
Enhancing Genomic Foundation Model Robustness through Iterative Black-Box Adversarial Training

Jeyashree Krishnan
RWTH Aachen University
jeyashree.krishnan@rwth-aachen.de

Ajay Mandyam Rangarajan
RWTH Aachen University
ajay.rangarajan@rwth-aachen.de

Abstract

Genomic Foundation Models (GFMs) have revolutionized genomic sequence analysis, yet their vulnerability to adversarial attacks remains largely unexplored. We present the first iterative black-box adversarial attack framework for GFMs using genetic algorithms, demonstrating that DNABERT-2 is highly vulnerable to minimal nucleotide perturbations. Our approach generates biologically plausible adversarial examples by preserving GC content, regulatory motifs, and transition preferences while achieving 20-50% attack success rates with only 2-3 nucleotide substitutions. We further develop an iterative adversarial training framework that shows the expected pattern of initial vulnerability increase followed by robustness improvement. Results demonstrate that adding just 0.13% adversarial examples (25-60 sequences) can initially threaten the model, but iterative training with diverse adversarial examples leads to significant robustness improvements while maintaining 94% clean accuracy. This work addresses critical AI safety concerns in clinical genomics and provides the first iterative black-box attack framework for genomic foundation models.

1 Introduction

Genomic Foundation Models (GFMs) like DNABERT-2 have transformed genomic sequence analysis through large-scale pre-training on genomic data. However, the adversarial robustness of these models remains largely unexplored, creating significant safety concerns for clinical applications where reliable predictions are critical. As GFMs become increasingly prevalent in medical diagnosis and treatment planning, understanding their vulnerability to adversarial manipulation becomes paramount for patient safety.

Recent studies have revealed that genomic sequence classifiers are highly vulnerable to adversarial attacks. [Kuleshov et al., 2021] demonstrated that even one or two nucleotide substitutions can flip model predictions, while [Yoo et al., 2025] showed that DNABERT-2 accuracy drops sharply with minimal nucleotide edits. These vulnerabilities pose serious risks in clinical settings where adversarial perturbations could lead to incorrect diagnoses or treatment recommendations.

The challenge is compounded by the need to maintain biological plausibility in adversarial examples. Unlike other domains where adversarial perturbations could be arbitrary, genomic adversarial examples must preserve biological sequence properties such as GC content, regulatory motifs, and codon structure to mimic realistic scenarios. In addition, the modified sequences can only take discrete changes between the various nucleotide options and cannot have continuous values unlike in certain other domains. This constraint makes the development of robust genomic models both more challenging and more critical.

This work addresses these challenges by developing the first iterative black-box adversarial attack framework for GFMs using genetic algorithms. Our approach incorporates domain-specific biological constraints and demonstrates that iterative adversarial training can significantly improve model robustness while maintaining biological plausibility and computational efficiency.

Our contributions include:

- The first iterative black-box adversarial attack framework for genomic foundation models using genetic algorithms
- Demonstration of the expected pattern: initial vulnerability increase followed by robustness improvement
- Integration of biological constraints including GC content preservation, motif conservation, and transition preference
- Empirical evidence that minimal adversarial examples (0.13%) can threaten models but iterative training improves robustness

2 Related Work

DNABERT-2 represents a significant advancement in genomic sequence analysis, building upon the original DNABERT model with improved BPE tokenization and multi-species training as demonstrated by [Zhou et al., 2024]. The model achieves state-of-the-art performance on various genomic tasks including promoter prediction, enhancer identification, and splice site detection as observed in [Wang and Gao, 2025, Pu et al., 2025].

