ADVANCED FUNCTIONAL MATERIALS
www.afm-journal.de

A Physically-Based Machine Learning Approach Inspires an Analytical Model for Spider Silk Supercontraction

Vincenzo Fazio, Ali D. Malay, Keiji Numata, Nicola M. Pugno, and Giuseppe Puglisi**

Scientific and industrial interest in spider silk stems from its remarkable properties, including supercontraction—an activation effect induced by wetting. Understanding the underlying molecular scale mechanisms is then also crucial for biomimetic applications. In this study, it is illustrated how the effective integration of physically-based machine learning with scientific interpretations can lead to significant physical insights and enhance the predictive power of an existing microstructure-inspired model. A symbolic data modeling technique, known as Evolutionary Polynomial Regression (EPR), is employed, which integrates regression capabilities with the genetic programming paradigm, enabling the derivation of explicit analytical formulas for deducing structure-function relationships emerging across different scales, to investigate the impact of protein primary structures on supercontraction. This analysis is based on recent multiscale experimental data encompassing a diverse range of scales and a wide variety of different spider silks. Specifically, this analysis reveals a correlation between supercontraction and the repeat length of the MaSp2 protein as well as the polyalanine region of MaSp1. Straightforward microstructural interpretations that align with experimental observations are proposed. The MaSp2 repeat length governs the cross-links that stabilize amorphous chains in dry conditions. When hydrated, these cross-links are disrupted, leading to entropic coiling and fiber contraction. Furthermore, the length of the polyalanine region in MaSp1 plays a critical role in supercontraction by restricting the extent of crystal misalignment necessary to accommodate the shortening of the soft regions. The validation of the model is accomplished by comparing experimental data from the Silkome database with theoretical predictions derived from both the machine learning and the proposed model. The enhanced model offers a more comprehensive understanding of supercontraction and establishes a link between the primary structure of silk proteins and their macroscopic behavior, thereby advancing the field of biomimetic applications.

1. Introduction

Due to their extraordinary properties, spider silks are among the most intensively studied materials, particularly in the field of biomimetics.[1] The advent of increasingly sophisticated experimental techniques over the last few decades has provided a deeper understanding of the complex, multiscale, and hierarchical structure underlying their remarkable mechanical behavior from both chemical and structural perspectives. However, many important phenomena governing their response to loading history, rate, temperature, and humidity effects remain unclear,[2] especially when considering multiscale effects.

A striking effect observed in spider silks is the so-called *supercontraction*, first reported in 1977,[3] which occurs when a spider dragline silk thread is exposed to humidity. Experiments show that, depending on the silk composition, there is a Relative Humidity (RH) threshold beyond which the fiber contracts to a length that may reach half of its initial (dry) length. This phenomenon also presents opportunities for employing supercontraction in the field of actuation.[4]

The amount of contraction depends on several factors, including spider species,[5] environmental conditions,[6] and the rate of hydration.[7]

The supercontraction of spider silk fibers, observed in biomimetic silks as well,[8] is a crucial phenomenon that

V. Fazio, G. Puglisi
Department of Civil Environmental Land Building Engineering and Chemistry
Polytechnic University of Bari
via Orabona 4, Bari 70125, Italy
E-mail: giuseppe.puglisi@poliba.it

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adfm.202420095

DOI: 10.1002/adfm.202420095

A. D. Malay, K. Numata
Biomacromolecules Research Team
RIKEN Center for Sustainable Resource Science
2-1 Hirosawa, Wako, Saitama 351-0198, Japan

K. Numata
Laboratory for Biomaterial Chemistry
Department of Material Chemistry
Graduate School of Engineering
Kyoto University
Kyoto-Daigaku-Katsura, Nishikyo-ku, Kyoto 615-8510, Japan

is considered closely linked to their outstanding mechanical properties.[9]

Therefore, fully understanding and modeling the molecular mechanisms underlying supercontraction, is crucial for engineering advanced biomimetic silk materials with tailored properties for various applications. Controlling supercontraction is key to unlocking the full potential of artificial spider silk as a high-performance, responsive biomaterial. An attempt to tune the artificial spider silks supercontraction was conducted by Greco et al.[8] based on protein engineering.

At the macromolecular scale, spider dragline silks are composed of protein chains giving rise to an amorphous matrix and pseudo-crystalline regions made up principally of polyalanine β-sheets[9,10] with dimensions ranging from 1 and 10 nm,[11] mostly aligned along the fiber direction.[12] It should be noted that the chemical and structural composition varies with the different types of silks produced by various glands[13] and different species. The main structural protein components of the different spider silk types are called spidroins. Here, we focus on the most performing and extensively studied type of silk, known as dragline silk, also referred to as Major Ampullate silk, which is mainly composed of two spidroin subtypes, the Major Ampullate Spidroin 1 and 2 (MaSp1 and MaSp2). The overall sequence architectures of these two spidroins are similar, featuring a highly repetitive core region flanked by small N-terminal and C-terminal domains (NTD and CTD, respectively, see **Figure 1**). The repetitive regions, which make up 90% of the primary structure, consist of alternating runs of polyalanine and multiple glycine-rich motifs arranged in tandem.

In recent years, research facilitated by advanced proteomics and sequencing technologies has revealed that the traditional two-component model of dragline silk, which included only MaSp1 and MaSp2, is overly simplistic to account for the true complexity of dragline silk. Further investigations have identified additional subtypes, such as MaSp3 in several species,[14,15] and up to eight distinct spidroin sub-types in *Trichonephila clavipes*.[16] Recent findings also provide evidence for spidroin cross-expression, where spidroins typically associated with one type of silk are found in another. For example, AcSp1 spidroin, which is usually found in prey-wrapping silk, has also been detected in dragline silk fibers.[14,17] Furthermore, non-spidroin components and post-translational modifications play crucial roles in the folding and function of these proteins.[18]

Despite this complexity, the main research effort is still mainly focused on MaSp1 and MaSp2, widely recognized as the primary proteins constituting spider silk and for which the largest amount of experimental data is available in the Silkome database.[19] Such a database, comprising experiments on silks from 1000 different spiders, recently proposed by some of the authors of this paper, provides the foundation for the symbolic data modeling analysis detailed later.

As anticipated, both MaSp1 and MaSp2 repetitive sequences consist of alternating sections of polyalanine (poly-Ala) and glycine-rich (Gly-rich) regions. The primary difference between the two lies in the Gly-rich region of MaSp2, which is enriched with proline and typically Gln–Gln motifs, absent in MaSp1, where the Gly-rich regions are generally shorter and more hydrophobic.[20] As both MaSp types contain poly-Ala sequences, it is reasonable to conclude that both contribute to the crystalline content of the silk fibers. However, it is not merely the presence of polyalanine segments or the formation of β-sheets, that predominantly governs the behavior of spider silk, but the development of aligned, 3D β-sheet crystallites as found in the silk fibers. In this context, the extent of crystallization is typically higher for MaSp1 due to several factors. First, MaSp1 contains a higher relative abundance of residues in its poly-Ala sections compared to the Gly-rich regions. Second, the MaSp1 Gly-rich region includes more hydrophobic elements (e.g., Leu residues), which may contribute to enhanced crystallinity and alignment of neighboring chains. Moreover, the role of proline induces the lower crystallinity observed in MaSp2, as evidenced by atomistic simulations on MaSp1 and MaSp2 protein segments.[11] Indeed, the abundance and regular spacing of proline residues in MaSp2 make it less prone to molecular alignment with other spidroin molecules at polyalanine sites, as the proline residues introduce a pronounced kink in the extended backbone chain, preventing the formation of stable β-sheet crystals.[21] This irregularity, caused by frequent proline residues, reduces the structural integrity of the crystals and results in their termination into a disordered matrix, leading to macromolecules with weaker crystal domains, typically in the form of (more hydrophilic) α-helices and β-turns playing a crucial role in the supercontraction phenomenon.[10,22–24]

The fiber's cross-section is organized radially,[23,25–27] with a protective skin that does not influence supercontraction or mechanical response,[28] therefore neglected here, and a core with a different external and internal composition. In particular, Brown et al.[29] propose a gradual transition between the inner and outer part of the core, associated with changes in spidroin composition along the fiber cross-section. On the other hand, Sponner et al.[10,23] indicate clear delineations between these regions, as evidenced also from a recent multi-omics study by Sonavane et al.,[27] which revealed distinct spatial segregation of spidroin components within the dragline fiber, with minimal mixing between them.

In the following, we formally distinguish an outer part mainly consisting of proteins MaSp1 organized into β-pleated sheets, referred to as a hard region, and an inner part, referred to as a soft region, primarily containing proteins MaSp2 with lower crystallinity and weaker crystal domains, typically in the form of α-helix and β-turns.[10,22–24] However, as detailed further, the axial (homogenized) response of the spider silk thread, as determined through our approach, is substantially independent of the precise spatial distribution of these material phases. The varying crystallinity has been shown to be possibly influenced by the shear stress at the spinning duct wall, which promotes the formation

N. M. Pugno
Laboratory for Bioinspired
Bionic, Nano, Meta Materials & Mechanics
University of Trento
Via Mesiano 77, Trento 38123, Italy
E-mail: nicola.pugno@unitn.it
N. M. Pugno
School of Engineering and Materials Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK

of harder crystal domains as β-sheets predominantly in the outer region.[29,30]

Malay et al.[18] propose that a closer examination of MaSp repetitive sequences could reveal important insights beyond the traditional "hard" poly-Ala and "soft" Gly-rich regions. Indeed, while proline's role in MaSp2,[31] particularly in supercontraction,[32] has been studied, the functions of other conserved residues remain underexplored. Recent findings, such as a potential link between tyrosine residues and supercontraction,[8] suggest that the conserved sequences in Gly-rich regions across spider species have significant biological relevance.

The supercontraction phenomenon has been extensively studied using various experimental techniques. Fourier-transform infrared (FTIR) spectroscopy, X-ray scattering techniques, and nuclear magnetic resonance (NMR) spectroscopy have demonstrated that water entering the fiber primarily affects the amorphous regions, possibly by breaking interchain hydrogen bonds.[28,33–39] However, due to the varying crystalline compositions, humidity affects the hard and soft fractions differently. Water has minimal impact on breaking the H-bonds within the compact β-sheet domains of the hard fraction.[28] However, in the hard fraction, an increase in humidity can cause a misalignment of crystals relative to the fiber direction.[39] As we show in the following, this misalignment may represent a crucial factor in determining the degree of supercontraction.

Conversely, water is well known to significantly affect the soft internal core, where it can more easily disrupt the hydrogen bonds between the chains of the hygroscopic amorphous phase.[40]

The primary sequence of the proteins in the soft region plays a significant role during supercontraction which will be discussed in detail later.

Finally, an important factor in the evolution of natural chain length (i.e. the end-to-end distance in the configuration that the chain assumes without any external applied force) is influenced by the stretch history (see ref. [41] and references therein for a detailed theoretical discussion of this phenomenon). As the end-to-end distance of the macromolecules changes, β-sheets may undergo unraveling, resulting in an increase in the number of available monomers, particularly in the hard fraction, as detailed in ref. [42].

Significant attempts have been made to model the supercontraction effect and gain a deeper understanding of its origin. A molecular dynamics analysis is presented in ref. [30], whereas in refs. [42, 43] microstructure-inspired and energy-based models that consider and track the microstructural evolution of the silk fiber as the relative humidity (RH) increases are outlined. Specifically, supercontraction was modeled as a loss of orientation and folding of the chains in the network due to the dissociation of intermolecular hydrogen bonds.

Recently, a physically-based machine learning approach has been proposed by some authors of this paper, to allow the deduction of new scientific knowledge based on database of material properties.[44] In particular, the adoption of a symbolic data modeling technique is proposed, namely the "Evolutionary Polynomial Regression" (EPR) which integrates regression capabilities with the genetic programming paradigm, enabling the derivation of explicit analytical formulas for deducing structure-function relationships emerging across different scales, in hierarchical materials.

The EPR algorithm explores possible polynomial models for calculating the target output based on predictive accuracy and parsimony. The models thus have a pseudo-polynomial structure, where each term is made up of a combination of candidate inputs, each given with its own exponent determined during the evolutionary search. Moreover, each polynomial term is multiplied by a constant coefficient, which is estimated by minimizing the error on the training data. A synthetic description of the EPR technique, along with relevant references, is provided in the following section.

