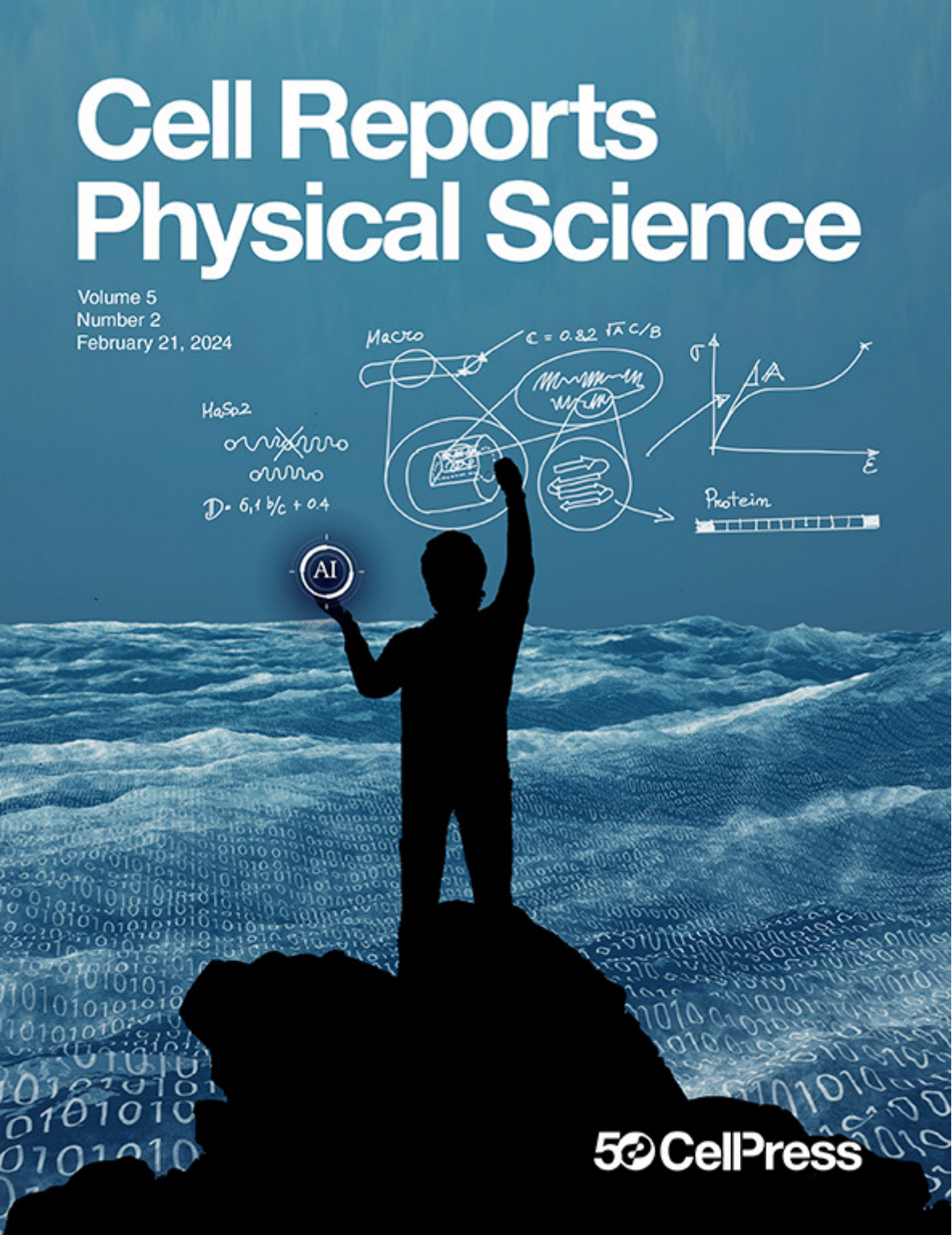
Cell Reports Physical Science

Volume 5
Number 2
February 21, 2024



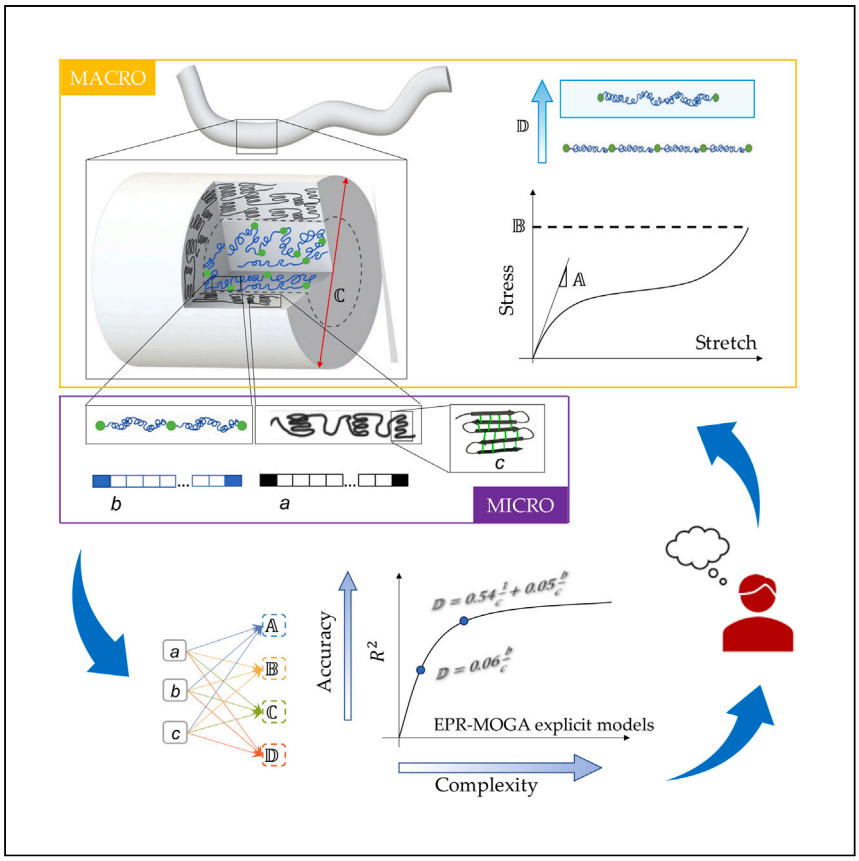
CellPress

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Physically based machine learning for hierarchical materials



Fazio et al. propose a multiscale data-driven strategy for hierarchical physical phenomena. An explicit analytical relationship to be easily interpreted is deduced and increasing complexity is considered only when needed, enabling a continuous interaction between data modeling and scientific knowledge. Spider silk is considered as an explicit proof of concept.

Vincenzo Fazio, Nicola Maria Pugno, Orazio Giustolisi, Giuseppe Puglisi

nicola.pugno@unitn.it (N.M.P.)
giuseppe.puglisi@poliba.it (G.P.)

Highlights

A methodological work for data-driven modeling of hierarchical phenomena

Simple, explicit analytical expressions to be interpreted by the scientist

Continuous interaction between data modeling and scientific interpretation

Spider silk case study from protein to macroscopic behavior

Fazio et al., Cell Reports Physical Science 5, 101790
February 21, 2024
https://doi.org/10.1016/j.xcrp.2024.101790

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Physically based machine learning for hierarchical materials

Vincenzo Fazio,[1] Nicola Maria Pugno,[1,2,*] Orazio Giustolisi,[3] and Giuseppe Puglisi[3,4,*]

SUMMARY

In multiscale phenomena, complex structure-function relationships emerge across different scales, making predictive modeling challenging. The recent scientific literature is exploring the possibility of leveraging machine learning, with a predominant focus on neural networks, excelling in data fitting, but often lacking insight into essential physical information. We propose the adoption of a symbolic data modeling technique, the "Evolutionary Polynomial Regression," which integrates regression capabilities with the genetic programming paradigm, enabling the derivation of explicit analytical formulas, finally delivering a deeper comprehension of the analyzed physical phenomenon. To demonstrate the key advantages of our multiscale numerical approach, we consider the spider silk case. Based on a recent multiscale experimental dataset, we deduce the dependence of the macroscopic behavior from lower-scale parameters, also offering insights for improving a recent theoretical model by some of the authors. Our approach may represent a proof of concept for modeling in fields governed by multiscale, hierarchical differential equations.

INTRODUCTION

Multiscale models play a crucial role in different fields of theoretical and applied science, especially due to the increasing possibility of experimental analyses and technologies working down to the micro and nano scales such as atomic force microscopy, optical tweezers, magnetic tweezers, etc.[1] As a matter of fact, in different fields, a huge experimental literature delivering big data libraries on hierarchical systems, starting from the nano and micro scales up to the macro scale, is now available. These experimental observations represent a potential fundamental new tool for a theoretical advancing in several fields. Such an advance requires the deduction of new numerical/theoretical tools delivering correct physical interpretation of impact in engineering,[2] medicine,[3] physiology,[4] biology,[5] and physics.[6]

Within this context, there is a growing debate concerning the need for effective methodologies capable of facilitating the interaction between theoretical insights and empirical data. In this perspective, machine learning approaches, in a broad sense, appear to be the most promising tools. However, it is important to point out that machine learning per se does not inherently possess the capability to automatically incorporate scientific knowledge, which is crucial for avoiding unphysical results. Indeed, in the digital age, the possibility of new instruments, such as unprecedented power of calculation and machine learning techniques, has opened up exciting possibilities for analyzing the vast amount of experimental data now accessible. However, such analysis can lead to a corresponding increase in the theoretical understanding and modeling of the resulting physical system only if adequate

[1]Laboratory for Bioinspired, Bionic, Nano, Meta Materials and Mechanics, University of Trento, Via Mesiano 77, 38123 Trento, Italy

[2]School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK

[3]Department of Civil Environmental Land Building Engineering and Chemistry, Polytechnic University of Bari, via Orabona 4, 70125 Bari, Italy

[4]Lead contact

*Correspondence: nicola.pugno@unitn.it (N.M.P.), giuseppe.puglisi@poliba.it (G.P.)

https://doi.org/10.1016/j.xcrp.2024.101790

numerical instruments of data modeling are available. On the other hand, as in every transition, the digital transition brings significant risks and drawbacks if not deeply analyzed in its possible effects. Thus, machine learning can lead to scientific knowledge growth or obscuration, rationalization or lack of clearness, access to deeper theoretical models, or reliance on purely data mining approaches.

Among data-driven techniques, many of which have been developed in recent years, the artificial neural network (ANN) is the most adopted to model complex, non-linear processes including multiscale hierarchical phenomena.[7] Loosely speaking, an ANN uses models consisting of multiple processing elements (neurons) connected by links of variable weights (parameters) to deduce typically "black box" representations of the analyzed systems. Learning in ANNs involves adjusting the parameters (weights) of interconnections in a highly parametrized system. In a few words, the main widely recognized disadvantages of ANN model construction are the curse of dimensionality, overfitting issues, and parameter estimation.[8,9] The well-known curse of dimensionality refers to the exponential increase in the need of parameters when the model input space grows. This means that the number of connections exponentially rises, and in a such widened space, the training set of input becomes more sparse, or the amount of data needed to preserve a constant level of accuracy increases exponentially. On the other hand, in such a way, an ANN acquires greater flexibility in mapping events with a complex structure. However, this leads often to overfitting problems, which is that an ANN tends to fit training data too precisely due to the large number of parameters, resulting in the propensity to generate poor predictions for events not close to the training dataset. A further disadvantage of using ANNs is the difficulty of incorporating knowledge derived from known physical laws into the learning process due to the inherent complexity of its framework. Despite these drawbacks, several significant results in this field have been reported. We may recall that good predictive performances were obtained by neural network methods in linking the elastic properties of composite materials to their meso-scale structure, in particular, the three-dimensional microstructure to its effective (homogenized) properties.[10] Recently, Linka et al.[11] adopted an ANN to choose, among an *a priori*-specified class of specific constitutive models depending on the right Cauchy-Green deformation tensor invariants, the model that best reproduces stress-strain behaviors under different classes of deformation. While the approach is interesting, it is highly oriented by the specific knowledge of the problem and restricted to the special case when the class of constitutive laws is already known, i.e., the stress dependence on the deformation invariants. With the aim of reducing the computational burden associated with the numerical solution of describing active force in the cardiac muscle tissue, ANNs were employed to build a reduced order model starting from high-fidelity mathematical models.[12] The implications are thus fundamental and let us obtain relevant information for problems that have long been theoretically unresolved, such as the recalled long-lasting problem of predicting the protein structures from amino acid sequences.[13] On the other hand, the main drawback in the perspective of extending the knowledge for the theoretical modeling of such phenomena is that an ANN leads to "black box" approaches. There is then a strong limitation on the "operational" advantages due to the lack of interpretability of the artificial intelligence results. Some very recent works address this issue,[14,15] but this is still an open problem[16,17] due to the intrinsic nature of the approach, as summarized above.

