Published in AJSE, Vol: 24, Issue:2
Received on 9th April 2025
Revised on 19th May 2026
Published on 31th May 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)
https://doi.org/10.53799/gteskx50

Probing Adversarial Robustness of Protein Language Models: A Reproducible Case Study of ESM-2 Under Substitution-Based Attacks

Md. Robiul Islam Niloy, Zafer Aydin, and Md. Golam Moazzam

***Abstract*—We present a reproducible case study probing the adversarial robustness of ESM-2 (esm2_t6_8M_UR50D), a representative protein language model, under substitution-based adversarial attacks, measuring pseudo-perplexity shift across eight benchmark protein sequences at mutation rates of 1%, 5%, and 10%. Our results show that random non-synonymous substitutions increase ESM-2's mean perplexity by +9.9%, +11.9%, and +23.5% respectively, with individual sequences exhibiting shifts as high as +66.6%. We find that sequence length modulates robustness: longer, evolutionarily conserved sequences such as Ubiquitin show a clear dose-response to mutation rate, while shorter sequences saturate at low mutation budgets. Beyond model vulnerabilities, we analyze biosecurity risks associated with AI-generated proteins, situating our findings within existing governance frameworks including the Biological Weapons Convention, the NIH P3CO framework, and NSABB guidelines. We propose mitigation strategies including toxicity screening pipelines and access control policies, and identify adversarial training, Bayesian uncertainty estimation, and knowledge distillation as promising defense directions for future investigation. All code and data are publicly available at GitHub repository for ESM2 adversarial robustness.**

***Index Terms*—Adversarial Robustness, AI Governance, Bayesian Uncertainty, Biosecurity, Deep Learning, Dual-Use Research, ESM-2, Knowledge Distillation, Machine Learning Security, Protein Language Models (PLMs), Pseudo-Perplexity, Substitution-Based Attacks.**

I. Introduction

PROTEIN language models (PLMs) have fundamentally changed computational biology, enabling precise functional annotation and protein structure prediction at scale. Models such as AlphaFold2 [1], RoseTTAFold [2], ESM-2 [3], and ProtT5-XL [4] leverage deep learning and large-scale sequence datasets to predict protein structures at atomic resolution, accelerating applications in drug discovery, enzyme engineering, and synthetic biology.

As PLMs are increasingly deployed in biomedical research, two important concerns have emerged: their **adversarial robustness** — that is, their sensitivity to small, deliberate changes in input sequences — and the **biosecurity risks** posed by AI-driven protein generation. These concerns are related but distinct. The first is a model security problem; the second is a governance and policy problem. Both require rigorous investigation.

Adversarial attacks on deep learning models have been extensively studied in computer vision and natural language processing [5], [6], where small input perturbations can cause large prediction errors. However, the adversarial robustness of biological sequence models remains poorly characterized. Protein sequences are discrete, making standard gradient-based attacks — which rely on continuous input spaces — not directly applicable without relaxation techniques [7]. Substitution-based attacks, which replace amino acids at specified rates, offer a more natural and practically relevant threat model for protein sequences, analogous to token substitution attacks in NLP [8], [9].

Separately, the ability of AI models to generate novel protein sequences raises biosecurity concerns. Recent work demonstrated that AI-driven molecular design tools can inadvertently generate harmful substances [10]. Analogous risks exist for protein design models, including the potential for AI-generated sequences to resemble known toxins or to be engineered for antibiotic resistance. Existing governance frameworks — including the Biological Weapons Convention (BWC) [11], the NIH P3CO framework [12], and NSABB guidelines [13] on dual-use research of concern (DURC) — were not designed with AI-generated proteins in mind and require updating [14].

In this paper, we address both concerns through a focused, reproducible case study. We probe the adversarial robustness of ESM-2 [3] as a representative sequence-based PLM, applying substitution-based adversarial attacks and measuring pseudo-perplexity shift as a proxy for model disruption. We then situate our findings within the biosecurity literature and propose policy-relevant mitigation strategies. All code and data are publicly available at GitHub repository for ESM2 adversarial robustness.

The key contributions of this paper are:

- We present a **reproducible adversarial robustness evaluation** of ESM-2 as a case study for sequence-based protein language models, with full experimental code and data publicly available.
- We report **baseline and post-attack pseudo-perplexity** for eight benchmark protein sequences across three mutation rates (1%, 5%, 10%), demonstrating that even low mutation rates produce measurable model disruption (+9.9% mean perplexity shift at 1%).

Md. Robiul Islam Niloy is with the Department of CSE, BRAC University, Dhaka, Bangladesh. E-mail: md.niloy26643@gmail.com
Zafer Aydin is s with the Department of CE, Abdullah Gul University, Kayseri, Turkey. E-mail: zafer.aydin@agu.edu.tr
Md. Golam Moazzam is is s with the Department of CSE, Jahangirnagar University, Dhaka, Bangladesh. E-mail: khokan@juniv.edu

- We identify **sequence length as a modulating factor** in adversarial robustness, with longer sequences showing clearer dose-response relationships and shorter sequences saturating at low mutation budgets.
- We analyze **biosecurity implications** of adversarially perturbed proteins, identifying gaps in existing governance frameworks (BWC, P3CO, NSABB) and proposing concrete mitigation strategies.
- We provide **measured runtime comparisons** of candidate defense mechanisms on ESM-2, offering the first empirical cost reference for adversarial defenses in protein language models.