In this work, we investigate the data-driven dependence of supercontraction on parameters describing the primary structure of the proteins that make up dragline silk, with the aim of gaining new insight to be benchmarked against microstructure-based models of supercontraction. Accordingly, here we identify some directions to extend an existing model for supercontraction to consider an additional length scale compared to the previous model, namely the primary structure of proteins, and describe their role within the supercontraction mechanism. The introduction of molecular-scale properties, particularly the amino acid sequences of MaSp1 and MaSp2 proteins, into a multiscale analytical model represents, in our opinion, a significant theoretical advancement. To the best of our knowledge, no other multiscale model has directly integrated these molecular-level properties. Specifically, in the microstructure-based model we aim to extend, proposed by some of the authors of this paper, the silk fiber is treated as a composite material with a hard external fraction of crystalline chains and a soft internal fraction of amorphous chains.[42] Both the hard and soft fractions of macromolecules are supposed to be aligned with the fiber axis and embedded in a tridimensional elastic matrix, describing the complex macromolecular network with inner and intrachains connections. In that context, the dependence of supercontraction on the soft amorphous proteins has been addressed, with proteins treated using the classical statistical mechanics approach to quantify the expected length of the protein's macromolecules depending on the silk hydration conditions. In particular, the variation of the end-to-end length of the soft chains with humidity was determined as a function of the bonds that are naturally present within the silk when spun and can be disrupted by hydration water molecules. This physical interpretation was based on the available experimental literature, which at that time did not include enough experimental observations regarding the supercontraction and the protein lengths. On the other hand, the Silkome database[19] provides an extensive amount of data, including primary sequences of the proteins and corresponding hydration properties of the silks, with limited insights regarding the modeling of such correlations.

A viable strategy for extracting information from a material properties database to inform theoretical modeling was outlined in ref. [44] using spider silk as a case study. This strategy considered three length scales–macro, meso, and micro (protein molecules)—to deduce macroscopic properties from lower-scale data. Notably, for supercontraction, a strong dependence on micro properties was observed, as opposed to other macroscopic properties that were more accurately predicted by mesoscale properties. In particular, the data-driven analysis identified a

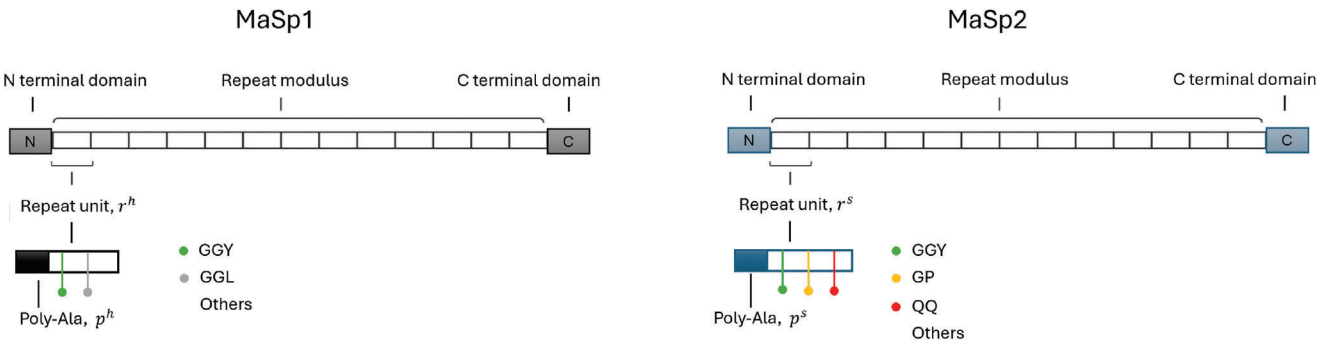

Figure 1. Scheme of the primary structure of the proteins MaSp1 ans MaSp2, both composed of a repeat modulus flanked by N and C-terminal domains. Each of the repeat units contains a polyalanine region. The repeat units of the MaSp1 and MaSp2 are on average composed of r^h and r^s amino acids respectively. The polyalanine regions of the MaSp1 and MaSp2 are on average composed of p^h and p^s alanine amino acids respectively. Additionally, the scheme highlights common motifs such as GGY, as well as exclusive motifs like GGL in MaSp1 and GP, QQ in MaSp2, which are responsible for their distinct structural behavior, including the role of the GP motif in preventing β-sheet formation, a critical aspect in supercontraction behavior.

linear dependence of supercontraction on the repetitive length of the MaSp2 protein and an inverse dependence on the MaSp1 polyalanine length.

In this work, in order to improve previous microscopically-motivated models for supercontraction, we analyze the recalled results of the machine learning models such as the ones found in ref. [44]. Specifically, based on the symbolic machine learning results, first we interpret and justify them through simple molecular scale models deduced based on classical considerations from rubber elasticity. This approach is enabled by the crucial feature of the machine learning technique proposed in ref. [44] which allows the possibility of deducing analytical relations that highlight the properties of the main material analytical functions. The new low scale model can then be implemented in previously adopted analytical model that can then be tested again in a comparison with the experimental behavior. As we show, the so enhanced microscopically-motivated model may outperform the original machine learning models. This is possible because the microstructure-inspired model considers more material properties relevant to supercontraction than those reported in the Silkome database, which was employed for the data modeling analysis.

As a byproduct result, new experiments are suggested by the proposed model for supercontraction, to measure material properties that appear to be relevant for determining the amount of silk contraction and were not measured in previous experimental campaigns.

A new research paradigm is therefore outlined by pursuing the proposed methodology. The classical paradigm, in which experimental observations inspire models of the phenomena, and models in turn suggest new experiments, is replaced by a strategy rooted on the increasing availability in many scientific fields of big experimental data. These data can be interpreted through physically-based machine learning algorithms that can provide interpretable models of data. Such data-driven models can be employed to enhance the scientific knowledge after a careful interpretation. This process may result in the generation of new experiments, thereby enabling a continuous cycle where machine learning is integrated within the well-established physically-based framework for modeling. This paradigm contrasts with the use of machine learning to deduce non-interpretable "black-box"models from data, which can provide even finer reproduction of data, but not a real knowledge advancement as extensively discussed in ref. [44] for the case of materials modeling.

Despite challenges, particularly regarding physical interpretability, the fitting performance and flexibility in mapping complex systems have allowed some black-box techniques based on artificial neural networks to achieve noteworthy results in linking the protein sequences of spider silk with their mechanical properties.[45,46] These results are also possible by the machine learning approach's ability to effectively integrate molecular data into macroscopic models, overcoming typical limitations of other computational methods such as Molecular Dynamics (MD) simulations and Finite Element Analysis (FEA). Specifically, MD simulations capture atomistic-level detail but are computationally expensive and challenging to scale for macroscale phenomena, while FEA is efficient for macroscopic modeling but may overlook microscopic details. On the other hand, our numerical approach bridges these scales with machine learning interpretable models that link molecular behavior to macroscopic material response.

We remark that, while many computational approaches typically operate separately from analytical models, our work uniquely integrates machine learning to directly enhance an analytical framework. This novel fusion of data-driven insights and theoretical modeling establishes a foundation for bridging the gap between computational techniques and physically-based approaches.

While finalizing this paper, we learned the news that the 2024 Nobel Prize in Physics has been awarded for contributions in the field of machine learning and artificial neural networks. This recognition underscores the significance of such approaches for scientific advancements, particularly in materials science. As reported in the recent paper,[47] Hinton, speaking by telephone during the prize announcement, stated that learning he had won the Nobel was "a bolt from the blue." "I'm flabbergasted, I had no idea this would happen," he said, adding that advances in machine learning "will have a huge influence, comparable to the industrial revolution. But instead of exceeding people in physical strength, it's going to exceed people in intellectual ability." We argue that some of the results presented in this paper represent a small step in this direction within the field of materials science.

2. Model/Method

We start by outlining a model of the spider silk, certainly simplified compared to the real complexity of the material, but retaining the components that we consider most effective in describing the behavior of interest.

Following the approach proposed by some of the authors of the present paper in ref. [42] and based on the previous description and referred literature, we consider the silk thread as a composite material composed of a hard external fraction of crystalline chains and a soft internal fraction of amorphous chains. The chains in the hard region are assumed to be ordered, aligned with respect to the fiber axis and therefore tightly packed. Both the hard and soft fractions of macromolecules are embedded in a 3D elastic matrix, which represents the complex macromolecular network with inter- and intrachain interactions.

According to the protein structure introduced above, we depict the schematic representation of the primary structure of the two considered proteins in Figure 1. The experimental (average) measure of the number of amino acids composing each repeat unit is available from the Silkome database and here is denoted by r^h and r^s for the MaSp1 and MaSp2 respectively.[19,48] Moreover, each repeat unit encompasses a polyalanine part which is responsible of composing the β-sheet secondary structure. The experimental (average) measure of the number of amino acids composing each polyalanine region is available from the Silkome database and here is denoted by p^h and p^s for the MaSp1 and MaSp2 respectively.

We note that in our analytical model we consider only two fractions: the soft and the hard one. More complex models could incorporate mixed phases, however, we believe that the proposed model is capable of capturing the main distinction between the proteins composing the silk, one of which (here referred as the hard fraction) represents the more crystalline domains. This simplification is an assumption of the model aimed at deducing fully analytical results while considering the main physical effects and material components.

Finally, we address the supercontraction effect, introduced before, which occurs when a spider silk thread is exposed to humidity, causing the fiber to contract to a length that may reach half of its initial dry length. To this respect, the Silkome database reports the maximum supercontraction, calculated as

$$sc := \frac{L_0 - L_f}{L_0} = 1 - \lambda_{sc} \qquad \text{where} \qquad \lambda_{sc} := \frac{L_f}{L_0} \tag{1}$$

where L_0 is the length in dry condition and L_f in fully wet conditions ($RH = 100\%$).[19]

2.1. EPR Analysis

As anticipated in the introduction, in the proof of concept work illustrating the potential of adopting physically-based machine learning to model the behavior of hierarchical materials, where the spider silk served as a case study, from preliminary EPR analysis resulted that the supercontraction can be predicted with relatively good accuracy from the amino acid number of the repetitive unit of the MaSp2 and the amino acid number of the poly wing region of the MaSp1. Interestingly, among other results that will be analyzed in future studies, the data optimization based on a hierarchical distinction in micro, meso and macro properties, compared with a direct deduction of macro properties from micro ones, evidenced a direct influence of MaSp1 and MaSp2 molecular scale properties directly on the macroscopic supercontraction behavior. This result suggested the need for a more in-depth investigation of the relationships among the involved quantities. Therefore, in this paper, we conduct a *de novo* specific analysis on the dependence of the supercontraction on the MaSp1 and MaSp2 properties, both from data analysis and modelling interpretation. The resulting knowledge is eventually adopted to improve the previously proposed multiscale model.[42]

We begin by recalling, for the reader's help, the main strategy of the EPR method and the modeling adopted in the previous work.[44]

2.1.1. EPR Modeling Strategy

Although the details regarding the main EPR paradigm are thoroughly explained in the reference works, refs. [49,50], and ref. [44] for the application of EPR with hierarchical materials, a brief summary is presented here. In the simplified setting considered in this paper, EPR constructs explicit mathematical expressions to model a set of data points, beginning from general model pseudopolynomial relations

$$Y = a_0 + \sum_{j=1}^{m} a_j \mathbf{X}_1^{\mathbf{ES}_{j1}} \mathbf{X}_2^{\mathbf{ES}_{j2}} \ \dots \ \mathbf{X}_k^{\mathbf{ES}_{jk}} \tag{2}$$

where Y is each considered output dependent variable, $\mathbf{X}$ is the vector of input variables and a_o is a bias term. In other words, we assume that Y can be expressed as a polynomial function composed of m terms, expressed as products of powers of the $\mathbf{X}_i$ generated by the algorithm. In the more general case, other elementary functions can be considered in place of the powers, as discussed in ref. [50]. Notice that each of the m terms is linearly dependent on the unknown parameters a_j. The power exponents denoted by $\mathbf{ES}_{ji}$ are selected from a defined set of values.

In a nutshell, EPR proceeds in two steps: structure identification and parameter estimation. The initial stage of the process involves simultaneously determining the optimal arrangement of the independent variables and the related exponents. This optimization is finalized using a multi-objective genetic algorithm known as OPTIMOGA (Optimized Multi-Objective Genetic Algorithm), based on the MOGA (Multi-Objective Genetic Algorithm) strategy, extensively described in refs. [50, 51].

It is important to note that since the set of candidate exponents is predefined by the user, possible negligible input variables are identified by including zero within the set. This is a fundamental option for determining the effective independent variables.

In the second stage, the values of the parameters a_j are determined using the linear Least Squares (LS) approach, which minimizes the Sum of Squared Errors (SSE). In addition to the usual LS estimation, this is performed with the constraint $a_j > 0$ to ensure positive parameter values. This choice helps in avoiding over-fitting, by excluding sequences of terms with negative/positive a_j values that may result from the modeling of the data noise,[52] however, it can be removed by the EPR algorithm.

Table 1. Equations of the Pareto Front returned by EPR.