In this work, we trace a rational way in the direction of deducing different tools for the modeling of multiscale phenomena based on machine learning techniques, with the potential to significantly advance scientific knowledge. Our approach is distinctive in

that it relies on the establishment of fundamental *analytical approximation relations*, crucial to achieve a fully effective and fruitful synergy between data and theoretical modeling. Thus, we recognize that the multiscale character typically corresponds to a hierarchical organization, involving a natural selection of dependent and independent variables. We adopt a genetic algorithm-based approach, which deduces analytical relations through a Pareto front-type optimization, i.e., the evolutionary polynomial regression (EPR) method. A comprehensive description of the history, concepts, and motivation is provided in a following dedicated section.

To analyze the efficiency of the proposed approach in treating complex multiscale hierarchical phenomena, we here consider the field of constitutive modeling of complex material behaviors. Specifically, we focus on the paradigmatic example of spider silk, one of the most studied and complex natural materials due to its extreme mechanical properties, particularly its strength and toughness. We base our analysis on the availability on recent experimental observations on a large number of silks from different spider species from all over the world, where several material properties at different involved scales have been cataloged for the first time in a comprehensive database.[18] In this respect, it is worth noting that previous data modeling results are founded only on statistical properties of the available data (statistical results based on correlation analysis),[18] and this allows for a very partial attainment of the potential impact of such experimental results.

This work aims to be general within the framework of a multiscale description of physical phenomena and the deduction of larger-scale properties from the structures at lower scales. Indeed, in the formulation of the specific case study here analyzed, we have considered three scales starting from the *micro* (protein) scale to the *macro* scale passing through the *meso* scale. We explicitly impose in our approach that these three scales interact with each other in a hierarchical way. In particular, we consider the three possibilities of deduction of the meso from the micro properties, a successive macro from meso, and eventually an interesting direct micro to macro deduction.

Our results show the effectiveness of the proposed method to deduce new physical knowledge on the studied phenomenon. In particular, regarding the considered example of spider silk, among other results, we deduce a functional relation between the thermal degradation temperature and the parameters describing the micro-scale protein structure, a very simple relationship between the diameter of the silk thread and the meso-scale properties, and finally, a straightforward and effective relationship that describes how to deduce the macro-scale supercontraction property as a direct function of micro-scale parameters. Additionally, we identify a meso-scale variable that does not depend on the considered micro variables, suggesting the importance of other micro variables in shaping the meso-scale structure of silk material. Thus, even the hierarchical structure of the involved variables results from the proposed approach, suggesting different micro-meso, meso-macro, and micro-macro relationships depending on the considered variables together with the determination of the effective dependent-independent variables.

We show that the proposed methodology also allows to enhance existing physical approaches by increasing the understanding of the underlying physical processes. In this respect, we have also identified possible directions for further investigating some relationships that have already been theorized. To this end, we interpret the obtained results in relation with a recent physically based model introduced by some of the authors.[19,20]

We argue that this is a first step toward a more effective adoption of the new availability of data and data modeling techniques that can be of fundamental help in several fields of multiscale phenomena compared with the diffuse ANN physically based approaches.

RESULTS AND DISCUSSION

Concepts and motivation of symbolic machine learning using EPR

Digital transition is defined as the review of processes, using products based on digital products, technologies (hardware), and strategies (software), to increase efficiency. *The simpler, more accessible, and representative collection and evaluation of data relating to processes is the knowledge base to provide useful information for efficiency. The process to be made efficient considered here is the scientific knowledge.*

The technological event, then, does not explain and is not alone the digital transition. In fact, at the basis of today's digitalization, there are always humans who developed the theories, paradigms, and concepts that generated the strategies, methods, and algorithms. The latter have not only allowed the development of digital technologies together with supporting the evolution of electronics, but strategies of the digital transition are also the basis to make efficient the processes themselves.

The scientific studies that have generated the possibility of today's *digital transition*, impacting definitively on the development of both *digital strategies*, pertain to the fields of mathematical logic and mathematics. They have developed throughout the last few centuries.

We report some fundamental stages that in the past have given rise to the science and conditions of the digital transition and the specific symbolic machine learning strategy here used, namely EPR.[21]

Alan Turing, already in the 1930s, had introduced the concepts of algorithms and calculating machines that would later lead to the development of computers. In fact, he is considered the father of information technology and of the concept of *machine learning*, which has nowadays entered everyday life with the idea of *artificial intelligence*. It is useful to clarify, without dwelling excessively on the subject, that *machine learning* or *data modeling* or *data driven* are more appropriate terms than the generic one of *artificial intelligence*. The latter is, in our opinion ,an abused "slogan," rooted often in motivations distant from scientific reality, which should not be used for at least two reasons.

- It is misleading with respect to the possibility of digital machines to make all the processes more efficient using an unsupervised strategy, i.e., without the intervention of the human reasoning. This is especially true when digitalization is used to support and make more efficient the progress of scientific knowledge.
- It is not yet known what human intelligence is, either at the mechanistic or biological level of brain functioning; therefore, it is impossible to build machines that simulate something that is unknown in the intrinsic mechanisms that, in addition, generate human consciousness, for example, a concept itself that is difficult to understand.

The idea of *artificial intelligence* was born with McCulloch and Pitts in 1943[22] when they published a work showing a simple system of artificial neurons able to perform

basic logical functions. At least in theory, this system could learn in the same way that humans learn by using experience through trial and error that strengthens or weakens the connections between neurons. *ANNs* are *machine learning* techniques based on this idea from McCulloch and Pitts. They brought to the today well-known *ANNs* that were already programmed in the first personal computers when Rumelhart, Hinton, and Williams developed the error back-propagation method,[23] in 1986, to train them or rather to calibrate the weights of the "synapses" that connect the neurons simulating, in a "very simplified way," the functioning of the human brain. Note that *ANNs* can be seen today as a category of *machine learning* strategies, which are based on the original paradigm with developments of the mathematical structure and learning strategies.

As said, we propose the *symbolic machine learning* strategy, in particular EPR, searching for models using a multi-objective strategy based on evolutionary optimization by genetic algorithms. Thus, as a summary for the reader, we report a brief history of the origin of evolutionary optimization to better understand the motivations of adopting EPR together with a Multi-Objective Genetic Algorithm (EPR-MOGA).[24]

The idea of *evolutionary optimization* was born in the last century and is nowadays a relevant component of process efficiency strategies, which in our context can be identified with the scientific knowledge. In 1973, Ingo Rechenberg was the pioneer of evolutionary calculation and artificial evolution,[25] whose theories were taken up again in 1975 by John Holland who developed the theory of *genetic algorithms* reported in the book *Adaptation in Natural and Artificial Systems*.[26] In 1989, David Goldberg, a student of John Holland and a hydraulic engineer, wrote a book[27] that became the milestone for the use of *genetic algorithms*.

The tools allowing for the optimization with *evolutionary calculation strategies* such as *genetic algorithms* are essential to attain system efficiency. Indeed, they allow cost-benefit (efficiency) problems to be solved by considering more than a single objective, contrary to most of the classical techniques. Moreover, these strategies allow for the adoption of the so-called *Pareto front of optimal or efficient solutions*[28] from the cost-benefit point of view or more solutions with different trade-offs based on the objective functions. These may become a *decision support* for the efficiency of any process, which in the case of this paper regards scientific knowledge.

In this respect, we remark that EPR-MOGA uses a *genetic algorithm* to search for *symbolic models* of data because the strategy is to search for the best trade-off models in complexity vs. data fitting. We point out that *symbolic modeling* of data is a specific strategy internal to the paradigm of *genetic programming*.

In 1992, John Koza developed the paradigm of *genetic programming*, showing[29] the possibility of creating machines that program themselves to solve problems postulated by humans. *Genetic programming* integrates *machine learning*, in a wider sense with respect to the original studies, with *evolutionary optimization* in an original way. Much of what is proposed today as *artificial intelligence* refers to the paradigm of *genetic programming*. *Symbolic modeling* is a specific application of Koza's paradigm to obtain models by means of the integration of *machine learning* and *genetic algorithms* in the form of *symbolic formulas* that can be evaluated as such by the expert. This is a paradigm alternative to that of *ANNs* (see the previous section), which are general mathematical structures characterized by the "universal" ability of interpolating data but, for this reason, they are not suitable

for the interpretation of the results with respect to the physical knowledge of the expert and the required cause-effect relation.

Roughly speaking, the key idea of EPR-MOGA is that the domain of formulas interpolating data is very wide also because of the unavoidable data errors. In other words, we can argue that many formulas might exist of different structures that interpolate the same data with a similar accuracy. However, a clear scientific interpretation is crucially favoured by simple polynomial structures because the parameters can be simply evaluated as being a problem of linear optimization; i.e., a single set of constants exists given a standard error function as opposed to an artificial network whose training depends on the initial "guess" of the weights (parameters).

In summary, EPR is a two-stage strategy: (1) a polynomial structure is selected, and (2) the constants are calculated. Each monomial is composed of one constant and the product of independent variables. If we assign to each of those independent variables an exponent, or they are an argument of logarithm or exponential functions, we obtain a very wide family of possible, non-linear, formula models with the same fundamental characteristics of being linear in parameters. Thus, EPR model's coding is through exponents and functions for independent variables and the maximum number of monomials (parameters). They are prior assumptions of the expert human, who is the only candidate for model building.

The model building is based on the evolution of polynomial structures, which are solutions of a genetic algorithm; each solution is assigned as a set of exponents (where the null exponent reduces the number of independent variables and of monomials) that determines the model structure and parameters.

Thus, as explicitly described in the following, a genetic algorithm determines evolving analytical solutions with the single objective of best fitting to data, possibly with constraints such as the statistical relevance of a monomial. Then the algorithm searches for the optimal values of the constant polynomial parameters, based on a sequence of linear optimizations.

The further development of EPR[24] was to use the MOGA strategy. In this way, the optimization searches for the best trade-off of model complexity versus fitting to data. The complexity is measured with two functions to be minimized, the number of monomials and the number of independent variables used. In this way, a Pareto set of symbolic models is obtained with two competing terms: the model parsimony and the fitting to data.

This is a very effective innovation with respect to standard machine learning, in addition to being useful in scientific knowledge support. In fact, the expert human assumes the exponents, functions, and independent variables, and the strategy returns a front of models, which is a decision support for the expert at the end of the model search. The symbolic structure of the Pareto front of models, together with possible recursive functional terms and independent involved variables, allows for the selection of the adopted analytical model by the expert in a more robust way with respect to a pure statistical assessment.

In a few words, EPR-MOGA returns the model formulas with the best trade-offs of complexity (parsimony) versus fitting to data. The expert chooses the best model looking at the whole Pareto front and its symbolic structure, also considering the increase of complexity versus the effective improvement in terms of fitting to data. In

the following section, we give, for the help of the reader, a brief introduction to the mathematical treatment of numerical optimization problems based on EPR algorithms. We refer to other works[21,24] for a detailed description of the method.

