Sequence-based data-constraint deep learning framework to predict spider dragline mechanical properties

Sinan Keten (✉️ s-keten@northwestern.edu)  Northwestern University  https://orcid.org/0000-0003-2203-1425

Akash Pandey  Northwestern University

Wei Chen  Northwestern University

Article

Keywords:

Posted Date: December 21st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3727903/v1

License: ☺️ ✂️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: There is NO Competing Interest.
Sequence-based data-constrained deep learning framework to predict spider dragline mechanical properties

Akash Pandey,† Wei Chen,† and Sinan Keten*,†,‡

†Department of Mechanical Engineering, Northwestern University, Evanston
‡Department of Civil and Environmental Engineering, Northwestern University, Evanston

E-mail: s-keten@northwestern.edu

Abstract

We establish a deep-learning framework for describing the mechanical behavior of spider dragline silks to clarify the missing link between the sequence and mechanics of this exceptionally strong and tough biomaterial. The method utilizes sequence and mechanical property data of dragline spider silk as well as enriching descriptors such as residue-level mobility (B-factor) predictions. Our sequence representation captures the relative position, repetitiveness, as well as descriptors of amino acids that serve to physically enrich the model. We obtain high Pearson correlation coefficients (0.76-0.88) for strength, toughness, and other properties, which show that our B-factor based representation outperforms pure sequence-based models or models that use other descriptors. We prove the utility of our framework by identifying influential motifs, and also by demonstrating how the B-factor serves to pinpoint potential mutations that improve strength and toughness, thereby establishing a validated, predictive, and interpretable sequence model for designing sustainable biomaterials with sequence-defined properties.
Introduction

Spider dragline silk is well known for its extraordinary strength and toughness; higher than other natural silks, Kevlar, and steel.\textsuperscript{1,2} While extensibility\textsuperscript{3} and tensile strength\textsuperscript{4–6} are diametrically opposed properties, they are uniquely both attained in spider silk’s lifeline, the dragline silk. Microbial fabrication of protein-based materials\textsuperscript{7–11} positions spider silk to show even greater potential in the fields of protective gear, textile engineering, medicine, and surgery by increasing yield and other desirable attributes.\textsuperscript{1,12} It has been found in the literature that the spider silk properties are highly dependent on its semicrystalline structure which in turn is dependent on the amino acid sequence.\textsuperscript{13} Therefore it is paramount to understand the influence of the sequence on properties.

Researchers have used experimental techniques like X-ray,\textsuperscript{14} solid-state NMR spectroscopy\textsuperscript{15,16} and Raman spectroscopy\textsuperscript{17} along with tensile tests\textsuperscript{18} to capture the structure-property relationships in the spider silk. Computationally, researchers have used Molecular Dynamics (MD) simulations to characterize the importance of \( \beta \)-sheet crystal confinement in the concerted failure of hydrogen bonds, which is partly facilitated by disorder-inducing Prolines.\textsuperscript{19} MD has also been used to study the impact of \( \beta \)-sheet nanocrystal size on mesoscale mechanical properties.\textsuperscript{20,21} Dissipative particle dynamics (DPD) and MD simulations are used complementarily to establish and validate relationships among experimental process parameters such as spinning solution concentration and spin speed\textsuperscript{22,23} and to understand the impact of spidroin hydrodynamic flow in the spider duct.\textsuperscript{24} Overall, MD simulations have served as an invaluable tool for establishing mechanistic insights into the molecular underpinnings of spider silk’s superb properties and other mechanisms.\textsuperscript{20,25} However, given the computationally intensive nature of MD calculations, establishing a rigorous relation between the primary sequence of the spider silk to the macroscopic mechanical properties has been elusive with MD and other physical models.

Noting the challenges of physics-based modeling for protein-based materials, there is a need for a predictive model to bridge the gap between the primary sequence and the
macroscale property. In case of the spider silk, such work is scarce. One such work for spider silk in the literature predicts the peak force obtained from the MD simulation based on the features obtained from the primary sequence of the repetitive region. However, since the data used is still based on nanoscale simulations at fast rates of deformation, the comparisons are largely qualitative. In another work, researchers have proposed a transformer-based generative model to produce spider silk for a given set of properties but have not yet reported the mechanical property prediction accuracy of the model. With the recent advances in the field of machine learning (ML), several groups have worked towards establishing a primary sequence-property relationship in other contexts either using a large sequence-based deep learning (DL) model or small ML models like XGBoost and random forest. Large DL models have the risk of overparameterization when trained on small (<500) data sets and small ML models require prior knowledge of the application to extract input features from the primary sequence. To fulfill the need for an experimentally validated, quantitative, and predictive model to link the sequence and mechanical properties of spider silk, we train a DL model for dragline spider silk on the recently published spider silkome database. Uniquely, we use limited data and no assumptions about the important motifs in spider silk. We then use our DL model to show the importance of residues’ dynamic property information (B-factor) for the prediction of mechanical properties of dragline spider silk; hence drawing a parallel with the findings that the Debye-Waller parameter is strongly correlated with the mechanical properties in polymers.

The layout of the study is as follows. First, we establish a DL model with less trainable parameters to predict the mechanical properties of spider dragline silk. For the DL model, we explore a sequence-length agnostic representation that captures the relative position, repetitiveness, as well as descriptors of amino acids that serve to physically enrich the model. We show the benefit of choosing the B-factor as the enriching descriptor. Subsequently, we use our framework to solve the large combinatorial problem of motif identification and establish some design rules for modulating different mechanical properties. We then carry out
a systematic study of potential mutations that can improve mechanical properties. Finally, we try to interpret these mutations from the viewpoint of molecular mobility, using the B-factor as a sequence-dependent local order parameter. We conclude with mutations and motifs that seem promising for future synthetic silk-like material designs.

Materials and Methods

Dataset
Kazuharu et al.\textsuperscript{31} have carried out extensive work to test and document various mechanical properties of dragline silks in the so-called Spider Silkom Database. The database consists of the mechanical properties of silks obtained from 446 species along with the respective repetitive regions of various spidroins (major and minor spidroins) that make up these specimens. In this work, the SSD database was utilized for training the model. The primary sequences were downloaded using the FASTA download option in SSD. The mean and the standard deviation of six different mechanical properties of the dragline of 446 species were obtained from the CSV file output of SSD. The properties investigated herein are strain at break ($\epsilon_f$), tensile strength ($\sigma_{UTS}$), toughness ($T$), Young’s modulus ($E$), and supercontraction ($\epsilon_{sup}$). For our DL model, we match the IDs of the primary sequence FASTA files and IDs in the CSV file to match the primary sequence information to its corresponding mechanical properties. After matching the IDs, we gather the primary sequence and corresponding properties of the dragline of 203 species only for all properties except $\epsilon_{sup}$. The reduction in the data set size is a result of the mismatch between IDs in the FASTA file and the CSV file. For $\epsilon_{sup}$, we gather data primary sequence and property data of 49 species. Since SSD only documents the primary sequence of the repetitive part of the spidroins, hereafter the word \textit{sequence} will be used to indicate the primary sequence of the repetitive part of the spider silk.
Model Architecture

