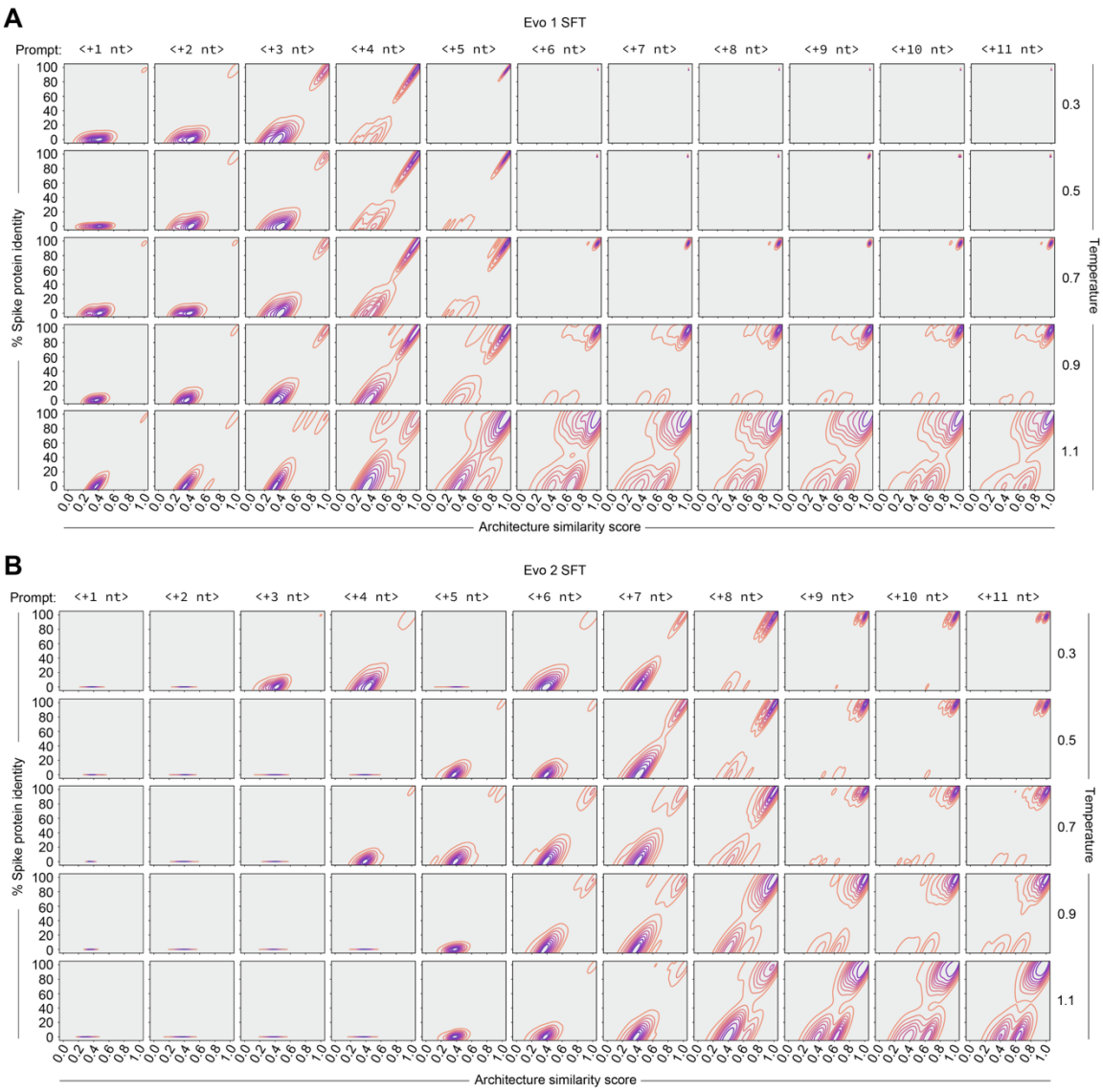

Fig. S6. Generation temperature and prompt sweep. (**A**) Percent predicted spike protein identity and architecture similarity score of Evo 1 SFT- and Evo 2 SFT-generated sequences with increasing prompt lengths and generation temperatures. N = 1000 sequences per parameter combination. nt, nucleotide.

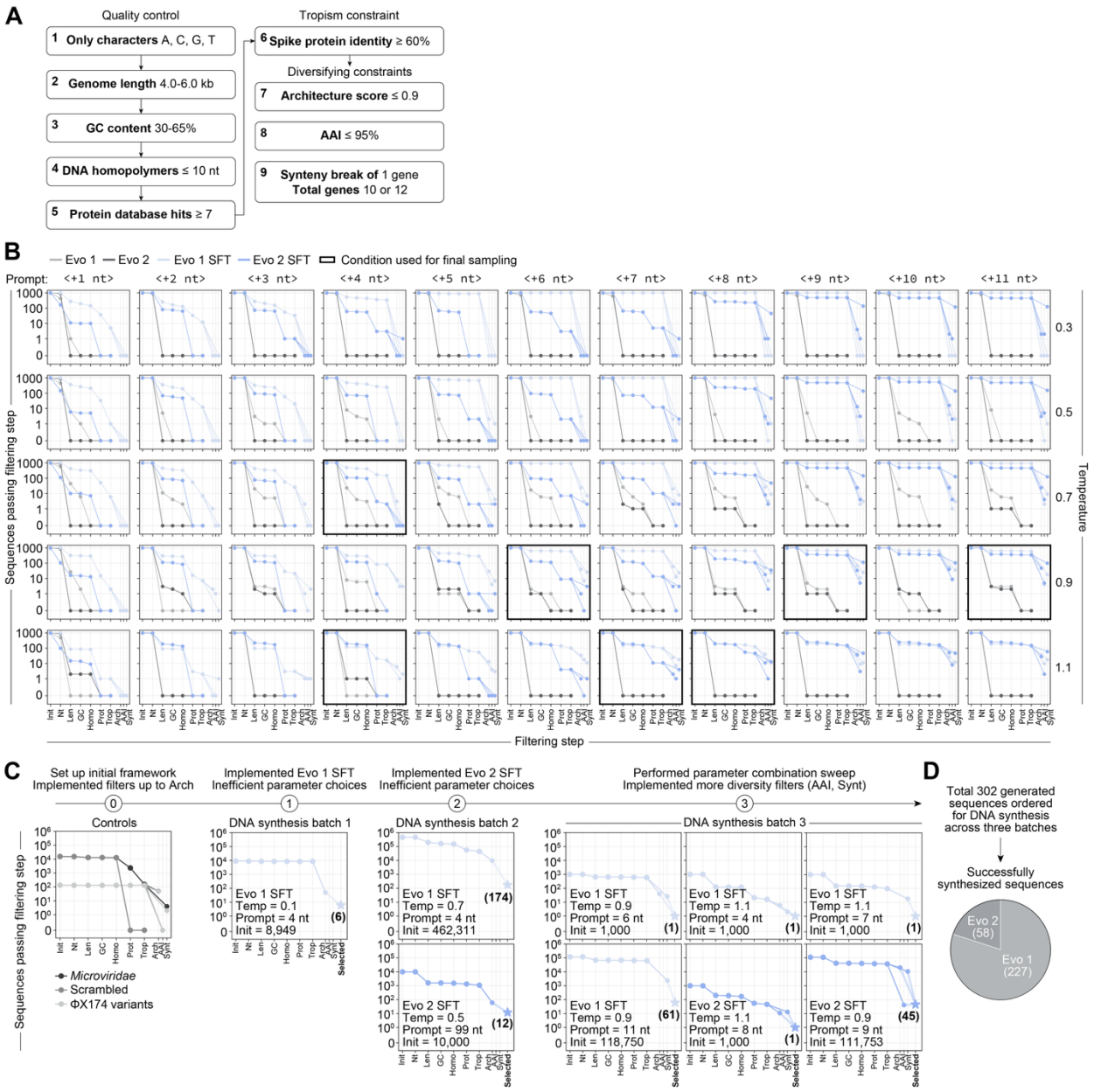

Fig. S7. Supplementary data for bacteriophage genome design filtering. (**A**) Final filtering criteria for generated phage genomes. Diversifying constraints were preferred but not always applied. kb, kilobase; nt, nucleotides; AAI, average amino acid sequence identity. (**B**) Prompt length and generation temperature sweep for sampling generated genomes with Evo 1 (light gray) and Evo 2 (dark gray) base models, or Evo 1 (light blue) and Evo 2 (dark blue) SFT models. The amount of sequences passing each filtering step are shown, with the three final diversification filters (Arch, AAI, Synt) separately applied to sequences passing the Trop filter. Bolded boxes, sampling conditions used for final candidates; Init, initial amount of sequences sampled; Nt, nucleotide character filter; Len, genome length filter; GC, percent GC content filter; Homo, DNA homopolymer filter; Prot, protein database hit count filter; Trop, tropism protein identity filter; Arch, architecture similarity score filter; AAI, average amino acid identity filter; Synt, syntenic gene count filter; nt, nucleotide. $N = 1000$ sequences sampled per parameter combination. (**C**) We

performed sequence sampling in three batches of DNA synthesis and experimental validation, where we improved our sampling and filtering strategy after each batch. Large amounts of sequences in early phases were sampled due to inefficient sampling parameter choices. We selected final sequences following optional diversification filters by manual inspection. Bolded number in parentheses and star, final number of sequences selected for DNA synthesis; Temp, generation temperature; nt, nucleotide; Init, initial number of sequences sampled. (**D**) Proportions of final sequence candidates generated by Evo 1 and Evo 2 that could be synthesized.

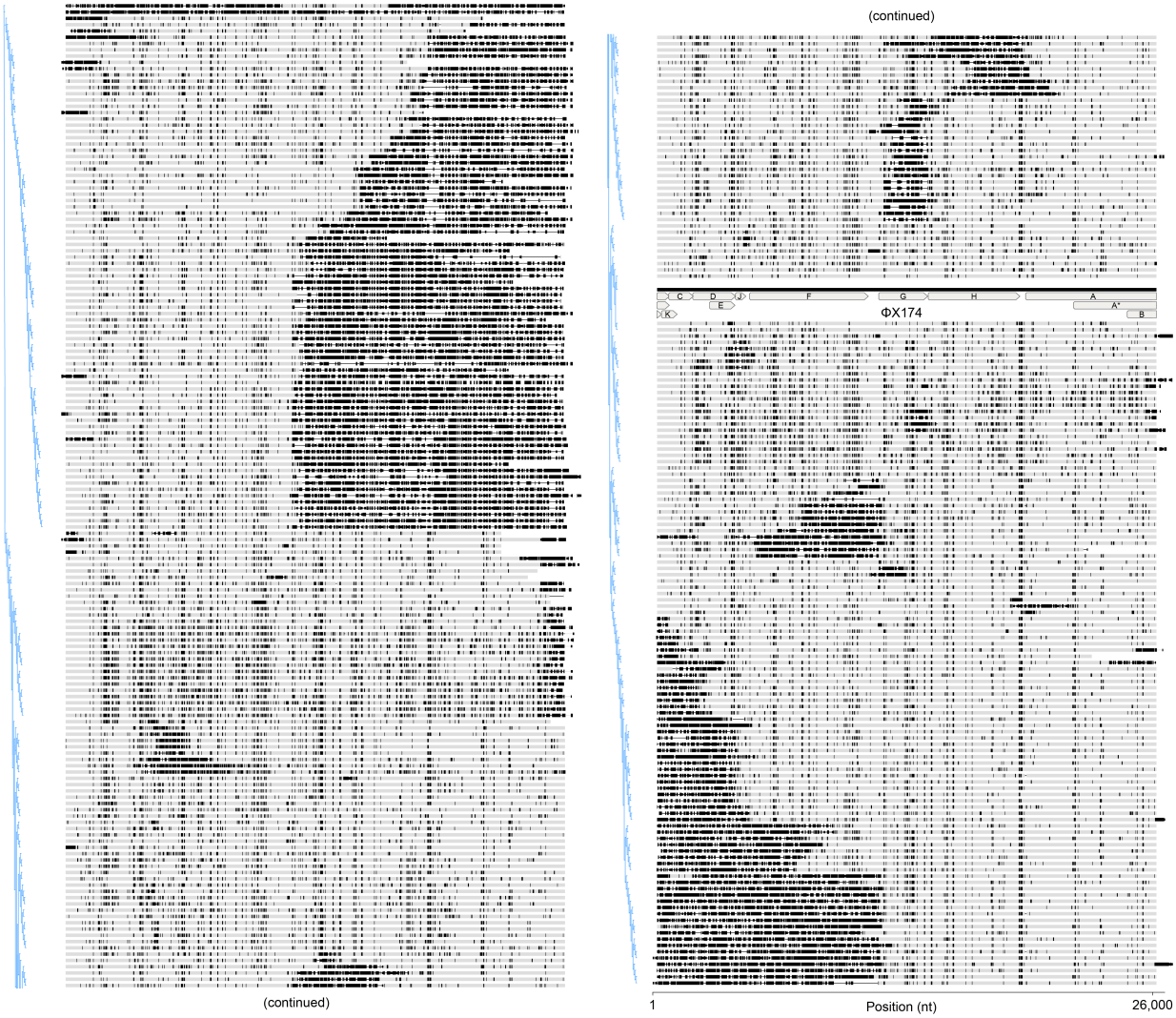

Fig. S8. Multiple sequence alignment of synthesized bacteriophage genome design candidates. Multiple sequence alignment of final bacteriophage genome candidates and ΦX174, ordered by a neighbor-joining phylogenetic tree. The x-axis represents alignment columns (including gap positions introduced during alignment), not native genome coordinates. Black, unaligned positions to ΦX174; gray, aligned positions to ΦX174; nt, nucleotide.

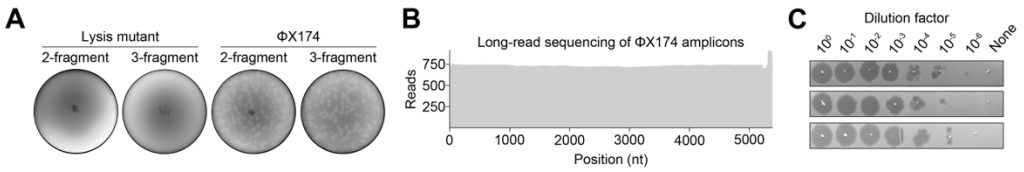

Fig. S9. Supplementary data for bacteriophage genome assembly and rebooting. (**A**) Plaque assays of ΦX174 wild-type or lysis mutant genome assemblies transformed into *E. coli* C show robust rebooting of wild-type phage. Two- and three- fragment Gibson assemblies were tested. (**B**) Long-read sequencing read counts of PCR amplicons from ΦX174 plaques. nt, nucleotide. (**C**) Titrations of purified ΦX174 particles on *E. coli* C.

