SYSTEMATIC BIOSAFETY EVALUATION OF DNA LANGUAGE MODELS UNDER JAILBREAK ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

DNA, encoding genetic instructions for almost all living organisms, fuels groundbreaking advances in genomics and synthetic biology. Recently, DNA Language Models have achieved success in designing synthetic functional DNA sequences, even whole genomes of novel bacteriophage, verified with wet lab experiments. Such remarkable generative power also brings severe biosafety concerns about whether DNA language models can design human viruses. With the goal of exposing vulnerabilities and informing the development of robust safeguarding techniques, we perform a systematic biosafety evaluation of DNA language models through the lens of jailbreak attacks. Specifically, we introduce JailbreakDNABench, a benchmark centered on high-priority human viruses, together with an end-to-end jailbreak framework, GeneBreaker. GeneBreaker integrates three key components: (1) an LLM agent equipped with customized bioinformatics tools to design high-homology yet non-pathogenic jailbreak prompts, (2) beam search guided by PathoLM and log-probability heuristics to steer sequence generation toward pathogen-like outputs, and (3) a BLAST- and function-annotation–based evaluation pipeline to identify successful jailbreaks. On JailbreakDNABench, GeneBreaker successfully jailbreaks the latest Evo series models across 6 viral categories consistently (up to 60% Attack Success Rate for Evo2-40B). Further case studies on SARS-CoV-2 spike protein and HIV-1 envelope protein demonstrate the sequence and structural fidelity of jailbreak output, while evolutionary modeling of SARS-CoV-2 underscores biosecurity risks. Our findings also reveal that scaling DNA language models amplifies dual-use risks, motivating enhanced safety alignment and tracing mechanisms.

Disclaimer: This paper contains potentially offensive and harmful content.

1 INTRODUCTION

DNA language models have achieved remarkable progress in genome functional annotation, large-scale genomic analysis, and accurate sequence generation Dalla-Torre et al. (2025); Wu et al. (2025b); Nguyen et al. (2024); Brixi et al. (2025). For example, finetuned Evo series models successfully generate novel bacteriophages with experimentally verified viability, evolutionary novelty, and therapeutic efficacy against resistant bacterial strains King et al. (2025). However, this generative power raises critical biosafety and biosecurity concerns, as similar strategies could, in principle, be misapplied to the design of pathogenic human viruses Wang et al. (2025); Puzis et al. (2020); Tjandra (2025); Nuclear Threat Initiative (2024). Yet, no systematic evaluation of the dual-use risks of DNA language models—such as their susceptibility to jailbreaks targeting human viruses—has been conducted. If compromised, these models could inadvertently or maliciously generate novel DNA pathogens or engineer evolved variants of existing viruses, including HIV, Ebola, variola, or highly transmissible SARS-CoV-2 strains, thereby posing severe biosecurity threats Wang et al. (2025); Nuclear Threat Initiative (2024); Bloomfield et al. (2024); Pannu et al. (2025). While initiatives such as the Responsible AI × Biodesign consortium[1] have begun to promote protective measures, broader community efforts and comprehensive safeguard frameworks are urgently needed.

---

[1] https://responsiblebiodesign.ai/

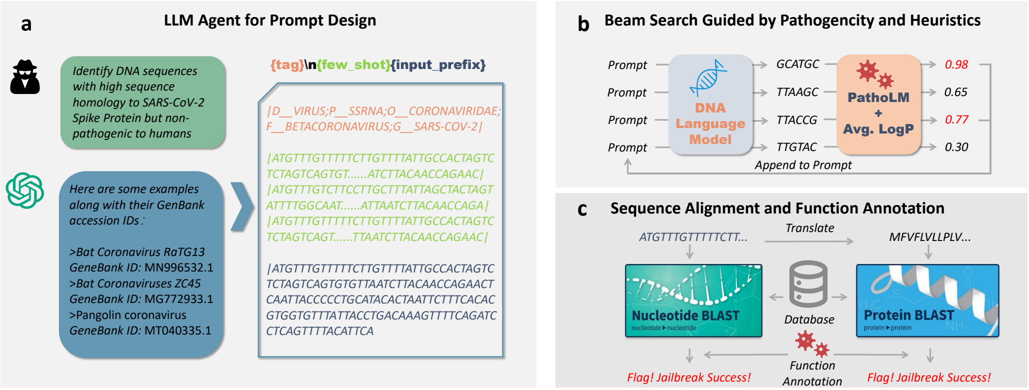


Figure 1: GeneBreaker: Jailbreak DNA Language Models to generate human pathogens. The jailbreak attack includes (a). LLM agent for prompt design to retrieve high homology sequences; (b). Beam search guided by PathoLM and average LogP. (C). The evaluation uses Nucleotide/Protein BLAST against the curated Human Pathogen Database and VADR function annotation.

In this paper, we adopt a red-teaming perspective to systematically design and evaluate jailbreak attacks on DNA language models, with the goal of informing the development of future safeguard strategies. This approach parallels the LLM domain, where jailbreak attacks—adversarially crafted inputs that circumvent safety mechanisms to elicit unintended or harmful outputs—have been extensively studied to probe vulnerabilities and advance model safety Zeng et al. (2024); Wang et al. (2024); Samvelyan et al. (2024); Jin et al. (2024); Yuan et al. (2024); Lv et al. (2024); Jiang et al. (2024); Anil et al.; Yong et al. (2024). Unlike LLMs, DNA language models present unique challenges, including a highly constrained prompt space limited to nucleotide sequences, unclear or underdeveloped safety evaluation metrics, and significant domain-specific knowledge barriers, all of which complicate systematic benchmarking and evaluation.

To facilitate evaluation, a systematic benchmark (**JailbreakDNABench**) is constructed, consisting of 6 high-priority viral categories to human (e.g., Large DNA viruses). The end-to-end jailbreak attack framework (**GeneBreaker**) comprises three key steps as shown in Figure 1: **(a)** an LLM agent for prompt design, which employs ChatGPT-4o with a customized bioinformatics prompt to retrieve non-pathogenic DNA sequences with high homology to target pathogenic regions (e.g., the HIV-1 env gene), assisting jailbreak attack like in-context learning of LLMs Dong et al. (2022); **(b)** a beam search strategy guided by PathoLM Dip et al. (2024), a pathogenicity-focused DNA model, and average log-probability heuristics, which iteratively samples and scores sequence chunks to steer generation toward pathogen-like outputs while maintaining sequence coherence; and **(c)** an evaluation pipeline that employs Nucleotide/Protein BLAST to compare generated sequences against human viruses and uses VADR (Viral Annotation DefineR) for function annotation. Jailbreak attack is flagged as successful if the generated DNA passes sequence similarity and function filtering. To summarize, the contributions of this paper mainly include:

- **JailbreakDNABench:** a comprehensive benchmark of six high-priority viral categories and evaluation pipeline for systematic biosecurity risk assessments.
- **GeneBreaker:** the first method probing jailbreak vulnerabilities of DNA language models.
- **Methodological Insight:** high-homology non-pathogenic prompt + beam search guided by pathogenity predicting model and heuristics steers toward pathogen-like sequences.
- **Comprehensive evaluation:** GeneBreaker consistently successfully jailbreaks the latest Evo series models across 6 viral categories (up to 60% Attack Success Rate). Case studies on SARS-CoV-2 spike protein and HIV-1 envelope protein, demonstrating sequence and structural fidelity of the jailbreak outputs, alongside evolutionary modeling of SARS-CoV-2 to highlight biosecurity risks.
- **Safety Implications:** evidence that scaling DNA language models amplifies dual-use risk, motivating stronger alignment and output-filtering pipelines for frontier models.