Predicting the mechanical properties of spider silk based on the primary sequence of various spidroins is a complex task due to the lack of knowledge about the microstructural organization of the silk and how that translates to constitutive relations. Additional complexity arises from the fact that different spidroins (MaSp1, MaSp2, MaSp3, MaSp, MiSp) are present in the spider silk and their molecular organization is largely unknown. This poses challenges to the physics-based modeling of silks, and as such makes data-driven methods that are agnostic to structural data advantageous for this purpose. This calls for a representation that can capture the effect of each spidroin on the properties of the spider silk. With the advancements in the field of deep learning, one common strategy is to use pre-trained models such as ProtBERT. ProtBERT has been used previously for predicting properties in other contexts. However, this method involved fine-tuning more than a million parameters. As discussed in the above section, this would approximately require 500 or more training data to train such a large model, making it infeasible for the SSD. For this reason, we present a new representation of the sequence that deals with the complexities and data constraints mentioned herein (see Sec. Representation).

It has been reported in the literature that certain motifs have a higher impact on the properties but the list of motifs is limited and needs a framework to identify important motifs for different properties. Ideally, a predictive model should be able to help ascertain which segments of the primary sequence (motifs) impact (positively or negatively) each property. Identifying influential motifs is crucial for designing new sequences that result in improved mechanical properties, for instance through microbial production of designer sequences to form a silk dope that can be spun into fibers. However, motif identification in protein is a large combinatorial problem. Therefore, our secondary goal with this study is to build a framework that can identify critical motifs.
**Representation**

We first establish a representation that deals with data constraints, namely small property datasets, incomplete sequence information, and the presence of multiple spidroins in each fiber, while also giving us interpretable features that can be used for motif identification later. In our study, we create representations only for spidroins that form the major fraction of the sequence, namely MaSp1, MaSp2, MaSp3, MaSp, and MiSp spidroins. For this purpose, we build our representation based upon the Quasi Residue Couple (RC) concept. The RC representation converts the primary sequence information of a spidroin to a three-dimensional array labeled as $P$. In its original form, the RC representation records the summation of the physicochemical properties of two residues being at a specific distance from each other in the primary sequence. For instance, adjacent residues in sequence space form a $20 \times 20$ matrix with matrix elements representing the average sum of physicochemical properties of pairs of residues adjacent to each other in the sequence. Similarly for a sequence distance of $m$ residues, a $20 \times 20$ matrix can be expressed using Eq. 1a & 1b:

$$P^d_m(i, j) = \frac{1}{N - m} \sum_{n=1}^{N-m} H_{i,j}(n, n + m, d), \quad i, j \in [1...20] \quad (1a)$$

$$H_{i,j}(a, b, d) = \begin{cases} 
  d_i + d_j, & \text{if at position } a \text{ amino acid } i \text{ and at position } b \text{ amino acid } j \text{ is present} \\
  0, & \text{otherwise}
\end{cases} \quad (1b)$$

The $20 \times 20$ representation for each $m$ value (1, 2,..., max(m)) can be stacked together to get a 3D representation of dimension $20 \times 20 \times \text{max(m)}$. Our representation $P$ also incorporates enriching descriptors of the amino acids in the spidroin which is assigned to the variable $d$. To study the effect of amino acid’s enriching descriptor $d$, we train and test our model using several different representations, which are repetition ($d = 1$), $d$ as B-factor,
It is important to note that the hydrophobicity value we use for each amino acid represents the solubility of the amino acids in the water at pH 7. Similarly, to account for the charge using a continuous variable, we use the isoelectric point (pI) value for each amino acid, which represents the pH of the solution at which the net charge of amino acids is zero. Of particular note, we consider the B-factor of individual residues as an enriching descriptor since it is a critical indicator of their dynamic properties and is similar to the Debye-Waller parameters in the polymers. In the literature, it has been shown that the Debye-Waller parameter is strongly correlated with the mechanical properties. Therefore, in our study, we have chosen the B-factor as one of the d’s. To calculate the B-factor based on the primary sequence of the spidroin, we use the deep learning model developed in a previous study by some of the authors. It should be noted that the deep-learning model for the B-factor outputs the normalized B-factor value with the mean and standard deviation of 0 and 1 respectively among all the residues in a protein.

In sum, \( P \) captures the property and the frequency of all possible pairs of amino acids (20*20) in the primary sequence. It captures the information about not only the pair of amino acids next to each other in the primary sequence but also at a distance of \( m \) from each other in the primary sequence. A major advantage of this representation is that the array size is independent of the sequence length provided and is a constant once \( m \) is set.

In Sec. Results and discussion, we discuss the effect of \( \text{max}(m) \) on the model quality. This distinction makes our representation advantageous over other models like ProtBERT as their output size is dependent on the length of the primary sequence and consequently requires large sequence-based deep learning models to post-process the data. Since we have less data (refer Sec. Dataset), our choice of representation is justified as it is agnostic to the length of the primary sequence and does not require sequence-based models to post-process them.

One of the numerical shortcomings of the \( P \) representation is that when amino acids
i and j have highly contrastive d, addition in Eq.1a can lead to loss of information due to averaging. Therefore, we introduce another representation (L) as shown in Eq.2a & 2b which has an extra dimension compared to P to store the values of di and dj separately as

\[ L^d_m(i, j, k) = \frac{1}{N - m} \sum_{n=1}^{N-m} H_{i,j}(n, n + m, k, d), \quad i, j \in [1...20], \quad k \in [1, 2] \quad (2a) \]

\[ H_{i,j}(a, b, k, d) = \begin{cases} 
  d_i, & \text{(if at position } a \text{ amino acid } i \text{ and} \\
  \text{at position } b \text{ amino acid } j \text{ is present) and (k=1)} \\
  d_j, & \text{(if at position } a \text{ amino acid } i \text{ and} \\
  \text{at position } b \text{ amino acid } j \text{ is present) and (k=2)} \\
  0, & \text{otherwise} 
\end{cases} \quad (2b) \]

This will resolve the numerical issue for the pairs of i and j with contrastive d values in P. It is important to note that for d=1, P, k=1 in L, and k=2 in L are exactly the same; hence for d=1, we will be using only P representation throughout the paper.