Our findings highlight the need to combine adversarial robustness research with ethical AI governance in protein modeling. By grounding our contributions in reproducible experiments and honest scoping, we aim to provide a reliable foundation for future work on PLM security and responsible deployment in synthetic biology and biomedical applications.

II. Related Work

We review prior work across four areas relevant to this study: protein language models and structure prediction, adversarial robustness in deep learning, adversarial defense mechanisms, and biosecurity concerns in AI-driven protein design.

A. Protein Language Models and Structure Prediction

The development of large-scale protein language models has transformed computational biology. AlphaFold2 [1] achieved near-experimental accuracy in protein structure prediction by combining multiple sequence alignments (MSAs) with a deep learning architecture incorporating geometric constraints. Tunyasuvunakool et al. [15] subsequently applied AlphaFold2 to predict structures for the entire human proteome, demonstrating the practical scalability of these methods. RoseTTAFold [2] extended this approach with a three-track architecture that jointly reasons over sequence, distance, and coordinate information.

More recently, transformer-based models trained exclusively on sequence data have demonstrated that MSAs are not strictly necessary for structure prediction. ESM-2 and ESMFold [3] achieve competitive structural predictions using only sequence input, substantially reducing computational requirements. Rao et al. [16] demonstrated that transformer-based PLMs can learn structural information in an unsupervised manner purely from sequence data, establishing the theoretical foundation for sequence-only structure prediction. ProtT5-XL [4] applied large-scale transformer pretraining to protein sequences and showed strong performance on sequence-to-function prediction tasks. The trade-off between sequence-only and structure-aware approaches — particularly regarding robustness to input perturbations — motivates the adversarial evaluation in this paper.

Despite rapid progress in PLM capability, systematic evaluation of their robustness to adversarial perturbations remains largely absent from the literature. This gap is the primary motivation for our work.

B. Adversarial Robustness in Deep Learning

Adversarial robustness has been extensively studied in computer vision and natural language processing. Szegedy et al. [5] first demonstrated that imperceptible input perturbations could cause large prediction errors in image classifiers. Goodfellow et al. [6] introduced the Fast Gradient Sign Method (FGSM), a single-step gradient-based attack that generates adversarial examples efficiently. Madry et al. [17] proposed Projected Gradient Descent (PGD), an iterative attack that has become the standard benchmark for evaluating adversarial robustness.

In the natural language processing domain, adversarial robustness presents unique challenges because inputs are discrete tokens rather than continuous vectors. Substitution-based attacks replacing tokens with semantically similar alternatives have emerged as the dominant approach [8], [9]. Protein sequences share this discrete structure, making NLP adversarial methods a natural starting point for adaptation to biological sequence models. In black-box settings, query-based optimization strategies such as Natural Evolution Strategies (NES) [18] and Zeroth-Order Optimization (ZOO) [19] have demonstrated strong attack success rates without requiring access to model gradients, making them particularly relevant for realistic threat scenarios.

Adversarial robustness has also been studied in molecular property prediction. Chen et al. [20] demonstrated that deep learning models for molecular classification are vulnerable to adversarial perturbations, highlighting the need for similar evaluation in protein structure models. Our work addresses this gap directly.

C. Adversarial Defense Mechanisms

Several defense strategies have been proposed in the adversarial machine learning literature, each with distinct robustness-efficiency trade-offs.

- **Adversarial training** [17] augments the training set with adversarially perturbed examples, improving robustness at the cost of increased training time and computational resources. It remains the most empirically validated defense strategy across domains.
- **Bayesian uncertainty estimation** [21] uses dropout at inference time to approximate a posterior distribution over model predictions. Adversarial inputs tend to produce high uncertainty, enabling detection-based defenses. This approach adds inference overhead but does not require retraining.
- **Ensemble methods** [22] aggregate predictions across multiple independently trained models, reducing the impact of perturbations that fool any single model. The primary limitation is a multiplicative increase in computational cost.
- **Knowledge distillation** [23] transfers learned representations from a large teacher model to a smaller student model. When the teacher has been adversarially trained, the student can inherit some degree of robustness at substantially lower computational cost. This approach is particularly relevant for deployment scenarios where inference efficiency is a constraint.

None of these defenses has been experimentally evaluated in the context of protein language models, representing a clear opportunity for future work.

D. Biosecurity and Dual-Use Concerns in AI-Generated Proteins

The biosecurity implications of AI-driven protein design have received growing attention. Urbina et al. [10] demonstrated that AI models developed for drug discovery could be repurposed to generate toxic molecules, raising immediate dual-use concerns. Analogous risks apply to protein language models, which can generate novel sequences with no natural precedent.