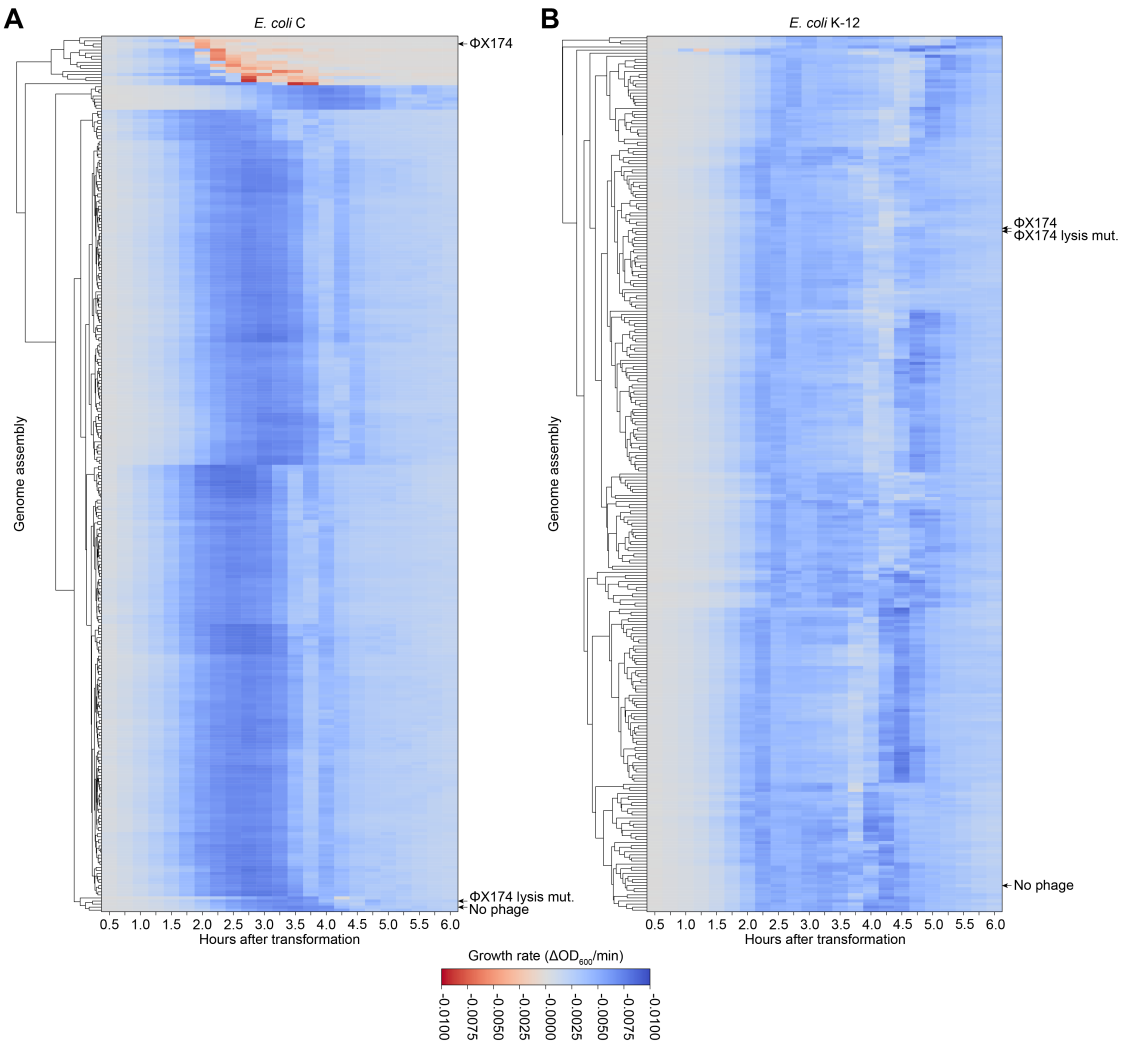

Fig. S10. Supplementary data for generated bacteriophage rebooting in *E. coli* C and *E. coli* K-12. (**A–B**) Heatmaps of growth rate of *E. coli* C (**A**) and *E. coli* K-12 (**B**) over six hours after transformation with assembled phage genomes. Growth rates of *E. coli* C transformed with viable generated phages clusters with ΦX174.

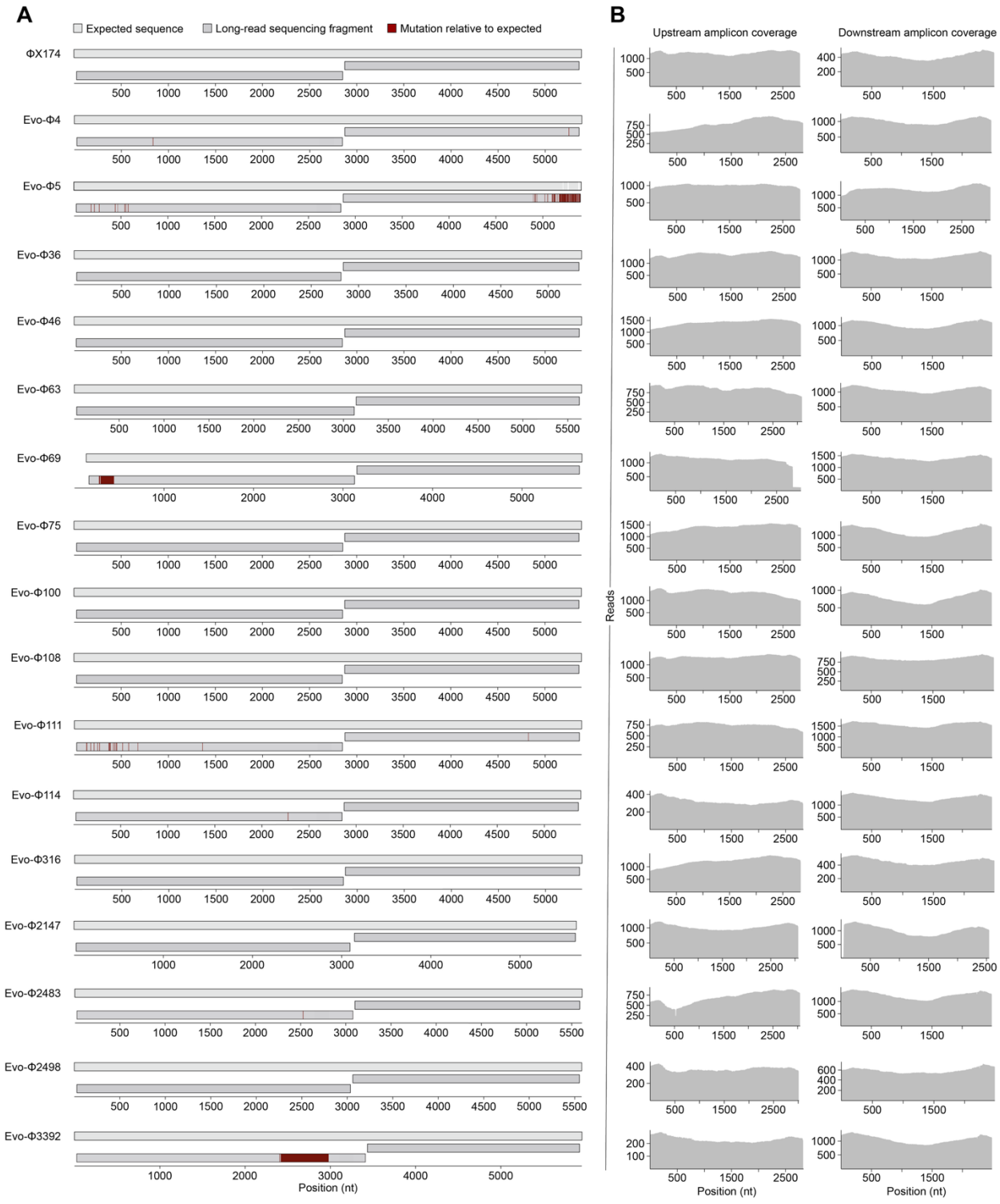

Fig. S11. Long-read sequencing of viable generated bacteriophage genomes. (**A**) Long-read sequencing of PCR amplicons (dark gray) from ΦX174 and generated phage stocks, aligned against the expected sequence (light gray). Red, mutations relative to the expected sequence; white, gaps in the alignment; nt, nucleotide. (**B**) Read counts from long-read sequencing.

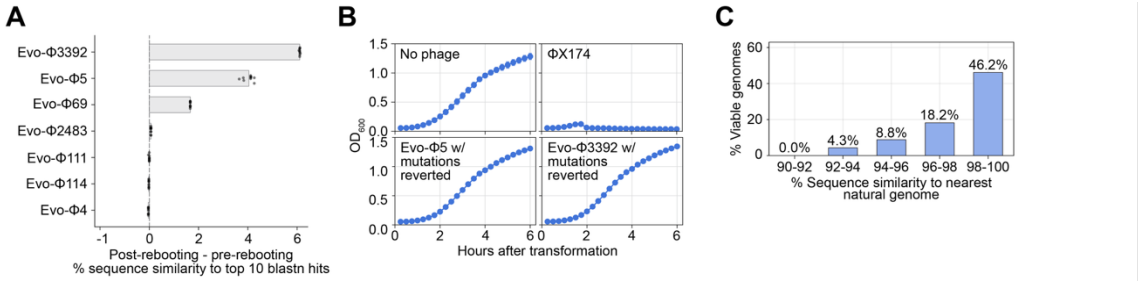

Fig. S12. Viability of natural-like mutations in generated phages. (**A**) Percent change in sequence similarity of generated genomes that acquired mutations post-rebooting compared to the original sequences ordered for DNA synthesis to their top 10 nucleotide blast (blastn) hits. Bar height, mean; data point, percent change in sequence similarity value. (**B**) Upon re-synthesizing the original sequences of Evo-Φ5 and Evo-Φ3392 (reverting the mutations they acquired during previous rebooting), they do not inhibit growth in *E. coli* C. (**C**) Viability rates of generated genomes tested, sorted by bins of percent sequence similarity to their nearest natural genome.

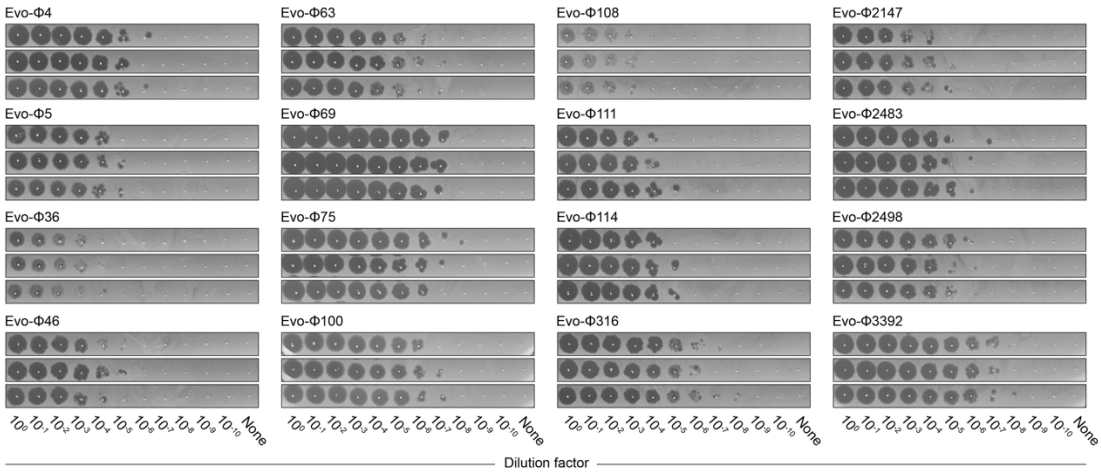

Fig. S13. Titration of viable generated bacteriophages. Plaque titrations of purified generated phages on *E. coli* C. Each row is a titration replicate.

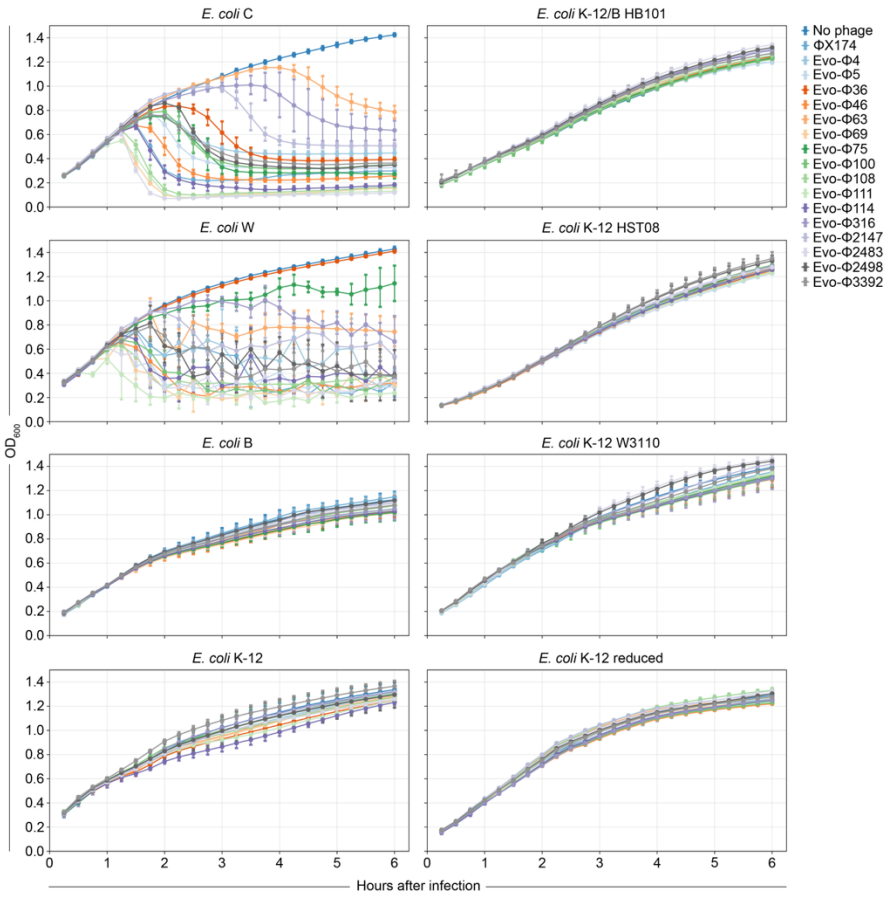

Fig. S14. Supplementary data for host tropism assay. Growth curves of eight different *E. coli* strains infected with generated phages and ΦX174. $N = 3$ infection replicates.

