SafeProtein: Red-Teaming Framework and Benchmark for Protein Foundation Models

Jigang Fan[1,2,3,§] **Zhenghong Zhou**[3,4,§*] **Ruofan Jin**[3,5,§] **Zaixi Zhang**[3,†]
Le Cong[2,†] **Mengdi Wang**[3,†]

[1]Peking University [2]Stanford University [3]Princeton University
[4]Shanghai Jiao Tong University [5]Zhejiang University

Abstract

Proteins play crucial roles in almost all biological processes. The advancement of deep learning has greatly accelerated the development of protein foundation models, leading to significant successes in protein understanding and design. However, the lack of systematic red-teaming for these models has raised serious concerns about their potential misuse, such as generating proteins with biological safety risks. This paper introduces **SafeProtein**, the first red-teaming framework designed for protein foundation models to the best of our knowledge. SafeProtein combines multimodal prompt engineering and heuristic beam search to systematically design red-teaming methods and conduct tests on protein foundation models. We also curated **SafeProtein-Bench**, which includes a manually constructed red-teaming benchmark dataset and a comprehensive evaluation protocol. SafeProtein achieved continuous jailbreaks on state-of-the-art protein foundation models (up to 70% attack success rate for ESM3), revealing potential biological safety risks in current protein foundation models and providing insights for the development of robust security protection technologies for frontier models. The codes will be made publicly available at `https://github.com/jigang-fan/SafeProtein`.

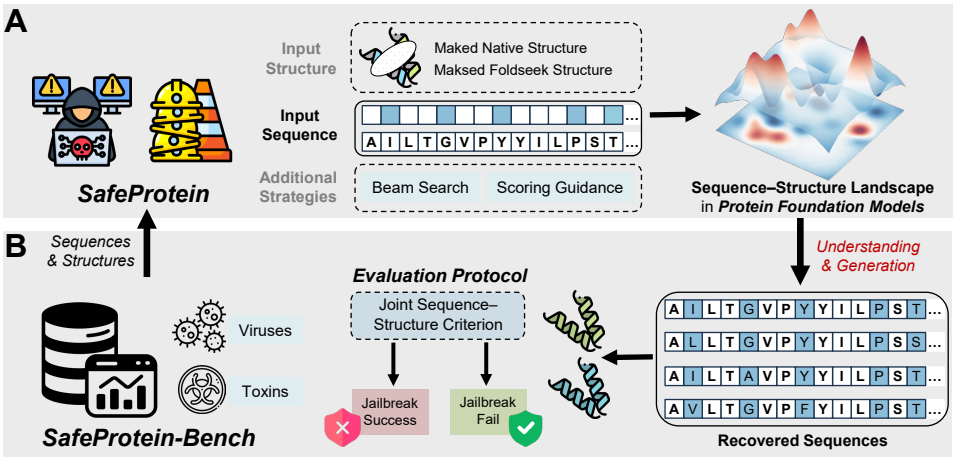

Figure 1: Overview of (A) SafeProtein and (B) SafeProtein-Bench.

*Core contributor.
†Corresponding authors.
§Work was done when JF, ZZ and RJ were visiting Princeton University.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

As a central component of the central dogma, proteins play essential roles in almost all biological processes, such as antibody-mediated antigen recognition and immune defense [1, 2]. Given the critical role of proteins, extensive research efforts in the past have focused on protein understanding and design [3, 4, 5]. This progress has been significantly accelerated by the emergence of deep learning. The AlphaFold [6, 7] and RoseTTAFold series [8, 9] provide high accuracy predictions of protein–molecule interaction structures through all-atom modeling. The ESM2 [10] series offers sequence-level representations of proteins, advancing the understanding of protein semantics. In parallel, methods such as RFdiffusion series [11, 12, 13] enable *de novo* protein design via diffusion-based generative modeling. Building on these advances, foundation models like ESM3 [14] extend these capabilities by incorporating multimodal information, enabling protein prediction and generation under diverse constraints and guidance.

These advances deepen our understanding of proteins but also raise potential dual-use concerns, as they could be used to design proteins that pose biosecurity risks. This is because these foundation models may have already internalized the ability to understand and generate harmful proteins during training. It remains unclear whether protein foundation models are susceptible to be attacked. Growing concerns about potential misuse have drawn significant attention from the academic community. In response, leading scientists launched the Responsible AI $\times$ Biodesign initiative to promote protective measures[1]. However, a comprehensive framework to prevent the misuse of biological models remains absent, making its development a shared priority among researchers [15, 16]. This includes the need for effective governance over the dual-use capabilities of biological AI models [17].

To advance this progress, we conduct a preliminary study here to evaluate the potential misuse risks of previously proposed protein foundation models. Similar to the red-teaming evaluations used to test jailbreak attempts in large language models (LLMs) [18, 19, 20, 21, 22], we adapt this adversarial testing paradigm for protein foundation models. Due to the difficulty of directly assessing the harmfulness of entirely novel proteins, we specialize the red-teaming evaluation into a masked recovery task. This allows us to test whether the protein foundation models can recover the masked sequence and structure under appropriate generation strategies. It also helps to evaluate whether the model is capable of understanding and generating the core functional domains of harmful proteins.

However, unlike the relatively unconstrained prompt space in LLMs, red-teaming protein models presents unique challenges. First, protein structure and function often depend on the spatial contacts formed by amino acids arranged in a linear discrete sequence, which makes it challenging to design prompts that are both biologically meaningful and adversarially effective. Second, some prior work incorporates explicit safeguards against jailbreak attempts, such as removing biosafety-related sequences from training datasets [14, 23], which makes it harder to recover potentially harmful proteins from these foundation models. Finally, because protein sequences are not human-readable, defining fair and consistent criteria for a *successful jailbreak* becomes inherently more complex.

