Variant Bias In Genomic Foundation Models for Red Teaming Biological Security Screeners [1]

Henry Wong
University of Washington

With
Apart Research

Abstract

Genomic foundation models (GFMs) are increasingly used to generate pathogenic sequences for red teaming biosecurity screening systems, yet their potential biases in sequence generation remain poorly characterized. If these models systematically favor certain pathogenic variants over others, red teaming exercises could leave critical blind spots in security evaluations. We developed a systematic framework to evaluate variant bias in GFMs by analyzing their ability to generate diverse SARS-CoV-2 spike protein sequences. Using EvoDiff and Evo2, we generated 200 and 222 sequences respectively, then assessed structural quality via ESMFold pLDDT scores, taxonomic classification via Kraken2, and variant diversity through comparison with known SARS-CoV-2 lineages. Both models exhibited significant bias toward the original 2019 SARS-CoV-2 variant, with EvoDiff producing only 37 recognizable variants from 200 generations and Evo2 showing similar limitations. We discovered that Evo2's generated sequences demonstrated low perplexity scores (<30), directly contradicting published safety claims that pathogenic sequences should exhibit elevated perplexity values. These findings reveal fundamental limitations in current GFM capabilities that could systematically compromise biosecurity evaluation protocols. Our work provides the first systematic characterization of variant bias in genomic AI and highlights the urgent need for more comprehensive approaches to evaluating both model capabilities and safety mechanisms in dual-use biological AI systems..

[1] Research conducted at the AIxBio Hackathon, April 2026

1. Introduction

The rapid advancement of genomic foundation models (GFMs) has created unprecedented capabilities for generating biological sequences, raising critical questions about their potential misuse in biosecurity contexts. While these models enable beneficial applications in protein design and drug discovery, they also pose risks if used maliciously to generate pathogenic sequences that could evade biosecurity screening systems. A fundamental concern is whether current red teaming approaches adequately test the robustness of biosecurity screening software when faced with AI-generated biological threats.

This work has immediate practical implications for biosecurity policy and AI safety. If GFMs exhibit systematic biases in the types of pathogenic variants they generate, current red teaming exercises could leave significant security blind spots undetected. Understanding these biases is essential for developing more comprehensive evaluation frameworks and ensuring that screening systems can defend against the full range of AI-generated biological threats.

The threat model we address centers on the systematic evaluation failure of biosecurity screening systems. If GFMs preferentially generate certain variants over others when producing pathogenic sequences, red teaming exercises will systematically undertest screening systems against less common but equally dangerous variants, creating exploitable blind spots.

Our main contributions are:

1. A novel framework for systematically evaluating variant bias in genomic foundation models, demonstrated through comprehensive analysis of SARS-CoV-2 spike protein generation across EvoDiff and Evo2
2. Empirical evidence that both EvoDiff and Evo2 exhibit significant bias toward generating the original 2019 SARS-CoV-2 variant, with limited diversity in variant generation that could compromise red teaming effectiveness
3. Discovery of a contradiction between Evo2's claimed safety properties and observed behavior, where generated sequences show low perplexity scores despite claims that pathogenic sequences should exhibit elevated perplexity

2. Related Work

The intersection of genomic foundation models and biosecurity has become an increasingly critical research area. Wittman et al. (2023) conducted foundational work using EvoDiff to generate pathogenic sequences for red teaming biosecurity screening software, demonstrating that GFMs could produce sequences that evaded detection by existing screening tools. Their work established the viability of using AI-generated sequences for security testing but focused primarily on evasion capabilities rather than the diversity or bias of generated sequences. Similarly, Evo2's developers (Nguyen et al., 2024) conducted safety evaluations showing that their model assigns high perplexity scores to pathogenic sequences, which they interpreted as a safety feature limiting the model's ability to generate dangerous content.

Broader work in genomic AI has established the capabilities of various foundation models. ESMFold (Lin et al., 2023) and OpenFold (Ahdritz et al., 2022) have advanced protein structure prediction, while ProteinMPNN (Dauparas et al., 2022) demonstrated protein design capabilities. However, these studies have primarily focused on model performance for beneficial applications rather than potential dual-use concerns. The biosecurity community has developed various screening tools and frameworks (SecureDNA Consortium, 2023; Diggans & Leproust, 2019), but comprehensive evaluation of these systems against AI-generated sequences remains limited. Our work addresses a critical gap by systematically examining whether GFMs exhibit bias in the types of pathogenic variants they generate, which has direct implications for the robustness of red teaming exercises used to evaluate biosecurity screening systems.