Equations of the Pareto front	R^2 (%)	
$sc = 0.32479$	1.7	(3.1)
$sc = 6.4 \times 10^{-5} (r^s)^2 + 0.20495$	23.1	(3.2)
$sc = 0.051494 (r^s)^{0.5}$	19.3	(3.3)
$sc = 0.46367 \frac{(r^s)}{(p^h)^2} + 0.017527$	47.1	(3.4)
$sc = 0.41551 \frac{(r^s)^{0.5}}{p^h}$	41.0	(3.5)
$sc = 0.08718 \frac{r^h (r^s)^{0.5}}{(p^h)^2}$	47.6	(3.6)
$sc = 0.073216 \frac{(r^h)^{0.5} r^s}{(p^h)^2} + 0.032358$	49.8	(3.7)
$sc = 2.5526 \frac{1}{(p^h)^2} + 0.070618 \frac{(r^h)^{0.5} r^s}{(p^h)^2}$	50.7	(3.8)
$sc = 3994.7427 \frac{1}{(r^h)^2 (p^s)^2} + 0.011144 \frac{r^h r^s}{(p^h)^2}$	53.0	(3.9)
$sc = 4705.107 \frac{1}{(r^h)^2 (p^s)^2} + 0.00381 \frac{r^h r^s (p^s)^{0.5}}{(p^h)^2}$	53.6	(3.10)
$sc = 11926.8478 \frac{1}{(r^h)^2 (p^h)^{0.5} (p^s)^2} + 0.0038596 \frac{r^h r^s (p^s)^{0.5}}{(p^h)^2}$	54.1	(3.11)
$sc = 288.5672 \frac{(r^s)^{0.5}}{r^h p^h (p^s)^2} + 0.0011151 \frac{r^h r^s p^s}{(p^h)^2}$	54.6	(3.12)

This can be a fortiori justified in our physical model of spider silks by referring to only positive values of the input and output variables.

Furthermore, the uncertainty of the coefficients (a_j) is evaluated during the search process. The distribution of estimated pseudo-polynomial coefficients is then used to eliminate those parameters whose value is not sufficiently larger than zero, as detailed in refs. [49, 53]. It can be argued that a low coefficient value compared to the variance of estimates indicates that the corresponding term likely represents noise rather than a meaningful component of the studied phenomenon.

The set of exponents considered in the proof-of-concept work[44] {0,1,–1,0.5,–0.5} has here been expanded to {0,1,–1,0.5,–0.5,2,–2} in order to investigate more detailed nonlinear dependencies between variables. Such extension aims to enhance the model's ability to capture complex relationships, allowing for the inclusion of quadratic growth and decay patterns in addition to linear, inverse, and fractional power dependencies. The maximum number of terms has been set to $m = 2$. Such an assumption is justified by comparing the expressions and their corresponding performances in **Table 1**. Indeed, we achieve near-maximal performance with just one terms (bias excluded), and even adding a second term does not significantly improve the fitting performance. Additionally, we include the possibility of a bias term, a_0, in the model expressions, as this element may compensate for the absence of relevant inputs in the model.

As anticipated above, we refer to the recent experimental campaign proposed in ref. [19] by some of the authors of this paper, where the authors analyzed the properties, at different scales, of approximately 1000 different silks. Among the material properties analyzed in the paper, the quantity of interest for this work are the supercontraction, a nondimensionional quantity ranging in (0,1), the number of amino acids composing the repeat length of the MaSp1 and MaSp2, namely r^h, r^s and the number of amino acids composing the polyalanine region in the MaSp1 and MaSp2, p^h and p^s respectively. As a primary indicator of accuracy, we report the Coefficient of Determination for

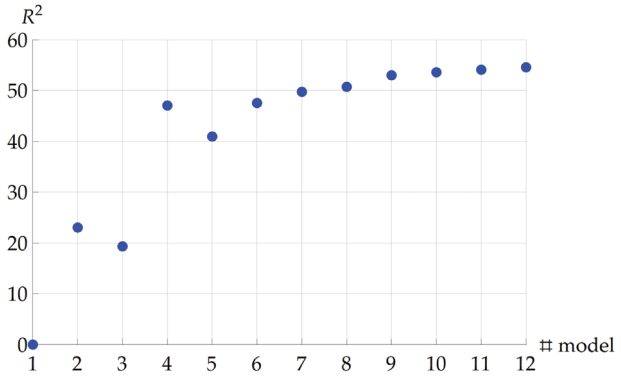

Figure 2. EPR model performance in terms of R^2 plotted against the increasing number of EPR equation.

the different EPR equations, considering the classical definition $R^2 = 1 - \sum_{i=1}^{N} \frac{(x_i^{num} - x_i^{exp})^2}{(x_i^{exp} - \bar{x}^{exp})^2}$, where the x_i^{num} are the output variables of the numerical test, x_i^{exp} are the corresponding experimental values and $\bar{x}^{exp} = \frac{1}{N} \sum_{i=1}^{N} x_i^{exp}$ their mean value, with $i = 1, \ldots N$, where N is the number of experimental observations considered as dependent variables. From the database,[19] we considered only the data where both the searched output and the relevant input are simultaneously reported. While the training set consists of 38 samples due to missing experimental data for certain silks, symbolic regression, by generating parsimonious models with fewer parameters, allows for reliable model training and prediction even with a limited number of data points (see ref. [44] for a more comprehensive discussion).

2.1.2. EPR Results and Discussion

The EPR technique has returned a series of polynomial expressions, reported in Table 1 for determining the supercontraction as a function of the variables considered as candidate input, namely r^h, r^s, p^h, p^s.

In **Figure 2** we report the variation of the accuracy of the analytical expressions in reproducing the experimental data. Thus, in a Pareto front approach, the model let us choose the best formulae considering parsimony (simple expression) and accuracy. In particular, the model simplicity is quantified by penalizing the number of inputs in the model expressions, controlling for constant values that may represent noise, and evaluating the variance of model terms against the noise variance in the data, as detailed by Giustolisi and Savic.[49] This approach prevents overfitting and ensures the selection of parsimonious yet accurate models.

From the Pareto front of expressions returned by EPR (Table 1), it is evident that the Equation (3.4), despite being rather simple, (as it includes, in addition to the bias, a single term consisting of a constant, and two of the four input parameters, i.e., the length of the MaSp2 repeat unit r^s and the length of the Poly-Ala segment of the MaSp1 p^h) achieves nearly maximum accuracy. In particular, the parameter $(r^s)^{0.5}$ is conserved from the two previous equations and is still present in the two following equations. This means that this dependence turns out to be particularly effective to produce a good estimation of the supercontraction value. The dependence of $1/(p^h)^2$ appears in the highlighted

expression and is of particular effectiveness as well because it does more than double the R^2 value. An inverse dependence of $1/p^h$ is still present in the following expression, whereas, the subsequent expressions restore the $1/(p^h)^2$ dependence. Accordingly, we can consider the dependence $1/(p^h)^2$ to be relevant for predicting supercontraction, although there is no fully clear dependence regarding the exponent 2. Compared to the equation found in ref. [44], i.e. $sc = 0.062 r^s/p^h + 0.00474$ with $R^2 = 43.4$, the exponent 2 allows for a slight increase in R^2.

Lastly, the expressions (3.6–3.12), while more complex, do not allow for a significant increase in R^2 compared to the highlighted expression, thus, as described above, they are likely describing noise in the data rather than important dependencies among the parameters considered in the analysis.

In the following we first try to deduce theoretically, based on molecular scale properties, the physical meaning of the determined analytical relations. We then implement the resulting models in the multiscale model we took as reference.

2.2. Enhancing the Original Microstructure Inspired Theoretical Model

As anticipated, the model we are based on and plan to extend in this work is the micro-structure inspired model proposed in ref. [42], where the authors describe the spider thread as being composed of a multi-phase material.

In this model, the stress is additively decomposed as the soft fraction contribution σ^s, the hard fraction contribution σ^h, and the elastic matrix contribution, representing the macromolecular network, modeled as neo-Hookean with modulus μ. Thus, the total stress-stretch constitutive response of the silk is given by

$$\sigma(\lambda) = \sigma^s(\lambda) + \sigma^h(\lambda) + \mu\left(\lambda - \frac{1}{\lambda^2}\right) \tag{4}$$

Notice that in the original model, the hard fraction σ^h is assumed active only for lengths larger than the natural one (see ref. [42] for details), whereas in this work such assumption will be modified as a consequence of the data-driven analysis to consider possible compression response due to the network interaction. Also the soft fraction behavior, dependent on RH as in the original paper, will be revisited in this article, where we further consider the primary structure of the protein constituting the soft fraction. The behavior of the soft fraction, which is dependent on RH as in the original paper, is also revisited in this article, where we further consider the primary structure of the protein constituting the soft fraction.

The chain of both the soft and hard fractions are modeled in accordance with the classical Statistical Mechanics approach,[54] showing that the expectation value of the end-to-end distance for ideal chains in the unloaded state is

$$L_n = < r^2 >^{1/2} = b\, n^{1/2} \tag{5}$$

where n is the number of Kuhn segments with length b. We refer to this length as "natural length," attained in the absence of any external forces. On the other hand, the chain contour length is simply $L_c = n\, b$.

Following,[55] the adopted energy density per unit chain contour length L_c is assumed of Worm Like Chain (WLC) type, $\varphi_e = \varphi_e(L, L_c) = \kappa \frac{L^2}{L_c - L}$ where $\kappa = \frac{k_B T}{4 l_p}$, T is the temperature, k_B the Boltzmann constant, and l_p the persistence length. This energy respects the limit extensibility condition, $\lim_{L \to L_c} \varphi_e(L, L_c) = +\infty$, and allows for explicit calculations. Moreover, the end-to-end distance L is decomposed into a variable (zero-force) natural length measured by (5) and the remaining elastic component $L_e = L - L_n$, as firstly proposed in ref. [56]. Thus the assumed energy and force-elongation laws for a single chain are

$$\begin{aligned} \varphi_e &= \kappa \frac{L_e^2}{L_c - L} = \kappa \frac{(L - L_n)^2}{L_c - L}, \\ f &= \frac{\partial \varphi_e}{\partial L} = \kappa \left[\left(\frac{L_c - L_n}{L_c - L} \right)^2 - 1 \right] \end{aligned} \tag{6}$$

with the force decreasing to zero as the length attains its natural length ($L = L_n$ or $L_e = 0$).

The macroscopic behavior of the silk thread is deduced following ref. [57] considering the classical *affinity hypothesis*[54] that identifies the macroscopic stretches with the macro-molecular ones. Hence, first the stretch measures of the different fractions are evaluated as

$$\begin{aligned} \lambda^i &= \frac{L}{L_0^i} && \text{total stretch,} \\ \lambda_e^i &= \frac{L_e^i}{L_0^i} && \text{elastic stretch,} \\ \lambda_n^i &= \frac{L_n^i}{L_0^i} && \text{permanent stretch,} \\ \lambda_c^i &= \frac{L_c^i}{L_0^i} && \text{contour stretch,} \end{aligned} \qquad i = h, s, m, t \tag{7}$$

with $L_o^i = b^i \sqrt{n_0^i}$ denoting the initial natural length of the macromolecule. Moreover, here and in the following we indicate by the apexes s, h, m, and t the soft, hard, matrix and homogenized (total) quantities. It is noteworthy that the permanent stretch takes the role of the plastic stretch in classical non-linear plasticity theories, as it measures the variation of the natural length (for a detailed theoretical discussion, see ref. [58]). We remark that in the original model,[42] the evolution of the hard region contour and permanent stretch is dependent on the load history, while the permanent stretch of the soft fraction is dependent on the silk hydration conditions.

2.2.1. Soft Region—MaSp2 Repetitive Region Length Effect

In this section, based on the theoretical model for the RH-induced evolution of the natural length of the soft fraction proposed in ref. [44], we give a physical interpretation of the previously deduced supercontraction dependence on the repeat length of the MaSp2 protein r^s as suggested by EPR by Equation (3.4). For further discussion, see Appendix A.

In ref. [42], the variation of the natural length of a "representative" soft chain with the humidity, has been determined by observing that, after spinning, several bonds among chains are naturally present within the silk, preventing the entropic coiling of the chains. These bonds keep the macromolecules in a glassy state in dry conditions, and the contraction of the silk occurs

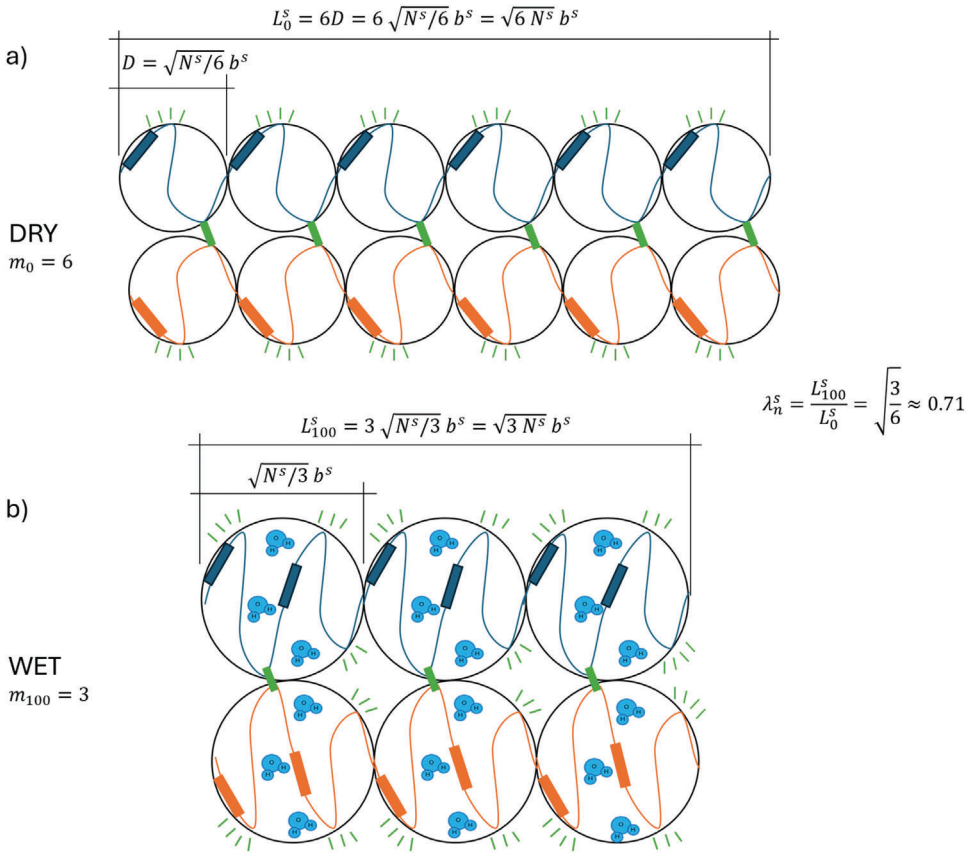

Figure 3. Schematic representation of two adjacent soft chains (blue and orange), each divided in $m_0 = 6$ domains in dry conditions due to the hydrogen bonds (green segments) among the chain keeping the chains in an extended "glassy" configuration. Each domain includes one repeat units composing the silk protein, which can occupy a space defined by a blob. The boxes represent the polyalanine region, whereas the wires represent the remaining parts of the repeat units. The green thin segments represent the sites where bonding with other chains may occurs. In wet conditions, some hydrogen bonds are broken by some hydration water molecules, e.g., three broken bonds and three intact ($m_{100} = 3$) in (b). As a consequence, a couple of blobs become a larger blob including two repeat units. The contraction stretch in this case is $\lambda_n^s = \sqrt{m_{100}/m_0} \approx 0.71$.

when they are disrupted by the interaction with the hydration water molecules. This simple idea allowed for a correct quantitative description of the thermo-hygro-mechanical behavior of spider silks in ref. [42].

