# UNVEILING THE ENTANGLED LANDSCAPE OF ARTIFICIAL KNOTTED PROTEINS

**Eva Klimentová**
Central European Institute of Technology
Masaryk University
Brno, 62500, Czech Republic
National Centre for Biomolecular Research, Faculty of Science
Masaryk University
Kamenice 5, 625 00 Brno, Czech Republic
469217@mail.muni.cz

**Petr Simecek**
Central European Institute of Technology
Masaryk University
Brno, 62500, Czech Republic
simecek@mail.muni.cz

## ABSTRACT

In this study, we delve into the generation of novel artificial knotted proteins, leveraging state-of-the-art computational techniques such as EvoDiff and RFdiffusion, in tandem with ProteinMPNN. Our aim is to broaden the spectrum of existing protein structures with novel knotted configurations, thereby deepening our insight into the intricate phenomenon of protein knotting. Our findings reveal that the generated artificial proteins closely mimic the natural occurrence of knotted proteins, with a comparable percentage exhibiting non-trivial topologies. Additionally, we introduce several knot types previously unobserved in natural proteins. At the heart of our study is the curated dataset of these artificial knotted proteins, aligned with their natural counterparts for comprehensive comparison. This dataset can serve as a benchmark, encouraging the development and application of new protein generation methodologies.

## 1 INTRODUCTION

Proteins are essential molecular machines whose functions are deeply intertwined with their unique three-dimensional (3D) structures. The quest to understand and predict these structures has significantly advanced with the advent of deep neural network models such as AlphaFold2 (AF) (Jumper et al., 2021) or OmegaFold (Wu et al., 2022). These tools have revolutionized the accuracy and speed of protein structure prediction.

Alongside progress in structure prediction, the field of artificial protein design has also advanced significantly. A notable method involves the integration of RFdiffusion (Watson et al., 2022) and ProteinMPNN (Dauparas et al., 2022) tools. This technique leverages the power of diffusion models, trained on a vast dataset of known protein structures, to generate diverse sequence-independent backbones. These backbones are then refined by incorporating sequence information to optimize structure and side-chain placements. Another exciting development is the EvoDiff method (Alamdari et al., 2023). Unlike the stepwise approach that typically involves predicting 3D structures followed by sequence deduction, EvoDiff directly targets the sequence space, bypassing structural predictions.

Within the vast diversity of protein structures, a particularly intriguing group stands out: the knotted proteins (Takusagawa & Kamitori, 1996). Knots, familiar to us from everyday life, can also occur within the polypeptide chain of proteins (Mishra & Bhushan, 2012), forming intricate loops and crossings (see Figure 1). Despite their rarity in nature, knotted proteins are a subject of increasing interest due to their distinct characteristics. The knotted arrangement is believed to confer enhanced stability and specific functional attributes to these proteins (Dabrowski-Tumanski et al., 2016), differentiating them from their unknotted counterparts. The presence of knot been shown to confer resistance to degradation and denaturation (Lim & Jackson, 2015), making knotted proteins attractive candidates for biotechnological applications (Xu & Zhang, 2018). Moreover, the unique structural features of knotted proteins have been implicated in specific biological functions, such as

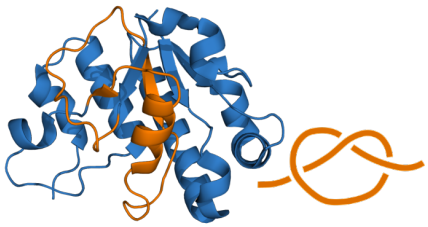

Figure 1: Human protein 5NFJ with a knotted backbone containing a highlighted $3_1$ knot and a simplified visualization of the knot.

modulation of enzymatic activity and regulation of ligand binding (Virnau et al., 2006). As a result, understanding the principles governing the folding and stability of knotted proteins could provide valuable insights into protein design and engineering, potentially leading to the development of novel enzymes, drug targets, or biomaterials with enhanced properties (Sulkowska, 2020) .

