SafeGenes: Evaluating the Adversarial Robustness of Genomic Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Genomic Foundation Models (GFMs), such as Evolutionary Scale Modeling (ESM), have demonstrated significant success in variant effect prediction. However, their adversarial robustness remains largely unexplored. To address this gap, we propose **SafeGenes**: a framework for Secure analysis of genomic foundation models, leveraging adversarial attacks to evaluate robustness against both engineered near-identical adversarial Genes and embedding-space manipulations. In this study, we assess the adversarial vulnerabilities of GFMs using two approaches: the Fast Gradient Sign Method (FGSM) and a soft prompt attack. FGSM introduces minimal perturbations to input sequences, while the soft prompt attack optimizes continuous embeddings to manipulate model predictions without modifying the input tokens. By combining these techniques, SafeGenes provides a comprehensive assessment of GFM susceptibility to adversarial manipulation. Targeted soft prompt attacks led to substantial performance degradation, even in large models such as ESM1b and ESM1v. These findings expose critical vulnerabilities in current foundation models, opening new research directions toward improving their security and robustness in high-stakes genomic applications such as variant effect prediction.

1 Introduction

Genomic Foundation Models (GFMs), such as Evolutionary Scale Modeling (ESM), have revolutionized the field of genomics by providing powerful tools for variant effect prediction and genomic sequence analysis. These models leverage large-scale data and sophisticated architectures to deliver high accuracy and generalizability across diverse tasks, offering transformative potential in biomedical applications. For instance, AlphaMissense (Cheng et al., 2023) integrates evolutionary conservation and structural modeling for clinical variant classification, while ESM1b (Brandes et al., 2023) demonstrates strong zero-shot performance on isoform-specific missense predictions using large-scale masked language modeling. Despite these advances, current GFMs are generally trained in a task-agnostic manner and lack disease-specific adaptation, limiting their direct applicability in clinical decision-making. Moreover, their deployment in critical domains such as clinical diagnostics necessitates rigorous evaluation of their reliability under adversarial conditions. In particular, it remains unclear whether GFMs are vulnerable to adversarial manipulations—whether

through perturbed input sequences or learned prompts—which could compromise model robustness and trustworthiness.

To explore this question in a clinically grounded setting, we fine-tune a range of GFMs for disease-specific variant effect prediction (VEP) using DYNA (Zhan et al., 2025)—a modular framework that adapts protein- and DNA-based GFMs to cardiac and regulatory genomics tasks. DYNA employs a Siamese neural network architecture (Figure 1a) and a pseudo-log-likelihood ratio (PLLR)-based scoring function (Figure 1b), enabling effective adaptation to domain-specific VEP with limited rare variant data. Using this framework, we fine-tune multiple GFM backbones, including ESM1b and ESM2 variants, on cardiomyopathy (CM) and arrhythmia (ARM) datasets. These fine-tuned models serve as the basis for evaluating adversarial robustness.

Recent studies have underscored the need to evaluate not only data privacy but also the adversarial robustness of genomic AI models—particularly those built on powerful pre-trained architectures. Our findings reveal that GFMs, while strong in generalization, remain vulnerable to attacks that exploit both surface-level perturbations and deeper latent representations. To systematically probe these vulnerabilities, we introduce a dual attack framework that combines perturbation-based and semantic-level adversarial strategies. First, as illustrated in Figure 1c, we use Fast Gradient Sign Method (FGSM) to probe GFM's susceptibility to adversarial sequences by injecting gradient-based noise into variant embeddings, assessing the impact of near-identical sequences on predictions. In addition to FGSM, a **soft prompt attack** is introduced, which operates in the model's embedding space rather than directly modifying input tokens. We consider two variants of the soft prompt attack: a confidence hijack, which reduces the margin between wild-type and variant predictions, and a targeted attack, which specifically pushes benign variants toward pathogenic predictions. As shown in Figure 1d, this method optimizes continuous prompt embeddings that, when prepended to the wild-type and variant inputs, guide the model toward incorrect predictions. Unlike token-level attacks, soft prompts leave the input sequence unchanged but manipulate internal representations to influence decision boundaries. This dual approach enables a comprehensive evaluation of GFM's adversarial robustness by combining input-space perturbations and embedding-space manipulations, revealing the model's susceptibility to both types of attacks.

Across both CM and ARM datasets, all evaluated GFMs exhibited susceptibility to adversarial perturbations, with consistent drops in AUC and AUPR under both input-space (FGSM) and embedding-space (soft prompt) attacks. The soft prompt attack with targeted optimization resulted in the most severe degradation, especially on smaller ESM2 models (Lin et al., 2023), where AUC and AUPR decreased by up to 10 and 6 percentage points, respectively. Even high-capacity models such as ESM1b (Rives et al., 2021) and ESM1v (Meier et al., 2021) were not robust, showing large performance drops under targeted manipulation. In contrast, the confidence hijack variant of the soft prompt attack led to more subtle but consistent degradation, reflecting the sensitivity of learned embeddings to adversarial manipulation. These findings demonstrate the urgent need to evaluate and improve the robustness of GFMs, particularly in clinical genomics where decision boundaries must remain stable under distribution shifts and intentional manipulation. Moreover, they demonstrate a critical limitation in current defense paradigms, which often assume that adversarial perturbations are easily detectable via prediction confidence or input-level cues. Our results suggest that effective defenses must instead operate in latent space—monitoring shifts in embedding geometry and prediction distributions to detect semantically aligned but deceptive adversarial behavior.