To get a quantitative measure of this effect, let us assume that, due to the presence of (RH-dependent) crystal domains in the soft region, the soft chain is divided into $m(RH)$ identical domain (see the scheme in **Figure 3**) able to independently recoil according to (5). Thus, let us indicate by N^s the number of Kuhn segments composing the whole soft chain. In the generic humidity state, the (mean) number of monomers in each domain is therefore $n^s(RH) = N^s/m(RH)$. In accordance with Equation (5), the natural length of the whole chain is then

$$L_n^s(RH) = m\sqrt{N^s/m(RH)}\,b^s = \sqrt{m(RH)\,N^s}\,b^s \tag{8}$$

In view of (5), the permanent stretch of the soft region can be deduced as

$$\lambda_n^s = \frac{L_n^s}{L_o^s} = \frac{\sqrt{N^s\,m(RH)}\,b^s}{\sqrt{N^s\,m_o}\,b^s} = \sqrt{\frac{m(RH)}{m_o}} \tag{9}$$

Since here, we limit our attention to the fully supercontracted silk, i.e., $RH = 100\%$ –the only case considered in the Silkome database[19]– we have

$$\lambda_n^s = \sqrt{\frac{m_{100}}{m_0}} \tag{10}$$

where $m_{100} \coloneqq m(100)$.

To interpret the scaling behavior of the supercontraction on r^s let us schematize, by following ref. [59] each chain as composed of a series of 'blobs' as shown in Figure 3, each of size D. To get this scaling, we identify each portion of the chain within a blob as a repeat unit of the MaSp2 and we assume that the blobs have a spherical shape with diameter calculated by employing Equation (5):

$$D = \sqrt{n^s}\,b^s \tag{11}$$

Since we may assume that $n^s \propto r^s$, we have

$$D = \sqrt{k_1\,r^s}\,b^s \tag{12}$$

with r^s the number of amino acids composing the repeat unit introduced before and experimentally measured in ref. [19].

Here, we simply assume that the bonds $m(RH)$ among the chains, limiting the entropic recoiling, are located at the surface of the blobs in which the chain is schematically divided. As a result, the number of bonds in dry conditions m_0 is proportional to the surface of the sphere that can be calculated as

$$S = 4\pi\left(\frac{D}{2}\right)^2 = \pi\, k_1\, r^s\, (b^s)^2 \tag{13}$$

Thus,

$$m_0 = k_2\, S = \pi\, k_1\, k_2\, r^s\, (b^s)^2 \tag{14}$$

On the other hand, m_{100} is considered to be a constant quantity as it is the number of bonds that remain intact even in the limit condition of $RH = 100\%$, i.e., the number of highly hydrophobic regions within the polymer chain.

Accordingly, Equation (10) can be rewritten as

$$\lambda_n^s = \sqrt{\frac{m_{100}}{m_0}} = k_3/\sqrt{r^s} \tag{15}$$

with $k_3 = \sqrt{\frac{m_{100}}{\pi\, k_1\, k_2\, (b^s)^2}}$.

The corresponding expression for the contour length is $L_c^s = m\frac{N^s}{m}\, b^s = N^s b^s$, so that the contour stretch of the amorphous part calculated by using Equation (7) is not dependent on the humidity:

$$\lambda_c^s = \frac{L_c^s}{L_o^s} = \frac{N^s\, b^s}{\sqrt{N^s\, m_o}\, b^s} = \sqrt{\frac{N^s}{m_o}} \tag{16}$$

In view of Equation (14), it can be rewritten as

$$\lambda_c^s = k_4/\sqrt{r^s} \tag{17}$$

with $k_4 = \sqrt{\frac{N^s}{\pi\, k_1\, k_2\, (b^s)^2}}$. We note that $k_4/k_3 = \sqrt{N^s/m_{100}}$.

The engineering (Piola Kirchhoff) stress is determined under an additive assumption, from Equations $(6)_2$ and (7) as

$$\sigma^s = E^s\left[\left(\frac{\lambda_c^s - \lambda_n^s}{\lambda_c^s - \lambda}\right)^2 - 1\right] \tag{18}$$

where $E^s = N_a\kappa$ is the elastic modulus of the soft fraction with N_a the number of chains per unitary reference area.

In view of Equations (15) and (17), the latter can be rewritten as

$$\sigma^s = E^s\left[\left(\frac{k_4/\sqrt{r^s} - k_3/\sqrt{r^s}}{k_4/\sqrt{r^s} - \lambda}\right)^2 - 1\right] \tag{19}$$

2.2.2. Hard Region—MaSp1 Polyalanine Region Length Effect

In this section we extend the same methodological approach of previous section also for the crystalline region, by considering the MaSp1 role, to physically interpret the scaling of supercontraction on p^h in Equation (3.4). The resulting relation is then inserted in the original model of ref. [42] to determine the stress in the hard crystalline fraction.

To this hand, we observe that in the original model,[42] the chains of the hard region were slack during supercontraction compression. These chains were active only for the successive extension, after supercontraction, when they reached the original natural length. On the other hand, this assumption would necessarily imply that the supercontraction is independent of MaSp1 properties, being this protein part of the hard fraction. Our data modeling results then suggested that the MaSp1 has a role also during the supercontraction. This is a further meaningful physical insight resulting from the EPR analysis.

In particular, since Equation (3.4) exhibits the dependence of the supercontraction $sc \propto 1/(p^h)^2$, i.e., on the length of the polyalanine region within the MaSp1, we outline a simple model to interpret this scaling based on previous knowledge of the microstructure of silk, which we briefly recall. The polyalanine regions in the hard region form β-sheets. We note that this occurs less frequently in MaSp2, here modeled as the soft region, as extensively discussed above. There is experimental evidence that during supercontraction, the β-sheets change their alignment with respect to the fiber axes.[28,39] In particular, while in the initial dry state they are mostly aligned, the misalignment increases as the humidity grows. Such humidity-induced modification in the crystals alignment was modeled in a phenomenological way in ref. [42]. Furthermore, it should be noted that water molecules can hardly disrupt H-bonds within the compact β-sheets domains, that for this reason are considered hydrophobic.[28]

We consider a single chain of MaSp1, and we assume that the length of the polyalanine region within such protein determines the size of the β-sheet in the direction of the fiber axes. We focus on a single repetitive region of MaSp1 and we consider that it corresponds to one β-sheet. During supercontraction, this chain would coil due to the interaction with the MaSp2, which undergoes entropic coiling driven by H-bond disruption, as described above.

The coiling of the MaSp1 chain results in a misalignment of the β-sheet with respect to the initial configuration, as evidenced by experiments.

By following a classical approach in rubber elasticity, describing the restricted motility of chains due to interacting network chains,[60] we assume that the hard chain can change its conformation without constraints only within a cylinder (see **Figure 4**). Thus, if we fix the tube diameter d and the β-sheet length l, the maximum misalignment angle $\alpha_{\max}$ that the β-sheet can reach is limited by the presence of the cylinder wall, according to the following equation:

$$d = l \sin \alpha_{\max} \tag{20}$$

In **Figure 5** we plot the angle of maximum misalignment with respect to the fiber axes with a tube diameter $d = 2$ nm, length of the β-sheet ranging within the interval (2.26 nm, 3.76 nm), based on the average length of the MaSp1 polyalanine region $\bar{p}^h = 8$ amino acids and its standard deviation of 2 amino acids and by considering the length of the alanine amino acid within a β-sheet configuration of $l_a = 0.376$ nm.

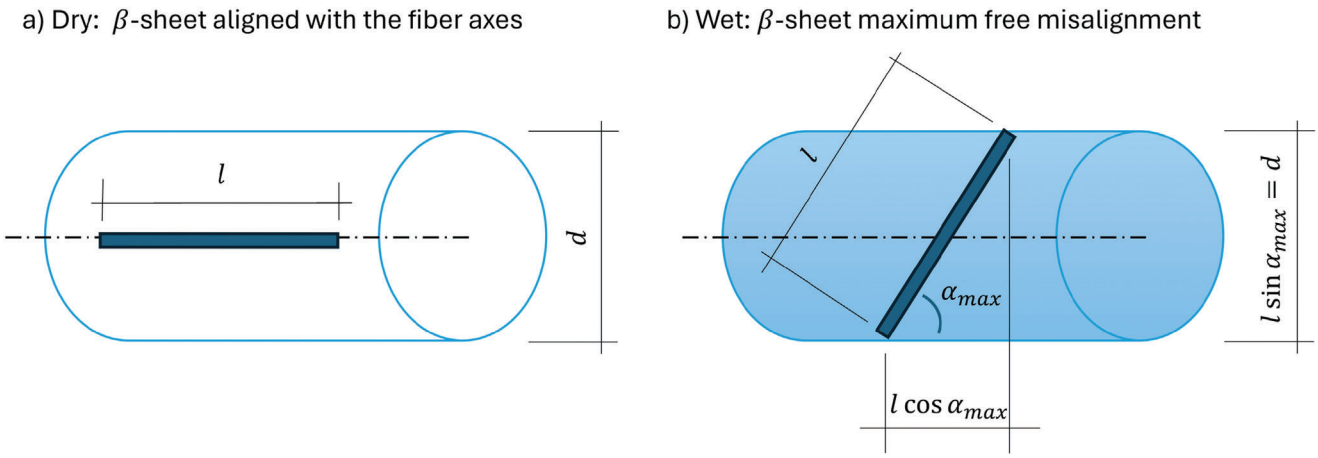

Figure 4. Humidity effect on β-sheets alignment: a) in dry conditions, the β-sheet is aligned with respect to the cylinder axes, b) in wet conditions the β-sheet results may freely misalign within a space confined by a tube representing the presence of other chains.

Accordingly, the contraction along the fiber axes due to the β-sheet misalignment can be computed by considering its initial length in dry conditions with crystal aligned with the axes and its projection on the fiber axes in wet condition, as

$$\lambda = \frac{l\cos\alpha_{\max}}{l} = \cos\alpha_{\max} \tag{21}$$

In view of Equation (20), we can determine the critical value of stretch where the β-sheet touch the cylinder wall as a function of the tube diameter and β-sheet-length:

$$\lambda_{cr} = \sqrt{1 - \left(\frac{d}{l}\right)^2} \tag{22}$$

with $l = l_a\, p^h$.

Due to the connection between poly-Ala and non poly-Ala regions, as schematized in **Figure 6** we may extend this result to the whole chain.

We then assume that after the β-sheet touches the tube wall, the tube (the Network) reacts with an elastic force $F_e = k\,y$ with $y = l\sin\alpha$.

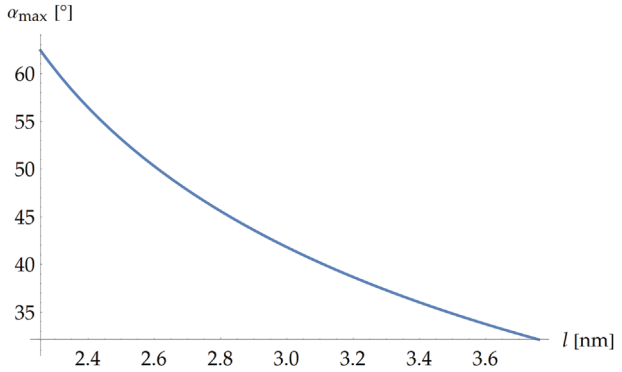

Figure 5. β-sheet maximum misalignment angle as a function of the β-sheet length. Here a cylinder diameter $d = 2$ nm was employed.