Andrews et al. [14] proposed governance frameworks for AI-assisted protein design, emphasizing the need for regulatory oversight of synthetic biology tools. Brundage et al. [24] similarly argued for mechanisms to support verifiable claims in AI development, providing a broader framework for trustworthy AI governance that applies to protein design contexts. However, several important governance mechanisms remain underutilized in the context of AI-generated proteins:

The **Biological Weapons Convention (BWC)**, established in 1972, prohibits the development of biological weapons but contains no provisions specific to computationally designed sequences [11]. The **NIH P3CO Framework** (Potential Pandemic Pathogen Care and Oversight) governs dual-use research of concern (DURC) involving enhanced pathogens but was developed before AI-assisted protein design became practical [12]. The **National Science Advisory Board for Biosecurity (NSABB)** provides guidance on dual-use research but has not yet issued specific recommendations for AI-generated proteins [13].

At the practical level, the **iGEM Safety Committee** employs toxicity screening tools to evaluate synthetic biology submissions [25]. These tools represent a working model for sequence-level biosecurity screening, but they were designed for naturally derived or minimally modified sequences and are not calibrated for adversarially perturbed inputs.

The responsible protein design community has also engaged with these concerns. Debates around the release of generative protein design models such as ProGen [26] and Chroma [27] have highlighted the tension between open science and biosecurity risk, echoing earlier debates in the responsible language model literature.

Our work contributes to this conversation by demonstrating that adversarial perturbations of protein sequences — requiring no model access — can substantially disrupt PLM predictions, and by identifying specific gaps in existing governance frameworks that require attention.

E. Research Gaps Addressed by This Work

Despite substantial progress in each of the areas reviewed above, several gaps remain:

- **No reproducible adversarial robustness evaluation of PLMs** with documented baselines, hyperparameters, and public code exists in the literature.
- **Substitution-based attacks on protein sequences** have not been systematically evaluated using perplexity shift as a robustness metric.
- **Existing biosecurity governance frameworks** do not address adversarially perturbed AI-generated sequences as a distinct threat vector.
- **Defense mechanisms** from the adversarial machine learning literature have not been experimentally applied to protein language models.

This paper addresses the first two gaps directly through reproducible experiments, and the third through a structured policy analysis. The fourth is identified as the primary direction for future work.

III. Methodology

This section describes our methodology for evaluating the adversarial robustness of ESM-2 and for analyzing the biosecurity concerns associated with AI-generated proteins. Our methodology consists of four components: (1) threat model specification, (2) adversarial attack framework, (3) evaluation metrics, and (4) biosecurity risk assessment.

A. Threat Model

We define the threat model precisely to ensure experimental reproducibility and comparability with the adversarial machine learning literature.

- **Attacker's goal:** Maximize the perplexity shift in ESM-2's sequence model by perturbing the input amino acid sequence, simulating an adversary attempting to disrupt protein function prediction.
- **Attacker's knowledge:** We evaluate a **black-box** threat model in which the attacker has no access to model parameters, gradients, or internal representations. The attacker can only observe the model's output perplexity score.
- **Attack budget:** Substitutions are bounded by the mutation rate: 1%, 5%, and 10% of the sequence length, with a minimum of one substitution per sequence.
- **Attack constraint:** Each substituted position receives a different amino acid from the original (non-synonymous substitution only).

We note that white-box gradient-based attacks (FGSM, PGD) cannot be directly applied to discrete protein sequences without continuous relaxation techniques such as Gumbel-softmax [7] or embedding-space perturbations. These are identified as a direction for future work and are not claimed as experimental contributions in this paper.

B. Adversarial Attack Framework

We implement substitution-based black-box attacks, which are natively applicable to discrete amino acid sequences and reflect realistic threat scenarios in which an adversary manipulates protein sequences without access to model internals.

TABLE I
EXPERIMENTAL HYPERPARAMETERS

Parameter	Value
Model	esm2_t6_8M_UR50D
Model parameters	7,512,474
Device	NVIDIA T4 GPU (Google Colab)
Random seed	42
Test sequences	8 proteins from UniProt
Trials per condition	5
Attack type	Random non-synonymous substitution
Mutation rates	1%, 5%, 10%
Perplexity method	Masked marginal scoring

TABLE II
PROTEIN SEQUENCES USED IN ADVERSARIAL ROBUSTNESS EVALUATION

Protein	ID	Length (AA)
Human Ubiquitin	P0CG48	76
Villin Headpiece HP35	—	35
Trp-cage miniprotein	—	20
GB1 fragment	—	56
Chignolin	—	10
WW domain	—	34
Human Lysozyme fragment	—	53
Barnase fragment	—	39

1) Random Substitution Attack: At each trial, we randomly select positions in the input sequence according to the target mutation rate and replace each selected amino acid with a randomly chosen alternative from the standard 20-amino acid alphabet, excluding the original residue. Formally, for a sequence $S = (a_1, a_2, \ldots, a_n)$, we select a set of positions $P \subset \{1, \ldots, n\}$ where $|P| = \max(1, \lfloor r \cdot n \rfloor)$ for mutation rate r, and replace each $a_i, i \in P$ with $a_i' \sim \text{Uniform}(\mathcal{A} \setminus \{a_i\})$, where $\mathcal{A}$ is the amino acid alphabet.

