# Exploring Adversarial Robustness in Classification tasks using DNA Language Models

Anonymous ACL submission


## Abstract

DNA Language Models, such as GROVER, DNABERT2 and the Nucleotide Transformer, operate on DNA sequences that inherently contain sequencing errors, mutations, and laboratory-induced noise, which may significantly impact model performance. Despite the importance of this issue, the robustness of DNA language models remains largely underexplored. In this paper, we comprehensivly investigate their robustness in DNA classification by applying various adversarial attack strategies: the character (nucleotide substitutions), word (codon modifications), and sentence levels (back-translation-based transformations) to systematically analyze model vulnerabilities. Our results demonstrate that DNA language models are highly susceptible to adversarial attacks, leading to significant performance degradation. Furthermore, we explore adversarial training method as a defense mechanism, which enhances both robustness and classification accuracy. This study highlights the limitations of DNA language models and underscores the necessity of robustness in bioinformatics.


## 1 Introduction

Transformer-based language models are increasingly being adopted in bioinformatics, leveraging NLP techniques to tackle sequence classification and functional prediction tasks, which have traditionally relied on alignment-based methods(Steinegger and Söding, 2017; Buchfink et al., 2021). Notably, language models such as DNABERT2 (Zhou et al., 2023), Nucleotide Transformer (Dalla-Torre et al., 2023), and GROVER (Sanabria et al., 2024) leverage large-scale genomic sequence data as textual data and are specialized for specific bioinformatics downstream tasks. By treating DNA sequences not merely as strings but as sequence data with contextual information, these models introduce a novel approach to solving bioinformatics problems.

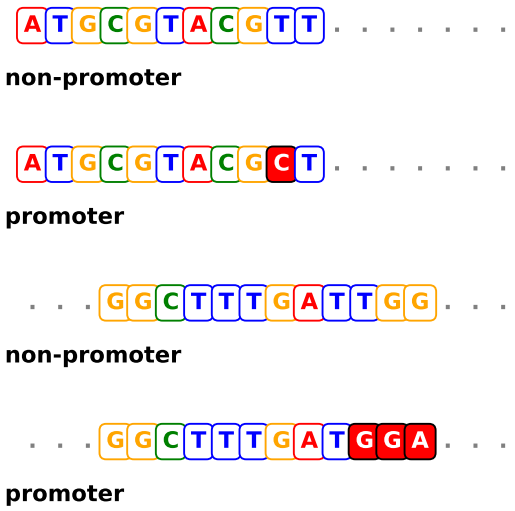


Figure 1: Adversarial examples with nucleotides and codon change, which will be misclassified by a non-promoter to a promoter.

However, despite these advancements, the robustness of DNA language models remains underexplored. In real-world biological environments, DNA sequences are susceptible to sequencing errors, mutations, and data noise introduced during the extraction process in laboratories (Ono et al., 2021; Ma et al., 2019). However, there is a lack of systematic research analyzing the impact of such variations on model performance. While robustness studies on language models have been active in text classification, research on DNA sequence classification models remains limited. Given the increasing adoption of DNA classification models in clinical and biotechnological applications, it is crucial to assess their reliability in real-world scenarios.

In this paper, we explore the adversarial robustness of DNA language models in classification tasks. In addition to existing character-level attacks (nucleotide-level attacks (Kuleshov

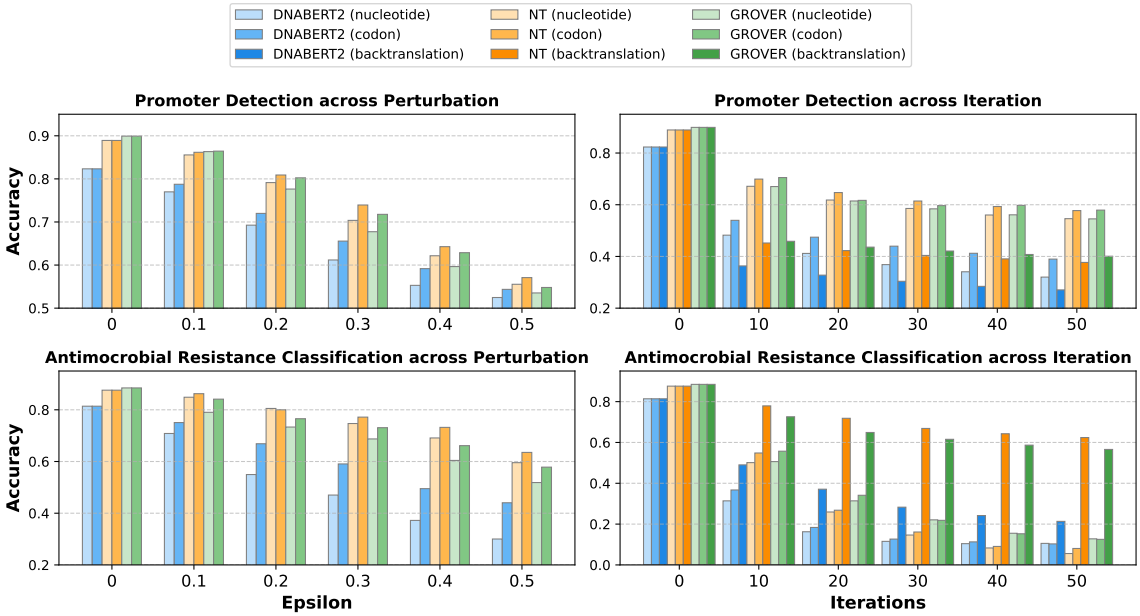


Figure 2: Anti-Microbial Resistance (AMR) gene Classification and Promoter Detection Results

et al., 2021)), we introduce word-level (codon-level) and sentence-level (backtranslation) attacks. We demonstrate that these attacks lead to significant performance degradation in Anti-Microbial Resistance (AMR) gene classification (Yoo et al., 2024) and promoter detection (Zhou et al., 2023).

Furthermore, we investigate whether a simple adversarial training strategy can mitigate the vulnerabilities of DNA language models. The results show that while its effectiveness varies by attack type and dataset, it enhances robustness and can occasionally improve overall model performance.