Since in the further sections, we will be studying the importance of different elements in the representations discussed above, it is important to introduce a nomenclature to represent each element. Any element can be named as "(Spidroin type)-(Amino Acid)i-(Amino Acid)j-m". The spidroin type represents the spidroin from which the feature is derived. (Amino Acid)i and (Amino Acid)j represent the amino acids which are represented by the position i and j in Eq.1a & 2a. As discussed in Sec. Representation, m represents the gap between amino acid i and j in the primary sequence as in Eq.1a & 2a. For example, let us consider a feature identified as MaSp1-Y-G-3. The notation indicates that this feature belongs to MaSp1 and that the amino acid G is 3 positions away from amino acid Y in the sequence.
Deep Learning Model

Once the representation is set, the next step involves choosing a deep-learning (DL) method to process this information for predicting properties. Before presenting the method, it is important to discuss the dimension as well as the total number of elements in the representation. Given that we have considered 5 types of spidroins in our study (refer Sec. Representation), for max(m) equal to 6, the total number of elements from $P_{20 \times 20 \times 6}$ and $L_{20 \times 20 \times 6 \times 2}$ are size($P$)*5 (12000 features) and size($L$)*5 (24000 features) respectively.

Processing such a high-dimensional representation using deep learning models such as 1D/2D convolution network (CNN) requires approximately a million parameters. However, the SSD is relatively small, so there are high chances of overfitting the data while training the model with millions of parameters. This issue can be addressed by using a filtering technique in which the important features are identified based on their Pearson correlation coefficient (PCC) with the output property (features with PCC>(user-defined cutoff value) are selected for the DL model) as shown in Fig.1. It is important to note that the cutoff value to select the features can vary based on the spidroin type to control the number of input features to the deep-learning model. Our choice for the cutoff values can be found in the code. Using all the features that pass this filtering, we make a vector of the input variables of length $T$ for each data point. For $N_e$ number of data points, the feature space $F$ can be written as a 2D tensor as shown in Eqn.3. Eqn.3 indicates a feature as $f_{ij}$ where $i$ indicates the example number and $j$ represents the $j^{th}$ feature in the $i^{th}$ example. This input vector is then fed into a feed-forward neural network (FFNN) which maps $f$ to predicted property $Y_p$. To optimize the parameters in FFNN, we use the Mean Square Error (MSE) between predicted ($Y_p$) and experimental ($Y_a$) properties as shown in Fig.1. The formula to calculate the MSE for any property with $N$ training example is shown in Eq.4. It should be noted that we use different FFNNs for different properties and they are optimized separately.
Figure 1: Model architecture for spider silk property prediction with representation $\mathcal{P}$

\[
\begin{align*}
\mathbf{F} &= [\mathbf{f}_1^1, \mathbf{f}_2^1, \cdots, \mathbf{f}_N^1] = \\
&= \begin{bmatrix}
f_1^1 & f_2^1 & \cdots & f_N^1 \\
\vdots & \vdots & \ddots & \vdots \\
f_j^1 & f_j^2 & \cdots & f_j^N \\
\vdots & \vdots & \ddots & \vdots \\
f_T^1 & f_T^2 & \cdots & f_T^N
\end{bmatrix}
\end{align*}
\] (3)

\[
MSE = \frac{\sum_{i=1}^{N}(Y_a - Y_p)^2}{N}
\] (4)

Training Details

For training and testing the model, we split the data as 80% train, 10% validation, and the other 10% test sets. We train the model using the train data set and check for overfitting using the validation data. Overfitting is the point in the training when training loss continues to decrease but validation loss starts increasing with the number of iterations. Then we use the trained model to predict on the test dataset, which is then used to judge the quality of the fit on the basis of two metrics: $R^2$ and PCC. $R^2$ is a common metric to study the quality of the fit for the regression problem. However, in the current study, we have seen that the experimental data has an average standard deviation of 20-30% of the mean value, and similar standard deviations in the experimentally measured mechanical properties are corroborated in the literature. Experimental conditions such as humidity or spinning process parameters like reeling speed can have a significant impact on the mea-
sured properties, which explains the high variability. Since the ground truth has such a high standard deviation, we also use PCC to study our model’s ability to capture the trend of the experimental data.

To study the generalizability of the model, we execute the training and testing process for 13 different seeds. These 13 seeds are chosen in such a way that every experimental point is used at least once in the test dataset. We report the mean and standard deviation of $R^2$ and PCC across these 13 seeds. Two different probabilities are used to compare models: 1) $P_r = P(R^2 \geq 0.5)$, and 2) $P_c = P(\text{PCC} \geq 0.7)$.

Having set the process for training and comparing two different models above, we perform a parametric study to arrive at the best representation ($\mathcal{P}$ or $\mathcal{L}$) and the max($m$) value for the same. Through the first parametric study in Sec. **Effect of representation size on the prediction** of the Supplemental information, we show that max($m$)=6 works reasonably well for all the mechanical properties. In another parametric study shown in Sec. **Effect of different representations on the prediction** of the Supplemental information, we conclude that the representation $\mathcal{L}$ is a better choice than $\mathcal{P}$ for our model.

**Feature Importance Analysis**

In this section, we discuss the first step toward motif identification - calculating feature importance. To calculate the feature’s importance, we follow the method called permutation feature importance algorithm. Additionally, we use test datasets to calculate the feature importance as the performance of the DL model on the test dataset reflects its true performance. To calculate the importance of $f_j$ based on this method, we do the following:

- Calculate the MSE ($e_o$) for test dataset with the features shown in Eq.3

- Freeze all the features $f_1^{1:N_e}$ to $f_T^{1:N_e}$ except $f_j^{1:N_e}$, then randomly shuffle feature $f_j^{1:N_e}$ in Eq.3

- Calculate the MSE ($e_p$) for the test dataset with the shuffled $f_j$ feature.
- Feature importance \((q_j)\) for \(f_j\) is calculated by \(|100^*(e_p - e_o)/e_o)|\.

The above steps are repeated for all the features in \(f\). At this point of analysis, we are only interested in the relative importance of all features; hence, we divide all the \(q'_j\)s with \(\max(q_1, q_2, \ldots, q_T)\). The normalized feature importance for \(f_j\) is denoted as \(\overline{q}_j\). Now the values of \(\overline{q}_j\) lie in the range [0,1] with 0 and 1 indicating the least and the most important feature respectively.

**Method for motif identification and quantifying their effect**

As a part of this study, we aim to identify some of the major motifs in spider silk that most strongly influence specific mechanical properties. For this purpose, we first choose the features with \(\overline{q}_j > 0.1\) using the model which is trained with \(d=1\). We choose the model with \(d=1\) because it purely captures the degree of occurrence of relative amino acid positioning in a given spider silk and this aligns with the aim for motif identification.