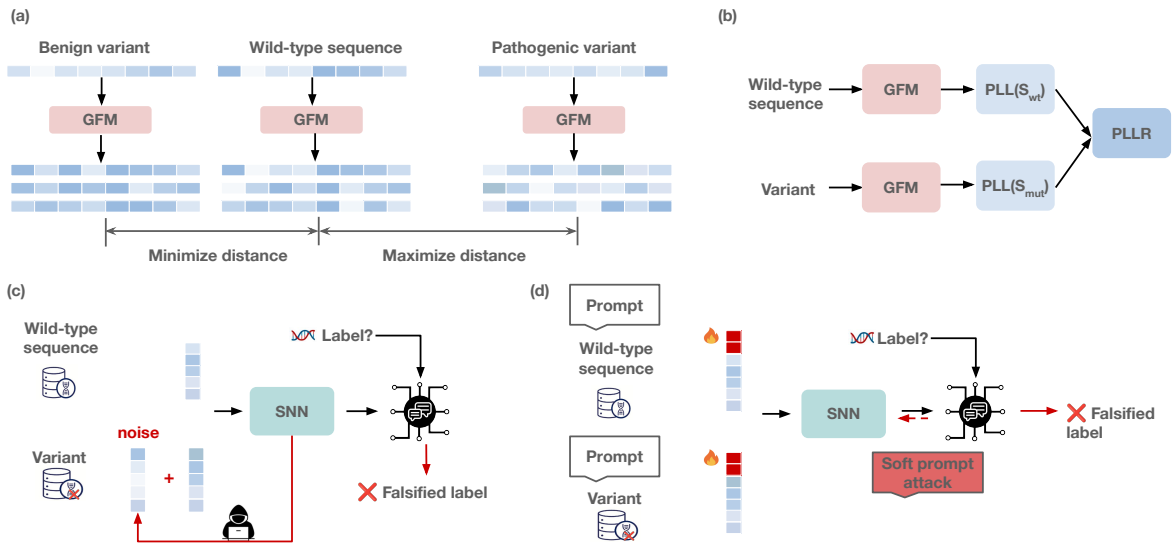


Figure 1: Illustration of adversarial sensitivity in GFMs for variant effect prediction. (a) Conceptual schematic showing wild-type and variant embeddings in GFM representation space. Pathogenic variants are expected to maximize distance from the wild-type, while benign variants minimize distance. (b) Overview of PLLR computation using a language model. The PLL for the wild-type and mutant sequence is compared to infer the variant label. (c) FGSM attack perturbs model embeddings to falsify pathogenicity predictions, shifting benign variants to appear pathogenic and vice versa. (d) Soft prompt attack: structured prompts induce the model to shift decision boundaries, leading to adversarial misclassification even when the original label is correct.

2 Related Works

Despite growing concerns around privacy and robustness in genomic machine learning, most prior efforts have focused narrowly on data anonymity or privacy-preserving mechanisms. For instance, Kuo et al. (2022) emphasized the field's primary concern as the protection of genomic identity, particularly in light of widespread services like 23andMe. Techniques such as differential privacy have been shown vulnerable to de-anonymization attacks (Chen et al., 2020), while others have explored secure data access methods to prevent malicious retrieval (Huang et al., 2015). However, these studies primarily target data privacy, not model integrity.

A separate line of work has emerged to explore attacks on the models themselves, with Montserrat & Ioannidis (2023) demonstrating gradient-based white-box attacks on deep learning models used for genomic classification. Feature Importance Model-agnostic Black-box Attack (FIMBA) builds on this direction by introducing a black-box, interpretability-driven framework to compromise gene expression models (Skovorodnikov & Alkhzaimi, 2024). By leveraging SHAP-based feature importance (Wang et al., 2024), FIMBA successfully perturbs top-ranked features to degrade classifier

accuracy, even without access to model gradients—marking an important advance in model-agnostic adversarial genomics.

Yet, these efforts still operate on conventional architectures (e.g., MLPs (Popescu et al., 2009; Riedmiller & Lernen, 2014), CNNs (Bouvrie, 2006)) and predominantly tabular gene expression data. In contrast, our work moves adversarial robustness research into the era of GFMs—large-scale pre-trained transformers capable of generalizing across diverse biological tasks. We propose two types of adversarial attacks:

- **FGSM perturbation:** a token-space attack that introduces gradient-based noise to the input embeddings, simulating small but adversarial changes to the biological sequence.
- **Soft prompt attack:** an embedding-space strategy that prepends learnable prompt tokens to the input, steering the model's predictions without modifying the original sequence.

Unlike FIMBA, which relies on perturbing high-importance input features derived from model interpretability tools, our methods operate at a deeper semantic level, targeting the latent representation space of GFMs. By crafting semantically aligned adversarial inputs—either through optimized gradient perturbations or soft prompt embeddings—we reveal vulnerabilities that are invisible to surface-level feature-based attacks. These latent space manipulations exploit the internal structure of GFMs, where meaning is encoded across multiple dimensions, making the attacks not only more effective but also more informative for understanding model behavior and failure modes. This aligns with recent findings in computational pathology, where vision transformers (Khan et al., 2022) were shown to exhibit more robust latent representations than CNNs under adversarial stress, suggesting that semantic structure in embedding space is a key factor in model resilience (Ghaffari Laleh et al., 2022). However, while that study focused on visual embeddings, our work offers a parallel in the genomic domain by directly targeting and probing vulnerabilities in the latent space of sequence-based foundation models. Additionally, our soft prompt attack mechanism enables task-specific manipulation without retraining or modifying model weights. This novel approach creates a flexible framework to test and stress the robustness of transformer-based models under adversarial prompts—advancing both our understanding of model generalization and our capacity to build defense mechanisms.

FGSM and the soft prompt attack serve complementary roles in this study. FGSM evaluates the model's sensitivity to small, direct perturbations in the input sequence, offering insights into its robustness against minimally altered genomic variants. In contrast, the soft prompt attack targets the model's internal representation space by optimizing a set of trainable prompt embeddings, effectively misleading the model toward confident yet incorrect predictions without modifying the input sequence. This targeted degradation of the confidence margin not only reveals latent vulnerabilities in GFMs, but also anticipates potential failure modes that are difficult to detect through surface-level perturbations. In particular, our observations align with recent work in vision-language models (Zhang et al., 2023), where aligning the confidence distributions between clean and adversarial samples has been proposed as an effective defense strategy. However, unlike prior work that assumes such distributions are already measurable or optimizable (Hendrycks & Gimpel, 2017; Zhang et al., 2023), our attacks directly create such pathological shifts in confidence from within the input space—offering a valuable adversarial probe that can expose and quantify semantic instability in GFMs. As such, our work not only contributes a new attack mechanism but also lays the

Table 1: Summary of key findings from adversarial robustness evaluation of genomic foundation models across CM and ARM datasets.

