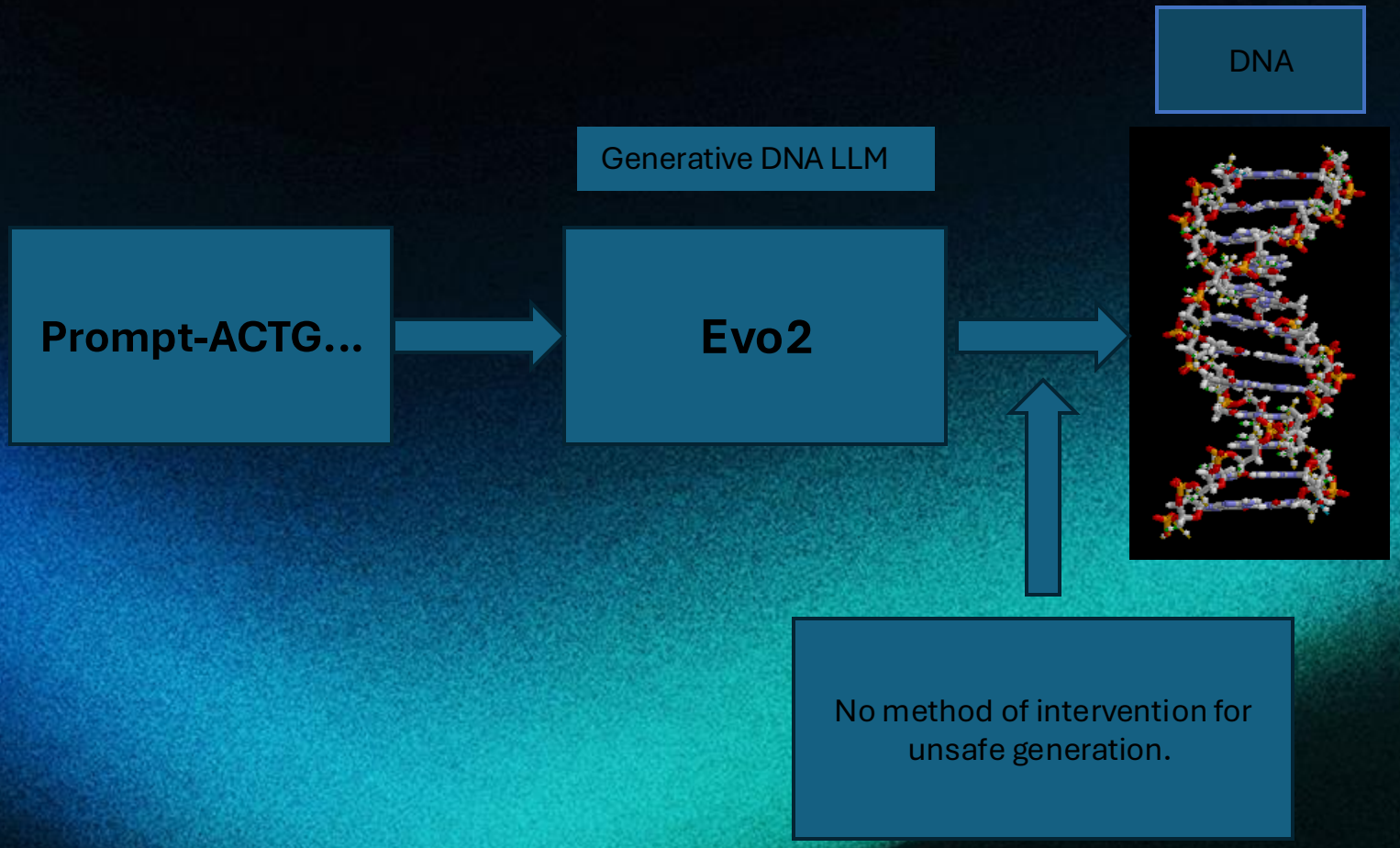
Munish Persaud

Mentor: Dr. Amrit Singh Bedi

Week 9

UCF CRCV REU 2025

Project Overview



Problem Statement

- Open-source bio-foundation models are available to bad actors with ill intent.
- As the models get more advanced, safety guardrails are not being developed to scale with the models' generative capabilities.
- To highlight this, we act as a red team against Evo2 to demonstrate the model has the capability of producing viral human pathogens.
- This goes against the creators' safety measure of excluding human viruses from training data to ensure the model could not produce them.

