Advancing Knotted Protein Design with ESM3: Guided Generation and Topological Insights

Petr Simecek [1] **Eva Marsalkova** [1 2]

Abstract

Knotted proteins represent a rare but functionally important class of proteins with complex topological features. While previous work demonstrated the generation of artificial knotted proteins using diffusion models and sequence design tools, these approaches suffered from low success rates (0.5%). We present a novel approach leveraging ESM3, a multi-modal protein language model, to achieve guided generation of knotted proteins with an 87% success rate. We introduce a continuous knot score metric that captures the robustness of protein knots, revealing that approximately 85% of a protein sequence must be altered to break its knot. Using ESM3 embeddings, we achieve 93% accuracy in knotted protein classification and demonstrate the ability to convert unknotted proteins to knotted variants through iterative modifications (31% success rate). Our work showcases the power of multi-modal models in tackling complex protein design challenges.

1. Introduction

Proteins with knotted topologies represent one of nature's most intriguing structural motifs, occurring in less than 1% of known protein structures (Virnau et al., 2006). Despite their rarity, knotted proteins exhibit enhanced stability and unique functional properties that make them attractive targets for protein engineering and drug design (Strassler et al., 2022). The ability to design and manipulate knotted proteins could unlock new therapeutic and biotechnological applications.

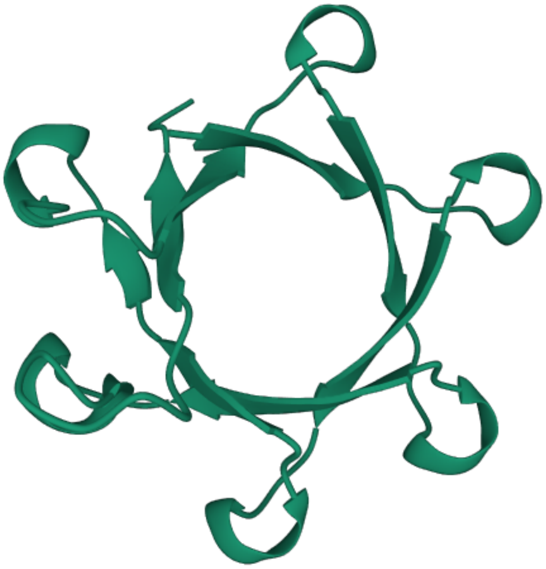

Figure 1. Example of a knotted protein: A0A165M1R5 protein (*Acidovorax* sp.), featuring a 7_1 knot, as predicted by ESMFold (Lin et al., 2023).

Knotted proteins are a unique class of proteins in which the backbone forms a knot-like structure, example on Figure 1. Imagine pulling on both ends of a protein chain like a piece of string—if the protein is knotted, it won't come apart (Takusagawa & Kamitori, 1996; Mishra & Bhushan, 2012). Such structures offer several advantages:

Enhanced stability: The formation of knots in proteins can significantly enhance the kinetic stability of proteins during folding and unfolding processes (Lua, 2012), (Beccara et al., 2013). Knotted proteins are often more resistant to mechanical degradation by proteolytic machines like ClpP/X, which could provide a survival benefit in various cellular contexts (Soler & Faisca, 2013).

Unique functional properties: The complex topology creates distinctive active sites and binding pockets, contributing to enzymatic activity and substrate specificity (King et al.,

[1] Central European Institute of Technology (CEITEC), Faculty of Science, Masaryk University, Brno, Czechia [2] National Centre for Biomolecular Research (NCBR), Faculty of Science, Masaryk University, Brno, Czechia. Correspondence to: Petr Simecek <simecek@mail.muni.cz>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

2007). For example, the enzyme N-acetylornithine transcarbamylase (AOTCase), which possesses a knot, showcases how topology can affect enzymatic pathways, leading to different substrate processing compared to its homologues without knot (Brems et al., 2022).

Evolutionary conservation: The preservation of knots across species and within protein families suggests functional importance, indicating these complex topologies provide selective advantages (Brems et al., 2022), (Begun et al., 2024), (Puri & Hsu, 2022).

Despite these intriguing properties, the rarity of knotted proteins in nature poses challenges for their study. Recent advances in AI-driven protein design have demonstrated the possibility of generating knotted proteins: (Klimentová & Simecek, 2024) combined RFdiffusion & ProteinMPNN together with EvoDiff, discovering three knot types (5_1, 7_4, and 8_{19}) not previously observed in natural proteins, (Watson et al., 2023), (Dauparas et al., 2022), (Alamdari et al., 2023). However, these approaches suffered from low success rates ($\sim 0.5\%$ success rate for generating knotted proteins) and lacked control over the generation process.

