Generative AI for Biomaterial Innovation: The Case of Spider Silk

Neeru Dubey*
KTH Royal Institute of Technology
Stockholm, Sweden
needub@kth.se

Abstract

Designing proteins with tunable mechanical properties remains a key challenge in biomaterials research. We present a generative modeling framework conditioned on experimentally measured properties to create protein sequences with targeted strength, extensibility, and toughness. By linking sequence motifs to macroscopic material behavior, this approach enables controllable, AI-driven design of next-generation functional biomaterials.

1 Introduction

Spider silk has long inspired biomimetic material science owing to its exceptional strength-to-weight ratio and elasticity. However, replicating its mechanical performance in synthetic analogues remains challenging for three main reasons. First, the relationship between sequence composition and mechanical behavior is still poorly understood. Second, the modular and highly repetitive architecture of spidroin proteins complicates modeling. Finally, experimental synthesis and testing are slow and resource-intensive.

Recent advances in protein language models (PLMs) present a new opportunity to learn sequence–property relationships directly from data. Yet, most existing PLMs are optimized for structural or functional prediction rather than mechanical property conditioning. In this work, we introduce a conditional generative framework that links spidroin sequence representations; particularly the highly repetitive core region shown in Fig. 1; to experimentally measured mechanical traits. This enables the inverse design of silk proteins with tailored properties derived from repeat-region composition and motif organization.

This work summarizes the core methodology and results on our recent publication [Dubey et al., 2025], where we first introduced the mechanical property–conditioned generative model for spidroin design.

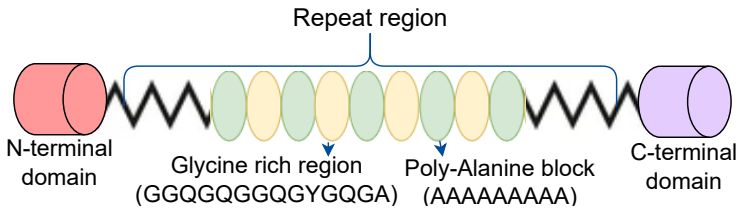

Figure 1: Schematic of a spidroin showing terminal domains and the repetitive region.

*Corresponding author: needub@kth.se

2nd Conference on Applied AI and Scientific Machine Learning (CASML 2025).

2 Methodology

We propose a two-stage framework for generating *MaSp* repeat sequences conditioned on mechanical properties and for predicting properties from given repeats. Unlike prior work on full-length proteins [Lu et al., 2024], our model focuses on the repetitive core regions of *MaSp* proteins that govern fiber mechanics. The approach begins by distilling ProtGPT2 [Ferruz et al., 2022] into a compact student model, *SpiderGPT*, pretrained on 100,000 spider proteins from UniRef50 [The UniProt Consortium, 2024]. *SpiderGPT* is then fine-tuned in two stages on the Spider Silkome dataset [Arakawa et al., 2022]: Level 1 uses 6,000 *MaSp* repeats to learn motif-level patterns, while Level 2 applies a custom five-fold cross-validation on 592 annotated sequences to model sequence–property relationships. This hierarchical setup enables *SpiderGPT* to handle both generative and predictive tasks with strong domain specificity and generalization.

To evaluate model performance, we used two datasets: a cross-validation test set of 185 instances from the original 592-sequence corpus to assess self-consistency, and a BLAST-based novelty set to measure out-of-distribution sequence diversity. Evaluation followed a dual-level analysis: at the sequence level, we examined molecular weight, instability index, isoelectric point, and motif frequencies (GGX, poly-Ala, YGQGG, and SV); for structural validation, we employed secondary structure prediction to verify α-helix, β-strand, and random-coil compositions characteristic of *MaSp*.

3 Results and Discussion

We evaluated *SpiderGPT* across two dimensions: sequence generation and sequence–property correlation. Generated sequences showed strong biological plausibility, closely matching natural proteins across nine physicochemical and structural metrics, including KL divergence, Hamming distance, molecular weight, isoelectric point, and instability index. Mean differences across features were below 5%, and secondary structure and motif patterns aligned with natural *MaSp* sequences, confirming biological fidelity. We further linked sequence features to mechanical traits—toughness, tensile strength, strain at break, and Young's modulus—and assessed prediction accuracy using Pearson's r, Spearman's ρ, MAE, RMSE, and cosine similarity. As summarized in Table 1, *SpiderGPT* effectively reproduces experimental property trends and captures key sequence–mechanics relationships. It also shows comparison with the closest current benchmark, *SilkomeGPT*.

Table 1: Comparison of SpiderGPT and SilkomeGPT on mechanical property prediction; bold values denote better scores.

Metric	**SpiderGPT (ours)**	**SilkomeGPT (baseline)**
Pearson Correlation (r)	**0.8884**	0.8349
Spearman Correlation (ρ)	**0.8343**	0.7798
Mean Absolute Error (MAE)	**0.0861**	0.0963
Root Mean Square Error (RMSE)	**0.1047**	0.1162
R^2 Score	**0.6383**	0.5861
Cosine Similarity	**0.9827**	0.9783

4 Conclusion and Next Steps

We aim to establish a holistic deep learning framework to understand, generate, and experimentally validate spidroins. Our findings lay the groundwork for next-generation synthetic biomaterials with potential applications across medicine, textiles, and engineering.

As next steps, we aim to generate new spidroin sequences using *SpiderGPT* and validate them through wet-lab synthesis in collaboration with the SLU biology team. This combined computational–experimental pipeline will test whether predicted mechanical properties translate to real fibers and iteratively refine model performance.

In parallel, we are expanding and annotating the spidroin dataset to enable automated classification of functional subtypes. These efforts will strengthen sequence–property modeling and support data-driven discovery of next-generation silk biomaterials.

Acknowledgments and Disclosure of Funding

This work was supported by grant WASP–DDLS 22:035 from the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation, and by the Swedish e-Science Research Centre (SeRC). The author also thanks Prof. Hedvig Kjellström for her valuable feedback on the poster.

References

Kazuharu Arakawa, Nobuaki Kono, Ali D Malay, Ayaka Tateishi, Nao Ifuku, Hiroyasu Masunaga, Ryota Sato, Kousuke Tsuchiya, Rintaro Ohtoshi, Daniel Pedrazzoli, et al. 1000 spider silkomes: Linking sequences to silk physical properties. *Science advances*, 8(41):eabo6043, 2022.

Neeru Dubey, Elin Karlsson, Miguel A. Redondo, Johan Reimegård, Anna Rising, and Hedvig Kjellström. Customizing spider silk: Generative models with mechanical property conditioning for protein engineering. *Transactions on Machine Learning Research*, 7, 2025.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Wei Lu, David L Kaplan, and Markus J Buehler. Generative modeling, design, and analysis of spider silk protein sequences for enhanced mechanical properties. *Advanced Functional Materials*, 34(11):2311324, 2024.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53 (D1):D609–D617, 11 2024. ISSN 0305-1048. doi: 10.1093/nar/gkae1010. URL `https://doi.org/10.1093/nar/gkae1010`.