# Structural Persistence Despite Sequence Redaction: A Biosecurity Evaluation of Protein Language Models

**Petr Simecek**[*]
Central European Institute of Technology, Masaryk University
Kamenice 753/5, 625 00 Brno
petr.simecek@ceitec.muni.cz

## Abstract

Redacting or deleting portions of sensitive protein sequences is a common practice intended to reduce dual-use risks when publishing or sharing biological data. However, recent advances in generative protein models challenge the effectiveness of this approach. We demonstrate that even after masking up to 85% of a protein's sequence, state-of-the-art multi-modal models (ESM3) can regenerate sequences that fold into nearly identical three-dimensional structures, preserving complex topological features. Using protein knots as a case study—structures requiring precise spatial arrangements—we show that partial sequence disclosure may not meaningfully reduce the ability of advanced models to reconstruct high-risk proteins. We further demonstrate that models can transform benign proteins into topologically complex variants through iterative modification (31% success rate). These findings highlight critical vulnerabilities in current biosecurity practices and underscore the need for AI-native defenses that address structural recovery directly. We propose our masking-recovery evaluation framework as a benchmark for assessing biosecurity risks in generative biological AI systems.

## 1 Introduction

The rapid advancement of generative AI models for protein design presents a fundamental challenge to biosecurity: how do we prevent the misuse of powerful capabilities while enabling beneficial research? Current biosecurity practices rely heavily on controlling access to complete sequences of potentially dangerous proteins through redaction, database restrictions, and publication guidelines Lewis et al. [2020], igs [2024], Schmidt [2009], for Biosecurity [2007]. These approaches assume that partial information significantly impedes reconstruction of functional designs.

We present evidence that this assumption is no longer valid. Using protein knots—complex topological structures found in less than 1% of known proteins Taylor [2000]—as a test case, we demonstrate that modern multi-modal protein language models can reconstruct complete, functional structures from heavily redacted sequences. Specifically, we show that ESM3 Hayes et al. [2025] can recover knotted protein structures even when 85% of the original sequence is deleted or obscured.

This "redaction failure" phenomenon has immediate implications for biosecurity. Information hazard mitigation strategies based on partial sequence disclosure may provide false security, as access controls that restrict complete sequences but allow partial access could be circumvented. Moreover, publication guidelines that recommend redacting sensitive regions may not prevent reconstruction of dangerous proteins by adversaries with access to modern AI tools.

Recent work has demonstrated the generation of artificial knotted proteins using diffusion models and sequence design tools Klimentová and Simecek [2024], Watson et al. [2023], Dauparas et al. [2022], Alamdari et al. [2023], though with extremely low success rates ( 0.5%). Our approach

[*] Corresponding author

leverages ESM3's guided generation capabilities to achieve dramatically improved recovery rates while highlighting biosecurity implications.

Our contributions are threefold: First, we quantify the robustness of structural features to sequence redaction, establishing that complex topologies persist despite extensive masking. Second, we demonstrate guided generation capabilities that achieve 87% success in recovering rare structural features, a 170-fold improvement over baseline approaches. Third, we propose a generalizable evaluation framework for assessing reconstruction risks in generative biological AI systems.

## 2 Background and Threat Model

### 2.1 Current Biosecurity Practices

Standard biosecurity measures for sequence data encompass multiple layers of control. DNA synthesis providers screen orders against databases of known pathogens igs [2024], while sensitive sequences remain restricted to authorized researchers through controlled access systems. Publications often employ redaction protocols, removing key regions from sequences before public release for Biosecurity [2007], Schmidt [2009], and function-based restrictions focus on preventing access to specific functional domains that could be weaponized. These measures collectively assume that incomplete information creates a meaningful barrier to recreating dangerous designs.

### 2.2 Capabilities of Modern Protein Models

Recent protein language models like ESM3 Hayes et al. [2025], AlphaFold3 Abramson et al. [2024], and Boltz Wohlwend et al. [2024] integrate multiple modalities in unprecedented ways. These models can predict structure from sequence Lin et al. [2023], perform inverse folding to design sequences from structural constraints Hsu et al. [2022], execute guided generation with custom objectives Ingraham et al. [2019], Rao et al. [2021], and complete masked sequences with high accuracy Rao et al. [2019]. These capabilities enable reconstruction pathways that bypass traditional safeguards. The ability to perform guided generation based on scoring functions makes these models particularly suitable for targeting rare protein features that might have dual-use potential.