Evaluation Aspect	**Observation and Implication**
Adversarial susceptibility (overall)	All GFMs exhibited consistent degradation in AUC and AUPR under adversarial perturbations, indicating a general lack of robustness across both CM and ARM tasks.
Impact of FGSM (input-space)	FGSM introduced minor but reproducible performance drops, showing that even simple gradient-based input perturbations can undermine prediction stability.
Impact of soft prompt attack (confidence hijack)	This embedding-space attack subtly collapsed confidence margins between wild-type and variant sequences, leading to systematic but less dramatic reductions in classification performance.
Impact of soft prompt attack (targeted)	Targeted prompt optimization produced the largest performance degradation, particularly in smaller models, with AUC reductions of up to 10 percentage points.
Relative robustness of larger models	Larger architectures such as ESM1b and ESM1v demonstrated better baseline performance and marginally greater resilience, though they remained vulnerable to targeted prompt-based attacks.
Relevance to clinical genomics	The observed vulnerabilities demonstrate the necessity of adversarial robustness evaluation in clinical genomic models, where decision reliability is critical for pathogenicity interpretation.

groundwork for developing future defenses that operate in latent and distributional space, rather than relying solely on input-level regularization.

In summary, while prior studies have primarily focused on protecting data privacy or exposing vulnerabilities in shallow genomic classifiers, our work systematically advances adversarial robustness research into the realm of GFMs. By introducing both gradient-based and prompt-based attack paradigms, we uncover failure modes at multiple semantic levels—ranging from input perturbations to latent representation manipulations. These insights directly address the emerging security challenges posed by powerful, general-purpose sequence models in genomics, and also provide actionable tools for evaluating the resilience of such models in clinical and biomedical applications.

Table 1 summarizes key findings from our evaluation of adversarial robustness in GFMs across CM and ARM datasets. The results highlight model vulnerabilities under different attack strategies and their implications for clinical deployment.

3 Methods

GFMs, including protein language models such as ESM1b, are typically trained using the Masked Language Modeling (MLM) objective. Within this MLM paradigm, specific amino acid residues in

protein sequences are "masked" or hidden, and the model is trained to predict the identity of these masked residues. As the model makes these predictions, it produces raw scores or predictions for each potential amino acid that could replace the masked residue, commonly referred to as "MLM logits". The logits predicted by a protein language model for observing the input amino acid s_i at position i given the sequence s are shown in red frames. When passed through an activation function like softmax, these logits provide probabilities over the possible amino acids, guiding the prediction process.

Pseudo-log-likelihood ratio computation In order to fine-tune GFMs on each pair of wild-type and mutant sequences, i,e., s^{WT} and s^{mut}, we create a siamese network with two weight-sharing protein language model branches (shown in Figure 1 (a) and (b)) to update the weights such that the produced MLM logits are both semantically meaningful and can be compared via pseudo-log-likelihood ratio (PLLR). For a sequence $s = s_1, ..., s_L$, the pseudo-log-likelihood is calculated as $\text{PLL}(s) = \sum_{i=1}^{L} \log P(x_i = s_i|s)$, where L denotes the sequence length, s_i represents the amino acid at position i, and $\log P(x_i = s_i|s)$ denotes the log-likelihood predicted by the protein language model when observing amino acid s_i at position i within sequence s. The PLLR between the s^{WT} and s^{mut} is then computed as:

$$\lambda = |\text{PLL}(s^{\text{WT}}) - \text{PLL}(s^{\text{mut}})|, \tag{1}$$

because a wild-type sequence typically tends to have a higher log-likelihood in a protein language model and a mutation disrupts the protein with a lower log-likelihood.

Classification objevtive function To perform the classification of pathogenic vs benign genetic variant via the siamese network, we will utilize a binary cross entropy loss. Binary cross entropy, often referred to as logarithmic loss or log loss, penalizes the model for incorrect labeling of data classes by monitoring deviations in probability during label classification. In order to fine-tune the siamese network using binary cross entropy loss, we calibrate the PLLR to a probability $\hat{\sigma}$ between 0 to 1 by: $\hat{\sigma}(\lambda) = 2\sigma(\lambda) - 1$, due to the sigmoid function σ is between 0.5 to 1 for $|\text{PLLR}|$ between 0 to $+\infty$. We then fine-tune the siamese network with the binary cross entropy loss as follows:

$$\mathcal{L}_{\text{BCE}} = y \cdot \log(\hat{\sigma}(\lambda)) + (1 - y) \cdot \log(1 - \hat{\sigma}(\lambda)). \tag{2}$$

Thus, the objective is to maximize the PLLR to distinct the MLM logits between s^{WT} and s^{mut} if the mutation is pathogenic and vice versa.

3.1 Attack Models

To evaluate the adversarial vulnerabilities of GFMs, we implement two distinct attack strategies: the FGSM and a soft prompt attack. While FGSM introduces direct input-space perturbations, the soft prompt attack operates in the embedding space and includes two variants: a confidence hijack that disrupts decision margins, and a targeted version that specifically shifts benign variants toward pathogenic predictions.

3.1.1 Fast Gradient Sign Method (FGSM)

The FGSM (Goodfellow et al., 2014) is a widely used adversarial attack that generates perturbed inputs by leveraging the gradient of the loss function with respect to the input. Given an original

input sequence $s \in \mathcal{A}^L$ and its corresponding label $y \in \{0, 1\}$, FGSM constructs an adversarial input s_{adv} using the following formula:

$$s_{\text{adv}} = s + \epsilon \cdot \text{sign}(\nabla_s \mathcal{L}(s, y)),$$

where $\mathcal{L}(s, y)$ is the classification loss, $\nabla_s \mathcal{L}$ denotes the gradient of the loss with respect to the input sequence embeddings, and ϵ is a small scalar controlling the perturbation magnitude. The sign function ensures that the perturbation is aligned with the direction that most increases the loss, while keeping the perturbation norm minimal.

In this study, FGSM is used to evaluate the robustness of the ESM-based variant effect predictor by introducing subtle perturbations to the wild-type and mutant sequences. We apply FGSM to the embedding space of each input and monitor changes in pseudo-log-likelihood ratio (PLLR) and downstream classification outcomes. This allows us to identify cases where the model is overly sensitive to near-identical input sequences, thereby revealing adversarial weaknesses in genomic prediction tasks.

3.1.2 Soft Prompt Attack

To assess the adversarial vulnerability of the GFMs, we implement a soft prompt attack in which a trainable embedding sequence is prepended to both wild-type and mutant sequences. In contrast to prompt-tuning with a frozen model, we allow full model fine-tuning during the attack to maximize adversarial impact. Both the soft prompt and the underlying protein language model are updated during optimization, enabling more effective perturbation of internal representations.