EPR algorithm

In the simple setting considered in this paper, EPR generates explicit mathematical expressions to fit a set of data points starting from the symbolic equation

$$Y = a_0 + \sum_{j=1}^{m} a_j \mathbf{X}_1^{\mathbf{ES}_{j1}} \mathbf{X}_2^{\mathbf{ES}_{j2}} \dots \mathbf{X}_k^{\mathbf{ES}_{jk}}, \qquad \text{(Equation 1)}$$

where Y is the considered output dependent variable, $\mathbf{X}$ is the vector of input variables, and a_o is a bias term. Thus, we assume that Y can be expressed as a polynomial function composed of m terms, here represented by products of powers of the $\mathbf{X}_i$ generated by the algorithm. Other simple functions can be considered instead of powers.[24] Observe that, as previously anticipated, each of the m terms is linearly dependent on the unknown parameters a_j. The power exponents $\mathbf{ES}_{ji}$ are selected from a predetermined set of values.

Synthetically, EPR is performed in two steps: (1) structure identification and (2) parameter estimation. The first stage entails simultaneously determining the best "arrangement" of the independent variables and the related exponents. A multi-objective genetic algorithm termed OPTIMOGA, which stands for OPTImized Multi-Objective Genetic Algorithm, is used to finalize this optimization. This algorithm is based on the MOGA strategy, introduced in the previous section and extensively described elsewhere.[24,30] We refer the readers to those papers for more detailed information.

We remark that, since the user defines *a priori* the set of candidate exponents, the possible negligible input variables are obtained by including zero among them. This represents a fundamental option for the important aspect, recalled in the introduction, of determining the effective independent variables. The values of the parameters a_j are determined in a second stage using the linear least squares (LS) approach, which minimizes the sum of squared errors (SSE). In addition to the usual LS search, the LS is typically performed by searching for only positive values (constraints $a_j > 0$). This choice can be removed by the EPR algorithm, and it can be *a fortiori* justified in our physical model of spider silks by referring to only positive values of the input and output variables. However, this choice helps in avoiding overfitting, by excluding sequences of terms with negative/positive a_j values that may result from the modeling of the data noise.[31]

Moreover, the uncertainty of the coefficients (a_j) is evaluated during the search, and the distribution of estimated pseudo-polynomial coefficients is used to eliminate those parameters whose value is not sufficiently larger than zero.[21,32] Indeed, it may be argued that a low coefficient value with respect to the variance of estimates relates to terms that describe noise rather than the underlying function of the phenomenon being studied.

The algorithm is depicted in the flowchart shown in Figure 1. As a starting point, the candidate independent variables, the general polynomial structure, the functions composing the monomials, the candidate exponents, and the maximum number of terms are assigned, possibly based on the initial knowledge of the physical phenomenon. The exponents can reflect the types of relationships between the inputs and

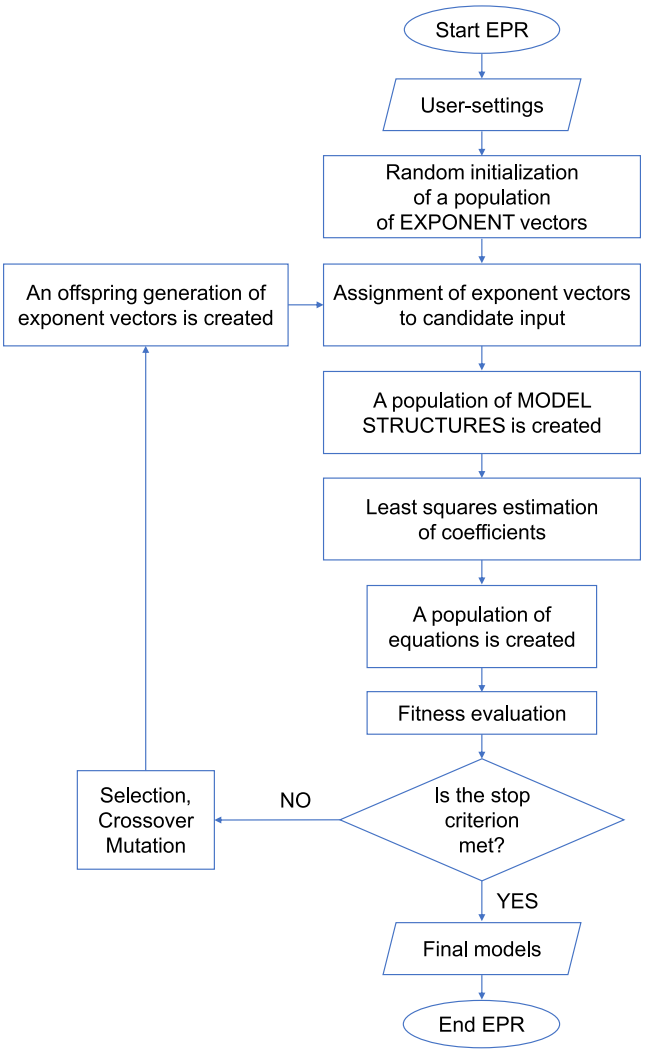


Figure 1. Flowchart of EPR working phases

Step-by-step evolutionary process for constructing the set of final models composing the Pareto dominance front complexity versus fitting.

output. For example, if the vector of candidate exponents is chosen to be $\mathbf{ES} = [-1, -0.5, 0, 0.5, 1]$, if the maximum number of terms is $m = 4$, and if the candidate independent input variables are $k = 3$, the polynomial regression problem is to find a matrix of exponent $\mathbf{ES}_{4\times3}$. In a first stage, an initial population of matrix of exponents is generated. An example of such a matrix is

$$\mathbf{ES}_{4\times3} = \begin{bmatrix} 1 & 0.5 & 0 \\ 0 & 0 & 1 \\ 0 & -0.5 & 1 \\ -1 & 0 & 0.5 \end{bmatrix}, \qquad \text{(Equation 2)}$$

so that the Equation 1 is

$$Y = a_o + a_1 X_1 X_2^{0.5} + a_2 X_3 + a_3 X_2^{-0.5} X_3 + a_4 X_1^{-1} X_3^{0.5} \qquad \text{(Equation 3)}$$

The adjustable parameters a_j are then computed by minimizing the SSE as a cost function. It follows the evaluation of the fitness function: if the termination criterion is satisfied, the output results are shown; otherwise, a new matrix of exponents is

generated through the genetic algorithm including crossover, mutation, and ranking selection.[21] Then, again, the adjustable parameters are calculated, and the fitness function is evaluated until the termination criterion is satisfied.

The equally performing models are those composing the Pareto dominance front,[24,33] and since EPR returns the whole set of formulas of the Pareto front, the final choice of the model among different possible relations can be based on physical considerations.[34] In this respect, we observe that genetic alghoritms generate formulas/models for f, coded in tree structures of variable size, performing a global search of the expression for f as symbolic relationships among $\mathbf{X}$, while the parameters a_j play a role only in the optimization process. On the other hand, the ANN goal is to map f, without focusing on the level of knowledge of the functional relationships among $\mathbf{X}$. This is why we argue that EPR represents a better tool for data-driven knowledge discovery.

Spider silk case study

Spider silk is one of the most studied natural materials due to its extreme mechanical properties, particularly its strength and toughness, which overcome many high-performance man-made materials. Furthermore, spider silks are regarded as a fundamental material for a new class of high-performance fibers in the context of biomimetics.[35,36] The availability of increasingly sophisticated experimental techniques has allowed for a deeper understanding—both chemically and structurally—of the complex multiscale, hierarchical material underlying their notable mechanical behavior. Despite this, many relevant phenomena governing their loading history dependence, rate, temperature, and humidity effects remain unknown.[37]

At the molecular level, spider silks are made up of an amorphous matrix of oligopeptide chains and pseudo-crystalline regions composed primarily of polyalanine β sheets[38,39] with dimensions ranging from 1 to 10 nm,[40] mostly oriented in the fiber direction.[41] The radial cross-section of the fiber is highly organized.[39,42,43] Furthermore, the chemical and structural composition varies according to the different silks produced by the different glands and, of course, the different species. Here, we focus on the most performing and extensively studied type of silk known as *dragline*.

Many biological examples of evolutionary material optimization suggest the possibility of obtaining unreached material performances at the macro scale, based on a clever, hierarchical organization of weak composing materials at the lower scales.[44] A further enrichment in biological structure is to possibly include different composing materials.[45] The analytical description of how the macroscopic performances result from these complex low-scale material organizations is far from being reached and represents a benchmark not only for their theoretical interest, but also in the crucial field of bioinspired material design.[36,46]

Spider dragline silk fibers (also known as major ampullate silk) are constituted by structural proteins called spidroins, which are divided into two major subtypes, MaSp1 and MaSp2. The overall sequence architectures of the two subtypes are similar, with a highly repetitive core region flanked by small N-terminal and C-terminal domains (NTD and CTD, respectively). The repetitive regions, which account for 90% of the primary structure, are composed of alternating runs of polyalanine and multiple glycine-rich motifs arrayed in tandem. Moreover, very recent studies, prompted primarily by advances in proteomics and sequencing technologies, paint a more complex picture of dragline silk composition than a simple MaSp1/MaSp2 dichotomy.[18] Despite this complexity, here, we only consider the proteins MaSp1

and MaSp2, which are widely recognized as the two main ones composing the spider silk. From the secondary structure point of view, the MaSp1 is mainly organized into pseudo-crystalline polyalanine β-pleated sheets.[42,47] On the other hand, the MaSp2 is mainly constituted by proteins with a proline content preventing the formation of β sheet crystals,[39] resulting in a structure with significantly lower crystallinity and macromolecules with weaker crystal domains, typically in the form of α helices and β turns.[39,48]

We remark that, as recognized in polymer mechanics[49] and described also for the spider silk case,[19] the number of monomers of the macromolecule (i.e., protein for the silk case) is fundamental for the mechanical behavior of the material. Based on the facts that (1) the mechanical behavior of the spider silk material is to be ascribed to the repetitive region features more than the terminal region of the protein,[50] and (2) the pseudo-crystalline β sheets, mainly present in the MaSp1, are recognized to be the most impactful feature in determining the exceptional strength of the spider silk,[51] here we consider the following three quantities describing the protein scale of the silk material.

- Length of the repetitive region of the protein MaSp1 in terms of number of amino acids.
- Length of the repetitive region of the protein MaSp2 in terms of number of amino acids.
- Length of the polyalanine β sheet in the protein MaSp1 in terms of number of alanine amino acids.

At the meso scale, we consider the proteins' secondary structure, how macromolecules are arranged in the fiber, and properties regarding the chemical and structural stability of the polymer. In particular, we analyze the following material properties.

- Birefringence. It reflects the degree of molecular orientation of silk protein chains. The birefringence of the dragline silk fiber was calculated from the retardation value and silk fiber diameter.[18]
- Degree of crystallinity. It was calculated based on wide-angle X-ray scattering analysis.[18] In particular, it was obtained as the ratio of the total area of the separated crystalline scattering components to that of the crystalline and amorphous scattering components as resulting from the 1D profile obtained by the 2D diffraction.
- Degradation temperature. This quantity gives a measure of the chemical and structural stability of the silk. In Arakawa et al.,[18] the thermal degradation temperature has been defined as the temperature that yielded 1% weight loss in the silk samples. Indeed, heating leads to changes of the molecular weight that in turn decreases the mass due to the production of gaseous by-products of the chemical reactions.

Spider silk is a very interesting material from the point of view of its mechanical performance at the macroscopic scale. In particular, here, we focus on the material's stiffness and strength. The Young's modulus, on the order of tens of GPa, is above man-made polymers and at the top among other natural materials. The strength is even more interesting, being comparable with high-strength steels (order of magnitude of 1 GPa) and with the best-performing man-made composites like carbon and kevlar reinforced composites.[52] The reason for these outstanding properties with respect to standard materials is not yet clear, with a relevant role also of the extremely small diameter of dragline spider silk.[53] For this reason, we also consider

the diameter in the properties at the macro scale. Finally, we address the very significant role of hydration in the material behavior of spider silks. Indeed, a striking effect observed in spider silks is the so-called *supercontraction*, addressed, to the knowledge of the authors, for the first time in 1977,[54] which occurs when a spider silk thread is exposed to humidity. Depending on the silk composition, the experiments show the existence of a relative humidity (RH) threshold beyond which the fiber contracts up to half of its initial (dry) length. This also results in the possibility of exploiting the supercontraction in the actuation field.[20] The experimentally observed contraction depends on several factors, including spider species,[55] type of silk (among the up to seven different ones that some spiders can produce[56,57]), environmental conditions,[58] and hydration rate.[59] The quantities we consider at the macro scale are therefore the following.