The filtered features are not the complete motifs. Referring to the above section it is clear from the feature name that there are gaps between the first and the last amino acid which needs to be filled. For example, the feature MaSp1-Y-G-3 looks like Y___G and there are 2 positions indicated with ___ that need to be filled to complete the motif. For general reference, let us indicate the incomplete motif as \(a_1 \ldots a_m + 1\), and the positions starting from left to right are filled using Eq.5 as:

\[
\begin{align*}
    a_k &= \arg \max_{AA} P(AA|a_1, \ldots a_{k-1}, a_{m+1}) \\
    k &\in [2, \ldots, m], \text{and } AA \in 20 \text{ amino acids}
\end{align*}
\]

The conditional probability in Eq.5 is calculated using all the examples obtained in Sec.\textbf{Dataset}. Basically to fill an empty position, Eqn.5 searches for the most commonly occurring amino acid at an empty position while considering the filled positions as a condition for the search through the dataset. Also, it is important to note that the positions are filled from left to right in the current scenario.
Once the motif is identified based on maximum likelihood, we consider it for further analysis only if it appears in more than 5% of the spider silks in the dataset. The number 5% was intuitively chosen based on the SSD paper in which they use the same percentage to study the correlation between the degree of occurrence of different motifs and measured mechanical properties.

Once all the motifs are identified using the above method, it is important to quantify their impact on the property. Before discussing the method for quantifying the effect of motifs, we want to introduce a variable to count the number of motifs in the sequence of the spidroin (MaSp1, MaSp2) in the spider silk. Since every spider silk has the sequence of varying lengths, it is important to introduce a variable that is independent of the length of the sequence. This is done by normalizing the number of motifs in the sequence with the number of repeat units in the sequence. Spider silk spidroins consist of several amorphous and crystalline segments in a sequence. The reference literature defines a repeat unit as starting from the crystal-forming segment and ending before the next crystal-forming segment in the sequence. Thus, the number of repeat units can be defined as the number of crystalline segments in a sequence. We adopt the definition of the crystalline segment from the reference literature as the continuous segments of S, A, and V amino acids with lengths of more than 5. Therefore, based on the above discussion, we define ($\theta_N$) as the normalized number of motifs as given by Eq.

$$\theta_N = \frac{\# \text{ of motifs in the sequence}}{\# \text{ repeat units}}$$  \hspace{1cm} (6)

For quantifying the effect of motifs, we first choose the best model we obtain from the parametric study across various max(m), representations ($\mathcal{P}$ & $\mathcal{L}$), and $d$. For this study, we select the data (test dataset) on which the model has not been trained; because the result on the test dataset reflects the true performance of the DL framework. To quantize the impact of the motif on the property, we do the following:
• Use the selected model to predict the property for all test datasets and store all the predicted results in an array $Y_p$.

• For each test example, calculate the $\theta_N$ value for the motif under the assessment and store it in an array $\theta_N$.

• For all the test examples, remove the motif under the assessment from the sequence and recalculate the features.

• Predict the property with the recalculated features and store the results in an array $Y_C$.

• Calculate the percentage impact ($P_m$) of the motif using Eq.7a & 7b sequentially. The max property value in Eq.7b for various mechanical properties is obtained from the SSD paper.\textsuperscript{31}

\begin{align*}
I_m &= \text{Mean} \left( \frac{Y_p - Y_c}{\theta_N} \right) \\
P_m &= \frac{100 \ast I_m}{\text{Max Property value}}
\end{align*} \tag{7a,b}

Table 1: Max values of various mechanical properties obtained from the literature for the dragline spider silk.\textsuperscript{31}

<table>
<thead>
<tr>
<th>Property</th>
<th>Max Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_f$</td>
<td>65.3%</td>
</tr>
<tr>
<td>$T$</td>
<td>0.425 GJ/m$^3$</td>
</tr>
<tr>
<td>$\sigma_{UTS}$</td>
<td>3.33 GPa</td>
</tr>
<tr>
<td>$E$</td>
<td>37 GPa</td>
</tr>
<tr>
<td>$\epsilon_{sup}$</td>
<td>49.2%</td>
</tr>
</tbody>
</table>

Above, the entire process is discussed with the motifs obtained from Eq.5 using all the examples obtained in Sec.\textbf{Dataset}. However, it is possible that the motifs in the dragline spider silk species with very high and low mechanical properties can be different. Therefore to study this difference, we first sort the dragline spider silk species in the dataset based on the respective mechanical properties. From the sorted list, we make two groups of dragline
silk species, one with the highest 10% and another with the lowest 10% mechanical properties. We then obtain the motif using Eq.\ref{eq:5} for both the highest and lowest 10% group of dragline spider silk separately. The process to calculate $P_m$ for these motifs remains the same as above.

To sum up the method for the motif identification, for every feature $f_i$ with $\bar{q}_i > 0.1$, we identify 3 types of motifs using Eq.\ref{eq:5} on the: 1) whole dataset ($\phi_w$), 2) dragline spider silk with the highest 10% property ($\phi_t$), and 3) dragline spider silk with lowest 10% property ($\phi_b$). Subsequently, we calculate the $P_m$ value of each motif identified to quantize their impact on the property.

**Mutation study**

After studying different motifs in the dragline spider silk, we will also study the effect of certain mutations on the mechanical properties. We represent all mutations as $Res_1 \rightarrow Res_2$. This means that we mutate all residues of type $Res_1$ to $Res_2$ in the specific spidroin of a given spider silk sample. The same mutation can have varying degrees of impact on different spider silk samples. Hence, to account for the variation in the impact, we calculate two probabilities: 1) the probability that the mutation will lead to an increase in the property by 20% ($P(\text{Increase}>20\%)$), and 2) the probability that the mutation will lead to a decrease in the property by 20% ($P(\text{Decrease}>20\%)$). Further to study if B-factor can explain the effect of mutation, we introduce a variable $\Delta$B-factor which is equal to ($Res_2$’s mean B-factor -$Res_1$’s mean B-factor) in $Res_1 \rightarrow Res_2$ mutation.

**Results and discussion**

**Choice of enrichment descriptor of amino acids**

In the Supplemental information, we have established appropriate $\text{max}(m)$ values and appropriate representations ($P$ or $L$) for all the properties. Next, we study the relative importance
of amino acid’s enriching descriptor \( d \). Fig.2 shows the comparison between different \( d \) for the best-chosen max(m) and representation for all the properties. The \( d=\text{B-factor} \) with \( \mathcal{L} \) representation clearly outperforms all other \( d \)’s for all properties except \( \epsilon_{\text{sup}} \), which is evident from \( P_r \) and \( P_c \) values given in Tab.2 and Tab.3 respectively. The B-factor renders features that are most informative for the prediction of mechanical properties as it performs as the best descriptor for all properties except \( \epsilon_{\text{sup}} \). This can be physically justified by the fact that the Debye-Waller factor of atoms in polymers is found to have an inverse correlation with its bulk modulus,\(^{32,48}\) and also serves as an indicator for glass-transition behavior. We also know that the cohesive energy, which is also related to the Debye-Waller factor, is used to compute the bulk modulus which governs all other mechanical properties in models like the group interaction model (GIM)\(^{49}\) and other constitutive laws.\(^{50}\) This implies that there is a correlation between molecular mobility and mechanical properties. The outperformance of the model using \( d=\text{B-factor} \) supports the notion that the segmental molecular mobility in proteins is strongly related to macroscale mechanical properties. For \( \epsilon_{\text{sup}} \), \( d=1 \) performs the best. This can either be due to the lack of data in the case of \( \epsilon_{\text{sup}} \) or the fact that \( \epsilon_{\text{sup}} \) majorly depends on just the occurrence of certain motifs in the spider silk. In the literature,\(^{31}\) it has been shown that the \( \epsilon_{\text{sup}} \) highly depends on the occurrence of poly-Alanine motifs.