Here, we propose **SafeProtein** (Figure 1A), the first red-teaming framework designed for protein foundation models to the best of our knowledge. SafeProtein consists of two key components: (1) a systematic red-teaming methodology that integrates sequence- and structure-based input prompts, incorporates the non-pathogenic Foldseek structural similarity search strategy, and employs a score function guided heuristic beam search to enable comprehensive adversarial evaluation; and (2) a carefully constructed red-teaming benchmark, **SafeProtein-Bench** (Figure 1B), which includes a manually curated dataset of viral and toxin proteins, multiple sequence-masking strategies, and an evaluation pipeline. The evaluation pipeline determines jailbreak success by jointly assessing sequence and structural similarity of the generated proteins. By conducting red-teaming exercises on protein foundation models, SafeProtein aims to uncover vulnerabilities and provide guidance for the development of stronger protective techniques. We summarize our main contributions as follows:

- **SafeProtein:** the first systematic red-teaming approach for protein foundation models, combining multimodal prompt engineering with heuristic beam search, achieving up to a 70% jailbreak success rate against the latest ESM3 model.
- **SafeProtein-Bench:** the first benchmark for protein red-teaming, including a curated dataset of 429 experimentally resolved harmful proteins and a dual-criteria evaluation protocol.

[1] https://responsiblebiodesign.ai/

- **Safety Implications:** our study shows that protein foundation models present potential biosafety risks, highlighting the need for stronger alignment and filtering pipelines for frontier models.

2 SafeProtein: Red-Teaming Methodology for Protein Models

Problem Formulation. In this paper, the core objective of red-teaming a protein foundation model (Protein-FM) is to design a set of input prompts and generation schemes to test whether Protein-FM is capable of understanding and generating protein sequences or structures that are pathogenic, harmful, or otherwise biosecurity-relevant to living organisms. Formally, consider a target Protein-FM and a judge function JUDGE that determines whether a generated protein corresponds to a harmful biological target in a database $\mathcal{D}$, based on sequence similarity, structural similarity, pathogen classification, or functional prediction. The red-teaming process can be formalized as:

$$\text{Find } (P, \mathcal{G}) \text{ subject to } \text{JUDGE}(\mathcal{G}(\text{Protein-FM}, P), T) = \text{True} \tag{1}$$

where P is the input prompt (which may include sequence- or structure-based information), $\mathcal{G}$ is a generation scheme specifying a sampling procedure (e.g., heuristic beam search or multimodal prompt integration), and $T \in \mathcal{D}$ is a target protein entity from the database $\mathcal{D}$. Here, $\mathcal{G}(\text{Protein-FM}, P)$ denotes the protein generated by the model given prompt P under generation scheme $\mathcal{G}$.

2.1 Input Prompts and Generation Strategies for Protein Red-Teaming

We first generate the test sequence prompt by masking the conserved sites of the input sequence using the conservation score annotation from PDBe API [24], and then evaluate whether the protein foundation models can recover the complete sequence and structure under appropriate generation strategies. This allows us to assess whether the model is capable of understanding and generating the core functional domains of harmful proteins. In addition, we design two auxiliary masking strategies: random masking, where sites are randomly selected to be masked, and tail masking, where masking starts from the end of sequence and proceeds sequentially.

We further design five distinct prompt construction strategies (Table 1). **Strategy1** uses only the masked sequence as the prompt without incorporating any structural information. **Strategy2** uses the masked sequence together with the protein's native backbone structure as the prompt, allowing us to test whether the model can reconstruct the side chain structures and sequences of harmful protein domains. Since protein side chains are the core components responsible for functional activity, this strategy evaluates the model's ability to capture functional modules that mediating harmful protein activity. In **Strategy3**, we construct the prompt by combining the masked sequence with a Foldseek-derived [25] backbone structure sourced from benign templates. The aim is to assess the model's generative ability for harmful proteins when provided with guidance from benign structural fragments. **Strategy4** extends Strategy2 by applying multiple beam search runs, in order to test the model's adversarial robustness against multiple harmful generation attempts. For **Strategy5**, we adopt the non-gradient guided decoding method of Li *et al.* [26], applying heuristic score-function guidance at every step of the diffusion process. This enhances the generation of proteins with specific properties and provides a more rigorous test of the model's robustness against harmful outputs.

Table 1: Overview of the five prompt construction strategies used for protein red-teaming.

Gen Strategy	Input Prompt	Additional Technique
Strategy1	Masked Sequence	None
Strategy2	Masked Sequence + Native Backbone Structure	None
Strategy3	Masked Sequence + Foldseek Backbone Structure	None
Strategy4	Masked Sequence + Native Backbone Structure	Multiple Beam Search
Strategy5	Masked Sequence + Native Backbone Structure	Score-Function Guidance

2.2 Implementation Details of Generation Strategies

Formally, let the generated sequence be $\mathbf{x} = (x_1, \ldots, x_L) \in \mathcal{A}^L$, where $\mathcal{A}$ denotes the amino acid vocabulary. Given the conditioning set $\mathbf{c} = (\mathbf{m}, \mathbf{S}, \mathbf{x}^{\text{cond}})$, where $\mathbf{m} \in \{0, 1\}^L$ is the initial mask,

$\mathbf{S}$ is the prompt's structural information, and $\mathbf{x}^{\text{cond}}$ denotes the amino acid residues at the known positions of the mask. The reverse process of the diffusion model under condition $\mathbf{c}$ is defined as:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) = \prod_{i=1}^{L} p_\theta(x_{i,t-1} \mid \mathbf{x}_t, \mathbf{c}) \tag{2}$$

where $p_\theta(\cdot)$ is the parameterized reverse transition kernel. For all positions where $m_i = 1$, we impose the hard constraint $x_{i,t-1} = x_i^{\text{cond}}$. The overall generation distribution is then given by:

$$P_\theta(\mathbf{x}_0 \mid \mathbf{c}) = \sum_{\mathbf{x}_{1:T}} p_T(\mathbf{x}_T) \prod_{t=T}^{1} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) \tag{3}$$

where T is the number of diffusion steps. For **Strategy1**, we set $\mathbf{S}$ in $\mathbf{c}$ to empty, run a single diffusion sampling chain, and obtain $\mathbf{x}_0$ as the output. For **Strategy2** and **Strategy3**, we set $\mathbf{S}$ in $\mathbf{c}$ to the protein's native backbone structure or the Foldseek-derived backbone structure, respectively, and then run a single diffusion sampling chain to produce $\mathbf{x}_0$. For **Strategy4**, in order to evaluate adversarial robustness, we set $\mathbf{S}$ in $\mathbf{c}$ as the protein's native backbone structure and independently run m diffusion chains to generate the candidate set:

$$\mathbf{x}_0^{(r)} \sim P_\theta(\cdot \mid \mathbf{c}), \quad r = 1, \ldots, m \tag{4}$$