- Young's modulus, obtained from the stress-strain curves determined through tensile tests of single dragline silk fibers conducted at 25°C and RH ≈50%.[18]
- Tensile strength, calculated as the breaking force determined by tensile test divided by the undeformed cross-sectional areas of the fiber samples determined by SEM observations.[18]
- Diameter, determined by SEM observations.[18]
- Maximum supercontraction, calculated as $(L_0 - L_f)/L_0$, where L_0 is the length in dry condition and L_f in fully wet conditions (RH = 100%).[18]

Modeling strategy

For all the EPR run, the maximum number of terms has been set to 3 and the chosen set of candidate exponents has been $[-1, -0.5, 0, 0.5, 1]$ to keep the expression as simple as possible, thus allowing their physical interpretability. Moreover, the expressions were optimized with a bias term a_o since this element may compensate for the possible lack of relevant inputs in the model.

The choice of the maximum number of terms is justified by comparing the expressions provided in Notes S1–S11 and their corresponding performances in Figure 4. Indeed, we achieve nearly maximal performance with just one or two terms, and adding a third term to the expression does not result in a significant improvement in fitting performance. In any case, this represents an optimization parameter that can be easily varied. Moreover, the choice of the candidate exponents can be considered as the simplest choice to consider the important possibility of determining the non-relevance of a candidate input (0), of a linear direct or inverse dependence (1 and −1), and just a simple non-linear direct or inverse dependence (0.5 and −0.5). Other richer choices, depending on the problem under investigation, could be considered. As a matter of fact, the proposed model, differently from the widely used ANN approaches, requires a systematic connection between the scientist and the machine learning results. All these choices are therefore guided by the specific physical problem at hand. Thus, they are part of the modeling and of the scientist's physically guided data preprocessing.

In this regard, we also remark that we considered the possibility of more complex elementary functions and verified the optimality of our choice of power functions. It's worth noting that such comparisons are computationally inexpensive compared to similar possibilities in an ANN, which is another notable advantage of the approach here proposed.

Once again, we refer the readers interested to the numerical performances of EPR to e.g., Giustolisi and Savic[21,24] and references therein. Here, we aim to focus on the

Table 1. Material properties considered for the data modeling case study divided by scales

Micro scale	a	length of the repetitive region of MaSp1
	b	length of the repetitive region of MaSp2
	c	length of the polyalanine β sheet in the MaSp1
Meso scale	A	crystallinity
	B	birefringence
	C	thermal degradation temperature (1% loss)
Macro scale	$\mathbb{A}$	Young's modulus
	$\mathbb{B}$	tensile strength
	$\mathbb{C}$	diameter
	$\mathbb{D}$	supercontraction

applicability of such an already tested numerical efficient approach to hierarchical problems in material science, a field of wide interest in the recent literature on physically based data modeling techniques.

As anticipated above, to exemplify the proposed approach, we focus on the challenging case of spider silks. In particular, we refer to the recently proposed experimental campaign,[18] where the authors analyzed the properties, at different scales, of approximately 1,000 different silks. From our perspective, this represents a significant opportunity for scientists interested in unraveling the "secrets" behind the remarkable mechanical properties of this material in relation to its hierarchical structure.

In the original paper, the authors already proposed a statistical correlation analysis, based on the classical study of the Pearson correlation coefficient. Here, we show how our data modeling approach, combined with our theoretical understanding of the model,[19,20] allows us to gain deeper physical insights in the considered experimental data.

Among the material properties analyzed in the paper, we chose the ones reported in Table 1 with the corresponding adopted symbols (the type of font distinguishes the scales). Notice that each quantity is considered with the unit of measurement reported in the original database, namely GPa for Young's modulus and limit stress, μm for the diameter, °C for the thermal degradation temperature. All the micro-scale properties are expressed in terms of the number of amino acids, whereas the supercontraction and the crystallinity are two non-dimensional quantities ranging in (0,1). As a result the parameters a_j (see Equation 1), estimated by means of the minimization of the SSE, can be dimensional quantities.

The role of these variables in the material hierarchical structure and response of spider silk is schematized in Figure 2.

As a main parameter of accuracy, we report for the different numerical results the coefficient of determination R^2. We recall the classical definition $R^2 = 1 - \sum_{i=1}^{N} \frac{(x_i^{num} - x_i^{exp})^2}{(x_i^{exp} - \overline{x}^{exp})^2}$, where the x_i^{num} are the output variables of the numerical test, and x_i^{exp} are the corresponding experimental values, with $i = 1, \ldots N$, where N is the number of experimental observations considered as dependent variables. Observe that EPR also considers other, not explicitly reported here, indicators of performance, e.g., the SSE. As a result, R^2 does not necessarily increase as the complexity of the expressions grows. The physical valence of the expressions found is discussed, by following Arakawa et al.,[18] also through the comparison with the

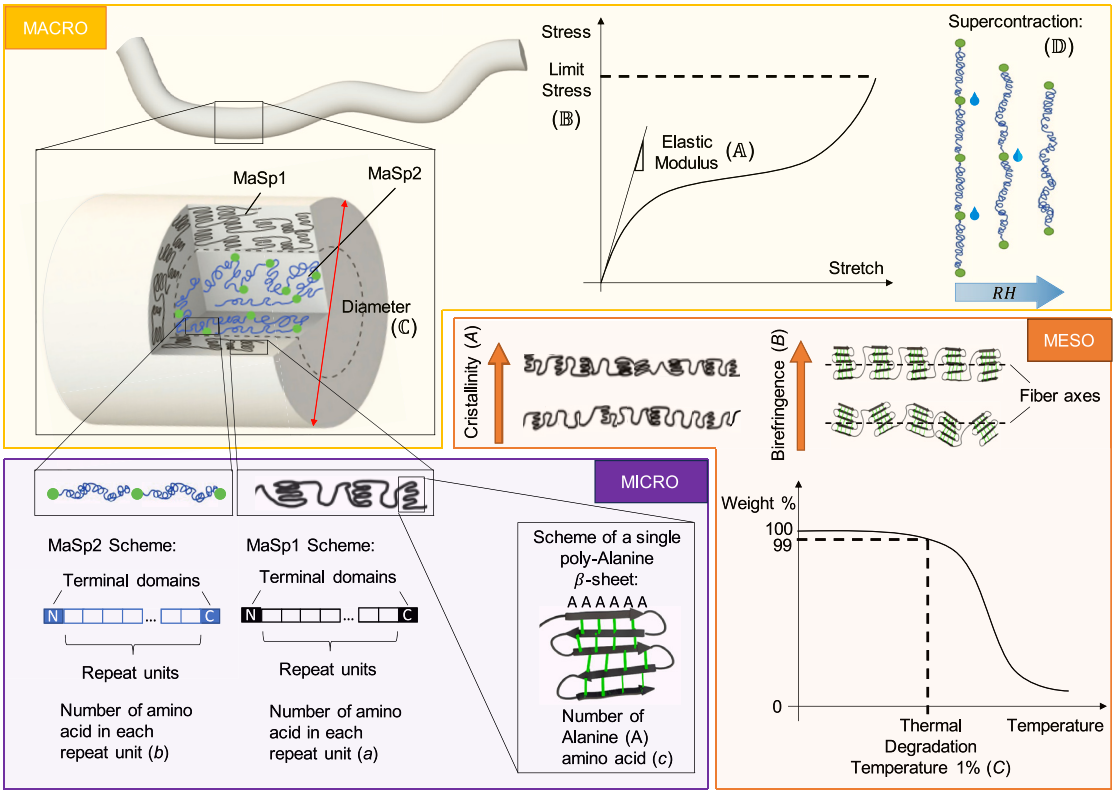


Figure 2. Scheme of the hierarchical structure of spider silks and of the considered variables at the different scales in the data modeling analysis

Macro scale: mechanical properties (elastic modulus and limit stress), morphology of the fiber (diameter), and the macroscopic behavior under humid environment (supercontraction), considered as a key characteristic of the spider silk. Meso scale: proteins' secondary structure (crystallinity), macromolecule alignment within the fiber (birefringence), and chemical/structural stability of the polymer (thermal degradation temperature). Micro scale: primary structure of the proteins, in particular length of the repetitive region of the proteins MaSp1 and MaSp2 in terms of number of amino acids and length of the polyalanine β sheet in the protein MaSp1 in terms of number of alanine amino acids.

correlation matrix represented in Figure 3 obtained by calculating the Pearson correlation coefficient between the different considered variables for the analyzed silks. We recall the classical definition for the Pearson coefficient, a measure of linear correlation between two sets of data $\{x_i, i = 1, \ldots, n\}$ and $\{y_i, i = 1, \ldots, n\}$ with n the number of data, defined as $\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$, where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the mean values.

Observe that, from the database by Arakawa et al.,[18] we considered only the data where the searched output and the considered input are reported simultaneously. Thus, since there are some experimental properties missing for some silks in the database, the number of silks composing the training set is different for each considered output. In doing so, for each target output, we consider the maximum possible available information in terms of the number of silks.

The EPR technique has returned a series of polynomial expressions for each dependent variable. In Figures 4D, 4E, and 4F, 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 formulas considering parsimony (simple expression) and accuracy. Observe that the analysis of the whole expression set,

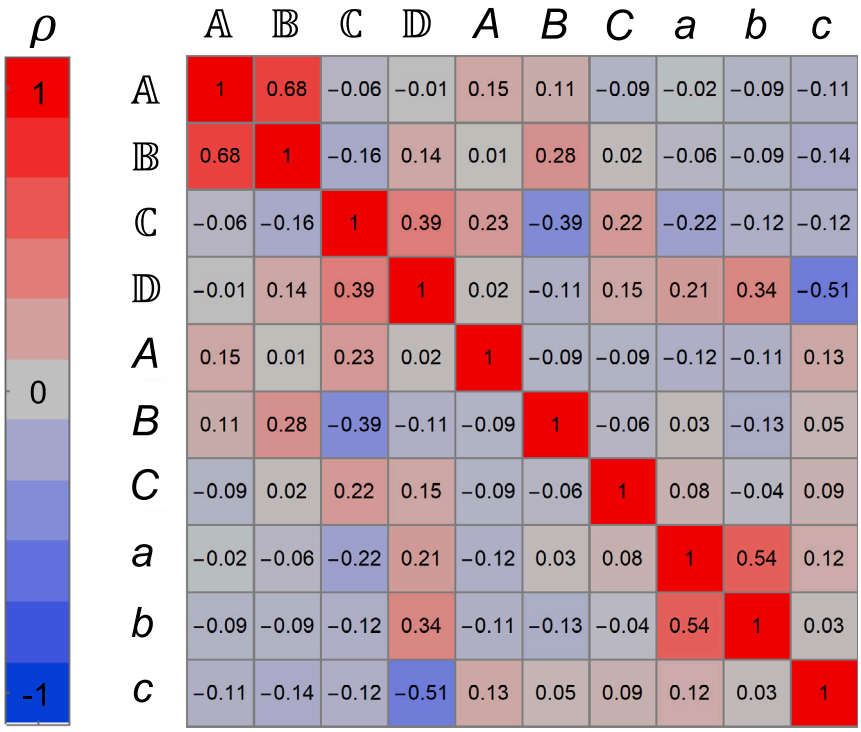


Figure 3. Experimental correlations

Pearson correlation coefficient for each pair of the multiscale properties, considering all the silk reported in the work of Arakawa et al.[18]

reported in the supplemental information, allows for a rational choice of the most suitable material relations (reported in Table 2) that can be selected considering not only parsimony and performance but also analyzing the physical interpretation of the experimental matrix correlations in Figure 3.

Multiscale deduction

In the following, we consider different possible data analyses. Specifically, we examine three potential deduction scenarios: deriving meso from micro properties, then macro from meso properties, and eventually a direct deduction from micro to macro. On one hand, this facilitates the identification of relationships between variables and the potential analytical forms of these relationships. On the other hand, it highlights the significant role of scales in the considered model and the analysis of the best-performing multiscale functional relationships.

