communications materials

Article

https://doi.org/10.1038/s43246-024-00519-y

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

Akash Pandey [1], Wei Chen [1] & Sinan Keten [1,2]

Spider dragline silk is known for its exceptional strength and toughness; hence understanding the link between its primary sequence and mechanics is crucial. Here, we establish a deep-learning framework to clarify this link in dragline silk. 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 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 tailored biomaterials.

Spider dragline silk is well known for its extraordinary strength and toughness; higher than other natural silks, Kevlar, and steel[1,2]. While extensibility[3] and tensile strength[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[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[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[13]. Therefore it is paramount to understand the influence of the sequence on properties.

Experimental techniques like X-ray[14], solid-state NMR spectroscopy[15,16] and Raman spectroscopy[17] along with tensile tests[18] have been able to capture the structure-property relationships in the spider silk. Computational models such as Molecular Dynamics (MD) simulations have characterized the importance of β-sheet crystal confinement in the concerted failure of hydrogen bonds, which is partly facilitated by disorder-inducing Prolines[19]. MD has also been used to study the impact of β-sheet nanocrystal size on mesoscale mechanical properties[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[22,23] and to understand the impact of spidroin hydrodynamic flow in the spider duct[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[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 properties. In case of spider silk, the quantitative link between the primary sequence and properties remains elusive. 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[26]. 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 design sets of de novo silk sequences[27] but the results of the generative model remains to be experimentally validated. With the recent advances in the field of machine learning (ML), several groups have worked toward establishing a primary sequence-property relationship in other contexts either using a large sequence-based deep learning (DL) models[28–30] or ML models like XGBoost

[1]Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA. [2]Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL 60208, USA. e-mail: s-keten@northwestern.edu

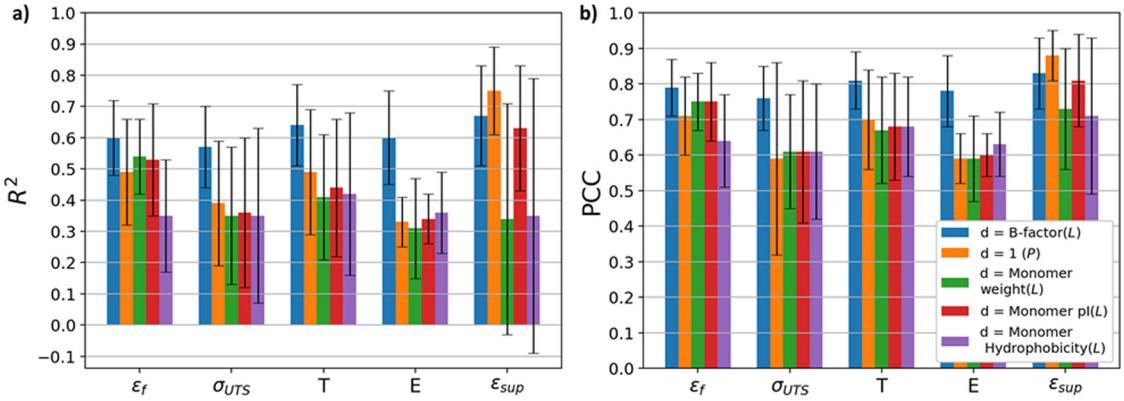

Fig. 1 | Effect of different enriching descriptor *d* on the model performance. a Shows the comparison of R^2, and **b** shows the comparison of PCC. The error bars in the figure indicate ±standard deviation.

and random forest[30]. DL models have the risk of overparameterization when trained on small (<500) data sets. On the other hand, ML models with small number of trainable parameters 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[31]. 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[32].

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.

Results and discussion

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[31]. 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[33]. ProtBERT has been used previously for predicting properties in other contexts. However, this method involved fine-tuning more than a million parameters[29,33]. Fine-tuning would approximately require 500 or more training data[29], making it infeasible for the Spider Silkome Database (SSD) (see "Dataset" section in "Methods"). For this reason, we present a new representation of the sequence that deals with the complexities and data constraints mentioned herein (see "Representation" section in "Methods"). The complete deep-learning framework used for the prediction of mechanical properties of the dragline spider silk is discussed in the "Deep learning model" section of "Methods".

It has been reported in the literature that certain motifs have a higher impact on the properties[31] 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[34]. Therefore, our secondary goal with this study is to build a framework that can identify critical motifs.

Choice of enrichment descriptor of amino acids

To use the representation of the primary sequence discussed in "Representation" section in "Methods" for the prediction of mechanical properties, it is necessary to first fix certain parameters through parametric studies. To establish the maximum distance (max(m)) between a pair of amino acids used to develop the representation, in Supplementary Notes 2 we have established appropriate max(m) values (refer Supplementary Fig. 2 and Supplementary Tables 1 and 2) for all properties. In Supplementary Note 3, we establish a choice between the two representations ($\mathcal{P}$ or $\mathcal{L}$; refer Supplementary Fig. 3) based on the best way to store the descriptors of amino acids in a pair. It is pointed out in the "Deep learning model" section of "Methods" that input features f_i's are downselected from representation $\mathcal{P}$ or $\mathcal{L}$ based on user-defined cutoff values. These cutoff values are heuristically chosen to keep the number of tunable parameters low. In Supplementary Fig. 4, we also show the impact of the cutoff value on the model performance. Next, we study the relative importance of amino acid's enriching descriptor *d*. Figure 1 shows the comparison between different *d* for the best-chosen max(m) and representation for all the properties. The *d*=B-factor with $\mathcal{L}$ representation clearly outperforms all other *d*'s for all properties except ϵ_{sup}, which is evident from P_r and P_c values given in Tables 1 and 2 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 ϵ_{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[32,35], 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)[36] and other constitutive laws[37]. This

Table 1 | P_r = P($R^2 \geq 0.5$) for different choice of enriching descriptor *d*

	Choice of *d*				
	B-factor ($\mathcal{L}$)	1	Monomer weight ($\mathcal{L}$)	Monomer pI value ($\mathcal{L}$)	Hydrophobicity ($\mathcal{L}$)
ϵ_f	**0.80**	0.5	0.63	0.56	0.20
σ_{UTS}	**0.70**	0.29	0.25	0.28	0.30
T	**0.86**	0.48	0.33	0.39	0.38
E	**0.75**	0.02	0.12	0.02	0.14
ϵ_{sup}	0.86	**0.96**	0.33	0.74	0.36

The highest probability (best scenario) is shown in bold.