3. Methods

We developed a systematic framework to evaluate variant bias in genomic foundation models by testing their ability to generate diverse pathogenic sequences. We selected EvoDiff and Evo2 as representative models and focused on SARS-CoV-2 due to its well-characterized variant diversity.We generated 200 sequences using EvoDiff and 222 sequences using Evo2, testing their ability to produce the C-terminal half (637 amino acids for EvoDiff, 1,911 nucleotides for Evo2) of the SARS-CoV-2 spike protein. The reference sequence was the original Wuhan-Hu-1 strain (NCBI: YP_009724390.1). We chose to split the protein in half rather than use masking, as many variant mutations occur in the latter portion of the spike protein. EvoDiff was run on a personal computer (AMD Ryzen 9 5950X, Nvidia 3080), while Evo2 required a rented H100 due to VRAM constraints.

For analysis, we used ESMFold to evaluate protein structure quality via pLDDT scores for EvoDiff sequences, and Kraken2 for taxonomic classification of Evo2 nucleotide sequences. Variant classification was performed by comparing generated sequences against known SARS-CoV-2 variants (Alpha, Beta, Gamma, Delta, Omicron), with sequences considered 'recognizable variants' if they showed $\geq$80% similarity to established variant profiles. Perplexity scores were extracted directly from Evo2 output and compared against the published safety threshold of 3.2 from Nguyen et al. (2024).

4. Results

In this section we describe the results of applying some of our methodology EvoDiff and Evo2. In doing so, we demonstrate the state of variant generation in both GFMs.

4.3 EvoDiff

In Figure 1, we see a distribution of our generated sequences comparatively to the actual spike protein of SARS-CoV-2. We use the pLDDT score given by ESMfold and typically other protein structure prediction models evaluate the quality of the predicted protein structure. We measure the

change in pLDDT between the original wild-type protein and the AI-generated synthetic variant. Therefore, Figure 1 shows a large number of the generations were predicted with lower confidence than the wild-type. However, by only a small amount (> 0.05) drawing conclusions EvoDiff generated sequences similar to the original wild-type. We do see some distribution among the generations . However, we consider these scores to be small enough to conclude EvoDiff did generate proteins similar enough to the original wild-type with little to no variation.

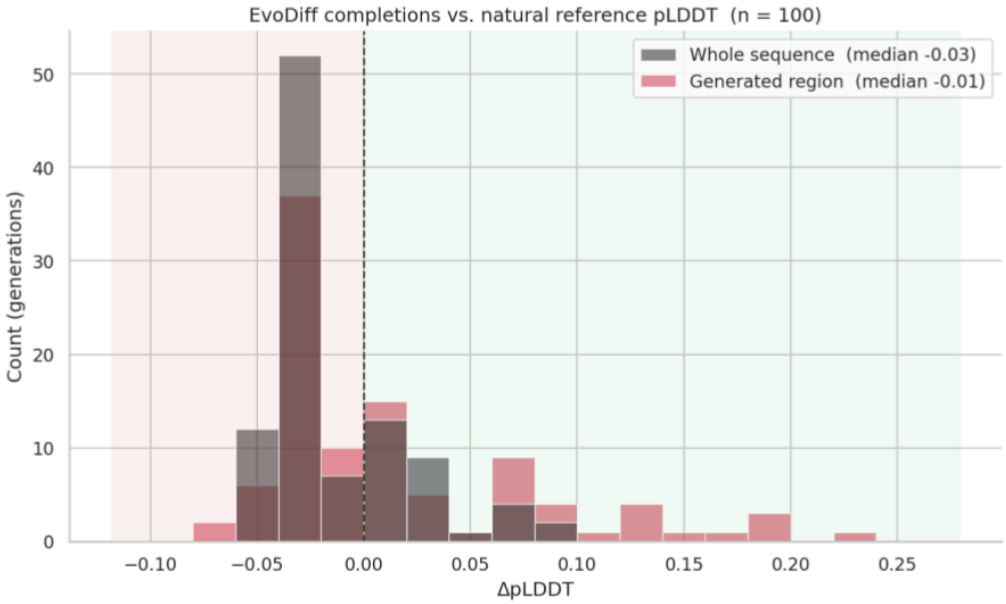

Figure 1. Distribution of ESMFold mean pLDDT for EvoDiff completions, relative to the natural reference (dashed line at 0). Whole-sequence (grey) and generated-region-only (red) values are shown.

