Customizing Spider Silk: Generative Models with Mechanical Property Conditioning for Protein Engineering

Anonymous authors
Paper under double-blind review

Abstract

The remarkable mechanical properties of spider silk, including its tensile strength and extensibility, are primarily governed by the repetitive regions of the proteins that constitute the fiber, the major ampullate spidroins (MaSps). However, establishing correlations between mechanical characteristics and repeat sequences is challenging due to the intricate sequence-structure-function relationships of MaSps and the limited availability of annotated datasets. In this study, we present a novel computational framework for designing MaSp repeat sequences with customizable mechanical properties. To achieve this, we developed a lightweight GPT-based generative model by distilling the pre-trained ProtGPT2 protein language model. The distilled model was subjected to multilevel fine-tuning using curated subsets of the Spider Silkome dataset. Specifically, we adapt the model for MaSp repeat generation using 6,000 MaSp repeat sequences and further refine it with 572 repeats associated with experimentally determined fiber-level mechanical properties. Our model generates biologically plausible MaSp repeat regions tailored to specific mechanical properties while also predicting those properties for given sequences. Validation includes sequence-level analysis, assessing physicochemical attributes and expected distribution of key motifs as well as secondary structure compositions. A correlation study using BLAST on the Spider Silkome dataset and a test set of MaSp repeats with known mechanical properties further confirmed the predictive accuracy of the model. This framework advances the rational design of spider silk-inspired biomaterials, offering a versatile tool for engineering protein sequences with tailored mechanical attributes.

1 Introduction

Recent advancements in protein design, particularly the integration of artificial intelligence (AI), have significantly enhanced our ability to engineer proteins with desired functions. Researchers have used deep learning techniques to improve the design of de novo proteins, achieving a tenfold increase in the success rates of target binding (Bennett et al., 2023). These innovations underscore the transformative potential of AI in protein engineering, paving the way for novel therapeutic interventions and biotechnological applications (Khakzad et al., 2023).

In parallel, the growing demand for sustainable, non-petroleum-based fibers has intensified interest in bio-derived alternatives. Spider silk, known for its exceptional mechanical properties and biodegradability, presents a promising candidate. However, efforts to develop artificial spider silk are hindered by limited knowledge of how the amino acid sequence of spider silk proteins (spidroins) influences the mechanical properties of the fibers. In this context, AI-driven protein engineering offers a powerful tool for designing spidroins that can be spun into fibers with customized performance characteristics.

Spiders spin up to seven different silk types, that all are composed of spidroins (Peakall, 1969). In this work, we focus on major ampullate silk, or dragline silk, renowned for exceptional mechanical properties - tensile strength comparable to steel (up to 1.3 GPa) with extensibility rivaling rubber (>30%) (Vollrath & Knight, 2001). This unique combination yields toughness exceeding both steel and Kevlar (Gosline et al.,

1999), making spider silk particularly attractive for applications ranging from biomedical sutures to high-performance textiles (Bourzac, 2015).

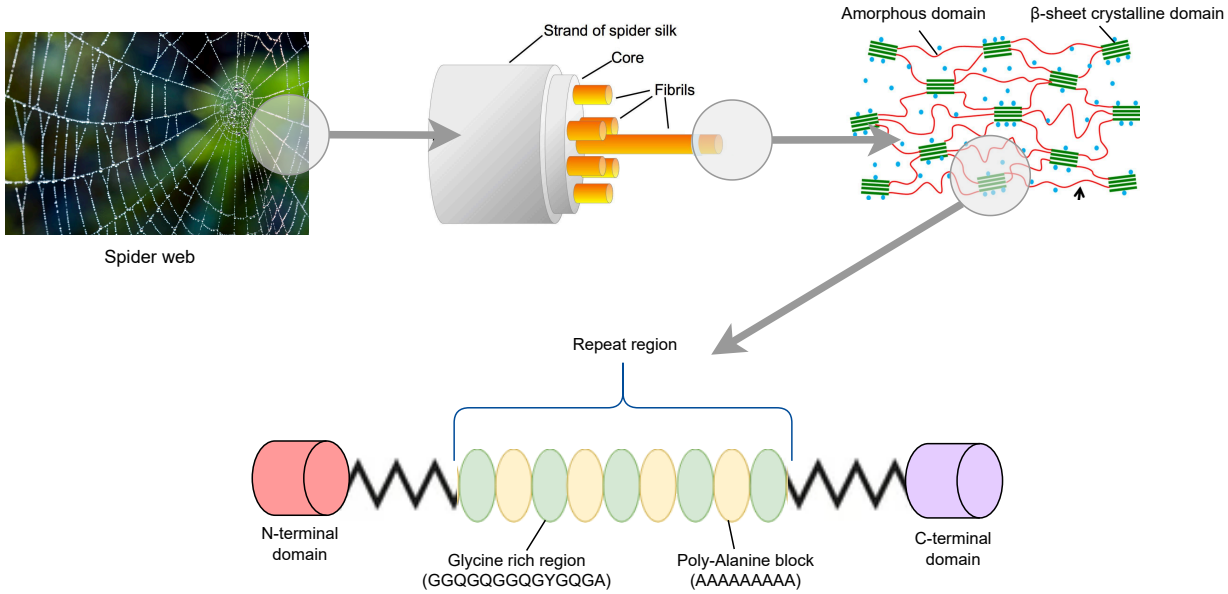

Figure 1: Hierarchical representation of the spider dragline silk fiber architecture, highlighting the schematic image of MaSp showing various sequential elements.

The molecular structure of MaSps comprises three primary regions: the globularly folded N-terminal and C-terminal domains, which are flanking the repetitive core region (Figure 1). The repetitive core region is believed to be the main contributor to the fiber mechanical properties, and in MaSps, this region is generally made up of poly-Ala blocks that are alternated with Gly-rich regions (Guerette et al., 1996). In the polymerized form, in the fiber, the poly-Ala blocks are predominantly arranged in nano-sized β-sheet crystals structures that contribute to the tensile strength of the silk fiber (Hijirida et al., 1996; Lewis, 1992; Yarger et al., 2018b). The poly-Ala β-sheet crystals are embedded in an amorphous matrix formed by the Gly-rich repeats, which is related to the extensibility of the fiber (Gosline et al., 1984). Although often referred to as amorphous, the Gly-rich repeats also form specific conformations like 3_1-helices, β-turns and β-spirals (Hayashi & Lewis, 1998; Hinman & Lewis, 1992; Van Beek et al., 2002).

Native major ampullate silk fibers from different spider species display large variability in mechanical properties. To elucidate the source of this large variability, Arakawa et al. (2022). undertook the significant challenge of sequencing the transcriptome of 1098 species and simultaneously determining the mechanical properties of major ampullate silks from 446 of these species. The results showed that there is a large interspecies difference in terms of mechanical properties. For example, the tensile strength of the major ampullate silks varies between 0.17 and 3.3 GPa (Arakawa et al., 2022). However, strong correlations between the amino acid sequence motifs in the repeat regions of the MaSps and the fiber mechanical properties could not be found (c.f. section 2.1).

Efforts to produce artificial spider silk have largely focused on replicating natural silk spinning processes through a variety of engineered approaches (Ferruz & Höcker, 2022; Koeppel & Holland, 2017). These methods majorly involve expression of engineered mini-spidroins in a heterologous host and subsequent spinning using an artificial spinning device (Schmuck et al., 2024). Although these methods have significantly advanced synthetic production, they often fail to provide a scalable means to tailor silk properties for specific

applications. This limitation further underscores the need for computational approaches to predict and design sequence–property relationships effectively.

In this study, we present a tailored multi-level strategy that addresses two key tasks: generating MaSp repeats customized for desired mechanical properties and predicting mechanical properties from given MaSp repeat sequences. Our model focuses exclusively on the repeat regions, distinguishing it from previous studies (Lu et al., 2024). Our methodology begins with knowledge distillation of ProtGPT2 (Ferruz et al., 2022), a pre-trained generative model, by training a lightweight student model on 100k spider protein sequences, a subset of the UniRef50 database (The UniProt Consortium, 2024). This distilled model then undergoes two-stage fine-tuning using the Spider Silkome dataset (Arakawa et al., 2022): first on 6,000 MaSp repeat sequences to learn general repeat patterns, followed by refinement on 592 MaSp repeat sequences with known fiber-level mechanical properties.

To evaluate the model's effectiveness, we employed two distinct datasets: a test set of 20 instances sampled from the original 592-instance Spider Silkome dataset for assessing self-consistency, and a BLAST set curated to determine sequence novelty within a broader protein sequence database. Our evaluation followed a comprehensive two-level analysis approach. At the sequence level, we analyzed key properties including molecular weight, instability index, isoelectric point, and the distribution of essential motifs (GGX, poly-Ala, YGQGG, and SV). For structural validation, we employed a secondary structure prediction tool to verify the composition of α-helices, β-strands and random coil characteristic of MaSps.

The model's ability to estimate the mechanical properties of silk fibers for a given MaSp repeat was evaluated by correlating generated and reference properties on a test set. The analysis yielded a cosine similarity of 0.9465 between the trend curves of generated and reference properties, indicating a strong alignment in predictive performance.

By combining generative modeling with biological validation, this work offers a robust computational framework for designing MaSp sequences with customizable mechanical properties. The findings hold promise for the advancement of synthetic biomaterial development and pave the way for applications in medicine, textiles, and engineering.

Our main contributions include:

- A novel computational framework that integrates knowledge distillation and multi-level fine-tuning to generate MaSp repeat sequences with targeted mechanical properties, addressing the challenge of limited mechanical property data.
- A dual-purpose model capable of both generating MaSp repeats based on desired mechanical properties and predicting mechanical properties from given sequences, offering flexibility for both design and analysis tasks.
- A comprehensive validation methodology that combines sequence-level analysis, structural prediction, and mechanical property correlation to ensure biological plausibility and functional relevance of generated sequences.
- Empirical demonstration of the model's effectiveness through statistically significant correlations between predicted and reference properties (cosine similarity of 0.94), while maintaining key structural motifs characteristic of spider silk proteins.
- A practical contribution to sustainable biomaterial development by providing a scalable approach for designing synthetic spider silk proteins with customizable mechanical properties.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive review of the literature that covers spider silk sequence-property relationships, protein design using machine learning, and current approaches to modeling mechanical properties of spider silk. Section 3 describes our proposed methodology, including the model architecture and multi-level fine-tuning strategy. Section 4 presents our experimental setup, dataset details, and training procedures. Section 5 discusses our results, including self-consistency assessment, sequence generation analysis, mechanical property predictions, and ablation studies. Finally, Section 6 concludes with a discussion of potential applications and future research directions.

2 Literature Review

2.1 Spider Silk: Sequence-property Relationships

Major ampullate silk is widely regarded a benchmark for high-performance biomaterials due to its unique combination of tensile strength, extensibility, and toughness (Vollrath & Knight, 1999; Gosline et al., 1999). The fiber is primarily composed of MaSps, and to date, five distinct MaSp classes (MaSp 1-5) have been described (Schmuck et al., 2024). The assignment of a MaSp to a specific class is determined by clustering of its terminal domains in phylogenetic analyses and the presence of characteristic amino motifs in the repetitive region.