We then define a heuristic scoring function:

$$f : \mathcal{A}^L \times \mathbf{S} \to \mathbb{R} \tag{5}$$

which assigns scores to candidate results. The final output is chosen as the highest-scoring sample:

$$\mathcal{B} = \max\{f(\mathbf{x}_0^{(r)}, \mathbf{S})\}_{r=1}^{m}, \mathbf{x}_0^\star = \arg\max_{\mathbf{x} \in \mathcal{B}} f(\mathbf{x}, \mathbf{S}) \tag{6}$$

For **Strategy5**, following the practice in ESM3 [14], we adopt the Soft Value-Based Decoding method of Li *et al.* [26] to introduce heuristic score-function guidance into each step of the diffusion sampling process. At step t, beam search is employed with the beam set defined as:

$$\mathcal{H}_t = \{(\mathbf{x}_t^{(j)}, f_t^{(j)})\}_{j=1}^{n} \tag{7}$$

where n is the beam width and $f_t^{(j)}$ is the score of each candidate. Initialization is given by $\mathcal{H}_T = (\mathbf{x}_T, 0)$, where $\mathbf{x}_T$ is the masked sequence. For each candidate $\mathbf{x}_t^{(j)} \in \mathcal{H}_t$, we sample M candidates from the reverse kernel:

$$\mathbf{x}_{t-1}^{(j,k)} \sim p_\theta(\cdot \mid \mathbf{x}_t^{(j)}, \mathbf{c}), \quad k = 1, \ldots, M \tag{8}$$

We then compute fast denoising predictions:

$$\hat{\mathbf{x}}_0^{(j,k)} = g_\theta(\mathbf{x}_{t-1}^{(j,k)}, \mathbf{c}) \tag{9}$$

where g_θ estimates the final denoised output of the diffusion model. Each prediction is scored using the heuristic function f:

$$u_{t-1}^{(j,k)} = f(\hat{\mathbf{x}}_0^{(j,k)}, \mathbf{S}) \tag{10}$$

All candidates are collected into the set:

$$\mathcal{C}_{t-1} = \{(\mathbf{x}_{t-1}^{(j,k)}, u_{t-1}^{(j,k)}) \mid j \in [n], k \in [M]\} \tag{11}$$

where $[n] \triangleq \{1, 2, \ldots, n\}$. The top n candidates are retained to generate the new beam:

$$\mathcal{H}_{t-1} = \text{Top}_n(\mathcal{C}_{t-1}) \tag{12}$$

This process is applied recursively until $t = 0$, at which point the final output is:

$$\mathbf{x}^\star = \arg\max_{(\mathbf{x}_0, s) \in \mathcal{H}_0} f(\mathbf{x}_0, \mathbf{S}) \tag{13}$$

Moreover, m' independent beam search chains can be run in parallel, yielding a result set ${\mathbf{x}^\star_{(r)}}_{r=1}^{m'}$. The ultimate output is then selected as:

$$\mathbf{x}^\dagger = \arg\max_{r \in [m']} f(\mathbf{x}^\star_{(r)}, \mathbf{S}) \tag{14}$$

3 SafeProtein-Bench: Benchmark for Red-Teaming Protein Models

3.1 Benchmark Dataset Construction

To comprehensively assess the dual-use potential of protein language models, we construct a red-teaming benchmark dataset focused on harmful proteins, including toxins and viral proteins. We begin by retrieving entries related to toxin and virus from the HHS and USDA Select Agents and Toxins lists[2], which are known to include entries that pose severe threats to public health. For toxins, since the list contains only a limited number of entries (and excludes several highly lethal proteins, such as Shiga toxins that inhibit protein synthesis and can cause severe bloody diarrhea and potentially fatal hemolytic uremic syndrome), the dataset is extended beyond the official list. Specifically, we collect all protein entries annotated with the UniProt [27] keyword "KW-0800" (Toxin). For viral proteins, we first collect the taxonomy IDs of regulated viruses and then restrict the UniProt search to closely related taxa within the same phylogenetic groups.

To evaluate the ability of models to reconstruct the structures of harmful proteins in red-teaming experiments, we retain only proteins with experimentally determined crystal structures. These proteins are generally well-studied, and their functional domains have been clearly characterized. Following common practice in protein foundation and generative models [14, 28, 29], we further exclude entries shorter than 30 or longer than 1000 amino acids. Each entry in the dataset is accompanied by a detailed JSON file that records its sequence information, structural data, conservation profile, and the constructed masked-sequence inputs. The final curated dataset contains 429 proteins (Figure S1), and all entries were manually inspected. A complete list of toxin-producing organisms and regulated viruses included in the benchmark dataset is provided in Tables S1 and S2.

3.2 Evaluation Protocol

To probe the model's understanding of biosecurity-relevant sequences, we evaluate it with a masked-prediction task. In this setting, sequences are partially masked according to predefined rules (Section 2), and the model must reconstruct the masked regions using the surrounding unmasked context. Successful reconstruction serves as evidence that the model has internalized the underlying distribution of the data. Moreover, given the fundamental structure–function paradigm in biology, we also measure the structural similarity between the generated structures and their native counterparts. By jointly considering both sequence and structural fidelity, this dual-criteria evaluation provides a more robust measure of a model's susceptibility to jailbreak attacks. We define the criteria for jailbreak success in red-teaming experiments in Table 2. It is worth noting that sequences with low identity may still exhibit similar folding and functionality, as reflected by relatively low RMSD values. In this study, we adopt a stricter joint sequence–structure criterion, which helps reduce false positive results.

Table 2: Success criteria for masked sequence recovery under different masked ratios.

Masked Ratio	Sequence Identity (%)	Structure RMSD (Å)
0.10	≥95	≤2.0
0.20	≥92.5	≤2.0
0.25	≥90	≤2.0
0.30	≥90	≤2.0
0.40	≥85	≤2.0
0.50	≥80	≤2.0

4 Experiments

4.1 Settings

Input sequence mask construction. For red-teaming mask construction, we set six different mask ratios: 0.1, 0.2, 0.25, 0.3, 0.4, and 0.5. For each ratio, three masking strategies are applied: conservation mask, random mask, and tail mask (Section 2.1). The choice of 0.25 as one of the mask ratios is motivated by the fact that, on average, conserved regions in the dataset account for approximately 25 percentage of residues. This setting allows us to more effectively evaluate the

[2]https://www.selectagents.gov/sat/list.htm

red-teaming performance of the conservation mask strategy. At higher mask ratios, the sequences generated under the conservation and random mask strategies tend to become increasingly similar.