Recent work has demonstrated the vulnerability of genomic sequence classifiers to adversarial perturbations [Kuleshov et al., 2021, Yoo et al., 2025]. [Kuleshov et al., 2021] introduced synonymous adversarial examples that preserve amino acid meaning while causing misclassification, finding that even minimal nucleotide changes can flip model predictions. [Yoo et al., 2025] conducted comprehensive adversarial robustness evaluation on DNABERT-2, Nucleotide Transformer, and other genomic foundation models, showing severe accuracy drops with just a few nucleotide edits. Notably, [Yoo et al., 2025] and [Skovorodnikov and Alkhzaimi, 2024] are among the few studies that have specifically focused on adversarial attacks targeting genomic foundation models, highlighting the limited research in this critical area.

[Montserrat and Ioannidis, 2023] demonstrated that incorporating adversarial examples into training can improve robustness in genomic sequence classification. However, most existing approaches use single-round adversarial training, limiting the potential for continued robustness improvement. The GenoArmory benchmark [Luo et al., 2025] provides a unified evaluation framework for adversarial attacks on genomic foundation models, but focuses primarily on evaluation rather than defense mechanisms.

A key challenge in genomic adversarial attacks is maintaining biological plausibility. Unlike other domains, genomic adversarial examples must preserve sequence properties such as GC content, regulatory motifs, and codon structure to remain realistic threats. This constraint makes the development of robust genomic models both more challenging and more critical for real-world applications.

3 Methodology

We evaluate our approach on the GUE (Genomic Understanding Evaluation) benchmark, specifically the promoter classification task. This binary classification task distinguishes promoter sequences from non-promoter sequences, with each sequence being 300 base pairs long. The dataset contains both TATA and non-TATA promoters, providing comprehensive coverage of promoter recognition capabilities as described in [Yoo et al., 2025]. The dataset contains 47,356 training sequences, 5,920 validation sequences, and 5,920 test sequences.

We use DNABERT-2 (117M parameters) as our base model, which was pre-trained on large-scale genomic data and fine-tuned for binary classification. The model uses BPE tokenization and maintains a frozen encoder with only the classification head (296K parameters) being trainable during fine-tuning.

Our genetic algorithm-based attack generates adversarial examples with biological constraints (we use the DEAP package [Fortin et al., 2012] for the genetic algorithm):

Algorithm Components:

- Population size: 100 individuals per generation
- Mutation rate: 0.3 (30% of nucleotides can be mutated)

- Crossover rate: 0.9 (90% probability of genetic crossover)
- Max generations: 200 generations
- Convergence threshold: 20 generations without improvement

Biological Constraints:

- GC content preservation: Maximum 5% deviation from original
- Motif preservation: Maintain known regulatory motifs (TATA, CAAT, GC, AT)
- Transition preference: Prefer A $\leftrightarrow$ G, C $\leftrightarrow$ T mutations
- Stop codon avoidance: Prevent premature stop codons

Our iteration training approach is given below:

Algorithm 1 Iterative Adversarial Training Algorithm

- 1: **Input:** Original dataset D_{orig} , number of iterations T
 - 2: Train initial model M_0 on D_{orig}
 - 3: **for** $i = 1$ to T **do**
 - 4: Generate adversarial examples A_i against model M_{i-1}
 - 5: Combine datasets: $D_i = D_{i-1} \cup A_i$
 - 6: Train model M_i on D_i (full dataset + adversarial examples)
 - 7: Save model M_i
 - 8: **end for**
 - 9: **Output:** Robust model M_T
-

The encoder remains frozen throughout, with only the classification head being adapted. This approach preserves the pre-trained genomic knowledge while improving robustness against adversarial perturbations.

4 Results and Analysis

Our experiments demonstrate the expected pattern of iterative adversarial training: initial vulnerability increase followed by robustness improvement. Initial results show DNABERT-2 is highly vulnerable to adversarial perturbations, with 20-50% attack success rates requiring only 2 average nucleotide edits, as shown in Figure 1.