![Figure 2: Effect of different enriching descriptor \( d \) on the model performance.](image-url)
Table 2: $P_r = P(R^2 \geq 0.5)$ for different choice of enriching descriptor $d$. The highest probability (best scenario) is shown in bold.

<table>
<thead>
<tr>
<th>Choice of $d$</th>
<th>B-factor ($\mathcal{L}$)</th>
<th>1</th>
<th>Monomer Weight ($\mathcal{L}$)</th>
<th>Monomer pI value ($\mathcal{L}$)</th>
<th>Hydrophobicity ($\mathcal{L}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_f$</td>
<td>0.80</td>
<td>0.5</td>
<td>0.63</td>
<td>0.56</td>
<td>0.2</td>
</tr>
<tr>
<td>$\sigma_{UTS}$</td>
<td>0.70</td>
<td>0.29</td>
<td>0.25</td>
<td>0.28</td>
<td>0.3</td>
</tr>
<tr>
<td>$T$</td>
<td>0.86</td>
<td>0.48</td>
<td>0.33</td>
<td>0.39</td>
<td>0.38</td>
</tr>
<tr>
<td>$E$</td>
<td>0.75</td>
<td>0.02</td>
<td>0.12</td>
<td>0.02</td>
<td>0.14</td>
</tr>
<tr>
<td>$\epsilon_{sup}$</td>
<td>0.86</td>
<td>0.96</td>
<td>0.33</td>
<td>0.74</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Table 3: $P_c = P(PCC \geq 0.7)$ for different choice of enriching descriptor $d$. The highest probability (best scenario) is shown in bold.

<table>
<thead>
<tr>
<th>Choice of $d$</th>
<th>B-factor ($\mathcal{L}$)</th>
<th>1</th>
<th>Monomer Weight ($\mathcal{L}$)</th>
<th>Monomer pI value ($\mathcal{L}$)</th>
<th>Hydrophobicity ($\mathcal{L}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_f$</td>
<td>0.87</td>
<td>0.54</td>
<td>0.73</td>
<td>0.67</td>
<td>0.32</td>
</tr>
<tr>
<td>$\sigma_{UTS}$</td>
<td>0.75</td>
<td>0.34</td>
<td>0.29</td>
<td>0.33</td>
<td>0.32</td>
</tr>
<tr>
<td>$T$</td>
<td>0.92</td>
<td>0.5</td>
<td>0.42</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>$E$</td>
<td>0.79</td>
<td>0.06</td>
<td>0.18</td>
<td>0.05</td>
<td>0.22</td>
</tr>
<tr>
<td>$\epsilon_{sup}$</td>
<td>0.90</td>
<td>0.99</td>
<td>0.57</td>
<td>0.80</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Based on the parametric studies presented for different representations, properties (d), and max(m) values (Sec. Best model details in the Supplemental Information), the best choice for each mechanical property is given in Tab.S3. Tab.S4 in the Supplemental Information gives details about the number of input features to FFNN and the number of trainable parameters in FFNN for each mechanical property. From the above discussions, we have shown that the deep learning model developed for the prediction of mechanical properties of the spider silk is robust and accurate, considering the high variability in experimental data as discussed in Sec. Training Details.

To further prove the robustness of our model, we test it against an experimental mutation study presented in the literature. One of the experimental studies shows that mutating Tyr (Y) to Phe (F) in MaSp1 of biomimetic spider silk decreases $\epsilon_{sup}$. Therefore, in our test dataset, we replace Y with F in MaSp1 and observe a mean decrease of 71% in $\epsilon_{sup}$.
Therefore, our model predicts the same trend as observed in the experiment, thereby validating our model.

**Motif identification**

Having proved the robustness of our deep-learning based model, in this section, we will discuss the motifs identified to be most influential for different mechanical properties. As the first step, we calculate the feature importance ($\overline{q}_i$) of all the features ($f_i$) considered for the prediction of the properties using the method discussed in Sec. Feature Importance Analysis. Subsequently, for the features with $\overline{q}_j > 0.1$, the 3 types of motifs ($\phi_w$, $\phi_t$, and $\phi_b$) are identified and their impact ($P_m$) is quantized as described in Sec. Method for motif identification and effect quantization. The complete information about the motifs and their impact is presented in Tab.S5-S9 in the Supporting Information. It can be observed from the tables that $P_m$ values can take on positive or negative values indicating a positive or negative correlation between the number of motifs ($\theta_n$ in Eq.6) and the property respectively. At this point, it is essential to physically interpret the impact magnitude $P_m$. To that extent, let us take motif LVSSGP (from MaSp1) for $\epsilon_f$ as an example. It is evident from Tab.S5 that LVSSGP contributes 0.61% of the max $\epsilon_f$ value per $\theta_n$. Now, if we want to increase the $\epsilon_f$ by 1.83% of max $\epsilon_f$ value, then we need to increase $\theta_n$ of LVSSGP by 3. Based on Eq.6, $\theta_n$ can be increased by either increasing the number of motifs or decreasing the number of repeat units in the sequence.

All the mechanical properties of the dragline spider silk are due to the collective effect of several motifs. This is evident from the fact that none of the $P_m$ values in Tab.S5-S9 are extremely high. The contribution of so many different motifs makes it very difficult to come up with one common design rule for optimizing any property. For example, increasing the LVSSGP motif in MaSp2 increases $\epsilon_f$ but it also leads to the increase of SS motif which has a negative impact on $\epsilon_f$. Hence, relationships like this need to be considered while designing fibrous protein-based materials. For the same reason, optimizing a primary sequence for
two properties at the same time will be more difficult especially when most motifs have contrasting effects on the two properties. For instance, it may be desirable to increase both $\epsilon_f$ and $\sigma_{UTS}$, however, motifs like SS have negative and positive impacts on both the properties as shown in Tab.S5 & S6 respectively.