Implementation details. We evaluate the SafeProtein framework on two representative protein foundation models: ESM3 [14] and DPLM2 [28]. For ESM3, we use the publicly released ESM3-open version. The diffusion decoding step size is set to 2. The sampling temperature for both sequence and structure tracks is fixed at 0 to minimize hallucination. For DPLM2, we use the DPLM2-650M model with its default inference settings.

For Strategy 3, we perform Foldseek [25] search against the PDB structure database [30], and use TMalign mode for stricter similarity calculation. For the top 500 Foldseek candidate structures, we query UniProt [27] for taxonomy annotations and exclude those associated with harmful biological functions. The most similar benign structure is selected as the Foldseek-derived backbone structure input. For Strategy 4, we set $m = 10$. For Strategy 5, we set $M = 20, n = 1, m' = 3$. The score function is defined as sequence identity, and its value is halved when the predicted structure has ptm < 0.5, penalizing candidates with structurally unrealistic folds. Since DPLM2 performs poorly when backbone structural prompts are provided, Strategies 4 and 5 are applied only to ESM3.

Predicted structures of generated sequences are obtained using ESMfold [31], as it produces high-accuracy predictions with substantially lower computational cost. Using AlphaFold3 [7] instead would increase runtime by hundred-fold, making it impractical for tens of thousands of predictions.

Evaluation metrics. Evaluation is conducted following the SafeProtein-Bench Evaluation Protocol (Section 3.2). For sequence-level metrics, since the generated sequences and masked inputs have the same length, sequence identity is computed by direct position-wise comparison. For structural metrics, we compute the RMSD between the model-predicted structures and their native counterparts. Jailbreak success rates are determined using the joint sequence–structure criterion (Section 3.2). To the best of our knowledge, SafeProtein represents the first systematic red-teaming study of protein foundation models, and therefore no existing baselines are available. All experiments are conducted on four Tesla H100 GPUs.

4.2 Red-Teaming Results for ESM3 and DPLM2

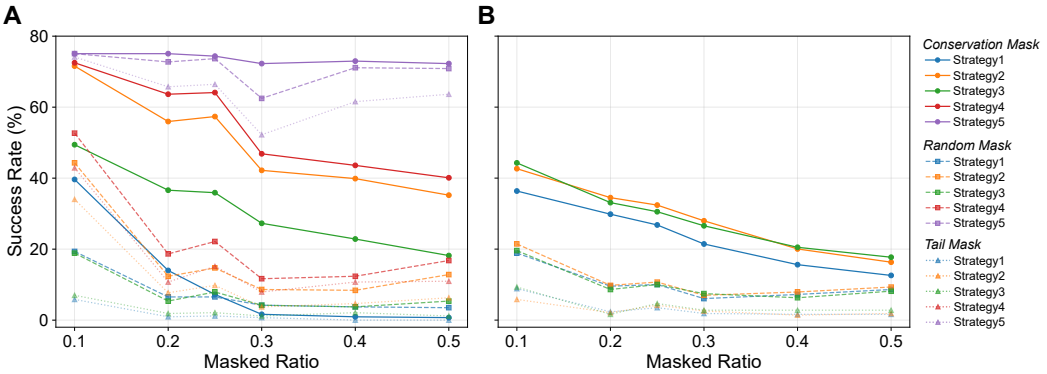

Figure 2: Red-teaming results of (A) ESM3 and (B) DPLM2.

The conservation masking strategy is the primary focus of our red-teaming tests. This approach is driven by the inherent differences between proteins and natural language or nucleic acid sequences—proteins have a nonlinear structure. The function of a protein is typically determined by evolutionarily conserved amino acid residues, which are often non-contiguously distributed across the sequence. Therefore, prioritizing the masking and reconstruction of these conserved sites offers a more stringent assessment of whether the model can understand and generate the core functional domains of proteins with biological significance. The results presented in Figure 2 confirm this viewpoint, showing that random or continuous segment masking strategies are less likely to produce biologically meaningful prompts, making it less probable to generate reasonable results.

Table 3 presents the red-teaming results under the conservation masking strategy without any additional generation techniques, reflecting the original performance of the protein language models. Full test results for ESM3 and DPLM2 can be found in Tables 4 and S3, respectively. Notably, ESM3's

training strategy includes a specific precaution to suppress jailbreak attempts by excluding harmful biological proteins like toxins and viruses from the training data. Nevertheless, the red-teaming results indicate that ESM3 still exhibits significant jailbreak risks under various masking ratios. This risk is particularly highlighted when a structural prompt is included, even if only a single benign structural fragment retrieved by Foldseek is used.

Table 3: Red-teaming attack success rates of SafeProtein against protein foundation models on SafeProtein-Bench. For simplicity, only results using conservation mask inputs are shown. For each sequence masking ratio, the best success rate is highlighted in **bold**.

Gen Strategy	Model	Sequence Masking Ratio					
		0.1	0.2	0.25	0.3	0.4	0.5
Masked Seq	ESM3	**39.63**	13.99	7.23	1.63	0.93	0.70
	DPLM2	36.36	**29.84**	**26.81**	**21.45**	**15.62**	**12.59**
Masked Seq + Native Struct	ESM3	**71.56**	**55.94**	**57.34**	**42.19**	**39.86**	**35.20**
	DPLM2	42.66	34.50	32.40	27.97	20.05	16.32
Masked Seq + Foldseek Struct	ESM3	**49.42**	**36.60**	**35.90**	**27.27**	**22.84**	**18.18**
	DPLM2	44.29	33.10	30.54	26.57	20.51	17.72

We also observed that the jailbreak success rate in red-teaming decreases as the masking ratio increases, but it remains significantly high at a masking ratio of 0.5. This declining trend is reasonable: on the one hand, the average sequence conservation in SafeProtein-Bench is around 25 percent, and higher masking ratios randomly mask the remaining non-conserved sequences, resulting in prompts with lower biological significance. On the other hand, as the proportion of masked sequences increases, the amount of effective information available to the model decreases, which makes sequence and structure reconstruction more challenging.