### 2.3 Threat Model

We consider an adversary who has access to partial sequence information from publications, leaks, or authorized databases, possesses computational resources to run open-source protein models, and seeks to reconstruct functional proteins with specific properties. Critically, we do not assume the adversary has expertise in protein engineering or access to specialized wet-lab facilities for initial design—only for downstream synthesis and testing. This threat model reflects the democratization of AI capabilities, where sophisticated protein design tools are becoming accessible to non-experts.

## 3 Methods

### 3.1 Evaluation Framework

We develop a systematic framework to evaluate structural reconstruction from partial sequences, as detailed in Algorithm 1. This framework systematically tests the ability of modern models to recover functional protein structures from increasingly redacted sequences.

### 3.2 Test Case: Protein Knots

We select protein knots as our test case for several strategic reasons. From a biosecurity perspective, protein knots provide an ideal model system that allows us to study reconstruction capabilities without directly working with dangerous sequences. While knots themselves are not inherently harmful, they share critical properties with potentially dangerous protein features: they require precise spatial arrangements that cannot form by chance Virnau et al. [2006], they are rare in nature (occurring in less than 1% of proteins), making successful generation non-trivial Taylor [2000], Dabrowski-Tumanski and Sułkowska [2017], and they have clear binary detection criteria using established algorithms Lua [2012], Mishra and Bhushan [2012].

**Algorithm 1** Redaction Robustness Evaluation

```
Input: protein sequence $s$, redaction levels $R$, trials $N$
Output: reconstruction success rates
for each redaction level $r \in R$ do
    Initialize $success\_count = 0$
    for $i = 1$ to $N$ do
        Randomly redact $r\%$ of sequence $s \rightarrow s_{redacted}$
        Generate missing regions using model $\rightarrow s_{recovered}$
        Predict structure of $s_{recovered}$
        Evaluate structural similarity to original
        If similarity $>$ threshold: $success\_count$++
    end for
    Record success rate = $success\_count/N$
end for
```

Most importantly, protein knots represent a level of structural complexity comparable to functional motifs found in dangerous proteins Orengo et al. [1997]. They exhibit enhanced stability and unique functional properties Strassler et al. [2022], King et al. [2007], Soler and Faisca [2013], and their evolutionary conservation across species suggests functional importance Brems et al. [2022], Begun et al. [2024], Puri and Hsu [2022]. This makes them an excellent proxy for studying how an adversary might reconstruct functionally important motifs from partial information, without the ethical concerns of working directly with pathogenic sequences.

### 3.3 Model and Generation Protocol

We employ ESM3-SM, the 1.4B parameter version of ESM3, which is currently the only publicly available variant of this powerful model family. While larger versions (7B and 98B parameters) exist and would likely show even stronger reconstruction capabilities, they remain restricted. Our use of ESM3-SM serves as a conservative lower bound for what will soon be possible when larger models become publicly accessible—a scenario that appears increasingly likely given the rapid democratization of AI capabilities Hayes et al. [2025].

For controlled reconstruction, we implement guided generation following:

$$p(x|c) \propto p(x) \cdot \exp(\lambda \cdot s(x)) \tag{1}$$

where $s(x)$ is a scoring function for desired structural features and $\lambda$ controls guidance strength. We use the Topoly package with Alexander polynomials for knot detection and scoring. See Figure 1 for an example of generated knotted protein.

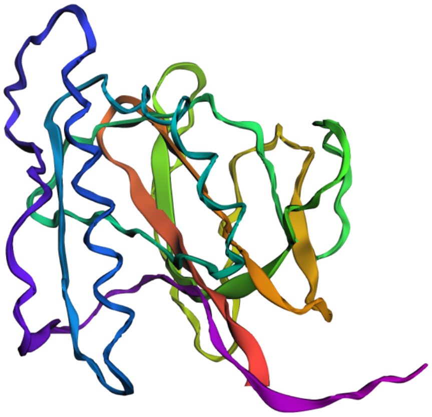

Figure 1: Example of a knotted protein generated through ESM3 guided generation. The structure maintains a stable knot topology while exhibiting realistic secondary structure elements.

### 3.4 Dataset and Evaluation

We analyze 1,000 knotted and 4,000 unknotted proteins from publicly available structural databases. For each protein, we apply varying levels of sequence redaction (5%-95%), attempt reconstruction using guided generation, verify topological preservation using established knot detection algorithms, and quantify structural similarity using TM-score. This comprehensive evaluation allows us to map the relationship between information retention and reconstruction success.