Table 2 | P_c = P(PCC $\geq$ 0.7) for different choice of enriching descriptor *d*

	Choice of *d*				
	B-factor ($\mathcal{L}$)	1	Monomer weight ($\mathcal{L}$)	Monomer pI value ($\mathcal{L}$)	Hydrophobicity ($\mathcal{L}$)
ϵ_f	**0.87**	0.54	0.73	0.67	0.32
σ_{UTS}	**0.75**	0.34	0.29	0.33	0.32
T	**0.92**	0.5	0.42	0.45	0.44
E	**0.79**	0.06	0.18	0.05	0.22
ϵ_{sup}	0.90	**0.99**	0.57	0.80	0.52

The highest probability (best scenario) is shown in bold.

implies that there is a correlation between molecular mobility and mechanical properties. The outperformance of the model using d = B-factor supports the notion that the segmental molecular mobility in proteins is strongly related to macroscale mechanical properties. For ϵ_{sup}, $d = 1$ performs the best. This can either be due to the lack of data in the case of ϵ_{sup} or the fact that ϵ_{sup} majorly depends on just the occurrence of certain motifs in the spider silk. In the literature[31] it has been shown that the ϵ_{sup} highly depends on the occurrence of poly-Alanine motifs. It is also clear from Fig. 1 and Supplementary Table 3 that the developed DL model works the best for ϵ_{sup} with mean $R^2 > 0.7$. This can be attributed to the fact that ϵ_{sup} follows a uniform distribution as shown in Supplementary Fig. 1 leading to a similar range of output values in the train and test dataset. Furthermore, it's important to recognize that the stress-strain curve of the protein relies on the intricate nanomechanics governing its unraveling process[38]. This complexity is further advanced when multiple proteins (such as spidroins in this scenario) are simultaneously subjected to a pulling force. Consequently, predicting the stress-strain curve from the primary sequence of spider silk is a highly non-linear problem. Therefore we observe $R^2 < 0.7$ for properties obtained from the stress-strain curve. We also present the comparison of results from the task-specific model and the model trained on all the properties (multi-task) simultaneously in Supplementary Note 6. The model architecture for multi-task learning is shown in Supplementary Fig. 5. From the results shown in Supplementary Fig. 6, it is clear that the task-specific model is a better option.

Supplementary Note 5 captures the details of best-performing models for all mechanical properties. Based on the parametric studies presented for different representations, properties (d), and max(m) values (see Supplementary Notes 2–4), the best choice for each mechanical property is given in Supplementary Table 3. Supplementary Table 4 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 spider silk is robust and accurate, considering the high variability in experimental data as discussed in the "Training details" section in "Methods".

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[39] shows that mutating Tyr (Y) to Phe (F) in MaSp1 of biomimetic spider silk decreases ϵ_{sup}. Therefore, in our test dataset, we replace Y with F in MaSp1 and observe a mean decrease of 71% in ϵ_{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 ($\bar{q}_i$) of all the features (f_i) considered for the prediction of the properties using the method discussed in the "Feature importance analysis" section in "Methods". Subsequently, for the features with $\bar{q}_j > 0.1$, the 3 types of motifs (ϕ_w, ϕ_t, and ϕ_b) are identified and their impact (P_m) is quantized as described in the "Method for motif identification and quantifying their effect" section in "Methods". The complete information about the motifs and their impact is presented in Supplementary Note 7. It can be observed from the Supplementary Tables 5–9 that P_m values can take on positive or negative values indicating a positive or negative correlation between the number of motifs (θ_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 ϵ_f as an example as it is one of the motifs with the highest positive impact on ϵ_f. It is evident from Supplementary Table 5 that LVSSGP contributes 0.61% of the max ϵ_f value per θ_n. Now, if we want to increase the ϵ_f by 1.83% of max ϵ_f value, then we need to increase θ_n of LVSSGP by 3. Based on Eq. (6), θ_n can be increased by either increasing the number of motifs or decreasing the number of repeat units in the sequence. It is also very interesting to note that the mean B-factor of LVSSGP motif in MaSp1 is 0.42 which is higher than the mean B-factor of all individual amino acids (Fig. 2a). This suggests that the LVSSGP segment exhibits greater mobility and flexibility within MaSp1, thereby positively impacting strain.

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 Supplementary Tables 5–9 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 MaSp1 increases ϵ_f but it also leads to the increase of SS motif which has a negative impact on ϵ_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 ϵ_f and σ_{UTS}, however, motifs like SS have negative and positive impacts on both properties as shown in Supplementary Tables 5 and 7 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[40]. The nomenclature for substitution is <Res_b><*pos*><Res_a> which means that amino acid Res_b is being replaced by amino acid Res_a at position *pos*. To indicate the deletion/insertion we use <Res_s><Res_s pos>-<Res_e><Res_e pos><*delins*< group of newly inserted *aminoacid* > as the nomenclature where Res_s and Res_e indicate the first and last amino acid deleted.

Based on the observations from Supplementary Tables 5–9, we present some mutation recommendations to increase the properties in Table 3. 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 β-sheets. Based on the

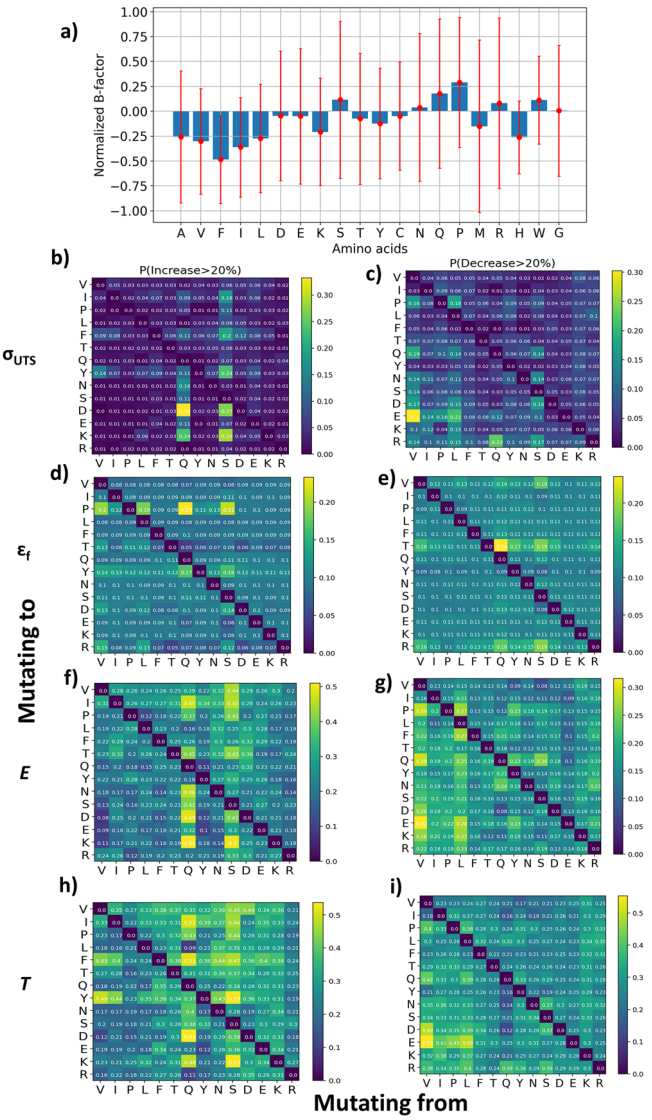

Fig. 2 | Effect of mutations in MaSp1. a Variation of normalized B factor prediction with respect to the amino acids in MaSp1. **b**, **d**, **f**, **h** Heatmap showing P(Increase > 20%) for σ_{UTS}, ϵ_f, E, T respectively. **c**, **e**, **g**, **i** Heatmap showing P(Decrease > 20%) for σ_{UTS}, ϵ_f, E, T respectively.