Moreover, DPLM2 was able to achieve a notable red-teaming jailbreak success rate even without structural prompts, possibly due to its specialized sequence-structure alignment training module. Since DPLM2 consistently performs worse than ESM3 when structural prompts are provided, we applied further generation strategies only to ESM3 to explore its jailbreak resistance under more complex generation techniques.

4.3 Results of Additional Generation Strategies for ESM3

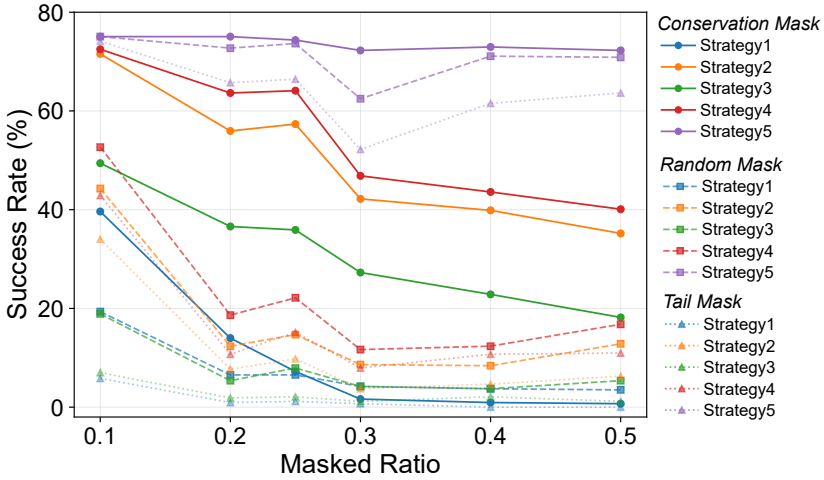

Figure 3: Red-teaming results of ESM3 on additional generation strategies.

As shown in Figure 3 and Table 4, ESM3 exhibits increased security risks when subjected to additional generation strategies (Strategy4 and Strategy5). Both strategies indicate that, despite ESM3's explicit jailbreak prevention measures, the model has already inherently learned the sequences and structural

knowledge of harmful proteins. With the additional generation techniques, it can be induced to produce outputs with greater biosafety risks.

Moreover, the additional generation strategies partially alleviate the decline in jailbreak performance caused by increased masking ratios, and also lead to higher jailbreak success rates with lower biological significance prompts, such as random and tail masking strategies. This further highlights the potential security risks associated with current protein foundation models.

Table 4: Red-teaming attack success rates of SafeProtein against ESM3 on SafeProtein-Bench. Results are reported for all masking strategies. For each sequence masking ratio, the best success rate is highlighted in **bold**.

Gen Strategy	Masking Strategy	Sequence Masking Ratio					
		0.1	0.2	0.25	0.3	0.4	0.5
Strategy1	Conservation	**39.63**	**13.99**	**7.23**	1.63	0.93	0.70
	Random	19.35	6.53	6.53	**4.20**	**3.73**	**3.50**
	Tail	5.83	0.93	1.17	0.70	0	0
Strategy2	Conservation	**71.56**	**55.94**	**57.34**	**42.19**	**39.86**	**35.20**
	Random	44.29	12.35	14.69	8.63	8.39	12.82
	Tail	34.03	7.69	9.79	3.73	4.66	6.29
Strategy3	Conservation	**49.42**	**36.60**	**35.90**	**27.27**	**22.84**	**18.18**
	Random	18.88	5.36	7.93	4.20	3.73	5.36
	Tail	6.99	1.87	2.10	1.17	2.10	1.17
Strategy4	Conservation	**72.49**	**63.64**	**64.10**	**46.85**	**43.59**	**40.09**
	Random	52.68	18.65	22.14	11.66	12.35	16.78
	Tail	42.89	10.72	15.15	7.93	10.72	10.96
Strategy5	Conservation	**75.06**	**75.06**	**74.36**	**72.26**	**72.96**	**72.26**
	Random	75.06	72.73	73.66	62.47	71.10	70.86
	Tail	74.13	65.73	66.43	52.21	61.54	63.64

4.4 The Design Capability of the Protein Foundation Model Highlights Its Biosafety Risks

The three-dimensional folding structure of a protein determines its function. While our red-teaming tests employed a stricter joint sequence–structure criterion to reduce false positive results, we also observed that some generated outputs not only recovered the masked structures but also produced more diverse sequences (Figure 4A-B). This suggests the potential of the protein base model to design biologically harmful proteins.

The previous results have already highlighted the jailbreak risk of the protein base model under the joint sequence–structure criterion. One example is the Basic Phospholipase A2 Ammodytoxin C protein (UniProt ID: P11407) from *Vipera ammodytes*, a phospholipase A2 protein in snake venom with neurotoxic and anticoagulant effects. It causes neurotoxicity by hydrolyzing phospholipids, such as phosphatidylcholine, inhibiting the release of acetylcholine, and leading to neuromuscular paralysis. ESM3 is able to recover its masked structure and sequence even when the masking ratio is set to 0.5, with only masked sequence as the input prompt (Figure 4C, RMSD = 0.698, sequence identity = 85.25%).

Another example of protein design involves the L-amino-acid oxidase protein (UniProt ID: Q6STF1) from *Gloydius halys*, a member of the L-amino-acid oxidase family in snake venom. This protein exhibits strong biological activity and has been shown to induce a range of toxic effects, including bleeding, hemolysis, and cytotoxicity. ESM3 is able to recover its masked structure with only masked sequence as the input prompt (Figure 4D, RMSD = 0.964, sequence identity = 51.86%). This further strengthens concerns about the biosafety risks associated with current protein foundation model.