## 4 Results

### 4.1 Structural Features Persist Despite Extensive Redaction

Our primary finding challenges the effectiveness of sequence redaction as a biosecurity measure. We demonstrate that protein structures, including complex topological features, can be recovered even when the majority of sequence information is withheld.

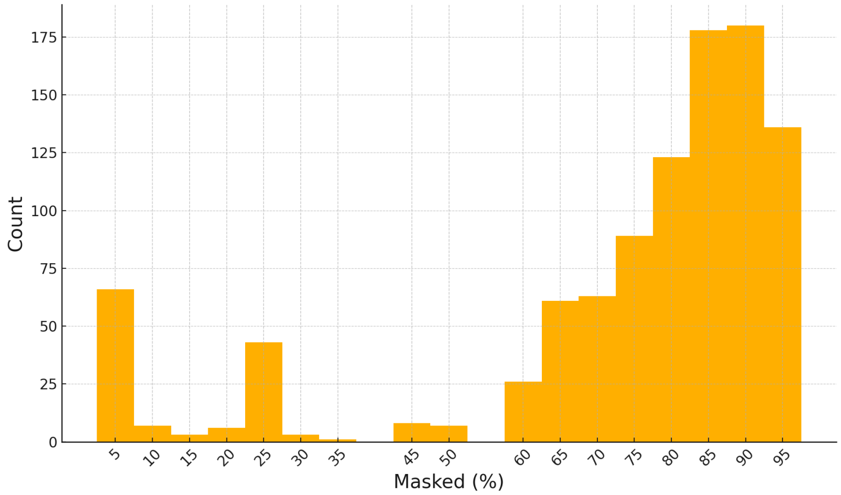

Figure 2: Distribution of minimum sequence retention needed to preserve structural topology. Most proteins maintain their knot structure until 80-85% of the sequence is redacted, indicating high robustness to information loss.

Figure 2 shows that on average, 85% of a protein sequence must be deleted before the topological structure is lost. This implies that retaining just 15-20% of a sequence may be sufficient for reconstruction—far less than what typical redaction protocols would remove.

### 4.2 Guided Generation Achieves High Recovery Rates

Using guided generation with structural objectives, we achieve an 87% success rate in generating knotted proteins from fully masked sequences, representing a 170-fold improvement over the 0.5% baseline rate observed with unguided generation. Additionally, we demonstrate a 31% success rate in converting unknotted proteins to knotted variants through iterative modification. These results demonstrate that modern models can be effectively steered toward specific structural outcomes, even extremely rare ones that occur in less than 1% of natural proteins.

### 4.3 Model Embeddings Capture Structural Information

Training a simple neural network classifier on ESM3 embeddings achieves 93% accuracy in detecting knotted proteins from sequence alone. This remarkable accuracy indicates that models learn implicit structural representations that capture complex topological features Bepler and Berger [2019], Rao et al. [2019]. The ability to infer topological features without explicit structural information suggests that partial sequences may contain sufficient signal for reconstruction, even when critical regions appear to be missing.

Table 1: Reconstruction success rates at different redaction levels

| Redaction Level | Success Rate | Structural Similarity |
|---|---|---|
| 25% | 98% | 0.95 |
| 50% | 94% | 0.91 |
| 75% | 87% | 0.84 |
| 85% | 71% | 0.76 |
| 90% | 42% | 0.63 |
| 95% | 18% | 0.41 |

### 4.4 Critical Threshold Behavior

We observe non-linear degradation in reconstruction success, as shown in Figure 3. Rather than a gradual decline, there exists a critical threshold where reconstruction capability drops sharply. This suggests that a critical amount of sequence information enables reconstruction, and below this threshold, reconstruction fails rapidly. Current redaction practices may unknowingly operate near this critical point, providing a false sense of security.