literature[41], it is understood that the amino acids Ala, Val, Ile, Tyr, Cys, Trp, Phe, and Thr are more likely to be found in β-sheet regions and amino acids Gly, Pro, Asn, and Ser are more likely to be found in the turns connecting two β-sheets. Further considering the likelihood/propensity of the amino acids to form β-sheet, they can be ranked as L, V > A > G[42]. From the perspective of major ampullate spidroins in spider silk, literature[43,44] highlights that Ala and Gly constitute their primary components. Most Ala residues are integrated into the β-sheet structure, underscoring their strong propensity for β-sheet formation. Gly is present in the β-sheets as poly(GA) and in the amorphous regions. Poly-valine is also found to form β-sheets within an amorphous network to improve toughness and strength[45]. These observations reinforce that A and V have higher β-sheet propensity than G. Additionally it is shown experimentally that Proline usually favors a more amorphous structure[46] or is present in β-turns as GPGXX[47]. Building upon the aforementioned insights regarding the propensity of various amino acids to form β-sheets, we will investigate the impact of mutations among different motifs. To examine this effect, we report the ΔP_m value which is defined as the difference between the higher and lower P_m values. In the literature, it has been established that the β-sheet represents a highly ordered domain within spider silk[17] and that the local order of a protein region correlates with its B-factor[48]. Consequently, in the next section, we investigate the impact of mutations on the mechanical properties from the perspective of the B-factor.

Taking the above facts into consideration, we study the effect of different mutations on different mechanical properties. We first start with ϵ_f and hypothesize that it increases 50% of the times when the mutations of the amino acids decrease the β-sheet propensity. Next, we examine σ_{UTS} and E and observe that 50% of times the mutation of amino acids that increase β-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 σ_{UTS} and ϵ_f. Since different mutations favor σ_{UTS} and ϵ_f, we cannot observe a clear trend of T with respect to the β-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.

Upon comparing the motifs documented in the SSD paper[31] with those presented in Supplementary Tables 5–9, we observe notable parallels. Specifically, in the case of ϵ_f, akin to the findings in the SSD paper, we identify that motifs such as SAAAAA and AS exert a negative influence on the property. Conversely, both studies concur that the motif GGAGQ within MaSp1 contributes positively to ϵ_f. Both our research and the SSD paper indicate that motifs QGPSG and YGPGS in MaSp2 impact ϵ_f positively and negatively, respectively. Furthermore, the motif GGPGGYG in MaSp2 affects ϵ_f negatively. In terms of σ_{UTS}, our observation regarding the adverse impact of a poly-Ala segment aligns with the findings in the SSD paper. However, augmenting the poly-Ala segment with Q and G demonstrates a positive effect on the property; for instance, the motif AGQGGA positively influences σ_{UTS}. Both the SSD paper and our study identify that motifs YGGL and GAGQGGY in MaSp1 positively impact σ_{UTS}. Additionally, in MaSp2, both studies ascertain that motifs PGGY and GPGGY positively affect σ_{UTS}. Concerning property E, both works indicate that motifs QGGQGG and AGQGGY within MaSp1 exhibit a positive impact. Furthermore, both studies highlight the recurrence of segments GQGG and GP in several motifs affecting E in MaSp1 and MaSp2 respectively. In the case of property T, both investigations reveal that motifs YGGL and YGG in MaSp1 have a positive influence. Moreover, they observe the segment GQ in many significant motifs for property T in MaSp1, while segments QGP and PG emerge in numerous impactful motifs for T in MaSp2. Overall, our approach has introduced an accelerated framework for identifying significant motifs by prioritizing feature importance, rather than relying on exhaustive motif searches. Additionally, we offer a more structured method to measure the influence of motifs through the computation of P_m.

In the case of supercontraction (ϵ_{sup}), it is evident from the P_m values in Supplementary Table 9 that the larger the length of the poly-Ala motif, the larger the decrease in ϵ_{sup}. This observation is backed by the literature study that shows that the ϵ_{sup} is positively correlated with the amorphous/poly-Ala region length ratio (PCC=0.53)[31]. Increasing the length of even one poly-Ala motif leads to the decreases in amorphous/poly-Ala region length ratio and subsequently the ϵ_{sup}. Building on this, we observe from Table 3 that mutating a larger poly-Ala to a smaller one increases the property. The role of poly-Ala blocks (4 or more Ala) is further highlighted by the presence of several poly-Ala blocks in the motifs reported in Supplementary Tables 5–9. The literature also emphasizes the significance of poly-Ala blocks in facilitating the formation of β-sheets[43,49,50]. The research[49] indicates that a minimum of three poly-Ala blocks is necessary for the formation of β-sheets in spider silk. Beyond three blocks of poly-Ala, an additional increase in the block count enhances crystallinity by 25–39%. It has also been shown experimentally in the literature that poly-Ala enhances the ability of the recombinant spider silk protein to form β-sheet structure, thereby increasing the σ_{UTS} and T[50].

Table 3 | Design rules for each property

Property	Before mutation	After mutation	Mutations	ΔP_m
ϵ_f	QVKT	QVNT	K3N	0.88
	GGAGQQ	GGYGPQ	A3Y, Q5P	0.35
	GGQGPYG	GGPGGYG	Q3P, P5G	0.35
	SAVST	SSGPT	A2_S4*delins*SGP	0.10
	AAAAGY	AAAGGY	A4G	0.26
	QGPGG	QGPSG	G4S	0.15
σ_{UTS}	AAAAAAAAA	AGQGGAGAA	A2_A7*delins*GQGGAG	0.46
	AGQGGA	AGAAAA	Q3_G5*delins*AAA,	0.43
	GAAAA	AAGGA	G7A, A9_A10*delins*GG	
	AGQGGA	AGAAAA	Q3_G5*delins*AAA,	0.43
	GAAAA	AAGGA	G7A, A9_A10*delins*GG	
	GGAGGAGQG	GAGAAAAAA	G2_G5*delins*AGAA,	0.27
	GLGSGQGY	GGAGQGGY	G7_G9*delins*AAA,	
			L11_Q15*delins*GAGQG	
	AAAAAAAGGS	AAAGGYGPGS	A4_G8*delins*GGYGP	0.16
	SGPGGYGPGS	SGPGGYGPAS	G9A	0.11
E	YGSA	YGPA	S3P	1.89
	AGQGGL	ALVHIL	G2_G5*delins*LVHI	1.1
	YQGP	YGAP	Q2_G3*delins*GA	0.60
	QGGYGGY	QGGQGGY	Y4Q	0.57
	VHILGSSSIGQVN	VHILGSANIGQVN	S7_S8*delins*AN	0.55
	SGGQGGY	SQNQGGY	G2_G3*delins*QN	0.32
T	PGGAGGSGPY	PGGQGPYGPG	A4Q, G6_S7*delins*PY,	4.7
	GPAASAA	AAAAAAA	Y10_P12*delins*GAA,	
	AAAGY	AAGGY	S15A, A20G	
	TSSNKLQA	TGAAAAAA	S2_Q7*delins*GAAAAA	0.99
	QGPGGAGQ	QGPSGPGA	G4S, A6P,	0.93
	QGPGGQ	YGPGSQ	Q8_Q9*delins*AY	
	QGPGQQG	QGPGGYG	Q5_Q6*delins*GY,	0.17
	PGGQGP	PGQQGP	G10Q	
ϵ_{sup}	GN	GL	N2L	0.83
	AAA	AA	A3del	0.38
	AAAAAA	AAAAA	A6del	0.37
	AAAAA	AAAA	A5del	0.28
	AAAA	AAA	A4del	0.26

Mutation from column 2 to column 3 increases the property.