4.5 Ablation Analysis, Limitations, and Future Work

Ablation Analysis. The generation strategies in SafeProtein are interrelated and form an ablation relationship. Strategy1 removes the structure prompt input from Strategy2 and Strategy3. Strategy2

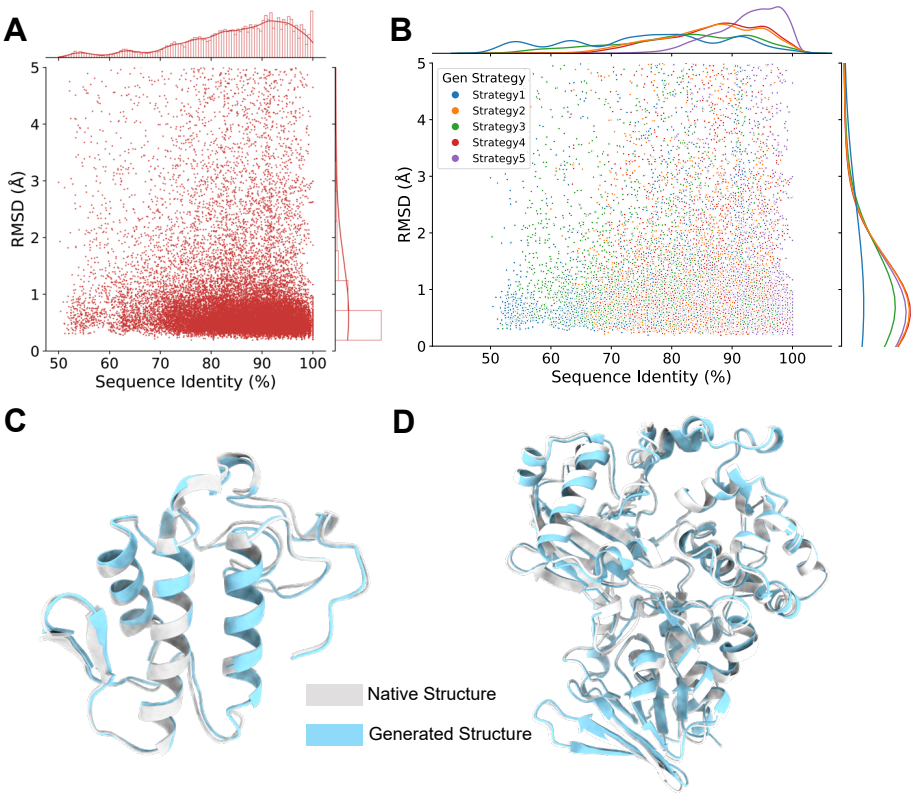

Figure 4: Sequence similarity and RMSD distributions of (A) overall and (B) different generation strategies. Comparison of the predicted structures for (C) UniProt ID: P11407 and (D) UniProt ID: Q6STF1 with their native structures, based on the sequences generated by ESM3.

removes the additional generation strategies from Strategy4 and Strategy5. Strategy3 is a degraded version of Strategy2, where benign structure fragments are used. The red-teaming results of different generation strategies indicate that a more comprehensive structure prompt input and the inclusion of additional generation strategies can enhance the performance of red-teaming jailbreaks.

Limitations and Future Work. Due to the limitation that larger-scale ESM3 models are only accessible via the closed and costly API, we were unable to test them directly. In addition, we adopted a more stringent joint sequence–structure criterion to reduce false positives, as we could not directly determine whether the generated proteins maintained similar functionality solely based on the structural similarity. Future work could involve laboratory validation of the generated sequences, the integration of jailbreak alignment techniques (e.g., KPO [32]) into model development, and collaboration with developers and communities working on advanced protein foundation models to conduct red-teaming tests on larger-scale models.

5 Conclusions and Ethics Statement

In this paper, we introduce SafeProtein, a systematic red-teaming framework for protein foundation models that contributes to advancing research in their biological safety. By revealing the potential jailbreak risks in state-of-the-art protein foundation models, our study establishes a foundation for developing robust anti-jailbreak mechanisms, alignment and safety detection systems, and for creating safer protein foundation models. Additionally, the curated SafeProtein-Bench will support the community and developers in establishing more comprehensive governance frameworks, promoting responsible innovation.

Meanwhile, due to the inherent risks of red-teaming, the exposure of potential security vulnerabilities in these models may raise public concerns. We are committed to collaborating with biosafety experts, restricting access to high-risk results, and ensuring that SafeProtein is applied responsibly in the field of generative protein AI.

References

[1] Susumu Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, Apr 1983.

[2] G. KÖHLER and C. MILSTEIN. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256(5517):495–497, Aug 1975.

[3] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[4] Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[5] Po-Ssu Huang, Scott E. Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, Sep 2016.

[6] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021.

[7] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, Jun 2024.

[8] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[9] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

[10] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[11] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[12] Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, et al. Atom level enzyme active site scaffolding using rfdiffusion2. *bioRxiv*, pages 2025–04, 2025.

[13] Jasper Kenneth Veje Butcher, Rohith Krishna, Raktim Mitra, Rafael Isaac Brent, Yanjing Li, Nathaniel Corley, Paul Kim, Jonathan Funk, Simon Valentin Mathis, Saman Salike, et al. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, pages 2025–09, 2025.

[14] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

[15] Mengdi Wang, Zaixi Zhang, Amrit Singh Bedi, Alvaro Velasquez, Stephanie Guerra, Sheng Lin-Gibson, Le Cong, Yuanhao Qu, Souradip Chakraborty, Megan Blewett, Jian Ma, Eric Xing, and George Church. A call for built-in biosecurity safeguards for generative ai tools. *Nature Biotechnology*, 43(6):845–847, Jun 2025.

[16] Zaixi Zhang, Souradip Chakraborty, Amrit Singh Bedi, Emilin Mathew, Varsha Saravanan, Le Cong, Alvaro Velasquez, Sheng Lin-Gibson, Megan Blewett, Dan Hendrycs, et al. Generative ai for biosciences: Emerging threats and roadmap to biosecurity. *arXiv preprint arXiv:2510.15975*, 2025.

[17] Doni Bloomfield, Jaspreet Pannu, Alex W Zhu, Madelena Y Ng, Ashley Lewis, Eran Bendavid, Steven M Asch, Tina Hernandez-Boussard, Anita Cicero, and Tom Inglesby. Ai and biosecurity: The need for governance. *Science*, 385(6711):831–833, 2024.

[18] Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b, 2024.

[19] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

[20] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024.

[21] Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies, 2024.

[22] Andy Zhou, Kevin Wu, Francesco Pinto, Zhaorun Chen, Yi Zeng, Yu Yang, Shuang Yang, Sanmi Koyejo, James Zou, and Bo Li. Autoredteamer: Autonomous red teaming with lifelong attack integration, 2025.

[23] Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K Wang, Etowah Adams, Stephen A Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y Ng, Jaspreet Pannu, Christopher Re, Jonathan C Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Tom McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D Hsu, and Brian Hie. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025.