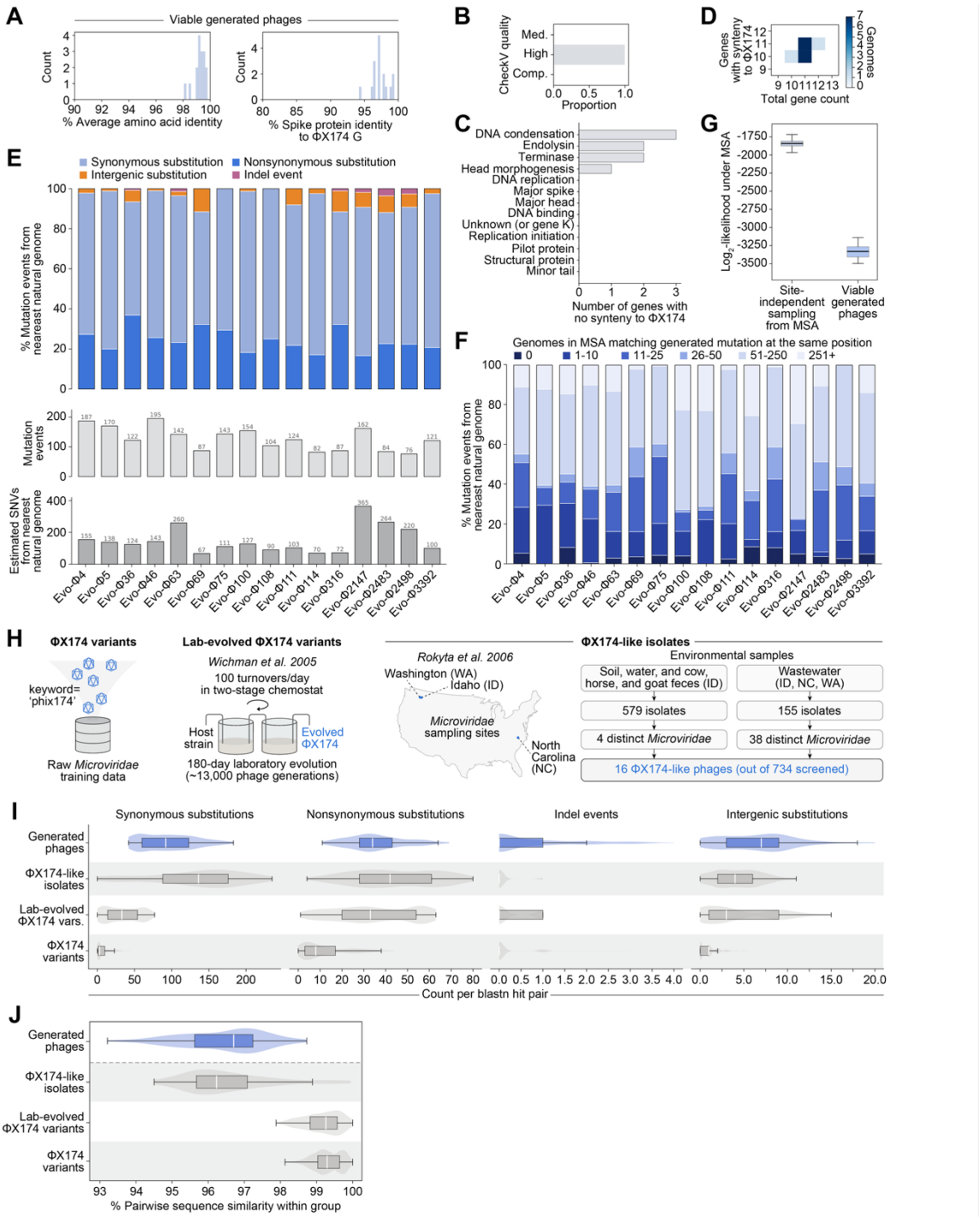

Fig. S15. Supplementary data for ΦX174-like bacteriophage sequence diversity. (**A–B**) Metrics for viable generated phages, including average amino acid identity and spike protein identity (**A**), and CheckV quality (**B**). Med., Medium; Comp. Complete. (**C**) Viable generated phages encode genes with synteny breaks to ΦX174. (**D**) Heatmap of total gene count versus number of syntenic genes shows that most viable phages contain one or no single-gene breaks in synteny to ΦX174. (**E**) Percentage of mutation events colored by mutation type (upper), their total counts (middle), and estimated single nucleotide variant (SNV) positions (lower) per generated

phage genome relative to their nearest natural genome. (**F**) Percentage of mutation events per generated phage genome relative to their nearest natural genome, colored by the number of genomes in the de-duplicated ΦX174-like clade multiple sequence alignment (MSA) with the same mutation. (**G**) Viable generated phage genomes are far less likely to be sampled by an MSA of natural ΦX174-like clade sequences compared to site-independent sampling of genomes from the same MSA. (**H**) We collected three genomic datasets of ΦX174 variants or ΦX174-like phages for comparison to generated phages: ΦX174 variants from our raw *Microviridae* dataset, queried by the keyword 'phix174'; ΦX174 variants evolved over 180 days of laboratory evolution (roughly equivalent to 13,000 phage generations) in a two-stage chemostat (*60*); and ΦX174-like isolates harvested from three locations across the United States through various environmental sampling methods (*61*). (**I**) Number of synonymous substitutions, nonsynonymous substitutions, indel events, and intergenic substitutions of genomes relative to their top 10 nucleotide blast (blastn) hits for each dataset of phages. Generated phages have lower synonymous and nonsynonymous substitution counts compared to ΦX174-like isolates, but more indel events and intergenic substitutions. They generally have more of all types of mutations compared to lab-evolved ΦX174 variants and natural ΦX174 variants. (**J**) Generated phages have broader sequence diversity amongst each other compared to the other ΦX174 variant and ΦX174-like datasets.

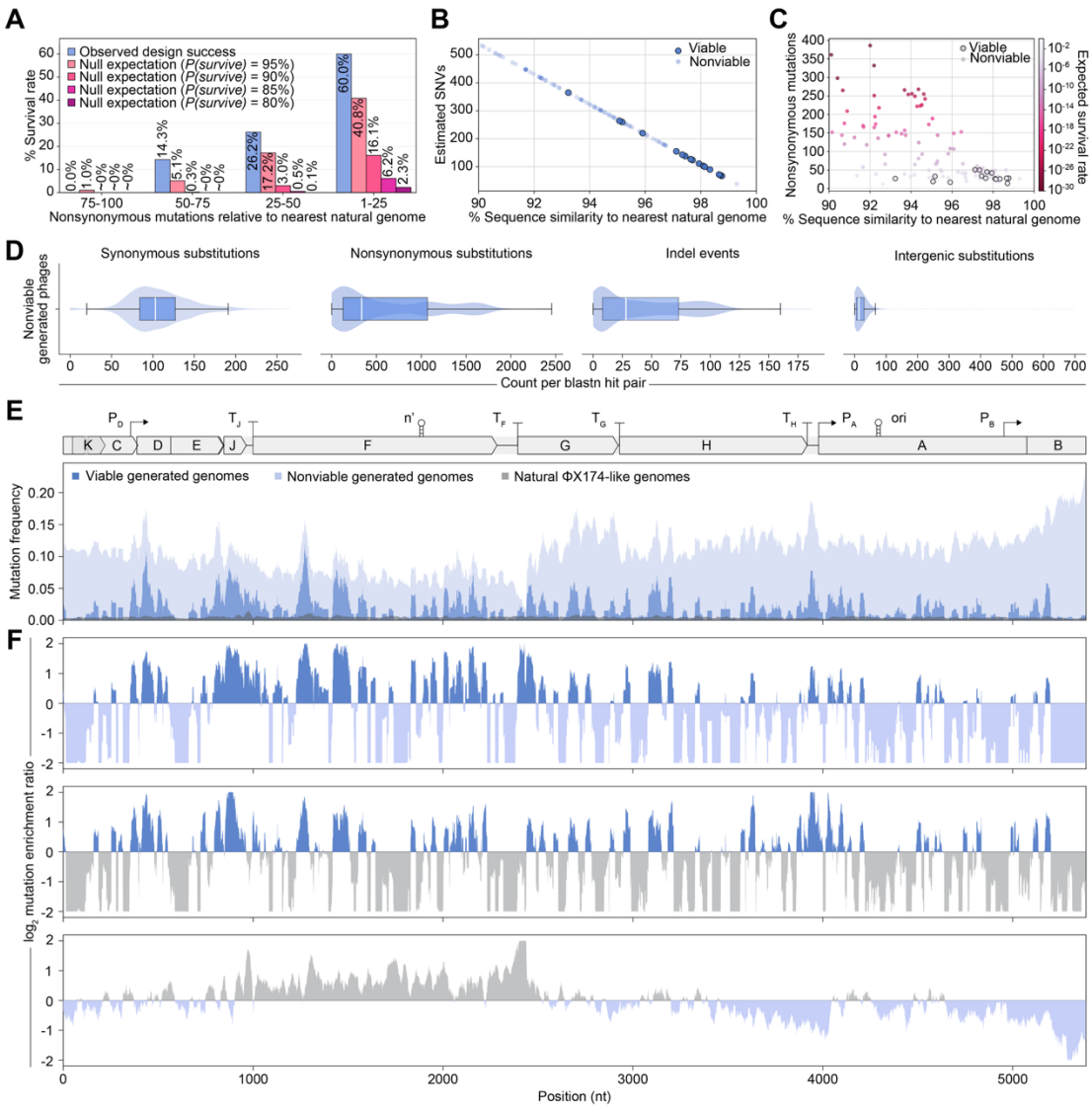

Fig. S16. Comparison of viable and nonviable generated mutations. (**A**) Observed survival rates (blue) of generated phages are higher than expected survival rates (pink) for nonsynonymous mutation counts. The null expectation is computed based on an empirically determined nonsynonymous mutation lethality of ~20% (i.e., a survival probability *P(survive)* = 80%) in ΦX174 (*16*; **Materials and Methods**), averaged across the nonsynonymous mutation counts of phage genomes in each bin. It is also computed for simulated lethality rates of 15% (*P(survive* = 85%), 10% (*P(survive)* = 90%), and 5% (*P(survive)* = 95%). (**B**) Estimated number of single nucleotide variant (SNVs) vs. percent sequence similarity to nearest natural genomes of viable (dark blue, outlined) and nonviable (light blue) generated genomes. Only genomes within 90–100% sequence similarity to their nearest natural genome are shown. (**C**) Number of nonsynonymous mutations and percent sequence similarity of viable and nonviable generated phages to their nearest natural genomes, colored by expected survival rate. (**D**) Number of synonymous substitutions, nonsynonymous substitutions, indel events, and intergenic substitutions of nonviable generated genomes relative to their top 10 nucleotide blast (blastn) hits per phage. (**E**) Number of per-position mutation events per genome for viable (dark blue) and

nonviable (light blue) generated genomes, and natural ΦX174-like genomes (dark gray) mapped onto reference positions of ΦX174 through a multiple sequence alignment (MSA) of ΦX174-like clade sequences (**Materials and Methods**). Schematic of ΦX174 genome is shown above. (**F**) Per-position mutation enrichment ratios of $\log_2$-transformed viable over nonviable generated genome mutations (upper), viable generated over natural ΦX174-like genome mutations (middle), and natural ΦX174-like over nonviable generated genome mutations (lower), after normalizing their total frequencies to 1. nt, nucleotide.

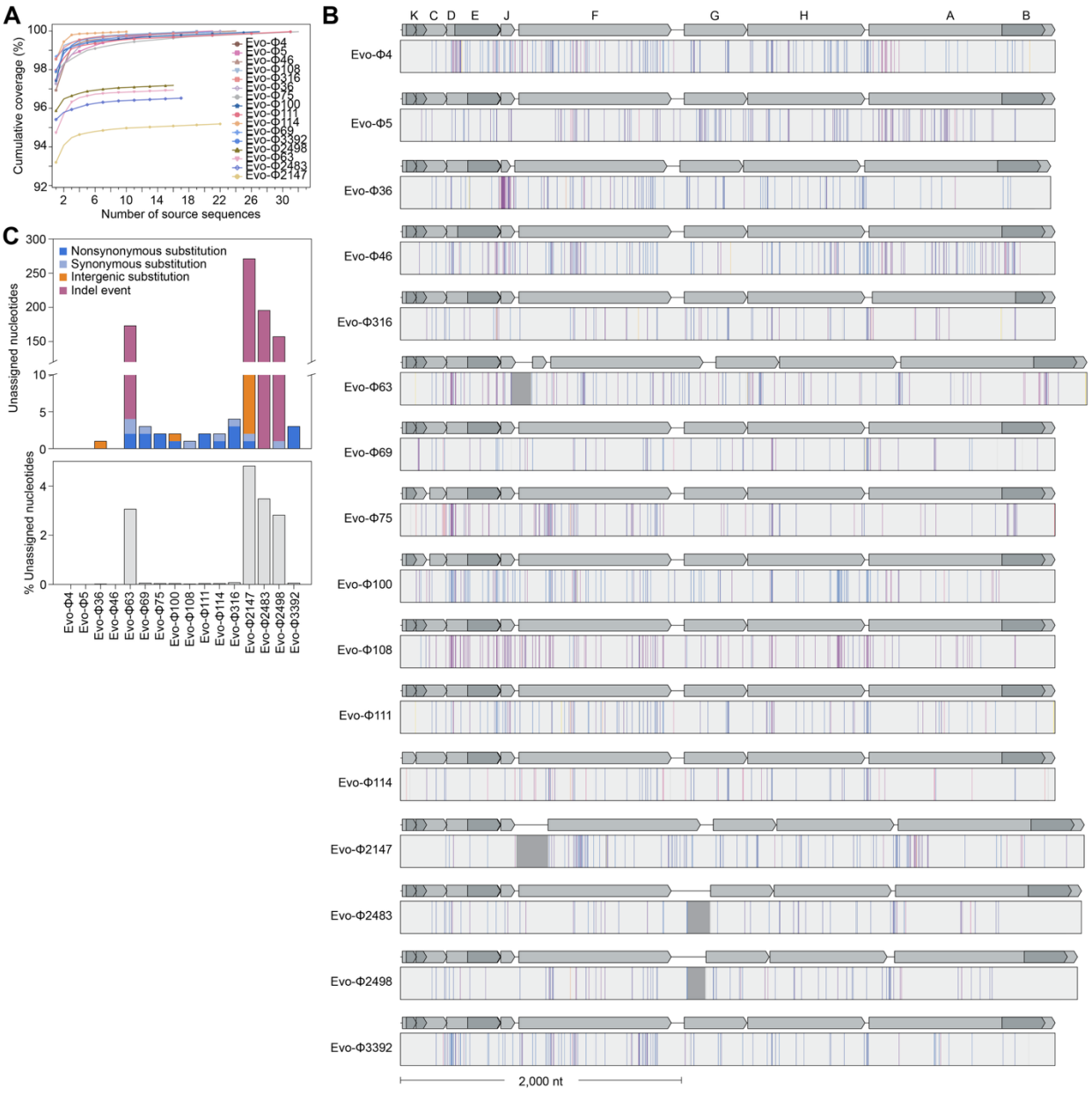

Fig. S17 continued on next page

Fig. S17. Cumulative genome attribution analysis. (**A–B**) Percent cumulative sequence coverage of generated phages highlights that mutations in most generated phages cannot be completely attributed to mutations seen in nature. The sequences were aligned by nucleotide BLAST to sequences in the core_nt database until all nucleotides were accounted for or there were no significant hits for remaining nucleotides. Nucleotides are colored as "Unassigned" or by their

top nucleotide BLAST hit in the core_nt database. Mutations in most generated phages cannot be completely attributed to mutations seen in nature. (**C**) Number of unassigned nucleotides (upper), colored by mutation type, and their percentage of the genome (lower) from the cumulative attribution analysis, for each generated phage.