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 results shown in Supplementary Fig. 7. For this study, we focus on σ_{UTS}, ϵ_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 β-sheet and amorphous region in MaSp1 spidroin respectively[51]. However, we will briefly discuss the impact of A and G on mechanical properties toward the end of this section. To understand the terms used for studying mutation refer to the "Mutation study" section in "Methods".

In this section, our analysis is focused on the effects of mutations in MaSp1 and MaSp2. However, we have chosen not to include MaSp3 and MiSp in the mutation study, based on the discussion below. We use test datasets for the mutation study as the performance of the DL model on the test dataset reflects its true performance. Out of 203 dragline species, only 22 dragline species have MaSp3 data document for them. We allocate 10% of the total examples as the test dataset, resulting in a statistically insignificant representation of dragline species with MaSp3 within the test dataset. This is the reason the mutation study on MaSp3 is not added to our work. As for MiSp, the reason for its inclusion to generate input features f_i is discussed in the "Representation" section in "Methods" even though it is primarily associated with auxiliary spiral silk[52]. Due to this, some of the input features derived from MiSp might just be a noise leading to some unrealistic mutation results. Hence, we have not included the mutation study for MiSp in our work.

From Fig. 2b, c, it can be observed that there are reasonably higher chances that Q → <D,K> and S → <K,D> in MaSp1 will increase σ_{UTS} whereas V → E in MaSp1 will lead to a decrease in σ_{UTS}. This can be very

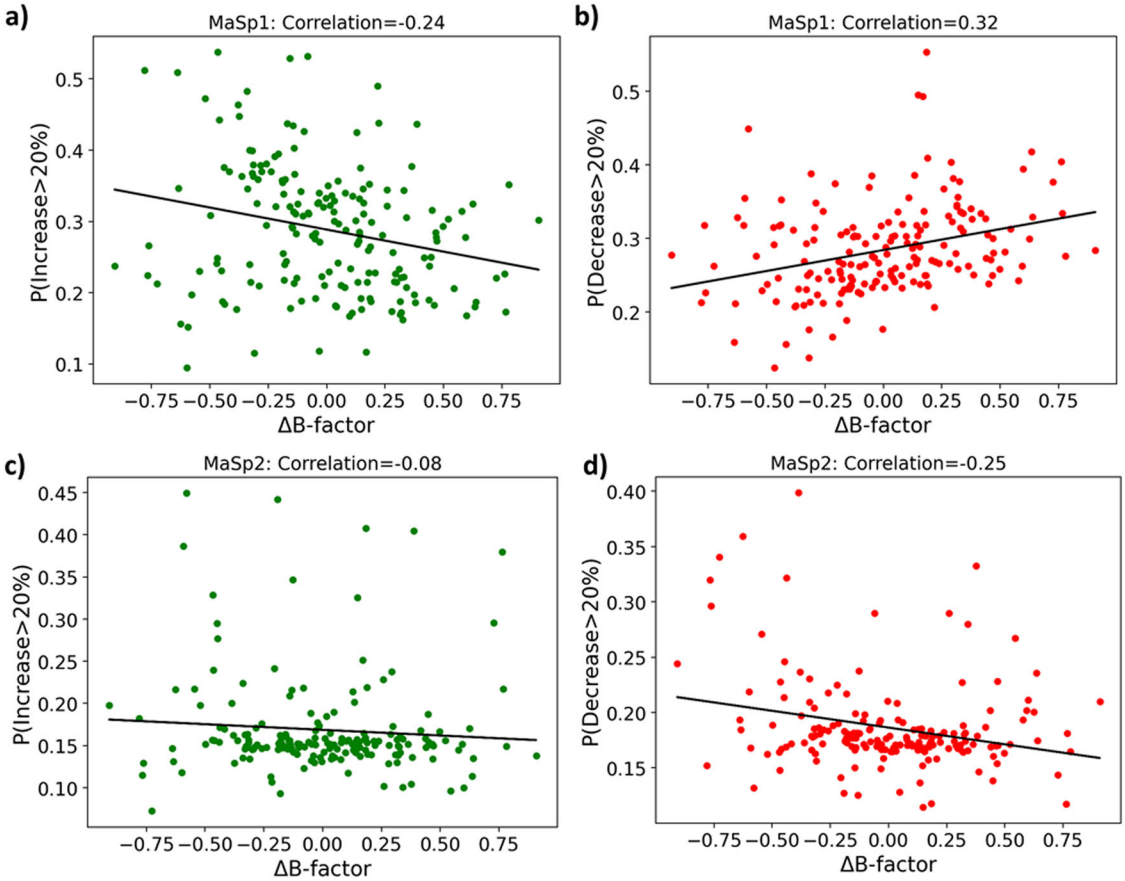

Fig. 3 | Variation of probabilities related to *T* with respect to ΔB-factor. **a**, **b** P(Increase > 20%) versus ΔB-factor and P(Decrease > 20%) versus ΔB-factor respectively in MaSp1, **c**, **d** P(Increase > 20%) versus ΔB-factor and P(Decrease > 20%) versus ΔB-factor respectively in MaSp2. The 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.

well explained using the B-factor values shown in Fig. 2a for all amino acids in MaSp1 spidroin. It is clear from Fig. 2a that Q and S have a higher B-factor compared to D and K. Even though D and K have more chances to exhibit higher B-factor[53], they exhibit lower B-factor than a polar amino acid S in MaSp1. This can be explained by the fact that in MaSp1 the amino acid A and T are the most frequent neighbors of D and K respectively. Amino acid A is mostly available in the crystalline part of the MaSp1 spidroin[13] and amino acid T has a higher propensity of forming β-sheets[54]. Thus implying the presence of D and K in a more crystalline/structured region of the MaSp1. On the other hand, amino acid S has the highest chance of being present next to amino acid G which is majorly present in the amorphous region in MaSp1[13]. This explains the high B-factor of S in MaSp1. Overall the lower B-factor of D and K implies their ability to form crystalline/structured regions in MaSp1; leading to an increase in σ_{UTS}. Conversely, amino acid E has a higher B-factor than V; hence V → E in MaSp1 leads to a decrease in σ_{UTS}. For ϵ_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 β-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. 2d, e that Q → P and Q → T in MaSp1 lead to an increase and decrease in ϵ_f respectively.