The rotational equilibrium of the β-sheet then gives

$$F_a\, l\sin\alpha = F_e\, l\cos\alpha \tag{23}$$

that can be written as

$$F_a\, l\sin\alpha = k\, l\sin\alpha\, l\cos\alpha \tag{24}$$

so that

$$F_a = k\, l\cos\alpha \tag{25}$$

Thus, in view of Equation (21)

$$F_a = k\, l\lambda = k\, \bar{l}_a\, p^h \lambda \tag{26}$$

Dividing by the area A on which F_a acts, we obtain

$$\sigma^h = -k_5\, p^h\, \lambda \tag{27}$$

with $k_5 = k\,\bar{l}_a/A$, where the negative sign reflects the fact that the hard fraction is under compression. The behavior of the hard region under tensile load was addressed in details in ref. [42].

Eventually, the hard region stress is

$$\sigma^h = \begin{cases} 0 & \text{if} \quad \lambda \geq \lambda_{cr} \\ -k_5\, p^h\, \lambda & \text{if} \quad \lambda < \lambda_{cr} \end{cases} \tag{28}$$

We note that the size of the β-sheet was determined based on the length of a poly-Ala region from a single MaSp1 chain, however β-sheets composed of poly-Ala segments from different chains may experience similar constraints.

While the real behavior of MaSp1 chains under compression involves intricate dynamics within the chain network, which could be addressed using a more complex statistical mechanics approach (e.g., ref. [61]), the proposed phenomenological model provides a simple yet effective explanation of the dependence of supercontraction on p^h, as suggested by the EPR results. Its effectiveness is discussed later through experimental comparison.

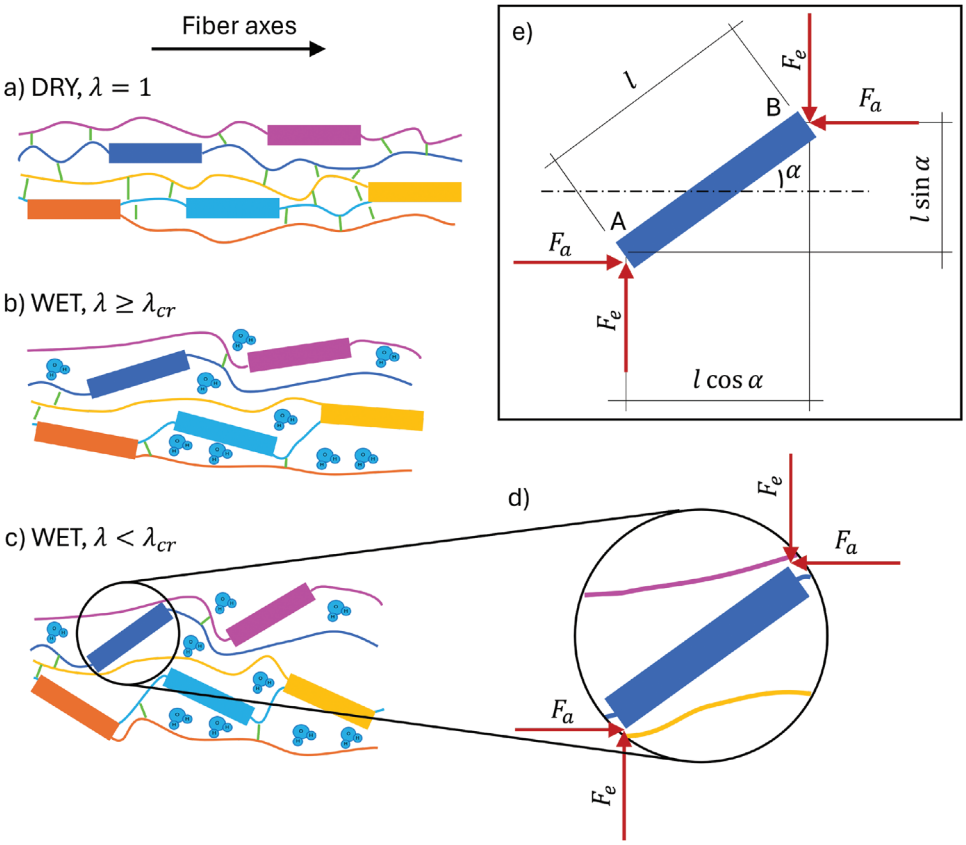

Figure 6. a) Initial dry configuration of the MaSp1 β-sheets and chains aligned along the fiber axes. Several Hydrogen bonds (green segments) prevent the chains entropic coiling. b) Final supercontracted configuration of the MaSp1 β-sheets and chains misaligned with respect to the fiber axes. Most of the H-bonds have been disrupt by the hydration water molecules (light blue). The silk stretch $\lambda \geq \lambda_{cr}$ means that the β-sheets freely rotate during supercontraction without touching the other chains. c) In the cases of most severe supercontraction ($\lambda < \lambda_{cr}$) the interaction with other chains limit the rotation of the β-sheet. d) Scheme of the forces acting on the misaligned β-sheet during the contact with other chains. e) Scheme of the geometry of the problem, where the β-sheet length l is the key variable.

2.2.3. Total Supercontraction

The interactions among the different regions of the silk thread are addressed through kinematic compatibility, meaning these regions must experience the same stretch. This approach allows for analytical results that balance model complexity and accuracy. The stress-stretch behavior of the material is described by Equation (4).

Here we focus on the case of unrestrained supercontraction, i.e., $\sigma = 0$ so that the supercontraction stretch corresponds to the equilibrium stretch uniquely defined by solving the equation

$$\sigma^s(\lambda_{sc}) + \sigma^h(\lambda_{sc}) + \mu\left(\lambda_{sc} - \frac{1}{\lambda_{sc}^2}\right) = 0 \tag{29}$$

where σ^s and σ^h are given by Equations (19) and (28) respectively.

We remark that, in the uniaxial extension case considered here, the precise spatial arrangement of MaSp1 and MaSp2 components does not affect the overall mechanical behavior, as their contributions are treated as additive.

While from a modeling perspective, the MaSp1 and MaSp2 are treated separately, as hard and soft region respectively, their primary sequences share some similarities, as described above. This may result in the presence of some amorphous portions in MaSp1 chains as well as the occurrence of β-sheets in MaSp2. Under the additive assumption here employed, the proposed model can effectively account for these cases. Specifically, if portions of MaSp1 assume an amorphous conformation along their length, these portions may behave similarly to MaSp2, undergoing entropic contraction due to hydrogen bond disruption during hydration. On the other hand, if some poly-Ala segments from MaSp2 form β-sheets, these β-sheets may experience similar limitations in their misalignment in response to hydration, as described here for the β-sheets of MaSp1. Although a more sophisticated model could explore hybrid chains in further detail, we believe that the current approach strikes a balance between simplicity and predictive capability.

3. Results and Discussion: Experimental Comparison

In order to assess the efficacy of the proposed model in providing a quantitative description of the experimental behavior, we compare it with the experimental data from the Silkome database[19] represented with dots in **Figure 7**. Specifically in Figure 7a, we analyze the effectiveness in reproducing the experimental

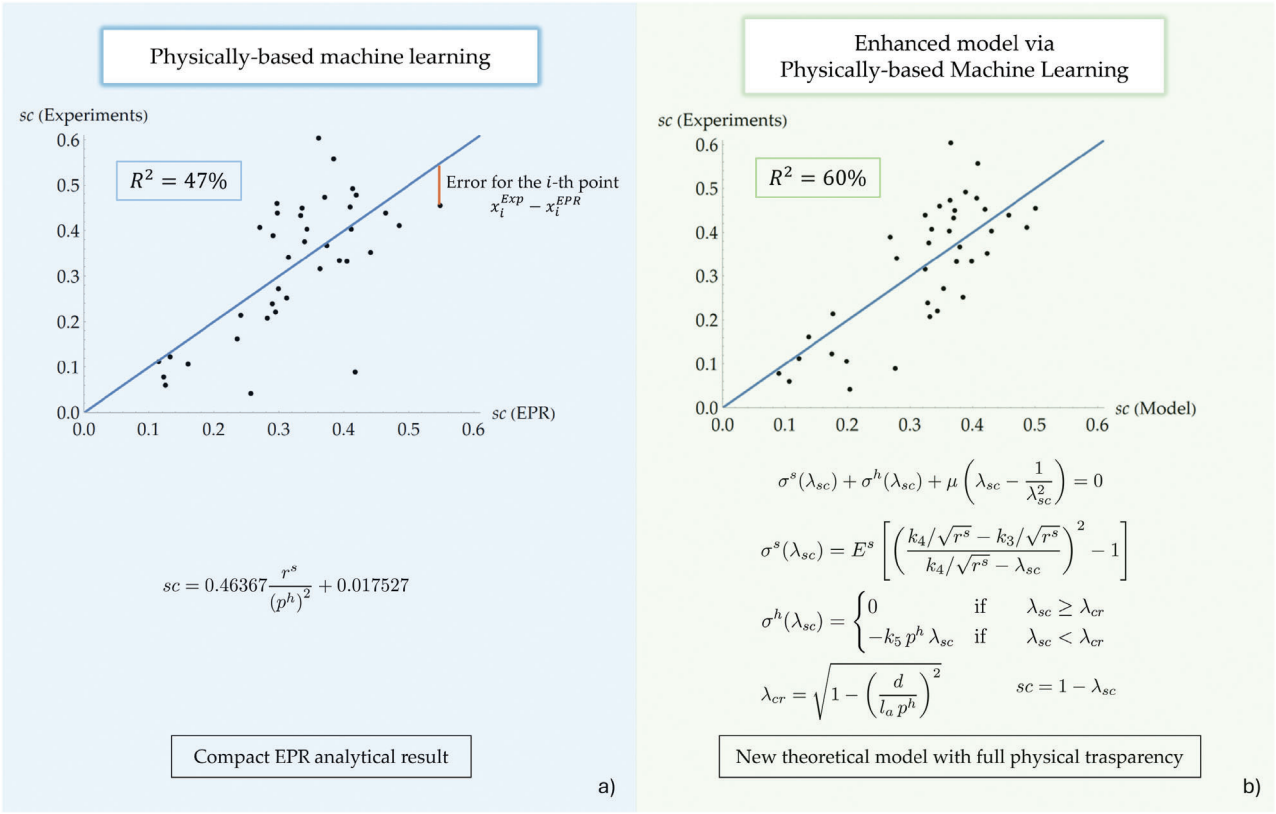

Figure 7. Comparison of supercontraction predictions ($sc = \frac{L_0 - L_f}{L_0}$) between the machine learning and the microstructure-inspired model enhanced via physically-based machine learning. a) The EPR values are calculated by using Equation (3.4). b) The model values are calculated by Equation (29), with $E^s = 10^7$ Pa the modulus of the soft region, $k_3 = 3.12$, $k_4 = 12$ to determine the natural and the contour stretch of the soft region chains respectively, $k_5 = 6.64 \times 10^5$ Pa the stiffness of the elastic response of the tube wall, $\mu = 10^5$ Pa the modulus of the elastic matrix, $d/l_a = 5.28$ (the tube diameter results $d = 2$ nm by considering the length of the alanine amino acid $l_a = 0.376$ nm). The predictive performance of the microstructure-inspired model enhanced via physically-based machine learning is higher than the physically-based machine learning model.

behavior of the analytical relation Equation (3.4) deduced by symbolic data modeling. In Figure 7b, we analyze instead the effectiveness in this quantitative reproduction by considering the theoretical model, emended through the new analytic dependence on MaSp1 and MaSp2 properties through Equation (29). In the figure we deduced the material parameters of the model and obtained the supercontraction by varying only r^s and p^h.

Interestingly, the efficacy of the model outlined in this article outperforms with $R^2 = 60\%$ as compared with the EPR results giving $R^2 = 47\%$. We remark that although the R^2 values obtained may seem lower compared to other modeling frameworks (e.g., ref. [62]), this is expected due to the high intrinsic variability of biological materials[63] including spider silks. Indeed, several factors, such as species,[5] reeling speed,[64] environmental,[6] and hydration conditions,[7] contribute to the observed mechanical property variability, even for silks taken from the same individual under similar conditions.[64] Despite this inherent variability, the results are considered satisfactory, demonstrating the robustness and feasibility of the proposed approach. Observe that the correlation coefficients obtained suggest that while the predictive models capture key aspects of the molecular-scale properties of the silks, other effects may also play a role in determining the spider silk supercontraction. These factors can be incorporated into the model following the same approach, once the necessary experimental data are available, and will be the subject of future extensions of our investigation.