Our main contributions can be summarized into three folds:

- We comprehensively explore the robustness of DNA language models through multi-granularity attack levels: **character-level (nucleotide-level)**, **word-level (codon-level)**, and **sentence-level (back-translation)**. Notably, we introduce **codon-level** and **back-translation** attacks, extending the scope beyond conventional character-level perturbations for a more thorough robustness evaluation.

- We demonstrate that increasing perturbation strength causes the most severe performance degradation in nucleotide-level perturbations,thought at the risk of disrupting biological context. In the AMR gene classification with drug labels task, our proposed back-translation-based perturbations effectively degrade classification performance while preserving semantic meaning.

- We show that even simple adversarial training effectively mitigates vulnerabilities and improves classification accuracy over the original model, highlighting its potential to enhance the real-world reliability of DNA classification models.

## 2 Related Work

**Adversarial Attacks in Text Classification** Adversarial attacks in text classification can be broadly categorized into character-level, word-level, and sentence-level attacks (Ebrahimi et al., 2018b). Character-level attacks involve random substitutions at the character level, or modifications at the word embedding or language model level (Jia and Liang, 2017; Alzantot et al., 2018). Word-level attacks follow a similar approach. Sentence-level attacks include paraphrasing, backtranslation (Ribeiro et al., 2018), and sentence embedding-level perturbations. (Zhang et al., 2019).

**Adversarial Attacks in DNA Sequence Classification** Kuleshov et al. (2021) and Montserrat and Ioannidis (2022) demonstrated that nucleotide-level adversarial attacks, including perturbations on single nucleotide polymorphism (SNP)-based ancestry classification models, significantly degrade the performance of DNA sequence classifiers. In

this paper, inspired by word-level and sentence-level perturbations, we further explore codon-level attacks and backtranslation-based DNA attacks.

**Adversarial Training** Several studies have shown that adversarial training enhances model robustness (Madry et al., 2018; Zhang et al., 2022; Liu and Sun, 2023), including in DNA sequence classification, where character-level adversarial examples are used to improve resilience (Kuleshov et al., 2021; Montserrat and Ioannidis, 2022).

## 3 Adversarial Attack on DNA Language Models

We examine the impact of adversarial attacks on DNABERT2 (Zhou et al., 2023), Nucleotide Transformer (Dalla-Torre et al., 2023), and GROVER (Sanabria et al., 2024) across three levels attack strategies: character-level (nucleotide substitutions (Kuleshov et al., 2021)), word-level (codon modifications), and sentence-level (backtranslation-based transformations). Our analysis focuses on DNA classification task using AMR gene (Yoo et al., 2024) and promoter detection (Zhou et al., 2023). Further implementation details and benchmark details are provided in Appendix A and Appendix B.

**Nucleotide-Level Attack** Nucleotides are the fundamental building blocks of nucleic acids, such as DNA and RNA, crucial for storing and transmitting genetic information. In DNA sequences, nucleotides are represented by single characters, making nucleotide-level attacks analogous to character-level attacks in text. We conducted nucleotide-level attacks using an enumeration approach combined with trial and error search methods (Iyyer et al., 2018; Ribeiro et al., 2018; Belinkov and Bisk, 2018).

**Codon-Level Attack** A codon is a sequence of three nucleotides in DNA or RNA that specifies a particular amino acid. Codons are essential for protein synthesis. Codon-level attack is similar to word-level attacks and is also performed using the enumeration search method including trial and error.

**Backtranslation Attack** We introduce backtranslation based attack method for DNA sequences. In biological translation, translation refers to converting mRNA sequences into protein amino acid sequences, and reverse translation refers to generating possible nucleotide sequences based on protein amino acid sequences.

## 4 Results

### 4.1 Attack Effectiveness and Context Preservation

In Figure 2, nucleotide-level attacks (Kuleshov et al., 2021) result in the most significant performance degradation. While these attacks are highly effective, they also pose a considerable risk of altering the DNA sequence's meaning, as discussed in Appendix D. Conversely, codon-level attacks are better at preserving sequence meaning since they group sequences based on meaningful units, such as codons. Backtranslation attacks preserve context the best but are relatively less effective in reducing accuracy. This trend is evident in the first graph, where there is a sharp decline in model accuracy with increasing epsilon values for nucleotide-level attacks, particularly affecting DNABERT2 (Zhou et al., 2023). In contrast, the Nucleotide Transformer maintaines relatively high accuracy even as epsilon increased.

### 4.2 Model Robustness Comparison

The Nucleotide Transformer (Dalla-Torre et al., 2023) and GROVER (Sanabria et al., 2024) demonstrate superior robustness compared to DNABERT2 (Zhou et al., 2023). Due to its larger model capacity and ability to handle longer sequences, the Nucleotide Transformer exhibits increased resistance to adversarial attacks. Compared to DNABERT2, GROVER employs an extensive BPE optimization process involving hundreds of iterations and incorporates an LSTM architecture, enhancing its ability to model long-range dependencies in DNA sequences. These architectural refinements are likely to improve its robustness against adversarial attacks. This robustness is particularly notable in experiments involving backtranslation attacks. The second graph in Figure 2 illustrates that DNABERT2's accuracy significantly drops with increasing iterations, whereas the Nucleotide Transformer and GROVER maintain high accuracy, demonstrating its robustness due to its larger model capacity or longer sequence processing capabilities.

Moreover, DNABERT2 promoter detection model demonstrates greater robustness compared to DNABERT2 AMR drug classification models, potentially due to the larger volume of training

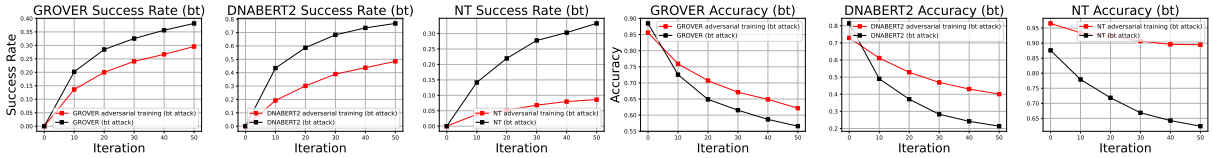


Figure 3: Comparison of success rates and accuracies between adversarial training and standard training in AMR gene Classification across increasing iterations of backtranslation attacks

data utilized. As illustrated in the third graph, with increasing epsilon values, the accuracy of DNABERT2 promoter detection model declines, though to a lesser extent than in AMR detection models. However, DNABERT2 still performs worse than the Nucleotide Transformer and GROVER under nucleotide-level attacks. The fourth graph further shows that, in AMR drug classification tasks, the Nucleotide Transformer and GROVER maintain the highest accuracy against backtranslation attacks, whereas DNABERT2 appears to be relatively more vulnerable to backtranslation attacks. We further address the robustness of general language model in Appendix E.