In Table 1, we also look at whether any of the synthetic homologs had generated a SARS-CoV2 variation. When looking at just the spike protein, we see some recognizable variants but not a proportionally large amount. Out of 200 generated proteins, only 37 were recognizable. Each generation we tested looked incredibly similar to the original wild-type. However, when EvoDiff is given the entire genome of SARS-CoV-2 without the spike protein, it struggles to complete a recognizable protein. Showing that EvoDiff is less accurate when asked to generate longer sequences.

Variant	Mutation Count	% of Synthetic Homologs
Omicron BA.1	10	5%
Gamma (P.1)	5	5%
Beta (B.1.351)	5	5%
Omicron BA.5	5	5%
Alpha (B.1.1.7)	4	2%
Delta (B.1.617.2)	4	2%
XBB.1.5	4	2%
Unidentified	163	81.5%

Table 1. Identified possible SARS-Cov2 variants in EvoDiff generated synthetic homologs.

4.2 Evo2

We briefly test Evo2 in this study by splitting the SARS-CoV2 spike protein in half and asking Evo2 to predict the second half. Instead of the protein, we looked at nucleotides. So instead of using ESMFold like with EvoDiff, we used Kraken2 to identify what was generated. We used Kraken2 due to its speed and database accessibility as we were running out time. Evo2 was used to generate 222 synthetic homologs. We would like to note, in the bias and safety study done by Evo2, they claim that the model was specifically not trained on pathogenic data and thus does not perform well on it. We ran all of our experiments on Evo2 40B on a single H100.

Kraken2 was able to classify 46.4% of the synthetic homologs but only 22.3% of those classified as SARS-CoV2. And only as the 2019 variant, which is the variant we gave Evo2. Of the classified 71.8% were classified as homo sapiens. Overall, we find that Evo2 has trouble generating SARS-CoV2. However, in Evo2, you can specify taxonomies to help steer Evo2 in the right direction. We did not add directions when generating our homologs but we leave that to further studies.

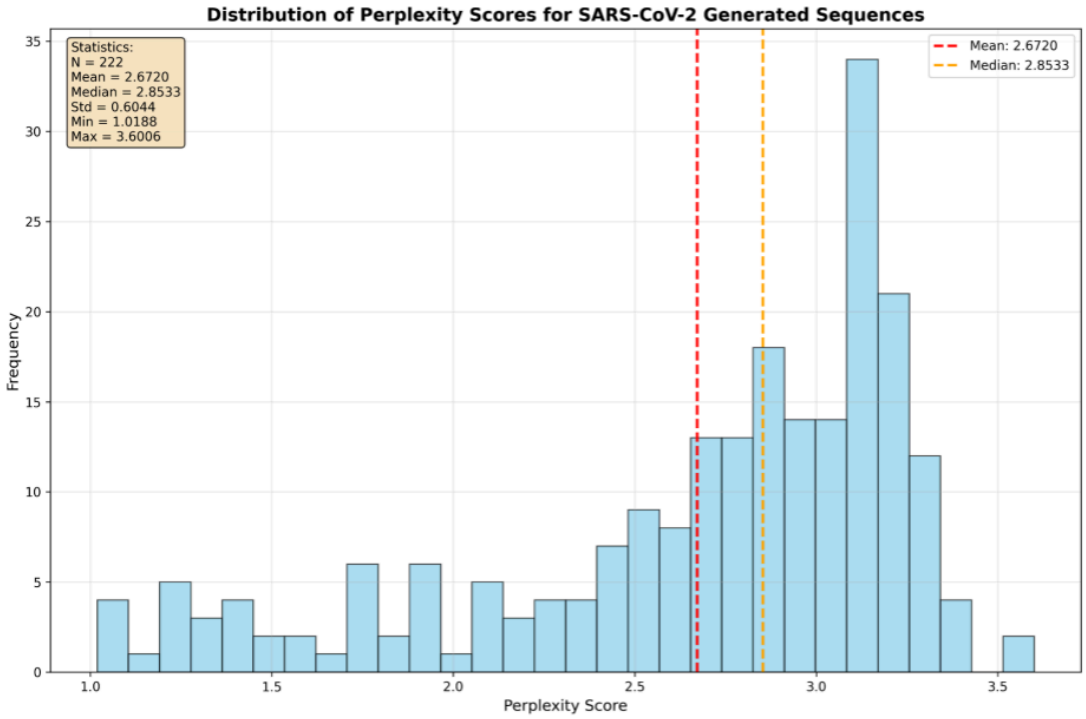

Figure 2. Distribution of perplexity scores for SARS-CoV-2 generated sequences.