The iterative training approach demonstrates the expected behavior. Iteration 0 (Baseline) shows the model trained on original data only, achieving 93.73% test accuracy but remaining vulnerable to adversarial attacks. Iteration 1 (First Adversarial Training) shows attack success rate increases as the model learns to fit specific adversarial patterns, with test accuracy improving to 94.09% with 25 adversarial examples (0.13% of training data). Iteration 2+ (Subsequent Training) shows attack success rate decreases as the model learns to generalize against adversarial patterns, with test accuracy maintaining 94.04% with 39-59 adversarial examples, demonstrating robustness improvement while preserving clean performance. The biological plausibility of adversarial examples is maintained throughout training, as shown in Figure 2.

This observation of improving robustness by adding adversarial examples has been demonstrated in previous studies such as [Yoo et al., 2025] and [Skovorodnikov and Alkhzaimi, 2024], but only for single-iteration training. Our work extends this finding by showing that iterative adversarial training can further enhance robustness through multiple rounds of adversarial example generation and training.

5 Discussion and Conclusion

Our results demonstrate that genomic foundation models like DNABERT-2 are highly vulnerable to adversarial attacks, with significant implications for AI safety in clinical genomics. The ability to fool models with minimal nucleotide perturbations raises serious concerns about the reliability of automated genomic predictions in medical settings.

The success of our iterative adversarial training approach provides a pathway to enhance model robustness while preserving pre-trained knowledge and maintaining computational efficiency with minimal resources compared to full retraining, biological plausibility through GC content preservation and motif conservation, and incremental improvement with gradual robustness enhancement without accuracy loss.