The property *E* is very similar to σ_{UTS} as it also increases with the formation of more β-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. 2f, g 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 σ_{UTS} and ϵ_f. Also from Fig. 2h, i it is difficult to hypothesize any pattern of *T* with respect to the B-factor prediction. Then we plot P(Increase > 20%) and P(Decrease > 20%) with respect to ΔB-factor as shown in Fig. 3a, b 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. 2h, i 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[55]. Therefore, we carry out a similar mutation study for MaSp2 spidroin as well, to see if there are contrasting 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 → I, P → I, and S → D in MaSp2 lead to an increase in σ_{UTS} due to the decrease in B-factor after mutation. From Fig. 4a, d, e, it can be inferred that the mutation of P → R in MaSp2 leads to a decrease in the B-factor, thus leading to a decrease in ϵ_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 → D, P → Q and S → Y in MaSp2), negatively impact ϵ_f. It is seen in the literature that S has a higher propensity of forming β-sheet[54] than D, and P has a higher probability of being present in the β-turns than Q[56]. The former observations from the literature can explain why mutations S → D and P → Q have higher chances of increasing ϵ_f. To explain the impact of S → Y, we find in the literature[31] that motifs such as GS, GGS, and AS more negatively impact ϵ_f than GY, GGY, and AY respectively. This explains the reason for the increase in ϵ_f after S → Y.

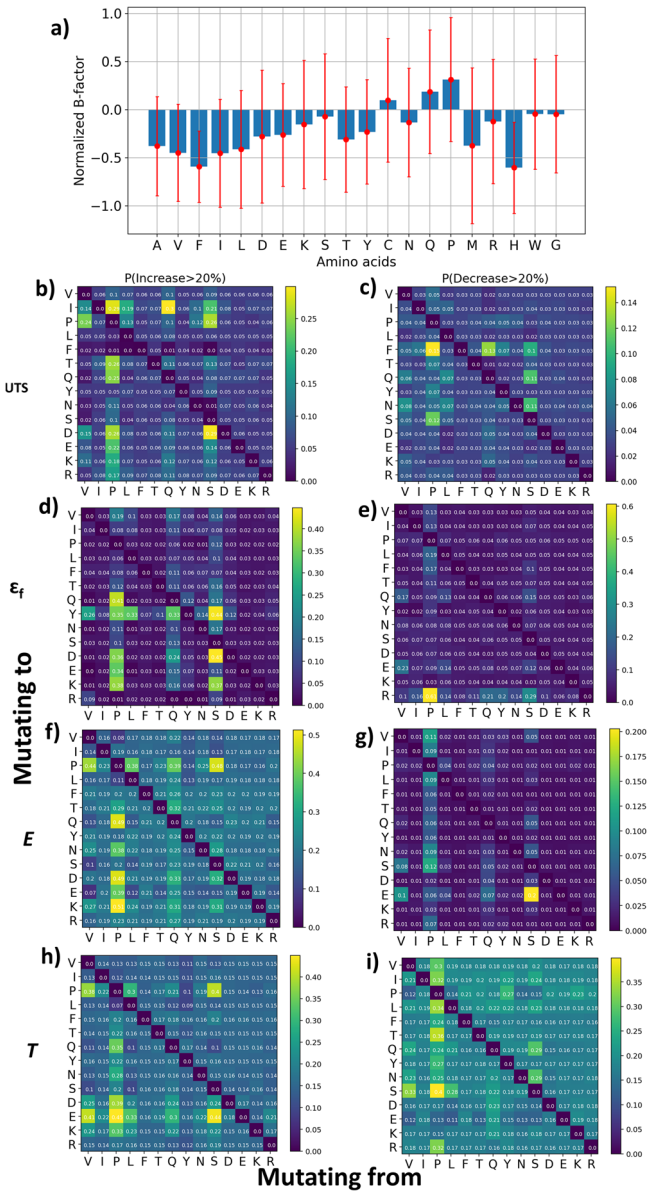

Fig. 4 | Effect of mutations in MaSp2. a Variation of normalized B factor prediction with respect to the amino acids in MaSp2. **b**, **d**, **f**, **h** Heatmap showing P(Increase > 20%) for σ_{UTS}, ϵ_f, E, T respectively. **c**, **e**, **g**, **i** Heatmap showing P(Decrease > 20%) for σ_{UTS}, ϵ_f, E, T respectively.

From Fig. 4f it can be observed that E increases due to the mutation of P → (Q, D, or K) in MaSp2 as these mutations lead to a decrease in the B-factor. But two of the mutations S → P and V → 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[55] that MaSp2 is a Proline-rich spidroin and this is important for the structure of MaSp2. Also, it has been observed in the literature[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(Increase > 20%) and P(Decrease > 20%) with respect to ΔB-factor as shown in Fig. 3c, d respectively. Figure 3c does not show any correlation between P(Increase > 20%) and ΔB-factor. But from Fig. 3d it can be observed that in MaSp2, an increase in the B-factor after mutation is favored. It is in accordance with the literature[57], where it has been shown that amino acid P participates in β-turns and contributes to the elasticity of the MaSp2 spidroin. It can be hypothesized from Fig. 4h, i 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 σ_{UTS}, E, and T with P(Increase > 20%) ≈ 0.6 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 σ_{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[55] with P majorly participating in the β-turns[57], 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 σ_{UTS}, T, and E and S → < any hydrophobic amino acid > increases T and E. Conversely, Q → R in MaSp1 worsens σ_{UTS} and T and V → E in MaSp1 worsens σ_{UTS} and E. Similarly, we find that P → Q mutation in MaSp2 has a high chance to increase ϵ_f, T, and E.

Conclusion

The outstanding mechanical properties of spider dragline silk surpass engineered polymers and give 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 0.76–0.88 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 probability P_r by an average value of 0.39 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 nearly half of the times the mutation of amino acids that increases β-sheet propensity increase ϵ_f but decreases σ_{UTS} and E. We also find that the length of the poly-Ala motif affects the ϵ_{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 ϵ_{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 σ_{UTS}, T, and E. Conversely, mutating from low to high B-factor like V → E in MaSp1 decreases σ_{UTS} and E. A similar mutation study in MaSp2 shows that P → Q mutation increases ϵ_f, σ_{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 establish 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.

Methods

Dataset

Kazuharu et al.[31] have carried out extensive work to test and document various mechanical properties of dragline silks in the so-called Spider Silkome 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 (ϵ_f), tensile strength (σ_{UTS}), toughness (T), Young's modulus (E), and supercontraction (ϵ_{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 ϵ_{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 ϵ_{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 *sequence* will be used to indicate the primary sequence of the repetitive part of the spider silk. We discuss the distribution of all the properties under consideration in Supplementary Note 1. It is evident from Supplementary Fig. 1 that all the properties except ϵ_{sup} follow skew-normal distribution whereas ϵ_{sup} exhibits uniform distribution.