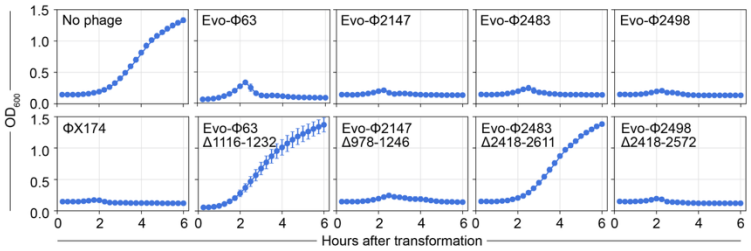

Fig. S18. Deletion mutant viability assay. Growth curves of *E. coli* C infected with viable generated phages containing large insertions in their genomes, and corresponding deletion mutants. The Evo-Φ63 and Evo-Φ2483 deletion mutants are nonviable while the others tolerated deletions. Data point, mean OD_{600} value; error bar, standard deviation; $N = 3$ growth replicates.

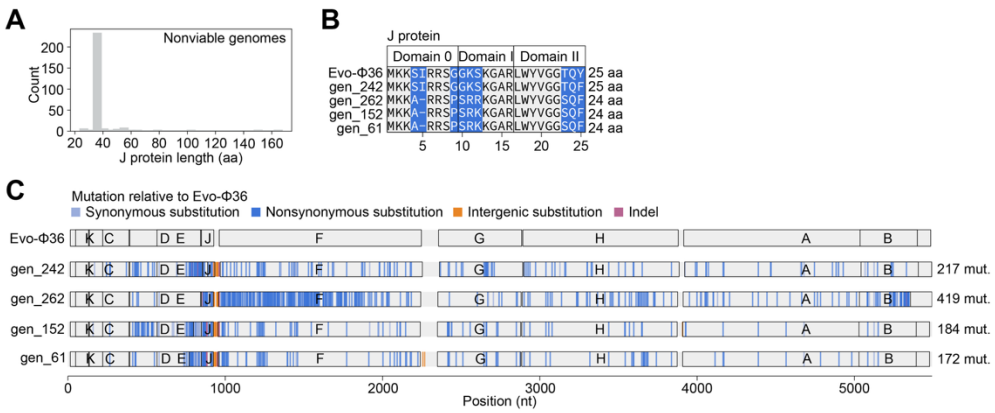

Fig. S19. Truncated J genes in nonviable generated genomes compared to Evo-Φ36. (**A**) Lengths of J proteins predicted in nonviable generated genomes show that the majority of generated J proteins are similar in length to the 38 amino acid (aa) ΦX174 J protein. (**B**) Only four nonviable genomes tested contain predicted J proteins similar in length to Evo-Φ36; however, they all contain at least one or more nonsynonymous mutations relative to Evo-Φ36 J (blue columns). Positions of J protein domains 0, I, and II are shown. (**C**) The four nonviable genomes with truncated J genes also contain hundreds of mostly nonsynonymous mutations relative to Evo-Φ36, indicating that Evo-Φ36 is a unique phage tested amongst all the generated phages.

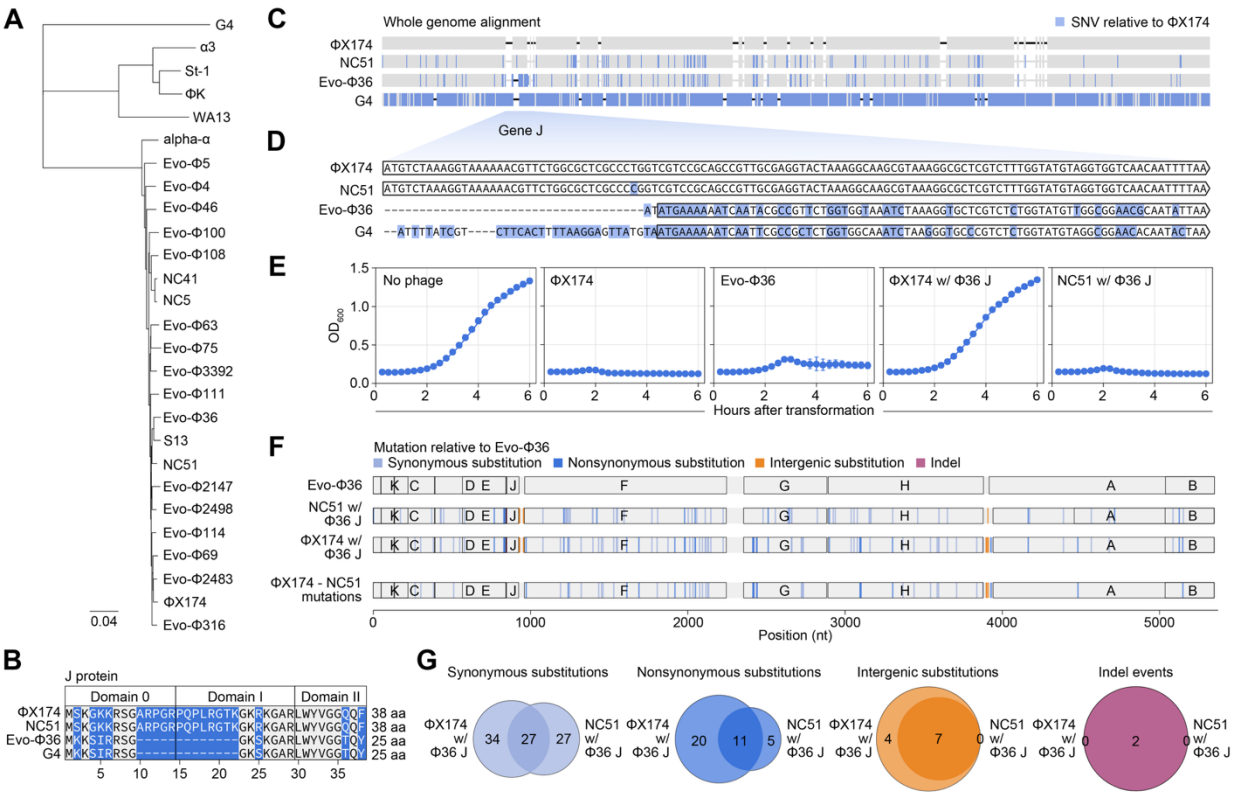

Fig. S20. Genomic context dependency analysis of Evo-Φ36 J gene. (A) Neighbor-joining phylogenetic tree of representative *Microviridae* phages and generated phages. (**B**) Protein sequences of the J gene of ΦX174, Evo-Φ36, and phage G4, with positions of domains 0, I, and II shown. Blue columns, nonsynonymous mutations or indels; aa, amino acids. **(C–D)** Whole genome alignments of ΦX174, NC51, Evo-Φ36, and G4 show that Evo-Φ36's J gene is highly similar to G4's J gene yet its genomic context is much more similar to its nearest natural genome NC51 and ΦX174. Blue, single nucleotide variant (SNV) positions relative to ΦX174. (**E**) Growth curves of *E. coli* C transformed with genomes of ΦX174, Evo-Φ36, ΦX174 with its J gene replaced by Evo-Φ36's J gene (ΦX174 w/ Φ36 J), and NC51 with its J gene replaced by Evo-Φ36's J gene (NC51 w/ Φ36 J) show that Evo-Φ36's J gene is viable in the context of NC51's genome but not in ΦX174's genome. Data point, mean OD_{600} value; error bar, standard deviation; $N = 3$ growth replicates. (**F–G**) ΦX174 contains unique SNV mutations relative to Evo-Φ36 and NC51 (**F**), which can be isolated to 58 synonymous, nonsynonymous, and intergenic substitutions which may be contributing to the incompatibility of Evo-Φ36's J gene in the context of ΦX174's genome (**G**).

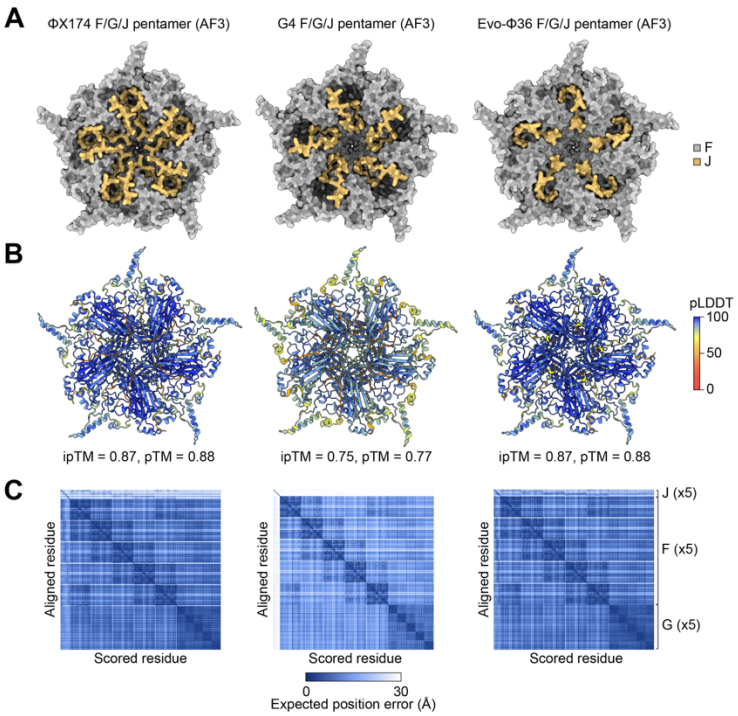

Fig. S21. Structure predictions of J protein in capsid pentamers. (**A**) AlphaFold 3 (AF3) predictions of capsid (F, gray), spike (G, not visible), and J protein (yellow) pentamers of ΦX174, G4, and Evo-Φ36. The interior face of the pentamers is shown, highlighting the notable differences in J protein interactions with capsid proteins. (**B**) AF3 predictions of F/G/J pentamers colored by predicted local distance difference test (pLDDT) score. ipTM, interface predicted template modeling score; pTM, predicted template modeling score. (**C**) Predicted aligned error (PAE) plots of the AF3 predictions.

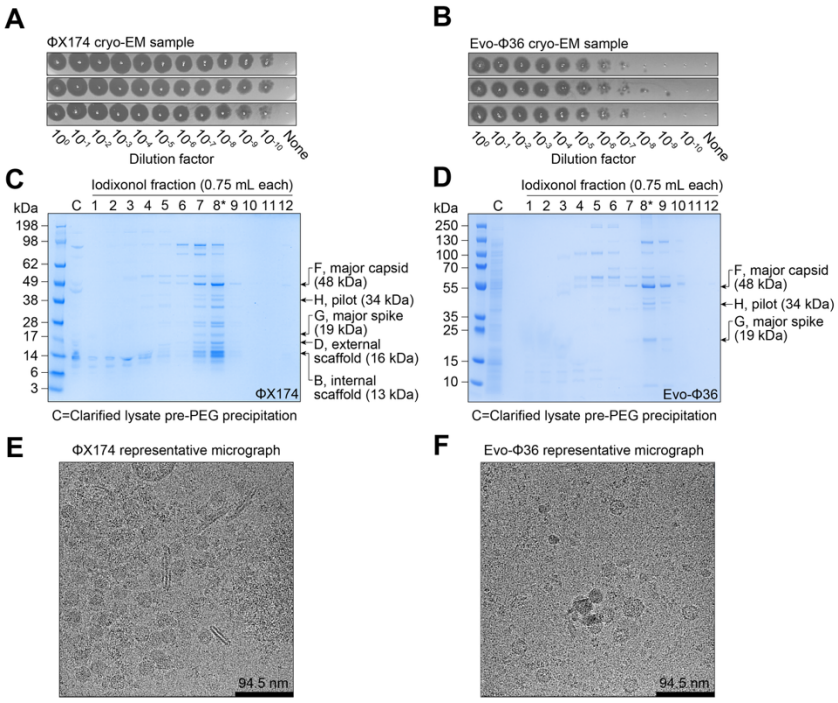

Fig. S22. Supplementary data for cryo-EM analysis. (**A–B**) Titrations of purified ΦX174 (**A**) and Evo-Φ36 (**B**) cryo-EM samples. (**C–D**) SDS-PAGE of ΦX174 (**C**) and Evo-Φ36 (**D**) iodixonol gradient fractions. *, fraction used for cryo-EM. (**E–F**) Representative cryo-EM micrographs of ΦX174 (**E**) and Evo-Φ36 (**F**) grids used for cryo-EM data collection.