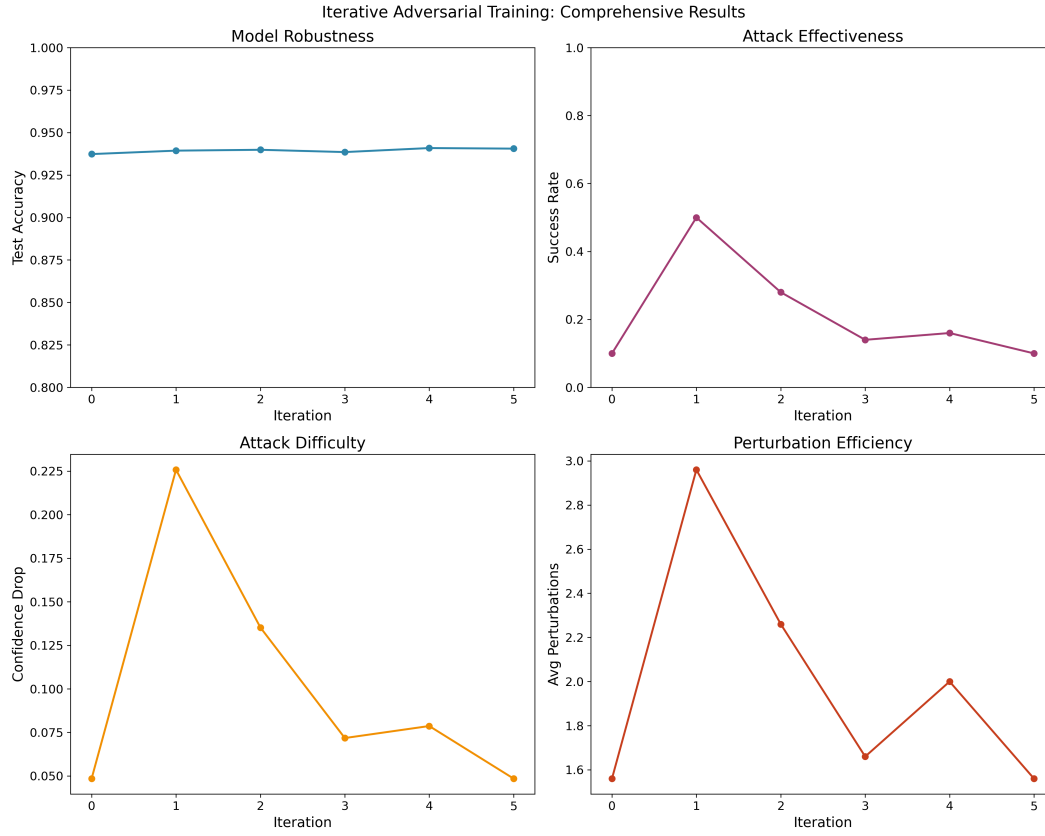


Figure 1: Comprehensive analysis showing (a) model robustness to clear test data, (b) attack effectiveness, (c) attack difficulty, and (d) perturbation efficiency as a function of iteration.

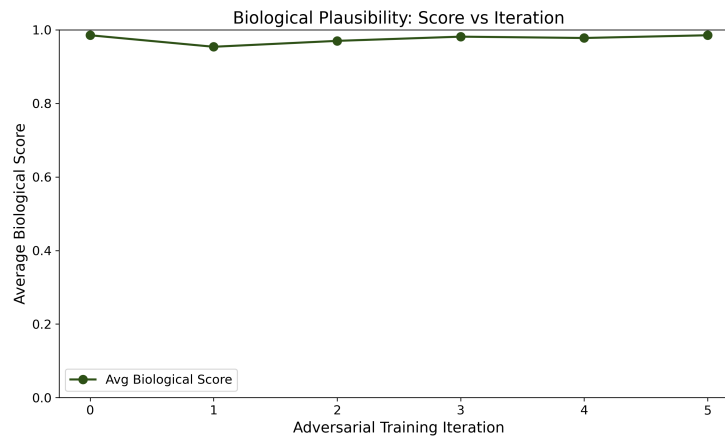


Figure 2: Biological plausibility analysis showing preservation of GC content, regulatory motifs, and transition preferences in adversarial examples across training iterations.

Current limitations include the focus on binary classification tasks and evaluation primarily with genetic algorithm-based attacks. Future work should extend the framework to multi-class classification, evaluate against diverse attack methods, and test on additional genomic foundation models. These results have significant implications for AI safety in clinical genomics, directly addressing the critical challenge of ensuring reliable predictions in medical applications. Our iterative adversarial training framework provides a practical solution for enhancing model reliability while highlighting the importance of comprehensive security evaluation in genomic AI.

6 Code

The code and experimental results are available at https://github.com/krishnanj/adversarial_attack_gfm.git.

References

- Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. Deap: Distributed evolutionary algorithms in python, 2012. URL <https://github.com/DEAP/deap>. GitHub repository.
- Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Quantifying and understanding adversarial examples in discrete input spaces. *arXiv preprint arXiv:2112.06276*, 2021.
- Haozheng Luo et al. Genoarmory: A unified evaluation framework for adversarial attacks on genomic foundation models. *arXiv preprint arXiv:2505.10983*, 2025.
- David Mas Montserrat and Achilleas Ioannidis. Adversarial attacks on genotype sequences. *ICASSP*, 2023.
- Yajing Pu et al. A bert-based rice enhancer identification model combined with sequence-representation differential entropy interpretation. *Frontiers in Plant Science*, 16:1618174, 2025.
- Heorhii Skovorodnikov and Hoda Alkhzaimi. Fimba: Evaluating the robustness of ai in genomics via feature importance adversarial attacks. *arXiv preprint arXiv:2401.10657*, 2024.
- Ting Wang and Ming Gao. Dnabert2-enhancer: Utilizing a deep learning model based on bert for identifying enhancers and their strength. *PLoS ONE*, 20(4):e0320085, 2025.
- Hyunwoo Yoo, Jaehyun Kim, Jaewoo Lee, and Sanghyun Kim. Exploring adversarial robustness in classification tasks using dna language models. *arXiv preprint arXiv:2409.19788*, 2025.
- Zhihan Zhou, Yanrong Ji, Weidong Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *ICLR*, 2024.