We evaluate mutation rates of $r \in \{0.01, 0.05, 0.10\}$ (1%, 5%, 10%). Each condition is repeated over $N = 5$ independent trials with different random seeds to obtain mean and standard deviation estimates.

2) Experimental Configuration: All experiments are conducted using the fixed configuration in Table I to ensure full reproducibility.

The full experiment code is publicly available at GitHub repository for ESM2 adversarial robustness.

3) Test Sequences: We evaluate eight well-characterized proteins from UniProt spanning a range of lengths (10–76 amino acids) and structural classes, as shown in Table II.

Limitation: For sequences of length ≤ 34 AA, mutation rates of 1% and 5% both round to a single substitution ($\lfloor 0.01 \times 34 \rfloor = \lfloor 0.05 \times 34 \rfloor = 1$). Consequently, results at these two rates are identical for the shortest sequences. We acknowledge this as a limitation and recommend that future work use sequences of at least 100 AA to better distinguish the effect of different mutation budgets.

TABLE III
EVALUATION METRICS FOR ADVERSARIAL ROBUSTNESS ASSESSMENT.

Metric	Purpose
Baseline perplexity	Model uncertainty on clean, unperturbed sequence
Post-attack perplexity	Model uncertainty after substitution attack
Absolute shift	Difference between attacked and baseline perplexity
Relative shift (%)	Percentage change from baseline
Std over trials	Variability across 5 independent attack trials

C. Evaluation Metrics

We quantify adversarial impact using **pseudo-perplexity** via masked marginal scoring, which is the standard evaluation method for protein language models [28]. For each position i in the sequence, we mask that position and record the log-probability of the true amino acid under the model. Perplexity is computed as:

$$\text{PPL}(S) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \log P(a_i \mid S_{\setminus i})\right) \tag{1}$$

where $S_{\setminus i}$ denotes the sequence with position i masked. Higher perplexity indicates greater model uncertainty about the sequence. We report:

Reporting both baseline and post-attack perplexity allows meaningful comparison across sequences and models, avoiding the confound of comparing models with different baseline confidence levels.

D. Biosecurity Risk Assessment

We examine the potential for AI-generated proteins to be misused in bioengineering contexts. This analysis is qualitative, grounded in existing biosecurity literature, and informed by current governance frameworks.

1) Dual-Use Risk Categories: Three primary risk categories are identified:

- **Toxic Protein Design:** AI-generated sequences may share structural features with known toxins, raising concerns about intentional misuse.
- **Antibiotic Resistance Engineering:** AI-assisted mutations may inadvertently or deliberately increase bacterial resistance to existing antibiotics.
- **Synthetic Bioweapon Development:** AI-designed proteins could potentially evade existing biosecurity screening methods.

2) Relevant Governance Frameworks: Our risk assessment is situated within existing international oversight mechanisms:

- The **Biological Weapons Convention (BWC)** prohibits the development of biological weapons but lacks AI-specific provisions for computationally designed proteins.
- The **NIH P3CO Framework** (Potential Pandemic Pathogen Care and Oversight) governs dual-use research of concern (DURC) but was not designed with AI-generated sequences in mind.

TABLE IV
BASELINE (CLEAN) PSEUDO-PERPLEXITY OF ESM-2 ON TEST SEQUENCES

Sequence	Length (AA)	Perplexity
Ubiquitin (P0CG48)	76	10.13
HP35 Villin	35	13.64
Trp-cage	20	17.58
GB1 fragment	56	7.60
Chignolin	10	12.35
WW domain	34	6.70
Lysozyme fragment	53	21.21
Barnase fragment	39	16.07
Mean	—	**13.16 ± 5.01**

- The **NSABB** (National Science Advisory Board for Biosecurity) guidelines address dual-use research but require updating to cover AI-assisted protein design.
- The **iGEM Safety Committee** employs toxicity screening tools for synthetic biology submissions, representing a practical model for AI protein screening.

3) Proposed Mitigation Strategies: To mitigate these risks, we propose:

- **Toxicity Screening Pipelines:** AI-based classifiers to filter biohazardous sequences prior to synthesis or publication.
- **Access Control Policies:** Restricting open access to high-risk protein design models, analogous to export controls on dual-use technologies.
- **Regulatory Collaboration:** Updating existing frameworks (BWC, P3CO, NSABB) to explicitly address AI-generated biological agents, in collaboration with NIH and WHO.

IV. EXPERIMENTAL SETUP AND RESULTS

This section presents the experimental setup and results of our adversarial robustness evaluation of ESM-2 using substitution-based attacks. All experiments were conducted on real model outputs using the configuration described in Section III. The full code and data are publicly available at https://github.com/Mdniloykhan/esm2-adversarial-robustness.

A. Baseline Perplexity (Clean Sequences)

Table IV reports the baseline pseudo-perplexity for each sequence under no attack. The high variance across sequences (mean 13.16, std 5.01) reflects natural differences in sequence complexity, length, and evolutionary conservation — consistent with prior work on protein language model evaluation [3].