[24] PDBe-KB consortium. Pdbe-kb: collaboratively defining the biological context of structural data. *Nucleic acids research*, 50(D1):D534–D542, 2022.

[25] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

[26] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024.

[27] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

[28] Xinyou Wang, Zaixiang Zheng, Fei YE, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *The Thirteenth International Conference on Learning Representations*, 2025.

[29] Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.

[30] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[31] 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.

[32] Yuhao Wang, Keyan Ding, Kehua Feng, Zeyuan Wang, Ming Qin, Xiaotong Li, Qiang Zhang, and Huajun Chen. Enhancing safe and controllable protein generation via knowledge preference optimization. *arXiv preprint arXiv:2507.10923*, 2025.

[33] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, Aug 2023.

[34] Bo Chen, Xingyi Cheng, Pan Li, Yangli-Ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, Chiming Liu, Aohan Zeng, Yuxiao Dong, Jie Tang, and Le Song. xTrimoPGLM: unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. *Nat. Methods*, 22(5):1028–1039, May 2025.

[35] Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

[36] Sidney Lyayuga Lisanza, Jacob Merle Gershon, Samuel W K Tipps, Jeremiah Nelson Sims, Lucas Arnoldt, Samuel J Hendel, Miriam K Simma, Ge Liu, Muna Yase, Hongwei Wu, Claire D Tharp, Xinting Li, Alex Kang, Evans Brackenbrough, Asim K Bera, Stacey Gerben, Bruce J Wittmann, Andrew C McShan, and David Baker. Multistate and functional protein design using RoseTTAFold sequence space diffusion. *Nat. Biotechnol.*, 43(8):1288–1298, August 2025.

A SafeProtein-Bench Dataset Details

Table S1: Details of toxin-producing organisms in the benchmark.

Domain	Organism Name
Bacteria	Aeromonas hydrophila, Aggregatibacter actinomycetemcomitans, Bacillus anthracis, Bacillus cereus, Bacillus thuringiensis, Bacteroides fragilis, Bordetella pertussis, Burkholderia pseudomallei, Caulobacter vibrioides, Chlamydia trachomatis serovar D, Clostridium botulinum, Clostridium perfringens, Clostridium sardiniense, Escherichia coli, Gardnerella vaginalis, Grimontia hollisae, Haemophilus ducreyi, Klebsiella pneumoniae, Listeria monocytogenes serotype 1/2a, Lysinibacillus sphaericus, Mycobacterium tuberculosis, Mycoplasma pneumoniae, Photorhabdus akhurstii, Photorhabdus laumondii subsp. laumondii, Pseudomonas aeruginosa, Pseudomonas entomophila, Salmonella typhi, Salmonella typhimurium, Serratia marcescens, Serratia proteamaculans, Shigella dysenteriae, Shigella flexneri, Shigella sonnei, Staphylococcus aureus, Streptococcus intermedius, Streptococcus mitis, Streptococcus pneumoniae, Streptococcus pyogenes, Streptococcus suis, Vibrio cholerae, Vibrio parahaemolyticus serotype O3:K6, Vibrio splendidus, Vibrio vulnificus, Yersinia pseudotuberculosis serotype O:3
Eukaryota	Abrus precatorius, Actinia equina, Actinia fragacea, Aedes aegypti, Agkistrodon contortrix contortrix, Agkistrodon contortrix laticinctus, Agkistrodon piscivorus piscivorus, Ancylostoma caninum, Anopheles albimanus, Anopheles darlingi, Anopheles stephensi, Bitis arietans, Boiga dendrophila, Boiga irregularis, Bothrops asper, Bothrops atrox, Bothrops brazili, Bothrops jararaca, Bothrops jararacussu, Bothrops leucurus, Bothrops moojeni, Bothrops pauloensis, Bothrops pirajai, Bougainvillea spectabilis, Bryonia dioica, Bungarus caeruleus, Bungarus multicinctus, Calloselasma rhodostoma, Canavalia brasiliensis, Canavalia ensiformis, Centrolobium tomentosum, Cerrophidion godmani, Cinnamomum camphora, Conus mucronatus, Conus striatus, Crateva tapia, Cratylia argentea, Crotalus atrox, Crotalus durissus terrificus, Cucurbita moschata, Culex quinquefasciatus, Daboia russelii, Daboia siamensis, Deinagkistrodon acutus, Dermacentor andersoni, Dianthus caryophyllus, Dioclea grandiflora, Dioclea guianensis, Dioclea lasiophylla, Dioclea virgata, Drimia maritima, Echis carinatus, Echis carinatus sochureki, Echis multisquamatus, Eisenia fetida, Gloydius halys, Gloydius intermedius, Haementeria officinalis, Hordeum vulgare, Hydra vulgaris, Iris hollandica, Ixodes ricinus, Loxosceles intermedia, Loxosceles laeta, Luffa acutangula, Luffa aegyptiaca, Lutzomyia longipalpis, Meccus pallidipennis, Metlapilcoatlus nummifer, Micrurus tener tener, Momordica balsamina, Momordica charantia, Musa acuminata, Naja atra, Naja sagittifera, Notechis scutatus scutatus, Ophiophagus hannah, Ornithodoros moubata, Phaseolus vulgaris, Phlebotomus duboscqi, Phlebotomus papatasi, Phytolacca acinosa, Phytolacca americana, Phytolacca dioica, Protobothrops flavoviridis, Protobothrops mangshanensis, Protobothrops mucrosquamatus, Pseudechis australis, Pseudechis porphyriacus, Pseudonaja textilis, Rhipicephalus microplus, Rhodnius prolixus, Ricinus communis, Sambucus nigra, Saponaria officinalis, Sicarius terrosus, Stichodactyla helianthus, Suregada multiflora, Tabanus yao, Trichosanthes kirilowii, Trichosanthes sp. Bac Kan 8-98, Trimeresurus stejnegeri, Triticum aestivum, Tropidechis carinatus, Vipera ammodytes ammodytes, Vipera ammodytes meridionalis, Vipera nikolskii, Viscum album, Zea mays
Viruses	Bovine rotavirus G10, Corynephage beta, Escherichia phage 933W, Rotavirus A, Rotavirus str. I321, Simian rotavirus A/SA11, Ustilago maydis P4 virus, Ustilago maydis P6 virus

Table S2: List of regulated viruses in the benchmark dataset.