The recent release of ESM3 (Hayes et al., 2025), a frontier-scale multi-modal protein language model, offers new opportunities for controlled protein design. ESM3 unifies multiple protein modalities—sequence, structure, and function—in a single framework, enabling capabilities that previously required multiple specialized tools: structure prediction (AlphaFold2), inverse folding (ProteinMPNN), and structure generation (RFdiffusion). Its ability to perform guided generation based on scoring functions makes it particularly suitable for targeting rare protein features like knots.

In this work, we leverage ESM3 to address the challenge of knotted protein design through three key contributions: (1) We introduce a continuous knot score metric that quantifies the robustness of protein knots, moving beyond binary classification; (2) We demonstrate ESM3-guided generation achieving 87% success rate in producing knotted proteins, a dramatic improvement over previous methods; (3) We show that ESM3 embeddings enable highly accurate classification and facilitate the transformation of unknotted proteins into knotted variants through iterative modifications.

2. Methods

We took advantage of previously published data on 1,000 knotted and 4,000 unknotted real proteins from `EvaKlimentova/Diffusion-all_knots` dataset stored at HuggingFace Hub[1]. The code used for the analysis was uploaded to GitHub repository https://github.com/ML-Bioinfo-CEITEC/KPDwESM3. Experiments were conducted on a virtual machine equipped with an NVIDIA A100 GPU, 16 CPU cores, and 64 GB of memory.

2.1. ESM3 Model and Guided Generation

We employed the open-source ESM3-SM (1.4B parameters) model. While larger models may exhibit higher fidelity in structure generation, the ESM3-SM provides an effective platform for demonstrating the key capabilities of the multi-modal architecture:

- **Structure prediction**: sequence → structure
- **Inverse folding**: structure → sequence
- **Masked sequence prediction**: partially masked sequence → sequence
- **Embeddings**: sequence → learned representations
- **Guided generation**: conditional sampling with custom scoring functions

The methods presented here are model-agnostic and are expected to show continued success when applied to larger-scale models as they become publicly available (like EMS3 7B and 98B models).

For guided generation of knotted proteins, we implement a scoring function $s(x)$ that evaluates the knottedness of generated structures using the Topoly Python package with Alexander polynomials https://github.com/ilbsm/topoly_tutorial. The generation process follows:

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

where $p(x)$ is the base ESM3 distribution, c represents conditioning information, $s(x)$ is the knot metric (either original or smoothed) and λ controls the strength of guidance.

2.2. Continuous Knot Score

Traditional knot detection provides binary labels (knotted/unknotted). To quantify knottiness, we first process Topoly output by extracting probabilities for all predicted topologies and calculating the proportion of trivial topology (0_1, unknotted).

We then propose a continuous knot score that captures the robustness of protein knots through randomized masking - see Algorithm 1.

For our analysis, we use $X = 10\%$ masking and $N = 16$ trials. The knot score is then defined as:

$$\text{knot\_score} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}[\text{topology}_i \neq 0\_1] \tag{2}$$

[1] https://huggingface.co/datasets/EvaKlimentova/Diffusion-all_knots

Algorithm 1 Randomized Knot Score Calculation

```
**Input:** protein sequence $s$, masking perc. $X$, trials $N$
**Output:** knot_score $\in [0, 1]$
Initialize $knotted\_sum = 0$
**for** $i = 1$ **to** $N$ **do**
    Randomly mask $X\%$ of sequence $s \rightarrow s_{masked}$
    Generate masked regions using ESM3 $\rightarrow s_{generated}$
    Predict structure of $s_{generated}$ using ESM3
    Calculate topology using Topoly
    $knotted\_sum = knotted\_sum + P(topology \neq 0\_1)$
**end for**
**Return** $knot\_score = knotted\_sum/N$
```

where $\mathbb{P}[\text{topology}_i \neq 0\_1]$ is the probability of knottiness. This score ranges from 0 (robustly unknotted) to 1 (robustly knotted), providing nuanced information about borderline cases.

2.3. Knot Stability Analysis

To assess knot stability, we systematically analyze how much sequence modification is required to break a knot. For each of 1,000 knotted proteins, we apply masking percentages from 5% to 95% in 5% increments. For each masking percentage, we perform 16 independent trials and calculate the randomized knot score. We define the "breaking point" as the masking percentage where the knot score falls below 0.75.

2.4. Guided Protein Generation and Transformation

We implement two distinct protocols for working with knotted proteins:

(a) De novo generation of knotted proteins: Starting from a fully masked sequence, we use guided generation with $\lambda = 1.0$ and the knot score as the objective function to directly generate knotted proteins.