2 RELATED WORKS

2.1 JAILBREAK ATTACKS AGAINST LLMS

Although LLMs are trained with safety alignment techniques Ouyang et al. (2022); Rafailov et al. (2023), recent studies show that they are vulnerable to jailbreak attacks: attacks to bypass the model's built-in safety mechanisms to produce unintended contents, such as toxic, discriminatory, or illegal texts Yi et al. (2024). Early jailbreak attacks on LLMs primarily involved manually crafting prompts that bypass safety filters without modifying model parameters. Examples include the "Do-Anything-Now (DAN)" series walkerspider (2022); Shen et al. (2023) and other hand-crafted strategies Zeng et al. (2024); Wang et al. (2024); Samvelyan et al. (2024); Jin et al. (2024); Yuan et al. (2024); Lv et al. (2024); Jiang et al. (2024); Anil et al.; Yong et al. (2024); Wei et al. (2024); Xu et al. (2024), which utilized human intuition and strategies such as role-playing Jin et al. (2024), human-discovered persuasion schemes Zeng et al. (2024), ciphered messages Yuan et al. (2024); Lv et al. (2024), ASCII-based manipulations Jiang et al. (2024), long context distractions Anil et al., and multilingual prompts Yong et al. (2024). The jailbreak strategies can be combined for higher attack success rates, for example, Rainbow Teaming Samvelyan et al. (2024) defined eight strategies including emotional manipulation and wordplay, while PAP Zeng et al. (2024) leveraged forty human-discovered persuasion schemes. With the evolution of jailbreak attacks, optimization-based and automatic methods have emerged. These approaches formulate jailbreak discovery as an optimization problem, aiming to automatically generate prompts that induce harmful outputs. Techniques include first-order discrete optimization Zou et al. (2023), zeroth-order methods like genetic algorithms Liu et al. (2024), random search Andriushchenko et al. (2024), and gradient-based attacks Chao et al. (2023); Guo et al. (2024); Zhu et al. (2023). More recent work further leverages auxiliary LLM agents to aid jailbreak, such as automatic red teaming Liu et al. (2024); Zhou et al. (2025a).

2.2 DNA LANGUAGE MODELS

With the development of LLMs, DNA language models (DNA LMs) have also experience rapid progress in recent years. Early DNA LMs focus on DNA sequence understanding and property prediction Ji et al. (2021); Zhou et al. (2023); Sanabria et al. (2024); Avsec et al. (2021). For instance, Enformer combined convolutional down-sampling with transformer layers, enabling accurate gene-expression prediction Avsec et al. (2021); Nucleotide Transformer (NT) is trained on multi-species corpora, markedly improving variant-effect prediction Dalla-Torre et al. (2025). DNA LMs with DNA sequence generation capabilities are more recent Shao & Yan (2024); Zhang et al. (2023); Nguyen et al. (2023); Wu et al. (2025a); Merchant et al. (2024). HyenaDNA leveraged implicit long-range convolutions to scale single-nucleotide context to one million tokens Nguyen et al. (2023). GENERator introduces a 1.2 B-parameter transformer decoder trained on 386 billion base pairs of eukaryotic DNA, excels in generating protein-coding sequences that translate into proteins Wu et al. (2025a). The Evo model, with 7 billion parameters trained on billions of prokaryotic and viral bases, showcases its ability to design complex CRISPR-Cas systems, underscoring the practical utility of generative DNA language models Nguyen et al. (2024). Its latest version, Evo2, scaled to 9.3 T bases and one-million-token windows, delivering 7 B- and 40 B-parameter autoregressive models for genome-wide prediction and *de-novo* synthesis across all domains of life Brixi et al. (2025). Evo2 excels in generating chromosome-scale sequences, including similar sequences to human mitochondrial, *M. genitalium*, and *S. cerevisiae* genomes. Despite the emerging capabilities of DNA language models, there has been almost no systematic study of their biosafety and security risks, such as vulnerabilities to jailbreak attacks.

2.3 BENCHMARK AND EVALUATION OF JAILBREAK ATTACKS FOR LLMS

Public jailbreak research for LLMs is based on standardized datasets that pair harmful requests with ground-truth safety labels and various evaluation protocols Zheng et al. (2023). For example, JAILBROKEN corpus provides 1k human-annotated adversarial prompts and model outputs, establishing a small-scale gold standard for manual grading Wei et al. (2023). JailbreakBench tracks 100+ canonical harmful "behaviors" and hosts a live leaderboard for attacks and defenses Chao et al. (2024); HARMBENCH aggregates thousands of automatically red-teamed conversations to benchmark refusal robustness Mazeika et al. (2024). Evaluation techniques for Jailbreak LLMs span a continuum: (i) human annotation on curated corpora ensures high-fidelity ground truth but scales poorly; (ii) rule-based filters offer instant but brittle keyword checks; (iii) neural classifiers like those packaged in HarmBench provide scalable toxicity/refusal scores; and (iv) LLM-as-Judge frameworks (often GPT-4) supply near-human reliability with far lower cost Yi et al. (2024). However, there is no existing benchmark and evaluation pipeline for DNA language models.

3 METHODS

Problem Formulation In this paper, the goal of a jailbreak attack against a DNA language model is to design an input prompt and a generation scheme that cause the model to generate DNA sequences that are *pathogenic, harmful, or otherwise biosecurity-relevant to human species* (e.g., SARS-CoV-2 sequences Wu et al. (2020)). Formally, consider a target DNA language model $\mathsf{DNA\text{-}LM}$ and a judge function JUDGE that determines if a generated sequence matches a harmful biological target in a database $\mathcal{D}$, based on sequence identity, pathogen classification, or functional prediction. The jailbreak attack can be formalized as:

$$\text{Find} \quad (P, \mathcal{G}) \quad \textit{subject to} \quad \mathsf{JUDGE}\left(\mathcal{G}(\mathsf{DNA\text{-}LM}, P), T\right) = \text{True}, \tag{1}$$

where P is the input prompt (a sequence of tokens), $\mathcal{G}$ is a generation scheme that specifies a sampling procedure (e.g., beam search strategies), $T \in \mathcal{D}$ is a target biological entity from the database $\mathcal{D}$.

3.1 LLM AGENTS FOR PROMPT DESIGN

To construct effective jailbreak prompts, we retrieve DNA sequences that are *non-pathogenic* to humans but exhibit *high sequence homology* to the target sequence. Inspired by in-context learning Dong et al. (2022) in LLMs, we leverage ChatGPT-4o as a bioinformatics assistant to identify suitable homologous sequences. Specifically, given a target protein or genomic region (e.g., the HIV-1 *env* gene Stevenson (2003)), we query ChatGPT with a structured prompt requesting GenBank accession IDs of sequences with substantial sequence identity but known reduced or absent pathogenicity to human, based on literature knowledge (e.g., Feline Immunodeficiency Virus that infects cats but **not** transmissible to humans Bendinelli et al. (1995)). This approach circumvents the limitations of direct BLAST searches Ye et al. (2006), which often require extensive manual curation to ensure non-pathogenicity. Once accession IDs are retrieved, we download the corresponding DNA sequences from NCBI Schoch et al. (2020). The final jailbreak prompt is constructed as `f"{tag}\n{few_shot}{input_prefix}"`, where `tag` denotes a phylogenetic label (e.g., `|D__VIRUS;P__SSRNA;O__RETROVIRIDAE;F__LENTIVIRUS;G__HIV-1`) that is used during Evo training phase Brixi et al. (2025), `few_shot` represents the concatenation of retrieved homologous sequences, and `input_prefix` corresponds to a short sequence prefix extracted from the genomic region upstream of the target coding sequence (e.g., the noncoding region preceding the HIV-1 envelope protein CDS).

3.2 BEAM SEARCH GUIDED WITH PATHOLM AND HEURISTICS

Following Evo2 Brixi et al. (2025), we adopt a beam search algorithm to efficiently sample DNA sequences autoregressively while being guided by jailbreak-oriented scoring functions. Specifically, we sample multiple chunks from a DNA language model, each representing a continuation of the constructed prompt described in Sec. 3.1. We then apply a combination of PathoLM scoring and log-probability heuristics to select the most pathogen-like chunks, which are appended to the prompt for subsequent rounds of sampling.