Analyses of the spidroin sequences in the Spider Silkome database have identified several recurring amino acid motifs within the repetitive regions of the MaSps (Arakawa et al., 2022). A number of these motifs have been highlighted in previous studies, including poly-Ala, GGX, YGQGG, SV, GPGXX, QQ, and AGQG (Arakawa et al., 2022; Keten & Buehler, 2010; Malay et al., 2017; Craig et al., 2020; Nakamura et al., 2024b; Kono et al., 2021b). Attempts to link the occurrence of these motifs individually with the mechanical properties of major ampullate silks have only revealed weak or no correlations (Pearson correlation coefficients < 0.6) (Arakawa et al., 2022). Noticeable among these are the YGQGG motif which is positively associated with toughness and the SV motif which is negatively correlated with this parameter (Arakawa et al., 2022). The poly-Ala, GGX, QQ and AGQG motifs occur frequently in MaSps, but the occurrences of these motifs were not strongly correlated to the fiber mechanical properties (Arakawa et al., 2022). This indicates that the sequence–function relationship of MaSps is complex and results from a synergistic interplay among multiple motifs.

In the present study, we leverage advancements in artificial intelligence to better understand the intricate sequence-to-property correlations, using data-driven models to unravel the hidden patterns governing silk mechanics.

2.2 Protein Design Using Machine Learning and Generative Models

Recent advances in artificial intelligence (AI) have significantly reshaped the landscape of protein sequence design, with generative pretrained language models (PLMs) playing a pivotal role. Models such as ProtGPT2 (Ferruz et al., 2022), ProGen (Madani et al., 2020), and ProtBERT (Brandes et al., 2022) have harnessed natural language processing (NLP) techniques to generate and interpret biologically meaningful protein sequences. These models are trained on extensive datasets of protein sequences, enabling them to discern intricate patterns and relationships that underpin biological functionality (Ferruz & Höcker, 2022). Their success spans the design of enzymes, antibodies, and other functional proteins, frequently surpassing the capabilities of traditional directed evolution methods by producing sequences with enhanced activity or stability. Additionally, structure-aware models like AlphaFold2 (Jumper et al., 2021) and ESMFold (Lin et al., 2022) have elevated sequence design by incorporating structural predictions, while emerging techniques, such as diffusion-based models (Anand et al., 2022), introduce innovative constraint-driven approaches for protein generation. In the context of spider silk proteins (spidroins), these generative PLMs are particularly promising due to their ability to capture the repetitive motifs and hierarchical organization inherent to spidroin sequences, though their application to optimizing mechanical properties remains an ongoing challenge.

2.3 Modeling Mechanical Properties of Spider Silk

Studies investigating the relationship between protein sequence and mechanical properties have primarily relied on sequence analysis and molecular simulations. For instance, (Tokareva et al., 2013) highlighted the importance of GGX and poly-Ala motifs in achieving spider silk's extensibility and tensile properties. Secondary structure prediction tools like PSIPRED (McGuffin et al., 2000) have been widely used to predict α-helices, β-strands and random coil arrangements,, which are critical for understanding silk mechanics (Buchan & Jones, 2019). However, these approaches are often retrospective and do not enable forward design of sequences with desired properties. Efforts to model mechanical properties using machine learning have shown promise. For instance, Kim et al. (2023) utilized neural networks to predict silk's mechanical properties

based on amino acid composition. While such methods enhance understanding of sequence-property relationships, they lack the generative capability required for designing novel sequences. Recent advancements in hybrid AI techniques have shown promise in overcoming these limitations, particularly for engineering silk-like proteins with desired mechanical characteristics (Lu et al., 2024; Ni et al., 2023; Fazio et al., 2023). Notably, Lu et al. (2024) introduced a generative modeling approach for spidroin sequence design, enabling the creation of synthetic spider silk proteins tailored to target mechanical properties. However, a key limitation of this approach is its lack of emphasis on MaSp repeat regions, which are the primary determinants of silk's mechanical behavior.

2.4 Our Proposed Approach in Context

Existing studies have made significant strides in predicting and designing silk proteins, yet they often overlook the critical role of MaSp repeat regions in governing mechanical properties. While predictive models offer insights into sequence-property relationships, they lack generative capabilities, and current generative approaches fail to focus specifically on the functional repeats of MaSp. To address these gaps, our proposed approach introduces a multi-level fine-tuning strategy tailored for MaSp repeat generation. We first train a specialized model to learn the underlying structure and composition of MaSp repeats, ensuring the generated sequences remain biologically relevant. In the second stage, we fine-tune the model to establish a direct correlation between sequence patterns and mechanical properties, enabling controlled generation of MaSp repeats with desired mechanical characteristics. By integrating domain-specific fine-tuning with generative modeling, our method advances beyond existing approaches, providing a more precise and biologically grounded framework for spider silk protein design.

The ability to design spider silk proteins with tunable mechanical properties holds immense potential for sustainable material innovation. Applications extend from eco-friendly, biodegradable textiles to high-performance biomedical devices, including tissue scaffolds and targeted drug delivery systems (De Giorgio et al., 2024). Moreover, leveraging computational and experimental methodologies in protein engineering can accelerate the discovery of novel biomaterials, driving advancements in synthetic biology and biomimetic material design.

3 Proposed Pipeline: A Multi-Stage Fine-Tuning Framework

The target dataset, which maps MaSp repeats to their mechanical properties, is extremely small, posing a significant challenge for training generative deep neural networks. To overcome this limitation, we propose a multi-stage fine-tuning approach that enables the model to generate novel MaSp repeats for desired mechanical properties while also predicting mechanical properties from a given MaSp repeat. The training process is structured into three distinct stages, as illustrated in Figure 2. The following sections provide a detailed breakdown of each stage.

3.1 Stage 1: Distillation of ProtGPT2 on Spider Sequences

ProtGPT2 (Ferruz et al., 2022)stands out among generative protein language models (PLMs) due to its specialized training and demonstrated ability to generate biologically plausible protein sequences. Unlike large-scale PLMs such as ESM-2 (Lin et al., 2023) and ProteinMPNN (Dauparas et al., 2022), which focus on structure prediction and sequence design constrained by existing protein scaffolds, ProtGPT2 is designed specifically for de novo protein generation. It has been trained on 10 million protein sequences from the UniRef50 dataset of the UniProt Knowledgebase (UniProtKB) (The UniProt Consortium, 2024) using a self-supervised learning approach. This extensive pre-training enables ProtGPT2 to predict the next amino acid in a sequence, effectively capturing the underlying "grammar" of protein structures.

However, while ProtGPT2 is a powerful model, it is not ideally suited for specialized tasks such as MaSp repeat generation, which requires the model to capture specific sequence patterns unique to spider silk proteins.

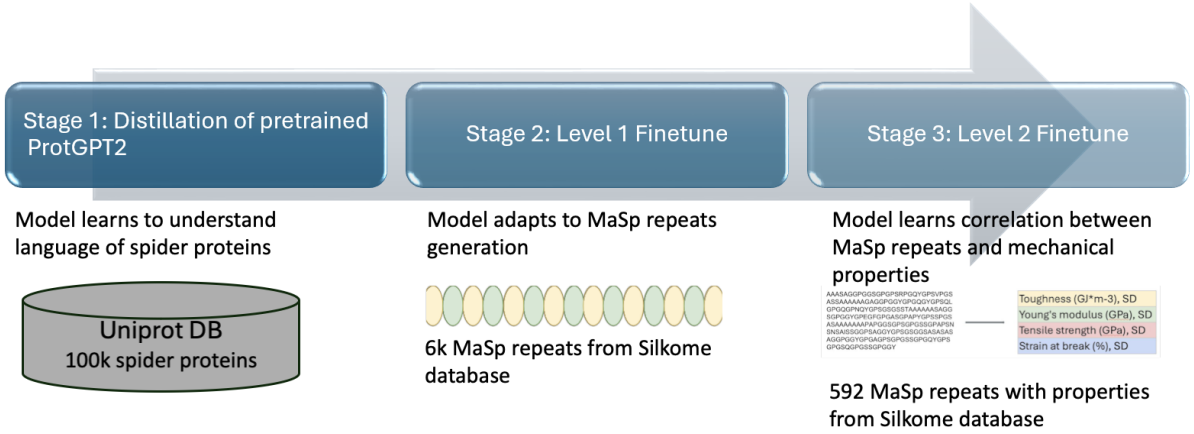

Figure 2: Illustration of the proposed methodology organized into three levels. Stage 1 involves training a distilled ProtGPT2 model using spider protein sequences from UniProtKB (The UniProt Consortium, 2024). In Stage 2, the model is fine-tuned on the repeat regions of MaSp to adapt to their unique patterns. Finally, Stage 3 further fine-tunes the model to capture correlations between MaSp repeats and their mechanical properties.

To address this challenge, we employ a multilevel strategy that leverages a pre-trained protein language model while optimizing it for our specific task. We begin with ProtGPT2. To enhance efficiency and adapt the model for MaSp repeat generation, we apply knowledge distillation (Hinton, 2015), creating a smaller, task-specific variant: SpiderGPT. This distilled model retains the essential knowledge of its teacher while significantly reducing model size and improving inference time, making it more practical for generating novel MaSp repeat sequences tailored to specific mechanical properties.

The distillation follows a teacher-student framework:

- **Teacher:** ProtGPT2, the pre-trained model, serves as the knowledge source.
- **Student:** SpiderGPT, a lightweight model trained to replicate the teacher's outputs.

Dataset: SpiderGPT was trained on a curated dataset of approximately 100k protein sequences obtained from UniProtKB. These sequences were specifically selected based on their taxonomic classification within *Araneae* (spiders) and an annotation score greater than 1. (Details of the procedure are provided in Appendix A.1.) This focused dataset enables the student model to specialize in spider proteins while leveraging the broader knowledge distilled from the teacher model.

Training Process: The distillation parameters include Temperature (T) to be 10, Alpha (α) to be 0.1, embedding dimension (n_{embd}) to be 512, number of transformer layers (n_{layer}) to be 6, number of attention heads (n_{heads}) to be 8. This reduced architecture was designed to strike a balance between computational efficiency and representational capacity, ensuring that the student model effectively captures domain-specific patterns. To facilitate the distillation process, both the teacher and student models were trained on the same tokenized dataset. The teacher model generated probability distributions over the vocabulary for each input sequence, serving as soft labels for the student model.

Our knowledge distillation approach resulted in a substantial reduction in model size and computational demands, without compromising the quality of sequence generation. A comprehensive analysis of the distillation process and its effects is presented in Section 5.3.

3.2 Stage 2: Level 1 Fine-Tuning on MaSp Repeat Regions

The SpiderGPT model, distilled on a spider protein dataset, learns to generate spider silk proteins. In Stage 2, fine-tuning on MaSp repeats refines its ability to recognize and generate MaSp-specific motifs and structures while preserving core protein language knowledge. This methodology demonstrates a sophisticated approach to transfer learning in computational protein design, which bridges the understanding of the fundamental protein language with specialized structural insights.

Dataset: For this study, we utilized the Spider Silkome dataset (Arakawa et al., 2022), a comprehensive resource cataloging silk gene sequences from 1,098 spider species and measuring mechanical, thermal, structural, and hydration properties for 446 species. It highlights the role of MaSp paralogs (MaSp1–MaSp3) in high-performance silk and identifies key amino acid motifs contributing to silk properties. This dataset serves as an open platform for advancing biomaterial research and innovation.

For level 1 fine-tuning, we curated a dataset of MaSp sequences, resulting in 6,000 instances. To focus on the functional repeat regions, we removed the first 150 and last 115 amino acid residues, which correspond to the N-terminal and C-terminal domains, respectively (De Oliveira et al., 2024). The repeat sequences of the same MaSp type were then concatenated for each species to create a structured dataset.