In particular, we point out that, based on the possibility of analyzing the functional relations at different scales, with immediate control over functionality and accuracy, we do not assume in advance that a sequential micro-meso-macro variable dependence is the most reasonable as typical of multiscale approaches. We instead suppose that also direct micro-macro variable relations can be observed.

In what follows, we show that this is actually the case, and we find a direct effective relation between a micro and a macro variable. Moreover, we obtain that one of the meso variables does not depend on the considered micro variables, thus suggesting the possibility that other micro variables could be important for the meso-scale structure of the silk material. These results exhibit the efficiency of the model in

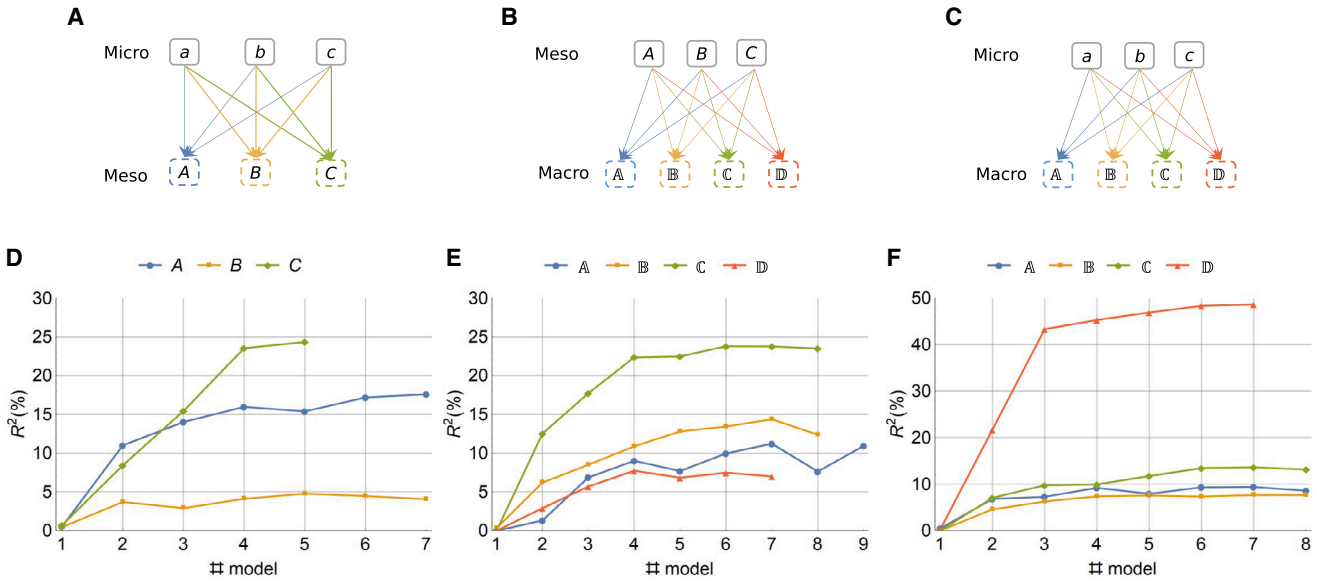


Figure 4. Deduction strategy and performance of the EPR models for each output variable

Prediction of material properties using two scales at a time: (A and D) meso from micro experimental properties, (B and E) macro from meso experimental properties, (C and F) macro from micro experimental properties. (A, B, and C) Scheme of the strategy to obtain each quantity: solid (dashed) box indicates experimental (obtained from EPR) quantities. (D, E, and F) EPR model performance in terms of R^2 plotted against the number of the found model.

selecting the correct functional dependence, the possibility of missing dependence among considered measured variables, and the role of the complex interactions among the different involved scales.

Meso-micro

Firstly, the meso-scale properties have been calculated using all the micro-scale quantities as independent variables (see Figure 4A) according to Equation 1. The results of the accuracy are reported in Figure 4D, and the resulting functional dependencies are reported in Table 2.

As a common property of the considered numerical tests (see in particular the variables A and C in Figure 4D), we may typically distinguish two regimes of the performance curves. In the first regime, the performance increases rapidly with the number of expressions and thus with the model's complexity. In the second regime, the performance curves stabilize in a saturation band. This indicates an easy way of selecting an optimal model complexity and to avoid overfitting due to possible noise of the considered data.

Regarding the selected functional dependence, first, we observe that the crystallinity A decreases with b, in accordance with the general correlation matrix (Figure 3). The presence of the bias term is coherent with the value of $R^2 = 11\%$, since, as recalled before, the bias may compensate for the lack of relevant inputs in the model.

The birefringence B shows a very low accuracy $R^2 < 5\%$, coherent with the experimental results that show a very low Pearson correlation between B and a, b, and c (see Figure 3). We remark that, in this case, the EPR method avoided data overfitting that could have resulted in better performing, but physically misleading, expressions deduced by other numerical approaches. We therefore conclude, in this case, that the considered meso-scale quantity, the birefringence B, cannot be predicted

Table 2. Prediction across two scales: selected explicit expressions

Scale	Expression	R^2(%)	Model number
Meso from micro	$A = 3.5562\frac{1}{b} + 0.10262$	11	2
	$C = 3,186.7046\frac{1}{a} + 0.86787\, a\, c^{0.5} + 45.2672$	23.54	4
Macro from meso	$\mathbb{A} = 0.091301\, \frac{B^{0.5}}{A} + 29.2668\, A$	9.01	4
	$\mathbb{B} = 0.013837\, \frac{B^{0.5}}{A} + 0.014276\, A\, C$	12.81	5
	$\mathbb{C} = 0.81928\, \frac{A^{0.5} C}{B}$	22.37	4
Macro from micro	$\mathbb{D} = 0.61926\frac{b}{c} + 0.0047393$	43.35	3

starting from the considered micro-scale properties, and we consider instead B as an independent variable to compute the macro-scale quantities in the following.

On the other hand, in the case of the thermal degradation temperature C, the EPR found expressions with higher R^2. In this case, the selected expression provides a quantitative estimate of the target quantity with a trend increasing with a and c, in accordance with the experimental correlation matrix in Figure 3.

Macro-meso

As a second data modeling analysis, we consider the possible functional dependence of the macro properties from the meso-scale quantities, now considered as independent variables (see Figure 4B). The results of the accuracy are reported in Figure 4E, and the resulting functional dependencies are reported in Table 2.

Regarding the Young's modulus $\mathbb{A}$, the chosen expression correctly reports the monotonic growth with crystallinity A, as can be immediately deduced by comparing the derivative of the expression for $A > 0$ with the experimental correlation matrix.

Regarding the limit stress $\mathbb{B}$, the selected expression correctly reports the highest experimental correlation, namely the positive one with the birefringence B.

The expression chosen for the diameter $\mathbb{C}$ has the highest accuracy among the macro-meso case ($R^2 = 22.37\%$) with a very simple expression composed of only a single term that includes all three variables at the meso scale. The correlation is positive for A and C and negative for B in accordance with the experiments.

For this last case, in Table 3, we report the complete Pareto front of formulas obtained as the output of the EPR method, along with the corresponding accuracy (R^2) for each expression. It is worth noting that the trend of accuracy concerning the model's complexity is illustrated in Figure 4E. By comparing the set of expressions with the accuracy of the 8 models found by EPR, we observe a rapid increase in accuracy ($R^2 = 0 \rightarrow R^2 \approx 12.5$) when the inverse dependence on variable B is introduced in model 2. A further significant increase in accuracy ($R^2 \approx 12.5 \rightarrow R^2 \approx 18$) is achieved by considering the dependency on $A^{0.5}$ in model 3. The last substantial accuracy improvement ($R^2 \approx 18 \rightarrow R^2 \approx 22.5$) is obtained by including the linear dependence on variable C in model 5. This expression is considered the most suitable for describing the relationship between the diameter and the meso-scale variables. It combines relatively high accuracy with a simple and interpretable structure. The models from 6 to 8, while more complex, do not significantly enhance predictive

Table 3. Pareto front of models for predicting the diameter ($\mathbb{C}$) from the meso-scale variables

Model number	Expression	R^2 (%)
1	$\mathbb{C} = 2$	0
2	$\mathbb{C} = 81.9474\frac{1}{B}$	12.53
3	$\mathbb{C} = 177.4203\frac{A^{0.5}}{B} + 0.051737$	17.73
4	$\mathbb{C} = 0.81928\frac{A^{0.5}C}{B}$	22.37
5	$\mathbb{C} = 0.00021001C + 0.80165\frac{A^{0.5}C}{B}$	22.49
6	$\mathbb{C} = 0.0037544\frac{C}{B^{0.5}} + 0.76892\frac{A^{0.5}C}{B}$	23.79
7	$\mathbb{C} = 0.011782\frac{C}{B} + 0.0025068\frac{C}{B^{0.5}} + 0.76003\frac{A^{0.5}C}{B}$	23.77
8	$\mathbb{C} = 0.028813\frac{C}{B} + 0.69464\frac{A^{0.5}C}{B} + 0.010027\frac{A^{0.5}C}{B^{0.5}}$	23.50

accuracy ($R^2 \approx 22.5 \rightarrow R^2 \approx 23.5$). This implies that the additional terms in these expressions, compared to the previous one, describe noise in the data rather than playing a real physical role in the considered phenomenon.

Finally, we note that the dependence on $1/B$ is maintained from model 2 to model 4, indicating the robustness of this relationship. It is considered reliable as it was preserved even when EPR attempted to reduce expression complexity. On the other hand, the analysis of the complete Pareto front permits the identification of terms that appear in only one model (see models 5 to 8); such terms are likely to be weakly related to the physical phenomenon but rather specific to the error present in the data. Similar considerations apply when examining the complete Pareto fronts for each considered output variable, as reported in Notes S1–S11. The Pareto front has been extensively discussed in this case, which is particularly suitable for explanatory purposes due to the shape of the Pareto front (Figure 4E), combining a rapid increase for the initial models and a clear performance saturation band for the more complex models.

We remark that the final choice of the appropriate equation requires an evaluation of the resulting physical consequences. This may necessitate further experimental and theoretical investigations, as is common in the analysis of any scientific open problem. In our opinion, it is only a continuous efficient interaction between data modeling with analytical formulas and scientific interpretation of them that can ensure the desired advancement of the understanding of the physical phenomena.

Eventually, we consider the selected expression for the supercontraction $\mathbb{D}$. In this case, we are not able to produce a good estimate of the target output from the meso variables ($R^2 < 8\%$).

Macro-micro

As anticipated, we now consider the possibility of direct dependence between macro and micro variables. Thus, the macro properties have been calculated also using all the micro-scale quantities as independent variables (see Figure 4C). The results of the accuracy are reported in Figure 4F, and the resulting functional dependencies are reported in Table 2.

In this case, regarding the Young's modulus ($\mathbb{A}$), the limit stress ($\mathbb{B}$), and the diameter ($\mathbb{C}$), the values of R^2 are generally low. On the other hand, the supercontraction $\mathbb{D}$ is

predicted with a relatively high accuracy ($R^2 > 40\%$), and the selected expression ($R^2 = 43.35\%$) provides a reasonably precise quantitative estimate of the supercontraction, higher than the ones deduced from the meso variables. This suggests the intriguing possibility of a direct influence of micro variables on the macroscopic supercontraction variable representing still a debated effect of spider silk behavior.[19,60] Moreover, we remark on the simplicity of the obtained analytical expression including the two most relevant experimental correlations between the supercontraction and the micro-scale properties, i.e., the positive one with b and the negative one with c. This result demands by itself a theoretical investigation, which will be the focus of our future studies.

In summary, we observe that by employing this direct macro-micro deduction, from one side, we obtain a relatively precise estimation of the supercontraction property that was missing from the macro-meso analysis, but from a modeling point of view, we deduce the possibility of modeling the supercontraction as a macro variable with a direct functional dependence from the micro ones. Moreover, the low accuracy in predicting the other macro variables ($\mathbb{A}$, $\mathbb{B}$, $\mathbb{C}$) directly from the micro ones enlightens on the importance of the meso-scale structures in generally determining the macro properties of the material, as expected from the classical hierarchical dependence. For the particular case of the spider silk, this reflects established results in the literature pointing out the dependence of the silk thread macroscopic behavior from the secondary structures of the proteins,[50,61] here described by the meso-scale variables.