### 4.3 Comparative Analysis of Classification Tasks

The experimental results indicate that the Nucleotide Transformer and GROVER possess higher robustness against adversarial attacks compared to DNABERT2, especially in the context of backtranslation attacks. Additionally, DNABERT2 trained with larger datasets, such as those used for promoter detection, demonstrates enhanced robustness. This comparative analysis highlights the importance of model capacity, training data size and The ability to process long sequences in improving the resilience of DNA sequence classification models against various adversarial attack methods.

### 4.4 Performance Against Defense Method

We further investigate adversarial training as a defense mechanism against adversarial attacks, incorporating adversarial examples into the training process to improve model robustness.

Figure 3 illustrates the changes in success rate and accuracy over iterations when applying the Back-translate attack in the AMR drug classification task. The results indicate that the success rate of adversarially trained models (red) increases at a slower rate compared to standard training models (black). This suggests that adversarial training enhances the model's robustness against attacks, leading to a relatively lower attack success rate. Furthermore, the accuracy graphs show that adversarially trained models (red) exhibit a smaller decline in accuracy compared to standard training (black). This observation implies that adversarial training improves the model's generalization performance and mitigates performance degradation when exposed to adversarial attacks. For GROVER and DNABERT2, the initial accuracy of adversarially trained models is lower than that of standard training models in Figure 5 and Figure 7. However, as iterations increase, adversarially trained models demonstrate relatively higher accuracy, indicating that repeated exposure to attacks reinforce the model's resilience. In contrast, Nucleotide Transformer exhibits superior robustness, with adversarially trained models outperforming standard training models in accuracy from the outset. Additionally, the accuracy degradation over iterations is the least pronounced among the three models. This suggests that the larger model capacity and ability to handle longer sequences of Nucleotide Transformer further amplify the benefits of adversarial training, making it more resistant to adversarial perturbations. Further details are described in Appendix C

## 5 Conclusion

In this study, we utilize three methods to generate DNA adversarial examples, proposing techniques for generating adversarial examples using nucleotide based attack, codon-based attacks and backtranslation. These methods successfully degrade the performance of AMR drug classification and promoter detection models. Nucleotide-level attacks are the most effective, although they risk disrupting the DNA context. Conversely, backtranslation better preserves meaning and context. Notably, adversarial training as a defense strategy proves effective in improving model robustness, and in some cases, even surpassing the performance of standard training methods. By leveraging adversarial training, models become more resilient against such attacks.

## Limitations

This study has several limitations. First, the datasets used are limited to AMR gene classification with drug classes which may restrict the generalization of the results. Expanding experiments to include datasets such as AMR gene families or AMR mechanisms would provide a more comprehensive evaluation of adversarial robustness.

The study evaluates adversarial attack methods on three distinct DNA language models such as DNABERT2, GROVER, and Nucleotide Transformer. While these models cover a range of architectures and sizes, extending the evaluation to additional DNA classification models could provide further insights into model-specific robustness.

Nucleotide-level attacks, though effective, risk altering the biological meaning of DNA sequences. Their impact is assessed through CG content, species recognition rate, and mapping rate via NCBI BLAST, demonstrating the effectiveness of our approach. While these metrics provide valuable insights, further validation in real biological experiments could strengthen the findings.

Backtranslation methods help preserve sequence meaning and context, as indicated by CG content ratios and species recognition, supporting their role in enhancing model robustness. While our evaluations offer a foundation, additional experimental verification could further solidify these conclusions.

## Ethical Considerations

This study focuses on adversarial robustness in DNA language models and does not involve human subjects or sensitive data. The adversarial attack methods used are intended for evaluating and improving model reliability in bioinformatics, not for malicious purposes. However, it is important to acknowledge the potential risks of adversarial attacks in bioinformatics applications, where model vulnerabilities could impact genomic analysis and medical decision-making.

## References


Brian P. Alcock, Amogelang R. Raphenya, Tammy T Y Lau, Kara K. Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V. Nguyen, Annie A. Cheng, Sihan Liu, Sally Y. Min, Anatoly Miroshnichenko, Hiu-Ki Tran, Rafik E. Werfalli, Jalees A. Nasir, Martins Oloni, David J. Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N. Sharma, Emily Bordeleau, Andrew C. Pawlowski, Haley L. Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L. Winsor, Robert G. Beiko, Fiona S L. Brinkman, William W L. Hsiao, Gary V. Domselaar, and Andrew G. McArthur. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 48(D1):D517–D525. Published online 2019 Oct 29.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Nathalie Bonin, Enrique Doster, Hannah Worley, Lee J. Pinnell, Jonathan E. Bravo, Peter Ferm, Simone Marini, Mattia Prosperi, Noelle Noyes, Paul S. Morley, and Christina Boucher. 2023. MEGARes and AMR++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic Acids Research*, 51(D1):D744–D752.

Benjamin Buchfink, Kerstin Reuter, and Hannes-Gerd Drost. 2021. Sensitive protein alignments at tree-of-life scale using diamond. *Nature Methods*, 18:366–368.

Charles E. Cook, Mary Todd Bergman, Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. 2016. The european bioinformatics institute in 2016: Data growth and integration. *Nucleic Acids Research*, 44(D1):D20–D26.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018a. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018b. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025. AAAI Press.

Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L. Madden. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36(suppl_2):W5–W9.

Volodymyr Kuleshov, Evgenii Nikishin, Shantanu Thakoor, Tingfung Lau, and Stefano Ermon. 2021. Quantifying and understanding adversarial examples in discrete input spaces. *arXiv preprint arXiv:2112.06276*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Jiajia Liu, Mengyuan Yang, Yankai Yu, Haixia Xu, Kang Li, and Xiaobo Zhou. 2024. Large language models in bioinformatics: Applications and perspectives. *arXiv*, Preprint(arXiv:2401.04155v1).

Tianyuan Liu and Yuqing Sun. 2023. End-to-end adversarial sample generation for data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11359–11368, Singapore. Association for Computational Linguistics.