Training Process: The fine-tuning process employed a causal language modeling objective, which fundamentally leverages the transformer architecture's auto-regressive nature. This approach ensures that the model learns to predict the next token in a sequence, creating a robust framework for understanding and generating protein sequences. To address the challenges inherent in working with a limited training dataset, we implemented Low-Rank Adaptation (LoRA), an innovative fine-tuning technique. LoRA introduces trainable low-rank matrices into the model's attention mechanisms, strategically reducing the number of trainable parameters while maintaining high performance. Our specific LoRA configuration included a low-rank dimension of 16, a scaling factor of 32, a dropout rate of 0.1, and a weight decay of 0.01, carefully calibrated to optimize the model's learning capabilities. The model was configured with a maximum sequence length of 512 tokens, providing sufficient flexibility to handle diverse input sequence lengths efficiently. We utilized a tokenizer consistent with the model's architecture, replacing padding tokens with end-of-sequence tokens to ensure seamless compatibility with the causal language modeling objectives. To mitigate potential overfitting and enhance performance on the limited dataset, we incorporated several advanced optimization techniques. Regularization strategies, including dropout and weight decay, were applied to prevent the model from becoming too specialized to the training data. An early stopping mechanism was implemented, monitoring performance on a validation subset and halting training when performance plateaued.

Additional training optimizations included a learning rate of 5×10^{-4} to encourage rapid convergence, a small per-device batch size of 4 to balance memory constraints and training dynamics, and 50 warmup steps to ensure stability during the initial training phases. The entire training process was conducted over 10 epochs, leveraging the Hugging Face Trainer API for streamlined implementation.

A sophisticated checkpoint system was employed to retain the best-performing model configurations, limiting storage requirements to the top two iterations. This approach ensures that we capture the most promising model states while maintaining computational efficiency.

By integrating these advanced techniques, we developed a robust and adaptable fine-tuning methodology that maximizes the potential of our limited dataset, creating a powerful computational tool for protein sequence analysis and generation.

3.3 Stage 3: Level 2 Fine-Tuning for Sequence-Property Associations

In this stage, the model undergoes Level 2 fine-tuning to establish robust associations between MaSp repeat sequences and their corresponding mechanical properties, such as toughness and extensibility. The fine-tuning objective serves a dual purpose: generating MaSp repeats with targeted mechanical properties and predicting mechanical properties from given MaSp repeat sequences.

This bidirectional approach enhances the model's ability to understand and predict sequence-property relationships, significantly advancing its utility in the design of synthetic spider silk proteins.

Dataset: For level 2 fine-tuning, we assembled MaSp sequences along with their corresponding mechanical properties from 293 spider species in the Spider Silkome database. From the MaSp repeat dataset, we filtered sequences that had corresponding mechanical property annotations in the Spider Silkome dataset, yielding 592 MaSp sequences specifically designed for learning sequence-property correlations. Of these, 572 instances were used for training, while 20 were reserved as a test set.

The mechanical property set includes toughness, young's modulus (E), tensile strength, and strain at break, along with their respective standard deviations. These eight values, used as conditioning features during fine-tuning and conditional generation, are represented as follows in equation 1:

$$\mathbf{P} = [\text{Toughness}, E, \text{Strength}, \text{Strain}], \quad \mathbf{SD} = [\sigma_{\text{Toughness}}, \sigma_E, \sigma_{\text{Strength}}, \sigma_{\text{Strain}}] \tag{1}$$

Since these properties span different value ranges, we applied Min-Max normalization to scale all values between 0 and 1 before training. This normalization facilitates faster convergence and simplifies value comparison in data analysis.

To facilitate bidirectional training, we curated the dataset to include a task-specific token that distinguishes between the forward and reverse tasks. This ensures the model effectively learns both sequence generation and property prediction within a unified framework. Model's input-output relationships during the bidirectional training process:

- Forward Task: Given a set of mechanical properties, the model generates a corresponding MaSp repeat sequence. Mentioned in equation 2

$$\text{Task Token}_{\text{EstimateProperty}} + \text{MaSp repeat} \rightarrow \text{Model} \rightarrow \text{Mechanical Properties}_{\text{8D}} \tag{2}$$

- Reverse Task: Given a MaSp repeat sequence, the model predicts its mechanical properties. Mentioned in equation 3

$$\text{Task Token}_{\text{GenerateSequence}} + \text{Mechanical Properties}_{\text{8D}} \rightarrow \text{Model} \rightarrow \text{MaSp repeat} \tag{3}$$

Training Process:

By explicitly incorporating task tokens, we enable the model to generalize across both tasks, reinforcing its ability to capture meaningful sequence-property relationships.

To optimize the model's performance on the limited dataset of 572 records, we employed the LoRA (Low-Rank Adaptation) approach for fine-tuning. We maintained the same configuration used in our level 1 fine-tuning phase, utilizing a low-rank dimension of 16 and a scaling factor of 32. For optimization, the model was trained with a learning rate of 1×10^{-4}, ensuring stable convergence while adapting to the limited data. We also adjusted the batch size to 8, optimizing memory usage while ensuring stable training dynamics. The model was trained for 5 epochs, ensuring adequate learning without overfitting.

To enhance generalization and avoid overfitting, regularization techniques such as dropout and weight decay were applied. A dropout rate of 0.2 and weight decay of 0.02 helped control model complexity. Additionally, an early stopping mechanism was implemented to halt training when performance on a validation subset plateaued, preventing unnecessary overfitting.

The refined model demonstrates exceptional capabilities in deciphering the complex relationships between protein sequences and their mechanical properties. By bridging computational modeling with experimental insights, this approach provides researchers and engineers with a sophisticated computational tool for designing synthetic spider silk proteins. The model's enhanced predictive accuracy represents a significant advancement in biomaterial design, offering unprecedented insights into the intricate connections between molecular structure and material performance. This methodology not only addresses the challenges of working with limited experimental data but also establishes a robust framework for computational protein engineering. The approach showcases the potential of machine learning techniques to extract meaningful insights from complex biological systems, opening new avenues for targeted material design and scientific innovation.

4 Experimentation

4.1 Model Architecture

The SpiderGPT model emerges as a strategically distilled version of the original ProtGPT2, representing a sophisticated approach to computational efficiency in protein sequence modeling. Developed through knowledge distillation, the model maintains core architectural principles while significantly reducing computational complexity. Architecturally, the SpiderGPT is designed with precise specifications that balance performance and efficiency. The model features an embedding dimension of 512, compared to the original model's 1280, and comprises 6 transformer layers against the original 36. This reduction is accompanied by a corresponding decrease in attention heads from 20 to 8, and a substantial reduction in total parameters from 738 million to approximately 50 million. The embedding layer continues to serve as a critical component, implementing a specialized protein sequence representation approach. By learning contextual representations of amino acid sequences, the model captures molecular structural information with a hidden dimension of 2048, enabling nuanced analysis of protein sequence characteristics while maintaining computational efficiency. The multi-head attention mechanism remains a key innovation, allowing parallel processing of sequence information. With 8 attention heads, the model can simultaneously analyze multiple sequence aspects, facilitating complex feature extraction and providing comprehensive insights into protein sequence relationships. This strategic model compression demonstrates a sophisticated approach to machine learning in protein sequence analysis. By preserving the core learning capabilities of the original ProtGPT2 while significantly reducing computational overhead, the SpiderGPT model offers researchers a more accessible and efficient tool for exploring protein sequence complexities. The model represents a critical advancement in computational protein modeling, bridging the gap between comprehensive sequence analysis and practical computational constraints. Its design reflects a nuanced understanding of both machine learning techniques and the intricate nature of protein sequence structures.

4.2 Setup

Our experimental evaluation focused on two critical aspects of the SpiderGPT model's performance in MaSp protein sequence analysis. The comprehensive assessment aimed to validate the model's capabilities in generating biologically meaningful sequences and understanding the intricate relationships between protein sequences and mechanical properties.

The first phase of experimentation concentrated on assessing the quality and biological plausibility of generated protein sequences. We employed a multi-metric approach to rigorously evaluate the generated sequences. This evaluation involved analyzing key molecular characteristics, including sequence composition, amino acid distribution, and structural coherence.

The second experimental phase delved into the complex relationship between MaSp repeat sequences and their mechanical properties. We sought to establish correlations between specific sequence characteristics and mechanical attributes such as toughness and elastic modulus. By systematically mapping sequence features to mechanical properties, we aimed to uncover the underlying molecular determinants that influence the material performance of spider silk proteins.

The experimental design was carefully constructed to provide insight into the predictive capabilities of the model and its potential to advance our understanding of protein sequence-structure-property relationships. By combining computational modeling with rigorous statistical analysis, we aim to bridge the gap between molecular-level sequence information and macroscopic material performance.

4.3 Unconditional Sequence Generation: Self-consistency Assessment

To evaluate the model's ability to generate biologically plausible sequences, we synthesized 200 sequences unconditionally, meaning no specific mechanical property constraints were applied (mentioned in supporting document SD1). Their alignment with natural MaSp sequences was analyzed by comparing key property distributions against 592 MaSp repeats used in level 2 fine-tuning. The results, illustrated in Figure 3,

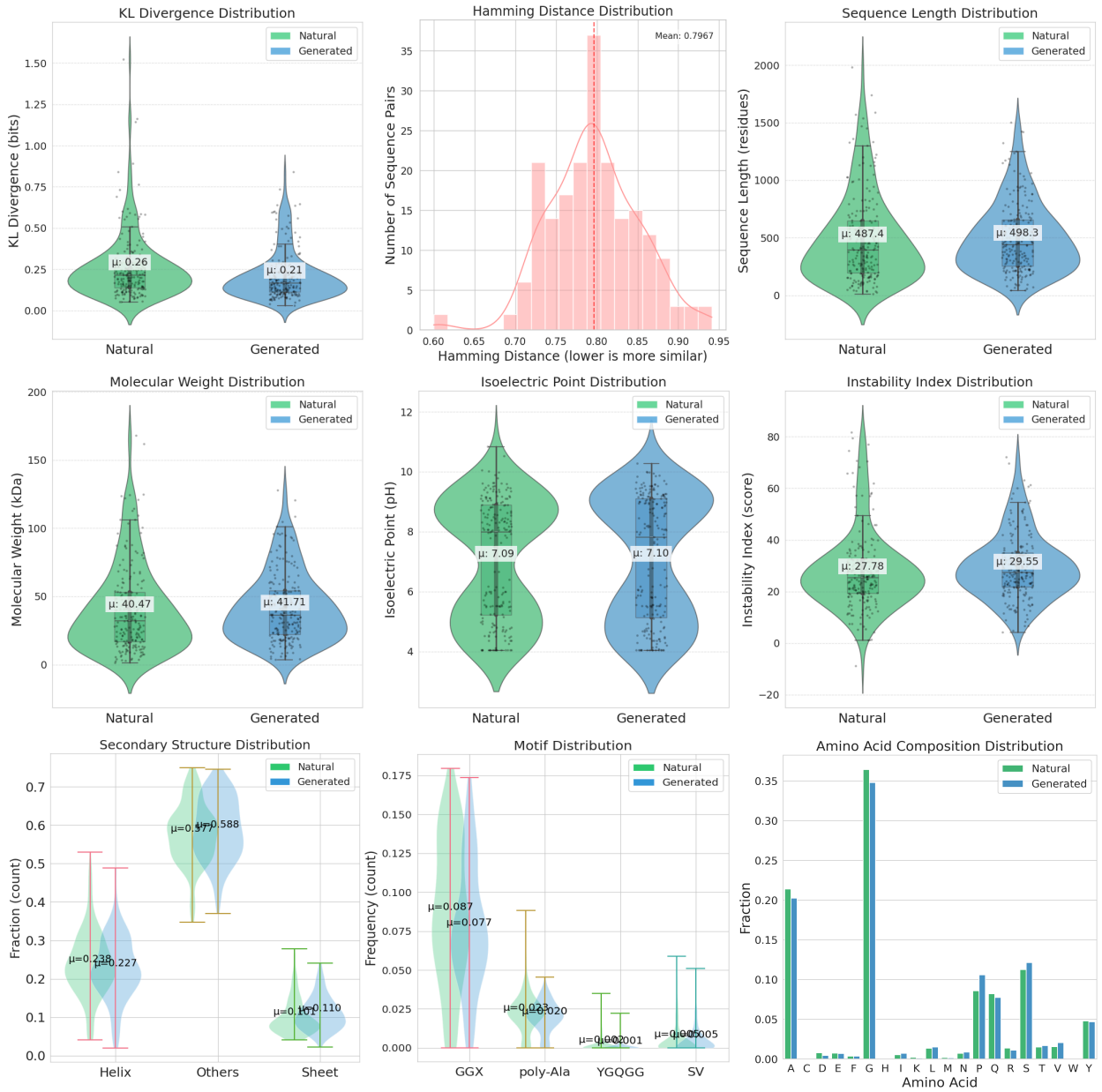

Figure 3: This figure presents a comprehensive comparison of nine key physicochemical and structural properties between naturally occurring (Natural) and computationally generated (Generated) proteins. The analysis includes distributions of KL divergence, Hamming distance, molecular weight, isoelectric point, instability index, sequence length, motif patterns, secondary structure elements, and amino acid composition. The plots demonstrate the degree of similarity between generated proteins and their natural counterparts across multiple biologically relevant parameters, providing insights into the fidelity of the protein design process.

provide a quantitative measure of sequence fidelity, offering insight into how well the model captures the fundamental characteristics of spidroins.