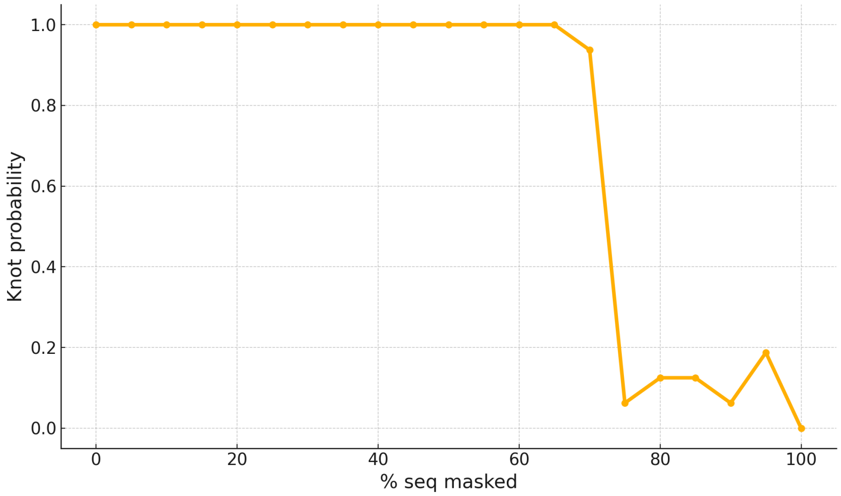

Figure 3: Knot probability after masking for one selected protein. Sharp transition in reconstruction success near the critical redaction threshold, indicating non-linear information requirements.

## 5 Biosecurity Implications

### 5.1 Vulnerability of Current Safeguards

Our findings reveal three critical vulnerabilities in current biosecurity practices. First, redaction is fundamentally insufficient: removing 50-75% of a sequence—far more than typically redacted in publications—still allows high-fidelity reconstruction. Second, if complex knots can be recovered from partial information, simpler functional domains such as binding sites or catalytic residues are likely even more robust to redaction. Third, the ability to use guided generation means adversaries don't need complete sequences—just enough information to steer the model toward desired structural outcomes.

### 5.2 Implications for Different Threat Scenarios

The implications extend across multiple threat scenarios. In toxin reconstruction, partial sequences of protein toxins published in scientific literature could be completed to functional designs using guided generation. For virulence factor engineering, key pathogenic domains could be reconstructed from fragments and integrated into novel protein contexts. In the context of immune evasion, antibody escape mutations could be explored even with incomplete structural data, potentially accelerating the development of vaccine-resistant variants Casadevall et al. [2014].

### 5.3 Recommendations for AI-Native Safeguards

Based on our findings, we recommend a fundamental shift in biosecurity strategy. First, policies should assume that any partial sequence disclosure enables full reconstruction—the traditional assumption that redaction provides meaningful protection is no longer valid. Second, models should incorporate function-based filtering that can detect and refuse to generate sequences with dangerous functions, regardless of input completeness. Third, continuous monitoring of model capabilities for reconstruction from partial information should become a standard component of safety evaluations. Finally, effective biosecurity will require layered defenses that combine model-level safeguards with downstream synthesis screening and anomaly detection systems.

## 6 Proposed Evaluation Benchmark

We propose RedactBench as a standardized evaluation suite for assessing biosecurity risks in generative biological AI systems. This benchmark tests reconstruction capability across different feature types, measures the minimum information needed for successful recovery, evaluates the effectiveness of proposed safeguards, and includes both benign test cases and dual-use relevant features. Key metrics include reconstruction rate at different redaction levels, information threshold for successful recovery, preservation of specific functional motifs, and the ability to bypass different defense mechanisms.

## 7 Limitations and Future Work

Our study has several limitations that should guide future research. We use topological features as a proxy for functional properties, though the relationship between topology and function may vary across protein families. Our experiments were conducted with the smallest publicly available ESM3 model; larger models will likely show even stronger reconstruction capabilities. Additionally, we focus on single-domain proteins, while many dangerous proteins contain multiple domains that might behave differently under redaction.

Future work should extend this evaluation to specific dual-use relevant features, though such research must be conducted with appropriate safeguards and oversight. Testing reconstruction of actual toxin or virulence sequences would provide direct evidence but raises ethical concerns that must be carefully managed. Development and evaluation of AI-native defense mechanisms represents a critical need, as does investigation of reconstruction from non-sequence modalities such as structure fragments or functional descriptions.

## 8 Conclusion

We demonstrate that sequence redaction—a cornerstone of current biosecurity practice—fails to prevent reconstruction of complex protein structures when advanced generative models are available. The ability to recover functional designs from as little as 15-20% of sequence information fundamentally challenges assumptions about information hazard mitigation in biology.

Our findings underscore the urgent need for new biosecurity approaches that account for AI capabilities. Rather than relying on information restriction alone, we must develop AI-native safeguards that detect and prevent generation of dangerous designs regardless of input completeness. The evaluation framework we present offers a systematic way to assess these risks and test proposed defenses.