On the accuracy of the EPR formulas for the spider silk case

A general comment is in order on the accuracy of the relationship found by EPR. The coefficient of determination of expressions found by the EPR method is generally low if compared with other frameworks where EPR was applied,[62–64] but this was expected for the study case of spider silks, as in biological materials, a high intrinsic variability for experimental observations is known.[65] Also, for this particular material, a meaningful variability of the mechanical property of silks taken from the same individual under similar conditions is well recognized (see e.g., Madsen et al.[66]). Further, the characteristics of the spider silks have high sensitivity to a large number of parameters, among which are starvation, reeling speed[66] other than the more expected spider species,[55] type of silk (among the up to seven different ones that some spiders can produce[56,57]), environmental conditions,[58] and hydration conditions.[59] In a very recent work,[67] the variability of spider silk properties has been directly compared to that of carbon fibers, and significantly higher variability in spider silk in all properties considered has been reported. For these reasons, even if the R^2 of the expressions found by means EPR is generally not as high as other frameworks, the performances of the data modeling strategies are considered satisfactory. Furthermore, we believe that the results we have described strongly demonstrate the feasibility of our proposed approach when compared to the more commonly used approaches, typically based on ANNs. This approach allows us to deduce both analytical results and important physical properties related to the problem at hand, thereby establishing a new way of investigation in the considered field.

Theoretical vs. experimental correlations

While the objective of this paper is general and mainly related to the exhibited possibilities of obtaining information on the considered physical properties, in this section, we show operatively this possibility by comparing experimental and theoretical results.

Going to the considered case of spider silks, we are now in the position of deducing the theoretical meso and macro response based on only the micro properties to compare with the experimental meso and macro response. Specifically, we have one set of data for which all properties are experimentally known, and on the other hand, we construct a set of theoretical data based solely on the experimentally determined properties at the micro scale. The properties at higher scales (i.e., meso and macro) of the theoretical dataset are then calculated using the explicit relationships selected in Table 2. The purpose is to demonstrate the applicability of EPR-derived relationships to predict macro-scale properties based on micro-scale knowledge. We have chosen to utilize pairwise correlations between variables as a means of comparison between experimental data and the relationships learned through EPR.

Coherently with the hierarchical assumption of our model, we first deduce the meso-scale variables by the micro ones, and then, based on previous analytical results, we deduce the macro variables. Accordingly, with previously described numerical data analysis, also the meso variable B (birefringence) is considered here as an independent variable. On the other hand, in the special case of the supercontraction $\mathbb{D}$, we assume that it directly depends on micro variables. Notice that all this relevant physical information has been deduced by previous data modeling.

Regarding the experimental data, we consider a subset of the silks analyzed by Arakawa et al.[18] and in particular only those for which all the 10 considered properties (see Table 1) are known simultaneously (the so-obtained subset consists of 35 silks). As a possible comparison between the theoretical and experimental datasets, we consider the Pearson correlation coefficients for each pair of properties. The comparisons of the correlation tables for theoretical and experimental results are reported in Figure 5. They show a satisfactory correspondence almost extensible to all the data and a satisfying result in terms of the values of the correlation coefficients. To get a global comparison, we also adopt a positive definite relative error:

$$e_r = \frac{|e_a|}{e_m}, \qquad \text{(Equation 4)}$$

where $e_a = \rho_t - \rho_e$ is the absolute error, and ρ_t and ρ_e are the theoretical and experimental Pearson coefficients, respectively. Here e_m is the mean error that since ρ_m and ρ_e range in the interval $(-1,1)$, we assume $e_m = 1$. The average value of the relative error by considering all the possible pairs of the correlation matrix $\overline{e_r} = 0.33$, with $0 < \overline{e_r} < 2$, indicates that the functional dependence found by the EPR method reproduces in a reasonably accurate way the experimental correlations among the properties of spider silks.

Eventually, as evidenced by Linka et al.,[11] an important extension of the proposed approach would be to consider a Bayesian framework for the uncertainty quantification in order to compute each output in terms of statistical distribution with a mean and a confidence interval by also taking into account the input data variability.

Toward integrating data-driven knowledge and physical modeling

The possibility of advancing, based on the proposed approach, the understanding of the underlying physical relationships can be exhibited by considering possible progresses in existing theoretical settings. To this end, we here explicitly refer to the recent works previously proposed by some of the authors,[19,20] where the dependence of supercontraction on the MaSp2 protein has been addressed. In that

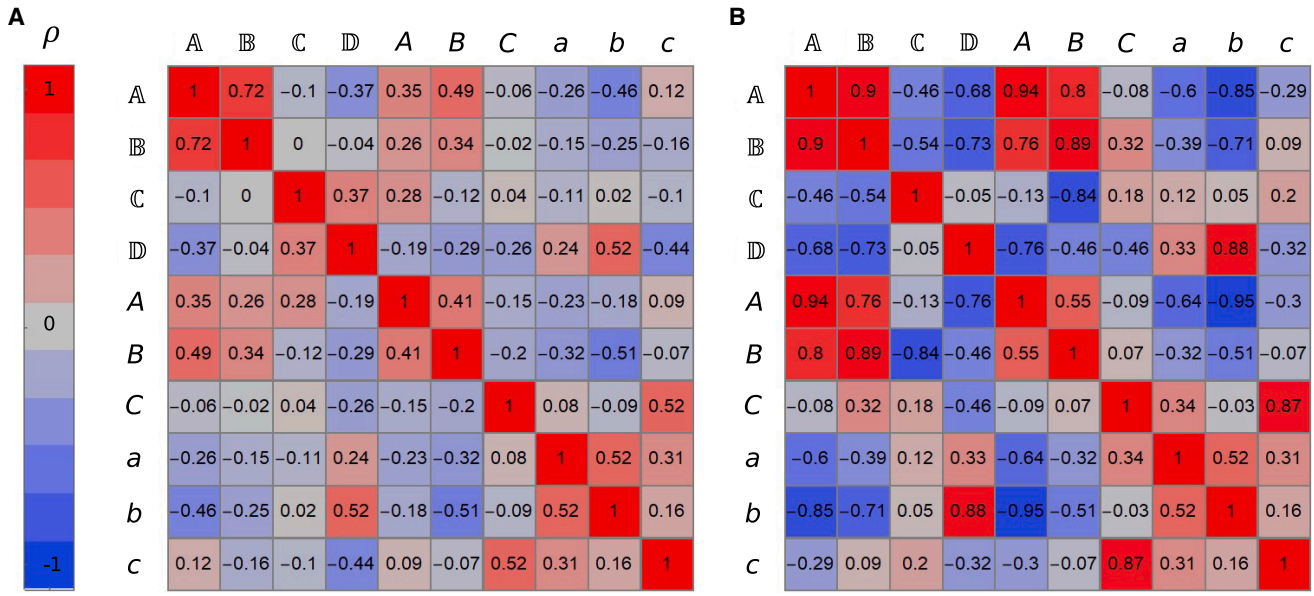


Figure 5. Pearson correlations among the material properties at the three scales

(A) Experimental correlations obtained considering a subset of silks for which all the analyzed properties are reported simultaneously.

(B) Correlations among the material properties obtained from the data modeling EPR approach (macro and meso) starting from the known micro experimental properties.

context, the MaSp2 protein was treated using the classical approach of multiscale analysis of soft macromolecular materials, based on a classical statistical mechanics approach to quantify the expected length of the protein's macromolecules. On the other hand, based on the EPR method, an explicit relationship between the length of the repetitive unit of the MaSp2 protein and supercontraction can be inferred. As an extension of this work, this relationship will be employed to enhance the micro-structure-based model considered in the previous works.[19,20] Presumably, from the EPR findings, it will be possible to establish a quantitative relationship between micro-scale variables describing the primary structure of the MaSp2 protein and supercontraction. In other words, owing to the interpretable relationships obtained through EPR, it is feasible to extend the prediction of macroscopic supercontraction behavior toward the precise primary structure of the involved proteins. Indeed, in the previously proposed theoretical multiscale approach, this prediction was based on more general properties of macromolecule behavior without specifying the detailed primary structure properties obtained in previous analysis. This example is just one illustration of our approach, but it serves as a representative instance of how our work intends to improve the theoretical understanding in material science. With each of the relationships considered in Table 2, and more broadly, as we explore the Pareto front of expressions, additional relationships discovered among variables are subjects of ongoing study by the authors. These investigations aim to contribute significantly to the expansion of our knowledge concerning the multiscale mechanisms responsible for the remarkable characteristics of spider silk. This extension is the subject of the forthcoming research of the authors.

Concluding remarks

We showed the possibility of adopting, based on a genetic programming approach, data modeling techniques, innovative in the field of material science, that are particularly suitable for the deduction of analytical models for multiscale problems. Our approach is based on the EPR method, which, as we showed, lets us deduce models

that are both accurate and simple and able to describe the dependence of macroscale variables from the ones at lower scales, with their hierarchical order itself deduced by a careful analysis of the data. The best-performing models are those located on the Pareto dominance front, which takes into account both accuracy and parsimony and are returned by the EPR algorithm. The final choice of the model can then be based on physical considerations.

To explicitly show the possibility of acquiring physical insight in a complex multiscale problem, and to evidence the key advantages of our multiscale approach compared to classical, non-physically based techniques, we referred to the materials science field and in particular to the complex case of spider silk: a biological material with exceptional properties hugely analyzed also in the spirit of bioinspiration. The choice of this specific case is due to the observation that such remarkable properties are strictly based on an evolutionary hierarchical optimization, and the macroscopic spider silk behavior is the result of noteworthy mesostructures emerging from the aggregation of amino acids at the molecular scale. For this intriguing and very complex material, many phenomena underlying the multiscale structure and the complex energetic exchanges among the scales ensuring their notable properties are still strongly unclear. We then used this paradigmatic example to show how the presented data modeling approach can be useful in several directions: to determine dependent and independent variables, to indicate their hierarchical organization, and to deduce explicit relations among different groups of variables. In this direction, we also want to remark that a possible important role can be attained by a following dimensional analysis (e.g., Buckingham theorem) that should be related to the possible absence of variables at the different scales. This is another important aspect that is beyond the scope of this paper and will be the subject of our future investigation.

Furthermore, we showed that the proposed approach lets us overcome the overfitting problem typically observed in the analysis of big data within the ANN framework diffusely adopted in this field. Based on this, new physical knowledge is acquired, which can be used as a starting point for determining new analytical models, suggesting new experiments, and applying more focused data modeling analysis. We also strive to enrich existing physical approaches by enhancing our comprehension of the underlying physical processes. In this context, we investigate the potential for enhancing some authors' previously introduced physically based theoretical analyses, leveraging the insights obtained from our current approach.

In this sense, we assert that machine learning or artificial intelligence can have a significant impact on scientific knowledge only if the data modeling approaches are in continuous synergy with the scientific interpretation of the results. We argue thus that a new mixed genetic programming-theoretical approach can be a fruitful approach in material science but also in fields as diverse as biology and medicine.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this paper is Giuseppe Puglisi (giuseppe.puglisi@poliba.it).

Materials availability

No materials were used in this work.

Data and code availability

All the data generated by analyses during this study are included in this published article. The software that supported this research was EPR, and it is available from the author O.G. (correspondence: orazio.giustolisi@poliba.it) with free-of-charge licensing.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2024.101790.

ACKNOWLEDGMENTS

Funding is as follows: V.F. and G.P. have been supported by GNFM and INdAM), G.P. has been 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 is funded by the European Union - Next Generation EU. G.P. has been supported by the Research Project Prin2022 of National Relevance 2022XLBLRX and Prin 2022 PNRR P2022KHFNB granted by the Italian MUR. N.M.P. has been supported by the European Commission under the FET Open "Boheme" grant no. 863179 and by the Italian Ministry of University MUR under PRIN-20177TTP3S.