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 spidroins[1,31,58]. In the SSD paper[31], it is shown that certain MiSp sequences are indistinguishable from major ampullate spidroins and vice versa. This assertion is further supported by the SSD paper, which demonstrates an overlap between major and minor spidroins when clustered based on repetitive and N-terminal sequences. Such overlaps can introduce ambiguity in labeling the spidroins. Hence, we also include MiSp for training, despite its primary occurrence in auxiliary spiral silk[52]. We build our representation for each spidroin based upon the Quasi Residue Couple (RC) concept[59]. The RC representation converts the primary sequence information of a spidroin of sequence length (N) 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. From this point forward, this distance between two residues in the primary sequence will be denoted by m. For instance, adjacent residues in sequence space form a 20×20 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 20×20 matrix can be expressed in terms of sequence length (N), amino acid descriptor (d) using Eqs. (1a) and (1b):

$$\mathcal{P}_m^d(i,j) = \frac{1}{N-m}\sum_{n=1}^{N-m} H_{i,j}(n, n+m, d),\ \ i,j \in [1 \ldots 20] \tag{1a}$$

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

The 20×20 representation for each m value (1,2,…, max(m)) can be stacked together to get a 3D representation of dimension $20 \times 20 \times$ 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, hydrophobicity[60], monomer charges[61], monomer weight[61]. 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[60]. 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[61]. 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[53]. In the literature, it has been shown that the Debye-Waller parameter is strongly correlated with the mechanical properties[62]. 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[53]. It should be noted that the deep-learning model[53] 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 the "Results and discussion" section, we discuss the effect of max(m) on the model quality. This distinction makes our representation advantageous over other models like ProtBERT[33] 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[29,63]. Since we have less data (refer to "Dataset" section in "Methods"), 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 $\mathcal{P}$ 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 ($\mathcal{L}$) as shown in Eqs. (2a) and (2b) which has an extra dimension compared to $\mathcal{P}$ to store the values of d_i and d_j separately as:

$$\mathcal{L}_m^d(i,j,k) = \frac{1}{N-m}\sum_{n=1}^{N-m} H_{i,j}(n, n+m, k, d),\ \ i,j \in [1 \ldots 20],\ \ k \in [1,2] \tag{2a}$$

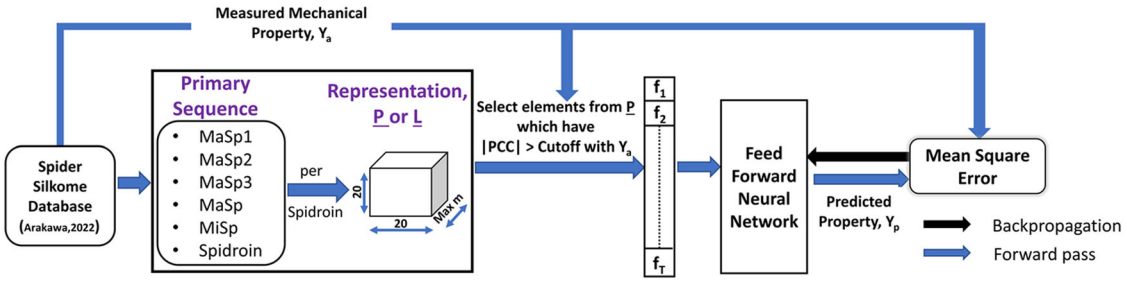

Fig. 5 | Model architecture for spider silk property prediction with representation $\mathcal{P}$. This figure showcases the entire deep-learning framework for predicting dragline spider silk properties, highlighting each step from data preparation and feature engineering to neural network design and property prediction.

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

This will resolve the numerical issue for the pairs of i and j with contrastive d values in $\mathcal{P}$. It is important to note that for $d = 1$, $\mathcal{P}$, k = 1 in $\mathcal{L}$, and k = 2 in $\mathcal{L}$ are exactly the same; hence for $d = 1$, we will be using only $\mathcal{P}$ 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 Eqs. (1a) and (2a). As discussed in the "Representation" section in "Methods", m represents the gap between amino acid i and j in the primary sequence as in Eqs. (1a) and (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 6 types of spidroins in our study (refer "Representation" section in "Methods"), for max(m) equal to 6, the total number of elements from $\mathcal{P}_{20*20*6}$ and $\mathcal{L}_{20*20*6*2}$ are size ($\mathcal{P}$)*6 (14,400 features) and size($\mathcal{L}$)*6 (28,800 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. 5. 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 $\mathbf{F}$ can be written as a 2D tensor as shown in Eq. (3). Equation (3) indicates a feature as f_j^i where i indicates the example number and j represents the jth feature in the ith example. This input vector is then fed into a feed-forward neural network (FFNN) which maps $\mathbf{f}$ 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. 5. 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. This is in accordance with the literature[64–67] where researchers have used different DL models for different properties emerging from the same source:

$$\mathbf{F} = [\mathbf{f}^1, \mathbf{f}^2, \cdots, \mathbf{f}^{N_e}] = \begin{bmatrix} f_1^1 & f_1^1 & \cdots & f_1^{N_e} \\ \vdots & \vdots & \vdots & \vdots \\ f_j^1 & f_j^1 & \cdots & f_j^{N_e} \\ \vdots & \vdots & \vdots & \vdots \\ f_T^1 & f_T^1 & \cdots & f_T^{N_e} \end{bmatrix} \tag{3}$$

$$\text{MSE} = \frac{\sum_{i=1}^{N} (Y_a - Y_p)^2}{N} \tag{4}$$

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[68]. 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[69,70]. Experimental conditions such as humidity or spinning process parameters like reeling speed can have a significant impact on the measured properties, which explains the high variability[71,72]. 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.

Table 4 | Details of the size of the train and test dataset

Property	Train data	Test data
ϵ_f, σ_{UTS}, T, E	182	21
ϵ_{sup}	39	10

Table 5 | Max values of various mechanical properties obtained from the literature for the dragline spider silk[31]

Property	Max value
ϵ_f	65.3%
T	0.425 GJ/m^3
σ_{UTS}	3.33 GPa
E	37 GPa
ϵ_{sup}	49.2%

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. The size of train and test data in each seed is given in Table 4. 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) $P_r = \mathrm{P}(R^2 \geq 0.5)$, and (2) $P_c = \mathrm{P}(\mathrm{PCC} \geq 0.7)$.

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[73]. 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:

$$a_k = \arg\max_{AA} P(AA|a_1, \ldots a_{k-1}, a_{m+1}) \quad k \in [2, .., m], \text{and } AA \in \text{ 20 amino acids} \tag{5}$$

The conditional probability in Eq. (5) is calculated using all the examples obtained in the "Dataset" section in "Methods". Basically, to fill an empty position, Eq. (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[31] 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 a 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[31]. Spider silk spidroins consist of several amorphous and crystalline segments in a sequence. The reference literature[31] 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[31] 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 (θ_N) as the normalized number of motifs as given by Eq. (6):

$$\theta_N = \frac{\#\text{ of motifs in the sequence}}{\#\text{ repeat units}} \tag{6}$$

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 $\mathbf{Y}_p$.
- For each test example, calculate the θ_N value for the motif under the assessment and store it in an array θ_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 $\mathbf{Y}_C$.
- Calculate the percentage impact (P_m) of the motif using Eqs. (7a) and (7b) sequentially. The max property value in Eq. (7b) for various mechanical properties is obtained from the SSD paper[31] and given in Table 5.

$$I_m = \text{Mean}\left(\frac{\mathbf{Y}_p - \mathbf{Y}_c}{\theta_N}\right) \tag{7a}$$

$$P_m = \frac{100 * I_m}{\text{Max Property value}} \tag{7b}$$

Above, the entire process is discussed with the motifs obtained from Eq. (5) using *all* the examples obtained in the "Dataset" section in "Methods". 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[31] 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. (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 $\overline{q}_i > 0.1$, we identify 3 types of motifs using Eq. (5) on the: (1) whole dataset (ϕ_w), (2) dragline spider silk with the highest 10% property (ϕ_t), and (3) dragline spider silk with lowest 10% property (ϕ_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(Increase > 20%), and (2) the probability that the mutation will lead to an *decrease* in the property by 20% (P(Decrease > 20%). Further to study if B-factor can explain the effect mutation, we introduce a variable ΔB-factor which is equal to the (Res_2's mean B-factor − Res_1's mean B-factor) in $Res_1 \rightarrow Res_2$ mutation. The process of selecting amino acids for the mutation study is given in Supplementary Note 8.