The fit parameters have been optimized, in terms of R^2, starting from the ones employed in ref. [42] to reproduce the stress-stretch behavior of fibers from the spider species *Argiope trifasciata*, where they were deduced based on their micro-mechanical interpretation. As a result, here we assume $E^s = 10^7$ Pa, $\mu = 10^5$ Pa for the elastic modulus of the soft region and the modulus of the Neo–Hookean matrix respectively, in place of $E^s = 1.35 \times 10^7$ Pa, $\mu = 1.4 \times 10^5$ Pa employed in ref. [42]. The contour stretch of the soft region λ_c^s here depends on the MaSp2 repeat length r^s and it results to be within the range (1.43,2.53), whereas in ref. [42] it was employed $\lambda_c^s = 1.62$. Similarly, the natural stretch in full wet conditions that here depends on r^s, is in the range (0.38,0.67), while $\lambda_n^s = 0.35$ in full wet conditions in ref. [42].

On the other hand, the parameter d, assigning the tube diameter for the hard region, to the knowledge of the authors, was never estimated before. Here we assume a value $d = 2$ nm, comparable with the chain primary structure thickness.[65] The value assumed for the alanine amino acid length within a β-sheet configuration is $l_a = 0.376$ nm.[66] Similar considerations apply to the stiffness of the elastic response of the tube wall that here is

estimated to be $k_5 = 6.64 \times 10^5$ Pa and interestingly it is in the order of magnitude of the modulus estimated for the elastic matrix.

Further, we provide some explicit examples by predicting the supercontraction of specific silks from several spiders, namely, species of *Argiope, Araneus, Trichonephila* and *Caerostris*. Here we describe in detail the procedure for calculating the amount of supercontraction based on the experimental values of r^s and p^h. We will examine step by step what assumptions need to be made and how to calculate the contributions of the various fractions composing the silk. Eventually, we will compare the calculated value with the experimental one for the corresponding silk.

For all the silks, we have employed the set of parameters found to best fit the data of the Silkome database, indicated in Figure 7. In particular, for the soft region contribution, explicitly given by Equation (19), we assumed $E^s = 10^7$ Pa, $k_3 = 3.12$, $k_4 = 12$. For the hard region contribution, given by (28), we assumed $k_5 = 6.64 \times 10^5$ Pa, and $d/l_a = 5.28$ to compute the critical value λ_{cr} from Equation (22). Additionally, we assumed $\mu = 10^5$ Pa for the elastic matrix. Eventually, we calculated the supercontraction of dragline silk from each spider by solving Equation (29), by employing the experimental values of p^h and r^s computed from the sequence in the Silkome database (reported in the Appendix B) for each considered spider genus (recalled in **Table 2**).

The obtained values predicting the supercontraction are reported in Table 2, together with the corresponding experimental result (from Silkome database) and the relative percentage error.

For all the cases, the supercontraction is calculated with satisfactory accuracy. We remark that once the "super parameters" reported in Figure 7 and recalled above have been determined, predicting supercontraction for a new spider silk requires only the values of p^h and r^s. Of course, for a more refined prediction, one may also use different values for one or more 'super parameters' if they are available, for instance, through experimental data or numerical simulations.

Furthermore, in **Table 3**, we report the stress in the maximum supercontracted configuration of each fraction in which we have schematized the silk thread. We note that in three cases out of four (*Argiope, Araneus, Caerostris*) $\lambda_{sc} < \lambda_{cr}$, this means that the hard region is actually contrasting with the shortening of the soft region, whereas for the case of *Trichonephila*, $\lambda_{sc} > \lambda_{cr}$ so the hard region does not come into play and the equilibrium is reached between the soft region and the matrix only.

Table 2. Comparison between the supercontraction values (*sc*) predicted by the model and experimentally measured for MA silks of different spider genera, with corresponding relative error. The predicted values were obtained by solving Equation (29), by employing the reported experimental average values of r^s and p^h and a fixed set of parameters: $E^s = 10^7$ Pa, $k_3 = 3.12$, $k_4 = 12$, $k_5 = 6.64 \times 10^5$ Pa, $d/l_a = 5.28$, $\mu = 10^5$. All the experimental values are taken from the Silkome database.[19]

Genus [Silkome ID]	p^h	r^s	*sc*	*sc*	Error [%]
	Experiments	Experiments	Model	Experiments	
Argiope [3653]	8	49	0.377	0.367	2.8
Araneus [4487]	8	47	0.361	0.333	8.5
Trichonephila [6762]	6	32	0.429	0.403	6.5
Caerostris [7439]	8	62	0.459	0.439	4.6

Table 3. Supercontraction stretch compared with the critical stretch for MA silks of the various spider genera considered as explicit examples and resulting stress contributions for the different fractions schematically composing the silk thread.

Genus [Silkome ID]	λ_{sc}	λ_{cr}	Soft [MPa]	Matrix [MPa]	Hard [MPa]
Argiope [3653]	0.623	0.751	3.50	−0.19	−3.31
Araneus [4487]	0.639	0.751	3.57	−0.18	−3.39
Trichonephila [6762]	0.571	0.474	0.25	−0.25	0
Caerostris [7439]	0.541	0.751	3.16	−0.29	−2.87

We note that in this work, we assume that the MaSp1 poly-Ala segments form β-sheet crystals, which cannot freely rotate during supercontraction due to the constraints imposed by other chains in the protein network. This is a complex phenomenon, so we introduced a simplified orientational effect, as illustrated in Figure 6. Additionally, we have modeled the MaSp2 chains as a series of spherical blobs, each representing a repeat unit of MaSp2, following previous models from De Gennes.[59] Bonds between the chains, which may be disrupted by hydration, are assumed to be located on the surface of the blobs. While these simplifying assumptions enable us to derive analytical results, it is important to highlight that they only phenomenologically address certain relationships between the primary protein structures and the observed behaviors of the silk fibers. Such relationships, derived from machine learning insights, were completely neglected in our previous analytical model,[42] and, to the best of our knowledge, in other existing models of spider silk behavior.

4. Conclusion

In this paper, we have conducted a comprehensive study of the supercontraction behavior of spider silk, building on previous models and incorporating new insights derived from a physically-based machine learning method, specifically, a symbolic data modeling technique.

We reviewed the microstructure of spider silk, examining how humidity affects its properties, and assessed various models for predicting supercontraction, including a microscopically motivated model, recently proposed by the authors, representing silk fiber as composite material composed of an elastic matrix embedding a hard external fraction of crystalline chains and a soft internal fraction of amorphous chains.

Next, we conducted more detailed physically-based machine learning analyses (EPR) compared to preliminary studies on the dependence of supercontraction on the protein primary sequence and provided a physical interpretation of the insights obtained through machine learning. We then extended the considered microstructure inspired model to incorporate these new data-driven insights, focusing specifically on the dependence of supercontraction on the repeat length of the MaSp2 protein as suggested by EPR. The specific role of the MaSp2 repeat length is to regulate the number of cross-link that keep the amorphous chain in a frozen state in dry conditions and that can be disrupted by the hydration water molecules leading to the entropic coiling of the chains and therefore to the fiber contraction. Furthermore, we addressed the significant dependence of supercontraction on

the length of the MaSp1 polyalanine region length, again suggested by the EPR analysis. According to the proposed model, the polyalanine region length, which is indicative of the β-sheet size, may regulate the contraction that the silk can undergo. This is because the misalignment of the crystals, which is required to accommodate the contraction of the soft region, may be limited due to the presence of other chains for long crystals. We then evaluated the supercontraction of the entire silk thread based on lower scale parameters including the new effects of the protein sequences properties, under unrestrained conditions.

To assess the efficacy of the proposed model, we compared experimental data from the Silkome database with predictions from both the physically-based machine learning derived equation and the proposed theoretical model emending the one proposed in ref. [42] by including the new physical considerations deduced based on EPR results. The latter demonstrated superior performance in quantitatively describing the experimental behavior, with many of the optimal fit parameters aligning with those used in previous studies for modeling the spider silks thermo-hygro-mechanicanical behavior. Furthermore, we have reported some examples of explicit calculations to obtain a prediction of the supercontraction amount. The predicted values were in agreement with the experimental ones, with errors of only a few percentage points. In summary, we employed an interpretable machine learning approach, i.e., EPR, to glean novel insights from a database of experimental observations, integrating these insights into a pre-existing physically-based model. This enhanced model provides a more accurate and comprehensive understanding of supercontraction, bridging the gap between the primary structure of silk proteins and their macroscopic behavior, thereby advancing the knowledge and the availability of new theoretical models.

Furthermore, a new research paradigm is outlined by the methodology proposed in this work. While, traditionally, experimental observations inspire models of phenomena, which then suggest further experiments, this strategy not only enables experiments to directly inspire models, but also employs physically-based machine learning algorithms to generate interpretable models of data. These data-driven models can enhance physically-motivated models or even lead to the creation of new ones. Consequently, this process fosters a continuous cycle in which machine learning is integrated within the established framework of physically-based modeling. This methodology differs from the prevailing practice of machine learning, which typically generates non-interpretable black-box models that may yield intriguing outcomes especially in the objective of data fitting, but provide little insight into their functioning. By integrating interpretable, physically-based machine learning models, such new research paradigm may open new avenues for both theoretical and experimental research in materials science.

Appendix A

The amorphous regions of the Spidroins are typically considered responsible for supercontraction. For this reason, we attempted to evaluate the lengths of the repetitive units excluding the polyalanine segments that may form β sheets, specifically the quantities $a^h = r^h - p^h$ and $a^s = r^s - p^s$. An EPR analysis was performed using the input candidates a^h, a^s, p^h, p^s, with the results presented in terms of Pareto fronts of models and their respective accuracies in Appendix A. As shown in the Table A1, the performance does not improve compared to the case where r^h, r^s, p^h, p^s were considered. Therefore, the latter set was considered in the analysis.

Table A1. Equations contained in the Pareto Front returned by EPR using candidate inputs a^h, a^s, p^h, p^s.

Equations on the Pareto front	R^2 (%)	
$sc = 0.32479$	0	(A.1)
$sc = 0.058026(a^s)^{0.5}$	25.1	(A.2)
$sc = 0.0060527a^s + 0.12523$	27.4	(A.3)
$sc = 0.48456\frac{a^s}{(p^h)^2} + 0.068986$	46.2	(A.4)
$sc = 0.46747\frac{(a^s)^{0.5}}{p^h}$	44.7	(A.5)
$sc = 9.3064\frac{1}{(p^h)^2} + 0.00017861a^h a^s$	48.5	(A.6)
$sc = 0.66812\frac{(a^h)^{0.5}(a^s)^{0.5}}{(p^h)^2}$	48.0	(A.7)
$sc = 6.8657\frac{1}{(p^h)^2} + 0.009555\frac{(a^h)^{0.5}a^s}{p^h}$	51.6	(A.8)
$sc = 6.8051\frac{1}{(p^s)^2} + 0.0096324\frac{(a^h)^{0.5}a^s p^s}{(p^h)^2}$	51.9	(A.9)
$sc = 181.6428\frac{1}{a^h(p^s)^2} + 0.0017683\frac{a^h a^s p^s}{(p^h)^2}$	55.1	(A.10)
$sc = 1188.4947\frac{1}{a^h p^h (p^s)^2} + 0.0018531\frac{a^h a^s p^s}{(p^h)^2}$	55.8	(A.11)

Appendix B

Amino Acid Sequences from Silkome Database[19]

Argiope aetheroides (ID Silkome 3653)

MaSp1 (C-terminal region)

AAGGQGGQGGYGGLGSQGAGGAGQGGYGGGQGGAGSAAAAAAAG
GQGGQGGYGGLGSQGAGGAGQGGSGAAAAAAAAGGAGGAGRGGLGS
GGAGQGYGGLGSQGAGGAGQGGYGAGQGGAGAAAAAAAAGGQGG
QGGYGGLGSQGAGGAGQGGYGGGQGGAGSAAAAAAAGGQGGQGG
YGGLGSQGAGGAGQGGSGAAAAAAAAGGAGGAGRGGLGSGGAGQG
YGGLGSQGAGGAGQGGYGAGQGGAGAAAAAAAAGGQGGQGGYGGL
GSQGAGGAGQGGYGAGQGGAGAAAAAASAGGQGGQGGYGGLGSQG
AGGAGQGGSGAAAAAAAAGGAGGAGRGGLGSGGAGQGYGGLGSQG
AGGAGQGGYGAGQGGAGAAAAAAAAGGQGGQGGYGGLGSQGAGG
AGQGGYGSGQGGAGAAAAAAAAGGQGGQGGYGGLGSQGAGQGGYG
GGAYGGQGAAASASAASAAASRLSSPGAASRVSSAVTSLVSSGGPTNS
AALSNTISNVVSQISASNPGLSGCDILVQALLEIVSALVHILGSANIGQVN
SSAAGQSASLVGQSVYQALS

MaSp2 (N-terminal region)