The probability distribution of amino acid occurrences in both natural and generated sequences was examined using KL divergence, which quantifies how one probability distribution deviates from another. Expressed in bits, KL divergence represents the additional information required to encode a sequence based on the

background amino acid distribution of MaSp repeats (Appendix A.3). Additionally, Hamming distance was used to assess diversity in the generated sequences by measuring the number of differing positions between sequences of equal length, providing insight into the variation and uniqueness of the generated sequences relative to natural MaSp repeats.

To further assess the biological plausibility of the generated sequences, we conducted a comprehensive analysis of key physiochemical attributes, including sequence length, molecular weight (Appendix A.2), instability index, and isoelectric point (Guruprasad et al., 1990). These properties were computed using ProtParam (Gasteiger et al., 2005), a widely used tool for protein sequence analysis, implemented via a Python-based tool package to ensure accuracy and reproducibility. The results indicate that the generated sequences exhibit similar distributions across these physiochemical attributes, aligning closely with natural MaSp sequences.

Additionally, secondary structure fractions were analyzed using the ProteinAnalysis module from Biopython (Cock et al., 2009), which follows established secondary structure prediction frameworks, including DSSP (Kabsch & Sander, 1983) and Chou-Fasman propensity scales (Chou & Fasman, 1974). The ProteinAnalysis.secondary_structure_fraction() function was used to compute the fractional composition of α-helices, strands and others/unstructured regions. This analysis enabled a quantitative comparison of secondary structure content between natural and generated sequences, providing insights into structural stability and folding tendencies.

Furthermore, the frequency of key motifs in the generated sequences was examined as part of the secondary structure evaluation. Figure 3 presents the distribution of poly-Ala, GGX, YGQGG, and SV motifs, offering insight into their role in sequential integrity. The formal definition of these motifs is provided in Equation 4. The selected set of motifs are known to impact mechanical properties of the silk fiber (Arakawa et al., 2022; Nakamura et al., 2024b) (refer section 2.1). The motif frequency distribution of natural and generated sequences, illustrated in Figure 3, suggests that the generated sequences reflect a similar distribution as natural sequences.

To further explore the structural properties of the generated sequences, we analyzed the amino acid composition and secondary structure fractions, which are critical determinants of protein function and mechanical behavior. Figures 3 present these comparisons in detail. It's worth mentioning that the standard secondary structure prediction methods are optimized for globular proteins, making them less reliable for structural proteins like MaSp repeats (Jumper et al., 2021).

4.4 MaSp Motif-Property Correlation Analysis

The relationship between MaSp motifs and mechanical properties in spider dragline silk offers critical insights into how specific protein sequence patterns influence silk performance. Recent high-throughput studies have quantified key mechanical metrics -including toughness, Young's modulus, tensile strength, and strain at break - of dragline silk from various spider species, identifying correlations between motifs in the repetitive core region of MaSps and variations in these properties (Arakawa et al., 2022).

Structural motifs such as poly-Ala (associated with β-sheet crystallization) and GGX (contributing to elastic β-turns) play a direct role in defining tensile strength, toughness, and extensibility. In addition to these well-characterized motifs, recent analyses have identified emerging motifs like YGQGG, QQ, GPGXX, AGQG and SV, which may further modulate mechanical behavior (refer section 2.1). The regular expression of each motif, which we used for current analysis, is mentioned in equation 4. Understanding these relationships is instrumental in the design of biomimetic silk materials with tailored mechanical properties.

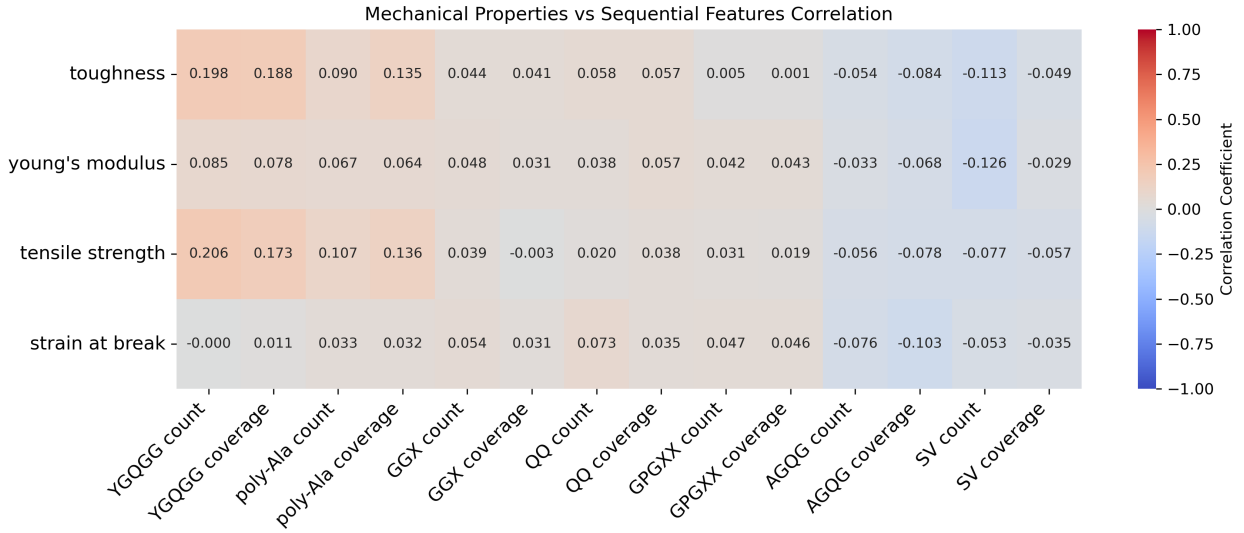

Figure 4: The heatmap illustrates the correlation between sequential features and the mechanical properties of the generated sequences. While some weak correlations are present, the overall low values suggest the need for a more advanced approach to better capture sequence-property relationships.

$$\text{motifs} = \begin{cases} \text{YGQGG:} & \text{YGQGG} \\ \text{poly-Ala:} & A\{3,\} \\ \text{GGX:} & GG[A-Z] \\ \text{QQ:} & QQ \\ \text{GPGXX:} & GPG[A-Z]\{2\} \\ \text{AGQG:} & AGQG \\ \text{SV:} & SV \end{cases} \tag{4}$$

In previous studies, correlation analyses have primarily focused on the distinct motifs present in different MaSp types (MaSp1, MaSp2, and MaSp3) and their specific influences on mechanical properties (Craig et al., 2020; Nakamura et al., 2024a; Kono et al., 2021a; Arakawa et al., 2022). By contrast, the current work considers MaSp more abstractly, providing an overarching view of how various motifs may affect mechanical performance.

For the correlation study, we refined the level 2 fine-tuning dataset by selecting instances with unique sets of mechanical properties. This resulted in a filtered subset of 294 instances. Using this refined dataset, we conducted a reverse generation task, where the model generated 294 MaSp repeats corresponding to these specific property sets. The generated sequences were then used for correlation analysis, allowing a structured evaluation of the relationship between protein motifs and mechanical properties. This study aims to determine whether the generated sequences accurately reflect the sequence-to-property correlations observed in previous research.

Figure 4 presents the correlation heatmap between sequential features and mechanical properties. The sequential features include the count and coverage ($\frac{\text{motif size} \times \text{count}}{\text{sequence length}}$) of the specified motifs. The heatmap indicates that correlation values remain relatively low. While no strong correlations were observed, this visualization is still informative, demonstrating that individual motifs alone cannot fully explain the relationship between sequence variation and mechanical properties. This highlights the intricate and multi-factorial nature of silk mechanics, where the combined influence of multiple motifs and structural contexts collectively shapes the observed mechanical behavior.

In the following we mention the row vise analysis of the heatmap:

Toughness (Energy to Break)

Toughness is the total energy absorbed before failure. Higher toughness in dragline silk is associated with motifs that enhance tensile strength and/or extensibility. The correlation heatmap reveals that YGQGG count (+0.198) and coverage (+0.188) both show moderate positive correlations with toughness. Poly-Ala coverage also correlates positively with toughness, though slightly less strongly (+0.135). The SV count showed negative correlation with toughness (-0.113).

Young's Modulus (Stiffness)

Young's modulus represents the initial stiffness of silk fibers and is influenced by motifs that foster β-sheet crystallinity. For example, poly-Ala count correlates positively with Young's modulus (+0.067). In our analysis, we also observed a positive impact of the count of YGQGG motif on Young's modulus (+0.085), whereas the count of AGQG (-0.033) and SV (-0.126) motifs showed negative correlations.

Tensile Strength (Ultimate Stress)

Tensile strength is the maximum stress that the silk fiber can withstand before failure. The correlation heatmap indicates that many of the same motifs influencing toughness also affect strength. The YGQGG motif count is positively correlated with tensile strength (+0.206) (Arakawa et al., 2022), making it one of the strongest sequence predictors of a stronger fiber in our analysis. Poly-Ala motifs also positively correlate with tensile strength (+0.107).

Strain at Break (Extensibility)

Strain at break is the maximal elongation (expressed as a fraction of its original length) that the silk fiber can achieve before breaking. The correlations highlight AGQG coverage (–0.103) is the most striking correlation overall (and the only moderately strong negative value). Other features in this row hover near zero, indicating minimal correlation. The results also show impact of QQ motif on strain at break. Other features in this row hover near zero, indicating minimal or no correlation.

In summary, the correlation between mechanical properties and individual motifs in the generated sequences follows patterns reported in previous studies (Arakawa et al., 2022; Nakamura et al., 2024b). However, the overall weak correlations suggest that individual motifs alone cannot account for the observed variations in fiber mechanical properties (Arakawa et al., 2022). This study builds upon existing sequence-property relationships and leverages machine learning to better capture the complex interplay between sequential motifs and mechanical properties, enabling the design of novel MaSp sequences.

5 Results and Discussions

The proposed framework performs both the forward task of generating MaSp repeats tailored to specific mechanical properties and the reverse task of predicting mechanical properties from a given MaSp repeat. This section presents the evaluations performed to assess the model's ability to generate biologically plausible sequences and accurately predict mechanical properties.