Code availability

The codes and files necessary for training as well as testing the model are available on https://osf.io/76e8z/?view_only=1322ee32d0204e55a3b65961da42c7f2 and https://github.com/pandeyakash23/spider_silk_codes.

Received: 20 December 2023; Accepted: 8 May 2024;
Published online: 25 May 2024

References

1. Gu, Y. et al. Mechanical properties and application analysis of spider silk bionic material. *e-Polym.* **20**, 443–457 (2020).
2. Heslot, H. Artificial fibrous proteins: a review. *Biochimie* **80**, 19–31 (1998).
3. Vollrath, F. & Edmonds, D. T. Modulation of the mechanical properties of spider silk by coating with water. *Nature* **340**, 305–307 (1989).
4. Vollrath, F. Strength and structure of spiders' silks. *Rev. Mol. Biotechnol.* **74**, 67–83 (2000).
5. Perez-Rigueiro, J., Viney, C., Llorca, J. & Elices, M. Mechanical properties of silkworm silk in liquid media. *Polymer* **41**, 8433–8439 (2000).
6. Andersson, M., Johansson, J. & Rising, A. Silk spinning in silkworms and spiders. *Int. J. Mol. Sci.* **17**, 1290 (2016).
7. Li, J. et al. Bi-terminal fusion of intrinsically-disordered mussel foot protein fragments boosts mechanical strength for protein fibers. *Nat. Commun.* **14**, 2127 (2023).
8. Li, J. et al. Microbially synthesized polymeric amyloid fiber promotes β-nanocrystal formation and displays gigapascal tensile strength. *ACS Nano* **15**, 11843–11853 (2021).
9. Dinjaski, N. & Kaplan, D. L. Recombinant protein blends: silk beyond natural design. *Curr. Opin. Biotechnol.* **39**, 1–7 (2016).
10. Bowen, C. H. et al. Recombinant spidroins fully replicate primary mechanical properties of natural spider silk. *Biomacromolecules* **19**, 3853–3860 (2018).
11. Roberts, E. G. et al. Fabrication and characterization of recombinant silk-elastin-like-protein (SELP) fiber. *Macromol. Biosci.* **18**, 1800265 (2018).
12. Salehi, S., Koeck, K. & Scheibel, T. Spider silk for tissue engineering applications. *Molecules* **25**, 737 (2020).
13. Römer, L. & Scheibel, T. The elaborate structure of spider silk: structure and function of a natural high performance fiber. *Prion* **2**, 154–161 (2008).
14. Asakura, T. Structure and dynamics of spider silk studied with solid-state nuclear magnetic resonance and molecular dynamics simulation. *Molecules* **25**, 2634 (2020).
15. Simmons, A., Ray, E. & Jelinski, L. W. Solid-state 13C NMR of nephila clavipes dragline silk establishes structure and identity of crystalline regions. *Macromolecules* **27**, 5235–5237 (1994).
16. Simmons, A. H., Michal, C. A. & Jelinski, L. W. Molecular orientation and two-component nature of the crystalline fraction of spider dragline silk. *Science* **271**, 84–87 (1996).
17. Wang, Q. et al. Protein secondary structure in spider silk nanofibrils. *Nat. Commun.* **13**, 4329 (2022).
18. Wang, M., Yang, Z., Wang, C. & Si, M. Exploration of the protein conformation and mechanical properties of different spider silks. *J. Mol. Struct.* **1270**, 133933 (2022).
19. Keten, S. & Buehler, M. J. Nanostructure and molecular mechanics of spider dragline silk protein assemblies. *J. R. Soc. Interface* **7**, 1709–1721 (2010).
20. Nova, A., Keten, S., Pugno, N., Redaelli, A. & Buehler, M. Molecular and nanostructural mechanisms of deformation, strength and toughness of spider silk fibrils. *Nat. Preced.* **10**, 1 (2010).
21. Keten, S., Xu, Z., Ihle, B. & Buehler, M. J. Nanoconfinement controls stiffness, strength and mechanical toughness of β-sheet crystals in silk. *Nat. Mater.* **9**, 359–367 (2010).
22. Rim, N.-G. et al. Predicting silk fiber mechanical properties through multiscale simulation and protein design. *ACS Biomater. Sci. Eng.* **3**, 1542–1556 (2017).
23. Yamane, T., Umemura, K., Nakazawa, Y. & Asakura, T. Molecular dynamics simulation of conformational change of poly(ala-gly) from silk I to silk II in relation to fiber formation mechanism of bombyx mori silk fibroin. *Macromolecules* **36**, 6766–6772 (2003).
24. Herrera Rodriguez, A. M. et al. The role of hydrodynamic flow in the self-assembly of dragline spider silk proteins. *bioRxiv* https://doi.org/10.1101/2022.10.25.513683 (2022).
25. Lin, S. et al. Predictive modelling-based design and experiments for synthesis and spinning of bioinspired silk fibres. *Nat. Commun.* **6**, 6892 (2015).
26. Kim, Y., Yoon, T., Park, W. B. & Na, S. Predicting mechanical properties of silk from its amino acid sequences via machine learning. *J. Mech. Behav. Biomed. Mater.* **140**, 105739 (2023).
27. Lu, W., Kaplan, D. L. & Buehler, M. J. Generative modeling, design, and analysis of spider silk protein sequences for enhanced mechanical properties. *Adv. Funct. Mater.* **34**, 2311324 (2024).
28. Yu, C.-H. et al. Colgen: an end-to-end deep learning model to predict thermal stability of de novo collagen sequences. *J. Mech. Behav. Biomed. Mater.* **125**, 104921 (2022).
29. Khare, E., Gonzalez-Obeso, C., Kaplan, D. L. & Buehler, M. J. Collagentransformer: end-to-end transformer model to predict thermal stability of collagen triple helices using an NLP approach. *ACS Biomater. Sci. Eng.* **8**, 4301–4310 (2022).
30. Detlefsen, N. S., Hauberg, S. & Boomsma, W. Learning meaningful representations of protein sequences. *Nat. Commun.* **13**, 1914 (2022).
31. Arakawa, K. et al. 1000 spider silkomes: linking sequences to silk physical properties. *Sci. Adv.* **8**, eabo6043 (2022).
32. Xia, W. et al. Energy-renormalization for achieving temperature transferable coarse-graining of polymer dynamics. *Macromolecules* **50**, 8787–8796 (2017).
33. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
34. Zaslavsky, E. & Singh, M. A combinatorial optimization approach for diverse motif finding applications. *Algorithms Mol. Biol.* **1**, 1–13 (2006).