(b) Unknotted-to-knotted transformation: We start with an unknotted protein sequence and iteratively modify it through guided generation. In each iteration, we randomly mask 5% of the sequence and apply guided regeneration to these masked regions using the knot score as the objective. This process continues for up to 10 iterations or until the protein achieves a knot score above 0.8, indicating a successful transformation to a knotted structure.

2.5. Knot Classification Using ESM3 Embeddings

To evaluate the discriminative power of ESM3 representations for knot detection, we train a neural network classifier. We use 1,000 knotted proteins from our dataset with sequences masked and regenerated from 5% to 95%, creating a dataset of 383,523 training proteins and 42,613 validation proteins. Different proteins were used for training and validation masking to prevent data leakage. Both datasets contain approximately 73% knotted proteins.

We extract ESM3 embeddings (dimension 1,536) using sequences only and train a fully connected neural network with two hidden layers (1,024 and 256 neurons) for binary classification (knotted vs. unknotted).

3. Results

3.1. ESM3 Embeddings Enable Accurate Knot Classification

We first evaluated whether ESM3's learned representations capture topological information. Training a simple neural network classifier on ESM3 sequence embeddings, we achieved 93% accuracy in distinguishing knotted from unknotted proteins, see Table 1. This high accuracy suggests that ESM3 implicitly learns topological features despite not being explicitly trained on knot labels.

Table 1. Confusion matrix on the validation set. The model achieves 93% accuracy in identifying knotted proteins.

TRUE VS. PREDICTED	UNKNOTTED	KNOTTED
UNKNOTTED	10644	866
KNOTTED	1924	29179

3.2. Knot Topology is Robust to Sequence Perturbation

Our analysis showed that knots are remarkably stable under sequence perturbation. On average, more than 80% of a protein's sequence must be altered to disrupt the knot (Figure 2), indicating that the core topological information is preserved within a fraction of the original sequence.

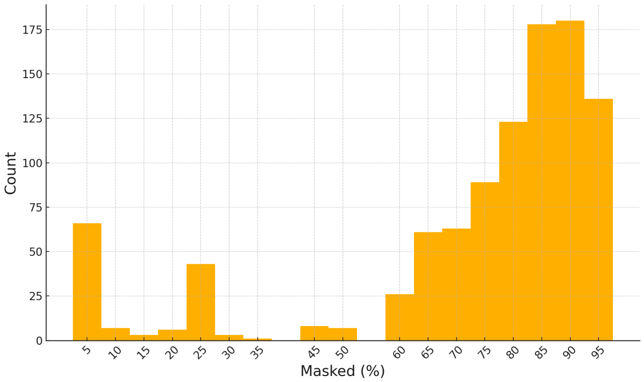

Figure 2. Histogram of percentage of sequence that must be masked to break the knot.

What we originally expected was a smooth decrease in the knot score from one to zero, as in Figure 3.

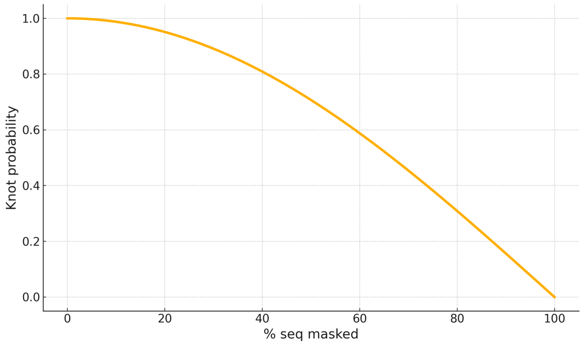

Figure 3. Expected dependence of knot probability on percentage of masking may look like.

In fact, the probability of maintaining the knot drops sharply around the breaking point, indicating that a critical threshold is needed for topological stability, see Figure 4.

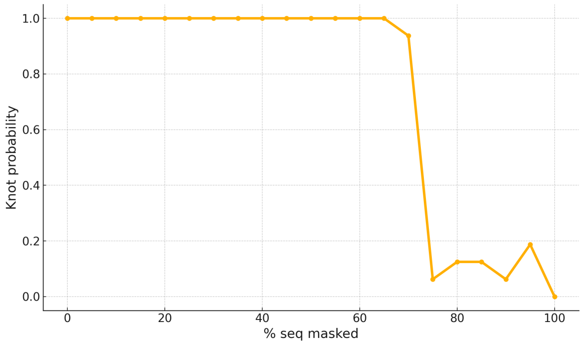

Figure 4. Observed dependence of knot probability on masking percentage for a representative protein.