5.1 Biological plausibility of generated sequences

We perform an in-depth investigation using two datasets: the test set and the BLAST set. The test set consists of sequences sampled from the Spider Silkome dataset, allowing us to assess the model's self-consistency and its ability to reproduce meaningful sequences conditioned on known mechanical properties. The BLAST set, on the other hand, is used to evaluate the novelty and classification of the generated sequences within broader protein sequence databases. By integrating these evaluations, we aim to ensure that the generative model produces biologically plausible sequences that captures key features that are essential for functional spidroin design.

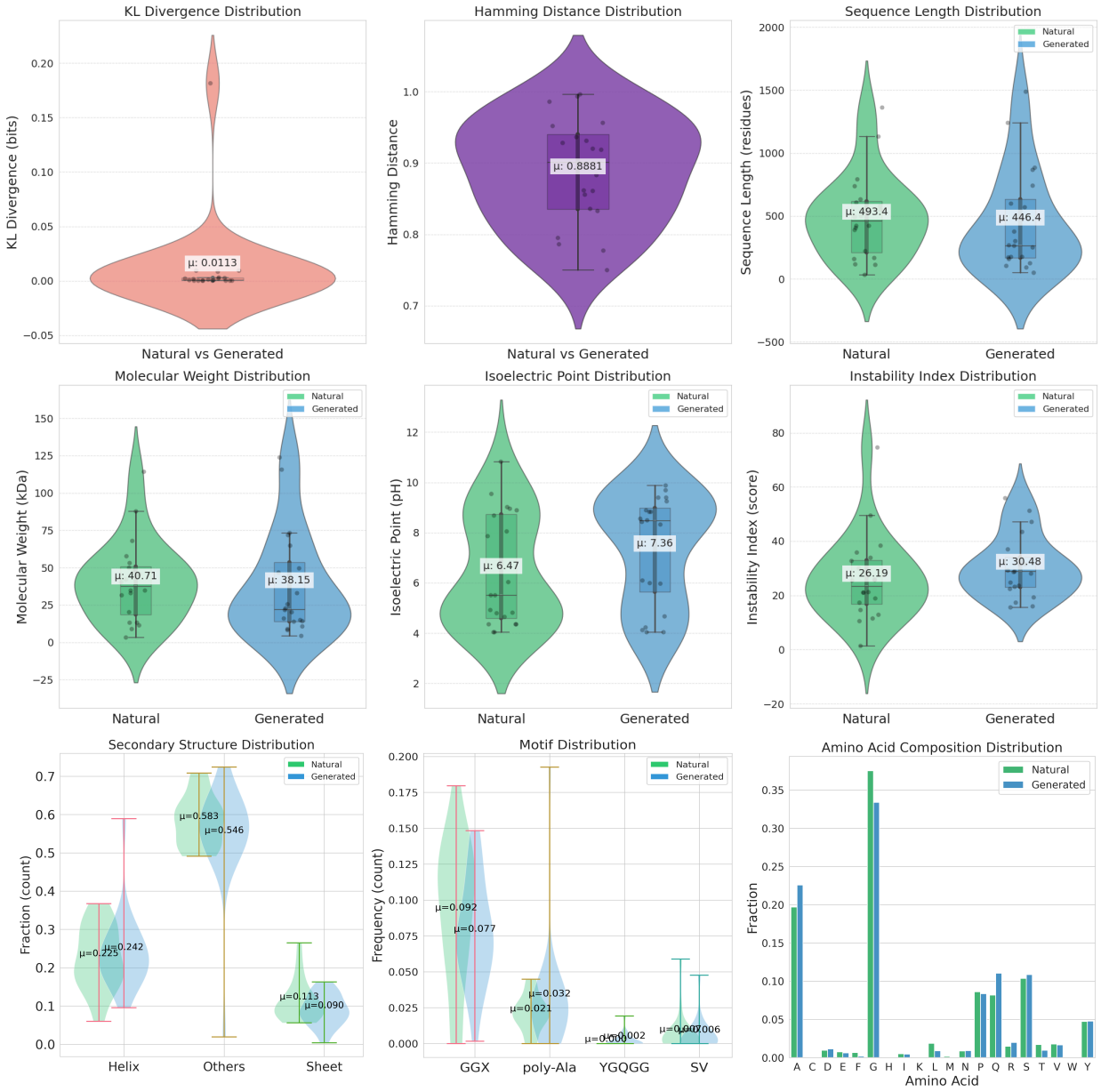

Figure 5: Comparison between original (natural) and generated sequences on the test set in terms of various matrices. (1) Sequence properties: sequence length, molecular weight, instability index, Isoelectric point. (2) Average amino acid frequency distribution grouped by physicochemical properties. The property consistency highlights the validity of the generated sequences in terms of their structural and biochemical features. Furthermore, the consistent alignment demonstrates the model's ability to effectively capture the key characteristics and properties of MaSp.

5.1.1 Conditional Generation Assessment: Evaluation on test dataset

To assess the self-consistency of the trained model, we curated a test set of 20 instances, sampled from the original 592-instance Spider Silkome dataset, which consists of MaSp repeats along with their corresponding mechanical properties. The details of the test set are mentioned in the supporting document SD2. The mechanical properties from the test set were used for conditional sequence generation, allowing for a direct comparison between the generated and natural sequences.

To evaluate the model's ability to generate sequences that correspond to a known set of mechanical properties, we analyzed the test set by comparing the generated MaSp repeats with their original counterparts. The evaluation results, presented in Figure 5, assess the generated sequences at both the sequence level and the structural level, providing insights into how well the model captures the underlying patterns of spider silk proteins.

The sequence-level evaluation assesses key physiochemical attributes, including molecular weight, instability index, and isoelectric point as well as motif occurrences, including poly-Ala, GGX, YGQGG, and SV. Figure 5 illustrates the experimental results. The high agreement between these attributes in the generated and original sequences suggests that the model effectively learns to generate biologically plausible MaSp repeats when conditioned on a specific set of mechanical properties.

Additionally, sequence similarity was analyzed using KL divergence and Hamming distance. KL divergence quantifies the deviation in amino acid probability distributions between original and generated sequences. As shown in Figure 5, KL divergence remains constrained within a low range, indicating that the model successfully captures the natural amino acid distribution of MaSp repeats. In contrast, Hamming distance evaluates sequence diversity at the character level. The high Hamming distance values demonstrate the model's ability to generate diverse sequences, avoiding excessive similarity to training data while preserving essential biological patterns.

The structural evaluation further validates the sequence fidelity by analyzing key secondary structure features— α-helices, β-strands and unstructured regions. The results indicate a strong structural resemblance of the secondary structure composition between generated and natural sequences, confirming that the model preserves critical biological motifs found in naturally occurring MaSp repeats (Figure 5).

The ability of the generative model to replicate both physiochemical and structural characteristics underscores its potential as a powerful tool for biomaterial design. By tailoring sequences to meet specific functional and mechanical property requirements, this approach offers a promising avenue for advancing the development of engineered biomaterials with enhanced properties.

5.1.2 Novelty Assessment: BLAST Evaluation

To evaluate the novelty of the generated MaSp repeats, we performed BLAST analyses against a broader Spider Silkome spidroin repeat database. The search space consists of 11K naturally occurring spidroin sequences. It includes seven primary types of spider silk proteins major ampullate spidroins (MaSp), flagelliform spidroins (Flag), minor ampullate spidroins (MiSp), aggregate spidroins (AgSp), pyriform spidroins (PySp), aciniform spidroins (AcSp), and cylindriform/tubuliform spidroins (CySp) (Yarger et al., 2018a). A subset of seven generated sequences was compared with the two most closely related natural sequences, designated BLAST1 (top match) and BLAST2 (second top match) based on BLAST results. The selection was based on high query coverage, percent identity, and sequence length similarity to the generated sequences. This dataset is referred to as the "BLAST set", with details of all sequences and their sources provided in the supporting document SD3. All selected BLAST matches had an expect value (E) of $E \leq e^{-10}$, which ensured statistically significant similarities. Furthermore, we selected matches spanning different MaSp subtypes (MaSp1, MaSp2, MaSp3) to provide a comprehensive comparative analysis. The expect value (E) quantifies the probability of obtaining a match by chance in a given database size, decreasing exponentially as the alignment score (S) increases.

To assess uniqueness, we follow established criteria where sequences with similarity values below 50–60% are considered novel (Quan et al., 2023). The generated sequences meet this threshold when compared to broader protein databases, demonstrating the model's ability to design distinct sequences that either do not exist in nature or have not yet been observed. Additionally, BLAST search results consistently classify the most similar existing sequences as MaSp, aligning with the intended spidroin type. This indicates that the generative modeling approach effectively captures the defining characteristics of specific spider silk proteins.

Figure 6 presents the comparison between the generated MaSp repeats and the closest matches, BLAST1 and BLAST2. The observed consistency highlights the ability of the model to accurately reproduce essential sequence features of spider silk proteins. The generated sequences maintain comparable molecular weight

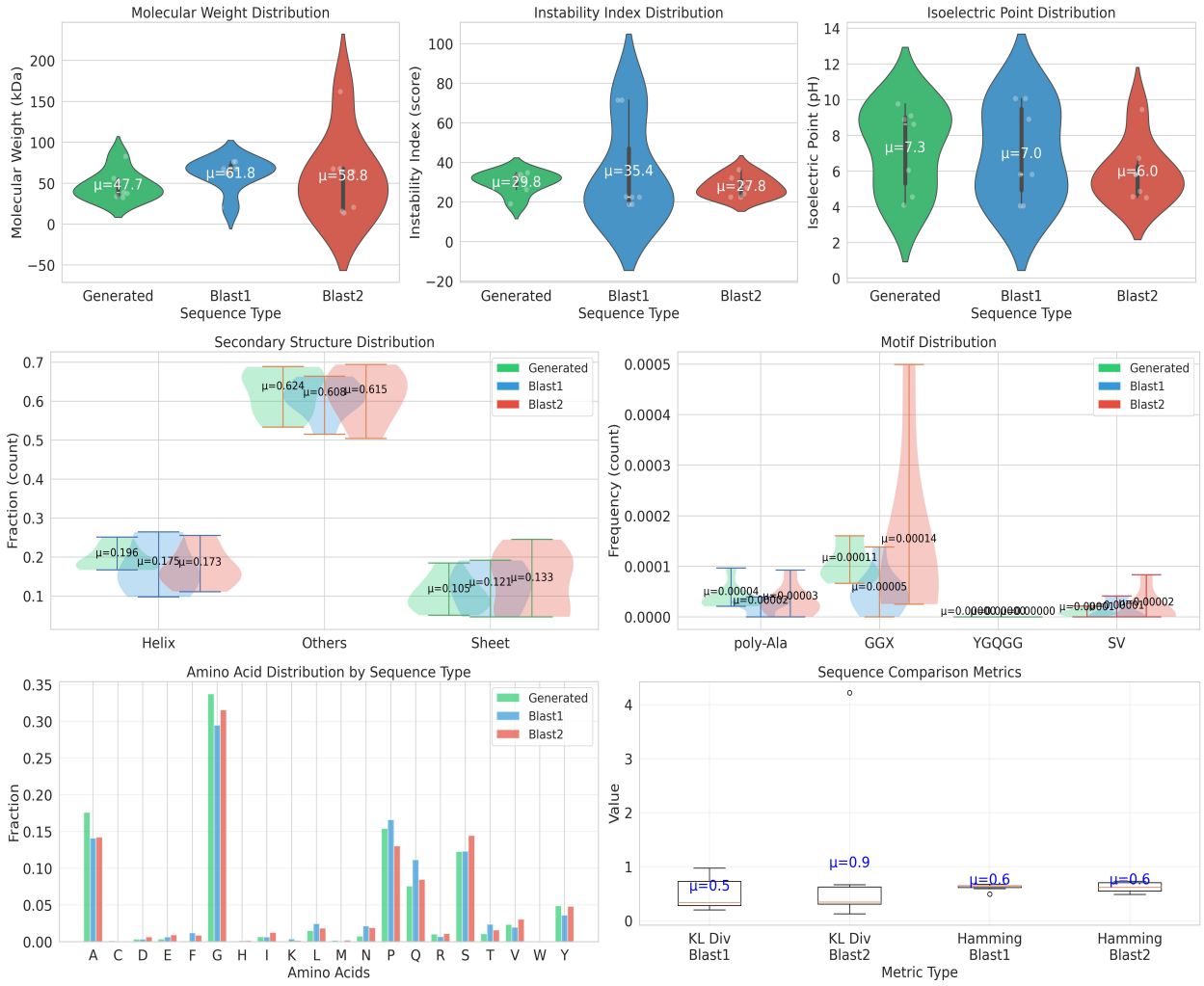

Figure 6: Comparison of sequence and structural properties of our generated protein sequences with BLAST1 and BLAST2 sequences. The analysis includes distributions of instability index, isoelectric points, molecular weight, motif prevalence, secondary structure fractions, amino acid composition, and sequence similarity metrics. This highlights the consistency of the generated sequences with natural proteins and their alignment across various properties