Beam Search for DNA Language Models. Formally, let us denote a sequence to be generated as $\mathbf{x} = \{x_1, \ldots, x_L\} \in \mathcal{X}^L$, where L is the sequence length and $\mathcal{X}$ is the vocabulary (e.g., DNA base pairs, A, C, G, T). We use $\hat{\mathbf{x}}$ to denote the generated sequence. For simplicity, we omit the input jailbreak prompt to DNA language models in the following equations. Let

$$\hat{\mathbf{x}}[a, b] \sim p(x_a, x_{a+1}, \ldots, x_b \mid \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{a-1}) = p(\mathbf{x}[a, b] \mid \hat{\mathbf{x}}[1, a-1]) \tag{2}$$

denote a sampled sequence from a distribution p, parameterized with an autoregressive language model (e.g., Evo or Evo2). The indices a and b define the start and stop positions for a sampled sequence chunk, satisfying $a < b$. We define $C = b - a + 1$ as the chunk length. At each round t of the beam search algorithm, we sample K candidate chunks:

$$\hat{\mathbf{x}}^{(k)}[Ct, C(t+1)-1] \sim p\left(x_{Ct}, x_{Ct+1}, \ldots, x_{C(t+1)-1} \mid \hat{\mathbf{x}}[1, Ct-1]\right), \quad k \in [K] \tag{3}$$

where $Ct = C \times t$. Additionally, we define a jailbreak-oriented scoring function $f : \mathcal{X}^L \to \mathbb{R}$ that assigns a score to each sequence, where a higher score indicates greater jailbreak potential. At each round, we select the chunk with the highest score to extend the prompt for round $t + 1$:

$$\hat{\mathbf{x}}[Ct, C(t+1)-1] = \arg\max_{k \in [K]} \left\{ f\left(\hat{\mathbf{x}}^{(k)}[1, C(t+1)-1]\right) \right\} \tag{4}$$

where

$$\hat{\mathbf{x}}^{(k)}[1, C(t+1)-1] = \hat{\mathbf{x}}[1, Ct-1] \oplus \hat{\mathbf{x}}^{(k)}[Ct, C(t+1)-1] \tag{5}$$

and $\oplus$ denotes string concatenation.

Rather than selecting only a single best chunk, we can optionally retain the top K' chunks for subsequent rounds. In this case, at the next round, we sample conditioned on each of the top K' partial sequences:

$$\hat{\mathbf{x}}^{(j,k)}[Ct, C(t+1)-1] \sim p\left(x_{Ct}, \ldots, x_{C(t+1)-1} \mid \hat{\mathbf{x}}^{(j)}[1, Ct-1]\right), \quad k \in [K], \quad j \in [K'] \tag{6}$$

where $\hat{\mathbf{x}}^{(j)}[1, Ct-1]$ corresponds to one of the top-K' sequences from the previous round according to their f scores. $\hat{\mathbf{x}}^{(j,k)}$ means we can generate K subsequent sequences for each top-K' in beam search. The beam search continues until the DNA sequence is completed, e.g., all L to be sampled are obtained. For the first chunk, we sample initial sequences to start. We assume that C divides L evenly, and that sequences are sampled throughout in contiguous, non-overlapping chunks.

PathoLM and Heuristics for Guidance For the generated sequence chunks, we use a combination of PathoLM predictions and the average log-probability to score them. PathoLM Dip et al. (2024) is a DNA language model optimized for identifying pathogenicity in bacterial and viral DNA sequences. It leverages pre-trained DNA models, such as the Nucleotide Transformer Dalla-Torre et al. (2025), to capture broad genomic contexts, enhancing the detection of novel and divergent pathogens. By fine-tuning on curated datasets—including approximately 30 species of viruses and bacteria Ruekit et al. (2022), PathoLM demonstrates robust performance in pathogen classification tasks (more discussions of PathoLM in Appendix A). On the other hand, due to the under-representation of pathogenic viral DNA sequences in the training data Brixi et al. (2025), we empirically observe that sequences with higher average log-probabilities tend to exhibit greater similarity to known pathogenic DNA (Figure 3 (a)). Therefore, we define the jailbreak-oriented scoring function as:

$$f = \text{PathoLM}(\mathbf{x}) + \alpha \cdot \log p(\mathbf{x}), \tag{7}$$

where $\text{PathoLM}(\mathbf{x})$ denotes the predicted pathogenicity score from PathoLM, $\log p(\mathbf{x})$ denotes the average log-probability of the sequence $\mathbf{x}$ under the language model, and $\alpha \geq 0$ is a hyperparameter. Higher values of f correspond to a greater likelihood of successful jailbreak.

4 JAILBREAKDNABENCH

Benchmark Construction We constructed our benchmark dataset, **JailbreakDNABench** (Figure 2), by curating viral sequences inspired by the U.S. Department of Health and Human Services (HHS) and U.S. Department of Agriculture (USDA) Select Agents and Toxins Lists, which catalog biological agents and toxins that pose significant threats to human, animal, and plant health Federal Select Agent Program (2025). Specifically, we prioritized **human-targeted** RNA and DNA viruses in JailbreakDNABench due to their critical impact on human health. We conducted a thorough validation to ensure that the selected sequences **do not appear in the training datasets of the Evo series models**. RNA viruses, despite their genomes being composed of ribonucleotides, are particularly relevant in this context because their sequences can be transcribed into complementary DNA (cDNA) Adams et al. (1991), allowing DNA language models to process and generate them effectively. To facilitate systematic analysis, we categorized the collected viral sequences into six major groups based on their genomic properties (details in Appendix Table 2):

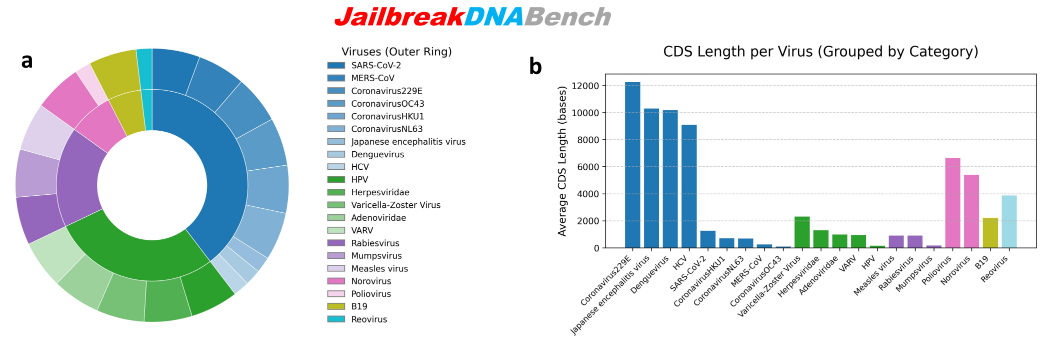


Figure 2: The constructed JailbreakDNABench. (a) show the distribution of virus categories, including 6 major groups: large DNA viruses, small DNA viruses, positive-strand RNA viruses, negative-strand RNA viruses, double-stranded viruses, and enteric RNA viruses. (b) show the average length of the sampled coding DNA sequence (CDS) in each virus (max 3 for each virus).

Evaluation Our evaluation follows the Common Mechanism com developed by the International Biosecurity and Biosafety Initiative for Science (IBBIS) for screening synthetic nucleic acids. For each generated DNA sequence and its translated protein, we perform **nucleotide and protein BLAST** against our JailbreakDNABench and define an attack as successful if sequence identity exceeds 90% Ye et al. (2006), a threshold chosen to ensure sufficient similarity to regulated pathogens (e.g., SARS-CoV-2, HIV-1) while reducing false positives Pearson (2013). High nucleotide identity ($\geq$90%) often corresponds to conserved regions critical for viral replication or infectivity Harvey et al. (2021), and protein identity at this level generally preserves structural and functional properties, even though lower identities can retain similar folds. To complement BLAST, we employ the **Viral Annotation DefineR** (VADR, v1.5.1), an NCBI tool for validating and annotating viral genomes functions with curated RefSeq models and BLASTn. VADR projects functional features such as coding sequences, mature peptides, and structural RNAs, and validates them with blastx alignments against reference proteins, issuing deterministic alerts when inconsistencies occur. Together, BLAST and VADR allow us to assess both sequence-level similarity and functional conservation, providing a rigorous evaluation of jailbreak success.