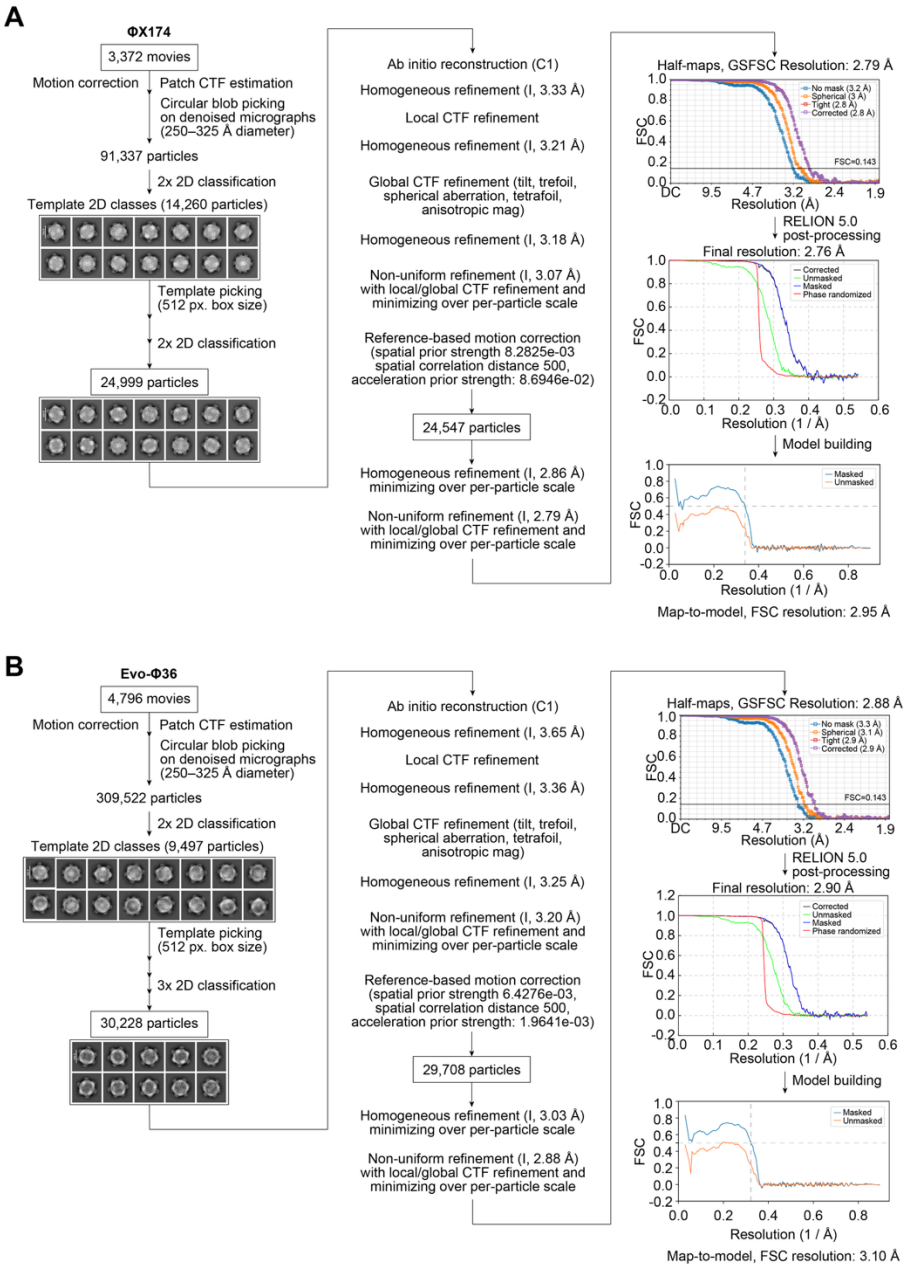

Fig. S23. Cryo-EM data processing workflow. (**A–B**) Cryo-EM data processing workflows in cryoSPARC for ΦX174 (**A**) and Evo-Φ36 (**B**) with half-map FSCs as reported by cryoSPARC without auto-tightening of the mask, FSCs from RELION with threshold-based mask, and map-to-model FSCs from Phenix with a threshold of FSC = 0.5. Masks for RELION post-processing were constructed using a low-pass filter of 20 Å, 1 pixel extension, and 8 pixel soft-edge with a binarization threshold of 0.015 and 0.01 for ΦX174 and Evo-Φ36, respectively.

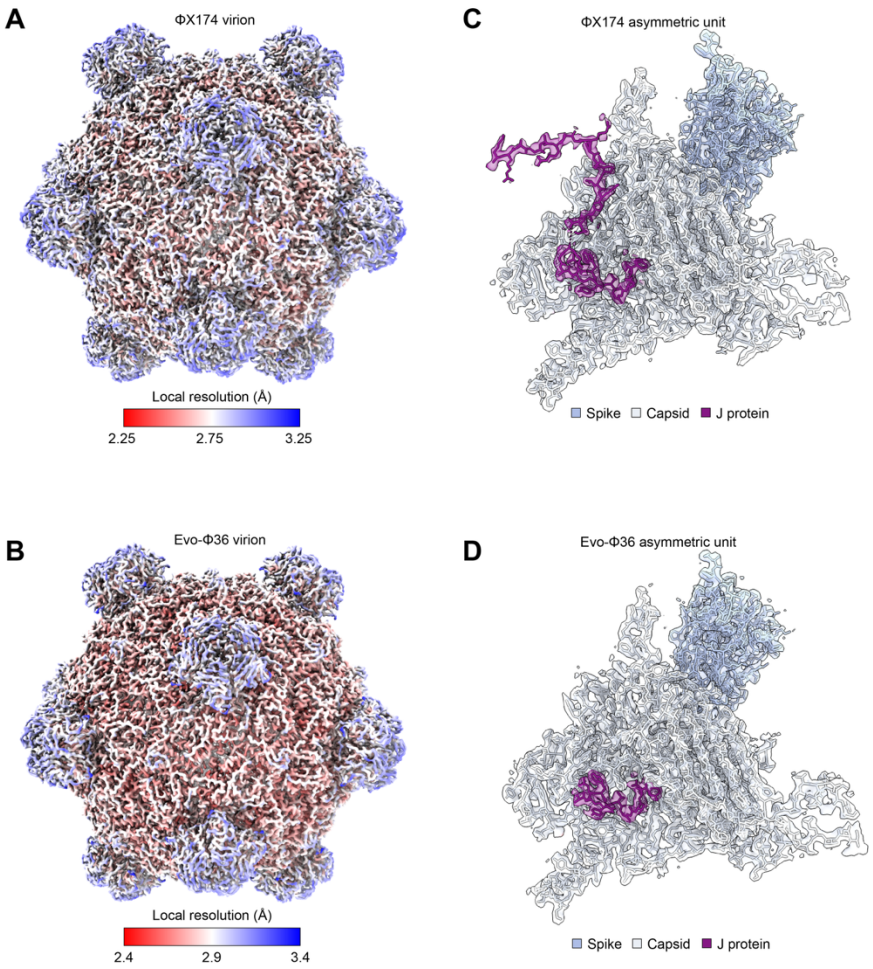

Fig. S24. Additional validation of cryo-EM maps and atomic models. (**A–B**) Local resolution estimation for ΦX174 (**A**) and Evo-Φ36 (**B**) as output by cryoSPARC (FSC = 0.5). (**C–D**) Visualization of atomic models fit into cryo-EM densities for an asymmetric unit of ΦX174 (**C**) and Evo-Φ36 (**D**).

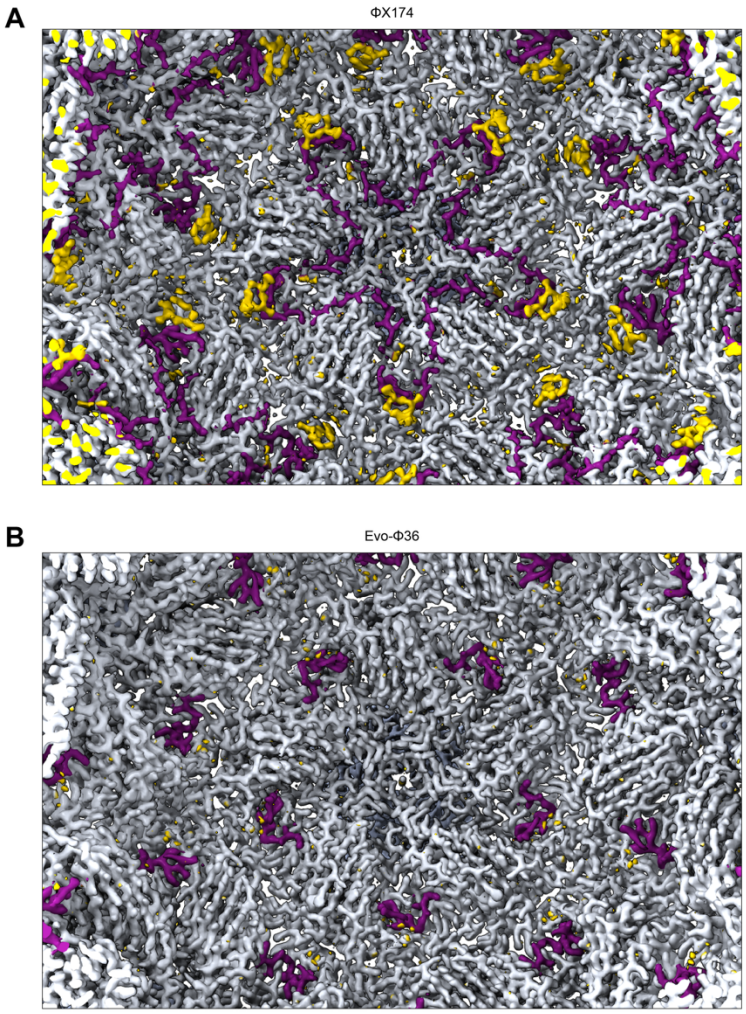

Fig. S25. Interior surface views of cryo-EM density maps. (**A–B**) Interior surface view of the cryo-EM density maps of ΦX174 (**A**) and Evo-Φ36 (**B**), colored by zone based on the model with capsid (F) in gray, spike (G) in light blue (largely not visible), and DNA packaging protein (J) in purple. Compare to **Figure 4J**. Gold corresponds to unmodeled densities, primarily of nucleic acids.

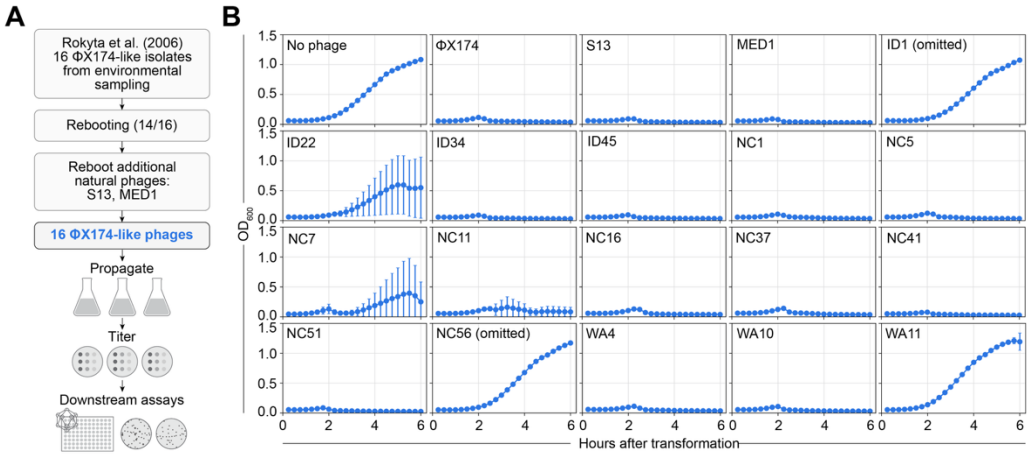

Fig. S26. Supplementary data for natural ΦX174-like bacteriophage rebooting. (**A–B**) We successfully rebooted 14 out of 16 ΦX174-like isolates in *E. coli* C (*61*) and additional ΦX174-like phages S13 and MED1. ID1 and NC56 were omitted from downstream assays due to failed rebooting. Data point, mean OD_{600} value; error bar, standard deviation; $N = 3$ growth replicates.

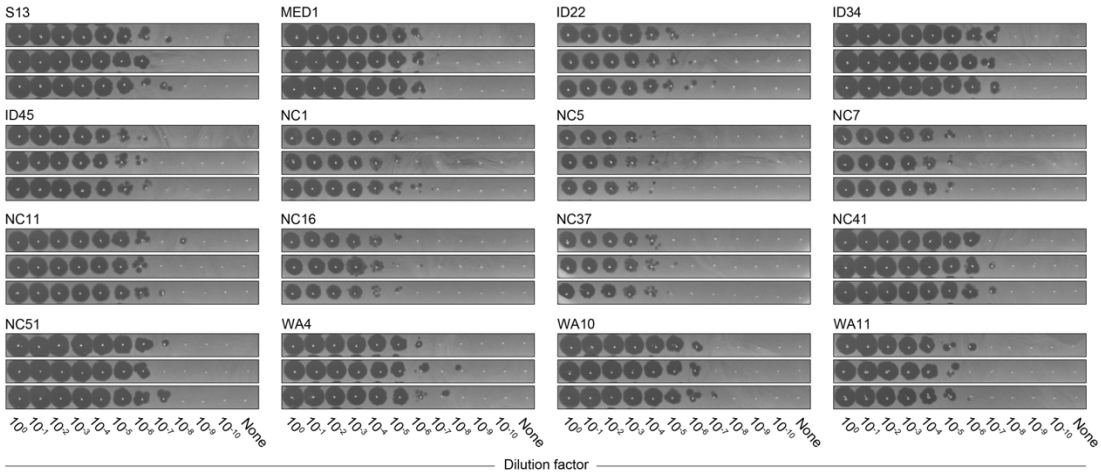

Fig. S27. Titration of natural ΦX174-like bacteriophages. Plaque titrations of purified natural ΦX174-like phages on *E. coli* C. Each row is a titration replicate.