Xiaotu Ma, Ying Shao, Liqing Tian, Diane A. Flasch, Heather L. Mulder, Michael N. Edmonson, Yu Liu, Xiang Chen, Scott Newman, Joy Nakitandwe, Yongjin Li, Benshang Li, Shuhong Shen, Zhaoming Wang, Sheila Shurtleff, Leslie L. Robison, Shawn Levy, John Easton, and Jinghui Zhang. 2019. Analysis of error profiles in deep next-generation sequencing data. *Genome Biology*, 20(50).

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Daniel Mas Montserrat and Alexander G. Ioannidis. 2022. Adversarial attacks on genotype sequences. *bioRxiv*.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Yukiteru Ono, Kiyoshi Asai, and Michiaki Hamada. 2021. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics*, 37(5):589–595. Published: 25 September 2020.

Karl Pearson. 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, 187:253–318.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

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

Oluwafemi A. Sarumi and Dominik Heider. 2024. Large language models and their applications in bioinformatics. *Computational and Structural Biotechnology Journal*, 23:3498–3505.

Martin Steinegger and Johannes Söding. 2017. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35:1026–1028.

Ramzan Umarov, Hiroyuki Kuwahara, Yu Li, Xin Gao, and Victor Solovyev. 2019. Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics*, 35(16):2730–2737. Published: 02 January 2019.

Hyunwoo Yoo, Bahrad Sokhansanj, James R. Brown, and Gail Rosen. 2024. Predicting anti-microbial resistance using large language models. *arXiv preprint arXiv:2401.00642*.

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An open-source textual adversarial attack toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371, Online. Association for Computational Linguistics.

Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. 2022. Improving the adversarial robustness of nlp models by information bottleneck. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3588–3598, Dublin, Ireland. Association for Computational Linguistics.

Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S. Dhillon, and Cho-Jui Hsieh. 2019. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations*. Huan Zhang and Hongge Chen contributed equally.

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

# A Implementation Details

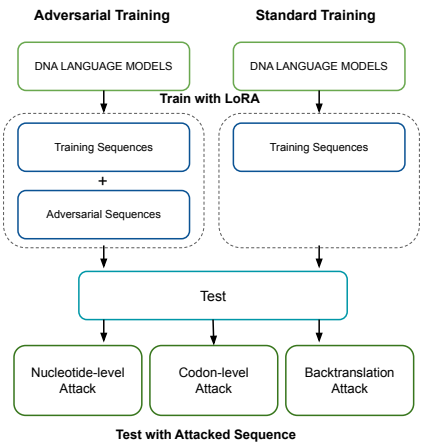


Figure 4: Overview of Adversarial Training(Fine-tuning) for DNA Language Models and Testing Adversarial Attacks

We utilize LoRA (Hu et al., 2022) to finetune the DNA language models for each benchmarks. The learning rate is set to 0.0005, the number of epochs is 2, the batch size is 64. Single A100 GPU is used, and and the entire process takes approximately one hour. The number of parameters for each model is as follows: DNABERT-2 (117M), Nucleotide Transformer (2.5B), and GROVER (90M). DNABERT-2 is available under the Apache License 2.0, and Nucleotide Transformer is distributed under the CC BY-NC-SA 4.0 License. The GROVER model, as hosted on Hugging Face (PoetschLab/GROVER), does not have a publicly specified license.

# B Details of Benchmarks

**Antimicrobial Resistance Classification** We employ the methodology outlined in Yoo et al. (2024) to integrate datasets from CARD v2 (Alcock et al., 2020) and MEGARes v3 (Bonin et al., 2023) for investigating antibiotic resistance. Classes with fewer than 15 instances are excluded. The remaining data is divided into training, testing, and validation sets, maintaining the same proportions as the study. Data integration is performed using the European Bioinformatics Institute Antibiotic Resistance Ontology (EBI ARO; Cook et al., 2016), with irrelevant classes excluded. Overall, the drug class classification dataset for AMR gene classification is split into 75% training data, 20% test data, and 5% validation data. There are a total of 9 drug classes. The MEGARes dataset is available under the MIT License, and the CARD dataset is distributed under the CC BY-NC 4.0 License.

**Promoter Detection** The promoter detection data from the Genome Understanding Evaluation (GUE) benchmark dataset introduced in the DNABERT-2 is used to detect promoters in gene sequences (Zhou et al., 2023). Promoters are DNA sequences that regulate the initiation of gene transcription and play a crucial role in gene expression regulation and understanding biological phenomena.

The data is extracted from gene sequences of various species, providing a broad context for biological research. Each data point includes both sequences that contain promoters and those that do not, enabling the model to learn how to distinguish promoters. This dataset includes complex promoter sequence elements known as MIX elements, which are found in specific genes. We utilize the prom 300 all dataset from the GUE benchmark for promoter detection. The dataset is splited into approximately 80% training, 10% development, and 10% test data, with 2 classes for binary classification. The GUE dataset is distributed along with DNABERT-2, which is available under the Apache-2.0 License.

# C Comparisons of Adversarial Training and Standard Training

Figure 5, Figure 6, Figure 7, and Figure 8 compare the performance differences between adversarially trained models and conventionally trained models (black) as the number of attack iterations and perturbation magnitude (epsilon) increase. Epsilon represents the rate of change in the sequence for the attack, indicating the proportion of the sequence that is altered. Iterations indicate the number of times the attack is performed. While all models experience a decrease in accuracy as the number of iterations and perturbation levels increase, adversarially trained models generally maintain higher accuracy and exhibit a more gradual decline in performance. In terms of attack success rate, conventionally trained models show a sharp increase in attack success as the number of iterations and perturbation magnitude grow, whereas adversarially trained models exhibit a more moderate increase, demonstrating greater resistance to adversarial attacks. These findings suggest that adversarial training effectively enhances model robustness against iterative attacks and higher perturbation magni-