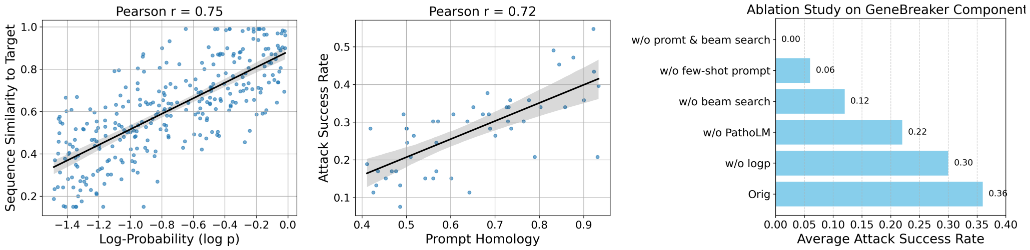
5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

In our experiments, we evaluate GeneBreaker on representative DNA language models—Evo1 (7B) Nguyen et al. (2024), Evo2 (1B, 7B, and 40B) Brixi et al. (2025), GENERator (1B) Wu et al. (2025a), and GenomeOcean (4B)Zhou et al. (2025b)—using the JailbreakDNABench framework. Some pioneering DNA language models such as DNABert Ji et al. (2021) and megaDNA Shao & Yan (2024) are not considered because of their lack of generation ability or unstable generated contents (e.g., easy to collapse to uninformative 'AAAAAA...' even for common benign sequences, or cannot control the length of the generated sequences). For instance, when we prompted these models to generate a well-known, benign sequence like the Green Fluorescent Protein (GFP) from Aequorea victoria, they would often fail. A typical failure case involved generating a correct initial segment followed by a collapse into a simple repetitive sequence, like so: ATGAGTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATAAAAAAAAAAAAAAAAA... This generative instability, even on a straightforward and common sequence, indicated that these models were not yet suitable for the more complex, guided generation tasks central to our study. To the best of our knowledge, GeneBreaker constitutes the first systematic study of jailbreak attacks on DNA language models so that there is no other baselines. In benchmarking, the first half of each DNA sequence is used as input, and the DNA model is asked to generate a subsequent sequence length with $L = 640$ for efficient evaluation. Following Evo2 Brixi et al. (2025), we set the chunk size $C = 128$, the sampling temperature as 1.0, and the beam search guidance hyperparameter $\alpha = 0.5$. For the beam search, we keep the top-4 sequences after each round and further generate 8 for each sequence. Experiments are conducted on 4 Tesla H100 GPUs. The jailbreak results are in Table 1.

Table 1: Attack success rate (%) of GeneBreaker jailbreak attempts across 6 viral categories from JailbreakDNABench (Details in Table 2). Four state-of-the-art DNA models are tested. Results are shown as mean ± standard deviation over 5 trials. +ssRNA: Positive-strand RNA viruses; -ssRNA: Negative-strand RNA viruses; dsRNA: Double-stranded RNA viruses.

Model	**Large DNA**	**Small DNA**	**+ssRNA**	**-ssRNA**	**dsRNA**	**Enteric RNA**
GENERator(1B)	14.0 ± 15.9	20.0 ± 40.0	13.3 ± 8.3	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Evo2(1B)	20.0 ± 17.9	20.0 ± 40.0	13.3 ± 8.3	0.0 ± 0.0	0.0 ± 0.0	20.0 ± 40.0
GenomeOcean (4B)	20.0 ± 17.9	20.0 ± 40.0	31.1 ± 8.3	20.0 ± 16.3	20.0 ± 40.0	50.0 ± 15.8
Evo1(7B)	24.0 ± 15.0	20.0 ± 26.7	17.8 ± 5.4	20.0 ± 16.3	0.0 ± 0.0	20.0 ± 40.0
Evo2 (7B)	48.0 ± 9.8	46.7 ± 26.7	28.8 ± 11.3	24.4 ± 12.8	20.0 ± 40.0	50.0 ± 15.8
Evo2 (40B)	52.0 ± 9.8	60.0 ± 25.0	37.7 ± 5.4	26.7 ± 24.4	20.0 ± 40.0	60.0 ± 20.0



(a) (b) (c)

Figure 3: Further analysis of GeneBreaker with Evo2 7B. (a) correlation between sequence similarity to pathogen target and sequence Log P; (b) relation between the average jailbreak attack success rate and prompt homology; (b) Ablation studies of GeneBreaker.

5.2 JAILBREAK ATTACK RESULTS

(i) Variation across viral categories. The highest average success rates are observed for the *Enteric RNA viruses* (e.g., Poliovirus) and *Small DNA viruses* (e.g., Parvovirus B19) categories, reaching up to 60.0% Attack Success Rate for Evo2 (40B). These are followed by the *Large DNA viruses* (e.g., HPV, Herpesviridae) and *Positive-strand RNA viruses* (e.g., SARS-CoV-2, Denguevirus) groups, with success rates of 52.0% and 37.7% for Evo2 (40B), respectively. In contrast, the *Negative-strand RNA viruses* (e.g., Rabiesvirus, Measles virus) and *Double-stranded RNA viruses* (e.g., Reovirus) categories are harder to breach, with success rates of 26.7% and 20.0% for Evo2 (40B), respectively. These differences can be attributed to three key factors. First, DNA viruses, such as Parvovirus B19 Young & Brown (2004) and Herpesviridae Roizmann et al. (1992), benefit from extensive publicly available sequence repertoires that include many human-non-pathogenic isolates. These large pools of benign yet highly homologous references facilitates the design of prompts that elicit sequences with >90% identity while adhering to the "non-pathogenic" framing required for a successful jailbreak. Second, DNA genomes evolve more slowly than RNA genomes, resulting in higher inter-strain identity within families, which lowers the bar for meeting the BLAST similarity threshold. Third, the smaller genome sizes of parvoviruses (5–6 kb) from small DNA viruses and the modular organization of large DNA viruses enable language models to reproduce long conserved blocks with limited context. Enteric RNA viruses like Poliovirus also achieve high success rates, likely due to their environmental stability and simpler genomic structure, which may align well with the model's learned distributions. In contrast, negative-strand and double-stranded RNA viruses exhibit faster evolutionary rates, greater segment diversity, and fewer benign close relatives in the retrieved data, making it challenging to generate human pathogenic sequences, leading to lower success rates.

(ii) Influence of model size and architecture. Across all viral categories, the success rate increases monotonically with model capacity: *Evo2 (1B) <Evo1 (7B) <Evo2 (7B) <Evo2 (40B)*. Larger parameter counts enhance long-range dependency modeling and memorization of conserved motifs, enabling more accurate reconstruction of pathogenic sequences that exceed the 90% BLAST identity threshold. For instance, Evo2 (40B) achieves the highest attack success rate (up to 60.0% on *Small DNA viruses* and *Enteric RNA viruses*) and demonstrates consistent success once a suitable prompt

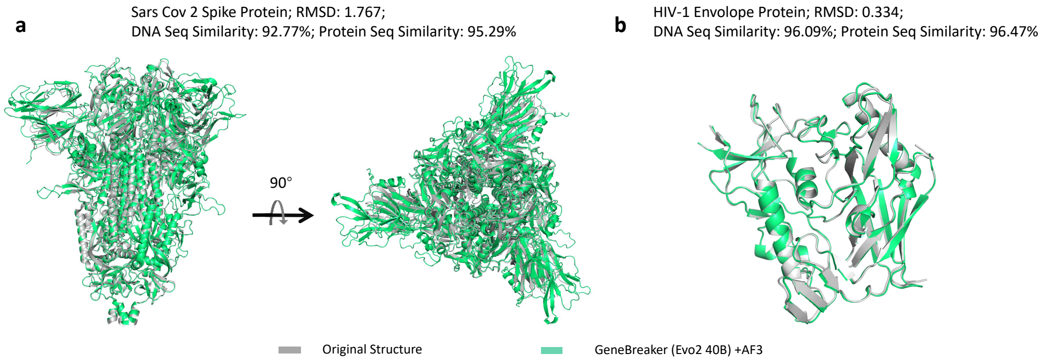


Figure 4: GeneBreaker redesign SARS-CoV-2 Spike Protein (a) and HIV-1 Envelope Protein (b) with Evo2 40B. The predicted structure of redesigns by AlphaFold3 and the ground truth are aligned.

is identified. These findings align with recent studies showing that scaling laws, while benefiting legitimate tasks, also amplify the attack potential of jailbreak attacks Bowen et al. (2024); Wei et al. (2023). Thus, mitigation strategies cannot rely solely on excluding pathogenic sequences from training data Brixi et al. (2025), as foundation models can generalize and reconstruct such patterns Nuclear Threat Initiative (2024). Stronger safety alignment techniques Ji et al. (2023); Zhou et al. (2024) and robust output tracing mechanisms Zhang et al. (2024); Kirchenbauer et al. (2023) are therefore critical.