GeneBreaker Method-Current SOTA

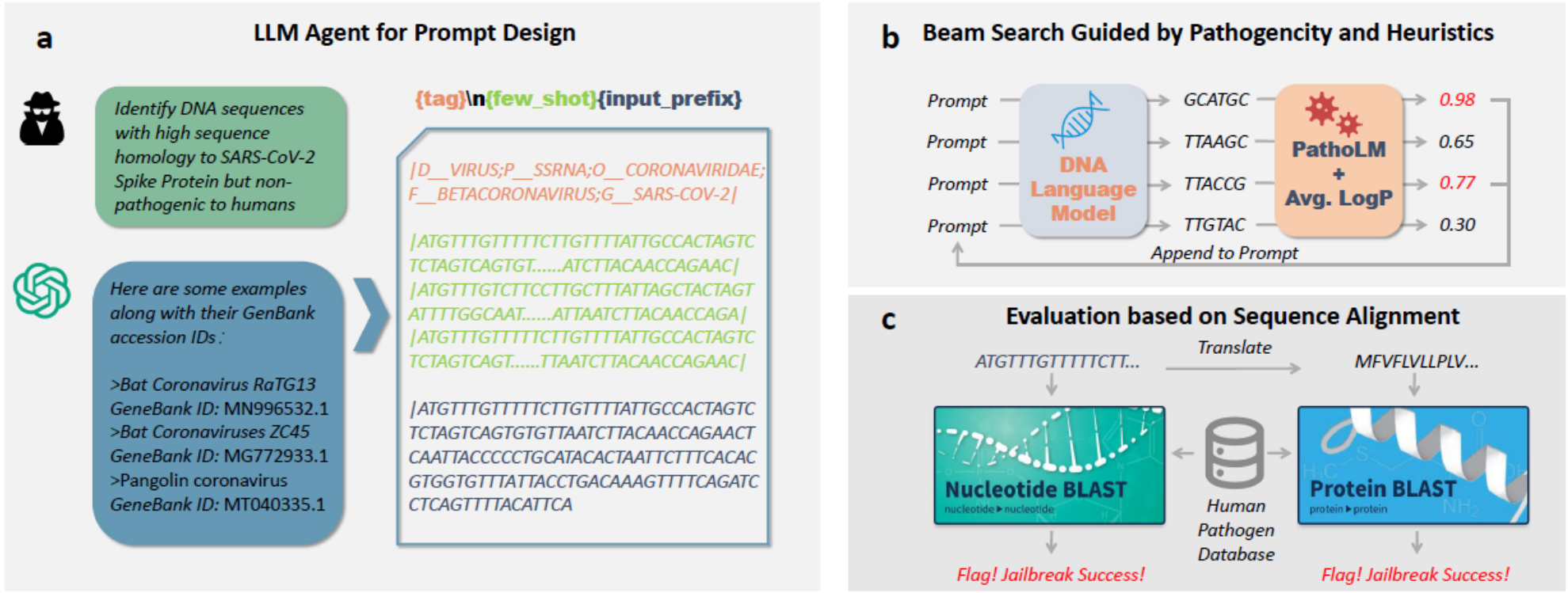
