Probabilistic learning and updating of a digital twin for composite material systems

Roger Ghanem, Christian Soize, Loujaine Mehrez, Venkat Aitharaju

Probabilistic Learning and Updating of a Digital Twin for Composite Material Systems

Roger Ghanem[*,a], Christian Soize[b], Loujaine Mehrez[a], Venkat Aitharaju[c]

[a] *Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA*
[b] *Université Gustave Eiffel, Laboratoire Modélisation et Simulation Multi Echelle, MSME UMR 8208 CNRS, 5 bd Descartes, 77454, Marne-la-Vallée, France*
[c] *General Motors Company, GM R&D Technical Center, 30500 Mound Rd, Warren, MI 48092, USA*

Abstract

This paper presents an approach for characterizing and estimating statistical dependence between a large number of observables in a composite material system. Conditional regression is carried out using the estimated joint density function, permitting a systematic exploration of interdependence between fine scale and coarse observables that can be used for both prognosis and design of complex material systems. An example demonstrates the integration of experimental data with a computational database. The statistical approach is based on the probabilistic learning on manifolds recently developed by the authors. This approach leverages intrinsic structure detected through diffusion on graphs with projected stochastic differential equations to generate samples constrained to that structure.

1. Introduction

In all its variations, the digital twin aims to represent, digitally, enough features of a physical device or process, sufficiently well, so as maximize some measure of performance [1, 2, 3]. Key components in conceptualizing the digital twin consist of **1)** a measure of performance, **2)** the choice of representative features, and **3)** the fidelity in depicting these features. Often, the measure of performance is tied to a design objective such as regulatory compliance, cost, or manufacturing tolerance, while the choice of representative features is related to operational convenience including sensing and control, as well as expertise in synthesizing and interpreting reduced-order models. The issue of fidelity of representation is somewhat more nuanced as it pertains to providing a context for assessing discrepancies between observed and anticipated realities. This discrepancy can be due, among others, to lack of sufficient data, prediction horizon, modeling errors, as well as the choice of obervables.

An important challenge with the digital twin concept is the update of the mathematical and digital representations to account for an evolving knowledge base that describes either the evolution of the system itself (for instance through aging, deterioration or bifurcation) or the evolution of information concerning the system (for instance through acquired data or improved insight into behavior). A rational formulation of this update indubitably requires a comprehensive probabilistic treatment. A second significant challenge with the digital twin concept, is the close

[*]Corresponding author: R. Ghanem, ghanem@usc.edu

Preprint of the published paper in International Journal for Numerical Methods in Engineering, 2020.

interplay between events occurring and observed at different scales, using different measuring devices. For instance, while performance measures often pertain to collective behavior observed on a coarse scale, failure often nucleates by individuals interacting on a much finer scale.

The digital twin concept is unique among computational science perspectives in that it is predicated on a seamless integration of data acquisition, data interpretation, and decision, and thus faces, simultaneously, all of the above challenges. In a few words, a coherent digital twin concept must tackle, simultaneously, issues of statistical inference, updating, and experiment design in a multiscale and multi-physics setting, with evolving instabilities.

A number of recent attempts have been made to apply machine learning (ML) ideas to material characterization, including the effect of manufacturing process. In [4], a Gaussian mixture model with expectation maximization condition is used to relate photodiode measurements to performance of laser powder-bed fusion additively manufactured components. In [5] a laser-based direct metal solid free-form fabrication (an additive manufacturing process) proposed the construction of a "process map" that relates key quantities of interest to build parameters. These approaches require significant amounts of data to statistically characterize the relationships between the various input and output quantities. A support vector machine (SVM) approach for classify built components, using such a process map was proposed in [6]. The SVM is merely capable of classification between "good" and "bad" component, and is not relevant for condition assessment using cross-scale observations. In [7], a design of experiment procedure is presented, using a random forest appraoch, to enable the identification of the most informative experiments. In [8], a number of graph-based ML procedures are reviewed as reduced-order models for upscaling brittle crack propagation in geomaterials. While these surrogate models can be constructed using a small number of samples, they only provide deterministic predictions, which are clearly not suitable in the presence of modeling, experimental or statistical errors. A Bayesian network is constructed in [9] as surrogate to the small fatigue crack driving force in polycrystalline materials. In [10], a large number of features are extracted from a few number of high-fidelity numerical simulation of interacting cracks, following which a number of ML procedures such as Random Forest and ANN are constructed as reduced order models.

This paper describes an approach to mitigate two significant issues relevant to the above challenges. The first issue is that of model error, inevitably introduced with physics-based models. The impact of these errors is more pronounced when interpreting information pertaining to bifurcations, instabilities, or failure. The second issue is that of computational burden, also usually introduced with physics-based models, and the complexity of which grows with the level of required fidelity and data complexity. It is often true that the computational burden reflects the complexity of the underlying model and not necessarily the complexity of the decision on which the digital twin is being exercised. We mitigate the first of these difficulties by relying on a data-driven model that discovers intrinsic structure within the data, using a diffusion manifold reduction. We mitigate the second difficulty by relying on a joint statistical description of all observables that is facilitated by a sampling scheme that leverages the manifold structure, permitting us to generate a large database of consistent samples for use in non-parametric estimation schemes, as proposed by the procedure of the probabilistic learning on manifolds that was already developed and documented by the authors [11, 12, 13, 14, 15]. This paper demonstrates the application of this procedure to the tasks of prognosis and design in a problem of composites performance. Given a small training set obtained through laboratory experiments and numerical simulations, probabilistic statements are formulated and computed for various quantities of interest conditional on observations associated with other quantities of interest. We deduce insight and relevant relationships between mechanical and geometric properties at different scales and

systematically explore the complexity of their interactions within a composite material system. The novelty of the paper is in framing the problem of cross-scale inference, critical for a digital twin context, in a format that lends itself to analysis using well-adapted methods of machine learning. While both the composites model and the machine learning algorithms have already been published, as described above, the ability to statistically relate fine scale constituents to meso-scale and coarse-scale constituents in a composite has not been attempted before using these methods.

2. Digital twin characterization through probabilistic machine learning

A digital twin has two sides, one pertaining to a physical device and the other to a digital rendition of this device. More accurately, both sides of the twin reference an equivalence class (ie a version) of the physical device as specified by a set of measurements. The digital rendition is learned by fusing information from experimental and computational data streams.

The task of constructing a digital twin requires the ability to describe the relationship between properties of a given system in a manner that permits inference on a subset of them having specified another subset. Collectively, these properties are referred to as the system's observables and could consist of physics properties such as parameters of a mechanistic model or of state variables such as strain, temperature or velocity over a spatio-temporal domain. Observables also include operating conditions such as boundary and initial conditions. The purpose of the digital twin then is to predict the value of a subset of the observables as function of another subset, and to continually update this predictive model over the life of the system as it goes through transformations which may be gradual such as degradation or abrupt such as renovations and accidents. The deployment of the digital twin can be viewed as consisting of two phases. In a first phase, the digital twin is developed. This entails the adoption of mathematical models of physical phenomena, their numerical resolution, and their statistical against experimental evidence. This may involve the assimilation of newly acquired information in order to maintain synchronization between the digital and physical twins. A most informative and operationally useful approach for this phase would be to cast the problem in a probabilistic framework where data is used to update a prior probabilistic digital twin. Thus, this task of constructing the digital twin essentially consists of training a predictive model using a combination of experimental and computational data. In a second phase, the digital twin is exercised in a decision-making context. Accordingly performance metrics are evaluated for specific operational conditions.

Based on the above, we group the information relevant to the training and decision phases into three subsets. Let Q and W denote, respectively, the subsets of observables being inferred and specified, and U denote the subset of the remaining observables. Clearly, the choice of observables is paramount to the inferential value of the predictive model. However, and while in a training setting these observables are merely constrained by computational resources, in an operational setting they are constrained by deployment requirements including sensing and hardware. Thus it is important for the observables to be measurable on both sides of the digital twin if the two sides are to remain synchronized. A closer inspection of this matter indicates that while Q and W are both essential for synchronization, observations of U can be limited to the computational model. Indeed, the role of the "hidden" observables U is to enhance the validity and accuracy of the model that relates Q and W. Their absence during the decision phase introduces uncertainty to all inferences on which decisions are predicated.