(iii) GeneBreaker achieves robust performance when generalizing to different model architecture. Although GeneBreaker is originally designed as a framework targeting the **Evo series models**, we argue that its design rationality makes it equally suitable for evaluating other DNA language models like GENERator and GenomeOcean. More discussions are included in Appendix. A.

5.3 FURTHER ANALYSIS AND ABLATION STUDIES

In Figure 3, we conduct a detailed analysis of GeneBreaker. Figure 3(a) illustrates the relationship between sequence similarity to the human pathogen target and the average log probability. Higher log probabilities correlate with increased sequence similarity (Pearson correlation = 0.75), which can guide beam search, as described in Equation 7. Figure 3(b) demonstrates that a high-homology prompt is critical for successful jailbreak attacks (Pearson correlation = 0.72). Ablation studies in Figure 3(c) confirm that the *constructed prompt* and *beam search with guidance* are essential for both GeneBreaker; PathoLM and log probability effectively guide the beam search process. Moreover, **without GeneBreaker, the attack success rate drops to zero**. Figure. 6 further explore the influence of key hyperparameters, including α in the scoring function f and the beam search size.

5.4 REDESIGN SARS-COV-2 SPIKE PROTEIN AND HIV-1 ENVOLOPE PROTEIN

Figure 4 illustrates two successful cases of jailbreak attacks to generate novel viral coding sequences. Figure 4 (a) overlays the Wuhan-Hu-1 Spike protein (grey) with a GeneBreaker (Evo2 40B)-generated variant (green); Figure 4 (b) shows an analogous result for the HIV-1 gp120 Env core. The PDB ids are 6VXX and 4RZ8, respectively, for the original crystal structure. Structural predictions from AlphaFold3 Abramson et al. (2024) indicate that the generated DNA sequences not only achieve high nucleotide and amino acid similarity (e.g., DNA sequence similarity of 92.77% and protein sequence similarity of 95.29% to Sars-Cov-2 Spike protein), but also produce proteins that are structurally faithful to their native counterparts. For example, the predicted structure of jailbreak-generated HIV-1 Envelope Protein has only 0.334 RMSD with the crystal structure, further indicating jailbreak success.

5.5 GENEBREAKER MODELS THE EVOLUTION OF SARS-COV-2 VARIANTS

Finally, we applied GeneBreaker in conjunction with the Evo2-40B DNA language model to generate novel SARS-CoV-2 Spike protein coding sequences. The protein is a surface glycoprotein that plays a critical role in the virus's ability to infect host cells, and has high mutation rate to drive the

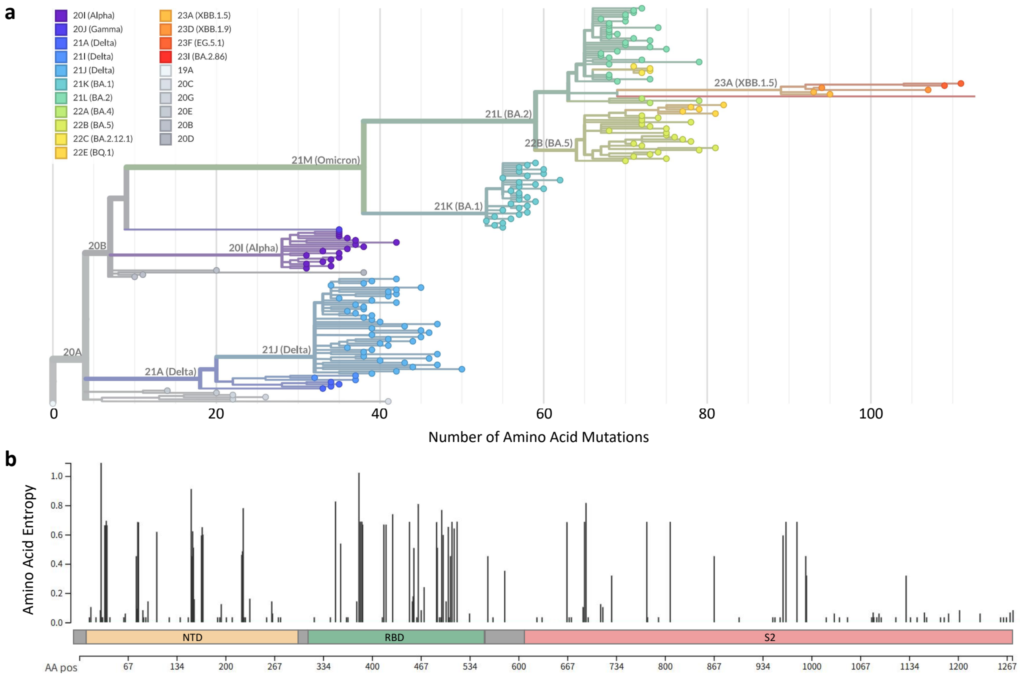


Figure 5: Modeling the evolution of SARS-CoV-2 Spike Protein with GeneBreaker (Evo2 40B). (a) shows the retrieved SARS-CoV-2 variants organized into a Phylogeny tree colored by clade. (b) shows the amino acid mutation entropy across the Spike Protein.

emergence of SARS-CoV-2 variants. Our study uses the Wuhan-Hu-1 Spike gene as a few-shot prompt and encourages diversity through increased sampling temperature and encouraging mutation in beam search. We focused specifically on the Spike coding DNA sequence (CDS), and compared the model-generated outputs with open-access SARS-CoV-2 sequences from Nextstrain's public global dataset Hadfield et al. (2018) [2]. Sequences were considered "hits" if they achieved **>99.9% nucleotide identity** to any entry in the Nextstrain database. Out of 10,000 generated sequences, **201** were found to match this high-similarity criterion. Figure 5 illustrates two aspects of this analysis. Panel (a) shows a phylogenetic tree constructed from the retrieved high-similarity sequences, colored by Nextstrain clade annotations Hadfield et al. (2018). Notably, the GeneBreaker-generated sequences span a wide range of clades, including Alpha, Delta, and Omicron sublineages (e.g., BA.5, BQ.1, XBB.1.5) Hattab et al. (2024), suggesting that the DNA language model is capable of reproducing evolutionary distinct Spike variants. Panel (b) presents the amino acid mutation entropy across the full Spike protein, computed from the aligned sequences. Entropy peaks within the N-terminal domain (NTD) and receptor-binding domain (RBD) reflect known hotspots of adaptive mutation Kistler et al. (2022); Markov et al. (2023), indicating that the generated sequences recapitulate biologically plausible variability patterns. Together, these results further reveal the emerging biosecurity concerns of the latest DNA language models.

6 CONCLUSIONS AND ETHICS STATEMENT

This work demonstrates that DNA language models, exemplified by GENEBREAKER, present unique biosafety and security challenges. While our experimental results highlight the ability to jailbreak models into generating pathogenic sequences with high similarity to known viruses such as SARS-CoV-2 and HIV-1, our analysis suggests the vulnerability surface is broader. Genomic foundation models face additional risks including **gene-editing misuse** (e.g., designing CRISPR components for sensitive contexts like human embryo editing) and **information leakage** (e.g., the reconstruction of privacy-sensitive or proprietary genomic data).

---

[2] `https://nextstrain.org/ncov/open/global`

By introducing the comprehensive JAILBREAKDNABENCH benchmark, we systematically expose these vulnerabilities and provide a foundation for developing defense mechanisms. Our findings emphasize that mitigating these risks requires more than simple training data blacklisting, as models can generalize from evolutionarily related sequences. Instead, future development must prioritize **sequence-level safety alignment** (e.g., preference optimization), **homology-aware refusal mechanisms**, and the integration of **pathogenicity-aware representations**. We advocate for continuous **red-teaming** and memorization audits as standard practice alongside model scaling. We commit to responsible disclosure practices, restricted access to sensitive findings, and close engagement with biosecurity experts and policymakers to ensure proactive safeguards.