The limited variety of natural knotted proteins, mostly confined to a few families (Šrámková et al., 2023), hints at a vast potential for artificial knotted proteins. While studying natural knotted proteins provides valuable insights, it will always be constrained by the narrow range of protein families in which they occur. In contrast, the universe of artificial proteins is much wider, offering the opportunity to explore a far greater diversity of knotted topologies and their associated functions. Using computational tools, we aim to create unprecedented knotted proteins, thereby expanding our knowledge and application of protein functionalities.

## 2 Results

### 2.1 Generated Sets of Artificial Proteins

Our generation efforts yielded $212,681$ structures using RFdiffusion, with $2,814$ $(1.3\%)$ displaying non-trivial topologies. Of these, ProteinMPNN successfully designed sequences for $1,037$ structures $(0.47\%)$. EvoDiff produced $433,992$ sequences, with $2,168$ $(0.50\%)$ forming knotted structures, closely aligning with the natural occurrence rate of $\sim 0.35\%$ as estimated in Šrámková et al. (2023). The predominant knot observed was the simple $3_1$, accounting for $89\%$ of cases. Remarkably, we identified three knot types not previously observed in nature (Figure 2). It's important to note that given the rarity of knotted proteins in nature, the performance of computational tools like ProteinMPNN and AlphaFold on these topologies should be validated in future with experimental structural determination of select designed proteins.

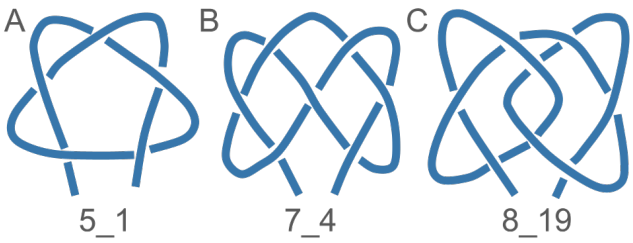


Figure 2: Novel knot types: (A) $5_1$ (eg. RF_3_251_1), (B) $7_4$ (eg. E_ID99324), (C) $8_{19}$ (eg. RF_3_153_1)

We created a dataset (available through `https://huggingface.co/datasets/EvaKlimentova/Diffusion-all_knots`) of randomly selected $1,000$ knotted proteins from each method together with a random sample of $1,000$ knotted proteins identified in AlphaFold2 database from Šrámková et al. (2023). As controls, we added $4,000$ unknotted proteins for each method, so the total number of proteins in the dataset is $3 * (1000 + 4000) = 15000$. See Table 1 for knot type distribution for each group.

Table 1: Knot type counts

| Tool | 3_1 | 3_1#3_1 | 4_1 | 5_1 | 5_2 | 6_1 | 6_3 | 7_4 | 8_19 | N/A |
|---|---|---|---|---|---|---|---|---|---|---|
| EvoDiff | 866 | 2 | 44 | 15 | 11 | 1 | 1 | 1 | 0 | 59 |
| RFdiffusion + MPNN | 950 | 5 | 7 | 26 | 1 | 0 | 0 | 0 | 3 | 8 |

### 2.2 Comparison to natural proteins

Our analysis aimed to evaluate the resemblance between the generated artificial proteins and their natural counterparts, focusing on both sequence and structural levels. Using BLAST for sequence alignment against the nr database, we found that a small but notable percentage of artificial proteins matched known natural proteins, see Figure 3 C. Specifically, the match rates for sequences generated by both methods and across knot statuses ranged from $7.3\%$ to $10.4\%$.