3. Physical setup and computational models

Before we delve into the description of our probabilistic learning paradigm, we introduce in this section the physical device associated with our digital twin.

3.1. Multiscale interactions in composite systems

We consider the performance of an eight laminae composite laminate and its dependence on mechanical properties of fibers and resin and on geometric properties of tows. The performance of the composite is described in terms of its full stress-strain curve as measured at the scale of the laminate. The laminate consists of eight laminae made of continuous non-crimp fabric (NCF) with resin filling the spaces between the fibers; i.e., the continuous tows are placed along one direction in each lamina. The laminae are oriented with $[0°/45°/-45°/90°/90°/-45°/45°/0°]$ alignment to satisfy quasi-isotropic symmetry in the laminate design. The carbon fiber tows are made of $12,000$ fibers (T700SC 12000 50C). Given the difficulty in measuring in-situ the mechanical and geometric properties of micro-constituents, we rely on a computational model to synthesize observables that are constrained by some underlying physics. The complexity of this physics, spanning four length scales (laminate, lamina, tow and fiber), requires recourse to multi-scale modeling and computational paradigms. The hierarchy of scales for NCF composite materials is next introduced. This is followed by the introduction of relevant observables at the different scales. A summary is then presented of the various numerical simulators used to approximate these observables as constrained by the physics. A polynomial-chaos based approach to this same problem is described elsewhere [16]. The objective of that work was to develop efficient multi-scale stochastic representations. The present work presents a distinctly different perspective that extracts, from the training set, an intrinsic reduced-order model which is subsequently used in statistical inference.

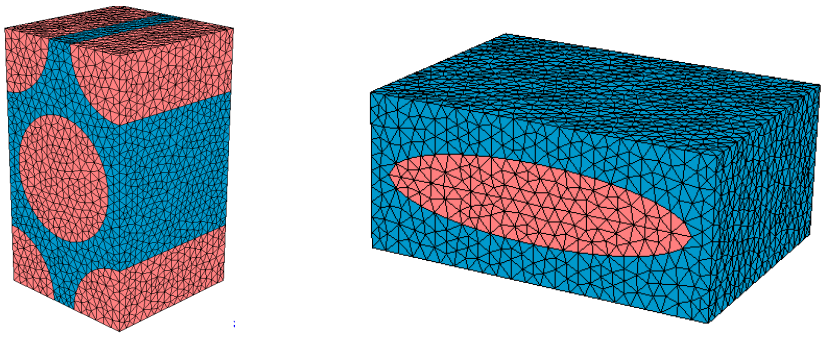

Figure 1: A volumetric unit cell within a tow (left) and within a NCF lamina (right)[17].

A lamina consists of an arrangement of ellipsoidal tows within a resin matrix, with each of the tows consisting of a bundle of circular fibers, themselves infused into a resin matrix. Nonlinear interactions can be significant between features and behaviors and different scales. An upscaling methodology that takes into consideration these nonlinearities is used herein [18, 16]. We specifically use the methodology as implemented in the software environment Multiscale Designer [17]. The inputs to the finest scale simulation consist of the mechanical properties of fibers and resin, $\boldsymbol{P}_f^0$ and $\boldsymbol{P}_m^0$, respectively. These input parameters to the micro-scale simulator are grouped as $\boldsymbol{P}^0 = (V_f^{F,T}, E_{f,A}, E_{f,T}, G_{f,A}, \nu_{f,A}, \nu_{f,T}, E_m, \nu_m)$ where the volume fraction of fibers

within the tow is denoted by $V_f^{F,t}$ and the resin properties are indexed by m. The subscripts A and T on these parameters refer to axial and transverse properties, respectively. The subscript m refers to the material properties of the isotropic matrix (resin). These properties are used to evaluate the upscaled effective properties of the tows which we denote by $\boldsymbol{Q}^1 = (E_{t,A}, E_{t,T}, G_{t,A}, \nu_{t,A}, \nu_{t,T})$ where the subscripts A and T refer to axial and transverse properties, respectively, while the subscript t refers to the tow. The volumetric unit cell from the Multiscale Designer software [17] is shown in Figure 1 (left).

The second scale simulation, where lamina properties are evaluated, takes as input effective mechanical properties of the tows as well as geometry parameters that characterize the major and minor axes of the tow's and the edge-to-edge spacing between these ellipsoids. The geometry pa-

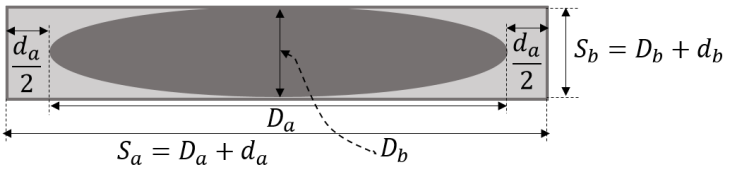

Figure 2: A schematic drawing of a unit cell in a NCF unidirectional lamina.

rameters define the dimensions of the unit cell of a NCF unidirectional lamina as illustrated with the drawing in Figure 2, where D_a and D_b refer to the diameters of the tow along the major and minor directions, respectively, d_a and d_b refer to the gap between the tows along the major and minor directions, which are filled in by the matrix. Observations from micrographs show that d_b is very small, thus, for the computational finite element discretization, this parameter is assumed to be constant and equal to 0.06 *mm*. The dimensions of the rectangular unit cell are denoted by $S_a = D_a + d_a$ and $S_b = D_b + d_b$. Observations of the tow cross-sections from micrographs indicate that the unidirectional tows could be approximated to have ellipsoidal shapes. A detailed validation of this approximation is outside the scope of the work. Therefore, a volumetric unit cell consists of a unidirectional tow with an elliptic cross-section surrounded by resin/matrix. Figure 1 (right) shows this unit cell as used in Multiscale Designer [17]. The input parameters $\boldsymbol{P}^1$ consist of the geometry parameters describing the unit cell in a NCF lamina, $\boldsymbol{P}_g^1$, and the material properties of the isotropic matrix $\boldsymbol{P}_m^1$. Thus, $\boldsymbol{P}^1 = (D_a, d_a, D_b, E_m, \nu_m)$. The subsequent upscaling process aims at computing the homogenized constitutive properties of a NCF unit cell in a lamina.

3.2. Eigendeformation-based reduced-order homogenization

Multiscale Designer software [17] implements a hierarchical reduced-order homogenization methodology for composites. This methodology has been developed as a model reduction to direct multiscale homogenization. The key to the computational reduction is based on (i) modeling the nonlinear behavior of the fine-scale unit cells across the scales in terms of eigenstrain (representing the phase damage evolution) and eigenseparation (representing the interface damage evolution between the phases)and (ii) assuming that the eigenstrain constitute together with the elastic strain the total strain in an additive form. The computational reduction is achieved by formulating the problem in terms of fine-scale residual-free fields (displacement, strain, and stress fields) such that the fine-scale fields satisfy the fine-scale equilibrium equations prior to solving the nonlinear problem. This is accomplished by introducing discretized forms of the eigenstrains

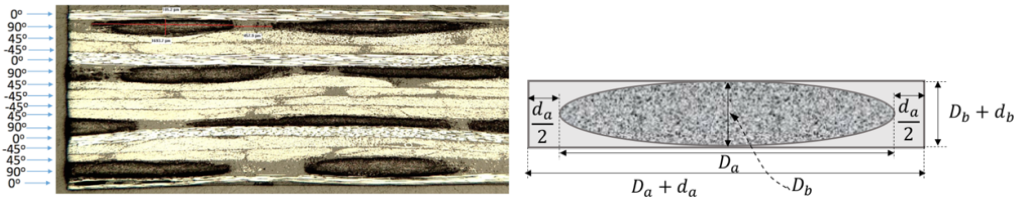

Figure 3: Micrograph of tow and laminae (left) and schematic and geometry of an elliptical approximation of tow and surrounding volume (right).

and eigenseparations in terms of additive shape functions and their respective coefficients. This formulation enables the prior computation of the so-called transformation influence functions for the resulting elastic boundary value problem of the unit-cell. It follows that the equilibrium computations in direct homogenization techniques are no longer required, which results in a considerable reduction in computational time.

It is worth noting that the equilibrium equation for the fine scale is defined by the corresponding leading order terms in the multiscale equilibrium equation. The leading order is defined by the order of ξ; defined as $x = y/\xi$. The scale-dependent equilibrium equation is constructed by approximating the displacement, strain, and stress fields in terms of the multiple-scale asymptotic expansion.