As generative models continue to improve, the gap between partial information and functional designs will only narrow. The biosecurity community must adapt quickly, developing new strategies that remain effective even as AI capabilities advance. Our work provides both a warning and a tool for meeting this challenge.

## Acknowledgments

We thank the anonymous reviewers for their insightful comments. The project was supported by the OPUS LAP program of the Czech Science Foundation, project no. 23-04260L ("Biological code of knots – identification of knotted patterns in biomolecules via AI approach"). Computational resources were provided by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

## References

International gene synthesis consortium harmonized screening protocol v3.0. Technical report, International Gene Synthesis Consortium (IGSC), 2024.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Bob Strome, Alan Moses, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pages 2023–09, 2023.

AM Begun, AA Korneev, and AV Zorina. The effect of a knot on the thermal stability of protein mj0366: Insights from molecular dynamics and monte carlo simulations. *arXiv preprint arXiv:2411.04390*, 2024.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using structural information. 2019.

Maarten A Brems, Robert Runkel, Todd O Yeates, and Peter Virnau. Alphafold predicts the most complex protein knot and composite protein knots. *Protein Science*, 31(8):e4380, 2022.

Arturo Casadevall, Terence S Dermody, Michael J Imperiale, Rozanne M Sandri-Goldin, and Thomas Shenk. On the need for a national board to assess dual use research of concern, 2014.

Pawel Dabrowski-Tumanski and Joanna I. Sułkowska. Topological knots and links in proteins. *Proceedings of the National Academy of Sciences*, 114(13):3415–3420, 2017. doi: 10.1073/pnas.1615862114.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

National Science Advisory Board for Biosecurity. Proposed framework for the oversight of dual use life sciences research: strategies for minimizing the potential misuse of research information, 2007.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, page eads0018, 2025.

Chloe Hsu et al. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022.

John Ingraham et al. Generative models for graph-based protein design. 2019.

Neil P King, Eric O Yeates, and Todd O Yeates. Identification of rare slipknots in proteins and their implications for stability and folding. *Journal of molecular biology*, 373(1):153–166, 2007.

Eva Klimentová and Petr Simecek. Unveiling the entangled landscape of artificial knotted proteins. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024.

Grace A. Lewis, Jacob L. Jordan, David A. Relman, Gregory D. Koblentz, Jade Leung, Allan Dafoe, Cassidy Nelson, Gerald L. Epstein, Rebecca Katz, Michael Montague, Ethan C. Alley, Claire Marie Filone, Stephen Luby, George M. Church, Piers Millett, Kevin M. Esvelt, Elizabeth E. Cameron, and Thomas V. Inglesby. The biosecurity benefits of genetic engineering attribution. *Nature Communications*, 11(1):6294, 2020. doi: 10.1038/s41467-020-19149-2.

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.

Rhonald C Lua. Pyknot: a pymol tool for the discovery and analysis of knots in proteins. *Bioinformatics*, 28(15):2069–2071, 2012.

Rama Mishra and Shantha Bhushan. Knot theory in understanding proteins. *Journal of mathematical biology*, 65(6):1187–1213, 2012.

Christine A. Orengo, Patrice Koehl, Michael Levitt, and Janet M. Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997. doi: 10.1016/S0969-2126(97)00227-3.

Sarita Puri and Shang-Te Danny Hsu. Elucidation of folding pathways of knotted proteins. *Methods in enzymology*, 675:275–297, 2022.

Roshan Rao et al. Evaluating protein transfer learning with tape. 2019.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International conference on machine learning*, pages 8844–8856. PMLR, 2021.

Markus Schmidt. Synthetic genomics: Options for governance. *Nature Reviews Genetics*, 10(7): 491–496, 2009. doi: 10.1038/nrg2624.

Miguel A Soler and Patricia FN Faisca. Effects of knots on protein folding properties. *PloS one*, 8(9): e74755, 2013.

Sarah E Strassler, Isobel E Bowles, Debayan Dey, Jane E Jackman, and Graeme L Conn. Tied up in knots: Untangling substrate recognition by the spout methyltransferases. *Journal of Biological Chemistry*, 298(10), 2022.

William R. Taylor. A deeply knotted protein structure and how it might fold. *Nature*, 406(6798): 916–919, 2000. doi: 10.1038/35022619.

Peter Virnau, Leonid A Mirny, and Mehran Kardar. Intricate knots in proteins: Function and evolution. *PLoS computational biology*, 2(9):e122, 2006.

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.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, pages 2024–11, 2024.