Structural alignment, conducted via Foldseek against the AlphaFoldDB, revealed more pronounced differences, see Figure 3 D. The EvoDiff-generated structures exhibited a lower similarity to known proteins, with only $27.2\%$ showing significant matches, compared to a much higher match rate of $92.2\%$ for structures generated using RFdiffusion combined with MPNN. This disparity was even more marked among knotted proteins, where match rates for EvoDiff and RFdiffusion + MPNN were $18.0\%$ and $96.8\%$, respectively, against $29.4\%$ and $99.8\%$ for their unknotted counterparts. These findings suggest that the structural configurations of knotted proteins generated by EvoDiff are more divergent from known protein structures, whereas RFdiffusion + MPNN tends to produce structures that more closely resemble existing proteins in databases.

To further assess the viability and fidelity of the generated proteins, we computed two metrics: average pLDDT and perplexity (see Methods). The pLDDT scores and perplexity values were plotted to identify trends in protein quality and complexity (Figure 3 A/B). Proteins situated in the bottom right corners of these subplots are considered most plausible, exhibiting high confidence in structural prediction and lower sequence complexity. Manual inspection of three such proteins from the EvoDiff set revealed two with high similarity to known proteins (E_ID119285, E_ID119291), and one characterized by repetitive sequences (E_ID25027).

### 2.3 Predictive Models

The T-SNE visualization of ProtBert-BFD embeddings for the sequences in the dataset (Figure 4A) corroborates our earlier findings, indicating significant differences between proteins generated by distinct methods. However, focusing on an individual method (Figure 4B-D), the differences between knotted and unknotted sequences are evident.

Given these observations, an intriguing question emerges: Is it possible to apply insights from artificially generated proteins to their natural counterparts? To investigate this, we developed two neural network classifiers, each trained on datasets from EvoDiff and RFdiffusion + MPNN, utilizing ProtBert-BFD embeddings as features. We allocated 20% of each dataset for validation purposes and employed a collection of real protein sequences as the test set.

Given the disproportionate number of unknotted proteins relative to knotted ones, we assessed our classifiers' performance using the odds-ratio statistic. An odds-ratio significantly above 1 indicates the model's proficiency in differentiating between knotted and unknotted proteins, with higher values reflecting greater discriminative power. The classifiers demonstrated robust performance on their respective validation sets, achieving odds-ratios of 8.82 for the EvoDiff-trained model and 8.04 for the RFdiffusion + MPNN-trained model. However, when applied to the test set of real proteins, the odds-ratios were 1.62 and 1.61, respectively, still statistically significant (z-score test, $p < 0.01$), hinting at the potential for knowledge transfer from artificial to real proteins, albeit with limitations.

## 3 Methods

We employed two strategies for generating knotted proteins: combining RFdiffusion (Watson et al., 2022) with ProteinMPNN (Dauparas et al., 2022), and using EvoDiff (Alamdari et al., 2023).