AUTHOR CONTRIBUTIONS

V.F., numerical analysis, methodology, and writing; O.G., methodology and writing; N.M.P., methodology and supervision; G.P., methodology, writing, and supervision.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: July 18, 2023
Revised: November 7, 2023
Accepted: January 5, 2024
Published: January 30, 2024

REFERENCES

1. Bustamante, C., Macosko, J.C., and Wuite, G.J. (2000). Grabbing the cat by the tail: manipulating molecules one by one. Nat. Rev. Mol. Cell Biol. *1*, 130–136.

2. Chen, Q., and Pugno, N.M. (2013). Bio-mimetic mechanisms of natural hierarchical materials: A review. J. Mech. Behav. Biomed. Mater. *19*, 3–33.

3. Ashley, E.A. (2016). Towards precision medicine. Nat. Rev. Genet. *17*, 507–522.

4. Zhang, P., Su, J., and Mende, U. (2012). Cross talk between cardiac myocytes and fibroblasts: from multiscale investigative approaches to mechanisms and functional consequences. Am. J. Physiol. Heart Circ. Physiol. *303*, H1385–H1396.

5. McLennan, R., Dyson, L., Prather, K.W., Morrison, J.A., Baker, R.E., Maini, P.K., and Kulesa, P.M. (2012). Multiscale mechanisms of cell migration during development: theory and experiment. Development *139*, 2935–2944.

6. Ji, H., Daughton, W., Jara-Almonte, J., Le, A., Stanier, A., and Yoo, J. (2022). Magnetic reconnection in the era of exascale computing and multiscale experiments. Nat. Rev. Phys. *4*, 263–282.

7. Alber, M., Buganza Tepole, A., Cannon, W.R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W.W., Perdikaris, P., Petzold, L., and Kuhl, E. (2019). Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. NPJ Digit. Med. *2*, 115.

8. Haykin, S.S. (1999). Neural Networks (A Comprehensive Foundation (Prentice Hall)).

9. Giustolisi, O., and Laucelli, D. (2005). Improving generalization of artificial neural networks in rainfall-runoff modelling. Hydrol. Sci. J. *50*, 439–457.

10. Cecen, A., Dai, H., Yabansu, Y.C., Kalidindi, S.R., and Song, L. (2018). Material structure-property linkages using three-dimensional convolutional neural networks. Acta Mater. *146*, 76–84.

11. Linka, K., and Kuhl, E. (2023). A new family of Constitutive Artificial Neural Networks towards automated model discovery. Comput. Methods Appl. Mech. Eng. *403*, 115731.

12. Regazzoni, F., Dedè, L., and Quarteroni, A. (2020). Machine learning of multiscale active force generation models for the efficient simulation of cardiac electromechanics. Comput. Methods Appl. Mech. Eng. *370*, 113268.

13. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang,

J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876.

14. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. *116*, 22071–22080.

15. Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. Commun. ACM *63*, 68–77.

16. Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. ArXiv. Prepr. ArXiv170208608.

17. Molnar, C. (2020). Interpretable Machine Learning (Lulu. com).

18. Arakawa, K., Kono, N., Malay, A.D., Tateishi, A., Ifuku, N., Masunaga, H., Sato, R., Tsuchiya, K., Ohtoshi, R., Pedrazzoli, D., et al. (2022). 1000 spider silkomes: Linking sequences to silk physical properties. Sci. Adv. *8*, eabo6043.

19. Fazio, V., De Tommasi, D., Pugno, N.M., and Puglisi, G. (2022). Spider silks mechanics: Predicting humidity and temperature effects. J. Mech. Phys. Solids *164*, 104857.

20. Fazio, V., Pugno, N.M., and Puglisi, G. (2023). "Water to the ropes": A predictive model for the supercontraction stress of spider silks. Extreme Mech. Lett. *61*, 102010.

21. Giustolisi, O., and Savic, D.A. (2006). A symbolic data-driven technique based on evolutionary polynomial regression. J. Hydroinformatics *8*, 207–222.

22. McCulloch, W.S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. *5*, 115–133.

23. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536.

24. Giustolisi, O., and Savic, D.A. (2009). Advances in data-driven analyses and modelling using EPR-MOGA. J. Hydroinformatics *11*, 225–236.

25. Rechenberg, I. (1971). Evolutionsstrategie. Optim. Tech. Syst. Nach Prinz. Biol. Evol. *(PhD Thesis)* (Reprinted by Fromman-Holzboog (1973)).

26. Holland, J.H. (1992). Adaptation in Natural and Artificial Systems (MIT press).

27. Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization, Machine Learning (Addison-Wesley).

28. Pareto, V. (1906). Manual of Political Economy (Oxford University Press).

29. Koza, J.R. (1992). Genetic Programming, on the Programming of Computers by Means of Natural Selection. A Bradford Book (MIT Press).

30. Giustolisi, O., Doglioni, A., Savic, D., and Laucelli, D. (2004). A proposal for an effective multi-objective non-dominated genetic algorithm: The optimised multi-objective genetic algorithm-OPTIMOGA. OPTIMOGA Rep *7*, 1–37.

31. Giustolisi, O., Doglioni, A., Savic, D.A., and Webb, B. (2007). A multi-model approach to analysis of environmental phenomena. Environ. Model. Softw. *22*, 674–682.

32. Giustolisi, O., and Savic, D. (2004). A novel genetic programming strategy: evolutionary polynomial regression. In Hydroinformatics (World Scientific), pp. 787–794.

33. Pareto, V. (1896). Cours d'Economie Politique (F. Rouge).

34. Giustolisi, O. (2004). Using genetic programming to determine Chezy resistance coefficient in corrugated channels. J. Hydroinformatics *6*, 157–173.

35. Greco, G., Arndt, T., Schmuck, B., Francis, J., Bäcklund, F.G., Shilkova, O., Barth, A., Gonska, N., Seisenbaeva, G., Kessler, V., et al. (2021). Tyrosine residues mediate supercontraction in biomimetic spider silk. Commun. Mater. *2*, 43.

36. Arndt, T., Greco, G., Schmuck, B., Bunz, J., Shilkova, O., Francis, J., Pugno, N.M., Jaudzems, K., Barth, A., Johansson, J., and Rising, A. (2022). Engineered Spider Silk Proteins for Biomimetic Spinning of Fibers with Toughness Equal to Dragline Silks (Adv. Funct. Mater. 23/2022. Adv. Funct. Mater. *32*, 2270134.

37. Pérez-Rigueiro, J., Elices, M., Plaza, G.R., and Guinea, G.V. (2021). Basic Principles in the Design of Spider Silk Fibers. Molecules *26*, 1794.

38. Elices, M., Plaza, G.R., Pérez-Rigueiro, J., and Guinea, G.V. (2011). The hidden link between supercontraction and mechanical behavior of spider silks. J. Mech. Behav. Biomed. Mater. *4*, 658–669.

39. Sponner, A., Vater, W., Monajembashi, S., Unger, E., Grosse, F., and Weisshart, K. (2007). Composition and Hierarchical Organisation of a Spider Silk. PLoS One *2*, e998.

40. Keten, S., and Buehler, M.J. (2010). Nanostructure and molecular mechanics of spider dragline silk protein assemblies. J. R. Soc. Interface *7*, 1709–1721.

41. Jenkins, J.E., Sampath, S., Butler, E., Kim, J., Henning, R.W., Holland, G.P., and Yarger, J.L. (2013). Characterizing the secondary protein structure of black widow dragline silk using solid-state NMR and X-ray diffraction. Biomacromolecules *14*, 3472–3483.

42. Li, S.F., McGhie, A.J., and Tang, S.L. (1994). New internal structure of spider dragline silk revealed by atomic force microscopy. Biophys. J. *66*, 1209–1212.

43. Eisoldt, L., Smith, A., and Scheibel, T. (2011). Decoding the secrets of spider silk. Mater. Today *14*, 80–86.

44. Giesa, T., Arslan, M., Pugno, N.M., and Buehler, M.J. (2011). Nanoconfinement of Spider Silk Fibrils Begets Superior Strength, Extensibility, and Toughness. Nano Lett. *11*, 5038–5046.

45. Bosia, F., Abdalrahman, T., and Pugno, N.M. (2012). Investigating the role of hierarchy on the strength of composite materials: evidence of a crucial synergy between hierarchy and material mixing. Nanoscale *4*, 1200–1207.

46. Liu, Y., Luo, D., and Wang, T. (2016). Hierarchical structures of bone and bioinspired bone tissue engineering. Small *12*, 4611–4632.

47. Brown, C.P., MacLeod, J., Amenitsch, H., Cacho-Nerin, F., Gill, H.S., Price, A.J., Traversa, E., Licoccia, S., and Rosei, F. (2011). The critical role of water in spider silk and its consequence for protein mechanics. Nanoscale *3*, 3805–3811.

48. Nova, A., Keten, S., Pugno, N.M., Redaelli, A., and Buehler, M.J. (2010). Molecular and Nanostructural Mechanisms of Deformation, Strength and Toughness of Spider Silk Fibrils. Nano Lett. *10*, 2626–2634.

49. Flory, P.J., and Erman, B. (1982). Theory of elasticity of polymer networks. 3. Macromolecules *15*, 800–806.

50. Hayashi, C.Y., Shipley, N.H., and Lewis, R.V. (1999). Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins. Int. J. Biol. Macromol. *24*, 271–275.

51. Yarger, J.L., Cherry, B.R., and van der Vaart, A. (2018). Uncovering the structure–function relationship in spider silk. Nat. Rev. Mater. *3*, 18008.

52. Ashby, M.F., and Johnson, K. (2013). Materials and Design: The Art and Science of Material Selection in Product Design (Butterworth-Heinemann).

53. Porter, D., Guan, J., and Vollrath, F. (2013). Spider Silk: Super Material or Thin Fibre? Adv. Mater. *25*, 1275–1279.

54. Work, R.W. (1977). Dimensions, Birefringences, and Force-Elongation Behavior of Major and Minor Ampullate Silk Fibers from Orb-Web-Spinning Spiders—The Effects of Wetting on these Properties. Text. Res. J. *47*, 650–662.

55. Boutry, C., and Blackledge, T.A. (2010). Evolution of supercontraction in spider silk: structure-function relationship from tarantulas to orb-weavers. J. Exp. Biol. *213*, 3505–3514.

56. Vollrath, F. (1992). Spider webs and silks. Sci. Am. *266*, 70–76.

57. Gosline, J., Pollak, C., Guerette, P., Cheng, A., DeMont, M., and Denny, M. (1994). Elastomeric Network Models for the Frame and Viscid Silks from the Orb Web of the Spider Araneus Diadematus.

58. Plaza, G.R., Guinea, G.V., Pérez-Rigueiro, J., and Elices, M. (2006). Thermo-hygro-mechanical behavior of spider dragline silk: Glassy and rubbery states. J. Polym. Sci. B Polym. Phys. *44*, 994–999.

59. Agnarsson, I., Boutry, C., Wong, S.-C., Baji, A., Dhinojwala, A., Sensenig, A.T., and Blackledge, T.A. (2009). Supercontraction forces in spider dragline silk depend on hydration rate. Zoology *112*, 325–331.

60. Cohen, N., Levin, M., and Eisenbach, C.D. (2021). On the Origin of Supercontraction in Spider Silk. Biomacromolecules *22*, 993–1000.

61. Yazawa, K., Malay, A.D., Masunaga, H., Norma-Rashid, Y., and Numata, K. (2020). Simultaneous effect of strain rate and humidity on the structure and mechanical behavior of spider silk. Commun. Mater. *1*, 10.

62. Berardi, L., Kapelan, Z., Giustolisi, O., and Savic, D.A. (2008). Development of pipe deterioration models for water distribution systems using EPR. J. Hydroinformatics *10*, 265–126.

63. Creaco, E., Berardi, L., Sun, S., Giustolisi, O., and Savic, D. (2016). Selection of relevant input variables in storm water quality modeling by multiobjective evolutionary polynomial regression paradigm. Water Resour. Res. *52*, 2403–2419.

64. Balf, M.R., Noori, R., Berndtsson, R., Ghaemi, A., and Ghiasi, B. (2018). Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers. J. Water Supply Res. Technol. *67*, 447–457.