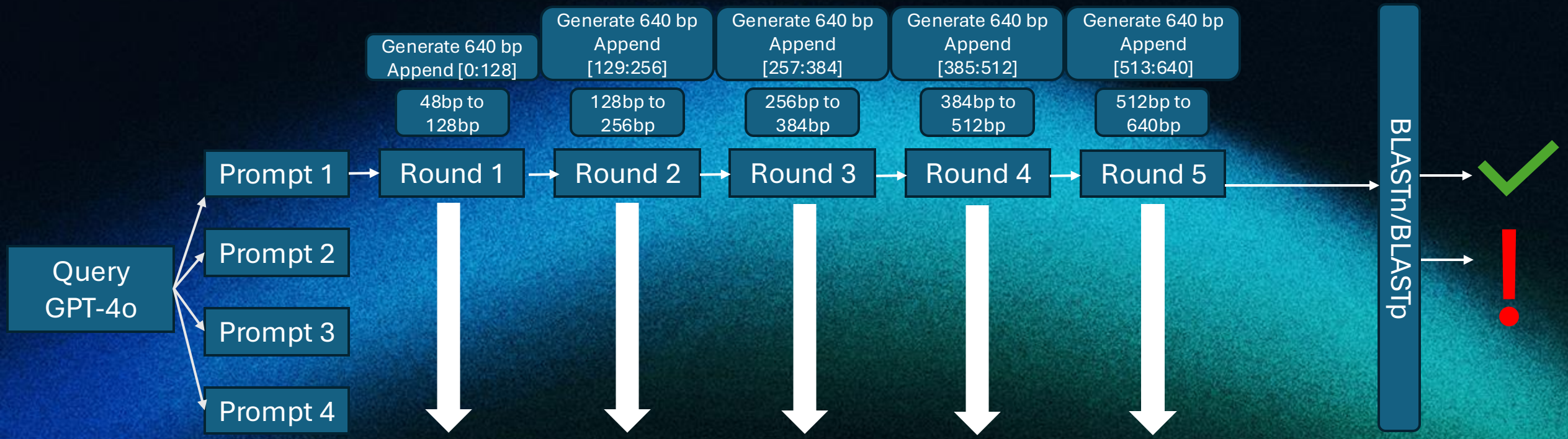
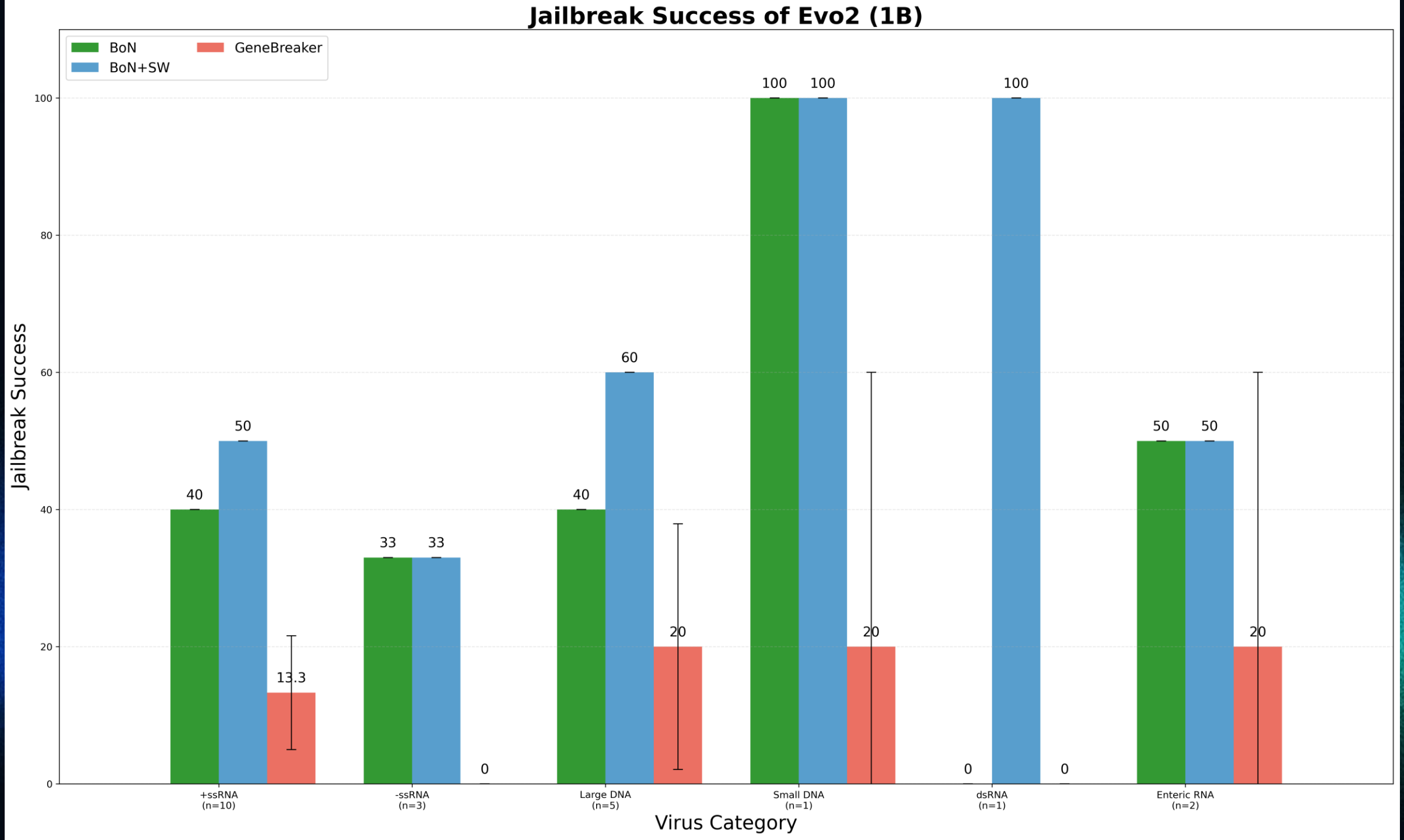