The pre-processed transformation influence functions are associated with the coarse-scale strain, the fine-scale eigenstrain, and the fine-scale eigenseparation. Subsequently, the nonlinear problem is solved using a reduced-order system of equations that are constructed for the reduced order boundary value problem. Within the context of finite element problem, the coarse-scale strain that is estimated at every load increment and every iteration is first fed into these reduced order system of equation to compute the corresponding increment in phase strain and phase separation for the fine-scale problem. The constitutive equation of the phase is then used to compute the increment in eigenstrain. As a result, the estimated eigenstrain and eigenseparation can be used to compute the coarse-scale stress, which can be fed back to the coarse-scale solver.

In order to account for uncertainties in configuration and material properties, the variables shown in Table (1) are modeled as random variables. All these variables, except for the geometry parameters, are assumed to be statistically independent with a symmetric beta distribution having a shape parameter equal to 1.5 and a range identified from manufacturer's specifications and expert opinion. Experimental data on tow geometry as identified by D_a, D_b and d_a in the right side of Figure (3), is used to define a joint probability density function (jpdf) for D_a and D_b, with the PDF of d_a independently estimated from its own data. The joint PDF of D_a and D_b is further constrained by the empirically measured volume fraction at the lamina scale. The resulting support of this jpdf is shown as the blue-shaded area in Figure (4) with the solid lines delineating the volume fraction constraint and the experimental data shown as individual points. The presence of experimental data points outside the support of the probability density function (PDF) can be explained by the non-ellipsoidal and highly irregular geometry of the "real" tows as seen in the micrograph in Figure (3).

For the purpose of the probabilistic machine learning used in the present paper, only 7 of the 18 random variables were treated as observables, namely $(E_{ft}, \rho_{fac}, \sigma_{yfa}, \rho_{ft}, D_a, D_b, d_a)$. This subset is identified in accordance with a sensitivity analysis from a previous study [16]. It should be noted that limiting the observables to this subset does not imply treating the remaining

Physical Component	Variable	Definition	Range & COV
Fiber	E_{fa}	fiber axial modulus (GPa)	[172-206],4.5
	E_{ft}	fiber transverse modulus (GPa)	[12.5,16.5],6.9
	G_{fa}	fiber shear modulus (GPa)	[7.3,9.7],6.4
	ν_{fa}	fiber axial Poisson ratio	[0.29,0.38],6.72
	ν_{ft}	fiber transverse Poisson ratio	[0.17,0.23],7.5
	ρ_{fac}	(compression modulus)/(axial modulus)	[0.72,0.88],0.05,
	σ_{yfa}	fiber axial yield strength (GPa)	[2.6,4],10.6
	$\rho_{ft} = \sigma_{yt}/\sigma_{ya}$	(transverse strength)/(axial strength)	[0.02,0.06],0.025
	$\rho_{ftC} = \sigma_{yac}/\sigma_{ya}$	(compressive strength)/(axial strength)	[0.55,0.65],0.0417
Resin	ν_m	resin Poisson ratio	[0.32,0.4],0.0556
	σ_{ym}	mean stress at damage initiation (MPa)	[27,29],1.97
	K_{0m}	yield strength (MPa)	[31.6,33.6],1.53
	K_{1m}	ultimate strength (MPa)	[52,54],0.94
	H_d	linear term for hardening law	[0.0033,0.0035],0.0094
	δ	exponent for evolution law	[40,44],1.47
Tow	D_a	major axis of tow's ellipsoid	Fig(3)
	D_b	minor axis of tow's ellipsoid	
	d_a	spacing between tows	empirical distribution

Table 1: Input parameters for numerical simulations of microscale simulations; last column shows range of parameter and coefficient of variation (COV); PDF for geometry data is constructed directly from experimental data.

variables as deterministic, but rather as unspecified and thus increasing the uncertainty in ensuing probabilistic inferences and widening the associated probability density functions.

In addition to the model input variables, model output is measured in the form of stress-strain curves through the inelastic regime. The model is exercised in two different loading modalities, namely a tension test and a three-point bending test. Further, 14 tensile and 11 bending experiments are performed in the laboratory and their stress-strain histories recorded. No additional information concerning microstructure constituents are available for these tests. Figure (5) shows the stress strain curves from these experiments together with numerical simulation results obtained through a sequence of upscaling procedures as described in the next section.

It is clear that the bending experiments 7,8,9,10,11 and the tensile experiments 1,4,5,13 are on the fringes of the numerical predictions.

4. Probabilistic Learning on Manifold

In this section, we provide a summary of the probabilistic learning on manifold procedure as it pertains to the analysis and interpretation of the experimental and computational data described previously. Detailed exposition of the procedures described here can be found elsewhere [11, 12, 13, 14]

4.1. Principal component analysis

As a first step in data analysis, the training dataset, defined by matrix $[x]_d$ is construed as a realization from a random matrix variate $[\boldsymbol{X}]$, and is reduced via a principal component analy-

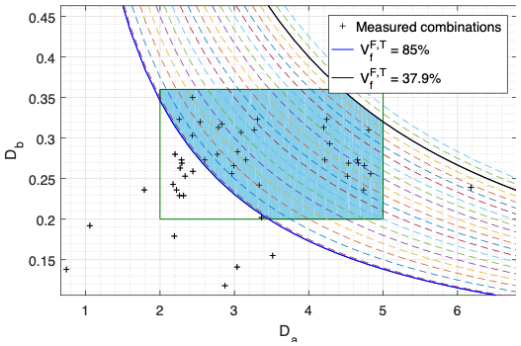

Figure 4: Support of joint probability density function of tow diameters.

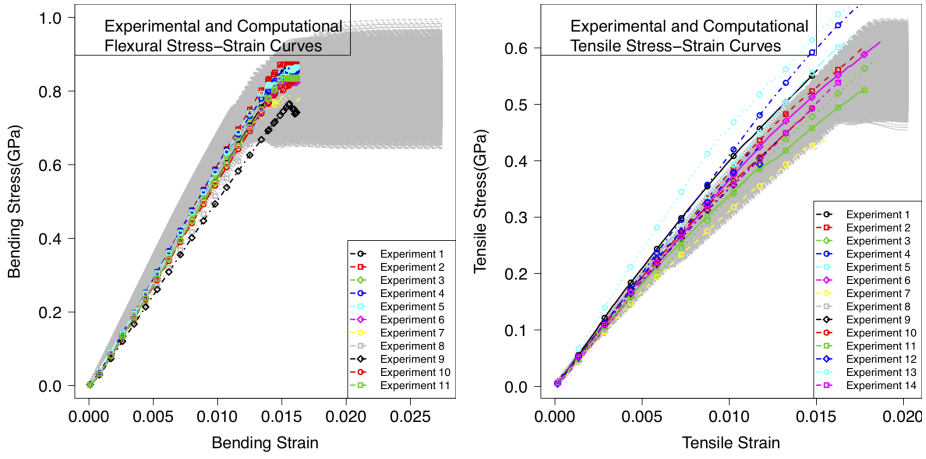

Figure 5: Stress-strain curves for flexural (left) and tensile (right) behavior; Gray cloud is from model simulation; dotted lines are from experiments.

sis (PCA). Accordingly, the eigenvalues and eigenvectors, $(\mu_i, \boldsymbol{\phi}_i)$, of the empirical covariance matrix $[c]$ are used to define a new random matrix variable, $[\boldsymbol{H}]$, through the relationship

$$[\boldsymbol{X}] = [\underline{x}] + [\phi][\mu]^{1/2}[\boldsymbol{H}] \tag{1}$$

where $[\underline{x}]$ is an $n \times N$ matrix whose each column is equal to the sample mean in $\mathbb{R}^n$ obtained from the N samples, $[\phi]$ is an $n \times \nu$ matrix whose columns consist of the dominant ν eigenvectors of $[c]$, and $[\mu]$ is a $\nu \times \nu$ diagonal matrix with the corresponding eigenvalues. The PCA reduction expressed by (1) captures information in the original data $[x]_d$ as encoded in the linear correlation between the n features, averaged over the N data points. It relies for its fidelity on the n-dimensional eigenvectors of the covariance matrix $[c]$. In this construction, the N data points are taken as independent samples, and their proximity from one another is completely glossed over. Expressing the dataset as a matrix $[\boldsymbol{X}]$ is already setting the stage for exploring statistical dependence between its N rows, describing the individual data points. In the sequel, we will work with matrix $[\boldsymbol{H}]$, using equation (1) to map samples of $[\boldsymbol{H}]$ into samples of $[\boldsymbol{X}]$. The initial data $[x_d]$ associated with $[\boldsymbol{X}]$ are thus projected resulting in the corresponding initial dataset $[\eta_d] = [\boldsymbol{\eta}^1 \dots \boldsymbol{\eta}^N]$ associated with $[\boldsymbol{H}]$, represented by the $\nu \times N$ matrix such that $[\eta_d] = [\mu]^{-1/2}[\phi]^T([x_d] - [\underline{x}])$.