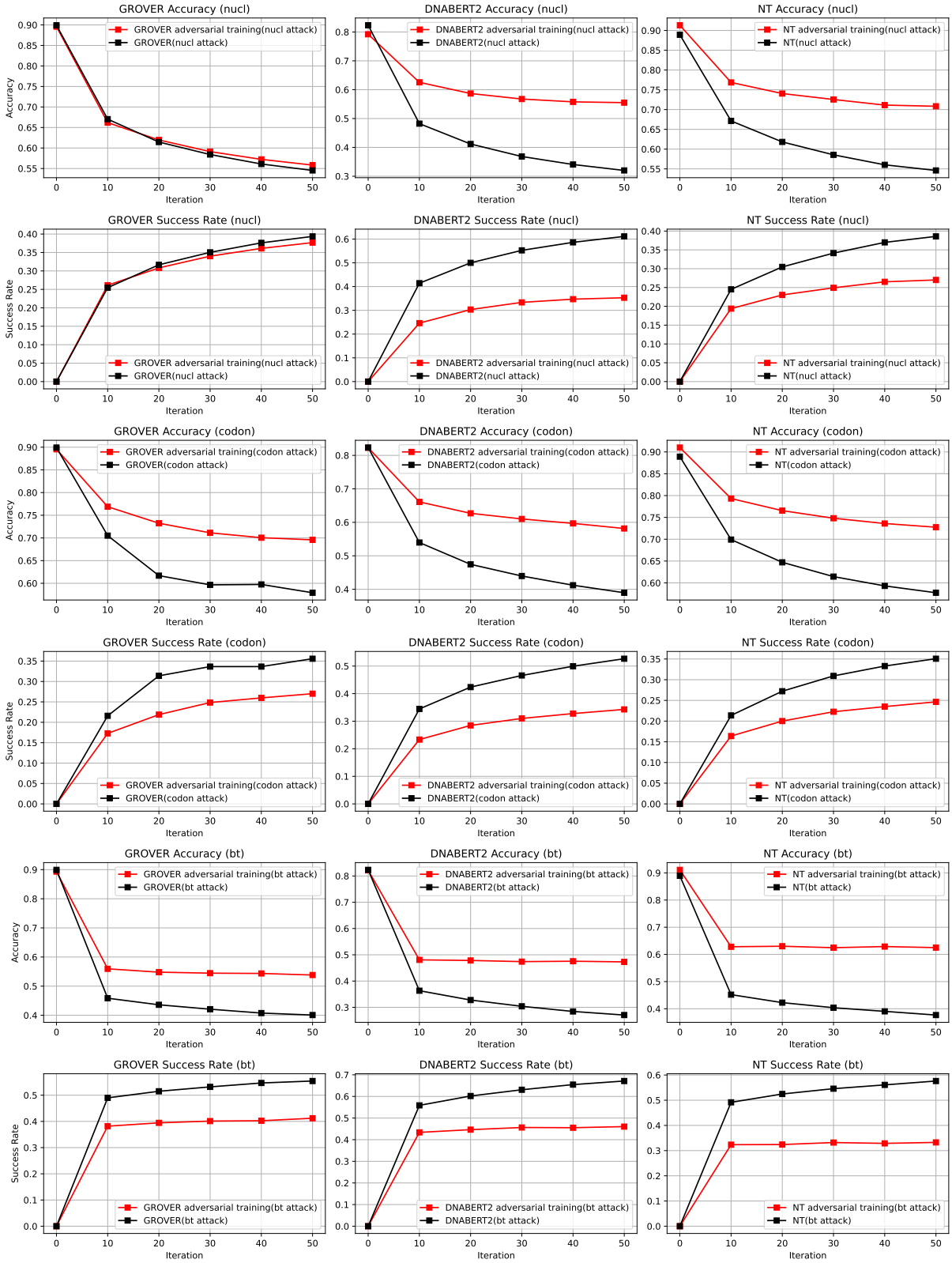


Figure 5: Comparison of Adversarial Training and Standard Training in Promoter Detection with Increasing Iterations