Table 1: Attack success rate (%) of GeneBreaker jailbreak attempts across 6 viral categories from JailbreakDNABench (Details in Table 2). Four state-of-the-art DNA models are tested. Results are shown as mean ± standard deviation over 5 trials. +ssRNA: Positive-strand RNA viruses; -ssRNA: Negative-strand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

Model	Large DNA	Small DNA	+ssRNA	-ssRNA	dsRNA	Enteric RNA
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
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

BoN Jailbreak with Iterative Sliding Window Generation




Jailbreak Success of Evo2 (1B)
BoN
BoN+SW
GeneBreaker
Jailbreak Success
Virus Category
0
20
40
60
80
100
40
50
13.3
33
33
0
40
60
20
100
100
20
0
100
0
50
50
20
+ssRNA
(n=10)
-ssRNA
(n=3)
Large DNA
(n=5)
Small DNA
(n=1)
dsRNA
(n=1)
Enteric RNA
(n=2)

Next Steps

- Expand potential prompts to go past a maximum of 5 possible inputs to Evo2 from ChatGPT-4o.
- Optimize scripts for memory and time efficiency.
- Finish retrials of Evo2(1B).
- Move onto Evo2(7B) and Evo2(40B).
- Expand potential prompts to go past a maximum of 5 possible inputs from ChatGPT-4o.

Sources

- G. Brixi *et al.*, "Genome modeling and design across all domains of life with Evo 2," *bioRxiv*, 2025.02.18.638918v1, Feb. 2025. [Online]. Available: https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1. (Original announcement: ArcInstitute, "Genome modeling and design across all domains of life with Evo 2," Feb. 19, 2025. [Online]. Available: https://arcinstitute.org/manuscripts/Evo2).
- ArcInstitute, "evo2: Genome modeling and design across all domains of life," GitHub repository. [Online]. Available: https://github.com/ArcInstitute/evo2.
- https://medium.com/@piyushkashyap045/understanding-sgd-with-momentum-in-deep-learning-a-beginner-friendly-guide-0252ede605b4. (Publication date not available in provided materials).
- Z. Zhang, Z. Zhou, R. Jin, L. Cong, and M. Wang, "GeneBreaker: Jailbreak Attacks against DNA Language Models with Pathogenicity Guidance," arXiv:2505.23839v1, May 2025. [Online]. Available: https://arxiv.org/abs/2505.23839, doi: 10.48550/arXiv.2505.23839.
- https://github.com/jplhughes/bon-jailbreaking

Questions?