1. **Confidence Hijack Objective.** In the full-class adversarial setting, we aim to mislead the model by increasing its confidence in incorrect predictions across both classes. Specifically, we retain the pseudo-log-likelihood ratio (PLLR) computation $\lambda = |\text{PLL}(s^{\text{WT}}) - \text{PLL}(s^{\text{mut}})|$, and define a calibrated probability $\hat{\sigma}(\lambda) = 2\sigma(\lambda) - 1$, where $\sigma(\cdot)$ is the sigmoid function. For a ground-truth label $y \in \{0, 1\}$, the adversarial objective flips the labels and maximizes the binary cross entropy loss as:

$$\mathcal{L}_{\text{attack}} = (1 - y) \cdot \log(\hat{\sigma}(\lambda)) + y \cdot \log(1 - \hat{\sigma}(\lambda)). \tag{3}$$

   This objective explicitly encourages benign variants ($y = 0$) to exhibit high PLLRs (pathogenic-like) and pathogenic variants ($y = 1$) to exhibit low PLLRs (benign-like), thereby increasing misclassification confidence.

2. **Targeted Soft Prompt Attack (Benign→Pathogenic).** In the targeted one-class attack scenario, we optimize the soft prompt solely to misclassify benign variants. Given a mask over benign examples ($y = 0$), we define the attack loss as:

$$\mathcal{L}_{\text{benign}} = -\log(\hat{\sigma}(\lambda)), \quad \text{for } y = 0. \tag{4}$$

   This targeted objective forces the model to output high PLLR values for benign variants, thereby mimicking confident pathogenic predictions. No gradients are applied to pathogenic examples, preserving their outputs while selectively increasing false positive rates.

Optimization and Evaluation During attack optimization, only the soft prompt parameters are updated via gradient descent. The model is evaluated under the original labels using ROC AUC, AUPR, and threshold-based metrics to assess the degradation in classification performance caused by the adversarial soft prompt. This setup allows us to isolate the effect of soft input perturbations on model robustness without altering biological content.

3.2 Settings

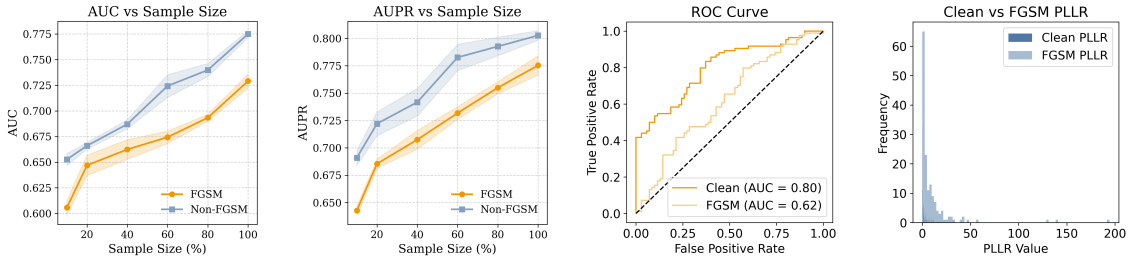
To evaluate the robustness of our variant effect prediction framework, we implemented two distinct adversarial strategies: FGSM and soft prompt attacks. For the FGSM setup, we applied perturbations directly to the input embeddings of the GFM using the Fast Gradient Sign Method, with an attack strength of $\epsilon = 0.01$. The gradient was computed with respect to the classification loss, and the perturbed embeddings were then passed through a frozen GFM encoder followed by a trainable classification head. In contrast, the soft prompt attack prepended $n = 10$ learnable prompt tokens to each input sequence. These prompt embeddings, initialized via Xavier uniform distribution (Glorot & Bengio, 2010), were optimized jointly with the classification head. For both experiments, we trained the models using the Adam optimizer with a learning rate of 1×10^{-4} (Kingma, 2014) and batch size of 4 over 10 epochs. The loss function used was binary cross-entropy, and training was conducted on a single A100 GPU. This design enables a direct comparison between perturbation-based (FGSM) and representation-based (soft prompt) adversarial robustness under a consistent training regime. This study focuses on clinically relevant variant sets associated with inherited CM and ARM. We draw on a previously curated collection of rare missense variants labeled as pathogenic or benign, organized using a cohort-driven framework for disease-specific analysis, as described by Zhang et al. (2021). The dataset is publicly accessible via `https://github.com/ImperialCardioGenetics/CardioBoost_manuscript`. To supplement this resource, compiled data are available on Zenodo at `https://zenodo.org/records/13397296` (Zhan & Zhang, 2024). The code is publicly available at: `https://anonymous.4open.science/r/SafeGenes-9086/`.

4 Results

4.1 Robustness Analysis on the Cardiomyopathy Dataset

To evaluate the impact of adversarial perturbations on variant effect prediction, we performed a robustness analysis using the CM dataset under both clean and FGSM conditions. Figure 2 provides a comprehensive evaluation of the adversarial robustness of PLLR-based variant effect prediction on the CM dataset.

Figure 2 (a) shows that AUC scores consistently improve as the training set size increases. However, across all sampling percentages, the model evaluated with FGSM-perturbed inputs exhibits a clear performance drop relative to its clean counterpart. For instance, at full data (100%), the clean model reaches an AUC of 0.78, while the FGSM model achieves only 0.73. This gap persists even at lower data fractions, revealing that adversarial inputs cause systematic degradation in discriminative power. Figure 2 (b) displays a similar pattern in AUPR, reflecting the precision-recall tradeoff. Without FGSM, the model reaches a peak AUPR of over 0.80, whereas the FGSM-exposed model underperforms by approximately 3–5 percentage points at every sample size. These results reinforce that the PLLR-based scoring is not only sensitive to adversarial noise but also exhibits increasing robustness with larger training sets. In conclusion, the consistent drop in both AUC and AUPR