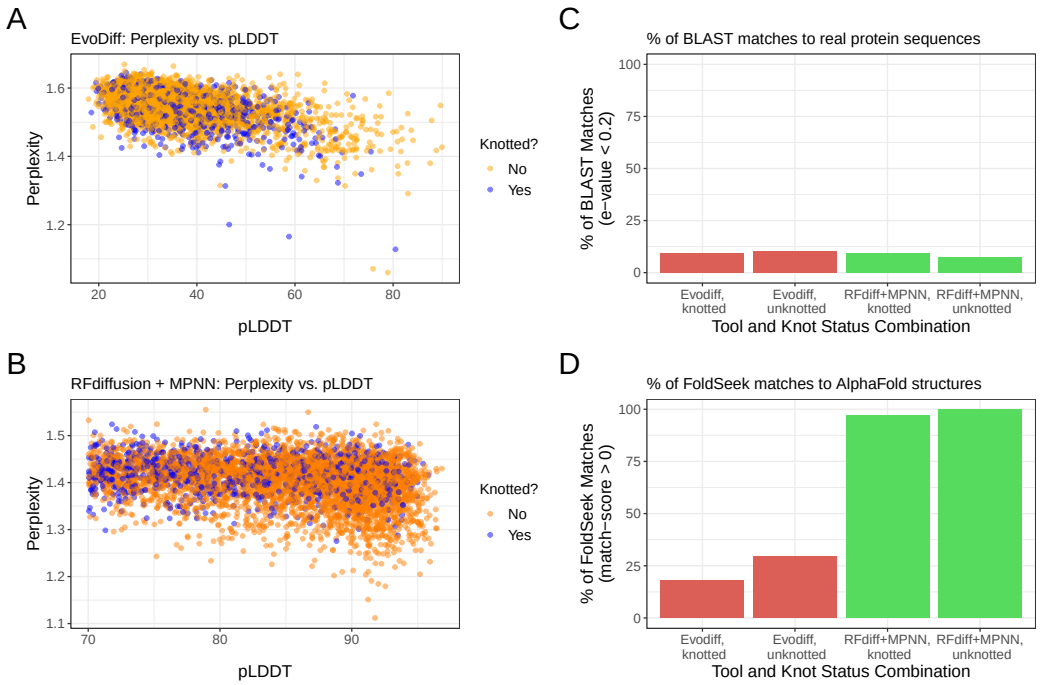


Figure 3: Viability of the generated proteins and the comparison to their natural counterparts. (A),(B) pLDDT vs Perplexity for EvoDiff and RFDiffusion+MPNN, respectively. (C) BLAST comparison to natural proteins' sequences. Only $e - value < 0.2$ counted as mach. (D) FoldSeek comparison to natural proteins' structures. Any $match\_score > 0$ counted as match.

### 3.1 Protein generation with RFdiffusion + MPNN

We configured RFdiffusion according to its GitHub repository guidelines, setting `contigmap.contigs=[100-500]` and `inference.num_designs=1` to generate protein structures ranging from 100 to 500 amino acids. The Topoly Python package, using the Alexander polynomial, evaluated the topology of these structures. Structures with an unknot probability below 0.5 and a specific knot probability above 0.4 were deemed knotted and advanced for sequence design via ProteinMPNN, following the ColabDesign GitHub guidelines. For each knotted backbone, up to eight sequences were designed with ProteinMPNN, but only the first sequence that achieved a pLDDT score of 70 or higher and retained the knotted topology in ColabFold predictions was included in our knotted designs dataset. Unsuccessful designs after eight attempts were discarded.

Unknotted designs from RFdiffusion underwent a similar process, filtering designs having an unknot probability above 0.5. They were complemented with ProteinMPNN sequences passing the criterion of Colabfold-predicted 3D structure being unknotted, having pLDDT greater or equal to 70 and RMSD of the RFdiffusion designed structure and Colabfold-predicted 3D structure being lower than 5.

### 3.2 Protein generation with EvoDiff

EvoDiff was configured according to its GitHub instructions to generate sequences of 100-500 amino acids using the default model (oa_dm_640M). Omegafold predicted the 3D structures of these sequences, categorizing them into knotted or unknotted sets based on their topology probabilities.