Domain	Virus Name
Viruses	African swine fever virus, Crimean-Congo hemorrhagic fever virus, Foot-and-mouth disease virus, Lake Victoria marburgvirus, Monkeypox virus, Orthonairovirus haemorrhagiae, Rift valley fever virus, Severe acute respiratory syndrome coronavirus 2, Sheeppox virus, Tick-borne encephalitis virus

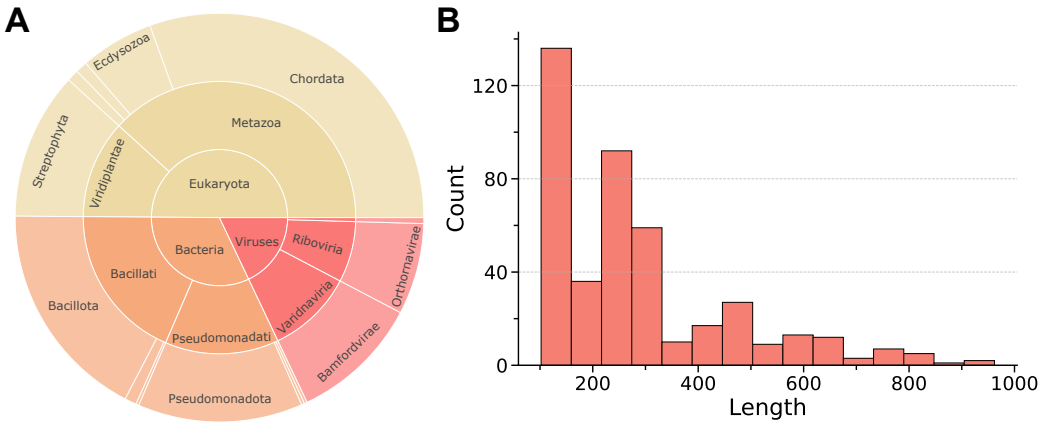

Figure S1: Details of SafeProtein-Bench dataset. (A) Taxonomic distribution of the test cases in our dataset. (B) Length distribution of the sequences in dataset.

B Related Work

B.1 Co-design of Protein Sequence and Structure

A deeper understanding of protein properties and the ability to design proteins with desired functions are central goals in protein engineering. Recent advances in deep learning have significantly accelerated this process. Early efforts, such as ProGen [33], adopted autoregressive language modeling to *de novo* generate protein sequences. However, the real-world applicability of these unimodal protein language models is limited by their lack of structural information, given the intrinsic connection between protein structure and function. This gap has motivated a paradigm shift toward integrating both sequence and structure modalities. For instance, xTrimoPGLM [34] demonstrated that large-scale protein language models can implicitly capture the mapping between sequence and structure. Nevertheless, it remains restricted to autoregressive sequence generation and does not explicitly incorporate structural information. To address this, MultiFlow [35] introduced flow-based generative modeling for protein design, enabling simultaneous sampling of sequences and their corresponding structures, thereby achieving stronger alignment between the two modalities. Similarly, ProteinGenerator [36] supports sequence–structure co-design and allows additional constraints such as amino acid preferences. However, these two approaches mainly emphasize *de novo* generation, with only limited capability to accommodate design tasks based on partial or customized sequence–structure inputs. Building on this direction, representative models such as ESM3 [14] further extend the scope of protein design by discretizing protein structures into tokens and employing multi-track architectures to unify representations across modalities. This enables flexible functionality: *de novo* sequence design, motif scaffolding, sequence-conditional structure prediction, or joint sequence–structure generation. Likewise, DPLM2 [28] achieves free-form cross-modal mapping by concatenating sequence and structure representations during training. Together, ESM3 and DPLM2 represent powerful tools for protein design, providing versatile frameworks tailed for real-world scenario application and is suitable to serve as target model in our work.

B.2 Benchmark and Evaluation of Jailbreak Attacks for LLMs

To rigorously evaluate the performance of a proposed red-teaming framework and to reasonably verify the vulnerabilities of large language models, it is essential to employ comprehensive testing benchmarks alongside efficient evaluation strategies. In the realm of large language models, numerous publicly available benchmarks already exist for this purpose. For instance, AdvBench [19] collected 500 harmful strings and 500 harmful behavior instructions across diverse unsafe themes, designed to test jailbreak attacks on both exact harmful outputs and compliance with unsafe requests. HarmBench [20] includes 510 harmful behaviors (400 textual and 110 multimodal) spanning seven semantic categories such as cybercrime, misinformation, and harassment to enable standardized, robust evaluation of jailbreak attacks. AIR-BENCH 2024 [21] includes 314 fine-grained risk categories derived from global regulatory frameworks and corporate policies, to provide a unified evaluation of AI risks. These open benchmarks underpin the foundation of LLMs jailbreak evaluation. To measure the jailbreak success rate, past works mainly rely on ways like rule-based or keyword matching systems or LLM-as-a-judge to evaluate the target model's outputs [18, 22]. These prior studies from LLM realm offer valuable insights for our work on jailbreaking protein foundation models. However, there are still some gaps to bridge, especially due to the lack of standard protein jailbreak benchmark as well as the differences between protein language and natural language.

C Red-teaming Results of DPLM2

Table S3: Red-teaming attack success rates of SafeProtein against DPLM2 on SafeProtein-Bench. Results are reported for all masking strategies. For each sequence masking ratio, the best success rate is highlighted in **bold**.

Gen Strategy	Masking Strategy	Sequence Masking Ratio					
		0.1	0.2	0.25	0.3	0.4	0.5
Strategy1	Conservation	**36.36**	**29.84**	**26.81**	**21.45**	**15.62**	**12.59**
	Random	18.88	9.56	10.02	6.06	7.23	8.63
	Tail	8.86	2.33	3.50	1.87	1.63	1.63
Strategy2	Conservation	**42.66**	**34.50**	**32.40**	**27.97**	**20.05**	**16.32**
	Random	21.45	9.79	10.72	6.99	7.93	9.32
	Tail	5.83	2.10	4.20	2.56	1.40	1.87
Strategy3	Conservation	**44.29**	**33.10**	**30.54**	**26.57**	**20.51**	**17.72**
	Random	19.58	8.63	10.02	7.46	6.29	8.16
	Tail	9.32	1.63	4.66	2.80	2.80	2.80