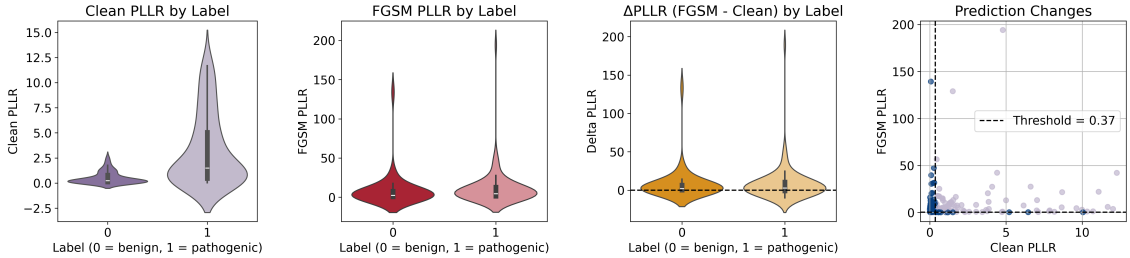


(a) AUC vs. Training Sample Size

(b) AUPR vs. Training Sample Size

(c) ROC curves under clean and FGSM conditions

(d) Histogram of PLLR under clean and FGSM conditions

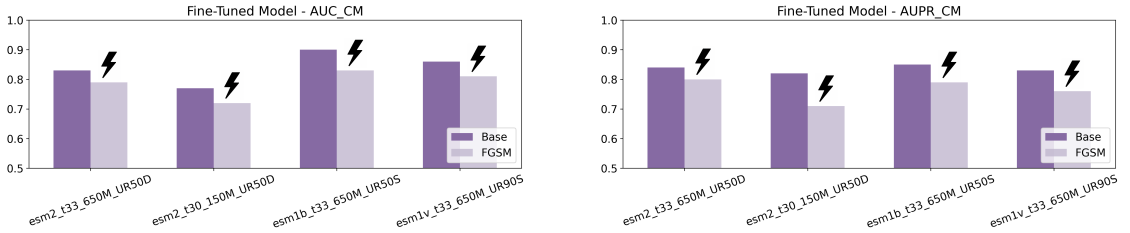


(e) Clean PLLR distribution by label

(f) FGSM PLLR distribution by label

(g) Δ PLLR distribution (FGSM - Clean)

(h) Prediction flips under FGSM perturbation



(i) AUC on fine-tuned models (clean vs. FGSM)

(j) AUPR on fine-tuned models (clean vs. FGSM)

Figure 2: Comprehensive robustness analysis of PLLR-based variant effect prediction on the cardiomyopathy (CM) dataset. (a–b) Performance degradation under FGSM across varying sample sizes. (c–d) ROC and histogram views show confidence collapse. (e–g) Violin plots reveal label-wise distribution shifts in PLLR. (h) Threshold crossing plot shows adversarially flipped predictions. (i–j) AUC and AUPR performance degradation across fine-tuned model variants under FGSM attack.

under adversarial evaluation suggests that models trained for variant effect prediction in clinical genomics must be stress-tested with adversarial scenarios to ensure reliability in real-world settings.

Figure 2(c–h) provides a comprehensive view of how FGSM perturbations affect model confidence and prediction stability. The ROC curves in Figure 2(c) show that performance degrades consistently under adversarial perturbation: while the clean evaluation yields an AUC of 0.80, FGSM reduces this to 0.62. This drop reflects a loss in discriminative capacity caused by small, structured input perturbations.

To further understand how prediction scores shift, Figure 2(d) compares the distributions of PLLR values under clean and adversarial conditions. Clean PLLR scores exhibit clear separation, with pathogenic variants having higher values. However, FGSM shifts the entire PLLR distribution downward, compressing the score range and blurring the margin between classes. Figures 2(e) and (f) break down PLLR distributions by label. Under clean conditions (e), pathogenic variants exhibit high PLLR with wide spread, while benign variants are tightly centered at low values. After FGSM, the pathogenic distribution becomes narrower and overlaps more with the benign class in Figure 2(f). This reduction in separation shows that FGSM weakens the signal used to distinguish variant pathogenicity. The degree of PLLR change is quantified in Figure 2(g), which plots the Δ PLLR (FGSM - clean) for each label. Pathogenic variants show a larger and more variable drop in PLLR than benign ones. This asymmetric sensitivity suggests that FGSM has a stronger effect on confident predictions—likely due to steeper gradient directions near high-PLLR regions.

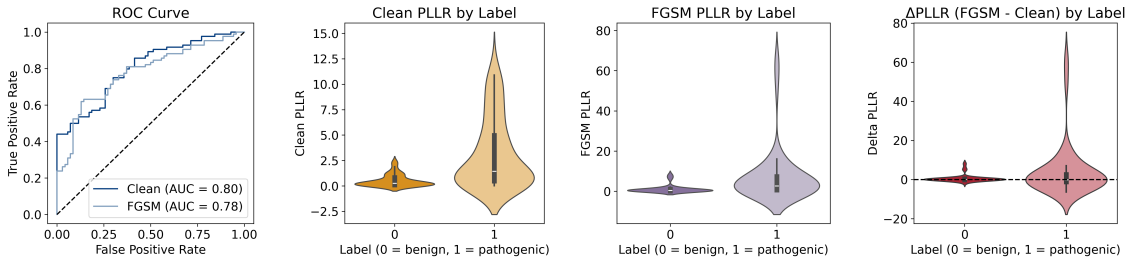
Figure 2(h) visualizes which samples switch predicted labels due to FGSM. Clean vs. FGSM PLLR values are plotted, and purple points mark samples that cross the decision threshold. Many lie just above the threshold under clean conditions and are pushed below it after attack. This illustrates that the model's decision boundary is not robust, especially for borderline cases.

Lastly, to evaluate whether the observed robustness patterns generalize across models, we further analyze AUC and AUPR performance of four fine-tuned variant effect predictors under clean and FGSM conditions. As shown in Figure 2(i–j), FGSM consistently reduces model performance across all architectures, confirming that adversarial vulnerability is not model-specific. The performance degradation is most pronounced in the AUPR metric, where even high-capacity models like ESM1b and ESM1v suffer drops of up to 5 percentage points. These results extend our earlier findings by demonstrating that adversarial perturbations universally reduce discriminative power—even in robust fine-tuned models—emphasizing the need for integrated robustness during model development.

Together, these analyses reveal that PLLR-based scoring is vulnerable to adversarial perturbations, particularly for confident pathogenic predictions. The consistent degradation in both AUC and score separability underscores the need for adversarial robustness in genomic language models.

4.2 Robustness Analysis on the Arrhythmia Dataset

We further evaluate adversarial robustness on the ARM dataset, with results shown in Figure 3. The ROC curve in Figure 3(a) indicates a moderate drop in classification performance under FGSM perturbation, where AUC declines from 0.80 to 0.78. Although this drop is smaller than on the CM dataset, it still reflects a sensitivity to input perturbations.