In the Evo2 bias and safety study, the authors claim that Evo2 assigns a high perplexity to eukaryotic-infecting viruses. However, when running out experiments, we noticed the perplexity scores. The figure in the Evo2 work (Extended Data Fig.2) shows eukaryotic hosts (pathogenic) having a minimum perplexity score of approximately 3.2. While Figure 2 shows that a large amount of the synthetic homologs generated have a perplexity score of less than 3.2. This contradicts their claims as SARS-CoV2 is a eukaryotic-infecting virus, however we believe further testing is needed as we believe our sample size is too small, and our methodology could be improved.

5. Discussion and Limitations

Our findings reveal critical gaps in current genomic foundation models that have significant implications for biosecurity. The observed bias toward generating only the 2019 SARS-CoV-2 variant suggests fundamental limitations in the diversity of pathogenic sequences these models can produce, potentially creating systematic blind spots in security evaluations. Most concerning is our discovery that Evo2's actual behavior contradicts its claimed safety properties, with generated sequences showing low perplexity scores despite published claims that pathogenic sequences should exhibit elevated perplexity. This discrepancy suggests that current safety evaluation frameworks for genomic AI may be insufficient.

Limitations

Our study focused exclusively on SARS-CoV-2 and examined only two genomic foundation models, limiting generalizability. The hackathon timeframe prevented more sophisticated evaluation metrics and larger sample sizes. Our approach of splitting rather than masking the spike protein may have influenced generation patterns, and we assumed that low pLDDT scores indicate poor protein quality, which may not hold for all pathogenic mechanisms.

Future Work

Next steps include expanding to pathogens from the US Select Agents list, implementing phylogenetic analysis, and exploring guided generation techniques. Developing standardized benchmarks for evaluating GFM bias in biosecurity contexts would benefit the broader research community. Most urgently, investigating potential mitigation strategies and collaborating with biosecurity screening developers to test whether observed biases translate to actual security vulnerabilities.

We also wish to expand bias work beyond BSS and examine ancestry bias. Evo2 previously has done a brief study on ancestry bias and found that Evo2 is no more biased than other variant effect predictors, however there is still a heavy bias towards European ancestry. We hope to establish a benchmark and standards for GFMs such that when GFMs are used for vaccine or drug discovery, certain populations are not left out.

6. Conclusion

This study reveals fundamental biases in genomic foundation models that have critical implications for biosecurity evaluation. Our analysis of EvoDiff and Evo2 demonstrates that both models exhibit strong bias toward generating the original 2019 SARS-CoV-2 variant, with limited diversity in pathogenic sequence generation. This bias pattern could systematically compromise red teaming exercises, leaving biosecurity screening systems untested against important variant classes that represent genuine security threats.

Perhaps most concerning is our discovery that Evo2's actual behavior contradicts its claimed safety properties, with generated sequences showing low perplexity scores despite published claims that pathogenic sequences should exhibit elevated perplexity. This finding suggests that current safety evaluation frameworks for genomic AI may be insufficient and highlights the urgent need for more rigorous, systematic approaches to evaluating both the capabilities and limitations of these powerful models. As genomic AI continues to advance, ensuring comprehensive and unbiased evaluation of biosecurity systems becomes essential for maintaining effective defenses against emerging biological threats.

Code and Data

We refrain from sharing any code right now due to most of the code being used to generate parts of SARS-CoV2, a pathogen. However, artifacts are available upon request.

References

1. Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xie, Q., Albrecht, B., ... & AlQuraishi, M. (2022). OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Nature Methods, 19(12), 1551-1556. https://doi.org/10.1038/s41592-022-01691-3
2. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., ... & Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science, 378(6615), 49-56. https://doi.org/10.1126/science.add2187
3. Diggans, J., & Leproust, E. (2019). Next steps for access to safe, secure DNA synthesis. Frontiers in Bioengineering and Biotechnology, 7, 86. https://doi.org/10.3389/fbioe.2019.00086
4. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130. https://doi.org/10.1126/science.ade2574
5. Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., ... & Massaroli, S. (2024). Evo: DNA foundation modeling from molecular to genome scale. bioRxiv. https://doi.org/10.1101/2024.02.27.582234
6. SecureDNA Consortium. (2023). Securing DNA synthesis with cryptographic approaches. Nature Biotechnology, 41(8), 1057-1060. https://doi.org/10.1038/s41587-023-01834-x
7. Wittman, M., Chong, S., Labianca, S., Yunez-Londono, C., McLaren, P., Gemmell, N.,&Maus, J. (2023). Evaluating genomic foundation models for pathogen identification and biosecurity screening. arXiv preprint arXiv:2311.00939. https://doi.org/10.48550/arXiv.2311.00939

LLM Usage Statement

Cladue was used in debugging code but not in development. LLM's were also used to generate python code to generate Figure 2, but all data analysis was done by hand. All ideation and brainstorming was done by hand. For writing Claude was used for spell check.