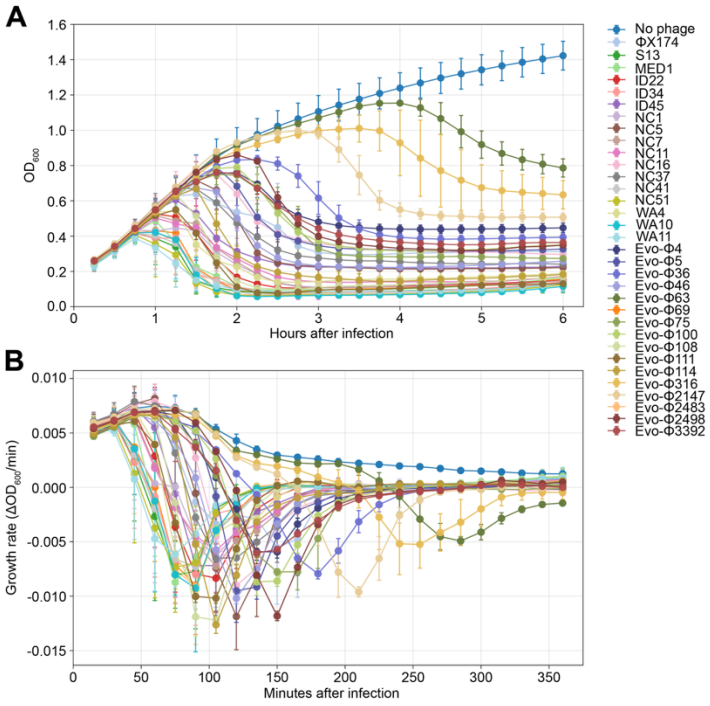

Fig. S28. Supplementary data for lytic assay. Growth rates of *E. coli* C infected by generated phages, natural ΦX174-like phages, and ΦX174, with no phage infection control. Data point, mean OD_{600} value; error bar, standard deviation; N = 3–12 infection replicates.

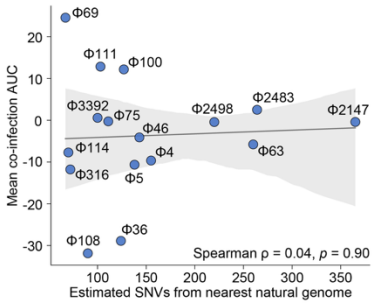

Fig. S29. Supplementary data for co-infection assay. Co-infection area under the curve (AUC) of generated phages is not correlated with their divergence from nearest natural genomes. Data point, mean AUC; solid line, regression line; shading, 95% confidence interval. Statistical significance was determined by Spearman rank correlation test ($\rho = 0.04$, $p = 0.90$).

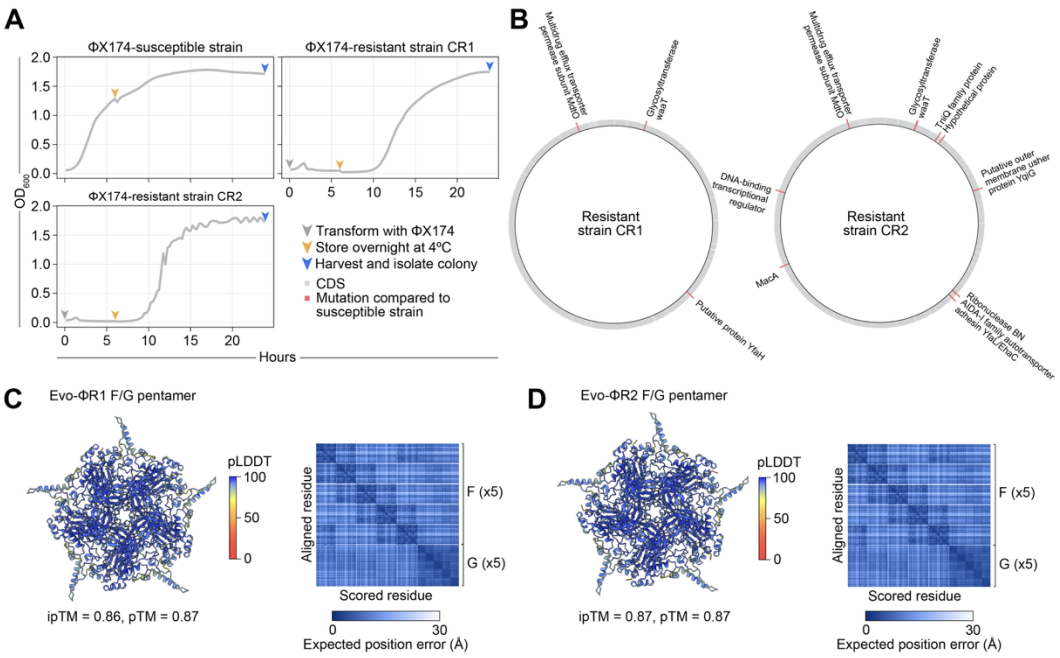

Fig. S30. Supplementary data for bacterial resistance assay. (**A**) Growth curves of *E. coli* C without infection by ΦX174 (ΦX174-susceptible strain) or transformed with ΦX174 genome and cultured over 24 hours to produce resistant populations (ΦX174-resistant strains). Arrowheads indicate key timepoints for producing and harvesting the different strains. (**B**) Whole-genome sequencing of resistant strains CR1 and CR2 show unique mutations compared to the susceptible strain. Gray, predicted coding sequence (CDS); red, point mutation relative to susceptible strain. (**C–D**) AlphaFold 3 (AF3) predictions of capsid (F) and spike (G) pentamers of Evo-ΦR1 and Evo-ΦR2. The exterior face of the pentamers is shown, colored by predicted local distance difference test (pLDDT) score. Predicted aligned error (PAE) plots of the AF3 predictions are shown on the right. ipTM, interface predicted template modeling score; pTM, predicted template modeling score.

Table S1

Cryo-EM data collection, refinement, and validation statistics

	#1 ΦX174 (PDB 36CQ) (EMDB 77390)	#2 Evo-Φ36 (PDB 36CR) (EMDB 77391)
Data collection and Processing		
Microscope	Thermo Scientific Glacios Cryo-TEM	
Voltage (keV)	200	
Camera	Falcon 4i	
Magnification	150000x	
Pixel size at detector (Å/pixel)	0.923	
Total electron exposure (e^{-}/Å^2)	41.15	50.00
Exposure rate (e-/pixel/sec)	11.69	13.19
Number of frames collected during exposure	40	43
Defocus range (μm)	-1.5 to -3.0	-1.5 to -2.5
Automation software	EPU	
Micrographs collected (no.)	3372	4796
Total extracted particles (no.)	24999	30228
Refined final particles (no.)	24547	29708
Symmetry-group	I	I
Resolution (global, Å)		
FSC 0.143 (unmasked/masked)	3.2 / 2.76	3.3 / 2.90
Map sharpening *B* factor (Å^2)	-73.5	-87.5
Map sharpening method	relion_postprocess	relion_postprocess
Model composition		
Atoms (non-hydrogen)	4996	4849
Protein residues	632	612
Ligands	0	0
RNA/DNA bases	0	0
Model Refinement		
Refinement package	phenix.real_space_refine	phenix.real_space_refine
- real or reciprocal space	Real space	Real space
- resolution cutoff	2.8	2.9
Model-Map scores		
- CC	0.88	0.85
- Map-model FSC=0.5	2.95 Å	3.10 Å
B factors (mean, Å^2)		
Protein residues	40.31	29.1
Ligands	0	0
RNA/DNA	0	0
R.m.s. deviations from ideal values		
Bond lengths (Å)	0.009	0.007
Bond angles (°)	1.155	1.108
Validation		
MolProbity score	0.70	0.95
CaBLAM outliers (%)	0.49	0.83
Clashscore	0.61	0.52
Rotamer outliers (%)	0.18	0.38
C-beta deviations (%)	0.00	0.00
Ramachandran plot		
Favored (%)	98.10	96.04
Outliers (%)	0.00	0.00

Legend for Data S1

Sheet 1. Generated bacteriophage genome candidates, metrics, assembly fragments, and viability.
Sheet 2. DNA oligo sequences.
Sheet 3. Bacteriophage genome amplicons for long-read sequencing.
Sheet 4. Viable sequencing-verified generated bacteriophage genomes and metrics.
Sheet 5. Top 10 nucleotide BLAST hits and mutation metrics for viable generated, nonviable generated, and natural bacteriophage genomes.
Sheet 6. Raw growth rate data for generated and natural bacteriophages.
Sheet 7. Lytic kinetics analysis data for generated and natural bacteriophages.
Sheet 8. Co-infection AUC of generated bacteriophages and ΦX174.
Sheet 9. Genome sequences of resistant generated bacteriophages.

References and Notes:

1. R. C. Edgar, B. Taylor, V. Lin, T. Altman, P. Barbera, D. Meleshko, D. Lohr, G. Novakovsky, B. Buchfink, B. Al-Shayeb, J. F. Banfield, M. de la Peña, A. Korobeynikov, R. Chikhi, A. Babaian, Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022). doi:10.1038/s41586-021-04332-2 Medline

2. D. G. Gibson, G. A. Benders, C. Andrews-Pfannkoch, E. A. Denisova, H. Baden-Tillson, J. Zaveri, T. B. Stockwell, A. Brownley, D. W. Thomas, M. A. Algire, C. Merryman, L. Young, V. N. Noskov, J. I. Glass, J. C. Venter, C. A. Hutchison 3rd, H. O. Smith, Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* **319**, 1215–1220 (2008). doi:10.1126/science.1151721 Medline

3. C. A. Hutchison 3rd, R.-Y. Chuang, V. N. Noskov, N. Assad-Garcia, T. J. Deerinck, M. H. Ellisman, J. Gill, K. Kannan, B. J. Karas, L. Ma, J. F. Pelletier, Z.-Q. Qi, R. A. Richter, E. A. Strychalski, L. Sun, Y. Suzuki, B. Tsvetanova, K. S. Wise, H. O. Smith, J. I. Glass, C. Merryman, D. G. Gibson, J. C. Venter, Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016). doi:10.1126/science.aad6253 Medline

4. S. M. Richardson, L. A. Mitchell, G. Stracquadanio, K. Yang, J. S. Dymond, J. E. DiCarlo, D. Lee, C. L. V. Huang, S. Chandrasegaran, Y. Cai, J. D. Boeke, J. S. Bader, Design of a synthetic yeast genome. *Science* **355**, 1040–1044 (2017). doi:10.1126/science.aaf4557 Medline

5. K. Wang, J. Fredens, S. F. Brunner, S. H. Kim, T. Chia, J. W. Chin, Defining synonymous codon compression schemes by genome recoding. *Nature* **539**, 59–64 (2016). doi:10.1038/nature20124 Medline

6. R. M. Dedrick, C. A. Guerrero-Bustamante, R. A. Garlena, D. A. Russell, K. Ford, K. Harris, K. C. Gilmour, J. Soothill, D. Jacobs-Sera, R. T. Schooley, G. F. Hatfull, H. Spencer, Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. *Nat. Med.* **25**, 730–733 (2019). doi:10.1038/s41591-019-0437-z Medline

7. M. G. Durrant, N. T. Perry, J. J. Pai, A. R. Jangid, J. S. Athukoralage, M. Hiraizumi, J. P. McSpedon, A. Pawluk, H. Nishimasu, S. Konermann, P. D. Hsu, Bridge RNAs direct programmable recombination of target and donor DNA. *Nature* **630**, 984–993 (2024). doi:10.1038/s41586-024-07552-4 Medline

8. D. J. Mandell, M. J. Lajoie, M. T. Mee, R. Takeuchi, G. Kuznetsov, J. E. Norville, C. J. Gregg, B. L. Stoddard, G. M. Church, Biocontainment of genetically modified organisms by synthetic protein design. *Nature* **518**, 55–60 (2015). doi:10.1038/nature14121 Medline

9. A. Nyerges, S. Vinke, R. Flynn, S. V. Owen, E. A. Rand, B. Budnik, E. Keen, K. Narasimhan, J. A. Marchand, M. Baas-Thomas, M. Liu, K. Chen, A. Chiappino-Pepe, F. Hu, M. Baym, G. M. Church, A swapped genetic code prevents viral infections and gene transfer. *Nature* **615**, 720–727 (2023). doi:10.1038/s41586-023-05824-z Medline

10. M. Costanzo, B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S. D. Lee, V. Pelechano, E. B. Styles, M. Billmann, J. van Leeuwen, N. van Dyk, Z.-Y. Lin, E. Kuzmin, J. Nelson, J. S. Piotrowski, T. Srikumar, S. Bahr, Y. Chen, R. Deshpande, C. F. Kurat, S. C. Li, Z. Li, M. M. Usaj, H. Okada, N. Pascoe, B.-J.

San Luis, S. Sharifpoor, E. Shuteriqi, S. W. Simpkins, J. Snider, H. G. Suresh, Y. Tan, H. Zhu, N. Malod-Dognin, V. Janjic, N. Przulj, O. G. Troyanskaya, I. Stagljar, T. Xia, Y. Ohya, A.-C. Gingras, B. Raught, M. Boutros, L. M. Steinmetz, C. L. Moore, A. P. Rosebrock, A. A. Caudy, C. L. Myers, B. Andrews, C. Boone, A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016). doi:10.1126/science.aaf1420 Medline

11. S. F. Elena, R. E. Lenski, Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390**, 395–398 (1997). doi:10.1038/37108 Medline

12. F. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961). doi:10.1016/S0022-2836(61)80072-7 Medline

13. B. G. Barrell, G. M. Air, C. A. Hutchison 3rd, Overlapping genes in bacteriophage phiX174. *Nature* **264**, 34–41 (1976). doi:10.1038/264034a0 Medline

14. C. A. Hutchison III, S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, J. C. Venter, Global transposon mutagenesis and a minimal Mycoplasma genome. *Science* **286**, 2165–2169 (1999). doi:10.1126/science.286.5447.2165 Medline

15. R. Sanjuán, A. Moya, S. F. Elena, The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8396–8401 (2004). doi:10.1073/pnas.0400146101 Medline

16. P. Domingo-Calap, J. M. Cuevas, R. Sanjuán, The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. *PLOS Genet.* **5**, e1000742 (2009). doi:10.1371/journal.pgen.1000742 Medline

17. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022). doi:10.1126/science.add2187 Medline

18. J. Fredens, K. Wang, D. de la Torre, L. F. H. Funke, W. E. Robertson, Y. Christova, T. Chia, W. H. Schmied, D. L. Dunkelmann, V. Beránek, C. Uttamapinant, A. G. Llamazares, T. S. Elliott, J. W. Chin, Total synthesis of Escherichia coli with a recoded genome. *Nature* **569**, 514–518 (2019). doi:10.1038/s41586-019-1192-5 Medline

19. T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. A. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, A. Rives, Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025). doi:10.1126/science.ads0018 Medline

20. B. L. Hie, V. R. Shanker, D. Xu, T. U. J. Bruun, P. A. Weidenbacher, S. Tang, W. Wu, J. E. Pak, P. S. Kim, Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **42**, 275–283 (2024). doi:10.1038/s41587-023-01763-2 Medline

21. J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, S. Tie, V. Xue, S. C. Cowles, A. Leung, J. V. Rodrigues, C. L. Morales-Perez, A. M. Ayoub, R. Green, K. Puentes, F. Oplinger, N. V. Panwar, F. Obermeyer, A. R. Root, A. L. Beam, F. J. Poelwijk, G. Grigoryan,

Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023). doi:10.1038/s41586-023-06728-8 Medline

22. K. Jiang, Z. Yan, M. Di Bernardo, S. R. Sgrizzi, L. Villiger, A. Kayabolen, B. J. Kim, J. K. Carscadden, M. Hiraizumi, H. Nishimasu, J. S. Gootenberg, O. O. Abudayyeh, Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science* **387**, eadr6006 (2025). doi:10.1126/science.adr6006 Medline

23. P. Srinivasan, C. D. Smolke, Biosynthesis of medicinal tropane alkaloids in yeast. *Nature* **585**, 614–619 (2020). doi:10.1038/s41586-020-2650-9 Medline

24. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models. arXiv:2001.08361 [cs.LG] (2020).

25. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need" in *Advances in Neural Information Processing Systems* (Curran Associates Inc., 2017), pp. 6000–6010.

26. G. Brixi, M. G. Durrant, J. Ku, M. Naghipourfar, M. Poli, G. Sun, G. Brockman, D. Chang, A. Fanton, G. A. Gonzalez, S. H. King, D. B. Li, A. T. Merchant, E. Nguyen, C. Ricci-Tam, D. W. Romero, J. C. Schmok, A. Taghibakhshi, A. Vorontsov, B. Yang, M. Deng, L. Gorton, N. Nguyen, N. K. Wang, M. T. Pearce, E. Simon, E. Adams, Z. J. Amador, E. A. Ashley, S. A. Baccus, H. Dai, S. Dillmann, S. Ermon, D. Guo, M. H. Herschl, R. Ilango, K. Janik, A. X. Lu, R. Mehta, M. R. K. Mofrad, M. Y. Ng, J. Pannu, C. Ré, J. St John, J. Sullivan, J. Tey, B. Viggiano, K. Zhu, G. Zynda, D. Balsam, P. Collison, A. B. Costa, T. Hernandez-Boussard, E. Ho, M. Y. Liu, T. McGrath, K. Powell, S. Pinglay, D. P. Burke, H. Goodarzi, P. D. Hsu, B. L. Hie, Genome modelling and design across all domains of life with Evo 2. *Nature* **652**, 1349–1361 (2026). doi:10.1038/s41586-026-10176-5 Medline

27. A. T. Merchant, S. H. King, E. Nguyen, B. L. Hie, Semantic design of functional de novo genes from a genomic language model. *Nature* **649**, 749–758 (2026). doi:10.1038/s41586-025-09749-7 Medline

28. E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, J. Sullivan, M. Y. Ng, A. Lewis, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu, B. L. Hie, Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024). doi:10.1126/science.ado9336 Medline

29. M. K. Kim, G. A. Suh, G. D. Cullen, S. Perez Rodriguez, T. Dharmaraj, T. H. W. Chang, Z. Li, Q. Chen, S. I. Green, R. Lavigne, J.-P. Pirnay, P. L. Bollyky, J. C. Sacher, Bacteriophage therapy for multidrug-resistant infections: Current technologies and therapeutic approaches. *J. Clin. Invest.* **135**, e187996 (2025). doi:10.1172/JCI187996 Medline

30. S. Kilcher, M. J. Loessner, Engineering bacteriophages as versatile biologics. *Trends Microbiol.* **27**, 355–367 (2019). doi:10.1016/j.tim.2018.09.006 Medline

31. D. P. Pires, S. Cleto, S. Sillankorva, J. Azeredo, T. K. Lu, Genetically engineered phages: A review of advances over the last decade. *Microbiol. Mol. Biol. Rev.* **80**, 523–543 (2016). doi:10.1128/MMBR.00069-15 Medline

32. B. Shao, J. Yan, A long-context language model for deciphering and generating bacteriophage genomes. *Nat. Commun.* **15**, 9392 (2024). doi:10.1038/s41467-024-53759-4 Medline

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

34. P. R. Jaschke, G. A. Dotson, K. S. Hung, D. Liu, D. Endy, Definitive demonstration by synthesis of genome annotation completeness. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 24206–24213 (2019). doi:10.1073/pnas.1905990116 Medline

35. D. Y. Logel, P. R. Jaschke, A high-resolution map of bacteriophage ϕX174 transcription. *Virology* **547**, 47–56 (2020). doi:10.1016/j.virol.2020.05.008 Medline

36. F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison III, P. M. Slocombe, M. Smith, Nucleotide sequence of bacteriophage φ X174 DNA. *Nature* **265**, 687–695 (1977). doi:10.1038/265687a0 Medline

37. J. Shlomai, A. Kornberg, An Escherichia coli replication protein that recognizes a unique sequence within a hairpin region in phi X174 DNA. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 799–803 (1980). doi:10.1073/pnas.77.2.799 Medline

38. P. C. Kirchberger, H. Ochman, Microviruses: A World Beyond *phi*X174. *Annu. Rev. Virol.* **10**, 99–118 (2023). doi:10.1146/annurev-virology-100120-011239 Medline

39. H. O. Smith, C. A. Hutchison 3rd, C. Pfannkoch, J. C. Venter, Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15440–15445 (2003). doi:10.1073/pnas.2237126100 Medline

40. M. Goulian, A. Kornberg, R. L. Sinsheimer, Enzymatic synthesis of DNA, XXIV. Synthesis of infectious phage phi-X174 DNA. *Proc. Natl. Acad. Sci. U.S.A.* **58**, 2321–2328 (1967). doi:10.1073/pnas.58.6.2321 Medline

41. P. R. Jaschke, E. K. Lieberman, J. Rodriguez, A. Sierra, D. Endy, A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. *Virology* **434**, 278–284 (2012). doi:10.1016/j.virol.2012.09.020 Medline

42. A. P. Camargo, S. Roux, F. Schulz, M. Babinski, Y. Xu, B. Hu, P. S. G. Chain, S. Nayfach, N. C. Kyrpides, Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024). doi:10.1038/s41587-023-01953-y Medline

43. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). doi:10.1186/1471-2105-10-421 Medline

44. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives,

Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). doi:10.1126/science.ade2574 Medline

45. P. Terzian, E. Olo Ndela, C. Galiez, J. Lossouarn, R. E. Pérez Bucio, R. Mom, A. Toussaint, M.-A. Petit, F. Enault, PHROG: Families of prokaryotic virus proteins clustered using remote homology. *NAR Genom. Bioinform.* **3**, lqab067 (2021). doi:10.1093/nargab/lqab067 Medline

46. J. R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource. *Nucleic Acids Res.* **43** (D1), D571–D577 (2015). doi:10.1093/nar/gku1207 Medline

47. A. Michel, O. Clermont, E. Denamur, O. Tenaillon, Bacteriophage PhiX174's ecological niche and the flexibility of its Escherichia coli lipopolysaccharide receptor. *Appl. Environ. Microbiol.* **76**, 7310–7313 (2010). doi:10.1128/AEM.02721-09 Medline

48. B. W. Wright, M. P. Molloy, P. R. Jaschke, Overlapping genes in natural and engineered genomes. *Nat. Rev. Genet.* **23**, 154–168 (2022). doi:10.1038/s41576-021-00417-w Medline

49. Y. Sun, A. P. Roznowski, J. M. Tokuda, T. Klose, A. Mauney, L. Pollack, B. A. Fane, M. G. Rossmann, Structural changes of tailless bacteriophage ΦX174 during penetration of bacterial cell walls. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 13708–13713 (2017). doi:10.1073/pnas.1716614114 Medline

50. R. K. Peet, The measurement of species diversity. *Annu. Rev. Ecol. Syst.* **5**, 285–307 (1974). doi:10.1146/annurev.es.05.110174.001441

51. S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, N. C. Kyrpides, CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021). doi:10.1038/s41587-020-00774-7 Medline

52. M. S. Faber, J. T. Van Leuven, M. M. Ederer, Y. Sapozhnikov, Z. L. Wilson, H. A. Wichman, T. A. Whitehead, C. R. Miller, Saturation Mutagenesis Genome Engineering of Infective ΦX174 Bacteriophage *via* Unamplified Oligo Pools and Golden Gate Assembly. *ACS Synth. Biol.* **9**, 125–131 (2020). doi:10.1021/acssynbio.9b00411 Medline

53. J. T. Van Leuven, M. M. Ederer, K. Burleigh, L. Scott, R. A. Hughes, V. Codrea, A. D. Ellington, H. A. Wichman, C. R. Miller, ΦX174 attenuation by whole-genome codon deoptimization. *Genome Biol. Evol.* **13**, evaa214 (2021). doi:10.1093/gbe/evaa214 Medline

54. W. Fiers, R. L. Sinsheimer, The structure of the DNA of bacteriophage φ-X174. III. Ultracentrifugal evidence for a ring structure. *J. Mol. Biol.* **5**, 424–434 (1962). doi:10.1016/S0022-2836(62)80031-X Medline

55. S. Wickner, J. Hurwitz, Conversion of phiX174 viral DNA to double-stranded form by purified Escherichia coli proteins. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4120–4124 (1974). doi:10.1073/pnas.71.10.4120 Medline

56. C. T. Archer, J. F. Kim, H. Jeong, J. H. Park, C. E. Vickers, S. Y. Lee, L. K. Nielsen, The genome sequence of E. coli W (ATCC 9637): Comparative genome analysis and an improved genome-scale reconstruction of E. coli. *BMC Genomics* **12**, 9 (2011). doi:10.1186/1471-2164-12-9 Medline

57. B. A. Fane, S. Head, M. Hayashi, Functional relationship between the J proteins of bacteriophages phi X174 and G4 during phage morphogenesis. *J. Bacteriol.* **174**, 2717–2719 (1992). doi:10.1128/jb.174.8.2717-2719.1992 Medline

58. E. T. Ogunbunmi, A. P. Roznowski, B. A. Fane, The effects of packaged, but misguided, single-stranded DNA genomes are transmitted to the outer surface of the φX174 capsid. *J. Virol.* **95**, e0088321 (2021). doi:10.1128/JVI.00883-21 Medline

59. A. P. Roznowski, S. M. Doore, S. Z. Kemp, B. A. Fane, Finally, a role befitting Astar: Strongly conserved, unessential microvirus A* proteins ensure the product fidelity of packaging reactions. *J. Virol.* **94**, 10–1128 (2020). doi:10.1128/JVI.01593-19 Medline

60. H. A. Wichman, J. Millstein, J. J. Bull, Adaptive molecular evolution for 13,000 phage generations: A possible arms race. *Genetics* **170**, 19–31 (2005). doi:10.1534/genetics.104.034488 Medline

61. D. R. Rokyta, C. L. Burch, S. B. Caudle, H. A. Wichman, Horizontal gene transfer and the evolution of microvirid coliphage genomes. *J. Bacteriol.* **188**, 1134–1142 (2006). doi:10.1128/JB.188.3.1134-1142.2006 Medline

62. D. Turner, A. M. Kropinski, E. M. Adriaenssens, A roadmap for genome-based phage taxonomy. *Viruses* **13**, 506 (2021). doi:10.3390/v13030506 Medline

63. R. A. Bernal, S. Hafenstein, R. Esmeralda, B. A. Fane, M. G. Rossmann, The phiX174 protein J mediates DNA packaging and viral attachment to host cells. *J. Mol. Biol.* **337**, 1109–1122 (2004). doi:10.1016/j.jmb.2004.02.033 Medline

64. G. N. Godson, B. G. Barrell, R. Staden, J. C. Fiddes, Nucleotide sequence of bacteriophage G4 DNA. *Nature* **276**, 236–247 (1978). doi:10.1038/276236a0 Medline

65. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024). doi:10.1038/s41586-024-07487-w Medline

66. R. McKenna, D. Xia, P. Willingmann, L. L. Ilag, S. Krishnaswamy, M. G. Rossmann, N. H. Olson, T. S. Baker, N. L. Incardona, Atomic structure of single-stranded DNA bacteriophage phi X174 and its functional implications. *Nature* **355**, 137–143 (1992). doi:10.1038/355137a0 Medline

67. W. Hu, Z. Liu, Y. Wei, Q. Bian, W. Lan, C. Fan, J. Song, Q. Sun, X. Zhang, Y. Liu, Y. Gao, Y. Chen, Structural basis for Salmonella infection by two Microviridae phages. *Commun. Biol.* **8**, 1166 (2025). doi:10.1038/s42003-025-08595-7 Medline

68. J. P. Pirnay, Phage therapy in the year 2035. *Front. Microbiol.* **11**, 1171 (2020). doi:10.3389/fmicb.2020.01171 Medline

69. J. Romeyer Dherbey, L. Parab, J. Gallie, F. Bertels, Stepwise evolution of E. coli C and ΦX174 reveals unexpected lipopolysaccharide (LPS) diversity. *Mol. Biol. Evol.* **40**, 154 (2023). doi:10.1093/molbev/msad154 Medline

70. A. L. V. Coradini, C. B. Hull, I. M. Ehrenreich, Building genomes to understand biology. *Nat. Commun.* **11**, 6177 (2020). doi:10.1038/s41467-020-19753-2 Medline

71. J. S. James, J. Dai, W. L. Chew, Y. Cai, The design and engineering of synthetic genomes. *Nat. Rev. Genet.* **26**, 298–319 (2025). doi:10.1038/s41576-024-00786-y Medline

72. A. Aoyama, M. Hayashi, Effects of genome size on bacteriophage phi X174 DNA packaging in vitro. *J. Biol. Chem.* **260**, 11033–11038 (1985). doi:10.1016/S0021-9258(17)39144-5 Medline

73. D. Endy, L. You, J. Yin, I. J. Molineux, Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5375–5380 (2000). doi:10.1073/pnas.090101397 Medline

74. P. W. Russell, U. R. Müller, Construction of bacteriophage luminal diameterX174 mutants with maximum genome sizes. *J. Virol.* **52**, 822–827 (1984). doi:10.1128/jvi.52.3.822-827.1984 Medline

75. M. Wang, Z. Zhang, A. S. Bedi, A. Velasquez, S. Guerra, S. Lin-Gibson, L. Cong, Y. Qu, S. Chakraborty, M. Blewett, J. Ma, E. Xing, G. Church, A call for built-in biosecurity safeguards for generative AI tools. *Nat. Biotechnol.* **43**, 845–847 (2025). doi:10.1038/s41587-025-02650-8 Medline

76. D. Bloomfield, J. Pannu, A. W. Zhu, M. Y. Ng, A. Lewis, E. Bendavid, S. M. Asch, T. Hernandez-Boussard, A. Cicero, T. Inglesby, AI and biosecurity: The need for governance. *Science* **385**, 831–833 (2024). doi:10.1126/science.adq1977 Medline

77. P. Berg, D. Baltimore, S. Brenner, R. O. Roblin 3rd, M. F. Singer, Asilomar conference on recombinant DNA molecules. *Science* **188**, 991–994 (1975). doi:10.1126/science.1056638 Medline

78. D. Bloomfield, M. S. Hanke, A. Maiwald, J. R. M. Black, T. Webster, T. Hernandez-Boussard, A. Berke, O. M. Crook, J. Pannu, Securing dual-use pathogen data of concern. arXiv:2602.08061 [cs.AI] (2026).