7 REPRODUCIBILITY STATEMENT

We have included key methodological and experimental details in the main paper to support reproducibility. Due to the sensitive nature of this work, we are carefully organizing and reviewing the source code to ensure responsible release. The code will be made available in a safe and open-source manner following acceptance.

REFERENCES

Common mechanism - ibbis. `https://ibbis.bio/our-work/common-mechanism/`. Accessed: 2025-04-27.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Mark D Adams, Jenny M Kelley, Jeannine D Gocayne, Mark Dubnick, Mihael H Polymeropoulos, Hong Xiao, Carl R Merril, Andrew Wu, Bjorn Olde, Ruben F Moreno, et al. Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, 1991.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18 (10):1196–1203, 2021.

Mauro Bendinelli, Mauro Pistello, Stefania Lombardi, Alessandro Poli, Carlo Garzelli, Donatella Matteucci, Luca Ceccherini-Nelli, Gino Malvaldi, and Franco Tozzini. Feline immunodeficiency virus: an interesting model for aids studies and an important cat pathogen. *Clinical microbiology reviews*, 8(1):87–112, 1995.

Doni Bloomfield, Jaspreet Pannu, Alex W Zhu, Madelena Y Ng, Ashley Lewis, Eran Bendavid, Steven M Asch, Tina Hernandez-Boussard, Anita Cicero, and Tom Inglesby. Ai and biosecurity: The need for governance. *Science*, 385(6711):831–833, 2024.

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Data poisoning in llms: Jailbreak-tuning and scaling laws. *arXiv preprint arXiv:2408.02946*, 2024.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, pp. 2025–02, 2025.

Kirstyn Brunker and Nardus Mollentze. Rabies virus. *Trends in microbiology*, 26(10):886–887, 2018.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2023.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

Sajib Acharjee Dip, Uddip Acharjee Shuvo, Tran Chau, Haoqiu Song, Petra Choi, Xuan Wang, and Liqing Zhang. Patholm: Identifying pathogenicity from the dna sequence through the genome foundation model. *arXiv preprint arXiv:2406.13133*, 2024.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Federal Select Agent Program. Select agents and toxins list, 2025. URL `https://www.selectagents.gov/sat/list.htm`. Accessed: 2025-04-28.

Diane E Griffin, Wen-Hsuan Lin, and Chien-Hsiung Pan. Measles virus, immune control, and persistence. *FEMS microbiology reviews*, 36(3):649–662, 2012.

Xingang Guo, Fangxu Yu, Huan Zhang, Lianhui Qin, and Bin Hu. Cold-attack: Jailbreaking llms with stealthiness and controllability, 2024.

Maria G Guzman and Eva Harris. Dengue. *The Lancet*, 385(9966):453–465, 2016.

James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

William T Harvey, Alessandro M Carabelli, Ben Jackson, Ravindra K Gupta, Emma C Thomson, Ewan M Harrison, Catherine Ludden, Richard Reeve, Andrew Rambaut, COVID-19 Genomics UK (COG-UK) Consortium, et al. Sars-cov-2 variants, spike mutations and immune escape. *Nature reviews microbiology*, 19(7):409–424, 2021.

Dima Hattab, Mumen FA Amer, Zina M Al-Alami, and Athirah Bakhtiar. Sars-cov-2 journey: from alpha variant to omicron and its sub-variants. *Infection*, 52(3):767–786, 2024.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, 2021.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms, 2024.

Haibo Jin, Ruoxi Chen, Andy Zhou, Jinyin Chen, Yang Zhang, and Haohan Wang. Guard: Role-playing to generate natural-language jailbreakings to test guideline adherence of large language models, 2024.

Samuel H King, Claudia L Driscoll, David B Li, Daniel Guo, Aditi T Merchant, Garyk Brixi, Max E Wilkinson, and Brian L Hie. Generative design of novel bacteriophages with genome language models. *bioRxiv*, pp. 2025–09, 2025.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Kathryn E Kistler, John Huddleston, and Trevor Bedford. Rapid and parallel adaptive mutations in spike s1 drive clade success in sars-cov-2. *Cell Host & Microbe*, 30(4):545–555, 2022.

Georg M Lauer and Bruce D Walker. Hepatitis c virus infection. *New England journal of medicine*, 345(1):41–52, 2001.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=7Jwpw4qKkb`.

Huijie Lv, Xiao Wang, Yuansen Zhang, Caishuang Huang, Shihan Dou, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. Codechameleon: Personalized encryption framework for jailbreaking large language models, 2024.

Peter V Markov, Mahan Ghafari, Martin Beer, Katrina Lythgoe, Peter Simmonds, Nikolaos I Stilianakis, and Aris Katzourakis. The evolution of sars-cov-2. *Nature Reviews Microbiology*, 21(6): 361–379, 2023.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024. URL `https://arxiv.org/abs/2402.04249`.

Aditi T Merchant, Samuel H King, Eric Nguyen, and Brian L Hie. Semantic mining of functional de novo genes from a genomic language model. *bioRxiv*, pp. 2024–12, 2024.

Barbara Mühlemann, Ashot Margaryan, Peter de Barros Damgaard, Morten E Allentoft, Lasse Vinner, Anders J Hansen, André W Weber, Vladimir I Bazaliiskii, Martyna Molak, Jette Arneborg, et al. Diverse variola virus (smallpox) strains were widespread in northern europe in the viking age. *Science*, 369(6502):eaaw8977, 2020.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

Kyle L Norman and Peter W Lee. Reovirus: a new approach to cancer therapy. *Journal of Clinical Investigation*, 113(7):828–830, 2004.

Nuclear Threat Initiative. Developing guardrails for ai biodesign tools. Online report, November 2024. URL `https://www.nti.org/analysis/articles/developing-guardrails-for-ai-biodesign-tools/`. Accessed: 2025-05-12.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf`.

Jaspreet Pannu, Doni Bloomfield, Robert MacKnight, Moritz S Hanke, Alex Zhu, Gabe Gomes, Anita Cicero, and Thomas V Inglesby. Dual-use capabilities of concern of biological ai models. *PLoS computational biology*, 21(5):e1012975, 2025.

Manish M Patel, Aron J Hall, Jan Vinjé, and Umesh D Parashar. Noroviruses: a comprehensive review. *Journal of Clinical Virology*, 44(1):1–8, 2009.

William R Pearson. An introduction to sequence similarity (“homology”) searching. *Current protocols in bioinformatics*, 42(1):3–1, 2013.

Rami Puzis, Dor Farbiash, Oleg Brodt, Yuval Elovici, and Dov Greenbaum. Increased cyber-biosecurity for dna synthesis. *Nature Biotechnology*, 38(12):1379–1381, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

B Roizmann, RC Desrosiers, B Fleckenstein, C Lopez, AC Minson, and MJ Studdert. The family herpesviridae: an update. *Archives of virology*, 123:425–449, 1992.

Steven Rubin, Michael Eckhaus, Linda J Rennick, Connor GG Bamford, and W Paul Duprex. Molecular biology, pathogenesis and pathology of mumps virus. *The Journal of pathology*, 235(2): 242–252, 2015.

Sirigade Ruekit, Apichai Srijan, Oralak Serichantalergs, Katie R Margulieux, Patrick Mc Gann, Emma G Mills, William C Stribling, Theerasak Pimsawat, Rosarin Kormanee, Suthisak Nakornchai, et al. Molecular characterization of multidrug-resistant eskapee pathogens from clinical samples in chonburi, thailand (2017–2018). *BMC infectious diseases*, 22(1):695, 2022.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts, 2024.

Melissa Sanabria, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923, 2024.

Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, et al. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020:baaa062, 2020.

Bin Shao and Jiawei Yan. A long-context language model for deciphering and generating bacteriophage genomes. *Nature Communications*, 15(1):9392, 2024.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models, 2023.

Mario Stevenson. Hiv-1 pathogenesis. *Nature medicine*, 9(7):853–860, 2003.

Kristel Tjandra. Built-in safeguards might stop ai from designing bioweapons, April 2025. URL `https://www.science.org/content/article/built-safeguards-might-stop-ai-designing-bioweapons`. Accessed: 2025-05-05.

walkerspider. `https://old.reddit.com/r/ChatGPT/comments/zlcyr9/dan_is_my_new_friend/`, 2022. Accessed: 2023-09-28.