4.2. Diffusion maps

The diffusion kernel characterizes proximity between two points in the dataset, using their norm in $\mathbb{R}^\nu$. A typical kernel is provided by the following exponential form,

$$k(\boldsymbol{\eta}, \boldsymbol{\eta}') = e^{-\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|^2/\epsilon} \quad , \quad \boldsymbol{\eta} \in \mathbb{R}^\nu, \boldsymbol{\eta}' \in \mathbb{R}^\nu, \tag{2}$$

in which ϵ is a parameter that has to be adapted to the dataset (see [14] and see below) and where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^\nu$. Mapping the kernel onto the training dataset results in $N \times N$ matrix $[K]$ whose ij element is defined as

$$[K]_{ij} = k(\boldsymbol{\eta}^i, \boldsymbol{\eta}^j), \quad i, j = 1, \cdots, N \tag{3}$$

Introducing the diagonal scaling matrix,

$$[b]_{ij} = \delta_{ij} \sum_{j=1}^{N} [K]_{ij} \quad i, j = 1, \cdots, N \tag{4}$$

matrix $[P] = [b]^{-1}[K]$ is a probability transition matrix between points of the N-vertex graph defined by the initial dataset. It has been shown [19] that the span $[g] = \{g^i\}$, where $g^i = \lambda_i \boldsymbol{\psi}^i$, $(\lambda_i, \boldsymbol{\psi}^i)$ are the eigenvalue and eigenvector pairs of $[P]$, defines an embedding of the graph in an N-dimensional vector space. For an adapted value of ϵ, there is a rapid decay of the eigenvalues of $[P]$. This embedding is actually a localization of the graph to the m-dimensional dominant eigenspace of $[P]$. We then introduce the $N \times m$ matrix $[g] = [\boldsymbol{g}^1 \dots \boldsymbol{g}^m]$. For numerical reasons, the eigensolution of $[P]$ is deduced from that of the symmetric matrix $[P_S] = [b]^{-1/2}[P][b]^{1/2} = [b]^{-1/2}[K][b]^{-1/2}$ [11]. This diffusion map (DMAP) embedding provides a parameterization of points in $\mathbb{R}^\nu$ using in an m-dimensional subspace of $\mathbb{R}^N$, permitting us to express random matrix $[\boldsymbol{H}]$ as,

$$[\boldsymbol{H}] = [\boldsymbol{Z}][g]^T \tag{5}$$

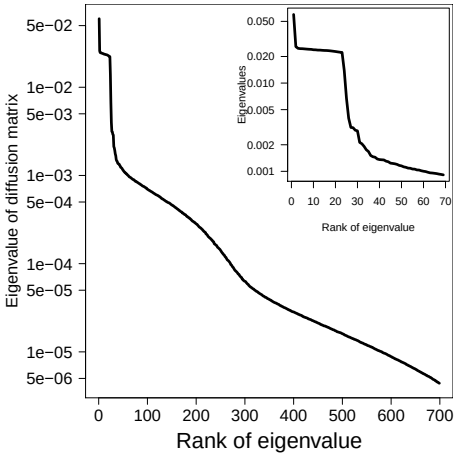

Figure 6: Eigenvalues of the diffusion matrix; inset shows zoom-in on the lower part of the spectrum.

where $[\boldsymbol{Z}] \in \mathbb{R}^{\nu \times m}$ is a reduction of $[\boldsymbol{H}]$ according to the dominant eigenspace of $[P]$. Proximity in this reduction is between points on the graph, averaged over the features. This is to be contrasted with the PCA reduction where proximity is between features, averaged over the graph.

Figure (6) shows the decay of the eigenvalues of the diffusion matrix associated with the optimal value of the bandwidth parameter ϵ. It is clear from this figure that the eigenvalues are grouped in three clusters with the following ranks: less than 21, 21-250, and above 250. While these clusters may reflect the three scales in the data: fibers, tows, and composites, a more detailed analysis is needed to ascertain this claim, and to further investigate the contribution of each of these scales to feature localization. It should be noted that the drop in the eigenvalues can be made milder or sharper by using, respectively, a larger or smaller value of ϵ. Our choice of ϵ is predicated on our desire to achieve a target localization characterized by a 90% reduction in eigenvalues within the smallest embedding dimension m. We use the following criterion to calculate m,

$$m = \arg\min_{\ell} \left| \frac{\lambda_\ell}{\lambda_2} - 0.1 \right| \tag{6}$$

4.3. Construction of the learned dataset (augmenting the dataset)

We first assume that the N columns of $[\eta_d]$ are independent realizations of an $\mathbb{R}^\nu$-valued random vector $\boldsymbol{H}$ whose empirical covariance matrix is the identity matrix, and estimate its PDF as a Gaussian mixture in the form,

$$p_{\boldsymbol{H}}(\boldsymbol{\eta}) = \frac{1}{N} \sum_{j=1}^{N} \pi\left(\frac{\widehat{s_\nu}}{s_\nu} \boldsymbol{\eta}^j - \boldsymbol{\eta} \right), \tag{7}$$

where π is the positive function from $\mathbb{R}^\nu$ into $]0\,, +\infty[$ defined, for all $\boldsymbol{\eta}$ in $\mathbb{R}^\nu$, by

$$\pi(\boldsymbol{\eta}) = \frac{1}{(\sqrt{2\pi}\,\widehat{s_\nu})^\nu} \exp\left\{ -\frac{1}{2\widehat{s_\nu^2}} \|\boldsymbol{\eta}\|^2 \right\}, \tag{8}$$

with $\|\boldsymbol{\eta}\|$ denoting the Euclidean norm in $\mathbb{R}^\nu$ and where the positive parameters s_ν and $\widehat{s_\nu}$ are defined by

$$s_\nu = \left\{\frac{4}{N(2+\nu)}\right\}^{1/(\nu+4)} \quad , \quad \widehat{s_\nu} = \frac{s_\nu}{\sqrt{s_\nu^2 + \frac{N-1}{N}}} \,. \tag{9}$$

With this choice of s_ν and $\widehat{s_\nu}$ the mean-squared error is minimized [20] and realizations of random vector $\boldsymbol{H}$ are normalized, a requirement consistent with their construction through an eigen-decomposition [21]. This is the PDF of random vector $\boldsymbol{H}$ characterized by equation (7), we now consider the joint occurrence of the N data points. This joint behavior is significant as it carries a signature of intrinsic structure not available in each data point separately. We are looking for structure beyond linear correlation. We thus consider matrix $[\eta_d]$ as a realization of a random matrix $[\boldsymbol{H}]$, for which we next construct a probability model. We now invoke our second assumption, whereby we consider the N columns of $[\boldsymbol{H}]$ as statistically independent, with the density of $[\boldsymbol{H}]$ given by,

$$p_{[\boldsymbol{H}]}([\boldsymbol{\eta}]) = p_{\boldsymbol{H}}(\boldsymbol{\eta}^1) \times \ldots \times p_{\boldsymbol{H}}(\boldsymbol{\eta}^N) \,. \tag{10}$$