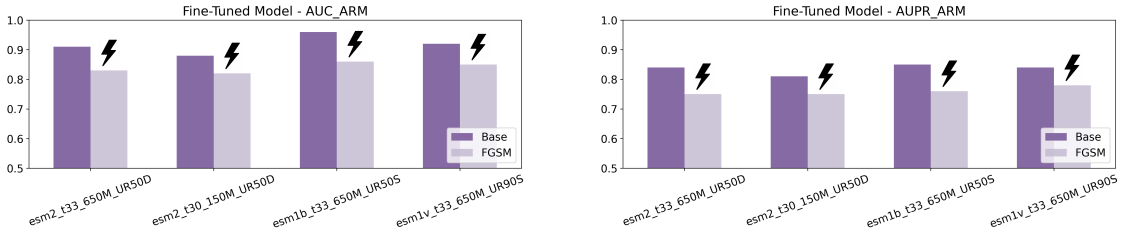


(a) ROC curve: clean vs FGSM conditions

(b) Clean PLLR distribution by label

(c) FGSM PLLR distribution by label

(d) Δ PLLR (FGSM - Clean) by label



(e) AUC on fine-tuned models (clean vs. FGSM)

(f) AUPR on fine-tuned models (clean vs. FGSM)

Figure 3: Evaluation on the arrhythmia (ARM) dataset under clean and adversarial (FGSM) conditions. (a) ROC performance slightly degrades under FGSM. (b–d) Violin plots show score distribution compression and asymmetric vulnerability across variant labels. (e–f) AUC and AUPR performance degradation across GFMs under FGSM perturbation.

Figures 3(b) and (c) show the PLLR distributions stratified by ground-truth labels. Under clean conditions (b), the model assigns clearly separable scores: benign variants center tightly near zero, while pathogenic variants occupy a broader, higher range. However, FGSM reduces the variance and separation in the pathogenic distribution (c), compressing the score space and pushing many predictions closer to the decision boundary.

The ΔPLLR violin plot in Figure 3(d) quantifies the impact of FGSM perturbations per label. Pathogenic variants (label=1) show a broader and more negative shift in PLLR, confirming that confident high-scoring predictions are more easily degraded. In contrast, benign predictions remain relatively stable. This pattern suggests that the model's internal representations for pathogenic variants are more brittle under adversarial conditions, consistent with findings from the CM dataset. Figures 3(e) and (f) summarize the impact of FGSM perturbations on AUC and AUPR across multiple GFMs evaluated on the ARM dataset. Across all model architectures, FGSM leads to a consistent drop in performance, confirming its effectiveness in disrupting decision boundaries even in models trained for clinical variant interpretation. Smaller models such as ESM2 with 150M parameters experience the largest degradation, while higher-capacity models like ESM1b and ESM1v, although more resilient, still exhibit large performance loss. These results reinforce that no

model is inherently robust and that adversarial perturbations can significantly impair the predictive reliability of GFMs in sensitive clinical contexts such as arrhythmia variant classification.

Overall, these results demonstrate that although the ARM model is slightly more robust than the CM model in AUC terms, it remains susceptible to adversarial degradation, especially for pathogenic predictions with initially high confidence.

4.3 Targeted Soft Prompt Attack (Benign→Pathogenic) on CM

We evaluate the effectiveness of a targeted soft prompt attack designed to selectively elevate the PLLR of benign variants, forcing them to be misclassified as pathogenic. Figure 4(a) shows that benign predictions shift significantly after the attack, with the PLLR distribution moving rightward while pathogenic scores remain largely unchanged. This is statistically confirmed by a paired t-test on benign PLLR scores before and after the attack ($p = 9.23 \times 10^{-4}$), annotated in the figure.

To quantify per-class vulnerability, Figure 4(b) presents the change in PLLR (ΔPLLR) for each label. As expected, the attack has a clear one-sided effect: benign variants exhibit a positive shift in PLLR, while pathogenic variants are unaffected. This validates that the soft prompt optimization objective successfully targets only the benign class.

The impact on downstream decision-making is illustrated in Figure 4(c), which plots the precision-recall (PR) curve before and after the attack. The area under the PR curve (AUPR) drops from 0.69 to 0.65, indicating a reduced ability to separate benign from pathogenic variants, driven by an increased false positive rate.

Finally, Figure 4(d) provides a sorted waterfall plot of ΔPLLR for benign variants. While the majority of benign variants exhibit modest changes, a subset shows large increases in PLLR, exceeding +1.0. These outliers demonstrate that the soft prompt is particularly effective at amplifying confidence on select benign cases, resulting in focused prediction flips.

Together, these results reveal the asymmetric nature of the targeted attack: it preserves the pathogenic class while systematically degrading the model's trustworthiness on benign samples. This highlights the importance of adversarial robustness measures in variant effect prediction.

4.4 Targeted Soft Prompt Attack (Benign→Pathogenic) on ARM

To evaluate generalizability, we applied the same targeted soft prompt attack to the ARM dataset. The attack again targets benign variants, aiming to increase their PLLR and induce misclassification. Figure 4(e) shows the PLLR distribution before and after the attack for both benign and pathogenic variants. As in CM, benign variants exhibit a rightward shift in PLLR, while the pathogenic distribution remains stable. This indicates that the attack retains its one-sided behavior across datasets. The ΔPLLR violin plot in Figure 4(f) further confirms this asymmetric effect. The benign class experiences a significant positive shift, while pathogenic variants show little change, consistent with the CM results. Figure 4(g) shows that the AUPR drops from 0.83 to 0.79, reflecting degraded performance due to the increased false positive rate. This consistent performance drop across datasets demonstrates that the attack generalizes. Finally, Figure 4(h) presents the ΔPLLR waterfall plot for benign ARM variants. Again, we observe that while many benign samples shift modestly, a subset displays extreme increases in PLLR—evidence that the soft prompt focuses on high-impact samples even in new distributions.