ranges, predicted secondary structure composition, and physiochemical properties, all of which are crucial for mimicking the functional behavior of natural spider silks. Furthermore, the model effectively balances critical sequence attributes, such as the instability index and isoelectric point, reinforcing the biological plausibility of the generated sequences.

Although Figure 6 confirms that the generated sequences share similarities in fundamental composition, structural, and sequence pattern with naturally occurring MaSps, reinforcing their biological relevance, variations in metric values may arise due to the selection of reference sequences for the BLAST comparison. This highlights the sensitivity of the evaluation process to the chosen dataset. Despite these variations, the model consistently captures the essential properties of spider silk proteins, further emphasizing its potential for biomaterial design and the generation of novel sequences with customized functionalities.

Table 1: Comparison of Generated and Reference Properties

Metric	Value
Pearson Correlation (r)	0.3911
Mean Absolute Error (MAE)	0.0587
Root Mean Square Error (RMSE)	0.0716
Cosine Similarity	0.9465

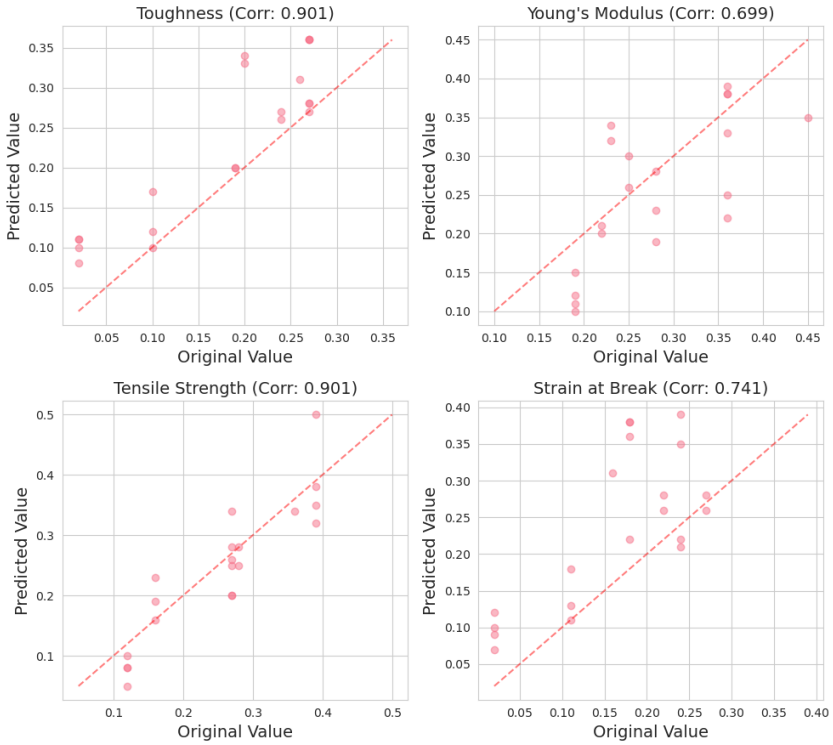

Figure 7: Comparison of predicted and original mechanical properties for MaSp repeats in the test set. Scatter plots illustrate the relationship between original and predicted values for toughness, Young's modulus, tensile strength, and strain at break. Property values are normalized between 0 and 1 with each plot displaying the corresponding Pearson correlation coefficient (r), which ranges from -1 to +1. The red dashed diagonal line represents the ideal 1:1 correlation, indicating perfect agreement between predicted and actual values. The observed trends suggest a significant correlation between generated and original mechanical properties, demonstrating the effectiveness of the prediction model.

5.2 Evaluation of Mechanical Property Prediction

To evaluate models' capability to estimate mechanical properties for a given MaSp repeat, we use the test set of 20 instances which we mentioned above. For each sequence in the test set, we run inference for the reverse task of the model and generate mechanical properties. The similarity between the generated and reference properties was assessed using four key statistical metrics: Pearson's correlation coefficient (r), mean absolute error (MAE), root mean square error (RMSE), and cosine similarity, as detailed in Table 1. These metrics provide quantitative insight into the degree to which the generated sequences replicate the structural and mechanical trends observed in the reference dataset.

The Pearson Correlation Coefficient (r) was found to be 0.3911, indicating a moderate positive correlation between the generated and reference properties. Although this suggests that the generated sequences capture some variations present in the reference data, the relationship is not strictly linear. The Mean Absolute Error

(MAE) of 0.0587 represents the average absolute difference between generated and reference values. The relatively low MAE suggests that the generated properties closely approximate the reference values, with minor deviations. The Root Mean Square Error (RMSE) of 0.0716 is slightly higher than the MAE, reflecting its sensitivity to larger deviations. This indicates that while most generated values align well with reference values, occasional discrepancies exist but remain within an acceptable range. Finally, the Cosine Similarity of 0.9465 demonstrates a high degree of directional similarity between the generated and reference properties. This suggests that while absolute values may not always align perfectly, the generated sequences effectively maintain overall structural and mechanical trends.

A property-level analysis is presented in Figure 7, further evaluating the accuracy of the estimated mechanical properties. Figure 7 (a) illustrates the correlation between original and generated values across all four properties, reaffirming the model's capability to estimate mechanical properties with reasonable accuracy. Overall, these findings highlight the generative model's effectiveness in capturing key mechanical characteristics, making it a reliable tool for designing bio-inspired materials.

5.3 Ablation Studies

In this section, we perform ablation experiments over a number of facets of the proposed methodology in order to better understand their relative importance.

5.3.1 Without Distillation

The first stage of the architecture pipeline involves distilling the ProtGPT2 model into a smaller SpiderGPT model. In this section, we explain the need for this technique and compare the ProtGPT2 and SpiderGPT models.

Pretrained protein language models (PLMs) like ProtGPT2 are large and designed for general protein generation tasks. While these models excel in broad protein generation, the current task is more specific. It involves generating MaSp repeats from a smaller dataset, where the sequences follow a distinct and specialized pattern. Given the focused nature of the task, a lighter model is more appropriate. It not only improves inference speed but also simplifies the overall architecture, making it more efficient for the specific requirements of this task.

Table 2 provides an architectural comparison between ProtGPT2 and SpiderGPT, highlighting that the SpiderGPT model is significantly more compact than its baseline counterpart, ProtGPT2. To assess their performance differences, we generated 200 protein sequences using each model. The reduced size of SpiderGPT not only simplifies the pipeline complexity but also accelerates inference, with the distilled SpiderGPT achieving a remarkable six-fold increase in inference speed compared to ProtGPT2. Evaluations conducted on these 200 sequences reveal that this boost in efficiency incurs only minimal performance trade-offs. Specifically, the student model, SpiderGPT, sustains perplexity levels comparable to those of the teacher model, ProtGPT2. For a more detailed visual comparison of the two models' performance, refer to Figure 8, which illustrates the outcomes of both the ProtGPT2 teacher model and the SpiderGPT model.

Attribute	**ProtGPT2**	**SpiderGPT**
Embedding Dim (n_{embd})	1280	512
Layers (n_{layer})	36	6
Hidden Dim	5120	2048
Attention Heads	20	8
Total Parameters	738M	50M

Table 2: Comparison of Teacher and Student Model Architectures

5.3.2 Without first level fine tuning

ProtGPT2 is a decoder-only transformer model that has been pre-trained on the protein space using the UniRef50 database (version 2021_04) (The UniProt Consortium, 2024), which contains 10 million protein

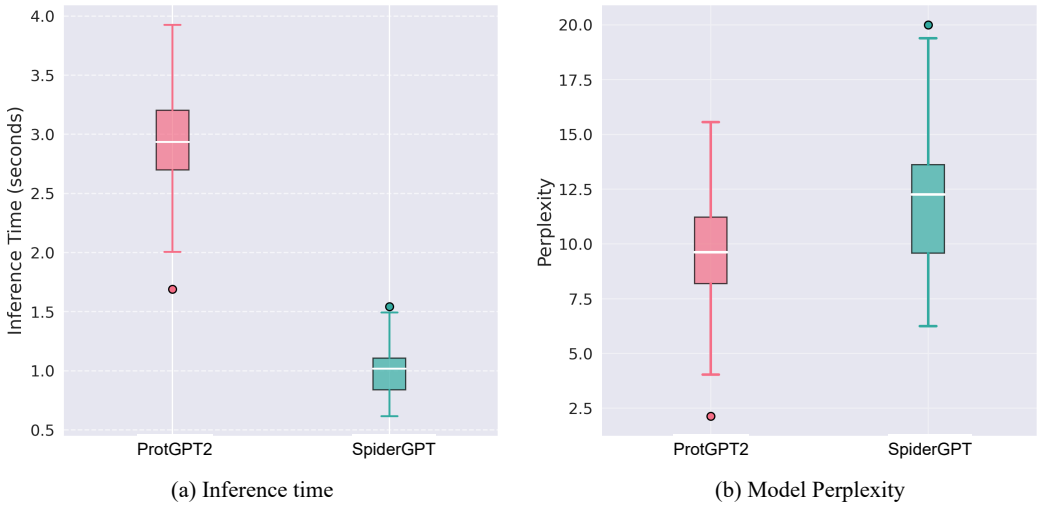

Figure 8: Comparison of performance of teacher (ProtGPT2) and student (SpiderGPT) models over generation of 200 novel protein sequences.

instances. However, the Sequence-to-Mechanical Property Correlation dataset, curated from the Spider Silkome database, contains only 592 instances. This small dataset is insufficient for training the model to generate MaSp repeat sequences while simultaneously learning the correlation between sequences and mechanical properties. To address this limitation, we split the training into two stages.

In the first stage, we focus on training the model to generate MaSp repeats. In the second stage, we train the model to learn the correlation between sequences and their corresponding mechanical properties. This staged approach allows the model to first specialize in generating MaSp repeats before learning to predict mechanical properties from these sequences.