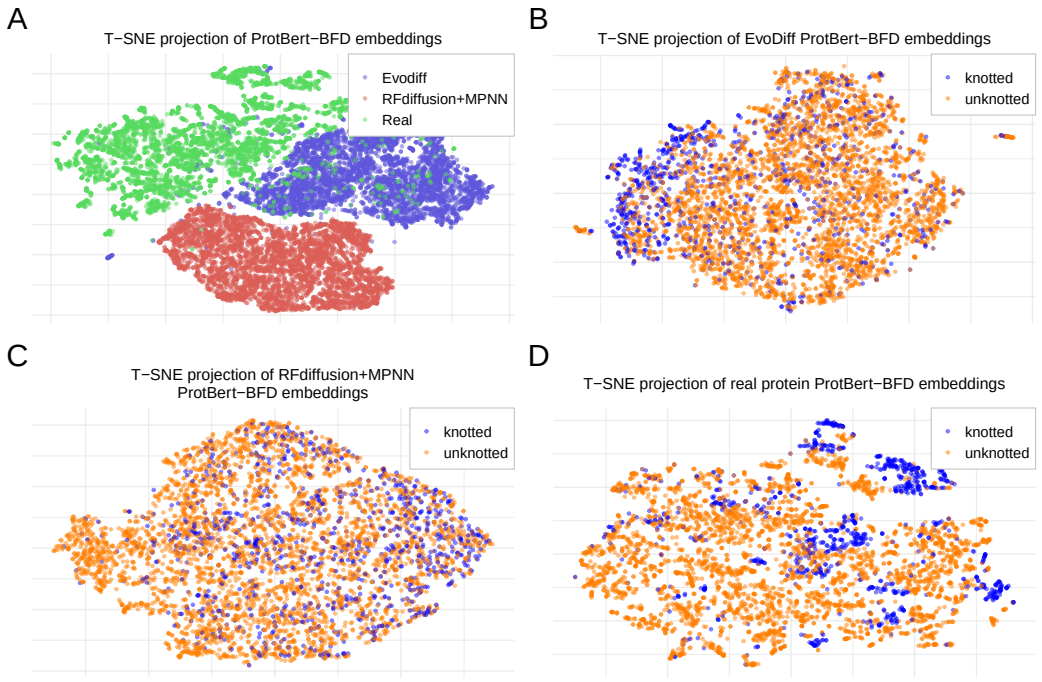


Figure 4: T-SNE projections of protein ProtBert-BFD embeddings. (A) Projection of EvoDiff-generated and RFdiffusion+MPNN-generated proteins and selection of real protein sequences. (B) Projection of EvoDiff-generated proteins coloured by knotting status. (C) RFdiffusion+MPNN-generated proteins coloured by knotting status. (D) Real protein sequences coloured by knotting status.

### 3.3 Analysis of the generated proteins

For the further analysis of the artificially generated proteins, $1,000$ knotted and $4,000$ unknotted proteins from both RFdiffusion + MPNN and EvoDiff datasets were randomly picked, making together a dataset of $10,000$ proteins.

**BLAST search** Using `NCBIWWW.qblast` function from Biopython (v 1.81) (Cock et al., 2009), we performed BLASTP (Altschul et al., 1990) searches against the NCBI non-redundant protein database (nr), reporting the top hit for each sequence.

**Foldseek search** We employed Foldseek (van Kempen et al., 2023) to search against AlphaFoldDB (Varadi et al., 2021) and PDB (Berman et al., 2000), reporting the top hits and their TM scores. The search was done through Foldseek's web page API with the TM-align algorithm and AlphaFold/UniProt50 and PDB100 databases.

**ProtBert-BFD embeddings and perplexity** The generated proteins' sequences were run through the ProtBert-BFD model (Elnaggar et al., 2022) and the last layer embeddings of the proteins were extracted and processed using average pooling to yield a fixed-size vector of 1024 dimensions. To assess the complexity of the generated sequences, the perplexity (Chen et al., 2008) of each sequence was computed with the ProtBert-BFD model by masking each amino acid with `[MASK]` token.

**Knot core determination** For knotted proteins, we identified the knot core by sequentially trimming residues and reassessing the topology until the knot was lost or changed, using the Topoly package's homfly method (Dabrowski-Tumanski et al., 2020).

**Neural Network Classifiers** for knot status prediction were trained using the `fastai` package's `tabular_learner` (Howard & Gugger, 2020) with two hidden layers (500 and 200 neurons). The training involved three epochs with a maximum learning rate of $0.005$, leveraging protein embeddings as input.