**Design rules**

One of the aims of this work is to find the mutations that are required for increasing the mechanical properties in the dragline spider silk. To discuss mutations, we introduce the nomenclature used to indicate substitution as well as deletion/insertion in proteins, which follows standard mutation nomenclature. The nomenclature for substitution is $<\text{Res}_b><\text{pos}><\text{Res}_a>$ which means that amino acid $\text{Res}_b$ is being replaced by amino acid $\text{Res}_a$ at position $\text{pos}$. To indicate the deletion/insertion we use $<\text{Res}_s><\text{Res}_e><\text{delins}<\text{group of newly inserted amino acid}>$ as the nomenclature where $\text{Res}_s$ and $\text{Res}_e$ indicate the first and last amino acid deleted.

Based on the observations from Tab.S5-S9, we present some mutation recommendations to increase the properties in Tab.4. Before interpreting these mutations, it is essential to recall that the spider silk structure consists of crystalline as well as amorphous regions. The crystalline region consists of groups of amino acids forming $\beta$-sheets. Based on the literature, it is understood that the amino acids Ala, Val, Ile, Tyr, Cys, Trp, Phe, and Thr are more likely to be found in $\beta$-sheet regions and amino acids Gly, Pro, Asn, and Ser are more likely to be found in the turns connecting two $\beta$-sheets. Further considering the likelihood/propensity of the amino acids to form $\beta$-sheet, they can be ranked as L, V>A>G. To study the effect of mutation, we report the $\Delta P_m$ value which is defined as the difference between the higher and lower $P_m$ values.

Taking the above facts into consideration, we study the effect of different mutations on different mechanical properties. We first start with $\epsilon_f$ and hypothesize that it increases 50% of the times when the mutations of the amino acids decrease the $\beta$-sheet propensity. Next,
we examine $\sigma_{UTS}$ and $E$ and observe that 50% of times the mutation of amino acids that increase $\beta$-sheet propensity also increases the property. The percentage 50% might look like a coin-toss probability but it is important to note in this case there are 3 possible mutations: high to low propensity, low to high propensity or the trend is not very clear such as for mutation G to N. The toughness ($T$) does not show any clear trend like other properties because $T$ is dependent on the area under the stress-strain curve. The area under the stress-strain curve is driven by high $\sigma_{UTS}$ and $\epsilon_f$. Since different mutations favor $\sigma_{UTS}$ and $\epsilon_f$, we cannot observe a clear trend of $T$ with respect to the $\beta$-sheet propensity like other properties. Higher toughness requires a higher area under the stress-strain curve which in turn requires higher maximum stress or the strain at break or both.

In the case of supercontraction ($\epsilon_{sup}$), it is evident from the $P_m$ values in Tab.S9 that the larger the length of the poly-Ala motif, the larger the decrease in $\epsilon_{sup}$. This observation is backed by the literature study that shows that the $\epsilon_{sup}$ is positively correlated with the amorphous/poly-Ala region length ratio (PCC=0.53). Increasing the length of even one poly-Ala motif leads to the decreases in amorphous/poly-Ala region length ratio and subsequently the $\epsilon_{sup}$. Building on this, we observe from Tab.4 that mutating a larger poly-Ala to a smaller one increases the property.
Table 4: Design rules for each property. Mutation from column 2 to column 3 increases the property.

<table>
<thead>
<tr>
<th>Property</th>
<th>Before Mutation</th>
<th>After Mutation</th>
<th>Mutations</th>
<th>(\Delta P_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\epsilon_f)</td>
<td>QVKT</td>
<td>QVNT</td>
<td>K3N</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>GGAGQQ</td>
<td>GGYGPQ</td>
<td>A3Y, Q5P</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>GGQGPYG</td>
<td>GGPQGGYG</td>
<td>Q3P, P5G</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>SAVST</td>
<td>SSGPT</td>
<td>A2, S5delinsSGP</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>AAAAGY</td>
<td>AAAGGY</td>
<td>A4G</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>QGPYG</td>
<td>QGPSG</td>
<td>G4S</td>
<td>0.15</td>
</tr>
<tr>
<td>(\sigma_{UTS})</td>
<td>AAAAAAAA</td>
<td>AGQQGAGAA</td>
<td>A2, A7delinsGQGGAG</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>AGQQGGA</td>
<td>AGAAAA</td>
<td>Q3, G5delinsAAA, G7A, A9A10delinsGG</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>GAAAAA</td>
<td>AGAAAA</td>
<td>Q3, G5delinsAAA, G7A, A9A10delinsGG</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>GAGGAGGQG</td>
<td>GAGAAAAA</td>
<td>G2, G5delinsAGAA, G7, G9delinsAAA, G7, G9delinsGG</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>GLGSGQGY</td>
<td>GAGQGGGY</td>
<td>G7, G9delinsGAGQQG</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>AAAAGGGS</td>
<td>AAAGGYGP</td>
<td>A4, G8delinsGGYGP</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>SGPGYGP</td>
<td>SGPGYGPAS</td>
<td>G9A</td>
<td>0.11</td>
</tr>
<tr>
<td>(E)</td>
<td>YGSA</td>
<td>YGPA</td>
<td>S3P</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>AGQQGL</td>
<td>ALVHIL</td>
<td>G2, G5delinsLVHI</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>YQGP</td>
<td>YGAP</td>
<td>Q2, G3delinsGA</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>QGGYGYY</td>
<td>QGQQGYY</td>
<td>Y4Q</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>VHILGSSIGQVN</td>
<td>VHILGSANIGQVN</td>
<td>S7, S8delinsAN</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>SGQGGY</td>
<td>SQNQGGY</td>
<td>G2, G3delinsQN</td>
<td>0.32</td>
</tr>
<tr>
<td>(T)</td>
<td>PGGAGGSGPY</td>
<td>PGGQGPGYPG</td>
<td>A4Q, G6, S7delinsPY, Y10, P12delinsGAA, S15A, A20G</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>GPAASAA</td>
<td>AAAAAA</td>
<td>A4Q, G6, S7delinsPY, Y10, P12delinsGAA, S15A, A20G</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>AAAAY</td>
<td>AAGGY</td>
<td>A4Q, G6, S7delinsPY, Y10, P12delinsGAA, S15A, A20G</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>TSSNLKQA</td>
<td>TGAAAAAA</td>
<td>S2, Q7delinsGAAAAA</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>QGPQGAGQ</td>
<td>QGPSQGPA</td>
<td>G4S, A6P, Q8, Q9delinsAY</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>QGPQQGQ</td>
<td>YPGPSQ</td>
<td>G4S, A6P, Q8, Q9delinsAY</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>PGQQQGP</td>
<td>PGQGGYG</td>
<td>Q5, Q6delinsGY, G10Q</td>
<td>0.17</td>
</tr>
<tr>
<td>(\epsilon_{sup})</td>
<td>GN</td>
<td>GL</td>
<td>N2L</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>AAA</td>
<td>AA</td>
<td>A3del</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>AAAAAA</td>
<td>AAAAA</td>
<td>A6del</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>AAAAA</td>
<td>AAAAA</td>
<td>A5del</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>AAAAA</td>
<td>AAA</td>
<td>A4del</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Can B-factor explain the effect of mutations? Based on the observations in the previous section, it is clear that mutations among the amino acids can have an impact on the mechanical properties. To clearly understand the
impact of the mutation of one amino acid to another, we first choose certain amino acids from hydrophobic, polar, and charged groups based on the study presented in Sec. Selecting amino acids for mutation study of the Supplemental Information. For this study, we focus on $\sigma_{UTS}$, $\epsilon_f$, $E$, and $T$ as they are all derived from the same stress-strain curve. We want to point out that we did not consider amino acids A and G for the mutation as they both are extremely important for the formation of $\beta$-sheet and amorphous region in MaSp1 spidroin respectively. But we will briefly see the impact of A and G on mechanical properties towards of end of this section. To understand the terms used for studying mutation refer Sec. Mutation study.