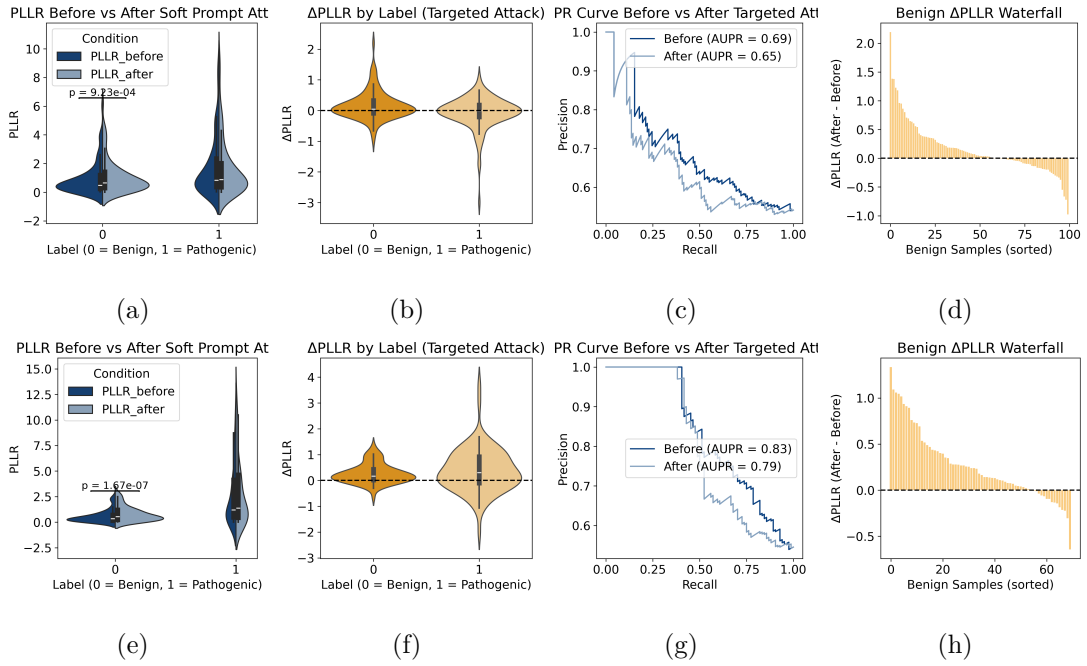


Figure 4: Targeted soft prompt attack results. (a–d) CM dataset; (e–h) ARM dataset. (a, e) PLLR before vs. after. (b, f) ΔPLLR by label. (c, g) PR curves. (d, h) Benign ΔPLLR waterfall plots.

These results demonstrate that the targeted attack is not only effective in CM but also generalizes to ARM, indicating a broader vulnerability in model confidence estimation across datasets.

4.5 Comparative Analysis of Attack Effectiveness

To quantify and compare the impact of different adversarial strategies, we evaluate three attack methods—FGSM, SPA_Confidence Hijack (soft prompt attack with confidence hijack), and SPA_Targeted Attack—across two benchmarks (CM and ARM) and two evaluation metrics (AUC and AUPR). Results are summarized in Figure 5(a–d). On both CM and ARM, we observe that all attack methods reduce model performance to varying degrees. FGSM consistently causes moderate drops in both AUC and AUPR, reflecting its role as a standard untargeted perturbation baseline. The confidence hijack attack (SPA_Confidence Hijack) leads to more pronounced performance degradation, especially in AUPR (Figure 5(b, d)), confirming its effectiveness at collapsing the confidence margin between classes. The most severe performance degradation is observed under the SPA_Targeted Attack condition, which directly optimizes the PLLR of benign samples to mimic pathogenic profiles. This method consistently yields the lowest AUC and AUPR across all model variants, particularly on the ARM benchmark (Figure 5(c–d)). This sharp performance decline highlights the increased vulnerability of models to targeted soft prompt optimization, especially under low-signal or low-data settings.

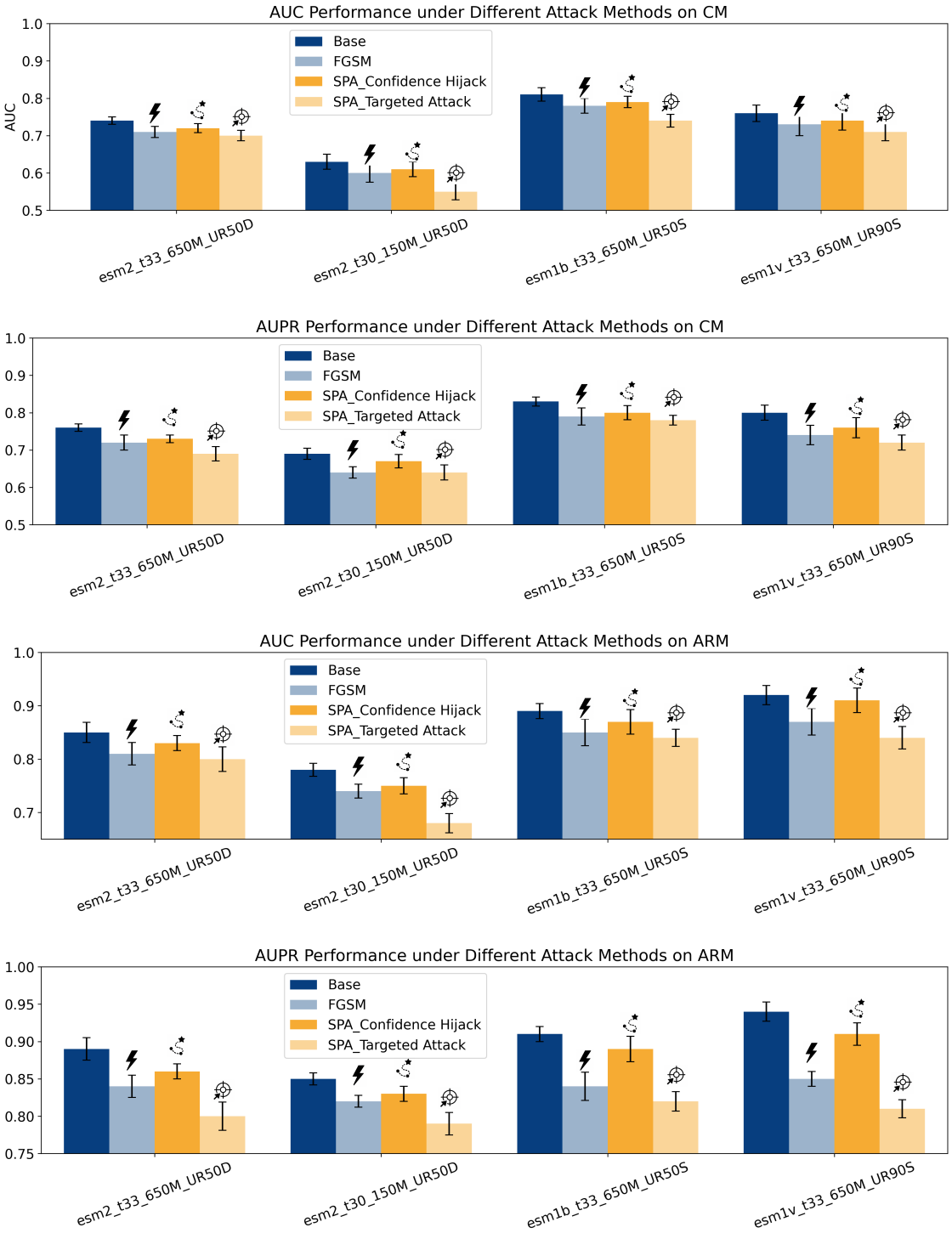


Figure 5: Performance under different adversarial attack methods across CM and ARM datasets. (a) AUC on CM. (b) AUPR on CM. (c) AUC on ARM. (d) AUPR on ARM. FGSM, SPA_Confidence Hijack, and SPA_Targeted Attack all degrade model performance, with the targeted attack showing the strongest effect.