MNWSIRLALLGFVVLSTQTVFAVGQAATPWENSQLAEDFINSFLRFIGQ
SGAFSANQLDDMSSIGDTLKTAIEKMAQSRKSSKSKLQALNMAFASSMAE
IAVAEQGGLSLQAKTDAIANALSSAFLETTGYVNQQFVNEIKSLIFMIAQA
SANEISGSYAAAGGSSGGGSYGGGGSGQGGYGQGAYASASAAVAYGSA
PQGAGGPAPQGPSQQGPVSQGPYGPGAAAAAAGTGGYGPGAGQQGQ
QGPGSQGPGSGGQQGPSQGQPYGPSAAAAAAAAGGYGPGAGQQGPG
GAGQQGPGSQGPGGAGQQGPGGQGPYGPGAAAAAAAAGGYGPGAGQ
QGPGSGGQQGPGGLGSGQQGPGGAGQQGPGGQGPYGPGAAAAAA
AAGGYGPGAGQQGPGSGGQQGPGGLGSGQQGPGGAGQQGPGGQGP
YGPGAAAAAAAAGGYGPGAGQQGPGGAGQQGPGGQGPYGPGAAAAA
AAAGGYGPGAGQQGPGSGGQQGSGGQGSGQQGPGGAGQQGPGGQ
GPYGPGAAAAAAAAGGYGPGAGQQGPGSQGPGGAGQQGPGGQGP
YGPGAA

MaSp2 (C-terminal region)

AAAAAAASGPGGYGPGSQGPSGPGGAGGYGPGSQGAGGAGAAAAAA
AASGPGGYGPGSQGPSGPGGYGPGSQGPSGPGGYGPGASGPGGAGGY
GPGSQGPGGASAAAAAAASGPGGYGPGSQGPSGSGSQGPSGSGGAGGYG
PGSQGPGGASAAAAAAAASGPGGYGPGSQGPSGPGSQGPSGPGGYGPGS
QGPSGPGGYGPGASGPGGAGGYGPGNQGSGGASAAAAAAAASGPGGYG

PGSQGPSGPGSQGPSGPGGYGPGASGPGGAGGYGPGSQGPGGASAAAA
AAASGPGGYGPGSQGPSGSGAYGPGSQGPSGYGGYGPSASASVSSA
ASRLSSPAASSRVSSAVSSLVSNGASNGASVSGALNGLVSQISSSNPGLSGC
DVLVQALMELVSALVAILGSASIGSVDYNSVGQTTQTISQYFS

Araneus macacus (ID Silkome 4487)
MaSp1 (N-terminal region)
MTWTARLALSLLAVICSQSLFALGQSPWQNARMAENFMSSFSSALGQ
SGAFSSDQMDDIMSICDSIQSGVDRMDRSGKTSANKLQAMNMAFASAV
AEIAIAEGGGQSAQVKTNAVADALASAFLQTTGVVNTQFVNEIRTLISMF
AQANVVSSSSASVSASTGGAGGYGPQAQGAASAAVSTSAQGGYG
PGPSGRGPQGPGPQTPGTASVSISTAAQGGYGQGPQSYGPGPQGPSGT
GQQGPYGPGPQGPSGPGPQGPGQSSYQPSISINSQSGSQGPSGGQRGY
GGGQGGAGSAAAAAAAGGAGQGGLGAGGAGQEYGAGLGGQGGVGQ
GGGAAAAAAAGGQGGQGGYGGLGSQGAGQGYGGGQGGAGSAAAA
AAAGGAGQGGLGAGGAGQGYGAGLGGQGGSGQGGAAAAAAAGGQ
GGQGLYGGLGSQGAGQGGYGDGQGGAGSAAAAAAGGAGGAGRGG
SGAGGAGQGYGAGLGGQGGAGQGGGAAAAAAAGGQGGQAG
YGGLGSQGAGQGYGGGQGGAGSAAAAAAAGGAGQGGLGAGGAGQGY
GAGLGGQGGSGQGGAAAAAAAGGQGGQGLYGGLGSQGAGQGGYGD
GQGGAGSAAAAAAAGGAGGAGRGGSGAGGAGQGYGAGLGGQGGA
GQSGGAAAA

MaSp1 (C-terminal region)
GYGGGQGGAGSAAAAAAAGGAGQGGLGAGGAGQGYGAGLGGQGG
SGQGGAAAAAAAGGQGGQGLYGGLGSQGAGQGGYGGGQGGAGSAA
AAAAAGGAGQGRLGSGGAGQGYGAGLGGQGGSGPGGAAAAGGQGG
QGGYGGLGSQGAGQGGYGGGQGGAGSAAAAAAAGGAGDAGRGGS
GSGGAGQGYGAGLGGQGLAGQGGGAAAAAAAGGQGGQGGYGGLG
SQGAGQGYGGGQGGAGSAAAAAGGAGQGGLGAGGAVQGYGAGLG
GQGGSGQGGSASAAAAVGGQGGQGGYGGLGSQLAG
QGGYGGRQGGAGSAAAAAAAGGAGQGGLGAGGAGQGYAAGLGG
QGGAGQGGAAAAAAAAGGQGGYGGLSSQGAGGAGQGGYGGGAYG
GQGAASSAAAASSAASRLSSPSAASRVSSAVSSLVSSGGPSSPAALS
STISNVVSQISASNPGLSGCDVLVQALLEIVSALVHILGSATIGQV
NSSAAGQSASLVGQSVYQALS

MaSp2 (N-terminal region)
MSSSRLALAFLALLCTNALFVAAGGPTPWDSPNMAEAFMNNFMSGIA
SSGAFSGGQMGDMQDIAGTMQDSVNKMASTGRSSKSKLQAMNMAFA
SSMAEIAAAEEGGSSMAAKTSAITNALRGAFLQTTGVSNEQFINEIATL
INLISQSNVNTVSASASSGGGGGGYGGPAYGPSSYGPSQGASSVSVSA
SAAGGGSGGQGPSGSPQQGPGGYGPSGSGASVAVAAAGGQGPSGPG
PQGPSGPGGYGPSSQGPSGPGASSAAAAASASGPSGYGPGSQGPSGPS
RTGGYGPGSQGPSGPSGPGASAAAAAAASGPGGYGSGSQGPSGPGGL
GSGSQGPSGPGGYGPGSQGPGGPGAAAAAAAAAASGPGGYGPGS
QGPSGPGGYGPGSQGPSGPGGPGASAAAAAASGPGGQGPSGPGG
PGGYGPGSQGPGGPGGYGPGSQGPGGPGASAAAAAAAAASGPGGYG
PGSQGPSGP

MaSp2 (C-terminal region)
GQQGPGQQGPGGQGPYGPGAAAAAAAAGGYGPGSGQQGPGQQG
PGQQGPGQQGPGQQGPGGQGPYGPGAAAAAAAAGGYGPGSGQQGP
GQQGPGQQGPGGQGPYGPGASAAAAAAGGYGPGSGQQGPGQQGP
GQQGPGQQGPGQQGPGQQGPGGQGPYGPGASAAAAAAGGYGPGP
GQQGPGQQGPGQQGPGGQGPYGPGAASAAVSVGGYGPQRSSAPVASA
AASRLSSPAASSRVSSAVSSLVSSGPANPAALSSTISNAVSQISASNPGLS
GCDVLVQALLEVVSALVHILGSSSIGQINYGASSQYAQMVGQSV
AQALG

Trichonephila plumipes (ID Silkome 6762)
MaSp1 (N-terminal region)
MTWTARLALSILAVLCTQGMFAQGQNTPWSSTELADAFINAF
MNEAGRTGAFTADQLDDMSTIGDTIKTAMDKMARSNKSSKGKLQA
LNMAFASSMAEIAAVEQGGLSVDAKTNAIADSLNSAFYQTTGAANPQFV
NEIRSLIKMFAQSSANEVSYGGGYGGGQGGQSAAAAAAVGSAGQGGY
GGLGSQGAGRGGYGGQGAGAAAAAAGGAGQGGQGLGGQGAGRGAG
AAAAAAGGAGQGGYGGLGGQGAGAAAAAGGAGQGGYGGQGAGRGAG
AAAAAAGGAGQGGLGGQGAGQGAGAAAAAAAAGGAGQGGYGGLG
SQGAGRGGYGGQGAGAAAAAAGGAGQGGQGLGGQGAGRGAGAAA
AAAGGAG

MaSp2 (N-terminal region)
MSWSTLALAILAVLSTQSIYASAQARSPWSDTATADAFIQNFLA
AVSGSGAFSSDQLDDMSTIGDTIMSAMDKMARSNKSSQHKLQALNMAF
ASSMAEIAAVEQGGMSMGVKTNAIADALNSAFYMTTGAANPQFVNEM
RSLISMISAASANEVSYGGGSSAASAAAAAGSYGQGPSGYGQGAS
VSSAAAAAPSGYGPSQQGPSGPAAAAAAAAGQAQQGPSAYGPSGQG
AAAAAASGPGSYGPQQGPGPQRPSGYGPSGPSGPSSAAAAAAAAGAGPG
GYGPQQQGPGQQGPSGYGPSGPSGPGSAAAAAASAGAGPGG
YGPQQQGPGQQGPSGYGPSGPSGPGSAAAAAAGAGPGGYGPG
QQGPGQQGPSGPGSAAAAAAAAGAGPGGYGAGQQGPGQQGPG
QQGPGRYGPSGPGSAAAAAA

MaSp2 (C-terminal region)
GPGGAAAAAAAAGPGGYGPGQQGPGQQGPAGYGPSGPGSAAAAA
AAAAGPGGYGPGQQGPSGPGGAAAAAAAAGPGGYGPGQQGPGQQ
GPAGYGPSGPGSAAAAAAAAAGPGGYGPGQQGPSGPGGAAAAAAAAA
GPGGYGPGQQGPGQQGPGQQGPAGYGPSGPGAAAAAAAAAAAGPGG
YGPGQQGPGQQGPAGYGPSGPGGAAAAAAAAGPGGFGPGQQGPGQQ
GPAGYGPSGPGGAAAAAAAAGPGGYGPGQQGPSGPGSAAAAAAAG
PGSYGPSQQGPARYGPSAPGSAAAAAAGAGTAGYG
PGAQASAAASRLASPDSGARVASAVSNLVSSGPTSSAALSSVIS
NAVSQIGASNPGLSGCDVLIQALLEIVSACVTILSSSSIGQ
VNYGAASQFAQVVSQSILSAF

Caerostris darwini (ID Silkome 7439)
MaSp1 (N-terminal region)
MTWTSRLALSLLVAICTQSMFALGQDNTPWSSTGTAESFMSSFMSAA
GNSGAFTADQLDDMNTITDTIRSAMDKMARSNKSSKSKLQAL
NMAFSSAMAEIAIDEGGQSVGYKTDAIADALSQAFLQ
TTGVVNGAFINEIRSLISMFAQNSANAIGSGGSSASVSVAASAGGGY
GGQGSYGPGPQGPSGPGPQGAGASSASAVSAASGPGGYAPG
PQGPSGPQGPGQSSYQYSVSISTQGGSQGGYGGQQGGAGQGGYGGG
LGGQGAGAAAAAAAAGGAGGLGGRGGGQGAGQGGYGSGQGGQ

MaSp1 (C-terminal region)
GQGAGGAGQGGYGSGLGGLGGGAAAAAAAAGGAGGLGGQGGG
QGAGQGGYGSGQGGQGAGSAAAAAAAGGAGGRGGYGGQGGQGAGGA
GQGGYGSGLGGLGGGAAAAAAAAGGAGGLGGQGGG
QGAGQGGYGSGQGGQGAGSAAAAAAAGGAGGRGGLGGQGGGQGAG
GAGQGGYGSGLGGLGGGAAAAAAAAGGAGGLGGQGGG
GGQGAGQGGYGSGQGGQGAGSAAAAAAAGGRGGYGGQGGGQGAGG
AGQGGYGSGLGGLGGGAAAAAAAAGGAGGLGGQGGGQ
VAGQGGYGSGQGGQGAGSAAAAAAAGGAGGRGGLGGQGG
GQGAGGAGQGGYGSGLGGLGGGASAAAAAAGGAGGLGGQGGG
QGAGQGGYGSGQGGQGAGSAAAAA
AAGGSGGLGGQGGYGGQGGYGGGYGGQQVAASATTASAAASRL
SSPAASSRVSSAVSSLVSSGPTSPAALSNTISNVVSQVGASNPGLSGCD
VLVQALLEIVSALIHILGSSSIGQVNYGATAQSTGIVSQSISQALG

MaSp2 (C-terminal region)
PYGPGGAAAAAAAAGGYAPAGQGQSGPGSQGQGQSGPGS
QGPGGAGPYGPGGAAAAAAAAGGYGPGGQGPSGPGSQGPGS
QGPSGPGLAAAAAAAGGYGPGGQGPSGSASQGPGGQGPYGPGGAAAA
AAAAGGYGPGGSGPSGPGSQGPSGPGSQGPGGAGPYGPGGAAAA
AAAAGGYGPGSQGPSGPGSQGPGGRGPSGPGSQGPGGAGPYGPGGA
AAAAAAAGGYGPGSQGPSGPGSQGPGGQGPYGPGGAAAAAAAAG
GYGPGSQGPSGPGSQGPGGQGPYGPSGAAAAAAAAGGYG
PAGQGPSGPGSQGPGGQGPSGPGGYGPSSAAAAFGGYGPSGQIPSAA
AAASRLSSPAVASRVSSTVSSLVSSGPTSQGALSNAISNAVSQISA
SNPGLSGCDVLVQALLEIVSALVHILGSSSVGQVSYNTAGQSAAVV
SQSISQALG

Acknowledgements

V.F. and G.P. were supported by GNFM and INdAM. G.P. and V.F were supported by the Project of National Relevance (PRIN), financed by European UnionEU - Next- GenerationEU - National Recovery and Resilience Plan - NRRP - M4C2 - I 1.1, CALL PRIN 2022 PNRR D.D. 1409 14-09-2022 - (Project code P2022KHFNB, CUP D53D23018910001)

granted by the Italian MUR. G.P. was supported by the Italian Ministry MIUR-PRIN project 2017KL4EF3 and by PNRR, National Center for HPC, Big Data and Quantum Computing (CN00000013) – Spoke 5 "Environment and Natural Disasters," G.P.'s research was funded by the European Union - Next Generation EU. G.P. was supported by the Research Project Prin2022 of National Relevance 2022XLBLRX granted by the Italian MUR. N.M.P. acknowledge the financial support of the European Union - Next Generation EU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 4 Componente 2, Investimento N. 1.1, Bando PRIN 2022 D.D. 104/02-02-2022 - (PRIN 2022 2022ATZCJN AMPHYBIA) CUP N. E53D23003040006. the Italian Ministry of University MUR under PRIN-20177TTP3S. K.N. was supported by The Ministry of Education, Culture, Sports, Science and Technology (MEXT): Data Creation and Utilization-Type Material Research and Development Project (Grant Number JPMXP1122714694).