From Fig. 3B and 3C, it can be observed that there are reasonably higher chances that Q→<D,K> and S→<K,D> in MaSp1 will increase $\sigma_{UTS}$ whereas V→E in MaSp1 will lead to a decrease in $\sigma_{UTS}$. This can be very well explained using the B-factor values shown in Fig.3A for all amino acids in MaSp1 spidroin. It is clear from Fig.3A that Q and S have a higher B-factor compared to D and K implying that D and K have higher tendencies of forming $\beta$-sheets in MaSp1; leading to an increase in $\sigma_{UTS}$. Conversely, amino acid E has a higher B-factor than V; hence V→E in MaSp1 leads to a decrease in $\sigma_{UTS}$. For $\epsilon_f$ to be higher, high extensibility and low stiffness are typically needed. This can be achieved by mutating to amino acids that have a higher B-factor as that can reduce the $\beta$-sheet regions in the spider silk. Since P has a higher B-factor and amino acid T has a lower B-factor compared to Q, we observe from Fig.3D and 3E that Q→P and Q→T in MaSp1 lead to an increase and decrease in $\epsilon_f$ respectively.

The property $E$ is very similar to $\sigma_{UTS}$ as it also increases with the formation of more $\beta$-sheet regions in the spider silk. Hence, mutating to amino acids with lower B-factor is beneficial for $E$. We observe exactly the same from Fig.3F and 3G that mutating Q→<D,K,N,S,I>, and S→<K,D,T,V,I,P> in MaSp1 leads to an increase in $E$. On the other hand, V→<E,Q,P> leads to a decrease in the property as they have a higher B-factor compared to V.
As discussed in the above section, higher $T$ needs higher $\sigma_{UTS}$ and $\epsilon_f$. Also from Fig.3H and 3I, it is difficult to hypothesize any pattern of $T$ with respect to the B-factor prediction. Then we plot $P(\text{Increase}>20\%)$ and $P(\text{Decrease}>20\%)$ with respect to $\Delta B$-factor as shown in Fig.5A and 5B respectively and also report the correlation (PCC) between the two variables. It is evident from the figures that in MaSp1, mutations that lead to the decrease in B-factor are favorable for toughness. It can also be hypothesized from Fig.3H and 3I that the presence of amino acids F and Y are favorable in MaSp1 for higher $T$.

It has been argued previously that MaSp1 is mostly responsible for the strength of the spider silk whereas MaSp2 is responsible for the elasticity and extensibility. Therefore, we carry out a similar mutation study for MaSp2 spidroin as well, to see if there are constrasting effects of mutations. The two probabilities discussed in Fig.4 are also calculated for mutations in MaSp2 and shown in Fig.4.

From Fig.4B it can be observed that mutations $Q \rightarrow I$, $P \rightarrow I$, and $S \rightarrow D$ in MaSp2 lead to an increase in $\sigma_{UTS}$ due to the decrease in B-factor after mutation. From Fig.4A, Fig.4D, and 4E, it can be inferred that the mutation of $P \rightarrow R$ in MaSp2 leads to a decrease in the B-factor, thus leading to a decrease in $\epsilon_f$ due to the formation of a more ordered region. But not all the mutations that lead to a decrease in the B-factor ($S \rightarrow D$, $P \rightarrow Q$ and $S \rightarrow Y$ in MaSp2), negatively impact $\epsilon_f$. It is seen in the literature that S has a higher propensity of forming $\beta$-sheet than D, and P has a higher probability of being present in the $\beta$-turns than Q. The former observations from the literature can explain why mutations $S \rightarrow D$ and $P \rightarrow Q$ have higher chances of increasing $\epsilon_f$. To explain the impact of $S \rightarrow Y$, we find in the literature that motifs such as GS, GGS, and AS more negatively impact $\epsilon_f$ than GY, GGY, and AY respectively. This explains the reason for the increase in $\epsilon_f$ after $S \rightarrow Y$.

From Fig.4F it can be observed that $E$ increases due to the mutation of $P \rightarrow (Q, D, \text{or } K)$ in MaSp2 as these mutations lead to a decrease in the B-factor. But two of the mutations $S \rightarrow P$ and $V \rightarrow P$ increase the B-factor and $E$ both. Therefore, this trend cannot be explained using the B-factor alone. However, it has been noted in the literature that MaSp2 is a
Proline-rich spidroin and this is important for the structure of MaSp2. Also, it has been observed in the literature\textsuperscript{31} that the increase in the occurrence of V in MaSp2 spidroin negatively impacts $E$. Thus, the increase in $E$ due to V→P mutation is supported by these prior findings.

Similar to MaSp1, we plot $P(\text{Increase}>20\%)$ and $P(\text{Decrease}>20\%)$ with respect to $\Delta B$-factor as shown in Fig.5C and 5D respectively. Fig.5C does not show any correlation between $P(\text{Increase}>20\%)$ and $\Delta B$-factor. But from Fig.5D it can be observed that in MaSp2, an increase in the B-factor after mutation is favored. It is in accordance with the literature,\textsuperscript{59} where it has been shown that amino acid P participates in $\beta$-turns and contributes to the elasticity of the MaSp2 spidroin. It can be hypothesized from Fig.4H and 4I that amino acids D, E, and P are favorable in MaSp2 for higher $T$.

As pointed out above, we did not consider A and G for the mutation study, but we performed A→G and G→A in MaSp1 and MaSp2 to stress test the model. We observe that G→A in MaSp1 strongly favors $\sigma_{UTS}$, $E$, and $T$ with $P(\text{Increase}>20\%)\approx0.6$ whereas A→G does not have a strong impact on any properties. The G→A mutation leads to a decrease in the B-factor; hence increasing the $\sigma_{UTS}$, $E$, and $T$. In MaSp2, A→G and G→A do not have a huge impact on any properties.