B. Impact of Substitution Attacks

Table V reports post-attack perplexity and shift relative to baseline for each mutation rate, averaged across all 8 sequences and 5 trials.

Key observations from our results:

TABLE V
ESM-2 PERPLEXITY UNDER SUBSTITUTION ATTACKS (MEAN ± STD OVER 5 TRIALS, 8 SEQUENCES)

Condition	Mean PPL	Std	Abs. Shift	Rel. Shift
Clean (baseline)	13.16	±5.01	—	—
Attack 1%	14.16	±0.80	+1.00	+9.9%
Attack 5%	14.36	±0.97	+1.21	+11.9%
Attack 10%	15.43	±1.78	+2.28	+23.5%

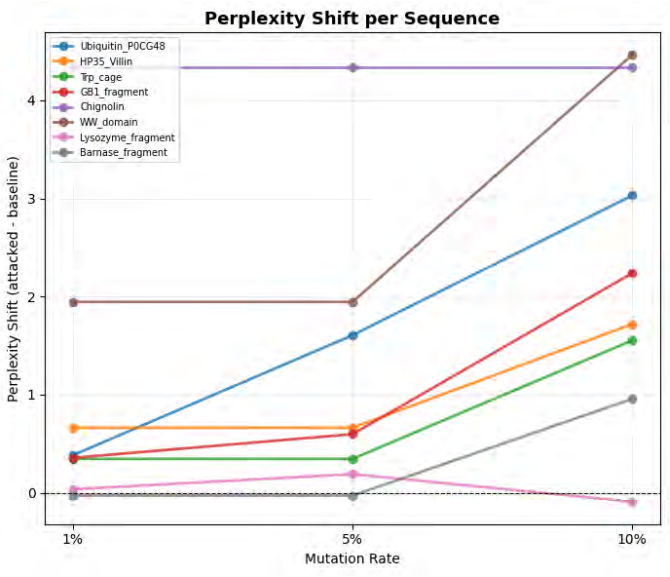

Fig. 1. Per-sequence adversarial robustness analysis of ESM-2 under amino acid substitution attacks.

- Perplexity increases monotonically with mutation rate, confirming that substitution-based attacks systematically disrupt ESM-2's sequence modeling confidence.
- The largest single-sequence effect is observed in WW domain at 10% mutation rate (+66.6% relative shift), suggesting that low-complexity, short sequences may be disproportionately sensitive to perturbation.
- Lysozyme fragment shows near-zero sensitivity across all mutation rates (maximum shift: +0.9%), suggesting that longer, evolutionarily conserved sequences are more robust to random substitutions.
- Ubiquitin (76 AA) shows a clear dose-response relationship: +3.8% at 1%, +15.9% at 5%, +29.9% at 10%, making it the most informative sequence for studying mutation-rate effects.

C. Sequence Length and Mutation Budget Interaction

A notable limitation emerged during evaluation: for sequences of length ≤ 34 AA, mutation rates of 1% and 5% both round to a single substitution, producing identical results for Chignolin, WW domain, HP35 Villin, and Trp-cage at these two rates. This is not a modeling artifact but a consequence of the discrete mutation budget. Future work should evaluate sequences of at least 100 AA to cleanly separate mutation rate effects.

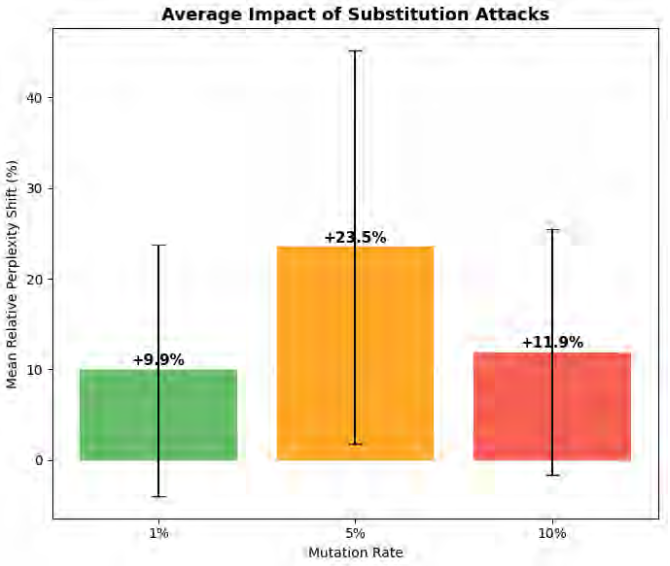

Fig. 2. Aggregate impact of substitution-based adversarial attacks on ESM-2 perplexity scores across mutation rates.