Overall, the comparative results demonstrate that soft prompt-based attacks can be more effective than input-space perturbations (like FGSM), particularly when targeted or designed to manipulate model confidence.

5 Conclusion

In this work, we present SafeGenes, a comprehensive framework for evaluating the adversarial robustness of GFMs. While GFMs such as ESM1b and ESM1v have demonstrated strong performance in variant effect prediction, our findings reveal that they remain vulnerable to both input-space and latent-space perturbations. Through a combination of FGSM-based attacks and novel soft prompt-based adversarial probes, we uncover consistent and significant performance degradation across multiple model architectures and disease-specific datasets. Our targeted soft prompt attack induces confident misclassification of benign variants by manipulating internal representations, without altering the input sequence. These semantic-level vulnerabilities expose a critical blind spot in current defense paradigms, which often rely on confidence scores or input-level cues. Our results suggest that effective genomic model defenses must operate within the latent space, tracking shifts in internal representations to detect adversarial manipulation. As genomic foundation models continue to be integrated into clinical workflows, ensuring their robustness under adversarial conditions is essential for safe and reliable deployment. SafeGenes lays the groundwork for both probing and improving the trustworthiness of sequence-based models in genomics and offers actionable insights for the design of future defense strategies in biomedical AI.

Broader Impact Statement

Advances in large language models have transformed the way we analyze language, images, and scientific data—and they are now reshaping biomedicine. GFMs, which apply language modeling principles to DNA and protein sequences, are emerging as powerful tools to predict the functional effects of genetic mutations. These models promise to accelerate the diagnosis of inherited diseases, personalize treatments, and uncover novel insights into human biology. However, with increasing adoption in clinical genomics, the question arises: how secure and trustworthy are these models when faced with unexpected or malicious inputs?

In this study, we present *SafeGenes*, a framework to evaluate the adversarial robustness of GFMs. We show that these models, despite their impressive performance, remain vulnerable to carefully crafted perturbations—either in the form of subtle changes to input sequences or imperceptible manipulations of their internal prompt structure. Our targeted soft prompt attacks can cause confident misclassification of benign variants, revealing latent vulnerabilities that are not captured by traditional robustness tests. These vulnerabilities are particularly concerning in medical contexts, where decisions influenced by model outputs can directly impact patient outcomes.

Beyond demonstrating specific failure cases, our findings have broader implications for the future of trustworthy AI in genomics. Current defense mechanisms, which rely on input-level checks or confidence scores, may not suffice when attacks operate within the semantic space of the model itself. Our work suggests a path forward: integrating robustness evaluation into the development cycle of genomic AI systems and designing future defenses that monitor changes in internal representation space. Just as medical instruments must pass rigorous safety tests, genomic models must be stress-tested for reliability before clinical use. *SafeGenes* is a step toward that future.

References

Jake Bouvrie. Notes on convolutional neural networks. 2006.

Nadav Brandes, Grant Goldman, Charlotte H Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, 2023.

Junjie Chen, Wendy Hui Wang, and Xinghua Shi. Differential privacy protection against membership inference attack on machine learning for genomic data. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pp. 26–37. World Scientific, 2020.

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

Narmin Ghaffari Laleh, Daniel Truhn, Gregory Patrick Veldhuizen, Tianyu Han, Marko van Treeck, Roman D Buelow, Rupert Langer, Bastian Dislich, Peter Boor, Volkmar Schulz, et al. Adversarial attacks and adversarial robustness in computational pathology. *Nature communications*, 13(1): 5711, 2022.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Zhicong Huang, Erman Ayday, Jacques Fellay, Jean-Pierre Hubaux, and Ari Juels. Genoguard: Protecting genomic data against brute-force attacks. In *2015 IEEE Symposium on Security and Privacy*, pp. 447–462. IEEE, 2015.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41, 2022.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Tsung-Ting Kuo, Xiaoqian Jiang, Haixu Tang, XiaoFeng Wang, Arif Harmanci, Miran Kim, Kai Post, Diyue Bu, Tyler Bath, Jihoon Kim, et al. The evolving privacy and security concerns for genomic data analysis and sharing as observed from the idash competition. *Journal of the American Medical Informatics Association*, 29(12):2182–2190, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Daniel Mas Montserrat and Alexander G Ioannidis. Adversarial attacks on genotype sequences. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8 (7):579–588, 2009.

Martin Riedmiller and A Lernen. Multi layer perceptron. *Machine learning lab special lecture, University of Freiburg*, 24, 2014.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Heorhii Skovorodnikov and Hoda Alkhzaimi. Fimba: Evaluating the robustness of ai in genomics via feature importance adversarial attacks. *arXiv preprint arXiv:2401.10657*, 2024.

Huanjing Wang, Qianxin Liang, John T Hancock, and Taghi M Khoshgoftaar. Feature selection strategies: a comparative analysis of shap-value and importance-based methods. *Journal of Big Data*, 11(1):44, 2024.

Huixin Zhan and Zijun Zhang. DYNA: Disease-Specific Language Model for Variant Pathogenicity [Data set], 2024. URL `https://doi.org/10.5281/zenodo.13397296`. Zenodo.

Huixin Zhan, Jason H Moore, and Zijun Zhang. A disease-specific language model for variant pathogenicity in cardiac and regulatory genomics. *Nature Machine Intelligence*, pp. 1–11, 2025.

Jiaming Zhang, Jitao Sang, Qi Yi, and Changsheng Xu. Introducing foundation models as surrogate models: Advancing towards more practical adversarial attacks. *arXiv preprint arXiv:2307.06608*, 2023.

Xiaolei Zhang, Roddy Walsh, Nicola Whiffin, Rachel Buchan, William Midwinter, Alicja Wilk, Risha Govind, Nicholas Li, Mian Ahmad, Francesco Mazzarotto, et al. Disease-specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. *Genetics in Medicine*, 23(1):69–79, 2021.