In conclusion, the B-factor can very well explain the effect of mutations on mechanical properties in MaSp1. In MaSp2, the B-factor can explain the effect of most of the mutations except for the few mutations involving Proline (P). This is due to the fact that MaSp2 is Proline (P) rich spidroin\textsuperscript{56} with P majorly participating in the $\beta$-turns,\textsuperscript{59} contributing to the elasticity of the dragline silk. Additionally, this study also highlights a few mutations that can improve or worsen a group of properties. For example, Q→D in MaSp1 increases $\sigma_{UTS}$, $T$, and $E$ and S→<any hydrophobic amino acid> increases $T$ and $E$. Conversely, Q→R in MaSp1 worsens $\sigma_{UTS}$ and $T$ and V→E in MaSp1 worsens $\sigma_{UTS}$ and $E$. Similarly, we find that P→Q mutation in MaSp2 has a high chance to increase $\epsilon_f$, $T$, and $E$. 
Figure 3: Effect of mutations in MaSp1. A) Variation of normalized B factor prediction with respect to the amino acids in MaSp1. B,D,F,H) Heatmap showing P(Increase>20%) for $\sigma_{UTS}$, $\epsilon_f$, $E$, $T$ respectively. C,E,G,I) Heatmap showing P(Decrease>20%) for $\sigma_{UTS}$, $\epsilon_f$, $E$, $T$ respectively.
Figure 4: Effect of mutations in MaSp2. A) Variation of normalized B factor prediction with respect to the amino acids in MaSp2. B,D,F,H) Heatmap showing P(Increase>20%) for $\sigma_{UTS}$, $\epsilon_f$, E, T respectively. C,E,G,I) Heatmap showing P(Decrease>20%) for $\sigma_{UTS}$, $\epsilon_f$, E, T respectively.
Figure 5: Variation of probabilities related to $T$ with respect to $\Delta B$-factor. A, B) $P($Increase$>20\%)$ versus $\Delta B$-factor and $P($Decrease$>20\%)$ versus $\Delta B$-factor respectively in MaSp1, C, D) $P($Decrease$>20\%)$ versus $\Delta B$-factor and $P($Decrease$>20\%)$ versus $\Delta B$-factor respectively in MaSp2. The green marker indicates that the higher values are favorable for $T$ and the red marker indicates the higher values are detrimental for $T$. A black line is fitted to the data to indicate the nature of the correlation.

Summary

The outstanding mechanical properties of spider dragline silk surpasses engineered polymers and gives us great inspiration for leveraging sequence-defined properties of proteins in materials science. However, due to its complex multi-phasic hierarchical structure, it is very difficult for any physics-based model to establish a relationship between the primary sequence of various spidroins in spider silk and its mechanical properties. A key contribution of this study is a new DL framework that predicts the mechanical properties of the
dragline spider silk from primary sequence information. Distinctly, our DL framework uses a sequence-length agnostic representation of the primary sequence, making it easier to train a deep-learning model with a small set (180) of training examples. We train and test our model for 5 different mechanical properties of the spider silk and obtain an $R^2$ and PCC in the range of 0.6-0.75 and 0.78-0.88 respectively. The reported values of the $R^2$ are remarkably good considering that process and environmental conditions are not factored into the model. We further show using our DL framework that enriching sequence information with the B-factor prediction tool trained on the protein data bank helps in improving the $R^2$ by an average value of 168% for all properties except supercontraction; highlighting the relationship between the dynamics and mechanical properties. We then use our developed framework to obtain certain important motifs along with their impact on the mechanical properties. We further use the motifs obtained to establish certain mutations that can help in modulating the properties. Through a mutation study, we find that 50% of times the mutation of amino acids that increases $\beta$-sheet propensity increase $\epsilon_f$ but decreases $\sigma_{UTS}$ and $E$. We also find that the length of the poly-Ala motif affects the $\epsilon_{sup}$ through the ratio of amorphous and pol-Ala region length. Increasing the length of the poly-Ala motif decreases the ratio, thereby decreasing the $\epsilon_{sup}$. We uncover that the sequence-defined B-factor values predicted by our model clarifies the impact of mutations on the mechanical properties. For example, mutating from a high B-factor amino acid (Q) to a low B-factor amino acid (D) in MaSp1 helps in $\sigma_{UTS}$, $T$, and $E$. Conversely, mutating from low to high B-factor like V → E in MaSp1 decreases $\sigma_{UTS}$ and $E$. A similar mutation study in MaSp2 shows that P → Q mutation increases $\epsilon_f$, $\sigma_{UTS}$, $T$, and $E$. We also identify that mutation to a higher B-factor in MaSp1 and MaSp2 is detrimental and beneficial to the toughness of spider silk respectively. Collectively, these observations based on our developed framework establishes that B-factor is a useful measure for identifying mutation-based rules for designing stronger protein-based biomaterials. This work sets the stage for two future directions of inquiry. The first is to make a generative model to guide the production of application-specific computationally-designed
fibrous proteins that utilize the model and design rules established in this study. The second direction is to explore representation learning for applications with limited data (the current application is an example) so that we can generate and use more domain-specific sequence-length agnostic representation rather than a generalized one. Our current work shows how effectively a sequence length-agnostic representation can be used for applications with scarce training data. We envision that the computational advances reported herein will be broadly useful for applications where data generation is expensive and time-consuming and where the input space is high dimensional, as in establishing sequence-property relationships in biological systems.

**Data and Code availability**

The authors did not generate any new data for the current study but rather used the data from PDB. The codes and files necessary for training as well as testing the model are available on [https://osf.io/76e8z/?view_only=1322ee32d0204e55a3b65961da42c7f2](https://osf.io/76e8z/?view_only=1322ee32d0204e55a3b65961da42c7f2) and [https://github.com/pandeyakash23/spider_silk_codes](https://github.com/pandeyakash23/spider_silk_codes).

**Acknowledgement**

The primary support for this work came from a National Science Foundation Growing Convergence Research Grant (award no. 2219149). Analysis methods for mutation studies were partially supported by the National Science Foundation’s MRSEC program (DMR-2308691) at the Materials Research Center of Northwestern University. The authors acknowledge support from the Department of Mechanical Engineering at Northwestern University. The authors also acknowledge Jacob Graham and Heather White for their valuable input regarding the preparation of this manuscript, and Dr. Wei (Wayne) Chen for the code review.
Authors Contribution

A.P. and S.K. conceived the idea. A.P. performed all implementations. A.P., S.K., and W.C. contributed to the manuscript writing.

Declaration of Interests

The authors declare no competing interests.

References


(38) Ouellette, R. J.; Rawn, J. D. *Organic chemistry study guide: Key concepts, problems, and solutions*; Elsevier, 2014.


(50) Alves, A. F. C.; Ferreira, B. P.; Pires, F. A. Constitutive modeling of amorphous thermoplastics from low to high strain rates: Formulation and critical comparison em-


Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BfactorSilk7.pdf