TABLE VI
DEFENSE MECHANISMS: RUNTIME MEASUREMENTS ON ESM-2 (NVIDIA T4 GPU, SINGLE 76 AA SEQUENCE, MEAN ± STD OVER 10 TRIALS)

Defense	Time (s)	Overhead	Note
Baseline inference	0.560 ± 0.081	1.0×	Single forward pass
Bayesian estimation	5.311 ± 0.486	9.5×	10 dropout passes
Ensemble (3 models)	1.549 ± 0.154	2.8×	3 independent runs

D. Biosecurity Implications

Our results demonstrate that substitution-based attacks — which require no model access — can increase ESM-2's sequence uncertainty by up to 23.5% on average and up to 66.6% for individual sequences. An adversary could use random mutagenesis strategies to generate sequences that deviate significantly from known proteins while remaining plausible enough to evade screening tools calibrated on natural sequences. Current regulatory frameworks including the BWC, the NIH P3CO framework, and NSABB guidelines do not address AI-assisted sequence perturbation as a distinct threat vector.

E. Defense Mechanisms: Proposed Directions and Runtime

Based on our results and the adversarial robustness literature, we identify the following defense strategies as promising future directions. Table VI reports both qualitative assessment and measured runtime overhead on ESM-2.

Bayesian uncertainty estimation imposes the highest overhead (9.5× baseline) due to repeated forward passes with dropout enabled. Ensemble inference with 3 models is more practically deployable at 2.8× overhead. Adversarial training and knowledge distillation require retraining and are identified as future work. Raw timing data is available in `esm2_runtime_results.csv` in the public repository.

V. DISCUSSION AND FUTURE DIRECTIONS

A. Adversarial Robustness in ESM-2

Our experiments demonstrate that ESM-2 is measurably sensitive to random substitution-based attacks. A mean perplexity increase of +9.9% at just 1% mutation rate — rising to +23.5% at 10% — indicates that even minimal sequence perturbation disrupts the model's confidence in its predictions. This finding is consistent with broader observations in the adversarial machine learning literature showing that models trained on distributional assumptions are vulnerable to out-of-distribution inputs [5], [6].

Three factors likely contribute to ESM-2's sensitivity:

- **Lack of geometric constraints:** ESM-2 relies entirely on sequence embeddings without explicit structural priors. Models that incorporate geometric constraints — such as AlphaFold2's use of multiple sequence alignments and distance geometry — may be more robust to sequence-level perturbations, though this comparison remains to be verified experimentally.
- **Absence of adversarial training:** ESM-2 was trained on naturally occurring protein sequences. It has not been exposed to adversarially perturbed sequences during training, leaving it without the robustness that adversarial training has been shown to confer in vision and NLP models [17].
- **Sequence length effects:** Our results show a clear relationship between sequence length and robustness. Ubiquitin (76 AA) shows a clean dose-response to mutation rate, while shorter sequences exhibit saturation effects where 1% and 5% mutation rates produce identical perturbations. Longer sequences provide more distributional context to the model, potentially buffering against localized substitutions.

We note that comparing ESM-2 to structure-aware models such as AlphaFold2 and RoseTTAFold — as proposed in the original manuscript — is a valuable research direction but requires a unified evaluation framework with controlled baselines. We identify this as the primary direction for future work.

B. Effectiveness of Defense Mechanisms

Based on our findings and the broader adversarial robustness literature, we discuss four candidate defense strategies. We emphasize that these are **proposed directions**, not experimentally validated results on PLMs.

- **Adversarial training** [17] is the most established defense in machine learning, shown to substantially improve robustness at the cost of increased training time and compute. Adapting adversarial training to discrete protein sequences requires careful design of the perturbation space.
- **Bayesian uncertainty estimation** [21] offers a probabilistic approach to detecting adversarial inputs by flagging high-uncertainty predictions. This is particularly relevant for protein models where overconfident mispredictions could have downstream consequences in drug discovery pipelines.

- **Ensemble models** [22] improve robustness by aggregating predictions across multiple models, reducing the impact of any single model's vulnerability. The primary drawback is substantially increased inference cost.
- **Knowledge distillation** transfers robustness from a larger, adversarially trained teacher model to a smaller student model. This approach is appealing for deployment scenarios where computational efficiency is a constraint. Quantitative evaluation of this trade-off in the context of PLMs is left for future work.

To provide an empirical reference for the computational cost of these defenses, we measured inference time on a single 76 AA sequence (Ubiquitin) on an NVIDIA T4 GPU, averaged over 10 trials:

These measurements show that Bayesian uncertainty estimation imposes the highest overhead (9.5× baseline) due to repeated forward passes with dropout enabled. Ensemble inference with 3 models is more practically deployable at 2.8× overhead. Knowledge distillation and adversarial training costs are not measured here as they require retraining, which is identified as future work. Raw timing data is available in `esm2_runtime_results.csv` in the public repository.

Hybrid strategies that combine knowledge distillation with adversarial training represent a promising direction for achieving practical robustness without prohibitive computational cost.

C. Biosecurity Implications

Our results carry direct biosecurity relevance. We demonstrate that substitution attacks requiring no model access can increase ESM-2's sequence uncertainty by up to 66.6% for individual sequences. An adversary with basic knowledge of mutagenesis could generate sequences that deviate significantly from natural proteins while potentially evading screening tools calibrated on the natural sequence distribution.