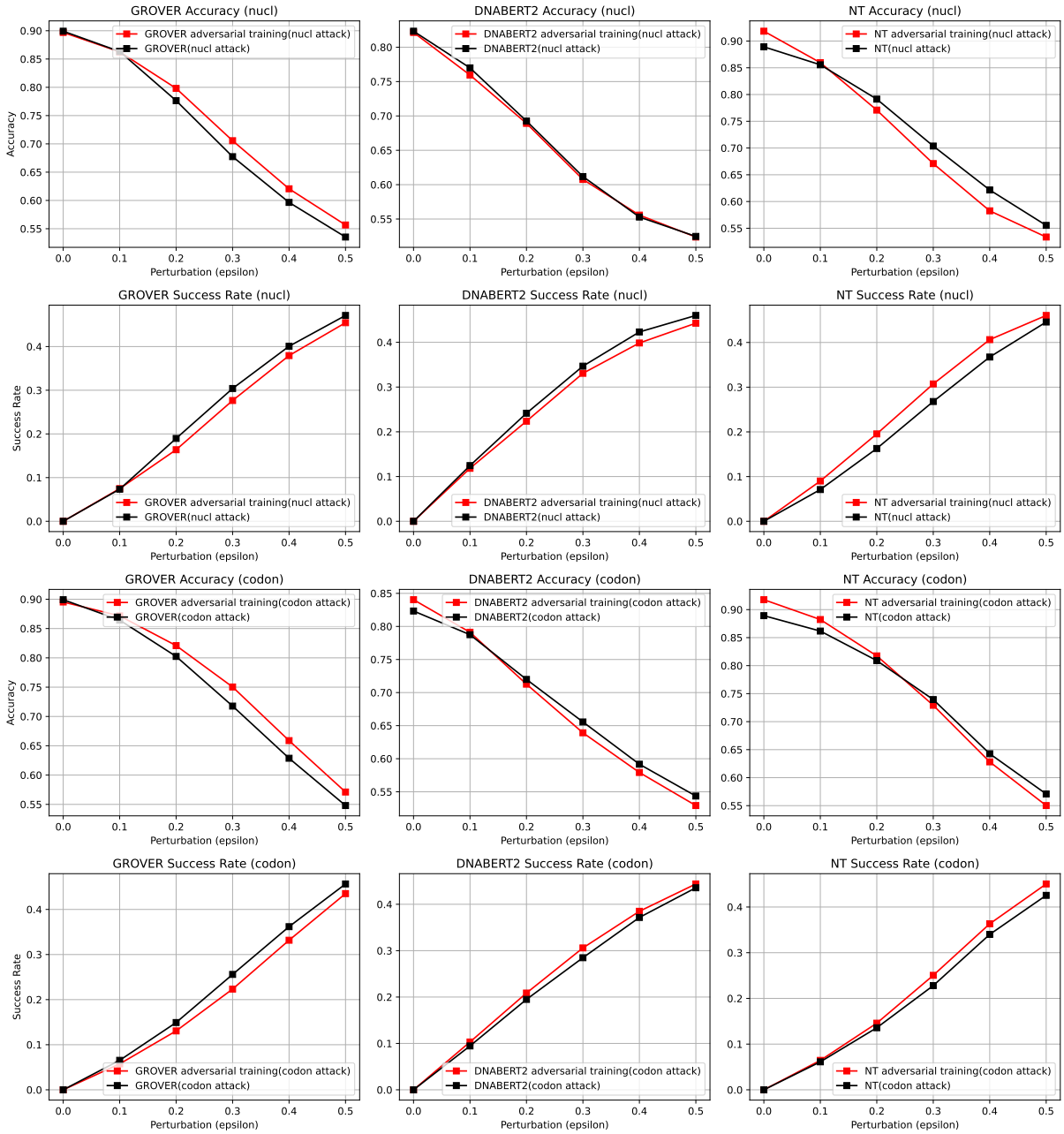


Figure 6: Comparison of Adversarial Training and Standard Training in Promoter Detection Across Different Perturbation Levels

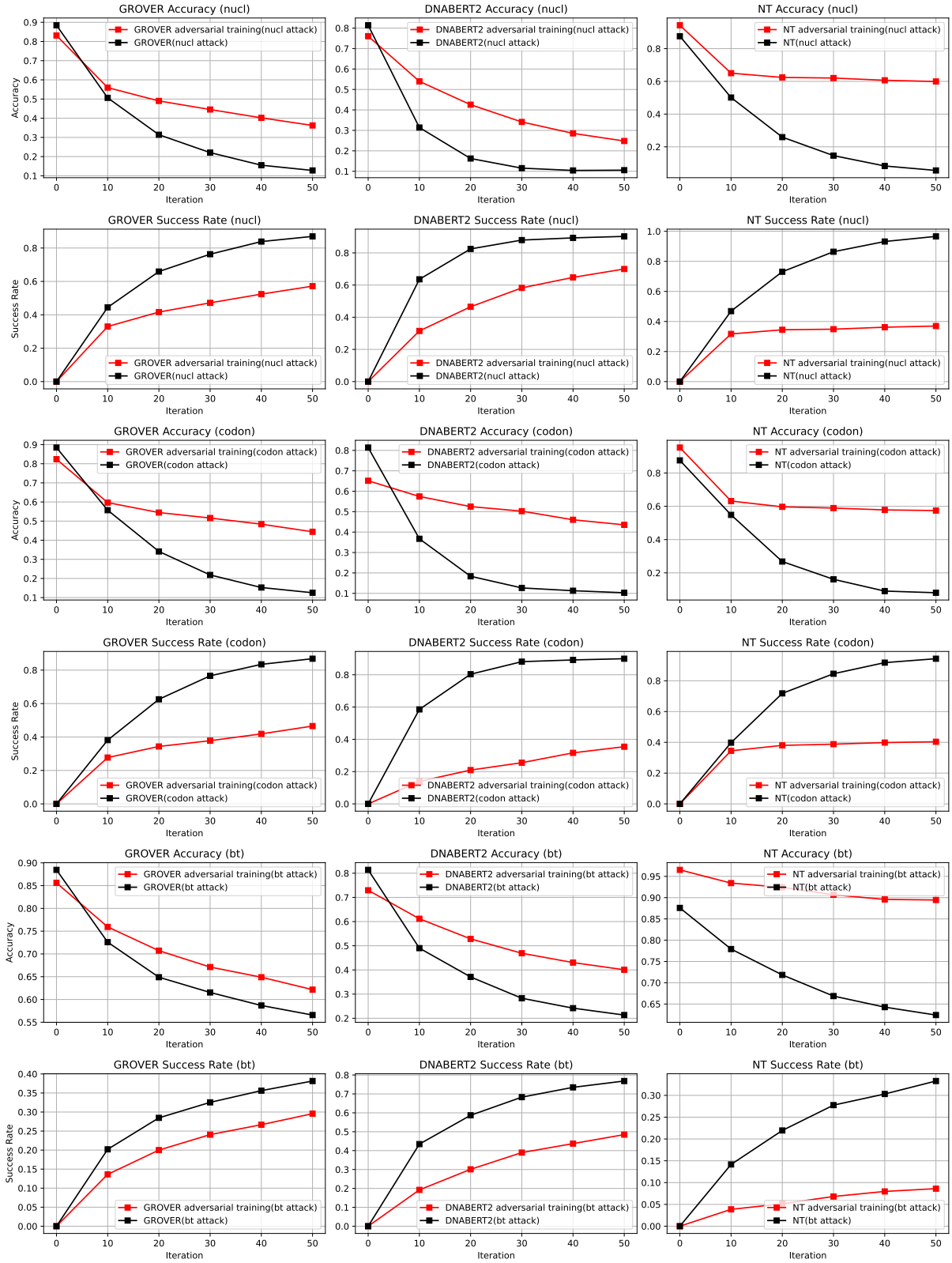


Figure 7: Comparison of Adversarial Training and Standard Training in Antimicrobial Resistance Drug Classification with Increasing Iterations