35. Riggleman, R. A. & de Pablo, J. J. Antiplasticization and local elastic constants in trehalose and glycerol mixtures. *J. Chem. Phys.* **128**, 224504 (2008).
36. Müller-Plathe, F. Group interaction modelling of polymer properties. By David Porter, Marcel Dekker, New York 1995, X, 512 pp., ISBN 0-8247-9599-7. *Adv. Mater.* **8**, 542–542 (1996).
37. 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 employing an optimization-based parameter identification. *Int. J. Solids Struct.* **273**, 112258 (2023).
38. Mora, M. & Garcia-Manyes, S. Protein nanomechanics: the power of stretching. *Europhys. N.* **51**, 24–27 (2020).
39. Greco, G. et al. Tyrosine residues mediate supercontraction in biomimetic spider silk. *Commun. Mater.* **2**, 43 (2021).
40. Ogino, S. et al. Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. *J. Mol. Diagn.* **9**, 1–6 (2007).
41. Zvelebil, M. & Baum, J. O. *Understanding Bioinformatics* (Garland Science, 2007).
42. Nowick, J. S. & Insaf, S. The propensities of amino acids to form parallel β-sheets. *J. Am. Chem. Soc.* **119**, 10903–10908 (1997).
43. Van Beek, J. D., Hess, S., Vollrath, F. & Meier, B. The molecular structure of spider dragline silk: folding and orientation of the protein backbone. *Proc. Natl Acad. Sci.* **99**, 10266–10271 (2002).
44. Craig, H. C., Piorkowski, D., Nakagawa, S., Kasumovic, M. M. & Blamires, S. J. Meta-analysis reveals materiomic relationships in major ampullate silk across the spider phylogeny. *J. R. Soc. Interface* **17**, 20200471 (2020).
45. Chan, N. J.-A. et al. Spider-silk inspired polymeric networks by harnessing the mechanical potential of β-sheets through network guided assembly. *Nat. Commun.* **11**, 1630 (2020).
46. Savage, K. N. & Gosline, J. M. The role of proline in the elastic mechanism of hydrated spider silks. *J. Exp. Biol.* **211**, 1948–1957 (2008).
47. Jenkins, J. E. et al. Solid-state NMR evidence for elastin-like β-turn structure in spider dragline silk. *Chem. Commun.* **46**, 6714–6716 (2010).
48. Radivojac, P. et al. Protein flexibility and intrinsic disorder. *Protein Sci.* **13**, 71–80 (2004).
49. Rabotyagova, O. S., Cebe, P. & Kaplan, D. L. Role of polyalanine domains in β-sheet formation in spider silk block copolymers. *Macromol. Biosci.* **10**, 49–59 (2010).
50. Tsuchiya, K., Ishii, T., Masunaga, H. & Numata, K. Spider dragline silk composite films doped with linear and telechelic polyalanine: effect of polyalanine on the structure and mechanical properties. *Sci. Rep.* **8**, 3654 (2018).
51. Malay, A. D., Craig, H. C., Chen, J., Oktaviani, N. A. & Numata, K. Complexity of spider dragline silk. *Biomacromolecules* **23**, 1827–1840 (2022).
52. Chen, G. et al. Full-length minor ampullate spidroin gene sequence. *PLoS ONE* **7**, e52293 (2012).
53. Pandey, A., Liu, E., Graham, J., Chen, W. & Keten, S. B-factor prediction in proteins using a sequence-based deep learning model. *Patterns* **4**, 100805 (2023).
54. Fujiwara, K., Toda, H. & Ikeguchi, M. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC Struct. Biol.* **12**, 1–15 (2012).
55. dos Santos-Pinto, J. R. A., Arcuri, H. A., Lubec, G. & Palma, M. S. Structural characterization of the major ampullate silk spidroin-2 protein produced by the spider nephila clavipes. *Biochim. Biophys. Acta Proteins Proteom.* **1864**, 1444–1454 (2016).
56. Halfmann, R. et al. Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins. *Mol. cell* **43**, 72–84 (2011).
57. Malay, A. D., Arakawa, K. & Numata, K. Analysis of repetitive amino acid motifs reveals the essential features of spider dragline silk proteins. *PLoS ONE* **12**, e0183397 (2017).
58. Porter, D., Vollrath, F. & Shao, Z. Predicting the mechanical properties of spider silk as a model nanostructured polymer. *Eur. Phys. J. E* **16**, 199–206 (2005).
59. Nanni, L., Lumini, A. & Brahnam, S. An empirical study on the matrix-based protein representations and their combination with sequence-based approaches. *Amino Acids* **44**, 887–901 (2013).
60. Monera, O. D., Sereda, T. J., Zhou, N. E., Kay, C. M. & Hodges, R. S. Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. *J. Pept. Sci.* **1**, 319–329 (1995).
61. Ouellette, R. J. & Rawn, J. D. *Organic Chemistry Study Guide: Key Concepts, Problems, and Solutions* (Elsevier, 2014).
62. Zheng, X., Guo, Y., Douglas, J. F. & Xia, W. Competing effects of cohesive energy and cross-link density on the segmental dynamics and mechanical properties of cross-linked polymers. *Macromolecules* **55**, 9990–10004 (2022).
63. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30 (2017).
64. Mohapatra, P., Pandey, A., Islam, B. & Zhu, Q. Speech disfluency detection with contextual representation and data distillation. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, 19–24 (2022).
65. Mohapatra, P., Pandey, A., Sui, Y. & Zhu, Q. Effect of attention and self-supervised speech embeddings on non-semantic speech tasks. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9511–9515 (2023).
66. Mohapatra, P., Islam, B., Islam, M. T., Jiao, R. & Zhu, Q. Efficient stuttering event detection using siamese networks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (IEEE, 2023).
67. Sukegawa, S. et al. Evaluation of multi-task learning in deep learning-based positioning classification of mandibular third molars. *Sci. Rep.* **12**, 684 (2022).
68. Casella, G. & Berger, R. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences (Thomson Learning, 2002).
69. Greco, G., Mirbaha, H., Schmuck, B., Rising, A. & Pugno, N. M. Artificial and natural silk materials have high mechanical property variability regardless of sample size. *Sci. Rep.* **12**, 3507 (2022).
70. Cook, D., Julias, M. & Nauman, E. Biological variability in biomechanical engineering research: significance and meta-analysis of current modeling practices. *J. Biomech.* **47**, 1241–1250 (2014).
71. Madsen, B., Shao, Z. Z. & Vollrath, F. Variability in the mechanical properties of spider silks on three levels: interspecific, intraspecific and intraindividual. *Int. J. Biol. Macromol.* **24**, 301–306 (1999).
72. Agnarsson, I. et al. Supercontraction forces in spider dragline silk depend on hydration rate. *Zoology* **112**, 325–331 (2009).
73. Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**, 1–81 (2019).

Acknowledgements

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.

Author contributions

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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43246-024-00519-y.

Correspondence and requests for materials should be addressed to Sinan Keten.

Peer review information *Communications Materials* thanks Keiji Numata and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jet-Sing Lee.