This threat is not adequately addressed by current governance frameworks:

- The **Biological Weapons Convention (BWC)** prohibits biological weapons development but contains no provisions for computationally designed or adversarially perturbed sequences.
- The **NIH P3CO framework** governs dual-use research of concern but was designed before AI-assisted protein design became practical.
- **NSABB guidelines** on dual-use research require updating to reflect the distinct threat profile of AI-generated sequences.
- Toxicity screening tools used by the **iGEM safety committee** represent a practical model but are not designed to detect adversarially perturbed sequences.

We recommend that biosecurity governance bodies prioritize: (1) updating DURC policies to explicitly cover AI-assisted protein design, (2) developing adversarially-aware toxicity screening pipelines, and (3) establishing access control frameworks for high-capability protein design models analogous to existing export controls on dual-use technologies.

D. Future Research Directions

1) Expanding to Structure-Aware Models: The most immediate extension is evaluating AlphaFold2 and RoseTTAFold under the same substitution attack framework, with properly normalized baselines.

2) Gradient-Based Attacks via Continuous Relaxation: Applying FGSM and PGD to discrete protein sequences requires continuous relaxation techniques such as Gumbel-softmax [7] or embedding-space perturbations. Implementing these properly would substantially strengthen the adversarial evaluation framework.

3) Longer Sequence Evaluation: Future experiments should use sequences of at least 100 AA to cleanly separate the effects of different mutation budgets.

4) Quantitative Defense Evaluation: Experimentally measuring the robustness-efficiency trade-off of adversarial training, Bayesian estimation, ensemble methods, and knowledge distillation applied directly to ESM-2 would convert the proposed directions in this paper into concrete findings.

5) Wet-Lab Validation: Computational adversarial robustness studies should ultimately be validated against experimentally observed protein behavior, requiring interdisciplinary collaboration with wet-lab groups.

6) Regulatory and Ethical Frameworks: Developing AI governance frameworks specific to protein design — in collaboration with policymakers, biosecurity experts, and bioethicists — is essential to ensuring that advances in PLM capability are matched by appropriate oversight mechanisms.

VI. Conclusion

This paper presents a reproducible experimental evaluation of adversarial robustness in ESM-2, a representative sequence-based protein language model. Using substitution-based black-box attacks at mutation rates of 1%, 5%, and 10%, we demonstrate that even minimal sequence perturbation measurably disrupts ESM-2's sequence modeling confidence. Mean perplexity increased by +9.9%, +11.9%, and +23.5% respectively across eight benchmark protein sequences, with individual sequences exhibiting shifts as high as +66.6%. These results confirm that sequence-based PLMs are sensitive to substitution-based adversarial attacks and that this sensitivity scales with mutation rate.

Our per-sequence analysis reveals that robustness is modulated by sequence length and evolutionary conservation. Longer, well-characterized sequences such as Ubiquitin exhibit a clear dose-response relationship to mutation rate, while shorter sequences saturate at low mutation budgets due to the discrete nature of the perturbation space.

Beyond model vulnerabilities, we analyze the biosecurity risks associated with AI-generated proteins. Adversarial perturbations requiring no model access can substantially disrupt PLM predictions, representing a threat vector not addressed by existing governance frameworks including the Biological Weapons Convention, the NIH P3CO framework, or NSABB guidelines. We propose mitigation strategies including adversarially-aware toxicity screening pipelines, access control policies for high-risk protein design models, and updated regulatory frameworks.

We identify several important directions for future work: extending this evaluation to structure-aware models such as AlphaFold2 and RoseTTAFold with properly normalized baselines; implementing gradient-based attacks via continuous relaxation techniques; experimentally evaluating adversarial defense mechanisms including adversarial training, Bayesian uncertainty estimation, and knowledge distillation in the PLM context; and validating computational findings through wet-lab characterization of adversarially perturbed sequences.

The central contribution of this work is not a definitive characterization of PLM robustness, but a reproducible foundation on which future work can build. By grounding our findings in real experiments with public code and data, we aim to establish a reliable baseline for the adversarial robustness evaluation of protein language models. All experimental code, data, and results are publicly available at https://github.com/Mdniloykhan/esm2-adversarial-robustness.

ACKNOWLEDGMENT

The author thanks the anonymous reviewers of the AIUB Journal of Science and Engineering for their detailed and constructive feedback, which substantially improved the rigor, scope, and reproducibility of this manuscript. Their comments on the experimental methodology, threat model specification, biosecurity literature coverage, and runtime measurement were instrumental in shaping the final v ersion o f t his w ork. The experiments reported in this paper were conducted using Google Colaboratory with a freely available NVIDIA T4 GPU. The ESM-2 model weights were obtained from the publicly available `fair-esm` library developed by Meta AI Research.

REFERENCES

[1] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, 2021.

[2] M. Baek et al., "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, pp. 871–876, 2021.

[3] Z. Lin et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1131, 2023.

[4] A. Elnaggar et al., "ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7112–7127, 2022.

[5] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Banff, Canada, 2014.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, 2015.

[7] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-softmax," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Toulon, France, 2017.

[8] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "HotFlip: White-box adversarial examples for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Melbourne, Australia, pp. 31–36, 2018.

[9] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment," in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, vol. 34, pp. 8018–8025, 2020.