Limitations and Dual-Use Considerations

Research involving AI-generated pathogenic sequences requires careful consideration of potential limitations, risks, and ethical implications. We address these concerns transparently to support responsible advancement of biosecurity research.

Limitations

False Positives/Negatives: Our variant classification system may misidentify sequences due to the 80% similarity threshold, potentially missing subtle but significant mutations or incorrectly flagging benign variations. The pLDDT structural quality assessment may not capture all pathogenic mechanisms, particularly those involving protein-protein interactions or conformational changes not reflected in static structure prediction.

Edge Cases and Scalability: Our framework currently handles only single-pathogen analysis and may not scale to multi-pathogen scenarios or complex biological systems. The computational requirements for large-scale analysis (particularly Evo2's H100 requirement) limit accessibility. Our approach may not capture emergent properties that arise from pathogen interactions with host systems or environmental factors.

Generalizability Constraints: Results from SARS-CoV-2 analysis may not generalize to other pathogen classes, particularly those with different evolutionary pressures, structural constraints, or mechanisms of action. The bias patterns we observed could be pathogen-specific or model-specific rather than representative of broader GFM limitations.

Dual-Use Risks

Security Vulnerability Disclosure: Our findings reveal systematic weaknesses in current biosecurity evaluation approaches that could be exploited by malicious actors. The discovery that GFMs exhibit predictable biases could enable adversaries to develop generation strategies that specifically target these blind spots.

Methodology Misuse: The systematic framework we developed could potentially be used to optimize pathogenic sequence generation for evasion rather than defensive evaluation. The identification of specific model limitations could guide adversarial approaches to circumvent safety mechanisms.

Information Hazard Considerations: While we avoid publishing generated sequences, our analytical approach and findings about model biases constitute information that could facilitate misuse. The contradiction we identified in Evo2's safety mechanisms could undermine confidence in current safeguards. However, we are not aware of any institution that uses these safeguards.

Responsible Disclosure Recommendations

We have followed responsible disclosure principles throughout this research:

• Pre-publication notification: We recommend notifying the developers of EvoDiff and Evo2 of our findings before broader publication, particularly regarding the perplexity safety mechanism discrepancy

• Controlled access: Generated sequences and detailed analysis code should be made available only to verified researchers with legitimate biosecurity research purposes

• Coordinated improvement: Engage with biosecurity screening software developers to develop mitigation strategies before public disclosure of specific vulnerabilities

Ethical Considerations

Benefit-Risk Assessment: We conducted this research with the primary goal of strengthening biosecurity defenses. The potential benefits of identifying systematic weaknesses in current evaluation approaches outweigh the risks of disclosure, provided appropriate safeguards are maintained.

Minimal Necessary Disclosure: We disclose only the information necessary to demonstrate the existence and significance of the bias problem without providing detailed instructions for exploitation. Specific generated sequences are not included in this publication.

Research Governance: This research was conducted with awareness of dual-use research of concern (DURC) principles. We recommend that follow-up studies involve institutional review and oversight appropriate to the sensitivity of the research.

Suggestions for Future Improvements

Enhanced Safety Mechanisms: GFM developers should implement more robust safety evaluation frameworks that go beyond single metrics like perplexity. Multi-faceted safety assessments including structural analysis, functional prediction, and evolutionary plausibility could provide more reliable safeguards.

Bias Mitigation Strategies: Techniques such as adversarial training, diverse data augmentation, or ensemble methods could help reduce systematic biases in pathogenic sequence generation. Regular bias auditing should become standard practice for GFMs used in security contexts.

Standardized Evaluation Frameworks: The biosecurity community should develop standardized benchmarks for evaluating both the capabilities and limitations of GFMs in security-relevant contexts. This includes diverse pathogen panels, variant coverage metrics, and bias detection protocols.

Collaborative Security Research: Establishing secure research environments where biosecurity researchers can safely conduct red teaming exercises without risk of information leakage or misuse. This includes developing protocols for responsible sharing of vulnerability discoveries and coordinated disclosure timelines.