## 4 DISCUSSION

In this study, we explored the design of artificial knotted proteins using two methodologies: EvoDiff and a combination of RFdiffusion with ProteinMPNN. While both methods showed promise in generating complex structures, they differed significantly in their outputs. The RFdiffusion+ProteinMPNN approach produced proteins more akin to known structures, as shown by structural similarity metrics. In contrast, EvoDiff was better at generating novel sequences, though these were less likely to resemble existing proteins. While our study demonstrates the potential of generative models to sample knotted protein topologies, further validation on more diverse test sets is needed to confirm the robustness of these findings.

To support further research in this area, we've made available a dataset of 10,000 generated proteins and the corresponding analysis code through `https://github.com/ML-Bioinfo-CEITEC/ArtificialKnottedProteinsPaper`. Our ongoing goal is to develop methods that not only create novel proteins but also accurately reflect natural protein knotting, enabling the transfer of insights from artificial to natural proteins. This balance between innovation and biological relevance is crucial for advancing our understanding of protein structures and their potential applications. We aspire to bridge the gap between the realms of artificial and natural proteins, fostering a deeper understanding of protein folding and knotting mechanism.

### ACKNOWLEDGMENTS

We are immensely grateful to Prof. Joanna Sułkowska, Agata Perlińska, and Maciej Sikora for their expert feedback and significant insights. The work was supported by the OPUS LAP program of the Grant Agency of Czech Republic (Reg. No. 204/07/1592 grant "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

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex X. Lu, Nicolo Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023. doi: 10.1101/2023.09.11.556673.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi: https://doi.org/10.1016/S0022-2836(05)80360-2.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235.

Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. Evaluation Metrics For Language Models. 1 2008. doi: 10.1184/R1/6605324.v1. URL `https://kilthub.cmu.edu/articles/journal_contribution/Evaluation_Metrics_For_Language_Models/6605324`.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 03 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163.

Pawel Dabrowski-Tumanski, Andrzej Stasiak, and Joanna I. Sulkowska. In search of functional advantages of knots in proteins. *PLOS ONE*, 11(11):1–14, 11 2016. doi: 10.1371/journal.pone.0165986.

Pawel Dabrowski-Tumanski, Pawel Rubach, Wanda Niemyska, Bartosz Ambrozy Gren, and Joanna Ida Sulkowska. Topoly: Python package to analyze topology of polymers. *Briefings in Bioinformatics*, 22(3), 09 2020. ISSN 1477-4054. doi: 10.1093/bib/bbaa196.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Jeremy Howard and Sylvain Gugger. Fastai: A layered api for deep learning. *Information*, 11(2): 108, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Nicole CH Lim and Sophie E Jackson. Molecular knots in biology and chemistry. *Journal of Physics: Condensed Matter*, 27(35):354101, 2015.

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

Denisa Šrámková, Maciej Sikora, Dawid Uchal, Eva Klimentová, Agata P. Perlinska, Mai Lan Nguyen, Marta Korpacz, Roksana Malinowska, Pawel Rubach, Petr Šimeček, and Joanna I. Sulkowska. Knot or not? sequence-based identification of knotted proteins with machine learning. *bioRxiv*, 2023. doi: 10.1101/2023.09.06.556468.

Joanna Ida Sulkowska. On folding of entangled proteins: knots, lassos, links and $\theta$-curves. *Current opinion in structural biology*, 60:131–141, 2020.

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

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, May 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061.

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, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, 2022. doi: 10.1101/2022.12.09.519842.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022. doi: 10.1101/2022.07.21.500999.

Lianjie Xu and Wen-Bin Zhang. Topology: a unique dimension in protein engineering. *Science China Chemistry*, 61(1):3–16, Jan 2018. doi: 10.1007/s11426-017-9155-2.

# A APPENDIX

**Supplementary data:** Dataset presented in this paper together with source code for its analysis are available through: `https://github.com/ML-Bioinfo-CEITEC/ArtificialKnottedProteinsPaper` and `https://huggingface.co/datasets/EvaKlimentova/Diffusion-all_knots`