79. B. D. Trump, S. E. Galaitsi, E. Appleton, D. A. Bleijs, M. V. Florin, J. D. Gollihar, R. A. Hamilton, T. Kuiken, F. Lentzos, R. Mampuys, M. Merad, T. Novossiolova, K. Oye, E. Perkins, N. Garcia-Reyero, C. Rhodes, I. Linkov, Building biosecurity for synthetic biology. *Mol. Syst. Biol.* **16**, e9723 (2020). doi:10.15252/msb.20209723 Medline

80. B. J. Wittmann, T. Alexanian, C. Bartling, J. Beal, A. Clore, J. Diggans, K. Flyangolts, B. T. Gemler, T. Mitchell, S. T. Murphy, N. E. Wheeler, E. Horvitz, Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science* **390**, 82–87 (2025). doi:10.1126/science.adu8578 Medline

81. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, S. Lu, W. Chen, LoRA: Low-rank adaptation of large language models. *Int. Conf. Learn. Represent.* (2022).

82. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, “Retrieval-augmented generation for

knowledge-intensive NLP tasks" in *Advances in Neural Information Processing Systems* (Curran Associates Inc., 2020), pp. 9459–9474.

83. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015). doi:10.1038/nature14236 Medline

84. T. Widatalla, R. Rafailov, B. Hie, Aligning protein generative models with experimental fitness via direct preference optimization. bioRxiv 2024.05.20.595026 [Preprint] (2024). https://doi.org/10.1101/2024.05.20.595026.

85. A. Levrier, I. Karpathakis, B. Nash, S. D. Bowden, A. B. Lindner, V. Noireaux, PHEIGES: All-cell-free phage synthesis and selection from engineered genomes. *Nat. Commun.* **15**, 2223 (2024). doi:10.1038/s41467-024-46585-1 Medline

86. J. M. Pryor, V. Potapov, K. Bilotti, N. Pokhrel, G. J. S. Lohman, Rapid 40 kb genome construction from 52 parts through data-optimized assembly design. *ACS Synth. Biol.* **11**, 2036–2042 (2022). doi:10.1021/acssynbio.1c00525 Medline

87. E. Shaer Tamar, R. Kishony, Multistep diversification in spatiotemporal bacterial-phage coevolution. *Nat. Commun.* **13**, 7971 (2022). doi:10.1038/s41467-022-35351-w Medline

88. J. Zheng, J. L. Payne, A. Wagner, Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science* **365**, 347–353 (2019). doi:10.1126/science.aax1837 Medline

89. V. K. Mutalik, A. P. Arkin, A phage foundry framework to systematically develop viral countermeasures to combat antibiotic-resistant bacterial pathogens. *iScience* **25**, 104121 (2022). doi:10.1016/j.isci.2022.104121 Medline

90. E. Wimmer, S. Mueller, T. M. Tumpey, J. K. Taubenberger, Synthetic viruses: A new opportunity to understand and prevent viral disease. *Nat. Biotechnol.* **27**, 1163–1172 (2009). doi:10.1038/nbt.1593 Medline

91. S. H. King, Training data for paper "Generative design of bacteriophages with genome language models," Zenodo (2025); https://doi.org/10.5281/zenodo.17101843.

92. S. H. King, Code for paper "Generative design of bacteriophages with genome language models," Zenodo (2026); https://doi.org/10.5281/zenodo.20484521.

93. T. A. Chang, B. K. Bergen, Language model behavior: A comprehensive survey. *Comput. Linguist.* **50**, 293–350 (2024). doi:10.1162/coli_a_00492

94. E. Nguyen, M. Poli, M. Faizi, A. W. Thomas, C. B. Sykes, M. Wornow, A. Patel, C. Rabideau, S. Massaroli, Y. Bengio, S. Ermon, S. A. Baccus, C. Ré, "HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution" in *Advances in Neural Information Processing Systems* (Curran Associates Inc., 2023), pp. 43177–43201.

95. D. Hyatt, G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010). doi:10.1186/1471-2105-11-119 Medline

96. A. A. Egorov, G. C. Atkinson, LoVis4u: A locus visualization tool for comparative genomics and coverage profiles. *NAR Genom. Bioinform.* **7**, lqaf009 (2025). doi:10.1093/nargab/lqaf009 Medline

97. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017). doi:10.1038/nbt.3988 Medline

98. N. A. O'Leary, E. Cox, J. B. Holmes, W. R. Anderson, R. Falk, V. Hem, M. T. N. Tsuchiya, G. D. Schuler, X. Zhang, J. Torcivia, A. Ketter, L. Breen, J. Cothran, H. Bajwa, J. Tinne, P. A. Meric, W. Hlavina, V. A. Schneider, Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **11**, 732 (2024). doi:10.1038/s41597-024-03571-y Medline

99. R. H. Wang, S. Yang, Z. Liu, Y. Zhang, X. Wang, Z. Xu, J. Wang, S. C. Li, PhageScope: A well-annotated bacteriophage database with automatic analyses and visualizations. *Nucleic Acids Res.* **52** (D1), D756–D761 (2024). doi:10.1093/nar/gkad979 Medline

100. N. Quinones-Olvera, Bacteriophage genome sizes, Github (2025); https://github.com/nataquinones/phage_genome_size/.

101. P. Kunzmann, K. Hamacher, Biotite: A unifying open source computational biology framework in Python. *BMC Bioinformatics* **19**, 346 (2018). doi:10.1186/s12859-018-2367-z Medline

102. K. McNair, C. Zhou, E. A. Dinsdale, B. Souza, R. A. Edwards, PHANOTATE: A novel approach to gene identification in phage genomes. *Bioinformatics* **35**, 4537–4542 (2019). doi:10.1093/bioinformatics/btz265 Medline

103. M. J. McGuffie, J. E. Barrick, pLannotate: Engineered plasmid annotation. *Nucleic Acids Res.* **49** (W1), W516–W522 (2021). doi:10.1093/nar/gkab374 Medline

104. A. L. Delcher, K. A. Bratke, E. C. Powers, S. L. Salzberg, Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007). doi:10.1093/bioinformatics/btm009 Medline

105. J. Besemer, M. Borodovsky, GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33** (Web Server), W451-4 (2005). doi:10.1093/nar/gki487 Medline

106. U. Singh, E. S. Wurtele, orfipy: A fast and flexible tool for extracting ORFs. *Bioinformatics* **37**, 3019–3020 (2021). doi:10.1093/bioinformatics/btab090 Medline

107. T. Mihara, Y. Nishimura, Y. Shimizu, H. Nishiyama, G. Yoshikawa, H. Uehara, P. Hingamp, S. Goto, H. Ogata, Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, 66 (2016). doi:10.3390/v8030066 Medline

108. T. U. Consortium; UniProt Consortium, UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53** (D1), D609–D617 (2025). doi:10.1093/nar/gkae1010 Medline

109. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-

Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature* **585**, 357–362 (2020). doi:10.1038/s41586-020-2649-2 Medline

110. I. Virshup, S. Rybakov, F. J. Theis, P. Angerer, F. A. Wolf, anndata: Access and store annotated data matrices. *J. Open Source Softw.* **9**, 4371 (2024). doi:10.21105/joss.04371

111. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018). doi:10.1186/s13059-017-1382-0 Medline

112. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

113. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004). doi:10.1101/gr.849004 Medline

114. I. Karcagi, G. Draskovits, K. Umenhoffer, G. Fekete, K. Kovács, O. Méhi, G. Balikó, B. Szappanos, Z. Györfy, T. Fehér, B. Bogos, F. R. Blattner, C. Pál, G. Pósfai, B. Papp, Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Mol. Biol. Evol.* **33**, 1257–1269 (2016). doi:10.1093/molbev/msw009 Medline

115. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021). doi:10.1002/pro.3943 Medline

116. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017). doi:10.1038/nmeth.4169 Medline

117. P. B. Rosenthal, R. Henderson, Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003). doi:10.1016/j.jmb.2003.07.013 Medline

118. D. Kimanius, L. Dong, G. Sharov, T. Nakane, S. H. W. Scheres, New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochem. J.* **478**, 4169–4185 (2021). doi:10.1042/BCJ20210708 Medline

119. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004). doi:10.1107/S0907444904019158 Medline

120. T. I. Croll, ISOLDE: A physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D Struct. Biol.* **74**, 519–530 (2018). doi:10.1107/S2059798318002425 Medline

121. D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L.-W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, P. D. Adams, Macromolecular structure determination using X-rays, neutrons and electrons:

Recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019). doi:10.1107/S2059798319011471 Medline

122. S. Seabold, J. Perktold, Statsmodels: Econometric and statistical modeling with python. *SciPy* **7**, 92–96 (2010). doi:10.25080/Majora-92bf1922-011

123. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt; SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020). doi:10.1038/s41592-019-0686-2 Medline

124. O. Schwengers, L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom, A. Goesmann, Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genom.* **7**, 000685 (2021). doi:10.1099/mgen.0.000685 Medline

125. "The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations" (The National Academies Press, 2025); https://doi.org/10.17226/28868.

126. "Doing diligence to assess the risks and benefits of life sciences gain-of-function research" (The White House: President Barack Obama, 2014); https://obamawhitehouse.archives.gov/blog/2014/10/17/doing-diligence-assess-risks-and-benefits-life-sciences-gain-function-research/.

127. "Framework for nucleic acid synthesis screening" (The White House: President Joseph Biden, 2024); https://aspr.hhs.gov/S3/Documents/OSTP-Nucleic-Acid-Synthesis-Screening-Framework-Sep2024.pdf.

128. "Improving the safety and security of biological research" (The White House: President Donald Trump, 2025); https://www.whitehouse.gov/presidential-actions/2025/05/improving-the-safety-and-security-of-biological-research/.

129. J. R. Black, M. S. Hanke, A. Maiwald, T. Hernandez-Boussard, O. M. Crook, J. Pannu, Open-weight genome language model safeguards: Assessing robustness via adversarial fine-tuning. arXiv:2511.19299 [cs.LG] (2025).

130. D. Baker, G. Church, Protein design meets biosecurity. *Science* **383**, 349–349 (2024). doi:10.1126/science.ado1671 Medline

131. "NIH guidelines for research involving recombinant or synthetic nucleic acid molecules" (NIH Office of Science Policy, 2024); https://osp.od.nih.gov/wp-content/uploads/NIH_Guidelines.pdf.

132. "Activating AI safety level 3 protections" (Anthropic, 2025); https://www-cdn.anthropic.com/807c59454757214bfd37592d6e048079cd7a7728.pdf.

133. X. Hou, Y. He, P. Fang, S.-Q. Mei, Z. Xu, W.-C. Wu, J.-H. Tian, S. Zhang, Z.-Y. Zeng, Q.-Y. Gou, G.-Y. Xin, S.-J. Le, Y.-Y. Xia, Y.-L. Zhou, F.-M. Hui, Y.-F. Pan, J.-S. Eden, Z.-H. Yang, C. Han, Y.-L. Shu, D. Guo, J. Li, E. C. Holmes, Z.-R. Li, M. Shi, Using

artificial intelligence to document the hidden RNA virosphere. *Cell* **187**, 6929–6942.e16 (2024). doi:10.1016/j.cell.2024.09.027 Medline

134. J. Ito, A. Strange, W. Liu, G. Joas, S. Lytras, K. Sato; Genotype to Phenotype Japan (G2P-Japan) Consortium, A protein language model for exploring viral fitness landscapes. *Nat. Commun.* **16**, 4236 (2025). doi:10.1038/s41467-025-59422-w Medline

135. S.-Y. Jiang, S.-S. Zhao, J.-Q. Wei, S. Zhang, Z. Zhao, Y. Tong, W. Liu, J. Wang, T. Jiang, J. Li, General intelligence framework to predict virus adaptation based on a genome language model. *Research* **8**, 0871 (2025). doi:10.34133/research.0871 Medline

136. C. Martin, A. Gitter, K. Anantharaman, Protein Set Transformer: A protein-based genome language model to power high-diversity viromics. *Nat. Commun.* **16**, 11123 (2025). doi:10.1038/s41467-025-66049-4 Medline

137. E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, A. Madani, ProGen2: Exploring the boundaries of protein language models. *Cell Syst.* **14**, 968–978.e3 (2023). doi:10.1016/j.cels.2023.10.002 Medline

138. C. Peng, J. Shang, J. Guan, D. Wang, Y. Sun, ViraLM: Empowering virus discovery through the genome foundation model. *Bioinformatics* **40**, 704 (2024). doi:10.1093/bioinformatics/btae704 Medline

139. M. Zvyagin, A. Brace, K. Hippe, Y. Deng, B. Zhang, C. O. Bohorquez, A. Clyde, B. Kale, D. Perez-Rivera, H. Ma, C. M. Mann, M. Irvin, D. G. Ozgulbas, N. Vassilieva, J. G. Pauloski, L. Ward, V. Hayot-Sasson, M. Emani, S. Foreman, Z. Xie, D. Lin, M. Shukla, W. Nie, J. Romero, C. Dallago, A. Vahdat, C. Xiao, T. Gibbs, I. Foster, J. J. Davis, M. E. Papka, T. Brettin, R. Stevens, A. Anandkumar, V. Vishwanath, A. Ramanathan, GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *Int. J. High Perform. Comput. Appl.* **37**, 683–705 (2023). doi:10.1177/10943420231201154

140. Z. Zhang, R. Jin, G. Xu, X. Wang, M. Zitnik, L. Cong, M. Wang, FoldMark: Safeguarding protein structure generative models with distributional and evolutionary watermarking. bioRxiv 2024.10.23.619960 [Preprint] (2025); https://doi.org/10.1101/2024.10.23.619960.