In this section, we evaluate the impact of the first level of fine-tuning, where the model is trained on 6,000 instances of MaSp repeats. To emphasize the significance of this stage, we also present the model's performance when fine-tuned directly on the 592 instances from the Spider Silkome dataset, which contains MaSp repeats along with their corresponding mechanical properties. This comparison highlights the importance of multi-level fine-tuning in achieving optimal results.

For consistency, we used the same training setup across both scenarios. The SpiderGPT model was fine-tuned with the following LoRA settings: low-rank dimension $(r) = 16$, scaling factor $(\alpha) = 32$, dropout $= 0.1$, and weight decay $= 0.01$. Training was conducted for 10 epochs with a learning rate of 5×10^{-4} and a per-device batch size of 4. Additionally, 50 warmup steps were used, and weight decay was applied at a rate of 0.01. Below we evaluate the impact of this ablation study on both forward (sequence generation) and reverse (property estimation) tasks.

Impact on Sequence Generation Quality

As discussed in previous sections, the primary objective of the model in the forward task is to generate MaSp repeats tailored to a given set of mechanical properties. In this section, we evaluate the impact of omitting the initial fine-tuning phase (level 1) on sequence generation quality. To this end, we generated 100 sequences using two models: one trained with both levels of fine-tuning and another trained without level 1 fine-tuning.

The omission of level 1 fine-tuning resulted in significant deviations in sequence generation, with the model frequently producing spider silk sequences containing non-repeat regions—an unintended outcome. To detect these deviations, a sliding window algorithm was employed to analyze the density of poly-alanine and glycine-rich regions, which are typically abundant in valid MaSp repeat sequences. In contrast, sequences containing

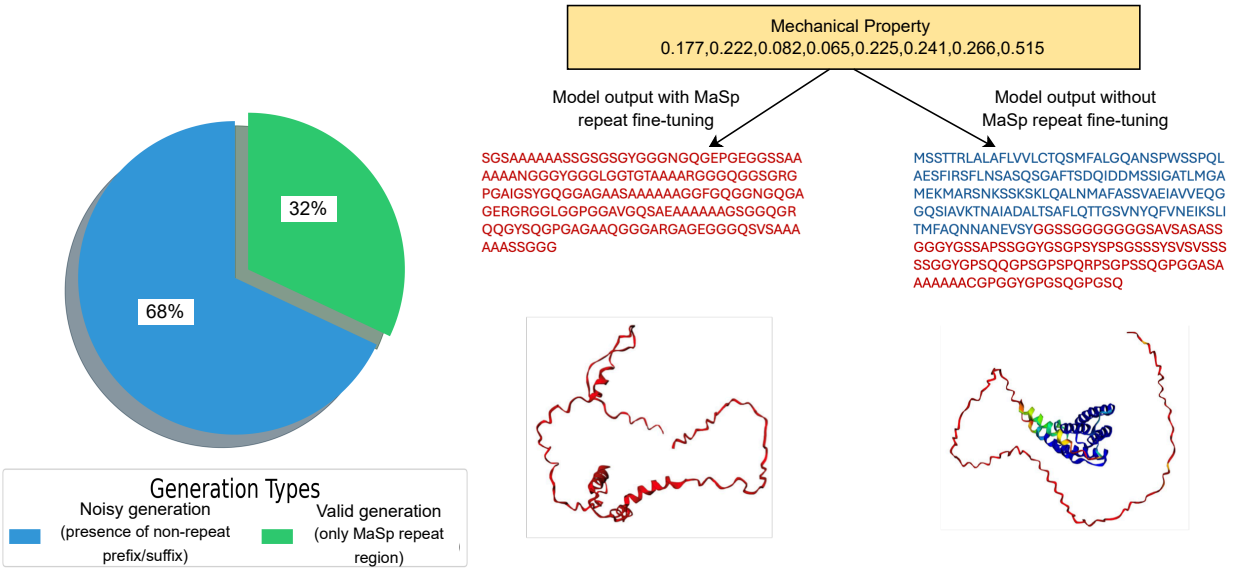

Figure 9: Impact of level 1 fine-tuning on sequence generation quality. Without MaSp fine-tuning, the model generates non-repeat prefix/suffix (68% occurrence), leading to structural deviations. The molecular structure predicted by Omegafold reveals helical formations, indicating the presence of terminal domains. In contrast, model achieves 100% valid sequence generation if we include MaSp repeat fine-tuning step in the methodology pipeline.

non-repeat regions exhibited almost null densities of these characteristic motifs. Among the 100 generated sequences, 68% included non-repeat prefixes or suffixes, appearing at the beginning or end of the sequences, respectively (Figure 9). This finding suggests that, without the initial fine-tuning phase, the pretrained model retains its original behavior, failing to specialize in generating only the repetitive core region of MaSp sequences.

To further investigate these discrepancies, we compare the structural predictions of sequences generated by a model trained with only one fine-tuning stage against those produced by a model trained with both levels of fine-tuning. Figure 9 illustrates the molecular configurations of the generated sequences, revealing structural differences arising from the absence of MaSp-specific fine-tuning. Structural predictions were obtained using Omegafold (Wu et al., 2022), providing insights into the conformational tendencies of the sequences. Notably, the generated sequences without level 1 fine-tuning exhibited helical structures, which are typically associated with non repeat regions like terminal domains (NTD/CTD). These unwanted elements further highlight the necessity of level 1 fine-tuning in ensuring the correct generation of MaSp repeats.

Impact on Property Prediction

Here, we analyze the effect of skipping level 1 fine-tuning on the reverse task of mechanical property estimation. To evaluate performance, we used the same test set of 20 instances previously employed to assess the proposed pipeline. Specifically, we provided the sequence generation prompts from the test set and compared the predicted properties from both model variants—one with level 1 fine-tuning and one without.

Table 3: Comparison of Generated and Reference Properties

Metric	**Value**
Pearson Correlation (r)	0.1210
Mean Absolute Error (MAE)	0.2512
Root Mean Square Error (RMSE)	0.3024
Cosine Similarity	0.6543

The results, presented in Table 3, show significant changes in the correlation metrics. The Pearson Correlation (r) dropped to 0.1210, indicating a poor linear relationship between the generated and reference mechanical properties. This suggests that without level 1 fine-tuning, the model struggles to capture the variations in the properties correctly. Additionally, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) have increased, indicating larger discrepancies between the generated properties and the reference values. This highlights that the model is less accurate in predicting mechanical properties when fine-tuning is omitted. Furthermore, the Cosine Similarity decreased to 0.6543, suggesting a diminished directional similarity between the two sets of properties. This reduction reflects the model's inability to maintain the overall trends and patterns of the reference properties without the fine-tuning step. These observations underscore the importance of level 1 fine-tuning in ensuring that the model performs optimally in predicting mechanical properties.

6 Conclusion and Future Work

This research presents a novel generative model based on GPT architecture, meticulously fine-tuned to leverage a dataset comprising 6,000 repeat regions derived from spider silk proteins. This initial training was subsequently enhanced through a secondary fine-tuning phase utilizing approximately 600 Major Ampullate Spidroin (MaSp) sequences, each annotated with well-characterized mechanical properties. Our innovative dual-level fine-tuning strategy has proven effective in producing synthetic sequences that incorporate essential structural motifs, such as the glycine-rich GGX repeats and poly-Ala stretches, which are fundamental to the remarkable extensibility and tensile strength exhibited by natural spider silk.

Detailed structural analyses of the generated sequences reveal a strong alignment with the anticipated secondary structure composition, such as α-helical and β-strand conformations, which are critical for replicating the functional attributes of spider silk. Furthermore, comparisons between the mechanical properties predicted by the model and established reference data underscore the reliability and promise of this approach for designing spidroins with tailored mechanical properties. By enabling the precise generation of sequences tailored to specific performance criteria, this work lays a robust foundation for the sustainable production of synthetic spider silk materials. The platform that we have developed is both scalable and highly customizable, offering significant potential to transform material science and synthetic biology. Its implications extend beyond spider silk, paving the way for the creation of next-generation bioinspired materials with applications in diverse fields such as tissue engineering, textiles, and environmentally friendly composites.

Future work will encompass a comprehensive experimental validation process for the generated sequences, which will involve synthesizing these sequences and subjecting them to rigorous mechanical testing. This step is crucial to verify that the predicted properties align with the actual performance characteristics observed under controlled conditions. The synthesis process will aim to accurately replicate the molecular structures proposed by the generative model, while the mechanical testing will evaluate key parameters such as tensile strength, extensibility, and toughness to ensure that the sequences meet the anticipated functional benchmarks.

In parallel, another key focus of future research will be the enhancement of the generative model itself. This will involve integrating significantly larger and more diverse datasets to improve the predictive accuracy and generalizability of the model. By scaling up the data input, the model can better capture the complex relationships between sequence composition and resultant material properties, thereby refining its ability to generate optimized designs. The current methodology, while effective, was specifically designed and optimized for a relatively small dataset. Moving forward, we plan to broaden the scope of this approach by expanding the spidroin-to-mechanical-properties dataset. This expansion will include a wider range of spidroin variants and their corresponding mechanical attributes, enabling a more robust mapping of sequence-to-function correlations and supporting the development of advanced materials with tailored performance characteristics.

Impact Statement

Our research on computational protein design holds transformative potential for developing advanced biomaterials with applications in medicine, sustainability, and biotechnology. By bridging machine learning with protein engineering, we offer a powerful computational framework that can accelerate the design of novel materials with precise mechanical properties. The methodology not only advances scientific understanding but also presents opportunities for addressing critical challenges in sustainable material development and medical innovation, while maintaining a responsible approach to technological advancement.

References

Namrata Anand, Raphael Eguchi, Irimpan I Mathews, Carla P Perez, Alexander Derry, Russ B Altman, and Po-Ssu Huang. Protein sequence design with a learned potential. *Nature communications*, 13(1):746, 2022.

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.

Nathaniel Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, Frank DiMaio, Steven Munck, Savvas Savvides, and David Baker. Improving de novo protein binder design with deep learning. *Nature Communications*, 14, 05 2023. doi: 10.1038/s41467-023-38328-5.

Katherine Bourzac. Spiders: Web of intrigue. *Nature*, 519:S4–6, 03 2015. doi: 10.1038/519S4a.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Daniel W. A. Buchan and David T. Jones. The psipred protein analysis workbench: 20 years on. *Nucleic Acids Research*, 47(W1):W402–W407, 2019. doi: 10.1093/nar/gkz297. URL `https://doi.org/10.1093/nar/gkz297`.

Peter Y Chou and Gerald D Fasman. Prediction of protein conformation. *Biochemistry*, 13(2):222–245, 1974.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

Hamish Craig, Dakota Piorkowski, Shinichi Nakagawa, Michael Kasumovic, and Sean Blamires. Meta-analysis reveals materiomic relationships in major ampullate silk across the spider phylogeny. *Journal of The Royal Society Interface*, 17, 09 2020. doi: 10.1098/rsif.2020.0471.

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.

Giuseppe De Giorgio, Biagio Matera, Davide Vurro, Edoardo Manfredi, Vardan Galstyan, Giuseppe Tarabella, Benedetta Ghezzi, and Pasquale D'Angelo. Silk fibroin materials: Biomedical applications and perspectives. *Bioengineering*, 11(2):167, 2024.

Danilo Hirabae De Oliveira, Vasantha Gowda, Tobias Sparrman, Linnea Gustafsson, Rodrigo Sanches Pires, Christian Riekel, Andreas Barth, Christofer Lendel, and My Hedhammar. Structural conversion of the spidroin c-terminal domain during assembly of spider silk fibers. *Nature Communications*, 15(1):4670, 2024.