Open access publishing facilitated by Universita degli Studi di Trento, as part of the Wiley - CRUI-CARE agreement.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Spider Silkome Database at https://www.science.org/doi/10.1126/sciadv.abo6043, reference number 19. These data were derived from the following resources available in the public domain: https://spider-silkome.org.

Keywords

data modeling, micromechanically-based models, spider silk, supercontraction

Received: October 22, 2024
Revised: December 4, 2024
Published online: December 26, 2024

[1] N. Zhao, Z. Wang, C. Cai, H. Shen, F. Liang, D. Wang, C. Wang, T. Zhu, J. Guo, Y. Wang, X. Liu, C. Duan, H. Wang, Y. Mao, X. Jia, H. Dong, X. Zhang, J. Xu, *Adv. Mater.* **2014**, *26*, 6994.
[2] J. Pérez-Rigueiro, M. Elices, G. R. Plaza, G. V. Guinea, *Molecules* **2021**, *26*, 1794.
[3] R. W. Work, *Text. Res. J.* **1977**, *47*, 650.
[4] V. Fazio, N. M. Pugno, G. Puglisi, *Extreme Mechanics Letters* **2023**, *61*, 102010.
[5] C. Boutry, T. A. Blackledge, *J. Exp. Biol.* **2010**, *213*, 3505.
[6] G. R. Plaza, G. V. Guinea, J. Pérez-Rigueiro, M. Elices, *J. Polym. Sci., Part B: Polym. Phys.* **2006**, *44*, 994.
[7] I. Agnarsson, C. Boutry, S.-C. Wong, A. Baji, A. Dhinojwala, A. T. Sensenig, T. A. Blackledge, *Zoology* **2009**, *112*, 325.
[8] G. Greco, T. Arndt, B. Schmuck, J. Francis, F. G. Bäcklund, O. Shilkova, A. Barth, N. Gonska, G. Seisenbaeva, V. Kessler, J. Johansson, N. M. Pugno, A. Rising, *Commun. Mater.* **2021**, *2*, 1.
[9] M. Elices, G. R. Plaza, J. Pérez-Rigueiro, G. V. Guinea, *J. Mech. Behav. Biomed. Mater.* **2011**, *4*, 658.
[10] A. Sponner, W. Vater, S. Monajembashi, E. Unger, F. Grosse, K. Weisshart, *PLoS One* **2007**, *2*, e998.
[11] S. Keten, M. J. Buehler, *J. R. Soc., Interface* **2010**, *7*, 1709.
[12] J. E. Jenkins, S. Sampath, E. Butler, J. Kim, R. W. Henning, G. P. Holland, J. L. Yarger, *Biomacromolecules* **2013**, *14*, 3472.
[13] S. W. Cranford, N. M. Pugno, M. J. Buehler, *Silk and Web Synergy: The Merging of Material and Structural Performance*, Springer, Netherlands, ISBN 9789400771192, **2013**, pp. 219–268.
[14] R. C. Chaw, S. M. Correa-Garhwal, T. H. Clarke, N. A. Ayoub, C. Y. Hayashi, *J. Proteome Res.* **2015**, *14*, 4223.
[15] M. A. Collin, T. H. Clarke, N. A. Ayoub, C. Y. Hayashi, *Int. J. Biol. Macromol.* **2018**, *113*, 829.
[16] P. L. Babb, N. F. Lahens, S. M. Correa-Garhwal, D. N. Nicholson, E. J. Kim, J. B. Hogenesch, M. Kuntner, L. Higgins, C. Y. Hayashi, I. Agnarsson, B. F. Voight, *Nat. Genet.* **2017**, *49*, 895.
[17] C. Larracas, R. Hekman, S. Dyrness, A. Arata, C. Williams, T. Crawford, C. Vierra, *Int. J. Mol. Sci.* **2016**, *17*, 1537.
[18] A. D. Malay, H. C. Craig, J. Chen, N. A. Oktaviani, K. Numata, *Biomacromolecules* **2022**, *23*, 1827.
[19] K. Arakawa, N. Kono, A. D. Malay, A. Tateishi, N. Ifuku, H. Masunaga, R. Sato, K. Tsuchiya, R. Ohtoshi, D. Pedrazzoli, A. Shinohara, Y. Ito, H. Nakamura, A. Tanikawa, Y. Suzuki, T. Ichikawa, S. Fujita, M. Fujiwara, M. Tomita, S. J. Blamires, J.-A. Chuah, H. Craig, C. P. Foong, G. Greco, J. Guan, C. Holland, D. L. Kaplan, K. Sudesh, B. B. Mandal, Y. Norma-Rashid, et al., *Sci. Adv.* **2022**, *8*, 41.
[20] A. D. Malay, K. Arakawa, K. Numata, *PLOS One* **2017**, *12*, 0183397.
[21] B. L. Thiel, K. B. Guess, C. Viney, *Biopolymers* **1997**, *41*, 703.
[22] J. D. van Beek, S. Hess, F. Vollrath, B. H. Meier, *Proc. Natl. Acad. Sci.* **2002**, *99*, 10266.
[23] A. Sponner, E. Unger, F. Grosse, K. Weisshart, *Nat. Mater.* **2005**, *4*, 772.
[24] A. Nova, S. Keten, N. M. Pugno, A. Redaelli, M. J. Buehler, *Nano Lett.* **2010**, *10*, 2626.
[25] S. Li, A. McGhie, S. Tang, *Biophys. J.* **1994**, *66*, 1209.
[26] L. Eisoldt, A. Smith, T. Scheibel, *Mater. Today* **2011**, *14*, 80.
[27] S. Sonavane, S. Hassan, U. Chatterjee, L. Soler, L. Holm, A. Mollbrink, G. Greco, N. Fereydouni, O. Vinnere Pettersson, I. Bunikis, A. Churcher, H. Lantz, J. Johansson, J. Reimegård, A. Rising, *Sci. Adv.* **2024**, *10*, 33.
[28] K. Yazawa, A. D. Malay, H. Masunaga, Y. Norma-Rashid, K. Numata, *Commun. Mater.* **2020**, *1*, 1.
[29] C. P. Brown, J. MacLeod, H. Amenitsch, F. Cacho-Nerin, H. S. Gill, A. J. Price, E. Traversa, S. Licoccia, F. Rosei, *Nanoscale* **2011**, *3*, 3805.
[30] T. Giesa, R. Schuetz, P. Fratzl, M. J. Buehler, A. Masic, *ACS Nano* **2017**, *11*, 9750.
[31] J. E. Jenkins, M. S. Creager, E. B. Butler, R. V. Lewis, J. L. Yarger, G. P. Holland, *Chem. Commun.* **2010**, *46*, 6714.
[32] K. N. Savage, J. M. Gosline, *J. Exp. Biol.* **2008**, *211*, 1948.
[33] F. Vollrath, D. P. Knight, *Nature* **2001**, *410*, 541.
[34] R. W. Work, *J. Exp. Biol.* **1985**, *118*, 379.
[35] A. D. Parkhe, S. K. Seeley, K. Gardner, L. Thompson, R. V. Lewis, *J. Mol. Recognit.* **1997**, *10*, 1.
[36] K. M. Bonthrone, F. Vollrath, B. K. Hunter, J. K. M. Sanders, *Proc. Royal Soc. London. Series B: Biologic. Sci.* **1992**, *248*, 141.
[37] G. P. Holland, R. V. Lewis, J. L. Yarger, *J. Am. Chem. Soc.* **2004**, *126*, 5867.
[38] R. Ene, P. Papadopoulos, F. Kremer, *Polymer* **2011**, *52*, 6056.
[39] P. T. Eles, C. A. Michal, *Macromolecules* **2004**, *37*, 1342.
[40] Y. Termonia, *Molecular Modeling of the Stress/Strain Behavior of Spider Dragline*, Elsevier, **2000**, pp. 337–349.
[41] G. Puglisi, D. De Tommasi, M. F. Pantano, N. M. Pugno, G. Saccomandi, *Phys. Rev. E* **2017**, *96*, 042407.
[42] V. Fazio, D. De Tommasi, N. M. Pugno, G. Puglisi, *J. Mech. Phys. Solids* **2022**, *164*, 104857.
[43] N. Cohen, M. Levin, C. D. Eisenbach, *Biomacromolecules* **2021**, *22*, 993.
[44] V. Fazio, N. M. Pugno, O. Giustolisi, G. Puglisi, *Cell Rep. Phys. Sci.* **2024**, *5*, 101790.
[45] W. Lu, D. L. Kaplan, M. J. Buehler, *Adv. Funct. Mater.* **2023**, *34*, 11.

[46] Y. Kim, T. Yoon, W. B. Park, S. Na, *J. Mech. Behav. Biomed. Mater.* **2023**, *140*, 105739.
[47] E. Gibney, D. Castelvecchi, *Nature* **2024**, *634*, 523.
[48] As reported in ref. [19], repetitive regions of spidroin sequences were extracted as the longest segments containing amino acid motifs of serine (S), alanine (A), or valine (V) that were longer than four residues. Such regions were divided into repeat units separated by SAV motifs longer than five residues. This SAV region was classified as the crystalline region, while the remaining amino acids within the repeat were designated as the amorphous region. Typically, the polyalanine region is characterized by stretches of multiple A, S, and V residues exceeding five amino acids, as these tend to substitute for polyalanine. Note that in the Silkome database, the "N-terminal" and "C-terminal" regions include portions of the repetitive sequence adjacent to their respective terminal domains.
[49] O. Giustolisi, D. A. Savic, *J. Hydroinformatics* **2006**, *8*, 207.
[50] O. Giustolisi, D. A. Savic, *J. Hydroinformatics* **2009**, *11*, 225.
[51] O. Giustolisi, A. Doglioni, D. Savic, D. Laucelli, *OPTIMOGA, Report* **2004**.
[52] O. Giustolisi, A. Doglioni, D. Savic, B. Webb, *Environm. Modell. Softw.* **2007**, *22*, 674.
[53] O. Giustolisi, D. Savic, *A Novel Genetic Programming Strategy: Evolutionary Polynomial Regression*, World Scientific, ISBN 9789812702838, **2004**, pp. 787–794.
[54] M. Rubinstein, *Polymer physics*, Oxford University Press, Oxford, **2003**.
[55] D. De Tommasi, N. Millardi, G. Puglisi, G. Saccomandi, *J. R. Soc., Interface* **2013**, *10*, 20130651.
[56] F. Trentadue, D. De Tommasi, G. Puglisi, *J. Mech. Behav. Biomed. Mater.* **2021**, *115*, 104277.
[57] D. Grubb, G. Ji, *Int. J. Biol. Macromol.* **1999**, *24*, 203.
[58] D. De Tommasi, G. Puglisi, G. Saccomandi, *J. Mech. Phys. Solids* **2015**, *78*, 154.
[59] P.-G. De Gennes, *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca, NY **1979**.
[60] G. Puglisi, G. Saccomandi, *Proc. Royal Society A: Math., Phys. Eng. Sci.* **2016**, *472*, 20160060.
[61] C. Miehe, S. Göktepe, F. Lulei, *J. Mech. Phys. Solids* **2004**, *52*, 2617.
[62] E. Creaco, L. Berardi, S. Sun, O. Giustolisi, D. Savic, *Water Resources Research* **2016**, *52*, 2403.
[63] D. Cook, M. Julias, E. Nauman, *J. Biomech.* **2014**, *47*, 1241.
[64] B. Madsen, Z. Z. Shao, F. Vollrath, *Int. J. Biol. Macromol.* **1999**, *24*, 301.
[65] E. Oroudjev, J. Soares, S. Arcidiacono, J. B. Thompson, S. A. Fossey, H. G. Hansma, *Proc. Natl. Acad. Sci.* **2002**, *99*, 6460.
[66] K. Numata, R. Sato, K. Yazawa, T. Hikima, H. Masunaga, *Polymer* **2015**, *77*, 87.