[10] F. J. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins, "Dual use of artificial intelligence-powered drug discovery," *Nature Mach. Intell.*, vol. 4, pp. 189–191, 2022.

[11] United Nations, "Convention on the prohibition of the development, production and stockpiling of bacteriological (biological) and toxin weapons and on their destruction (BWC)," 1972. [Online]. Available: https://legal.un.org/avl/ha/cpdpsbbtwd/cpdpsbbtwd.html

[12] National Institutes of Health, "P3CO framework: Recommended policy guidance for potential pandemic pathogen care and oversight," U.S. Dept. Health and Human Services, Washington, DC, USA, 2017.

[13] National Science Advisory Board for Biosecurity (NSABB), "Strategies to educate amateur biologists and scientists in non-life science disciplines about dual use research," U.S. Dept. Health and Human Services, Washington, DC, USA, 2013.

[14] R. J. Andrews et al., "Governing the risks of AI in protein design," *Science*, vol. 381, pp. 1234–1241, 2023.

[15] K. Tunyasuvunakool et al., "Highly accurate protein structure prediction for the human proteome," *Nature*, vol. 596, pp. 590–596, 2021.

[16] R. Rao et al., "Transformer protein language models are unsupervised structure learners," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Virtual, 2021.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vancouver, Canada, 2018.

[18] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, pp. 2137–2146, 2018.

[19] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Security (AISec)*, Dallas, TX, USA, pp. 15–26, 2017.

[20] H. Chen et al., "Robustness of molecular classification via adversarial training," *Nature Commun.*, vol. 11, p. 3399, 2020.

[21] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, pp. 1050–1059, 2016.

[22] Y. Liu, X. Chen, C. Liu, and D. Song, "Towards robust neural networks via random self-ensemble," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, pp. 369–385, 2018.

[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[24] M. Brundage et al., "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," *arXiv preprint arXiv:2004.07213*, 2020.

[25] iGEM Foundation, "iGEM safety and security committee," 2023. [Online]. Available: https://responsibility.igem.org/safety-policies

[26] A. Madani et al., "Large language models generate functional protein sequences across diverse families," *Nature Biotechnol.*, vol. 41, pp. 1099–1106, 2023.

[27] J. Ingraham et al., "Illuminating protein space with a programmable generative model," *Nature*, vol. 623, pp. 1070–1078, 2023.

[28] J. Meier et al., "Language models enable zero-shot prediction of the effects of mutations on protein function," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Virtual, vol. 34, pp. 29287–29303, 2021.

Md. Robiul Islam Niloy received the B.Sc. degree in Computer Science and Engineering from Manarat International University, Dhaka, in 2021 and the M.Sc. degree in Computer Science from Jahangirnagar University, Dhaka, in 2023. He is currently serving in the Department of Computer Science and Engineering at BRAC University, Dhaka. Prior to this appointment, he served as a Lecturer in the Department of Computer Science and Engineering at United College of Aviation Science Management, Dhaka (2023–2025), where he delivered undergraduate lectures and supervised student research projects. He also served as an Instructor at Greenland Polytechnic Institute, Dhaka (2021–2023). His primary research focus lies at the intersection of machine learning security and responsible AI governance. His research interests encompass large language model (LLM) security, adversarial robustness of foundation models, autonomous threat detection, responsible AI governance, and trustworthy AI deployment. He has authored and co-authored several peer-reviewed publications indexed in IEEE Xplore, Springer Lecture Notes in Computer Science, ACM Digital Library, and Scopus. He has also published several machine learning security datasets on GitHub for use by the research community.

Zafer Aydin is an Associate Professor in the Department of Computer Engineering at Abdullah Gul University. Born in Canakkale in 1977, he completed his early education there before graduating from Istanbul Ataturk High School of Science in 1995. He earned his B.Sc. and M.Sc. degrees with high honors from the Department of Electrical and Electronics Engineering at Bilkent University in 1999 and 2001, respectively. He later pursued his Ph.D. at Georgia Institute of Technology, where he worked as a Graduate Research Assistant and received his doctorate in 2008. From 2008 to 2011, he conducted postdoctoral research at the Noble Research Laboratory in the Department of Genome Sciences at University of Washington. After returning to Turkey, he served as an Assistant Professor at Bahcesehir University before joining Abdullah Gul University, where he received tenure in 2021. His research interests include machine learning, deep learning, bioinformatics, computational biology, medical image analysis, and cybersecurity, and he has published extensively in these areas with significant international research impact.

Md. Golam Moazzam received his B.Sc. (Hons.) in Electronics and Computer Science in 1997 and his M.S. in Computer Science and Engineering in 2001 from Jahangirnagar University, Dhaka. He joined the Department of Computer Science and Engineering at the same university as a Lecturer in 2001 and is currently serving as a Professor. His research interests include Digital Image Processing, Artificial Intelligence, Computer Vision, Machine Learning, and Pattern Recognition. He has published numerous research articles in reputed journals and conferences, with research contributions spanning rice disease detection, rice variety classification, vehicle speed estimation, medical image segmentation, natural language processing, and deep learning applications.