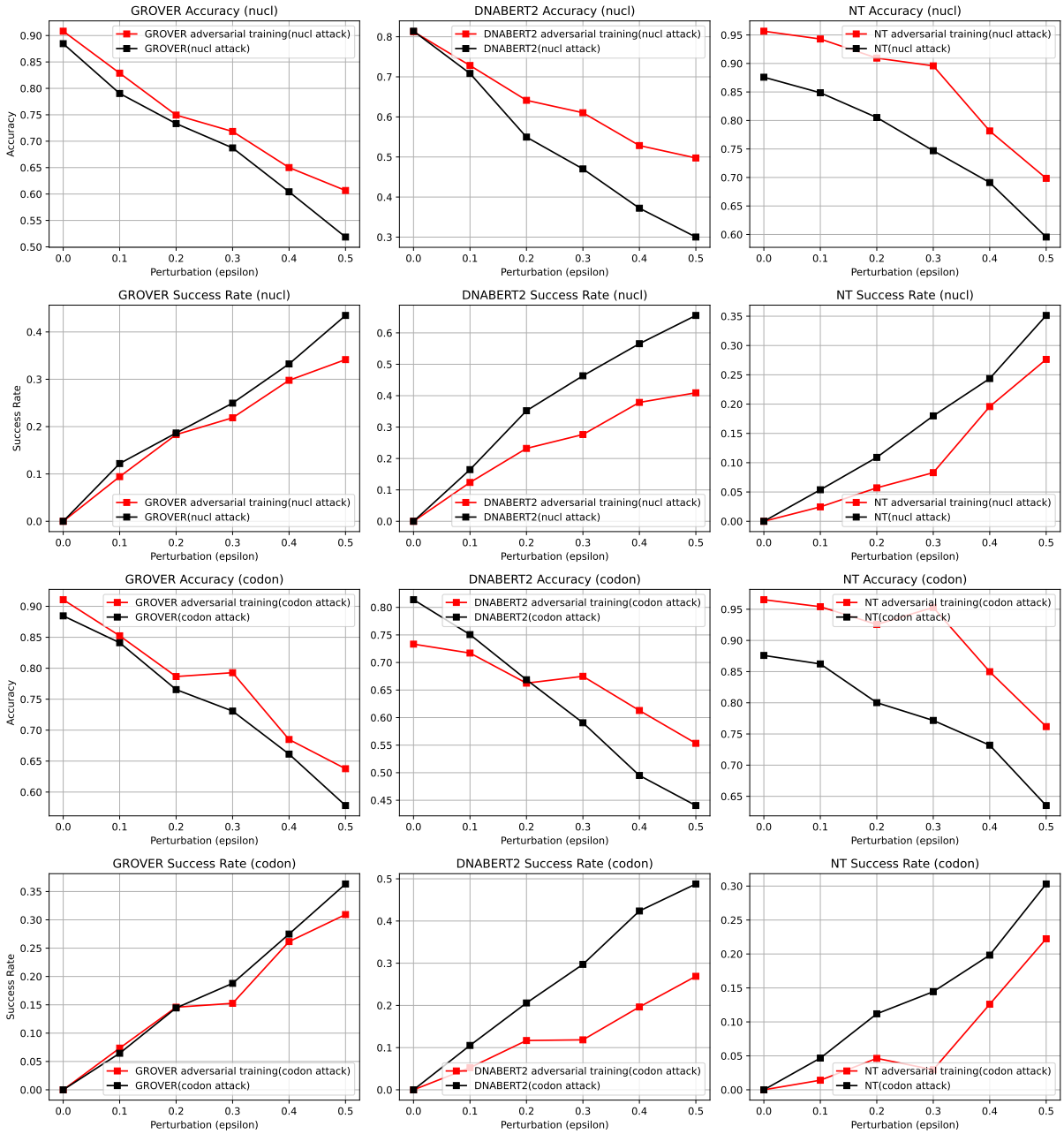


Figure 8: Comparison of Adversarial Training and Standard Training in Antimicrobial Resistance Drug Classification Across Different Perturbation Levels

tudes while mitigating performance degradation. Figures present accuracy and success rate trends under different perturbation settings. However, all results are based on a single run.

# D Adversarial DNA Sequence validation

## D.1 Promoter Detection Sequence validation

$$GC\% = \frac{G + C}{A + T + G + C} \times 100 \quad (1)$$

To validate adversarial sequences in promoter detection, we compared the GC content of the original sequences and the adversarially generated sequences. Prior research has highlighted the strong correlation between promoter regions and GC content (Umarov et al., 2019). Using this insight, we assess whether the adversarial sequences maintain a similar GC content distribution to the original sequences. We calculate the GC content percentage using Equation 1.

Specifically, we measure the Pearson correlation coefficient (Pearson, 1896) between the GC content of original and adversarial sequences to quantify their similarity. This approach allows us to evaluate whether the adversarial sequences retain key GC content characteristics, providing an initial validation measure.

Table 1 demonstrates a decreasing trend in the Pearson correlation coefficient of GC content between the original and perturbed sequences as the perturbation level increases. This trend is observed in both nucleotide-level and codon-level attacks, where a higher perturbation parameter ($\epsilon$) results in a lower Pearson correlation coefficient. Table 2 presents the Pearson correlation coefficient between the original sequences and perturbed sequences as the number of iterations increases for nucleotide-level, codon-level, and backtranslation-based attacks. Across all attack methods, the Pearson correlation coefficient decreases as the number of iterations increases. However, in the case of backtranslation, the decline is less pronounced compared to nucleotide-level and codon-level attacks, which may suggest that backtranslation better preserves the contextual integrity of the original sequences.

## D.2 Antimicrobial Resistance Drug classification Sequence validation

To evaluate the effectiveness of adversarial DNA sequences for attacks as the number of iterations and the magnitude of perturbations increase, we utilize the BLAST (Basic Local Alignment Search Tool(Altschul et al., 1990)) database mapping API provided by the National Center for Biotechnology Information (NCBI)(Johnson et al., 2008).

Analysis of the BLAST mapping results reveal a gradual decline in the mapping rate of adversarially perturbed sequences as the same species as the original test dataset, as the intensity of adversarial perturbations increased. Specifically, when the perturbation magnitude is small, the perturbed sequences are mapped as the same species as the original sequence. However, as the perturbation magnitude increases, the likelihood of being classified as a different species increases. This indicates that adversarial attacks effectively induce classification perturbations in the intended direction. However, if the perturbed sequences are classified as entirely different species, the effectiveness of the attack sequences may be somewhat diminished. Ideally, a adversarial sequence should not excessively distorted the biological properties of the sequence. Excessive perturbation, however, may result in adversarial sequences being recognized as entirely different species, which could undermine their intended effectiveness.