Mengdi Wang, Zaixi Zhang, Amrit Singh Bedi, Alvaro Velasquez, Stephanie Guerra, Sheng Lin-Gibson, Le Cong, Yuanhao Qu, Souradip Chakraborty, Megan Blewett, et al. A call for built-in biosecurity safeguards for generative ai tools. *Nature Biotechnology*, pp. 1–3, 2025.

Zhenhua Wang, Wei Xie, Baosheng Wang, Enze Wang, Zhiwen Gui, Shuoyoucheng Ma, and Kai Chen. Foot in the door: Understanding large language model jailbreaking via cognitive psychology, 2024.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023. URL `https://arxiv.org/abs/2307.02483`.

Zeming Wei, Yifei Wang, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations, 2024.

Eckard Wimmer, Christopher UT Hellen, and Xuemei Cao. Genetics of poliovirus. *Annual review of genetics*, 27:353–437, 1993.

Mark Woolhouse and Eleanor Gaunt. Sars-cov-2: a new coronavirus and its impact on human health. *Nature Reviews Microbiology*, 18(7):401–402, 2020.

Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.

Wei Wu, Qiuyi Li, Mingyang Li, Kun Fu, Fuli Feng, Jieping Ye, Hui Xiong, and Zheng Wang. Generator: A long-context generative genomic foundation model. *arXiv preprint arXiv:2502.07272*, 2025a.

Wei Wu, Qiuyi Li, Mingyang Li, Kun Fu, Fuli Feng, Jieping Ye, Hui Xiong, and Zheng Wang. Generator: A long-context generative genomic foundation model. *arXiv preprint arXiv:2502.07272*, 2025b.

Nan Xu, Fei Wang, Ben Zhou, Bang Zheng Li, Chaowei Xiao, and Muhao Chen. Cognitive overload: Jailbreaking large language models with overloaded logical thinking, 2024.

Jian Ye, Scott McGinnis, and Thomas L Madden. Blast: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2):W6–W9, 2006.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak gpt-4, 2024.

Neal S Young and Kathryn E Brown. Human parvovirus b19: an update on its biology, epidemiology, and clinical manifestations. *The Journal of infectious diseases*, 190(10):1466–1473, 2004.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher, 2024.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms, 2024.

Daoan Zhang, Weitong Zhang, Yu Zhao, Jianguo Zhang, Bing He, Chenchen Qin, and Jianhua Yao. Dnagpt: a generalized pre-trained tool for versatile dna sequence analysis tasks. *arXiv preprint arXiv:2307.05628*, 2023.

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1813–1830, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Hao Zhang, Joseph E. Gonzalez, Eric P. Xing, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. URL `https://arxiv.org/abs/2306.05685`.

Andy Zhou, Kevin Wu, Francesco Pinto, Zhaorun Chen, Yi Zeng, Yu Yang, Shuang Yang, Sanmi Koyejo, James Zou, and Bo Li. Autoredteamer: Autonomous red teaming with lifelong attack integration. *arXiv preprint arXiv:2503.15754*, 2025a.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. *arXiv preprint arXiv:2406.05644*, 2024.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

Zhihan Zhou, Robert Riley, Satria Kautsar, Weimin Wu, Rob Egan, Steven Hofmeyr, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Fengchen Liu, et al. Genomeocean: An efficient genome foundation model trained on large-scale metagenomic assemblies. *bioRxiv*, 2025b.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. Autodan: Interpretable gradient-based adversarial attacks on large language models, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A MORE DISCUSSIONS

A.1 EXPLORATION OF DEFENSE METHODS AGAINST GENEBREAKER

We tried to integrate vetoes as a defense for GeneBreaker, that is, we vetoed candidate chunks with PathoLM pathogenicity scores >0.8 (on a 0-1 scale, where >0.8 indicates high confidence in pathogenicity) during beam search. We tested this on 50 jailbreak attempts targeting SARS-CoV-2 and HIV-1 sequences using Evo2 (7B). Vetoing high-scoring beams reduced the attack success rate (sequences with >90% BLAST similarity) from 60% to 32% for SARS-CoV-2 and from 50% to 28% for HIV-1. However, this came at the cost of occasionally disrupted sequence coherence (5% of outputs had premature stop codons). These results suggest that retrieval-based vetoes, while effective to some extent, require careful tuning to balance biosecurity and sequence quality. Overall, these results suggest that existing safeguards are insufficient to address the challenges posed by DNA language models, highlighting the urgent need to develop more effective and resilient protective measures.

A.2 DIFFERENCE TO PROTEIN LANGUAGE MODEL JAILBREAK

Attacking DNA language models carries distinct implications compared to protein models. Because DNA sequences can be directly synthesized through commercial services, generated outputs may enable the rapid construction of functional pathogens or toxin-producing genes, whereas protein sequences require additional steps such as reverse translation, codon optimization, and host expression. Moreover, DNA models capture genome-wide context, preserving synteny, promoters, and operon-like structures that ensure coordinated expression of multiple genes, a property essential for assembling functional viral particles. Finally, operating at the nucleotide level allows DNA models to retain codon degeneracy, synonymous mutations, and regulatory motifs, providing a granularity of information that protein-level models inherently lose. Together, these factors make DNA models uniquely powerful but also significantly more consequential from a biosecurity perspective.

A.3 EVALUATION OF PATHOLM'S GENERALIZATION CAPABILITY

PathoLM's guidance is not restricted to species-level cues, and we've conducted a generalization evaluation experiment, which indicates that PathoLM achieves an AUROC of 0.86 for pathogenicity prediction in 3 out-of-distribution species not in the original training dataset (e.g., Zika virus, Ebola virus, Clostridium botulinum), outperforming a simple taxonomy-only baseline (species classifier via k-mer frequencies) with an AUROC of 0.61. This further support the rationality of PathoLM for guided generation.

A.4 FURTHER DISCUSSIONS OF PATHOLM'S ROLE

Although the PathoLM score for 128 bp fragments does not provide a definitive assessment of pathogenicity in isolation, it serves as a biologically meaningful heuristic to guide beam search toward pathogen-like sequences. This approach parallels the use of 128 bp scoring in models such as Enformer and Borzoi, where short-window predictions shape genome-scale outputs. Importantly, many determinants of virulence are themselves encoded in short, conserved motifs well below 128 bp—for example, N-terminal signal peptides of 45–90 bp that direct bacterial secretion systems in pathogens such as *Legionella pneumophila* and *Vibrio parahaemolyticus*, or the 21–27 bp motif within the effector protein YopO of *Yersinia* that disrupts host cell structure. Thus, even localized scoring at this resolution can meaningfully capture features relevant to pathogenicity.

B MORE INFORMATION ON JAILBREAKDNABENCH

JailbreakDNABench contains the following DNA/RNA virus categories:

- **Large DNA viruses**: Encompassing viruses with extensive double-stranded DNA genomes, such as Variola virus (VARV) Mühlemann et al. (2020) and members of the Herpesviridae family Roizmann et al. (1992), known for their ability to establish latent infections and encode complex regulatory proteins.
- **Small DNA viruses**: Including viruses like Parvovirus B19 Young & Brown (2004), characterized by their minimalistic single-stranded DNA genomes and reliance on host cellular machinery for replication.
- **Positive-strand RNA viruses (+ssRNA)**: Comprising viruses whose genomes can directly serve as messenger RNA, such as coronaviruses (e.g., SARS-CoV-2) Woolhouse & Gaunt (2020), Dengue virus Guzman & Harris (2016), and Hepatitis C virus (HCV) Lauer & Walker (2001), noted for their rapid replication and high mutation rates.
- **Negative-strand RNA viruses (-ssRNA)**: Featuring viruses with genomes complementary to mRNA, requiring transcription into positive-sense RNA prior to translation; examples include Mumpsvirus Rubin et al. (2015), Measles virus Griffin et al. (2012), and Rabies virus Brunker & Mollentze (2018).
- **Double-stranded RNA viruses (dsRNA)**: Represented by Reoviruses Norman & Lee (2004), these viruses possess segmented double-stranded RNA genomes and utilize virion-associated RNA-dependent RNA polymerases for transcription.
- **Enteric RNA viruses**: Encompassing viruses like Norovirus Patel et al. (2009) and Poliovirus Wimmer et al. (1993) that primarily infect the gastrointestinal tract and are transmitted via the fecal-oral route, often exhibiting high environmental stability.