Vincenzo Fazio, Nicola Maria Pugno, Orazio Giustolisi, and Giuseppe Puglisi. Hierarchical physically based machine learning in material science: the case study of spider silk. *arXiv preprint arXiv:2307.12945*, 2023. URL `https://arxiv.org/abs/2307.12945`.

Noelia Ferruz and Birte Höcker. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532, 2022.

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

Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, S'everine Duvaud, Marc R Wilkins, Ron D Appel, and Amos Bairoch. *Protein identification and analysis tools on the ExPASy server*. Springer, 2005.

John M Gosline, Mark W Denny, and M Edwin DeMont. Spider silk as rubber. *Nature*, 309(5968):551–552, 1984.

John M Gosline, PA Guerette, CS Ortlepp, and KN Savage. The mechanical design of spider silks: from fibroin sequence to mechanical function. *Journal of Experimental Biology*, 202(23):3295–3303, 1999.

Paul A Guerette, David G Ginzinger, Bernhard HF Weber, and John M Gosline. Silk properties determined by gland-specific expression of a spider fibroin gene family. *Science*, 272(5258):112–115, 1996.

Kunchur Guruprasad, BV Bhasker Reddy, and Madhusudan W Pandit. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection*, 4(2):155–161, 1990.

Cheryl Y Hayashi and Randolph V Lewis. Evidence from flagelliform silk cdna for the structural basis of elasticity and modular nature of spider silks. *Journal of molecular biology*, 275(5):773–784, 1998.

David H Hijirida, Kinh Gian Do, Carl Michal, Shan Wong, David Zax, and Lynn W Jelinski. 13c nmr of nephila clavipes major ampullate silk gland. *Biophysical journal*, 71(6):3442–3447, 1996.

Michael B Hinman and Randolph V Lewis. Isolation of a clone encoding a second dragline silk fibroin. nephila clavipes dragline silk is a two-protein fiber. *Journal of Biological Chemistry*, 267(27):19320–19324, 1992.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

Sinan Keten and Markus J Buehler. Nanostructure and molecular mechanics of spider dragline silk protein assemblies. *Journal of the Royal Society Interface*, 7(53):1709–1721, 2010. doi: 10.1098/rsif.2010.0149.

Hamed Khakzad, Ilia Igashov, Arne Schneuing, Casper Goverde, Michael Bronstein, and Bruno Correia. A new age in protein design empowered by deep learning. *Cell Systems*, 14(11):925–939, 2023.

Yoonjung Kim, Taeyoung Yoon, Woo B Park, and Sungsoo Na. Predicting mechanical properties of silk from its amino acid sequences via machine learning. *Journal of the Mechanical Behavior of Biomedical Materials*, 140:105739, 2023.

Andreas Koeppel and Chris Holland. Progress and trends in artificial silk spinning: A systematic review. *ACS Biomaterials Science & Engineering*, 3, 01 2017. doi: 10.1021/acsbiomaterials.6b00669.

Nobuaki Kono, Hiroyuki Nakamura, Masaru Mori, Yuki Yoshida, Rintaro Ohtoshi, Ali D Malay, Daniel A Pedrazzoli Moran, Masaru Tomita, Keiji Numata, and Kazuharu Arakawa. Multicomponent nature underlies the extraordinary mechanical properties of spider dragline silk. *Proceedings of the National Academy of Sciences*, 118(31):e2107065118, 2021a.

Nobuaki Kono, Rintaro Ohtoshi, Ali D Malay, Masaru Mori, Hiroyasu Masunaga, Yuki Yoshida, Hiroyuki Nakamura, Keiji Numata, and Kazuharu Arakawa. Darwin's bark spider shares a spidroin repertoire with caerostris extrusa but achieves extraordinary silk toughness through gene expression. *Open biology*, 11 (12):210242, 2021b.

Albert L. Lehninger, David L. Nelson, and Michael M. Cox. *Principles of Biochemistry.* W.H. Freeman and Company, 5th edition, 2008.

Randolph V Lewis. Spider silk: the unraveling of a mystery. *Accounts of chemical research*, 25(9):392–398, 1992.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *Science*, 378(6617):eadf2037, 2022.

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.

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.

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

Ali D Malay, Kazuharu Arakawa, and Keiji Numata. Analysis of repetitive amino acid motifs reveals the essential features of spider dragline silk proteins. *PLoS One*, 12(8):e0183397, 2017. doi: 10.1371/journal.pone.0183397.

Liam J McGuffin, Kevin Bryson, and David T Jones. The psipred protein structure prediction server. *Bioinformatics*, 16(4):404–405, 2000.

Hiroyuki Nakamura, Yusuke Ito, Ryota Sato, Hongfang Chi, Chikako Sato, Yasuha Watanabe, and Kazuharu Arakawa. Correlating mechanical properties and sequence motifs in artificial spider silk by targeted motif substitution. *ACS biomaterials science & engineering*, 10, 11 2024a. doi: 10.1021/acsbiomaterials.4c01389.

Hiroyuki Nakamura, Yusuke Ito, Ryota Sato, Hongfang Chi, Chikako Sato, Yasuha Watanabe, and Kazuharu Arakawa. Correlating mechanical properties and sequence motifs in artificial spider silk by targeted motif substitution. *ACS Biomaterials Science & Engineering*, 10(12):7394–7403, 2024b.

Bo Ni, David L Kaplan, and Markus J Buehler. Forcegen: End-to-end de novo protein generation based on nonlinear mechanical unfolding responses using a protein language diffusion model. *arXiv preprint arXiv:2310.10605*, 2023. URL `https://arxiv.org/abs/2310.10605`.

Dan Ofer, Nadav Brandes, and Michal Linial. The language of proteins: Nlp, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758, 2021.

David B Peakall. Synthesis of silk, mechanism and location. *American Zoologist*, 9(1):71–79, 1969.

Lijun Quan, Tingfang Wu, and Qiang Lyu. Computational de novo protein design: from secondary to primary, then toward tertiary structures. *Chem*, 9(7):1625–1627, 2023.

Benjamin Schmuck, Gabriele Greco, Tomas Bohn Pessatti, Sumalata Sonavane, Viktoria Langwallner, Tina Arndt, and Anna Rising. Strategies for making high-performance artificial spider silk fibers. *Advanced Functional Materials*, 34(35):2305040, 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`.

Olena Tokareva, Valquíria A Michalczechen-Lacerda, Elíbio L Rech, and David L Kaplan. Recombinant dna production of spider silk proteins. *Microbial biotechnology*, 6(6):651–663, 2013.

JD Van Beek, S Hess, F Vollrath, and BH124902 Meier. The molecular structure of spider dragline silk: folding and orientation of the protein backbone. *Proceedings of the National Academy of Sciences*, 99(16): 10266–10271, 2002.

Fritz Vollrath and David P Knight. Liquid crystalline spinning of spider silk. *Nature*, 410(6828):541–548, 2001.

Fritz Vollrath and DP Knight. Structure and function of the silk production pathway in the spider nephila edulis. *International Journal of Biological Macromolecules*, 24(2-3):243–249, 1999.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022.

Jeffery L Yarger, Brian R Cherry, and Arjan van der Vaart. Structure and dynamics of silk proteins. *Nature Reviews Materials*, 3(3):18008, 2018a.

Jeffery L Yarger, Brian R Cherry, and Arjan Van Der Vaart. Uncovering the structure–function relationship in spider silk. *Nature Reviews Materials*, 3(3):1–11, 2018b.

A Appendix

A.1 Distillation Data Acquisition

Protein sequences were retrieved from the UniProt Knowledgebase (UniProtKB) using the UniProt REST API. To obtain high-quality sequences, we filtered entries based on taxonomy ID 6893 and annotation scores of 2, 3, 4, or 5. The following query was used to download the dataset in FASTA format:

```
https://rest.uniprot.org/uniprotkb/stream?compressed=true&format=fasta&
query=(taxonomy_id:6893) AND (annotation_score:2 OR annotation_score:3 OR
    annotation_score:4 OR annotation_score:5)
```

The dataset comprises protein sequences specific to the selected taxonomy, ensuring relevance for further computational analyses.

A.2 Molecular Weight Calculation

Molecular weight determination is a critical parameter in protein characterization, providing essential insights into protein structure and function. In our research, we employed a standardized approach to molecular weight calculation using established biochemical methods.

The molecular weight of protein sequences was calculated using the standard amino acid molecular weights, accounting for the molecular mass of each amino acid and subtracting the mass of water molecules released during peptide bond formation. Specifically, we utilized the average molecular weights of amino acids as defined by the International Union of Pure and Applied Chemistry (IUPAC) standard (Lehninger et al., 2008).

The calculation followed the formula:

$$\text{Molecular Weight} = \sum(\text{Amino Acid Molecular Weight}) - (n-1) * 18.015 \tag{5}$$

Where:

- $\sum$ represents the sum of individual amino acid molecular weights
- n represents the number of amino acids in the sequence
- 18.015 accounts for the water molecule lost during peptide bond formation

Our computational approach leveraged established bioinformatics libraries to ensure precise and consistent molecular weight calculations across diverse protein sequences.

A.3 Standard Amino Acids with their Background Frequency & KL Divergence for MaSp Repeats

Like human language, protein sequences can be represented as strings of letters, where the protein alphabet consists of 20 standard amino acids (AAs), excluding rare and unconventional ones. Similarly, naturally evolved proteins are composed of modular elements with slight variations, which can be rearranged and assembled hierarchically. In this analogy, common protein motifs and domains—fundamental functional units of proteins—are akin to words, phrases, and sentences in human language (Ofer et al., 2021).

The twenty amino acids (that make up proteins)each have assigned to them both three-letter (can be upper or lower case) and one-letter codes (upper case). This makes it quicker and easier for notation purposes and are worth learning. Table 4 gives these notations.

Table 4: Amino Acids, Their Codes, and Background Frequency Distribution for MaSp Repeats

Amino Acid Name	**3-Letter Code**	**1-Letter Code**	**Background Frequency**
Alanine	Ala	A	0.2232
Arginine	Arg	R	0.0129
Asparagine	Asn	N	0.0070
Aspartic Acid	Asp	D	0.0078
Cysteine	Cys	C	0.0002
Glutamine	Gln	Q	0.0850
Glutamic Acid	Glu	E	0.0069
Glycine	Gly	G	0.3766
Histidine	His	H	0.0003
Isoleucine	Ile	I	0.0050
Leucine	Leu	L	0.0138
Lysine	Lys	K	0.0017
Methionine	Met	M	0.0014
Phenylalanine	Phe	F	0.0038
Proline	Pro	P	0.0788
Serine	Ser	S	0.1004
Threonine	Thr	T	0.0123
Tryptophan	Trp	W	0.0002
Tyrosine	Tyr	Y	0.0485
Valine	Val	V	0.0141

For evaluation purposes, we established a background amino acid frequency distribution based on the training dataset of MaSp repeats (6K instance which were used in Stage 2). This distribution represents the mean occurrence of each amino acid across all sequences in the dataset. The calculated background frequencies are shown in Table 4.

This frequency distribution serves as a reference baseline for KL divergence calculations, enabling a quantitative comparison between generated sequences and naturally occurring MaSp repeats.

To assess sequence similarity, we calculate the Kullback-Leibler (KL) divergence between the amino acid composition of generated sequences and the background distribution of MaSp repeats. The background frequencies of amino acids were derived from a dataset of 6,000 MaSp repeat sequences, providing a reference probability distribution. Given a sequence S, its amino acid probability distribution P is compared against the background distribution Q using:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{6}$$

where $P(i)$ and $Q(i)$ represent the probabilities of amino acid i in the generated sequence and background dataset, respectively. This metric quantifies the deviation of generated sequences from natural MaSp repeats, aiding in model evaluation.