Additionally, we examine cases where BLAST fails to clearly match the input sequence to any species in the existing database (i.e., cases where BLAST does not recognize the sequence). The results indicate that as the intensity of the adversarial attack increases, the proportion of sequences that BLAST fails to recognize rose sharply. This suggests that larger perturbations lead to greater distortion of the biological properties of the original DNA sequences. In particular, for iterations, both the mapping rate and recognition rate are significantly low, suggesting that the results may not fully align with real-case scenarios. However, similar to the validation of promoter detection sequences, applying backtranslation results in higher mapping and recognition rates compared to nucleotide-level and codon-level attacks. This may indicate that backtranslation relatively better preserves the contextual integrity of the sequences. Consequently, in such cases, the effectiveness of the adversarial sequences may also be reduced, further distancing them from the natural noise present in real-world conditions. This is particularly evident in multi-iteration nucleotide and codon-level attacks, as well as backtranslation, where recognition and mapping rates in the BLAST database are significantly low, making them a form of worst-case performance

| $\epsilon$ | Nucleotide-level Attack | Codon-level Attack |
|---|---|---|
| 0.1 | 0.9910 | 0.9903 |
| 0.2 | 0.9788 | 0.9776 |
| 0.3 | 0.9614 | 0.9592 |
| 0.4 | 0.9361 | 0.9327 |
| 0.5 | 0.8990 | 0.8970 |

Table 1: Pearson Correlation Coefficient of GC content based on perturbation ($\epsilon$) for different attack strategies.

| Iteration | Nucleotide-level Attack | Codon-level Attack | Back-translation Attack |
|---|---|---|---|
| 10 | 0.9525 | 0.9481 | 0.9303 |
| 20 | 0.8911 | 0.8925 | 0.9271 |
| 30 | 0.8512 | 0.8498 | 0.9264 |
| 40 | 0.8236 | 0.8193 | 0.9250 |
| 50 | 0.8068 | 0.7987 | 0.9270 |

Table 2: Pearson Correlation Coefficient of GC content based on iteration for different attack strategies.

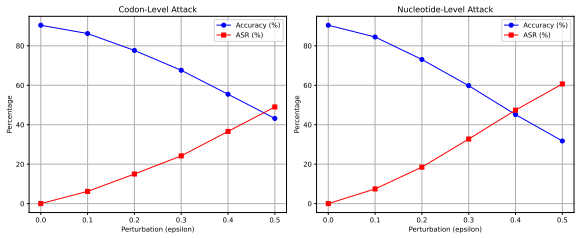


Figure 9: Accuracy and attack success rate (ASR) under codon-level and nucleotide-level perturbations in Antimicrobial Resistance Drug Classification Across Different Perturbation Levels

exploration.

## E Adversarial Evaluation of GPT-4o on DNA AMR Classification

We present additional analysis on the adversarial robustness of `gpt-4o-mini-2024-07-18` fine-tuned for DNA AMR classification. This study has demonstrated that DNA Language Models are vulnerable to adversarial attacks, and we extend this investigation to examine whether general-purpose LLMs, when adapted for biological sequence classification, exhibit similar weaknesses.

By applying codon-level and nucleotide-level perturbations, we observe a sharp decline in model accuracy as perturbation intensity increased (dropping to 31.76% at the highest level), while the attack success rate (ASR) rises to a maximum of 60.67% in Figure 9.

With the growing interest in applying LLMs to DNA sequence analysis (Sarumi and Heider, 2024; Liu et al., 2024), ensuring robustness remains a critical challenge. Adversarial training and domain-specific robustness strategies may be necessary to enhance the reliability of such models in biological sequence tasks.

## F Effect of Realistic Sequencing Errors on Robustness

To evaluate model robustness under more realistic sequencing scenarios, we simulate sequencing errors using PBSIM2 (Ono et al., 2021), incorporating real-world error patterns observed in PacBio and Oxford Nanopore sequencing. Sequencing Error Simulation is a technique used to model errors occurring during the genome sequencing process, playing a crucial role in genome analysis, algorithm evaluation, and robustness assessment in bioinformatics. PBSIM2, widely used for this purpose, simulates insertion, deletion, and substitution errors, as well as the length distribution observed in long-read sequencing technologies such as PacBio and Oxford Nanopore, enabling the generation of reads that closely resemble real sequencing data. The simulated sequencing error data can be utilized to test model robustness by incorporating noise similar to that encountered in real sequencing experiments. This holds significant importance from the perspective of adversarial training and robustness assessment, as it helps deep learning-based genomic analysis models maintain predictive accuracy while improving their resilience to sequencing errors. Our experiments show that the model's performance degrades under these conditions, indicating that realistic sequencing errors can pose a challenge to robustness. In the promoter detec-

| $\epsilon$ | **Nucleotide-level Attack** | **Codon-level Attack** |
|---|---|---|
| 0.1 | 93.05 | 92.60 |
| 0.2 | 70.35 | 85.98 |
| 0.3 | 16.25 | 75.81 |
| 0.4 | 1.36 | 46.15 |
| 0.5 | 0.50 | 13.03 |

Table 3: Species Matching rate (%) based on perturbation ($\epsilon$) for different attack strategies.

| $\epsilon$ | **Nucleotide-level Attack** | **Codon-level Attack** |
|---|---|---|
| 0.1 | 99.75 | 99.88 |
| 0.2 | 84.49 | 99.88 |
| 0.3 | 25.56 | 96.40 |
| 0.4 | 2.61 | 68.98 |
| 0.5 | 0.62 | 24.81 |

Table 4: Recognition Rate (%) based on perturbation ($\epsilon$) for different attack strategies.

tion task with Nucleotide Transformer, accuracy dropped from 0.8892 to 0.8608.

| Iteration | Nucleotide-level Attack | Codon-level Attack | Back-translation Attack |
|---|---|---|---|
| 10 | 0.25 | 1.24 | 1.99 |
| 20 | 0.25 | 0.25 | 2.73 |
| 30 | 0.25 | 0.25 | 3.47 |
| 40 | 0.25 | 0.25 | 1.61 |
| 50 | 0.25 | 0.25 | 3.10 |

Table 5: Species Matching rate (%) based on iteration for different attack strategies.

| Iteration | Nucleotide-level Attack | Codon-level Attack | Back-translation Attack |
|---|---|---|---|
| 10 | 0.25 | 1.99 | 15.63 |
| 20 | 0.25 | 0.25 | 16.63 |
| 30 | 0.25 | 0.25 | 18.11 |
| 40 | 0.25 | 0.25 | 18.73 |
| 50 | 0.37 | 0.25 | 16.63 |

Table 6: Species Recognition rate (%) based on iteration for different attack strategies.