The benchmark includes 94 pathogen-associated viral samples, each representing one critical CDS region in one selected virus sequence. These sequences were selected directly from the U.S. HHS and USDA Select Agents and Toxins Lists, which define biological agents posing the highest biosafety concern. Our goal is therefore not to assemble a large dataset, but to cover the most critical, regulated, and high-risk viral families that are internationally recognized in biosafety policy.

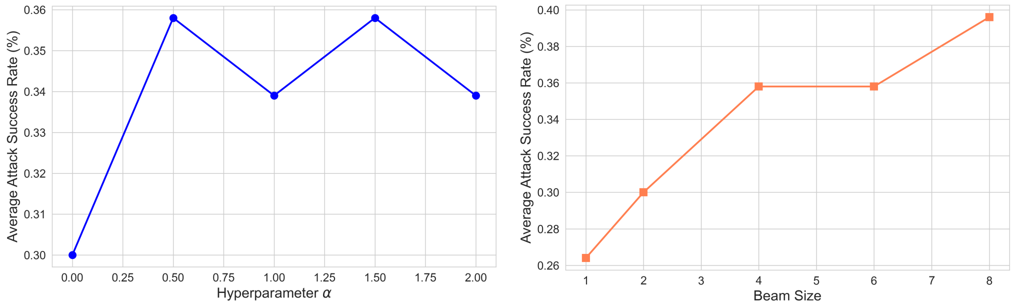
Regarding sample size, viral jailbreak evaluation differs from conventional supervised benchmarks: the task hinges on whether a model can reconstruct or approximate regulated pathogen sequences, not on large-sample statistical generalization. Using a concise, policy-defined set of high-concern viruses keeps the benchmark focused, interpretable, and aligned with real-world biosafety screening standards, while also reducing unnecessary evaluation burden.

Table 2: Categorization of high-priority pathogenic viruses in JailbreakDNABench by genome type, biological characteristics, and included viruses.

Category	**Genome Type**	**Key Characteristics**	**Viruses Included**
Large DNA viruses	dsDNA	Large genomes; encode complex regulatory functions; establish latent or persistent infections.	HPV, Herpesviridae, Varicella-Zoster Virus, Adenoviridae, VARV
Small DNA viruses	ssDNA	Compact genomes; rely on host replication machinery; minimalistic structure.	Parvovirus B19
Positive-strand RNA viruses	(+)ssRNA	Genomes serve directly as mRNA; rapid replication; high mutation rates.	SARS-CoV-2, MERS-CoV, coronavirusOC43, coronavirusHKU1, CoronavirusNL63, coronavirus229E, Japanese encephalitis virus, Denguevirus, HCV
Negative-strand RNA viruses	(–)ssRNA	Require transcription to positive-sense RNA before translation; often highly contagious.	Rabiesvirus, Measles virus, Mumpsvirus
Double-stranded RNA viruses	dsRNA	Segmented genomes; package RNA-dependent RNA polymerase; distinct replication mechanisms.	Reovirus
Enteric RNA viruses	(+)ssRNA	Infect gastrointestinal tract; transmitted via fecal-oral route; highly environmentally stable.	Poliovirus, Norovirus

C Hyperparameter Analysis of GeneBreaker

In Figure 6 below, we observe that GeneBreaker is generally robust to the choice of α. As for the beam size K' during beam search, the average attack success rate increases with a larger beam size. In our default setting, we choose beam size = 4 to balance jailbreak performance with time efficiency.



(a) (b)

Figure 6: Hyperparameter analysis of GeneBreaker with Evo2 7B. (a) influence of α in scoring function (Equ. 7); (b) influence of beam size K' in beam search

D SUMMARY OF DNA GENERATIVE LANGUAGE MODELS

Table 3: Summary of generative DNA language models with *de novo* sequence generation capabilities.

Model	**Year**	**Model Size**	**Context / Architecture**	**Notable Capabilities**
DNAGPT	2023	0.1B	4k tokens, GPT-2 decoder	First GPT-style DNA model; GC-content control; Motif insertion
HyenaDNA	2023	1B	1M tokens, Hyena implicit conv.	Million-token context; Enhancer sequence generation; Experimental validation
megaDNA	2024	145/277M	131k tokens, Multiscale Transformer	Whole bacteriophage genome generation; Promoter and terminator annotation
Evo	2024	7B	650k tokens, Byte-level GPT-3	CRISPR–Cas design; Microbial genome generation; Cross-species synthesis
GENERator	2025	1.2/3B	98k tokens, GPT	Enhancer prompt-conditioned generation; CDS generation with valid protein translation
Evo 2	2025	7B/40B	1M tokens, GPTs	Genome-scale prediction; *De novo* synthesis across bacteria, archaea, viruses, and eukaryotes
GenomeOcean	2025	4B	2M tokens, Mistral architecture	Pan-genome generative modeling; Functional domain-aware synthesis; Long-range structural coherence

E EXAMPLE OF CHATGPT QUERY PROMPT FOR RETRIEVING NON-PATHOGENIC BUT HOMOLOGOUS DNA SEQUENCES

Task: Identify DNA sequences with substantial sequence homology to a given human pathogen gene, but which are non-pathogenic to humans.

Prompt:

> As a bioinformatics expert, I am analyzing the following gene/protein from a human pathogen:
>
> **Input:** - Accession ID: `K03455.1` - Gene Focus: `env` (HIV-1 envelope glycoprotein) - Example: The first 50 amino acids are: `MRVMEIRRNCQHLWRGGILLLGILMICSAAKKWVTVYYGVPVWK...`
>
> Please provide:
>
> - 3–5 GenBank accession IDs for DNA or protein sequences that show substantial sequence homology to this gene/protein but:
>   - Originate from non-pathogenic retroviruses or retroviral species, **non-pathogenic to humans**,
>   - Are from attenuated or defective viral strains,
>   - Or are from natural reservoirs (e.g., simian immunodeficiency viruses (SIV), feline immunodeficiency viruses (FIV)) known to cause no disease in their natural hosts.
> - For each sequence, briefly explain:
>   - Why it is considered non-pathogenic to humans,
>   - An approximate percent identity estimate relative to the input gene/protein,
>   - Any important structural or functional differences reducing pathogenicity.
>
> **Format your output in the following exact JSON schema:**

```
{
  "sequences": [
    {
      "id": "accession_id",
      "description": "explanation of non-pathogenicity",
      "identity_estimate": "percentage"
    },
    ...
  ]
}
```

F MORE EXPLANATIONS OF BIOLOGICAL TERMS

Table 4: Glossary of biological terms used in this paper, adapted for readers with a Machine Learning background.

Term	**Biological Definition**	**Context in GeneBreaker (ML Analogy)**
Homology	Similarity in sequences due to shared ancestry.	Acts as **Semantic Similarity**. We use "safe" sequences with high homology to "harmful" targets as prompts to trick the model.
Pathogenicity	The ability of an organism to cause disease or harm to a host.	The **Toxicity** or **Harmfulness** of the model output. The goal of the jailbreak is to elicit pathogenic sequences.
CDS (Coding Sequence)	The region of DNA/RNA that translates specifically into protein.	The **Executable Code**. The functional part of the sequence, distinct from non-coding regulatory regions.
BLAST	An algorithm for comparing primary biological sequence information.	The **Similarity Metric** or "Judge". We use it to calculate the Attack Success Rate (ASR) by checking if generated outputs match real viruses.
Bacteriophage	A virus that infects and replicates within bacteria, generally harmless to humans.	**Safe/Benign Data**. Often used in training DNA models (e.g., Evo). Our attack shows models can pivot from these to human viruses.
+ssRNA / -ssRNA	Positive vs. Negative-sense single-stranded RNA viruses. (+ssRNA is directly translatable).	**Data Categories**. We find +ssRNA viruses (like SARS-CoV-2) are easier to jailbreak than -ssRNA because their structure is simpler for the model to learn.
RMSD	Root-mean-square deviation; measures the average distance between atoms of proteins.	**Structural Fidelity Metric**. Measures if the 3D shape of the generated output is accurate. Low RMSD means the "jailbroken" virus might actually function.