65. Cook, D., Julias, M., and Nauman, E. (2014). Biological variability in biomechanical engineering research: Significance and meta-analysis of current modeling practices. J. Biomech. *47*, 1241–1250.

66. Madsen, B., Shao, Z.Z., and Vollrath, F. (1999). Variability in the mechanical properties of spider silks on three levels: interspecific, intraspecific and intraindividual. Int. J. Biol. Macromol. *24*, 301–306.

67. Greco, G., Mirbaha, H., Schmuck, B., Rising, A., and Pugno, N.M. (2022). Artificial and natural silk materials have high mechanical property variability regardless of sample size. Sci. Rep. *12*, 3507.

Cell Reports Physical Science, Volume *5*

Supplemental information

Physically based machine learning for hierarchical materials

Vincenzo Fazio, Nicola Maria Pugno, Orazio Giustolisi, and Giuseppe Puglisi

SUPPLEMENTAL INFORMATION

EPR Expressions

Note S1. Cristallinity from micro properties

$$A = 0.19253 \tag{S1.1}$$

$$A = 3.5562\frac{1}{b} + 0.10262 \tag{S1.2}$$

$$A = 3.9403\frac{1}{b} + 0.0097339c + 0.013178 \tag{S1.3}$$

$$A = 0.0010868b + 2.0441\frac{c^{0.5}}{b} \tag{S1.4}$$

$$A = 0.007611\frac{b}{a^{0.5}} + 1.961\frac{c^{0.5}}{b} \tag{S1.5}$$

$$A = 0.0077062\frac{b}{a^{0.5}} + 2.4415\frac{1}{b} + 0.38037\frac{c}{b} \tag{S1.6}$$

$$A = 0.0079125\frac{b}{a^{0.5}} + 10.8129\frac{1}{cb} + 0.50353\frac{c}{b} \tag{S1.7}$$

Note S2. Birefringence from micro properties

$$B = 45.0698 \tag{S2.1}$$

$$B = 88.7638\frac{1}{b^{0.5}} + 31.2338 \tag{S2.2}$$

$$B = 188.8929\frac{1}{a} + 54.9397\frac{1}{b^{0.5}} + 31.4271 \tag{S2.3}$$

$$B = 571.6574\frac{1}{a} + 1.7581\frac{a}{b^{0.5}} + 19.1995 \tag{S2.4}$$

$$B = 120.3924\frac{b^{0.5}}{a} + 18.2244\frac{a}{b} + 7.3676 \tag{S2.5}$$

$$B = 135.0812\frac{b^{0.5}}{a} + 19.3865\frac{a}{b} + 0.036473a + 2.329 \tag{S2.6}$$

$$B = 141.865\frac{b^{0.5}}{a} + 17.6973\frac{a}{b} + 0.69564\frac{a}{b^{0.5}} \tag{S2.7}$$

Note S3. Thermal degradation temperature from micro properties

$$C = 226.8169 \tag{S3.1}$$
$$C = 3.1984c + 201.1538 \tag{S3.2}$$
$$C = 785.1137\frac{1}{c} + 13.9926c + 13.4973 \tag{S3.3}$$
$$C = 3186.7046\frac{1}{a} + 0.86787ac^{0.5} + 45.2672 \tag{S3.4}$$
$$C = 2482.7659\frac{1}{a} + 383.6809\frac{1}{c} + 0.27232ac + 25.2964 \tag{S3.5}$$

Note S4. Young's Modulus from meso properties

$$\mathbb{A} = 9.4674 \tag{S4.1}$$
$$\mathbb{A} = 0.86198B^{0.5} + 3.7949 \tag{S4.2}$$
$$\mathbb{A} = 0.60871\frac{1}{A} + 29.105A \tag{S4.3}$$
$$\mathbb{A} = 0.091301\frac{B^{0.5}}{A} + 29.2668A \tag{S4.4}$$
$$\mathbb{A} = 1.3608\frac{B^{0.5}}{AC^{0.5}} + 29.2732A \tag{S4.5}$$
$$\mathbb{A} = 20.1089\frac{B^{0.5}}{AC} + 1.9687AC^{0.5} \tag{S4.6}$$
$$\mathbb{A} = 2.3655\frac{B}{AC} + 0.098189\frac{1}{A} + 2.0067AC^{0.5} \tag{S4.7}$$
$$\mathbb{A} = 27.51\frac{1}{AC} + 2.2479\frac{B}{AC} + 2.0041AC^{0.5} \tag{S4.8}$$
$$\mathbb{A} = 10.3522\frac{B^{0.5}}{AC} + 1.3585\frac{B}{AC} + 1.9953AC^{0.5} \tag{S4.9}$$

Note S5. Limit Stress from meso properties

$$\mathbb{B} = 1 \tag{S5.1}$$

$$\mathbb{B} = 0.14344B^{0.5} + 0.2349 \tag{S5.2}$$

$$\mathbb{B} = 0.09374\frac{1}{A} + 3.1305A \tag{S5.3}$$

$$\mathbb{B} = 0.013904\frac{B^{0.5}}{A} + 3.1765A \tag{S5.4}$$

$$\mathbb{B} = 0.013837\frac{B^{0.5}}{A} + 0.014276AC \tag{S5.5}$$

$$\mathbb{B} = 0.011517\frac{B^{0.5}}{A} + 0.0017796ACB^{0.5} + 0.20612 \tag{S5.6}$$

$$\mathbb{B} = 0.021735\frac{1}{A} + 0.0090613\frac{B^{0.5}}{A} + 0.001894ACB^{0.5} + 0.14047 \tag{S5.7}$$

$$\mathbb{B} = 0.012666\frac{B^{0.5}}{A} + 0.0048689AC + 0.001261ACB^{0.5} + 0.09517 \tag{S5.8}$$

Note S6. Diameter from meso properties

$$\mathbb{C} = 2 \tag{S6.1}$$

$$\mathbb{C} = 81.9474\frac{1}{B} \tag{S6.2}$$

$$\mathbb{C} = 177.4203\frac{A^{0.5}}{B} + 0.051737 \tag{S6.3}$$

$$\mathbb{C} = 0.81928\frac{A^{0.5}C}{B} \tag{S6.4}$$

$$\mathbb{C} = 0.00021001C + 0.80165\frac{A^{0.5}C}{B} \tag{S6.5}$$

$$\mathbb{C} = 0.0037544\frac{C}{B^{0.5}} + 0.76892\frac{A^{0.5}C}{B} \tag{S6.6}$$

$$\mathbb{C} = 0.011782\frac{C}{B} + 0.0025068\frac{C}{B^{0.5}} + 0.76003\frac{A^{0.5}C}{B} \tag{S6.7}$$

$$\mathbb{C} = 0.028813\frac{C}{B} + 0.69464\frac{A^{0.5}C}{B} + 0.010027\frac{A^{0.5}C}{B^{0.5}} \tag{S6.8}$$

Note S7. Supercontraction from meso properties

$$\mathbb{D} = 0.31695 \tag{S7.1}$$

$$\mathbb{D} = 1.4429\frac{1}{B^{0.5}} + 0.093513 \tag{S7.2}$$

$$\mathbb{D} = 11.8488\frac{A^{0.5}}{B} + 0.18064 \tag{S7.3}$$

$$\mathbb{D} = 0.00085048C + 10.8458\frac{A^{0.5}}{B} \tag{S7.4}$$

$$\mathbb{D} = 0.00086571C + 158.9386\frac{A^{0.5}}{C^{0.5}B} \tag{S7.5}$$

$$\mathbb{D} = 0.00050622C + 3.8693e-05CB^{0.5} + 12.6449\frac{A^{0.5}}{B} \tag{S7.6}$$

$$\mathbb{D} = 0.00067818C + 2.1548e-05CB^{0.5} + 172.7934\frac{A^{0.5}}{C^{0.5}B} \tag{S7.7}$$

Note S8. Young's Modulus from micro properties

$$\mathbb{A} = 8.5718 \tag{S8.1}$$

$$\mathbb{A} = +150.4503\frac{1}{b} + 4.8595 \tag{S8.2}$$

$$\mathbb{A} = +103.8759\frac{1}{a} + 93.8966\frac{1}{b} + 3.4673 \tag{S8.3}$$

$$\mathbb{A} = +18.4325\frac{c}{a} + 36.55\frac{1}{c} \tag{S8.4}$$

$$\mathbb{A} = +111.3405\frac{c}{ab^{0.5}} + 37.8293\frac{1}{c} \tag{S8.5}$$

$$\mathbb{A} = +712.237\frac{c}{ab} + 5.6688\frac{b^{0.5}}{c} \tag{S8.6}$$

$$\mathbb{A} = +120.4316\frac{c}{ab^{0.5}} + 21.7225\frac{1}{c} + 2.2107\frac{a^{0.5}}{c} \tag{S8.7}$$

$$\mathbb{A} = +121.6927\frac{c}{ab^{0.5}} + 0.40972\frac{b}{c} + 114.29\frac{a^{0.5}}{cb} \tag{S8.8}$$

Note S9. Limit Stress from micro properties

$$\mathbb{B} = 1.2533 \tag{S9.1}$$

$$\mathbb{B} = +17.8802\frac{1}{a} + 0.77347 \tag{S9.2}$$

$$\mathbb{B} = +5.1279\frac{1}{a^{0.5}} + 3.14\frac{1}{c} + 0.025086 \tag{S9.3}$$

$$\mathbb{B} = +146.7895\frac{1}{a^{0.5}b} + 0.014999b + 0 \tag{S9.4}$$

$$\mathbb{B} = +0.26238\frac{b}{a^{0.5}c^{0.5}} + 24.7876\frac{1}{b} + 0 \tag{S9.5}$$

$$\mathbb{B} = +0.29999\frac{b}{a} + 0.056394\frac{b}{c} + 24.7595\frac{1}{b} + 0 \tag{S9.6}$$

$$\mathbb{B} = +132.185\frac{1}{a^{0.5}b} + 0.052774\frac{b}{a^{0.5}} + 0.056459\frac{b}{c} + 0.037522 \tag{S9.7}$$

$$\mathbb{B} = +0.44217\frac{b}{a^{0.5}c} + 49.6034\frac{c^{0.5}}{a^{0.5}b} + 0.054095\frac{b}{c} \tag{S9.8}$$

Note S10. Diameter from micro properties

$$\mathbb{C} = 1.5415 \tag{S10.1}$$

$$\mathbb{C} = +44.5973\frac{1}{b} + 0.43669 \tag{S10.2}$$

$$\mathbb{C} = +1030.4476\frac{1}{ab} + 0.82562 \tag{S10.3}$$

$$\mathbb{C} = +1367.643\frac{1}{ab} + 0.013763b \tag{S10.4}$$

$$\mathbb{C} = +8702.7135\frac{1}{acb} + 0.01363b + 0.17322 \tag{S10.5}$$

$$\mathbb{C} = +9507.752\frac{1}{acb} + 0.0019587cb \tag{S10.6}$$

$$\mathbb{C} = +9491.6515\frac{1}{acb} + 0.001894cb + 7.7482e-05ac \tag{S10.7}$$

$$\mathbb{C} = +9609.6043\frac{1}{acb} + 0.0014633cb + 7.4705e-05a^{0.5}cb \tag{S10.8}$$

Note S11. Supercontraction from micro properties

$$\mathbb{D} = +0.32479 \tag{S11.1}$$

$$\mathbb{D} = +0.0055013b + 0.097111 \tag{S11.2}$$

$$\mathbb{D} = +0.061926\frac{b}{c} + 0.0047393 \tag{S11.3}$$

$$\mathbb{D} = +0.53648\frac{1}{c} + 0.050578\frac{b}{c} \tag{S11.4}$$

$$\mathbb{D} = +0.8009\frac{1}{c} + 0.0072816\frac{a^{0.5}b}{c} \tag{S11.5}$$

$$\mathbb{D} = +15.9524\frac{1}{cb} + 0.008755\frac{a^{0.5}b}{c} \tag{S11.6}$$

$$\mathbb{D} = +430.4092\frac{1}{acb} + 0.0090635\frac{a^{0.5}b}{c} \tag{S11.7}$$