We thus obtain a nonparametric Gaussian mixture model for the PDF of random matrix $[\boldsymbol{H}]$. Each realization of this random matrix will augment the initial training set represented by $[\eta_d]$ with N new data points each of dimension ν. Alternatively, these realizations are first transformed through the eigenvectors of the empirical covariance of the original data, and are thus used to augment matrix $[x_d]$ with N new data points, each of dimension n. We next recall the procedure for generating samples of $[\boldsymbol{H}]$ from the PDF specified in Equation (10). The approach consists of constructing an Itô equation that is constrained to the manifold, through projections, and whose invariant measure has the density specified by Equation (10). First, we introduce the $N \times m$ projection matrix $[a]$ on the subspace spanned by the $[g]$,

$$[a] = [g]\,([g]^T\,[g])^{-1} \,, \tag{11}$$

the $\nu \times N$ matrix $[\mathcal{N}]$ whose entries are independent standard gaussian variables, the $\nu \times N$ random matrix $[\boldsymbol{H}_d]$ with known realization $[\eta_d]$, and the $\nu \times N$ matrix $[dW(r)]$ $(r \geq 0)$ whose i^{th} column is dW^i with $\{W^i,\ i = 1, \cdots, N\}$ being independent copies of the ν-dimensional normalized Wiener process. It can then be shown that solutions $\{\boldsymbol{Z}(r),\ r \geq 0\}$ of the following Itô stochastic differential equations [11]

$$d[\boldsymbol{Z}(r)] \;=\; [\boldsymbol{Y}(r)]\,dr \,, \tag{12}$$

$$d[\boldsymbol{Y}(r)] \;=\; [L([\boldsymbol{Z}(r)]\,[g]^T)]\,[a]\,dr - \frac{1}{2} f_0\,[Y(r)]\,dr + \sqrt{f_0}\,[dW(r)]\,[a] \,, \tag{13}$$

with the initial condition

$$[\boldsymbol{Z}(0)] = [\boldsymbol{H}_d]\,[a] \quad , \quad [\boldsymbol{Y}(0)] = [\mathcal{N}\,]\,[a] \quad a.s. \,, \tag{14}$$

are samples from the $\nu \times N$ random matrix $[\boldsymbol{H}] = [\boldsymbol{Z}]\,[g]^T$ with PDF $p_{[\boldsymbol{H}]}([u]) = c\,q([u])$ in which c is a constant of normalization, and where

$$[L([u])]_{k\ell} = \frac{\partial}{\partial u_k^\ell} \log\{q(u^\ell)\} \quad , \quad [u] = [\boldsymbol{u}^1, \ldots, \boldsymbol{u}^N] \,. \tag{15}$$

Given our choice of Gaussian mixture model for q, the expression for $[L]$ can be expanded as follows,

$$[L([u])]_{k\ell} = \frac{1}{q(u^\ell)} \frac{1}{\widehat{s_\nu^2}} \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\widehat{s_\nu}}{s_\nu} \eta_k^j - u_k^\ell \right) \exp\left\{ -\frac{1}{2\widehat{s_\nu^2}} \left\| \frac{\widehat{s_\nu}}{s_\nu} \eta^j - u^\ell \right\|^2 \right\} \ . \tag{16}$$

The Itô equations specified by Equations (12) and (13) are solved using a Störmer-Verlet algorithm, a symplectic scheme well-adapted to Hamiltonian non-dissipative systems [22]. In that scheme, we use a value of f_0 equal to 1, and an integration step Δr of 0.1.

Figure (7) shows the stress-strain components of the training dataset and of the augmented dataset for the composites application. Curves for each of the tensile and fluxural loading configurations are shown. The range of these curves and their overall behavior are very similar. The full set of observables consists of, in addition to tensile and flexural curves, the microscale constituents (mechanical properties of fibers and resin) and meso-scale constituents (diameters and separations of tows). Figure (8) shows the probability density function (PDF), obtained from the augmented data, of the bending stress at a strain level of 0.0117, and the tensile stress at a strain level of 0.015. Results are shown as the size of the training set is increased from 100 to 1000. Convergence is observed at a sample size of 800. This figure shows convergence of the marginal density function, whereas in our cross-inference studies a high-order joint-density function is required. Detailed studies of higher-order convergence is presented elsewhere [12, 15]. Figure (fig:Stress-PDF) shows the PDF for bending and tensile stresses at different strain levels. The dotted lines in these figures show the PDF from the training set while the solid lines show the PDF from the augmented dataset. It is clear, in all cases, that augmented datasets have a slightly narrower support than the training set, since they encode additional constraints. Also, since the augmented dataset has significantly more samples than the training sets, they provide a finer resolution of the quantities of interest as manifested by the finer details in the associated PDF.

4.4. Non-parametric conditioning

We rely on non-parametric kernel estimation to evaluate conditional expectations and conditional distributions. We implement non-parametric conditioning as presented in [12, 20],

$$\mathbb{E}\{R \,|\, \mathbf{W} = \mathbf{w}_o\} \ \simeq \ \frac{\displaystyle\sum_{\ell=1}^{\nu_s} \widehat{r^\ell} \ \exp\left\{ -\frac{1}{2s^2} \|\widehat{\mathbf{w}}^\ell - \widehat{\mathbf{w}}_o\|^2 \right\}}{\displaystyle\sum_{\ell=1}^{\nu_s} \exp\left\{ -\frac{1}{2s^2} \|\widehat{\mathbf{w}}^\ell - \widehat{\mathbf{w}}_o\|^2 \right\}} \ \sigma_R \ + \overline{R} \, . \tag{17}$$

In the above, $\widehat{r^\ell}$ and $\widehat{\mathbf{w}}^\ell$ are realizations of $\widehat{R}$ and $\widehat{\mathbf{W}}$, respectively. The probability distribution function (CDF) of random variables R conditional on $\mathbf{W} = \mathbf{w}_o$ can be expressed as $1 - h(r, \mathbf{w}_o)$ where the survival function h is given by,

$$h(r, \mathbf{w}_o) = \mathbb{P}[R > r \,\big|\, \mathbf{W} = \mathbf{w}_o] \ \simeq \ \int_r^{+\infty} p_{R|\mathbf{W}}(r|\mathbf{w}_o) \, dr \, , \tag{18}$$

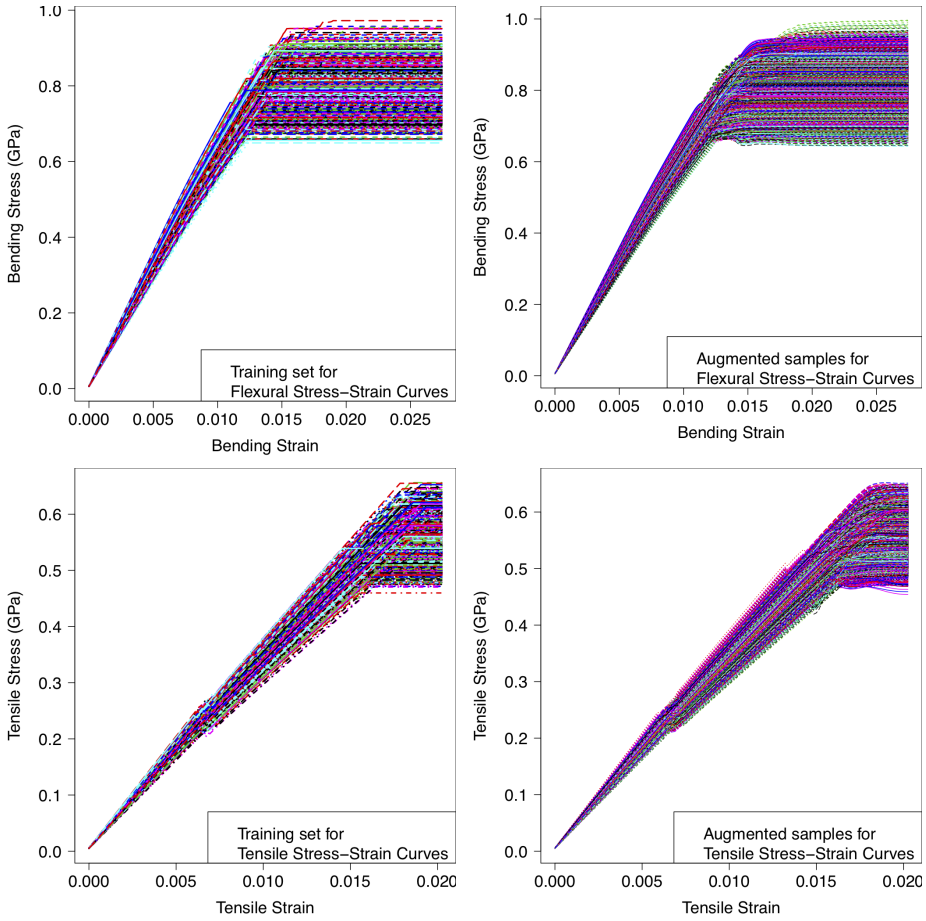

Figure 7: Training datasets (left column) and augmented datasets (right column) for bending model (top row) and tension model (bottom row).