3.3. Guided Generation Achieves 87% Success Rate

Using ESM3's guided generation capabilities, we achieved an 87% success rate in generating knotted proteins; see the example in Figure 5.

This represents $\sim$ 170-fold improvement over the unguided approaches ($\sim$ 0.5% success rate) used in previous work.

Our iterative modification protocol successfully converted 31% of unknotted proteins into knotted variants. Successful transformations typically required 3-5 iterations.

4. Discussion and Conclusion

This work demonstrates the transformative potential of multi-modal foundation models like ESM3 for complex protein design tasks. By unifying sequence, structure, and function modeling in a single framework, ESM3 enables capabilities that previously required multiple specialized tools, while adding powerful new features like guided generation.

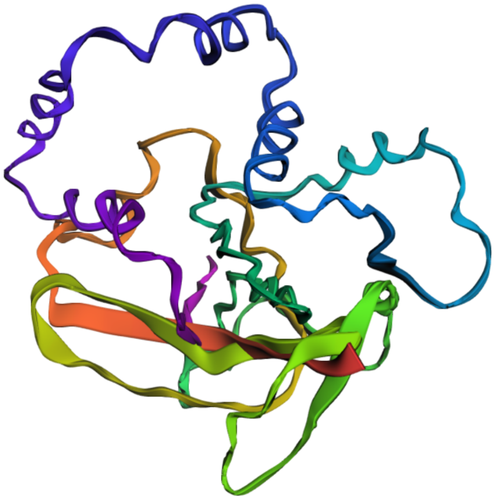

Figure 5. Example of generated knotted protein.

The finding that a large majority of sequence must be modified to break the knot has implications for understanding knot evolution and designing stable knotted proteins for applications. We verified that ESM3 is not simply memorizing sequences and the generated knotted proteins differ significantly from their original counterparts.

The dramatic improvement in generation success rate (from 0.5% to 87%) showcases the power of guided generation for targeting rare protein features. This success was achieved despite several simplifications in our approach. First, our random masking strategy did not distinguish between the knot core and terminal regions. Second, we did not incorporate secondary structure information or spatial proximity relationships in our generation process. A more sophisticated masking strategy that accounts for topological importance could potentially improve both generation success rates and the functional relevance of designed proteins.

While we acknowledge the limitations of using the smaller open-source ESM3 model, the techniques presented here are model-agnostic. Their relevance is underscored by the emergence of powerful new foundation models, such as Boltz (Wohlwend et al., 2024) and AlphaFold3 (Abramson et al., 2024), which are increasingly capable of integrating diverse biological data. The high accuracy (93%) we achieved in knot classification using only sequence embeddings suggests that these larger models will likely capture even more subtle topological features with greater fidelity.

Impact Statement

Our results highlight how multi-modal foundation models are reshaping protein design, enabling precise control over complex structural features that were previously accessible only through undirected sampling. As these models continue to scale and improve, we anticipate even greater capabilities for engineering proteins with desired topological and functional properties. The ability to generate genuinely novel knotted proteins—rather than minor variants of known sequences—opens new avenues for creating proteins with enhanced stability and unique functional properties for biotechnological applications.

Acknowledgments

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 supplied by the project ”e-Infrastruktura CZ” (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Alamdari, S., Thakkar, N., van den Berg, R., Tenenholtz, N., Strome, B., Moses, A., Lu, A. X., Fusi, N., Amini, A. P., and Yang, K. K. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pp. 2023–09, 2023.

Beccara, S., Škrbić, T., Covino, R., Micheletti, C., and Faccioli, P. Folding pathways of a knotted protein with a realistic atomistic force field. *PLoS computational biology*, 9(3):e1003002, 2013.

Begun, A., Korneev, A., and Zorina, A. 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.

Brems, M. A., Runkel, R., Yeates, T. O., and Virnau, P. Alphafold predicts the most complex protein knot and composite protein knots. *Protein Science*, 31(8):e4380, 2022.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

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

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

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

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

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

Puri, S. and Hsu, S.-T. D. Elucidation of folding pathways of knotted proteins. *Methods in enzymology*, 675:275–297, 2022.

Soler, M. A. and Faisca, P. F. Effects of knots on protein folding properties. *PloS one*, 8(9):e74755, 2013.

Strassler, S. E., Bowles, I. E., Dey, D., Jackman, J. E., and Conn, G. L. Tied up in knots: Untangling substrate recognition by the spout methyltransferases. *Journal of Biological Chemistry*, 298(10), 2022.

Takusagawa, F. and Kamitori, S. A real knot in protein. *Journal of the American Chemical Society*, 118(37):8945–8946, 1996. doi: 10.1021/ja961147m.

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

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, pp. 2024–11, 2024.