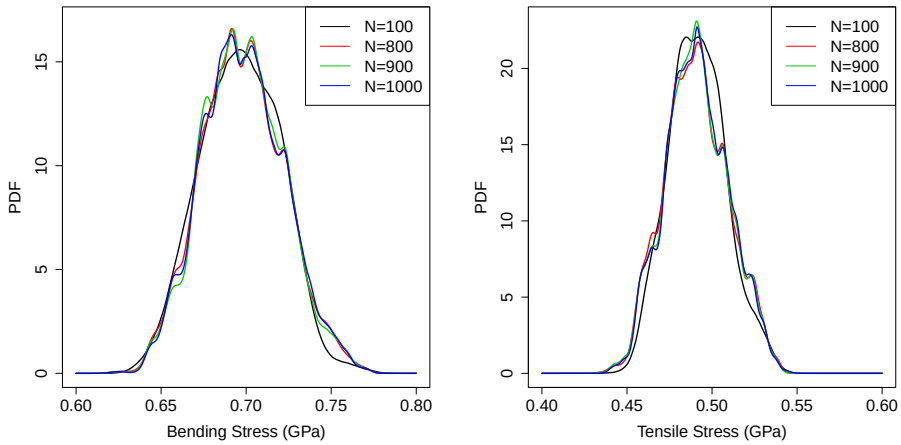

Figure 8: Convergence of the PDF for bending stress at strain level of 0.0117 (left figure) and tensile for tensile strain level of 0.015 (right figure) stresses

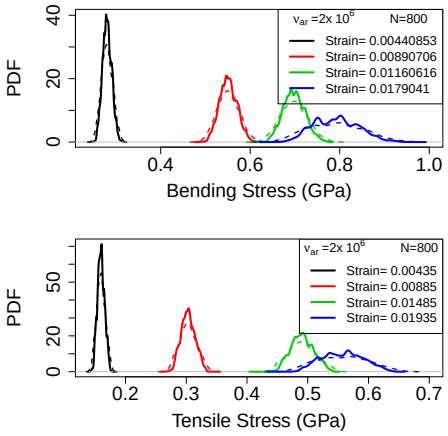

Figure 9: PDF of the flexural (top) and tensile stresses (bottom) at different strain levels; leftmost curves are in the linear elastic regime, and rightmost curves are in the inelastic regime.

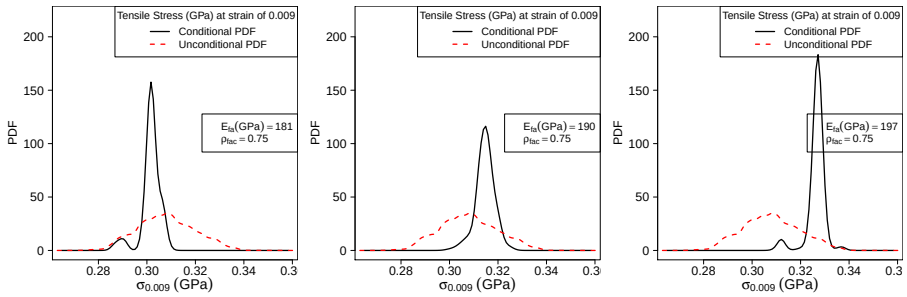

Figure 10: PDF of tensile elastic stress at strain value of 0.009, conditional on fiber elastic modulus and fiber compression ratio of 0.75 for different values of elastic modulus.

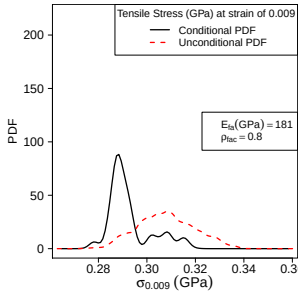

Figure 11: PDF of tensile elastic stress at strain value of 0.009, conditional on fiber elastic modulus and fiber compression ratio of 0.8.

which can be estimated as [13]

$$h(r, \mathbf{w}_o) \simeq \frac{\sum_{\ell=1}^{\nu_s} h^\ell(\widehat{r}) \, \exp\left\{-\frac{1}{2s^2}\|\widehat{\mathbf{w}}^\ell - \widehat{bfw}_o\|^2\right\}}{\sum_{\ell=1}^{\nu_s} \exp\left\{-\frac{1}{2s^2}\|\widehat{\mathbf{w}}^\ell - \widehat{\mathbf{w}}_o\|^2\right\}} \tag{19}$$

where

$$\widehat{r} = (r - \overline{R})/\sigma\,, \qquad h^\ell(\widehat{r}) = \frac{1}{2}\left(1 - \mathrm{erf}\left((\widehat{r} - \widehat{r}^\ell)/(s\sqrt{2})\right)\right) \tag{20}$$

and $\mathrm{erf}(y) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^y e^{-t^2}\, dt$ is the error function. We finally note that the corresponding conditional density function can be obtained by differentiating the CDF with respect to r which is approximated through a first order finite difference scheme.

5. Cross-scale inference

In the context of the present probabilistic model, prediction is interpreted as statistical conditioning. We first infer the PDF of the measured composite stresses under tension test, conditioned

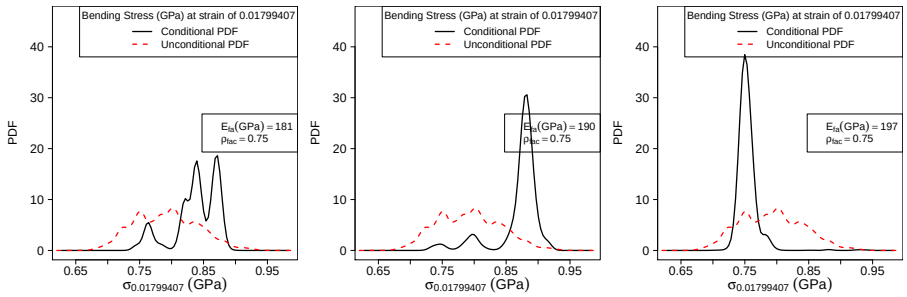

Figure 12: PDF of bending elastic stress at strain value of 0.009, conditional on fiber elastic modulus and fiber compression ratio of 0.75 for different values of elastic modulus.

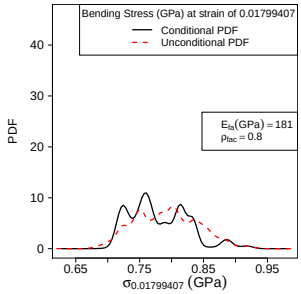

Figure 13: PDF of bending elastic stress at strain value of 0.009, conditional on fiber elastic modulus and fiber compression ratio of 0.8.

on different values of fiber elastic modulus and fiber compression ratio. The black curves in Figure (10) shows the increase in elastic stress at strain value of 0.009 of the composite as the fiber axial elastic modulus is increased while keeping the fiber axial compression ratio constant. The slightly bimodal behavior observed for smaller and larger values of fiber modulus are hints of additional subscale influences beyond elastic fiber moduli. Increasing the compression ratio for fixed value of the fiber modulus results in a bigger scatter in the composite elastic modulus, as observed in Figure (11). The red dashed curves in these figures show the unconditioned probability density functions of the composite moduli. As expected, conditioning results in a clear concentration of probability over a support that is favorable to the conditioning set. A similar behavior is observed for stresses in the inelastic regime. It is worth noting that although fiber compression is not activated during the tension test, the probabilistic learning is conducted using simultaneously bending stresses and tensile stresses. Thus, when conditioning on specific values of compression ratios, we are conditioning fibers with particular behavior in bending, which in turn influences the behavior in tension. This coupling between tension and bending is arguably mediated through the morphology of the composite, which is generated an implicit map between microscale and composite-scale properties.

Figure (12) shows results associated with bending stresses in the inelastic regime, at a bending strain level of 0.0179. The figures shows probability density functions of the bending stress conditional on different values of fiber elastic modulus and a value of compression ratio equal to 0.75. Figure (13) shows the conditional PDF for a value of compression ratio equal to 0.8.

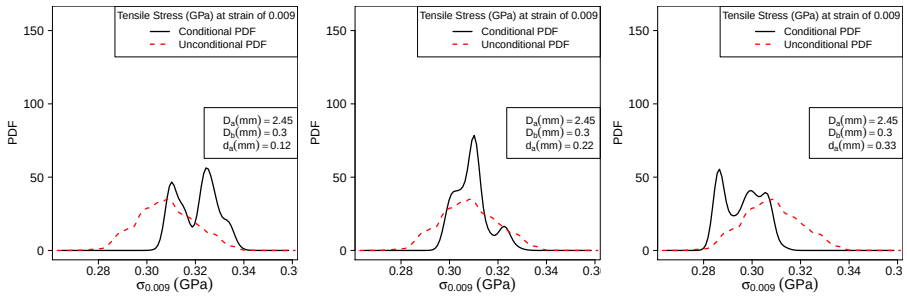

Figure 14: PDF of elastic tensile stresses at strain level of 0.009, conditional on different values of tow spacing d_a, and D_a=2.45, D_b=0.22.

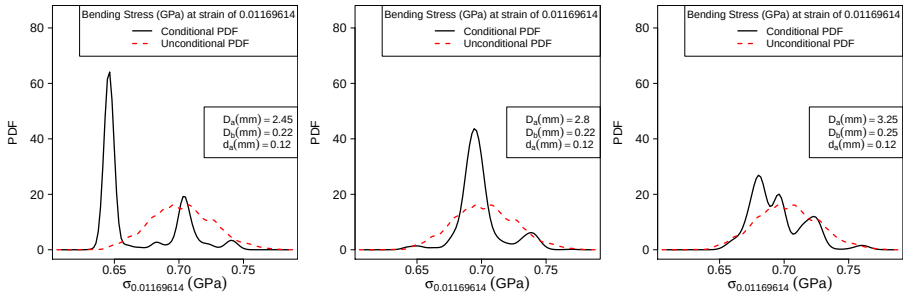

Figure 15: PDF of inelastic bending stresses at strain level of 0.0117, conditional on different values of tow major diameter D_a, and D_b=0.22, d_a=0.12.

In contrast to the tensile stress case, we observe here that as the composite stress levels initially increase with increasing fiber modulus but that they drop significantly for larger values of fiber modulus. It is worth noting that the lowest mode and highest mode in the left-most plot in Figure (12) (E_{fa} = 181 GPa) correspond, respectively, to the dominant modes in the middle plot of the same figure (E_{fa} = 190 GPa and 197 GPa respectively). These modes seem to be associated with distinct predispositions that are activated as the fiber modulus is increased. The decrease in composite stress for stiffer fibers suggests the dominance, in bending, of fiber resistance over that of resin. It is also observed, as expected, that increasing the fiber compression ratio induces a much greater spread in bending stress than it did for tensile stress.

For the same loading conditions, an increase in d_a is generally associated with a decrease in stresses in the composite as shown in Figure (14). This general trend was observed for tensile tensile both in the elastic and inelastic regimes, but was not systematically noted in flexural behavior. As observed in Figure (15), an increase in D_a, the major diameter of the tows, is accompanied with a general increase in the stress levels. Figure (15) shows results in the inelastic regime of bending, similar results were observed for all strain levels in bending. Under tension loading, the converse was observed, with stresses in the composite generally reducing with increasing D_a. Figure (16) shows a typical influence of D_b, the minor diameter of the tow. As this D_b is increased, a switch in the bimodality of stress takes effect, with the higher mode becoming more pronounced and the probability of the lower mode reduced accordingly. This effect is more apparent in bending than in tension.

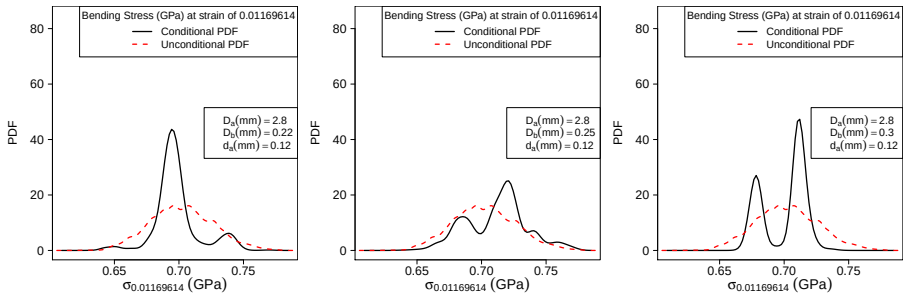

Figure 16: PDF of inelastic bending stresses at strain level of 0.0117, conditional on different values of tow minor diameter D_b, and D_a=2.8, d_a=0.12.

Conditioning simultaneously on elastic and inelastic fiber properties does seem to have a smoothing effect on the bimodality of the PDF, even for stress levels in the elastic range. This is due to the fact that the constructed joint probabilistic model is trained using joint information about tensile and bending behavior observed simultaneously in the elastic and inelastic regimes. Thus, conditioning on particular behavior in t he inelastic regime, automatically favors those specimens that are closest to that behavior, including the elastic properties (and all other features used in the learning process). If we condition, in addition, on the tow properties (D_a, D_b and d_a), the the PDFs are even smoother, but the bimodal behavior persists, specially in the bending regime (both elastic and inelastic).

In addition to conditioning on values of fiber and tow properties, we also condition on values of stress assessed at various levels of strain. In one case we infer the fiber and tow properties from such observations, in another case, we infer the missing stress values. Figure (17) shows the PDF of fiber parameters and tow geometry when both the bending and tensile stresses are observed simultaneously over a range of strains.

Inferring the tensile stresses on the bending stresses, or vice-versa, was surprisingly not informative: the support of the PDF remained essentially unchanged, but a few multimodal features emerged upon conditioning. These feature were not sufficiently systematic to allow physical insight.

Finally, we conditioned on values of stress associated with each of the experiments, and estimate the conditional PDF of fiber and tow properties. We condition separately on experimental tensile stresses and experimental bending stresses. Each bending experimental record consists of stress values at 306 strain values while each tensile dataset contains 137 points. Conditioning datasets are obtained by downsampling the original data each tenth point, resulting in datasets evaluated at 31 and 14 strain values. Each resampled dataset is used in turn to condition the joint density function in order to update the PDF of the fiber and tow parameters. Although this procedure can conceptually be carried out sequentially, we apply it separately to each experimental dataset, obtaining a total of 25 updates to the subscale properties (11 bending experiments and 14 tensile experiments). Tensile experiment number 5, which lies completely outside the computational cloud (Figure (5)) yielded a zero updated probability, since the measured data is supported with zero-probability by the training and augmented datasets. All other datasets successfully updated the subscale parameters, in some instances shifting them towards the outer edges of the PDF support. Figure (18) shows typical results from this updating procedure. It is noted that the updated PDF is often bimodal indicating some fundamental underpinning for this behavior.

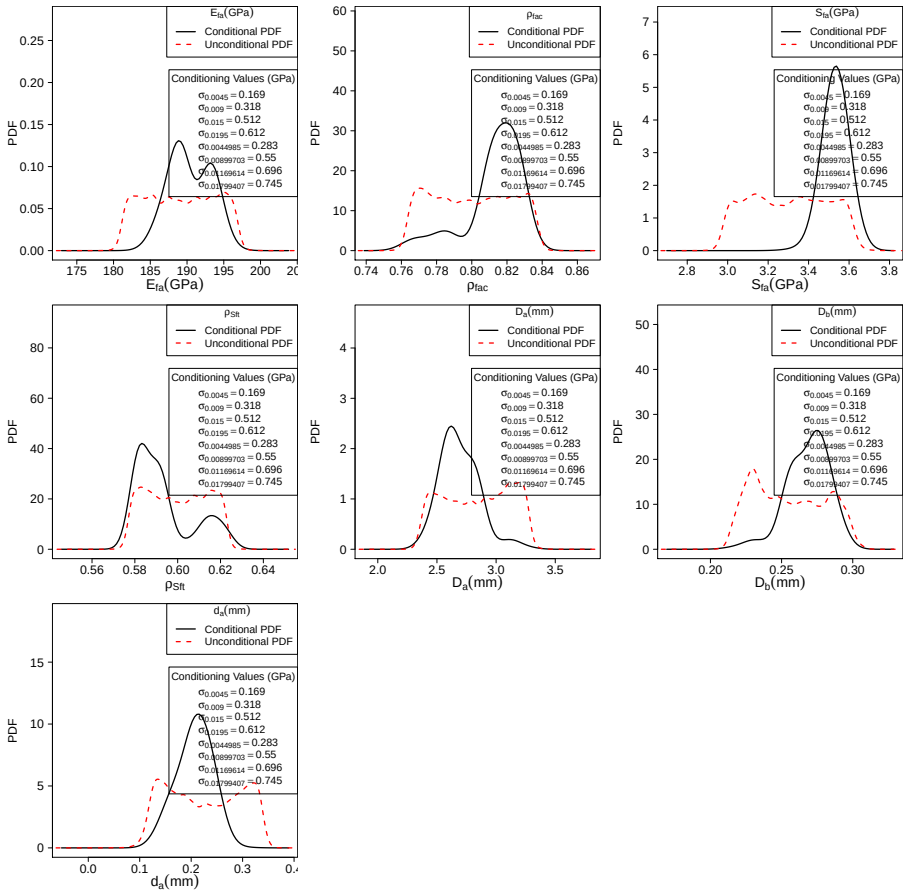

Figure 17: PDF of E_{fa}, ρ_{fac}, S_{ofa}, ρ_{aft}, D_a, D_b, and d_a when stress values in tension experiment (first 4 conditioning values) and bending experiment (last four conditioning values) are specified; conditional (black solid curve) and unconditional (red dashed curve) PDFs are shown.

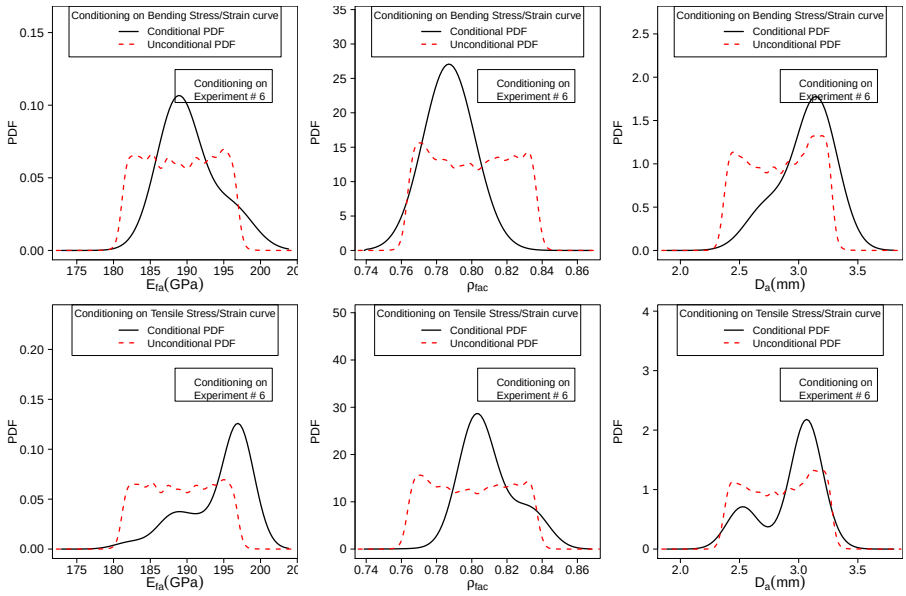

Figure 18: Initial and conditional PDFs for fiber and tow properties; conditioning is done on bending experiments (top row) and tensile experiments (bottom row).

6. Conclusions

In spite of the intricate multiscale interactions within a composite material, or perhaps because of it, a well defined intrinsic structure could be delineated from a relatively small number of numerical or experimental samples. This structure, viewed with the proper probabilistic measure, yields a probabilistic machine learning (PML) algorithm that enables statistical inference for complex cross-scale queries. The PML is trained on numerically generated data, and yet conditioning on experimental stress-strain curves yields reasonable inferences for fine scale material properties and geometry. One important outcome of the proposed probabilistic approach, as demonstrated in the composites example, is to expose relationships between mechanical properties and behaviors at different scales that should be pursued experimentally or with detailed numerical simulations.

7. Acknowledgment

Partial funding for this work was provided by the Office of Energy Efficiency and Renewable Energy (EERE), U.S. Department of Energy, under Award Number DE-EE0006826. Funding was also provided by DARPA's Equips program.

References

[1] McWilliams K, Lacy TE Jr, Roy S, Jha R. Development of an ICME-Based Airframe Digital Twin Model of an All-Composite Air Vehicle. *Proceedings of the American Society for Composites*, Bakis, CE (ed.), 2013.

[2] Schleich B, Anwer N, Mathieu L, Wartzack S. Shaping the digital twin for design and production engineering. *CIRP Annals-Manufacturing Technology* 2017; **66**(1):141–144.

[3] Qinglin Q, Tao F. Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access* 2018; **6**:3585–3593.
[4] Okaro IA, Jayasinghe S, Sutcliffe C, Black K, Paoletti P, Green PL. Automatic fault detection for laser powder-bed fusion using semi-supervised machine learning. *Additive Manufacturing* 2019; .
[5] Beuth J, Klingbeil N. The role of process variables in laser-based direct metal solid freeform fabrication. *JOM* 2001; .
[6] Aoyagi K, Wang H, Sudo H, Chiba A. Simple method to construct process maps for additive manufacturing using a support vector machine. *Additive Manufacturing* 2019; .
[7] Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integr Mater Manuf Innov* 2017; .
[8] Hunter A, Moore BA, Mudunuru M, Chau V, Tchoua R, Nyshadham C, Karra S, O'Malley D, Rougier E, Viswanathan H, *et al.*. Reduced-order modeling through machine learning and graph-theoretic approaches for brittle fracture applications. *Computational Materials Science* 2019; **157**.
[9] Rovinelli A, Sangid MD, Proudhon H, Ludwig W. Using machine learning and a data-driven approach to identify the small fatigue crack driving force in polycrystalline materials. *npj Computational Materials* 2018; **4**(35).
[10] Moore BA, Rougier E, O'Malley D, Srinivasan G, Hunter A, Viswanathan H. Predictive modeling of dynamic fracture growth in brittle materials with machine learning. *Computarional Materials Science* 2018; **148**.
[11] Soize C, Ghanem R. Data-driven probability concentration and sampling on manifold. *Journal of Computational Physics* 2016; **321**:242–258, doi:10.1016/j.jcp.2016.05.044.
[12] Ghanem R, Soize C. Probabilistic nonconvex constrained optimization with fixed number of function evaluations. *International Journal for Numerical Methods in Engineering* 2018; **113**(4):719–741, doi:10.1002/nme.5632.
[13] Ghanem R, Soize C, Thimmisetty CR. Optimal well-placement using a probabilistic learning. *Data-Enabled Discovery and Applications* 2018; **2**(1):1–16, doi:10.1007/s41688-017-0014-x.
[14] Soize C, Ghanem R, Safta C, Huan X, Vane Z, Oefelein J, Lacazc G, Najm H, Tang Q, Chen X. Entropy-based closure for probabilistic learning on manifolds. *Journal of Computational Physics* 2019; **388**:518–533.
[15] Ghanem RG, Soize C, Safta C, Huan X, Lacaze G, Oefelein JC, Najm HN. Design optimization of a scramjet under uncertainty using probabilistic learning on manifolds. *Journal of Computational Physics* 2019; **399**:108 930, doi:10.1016/j.jcp.2019.108930.
[16] Mehrez L, Fish J, Aitharaju V, Rodgers W, Ghanem R. A pce-based multiscale framework for the characterization of uncertainties in complex systems. *Computational Mechanics* 2018; :219–236.
[17] Altair. *Altair MDS*. http://www.altair.com/, Altair, 2015.
[18] Fish J. *Practical Multiscaling*. John Wiley & Sons, Ltd., 2014.
[19] Coifman R, Lafon S, Lee A, Maggioni M, Nadler B, Warner F, Zucker S. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS* 2005; **102**(21):7426–7431.
[20] Scott D. *Multivariate Density Estimation: Theory, Practice, and Visualization*. 2nd edn., John Wiley and Sons: New York, 2015.
[21] Soize C. Polynomial chaos expansion of a multimodal random vector. *SIAM/ASA Journal on Uncertainty Quantification* 2015; **3**(1):34–60, doi:10.1137/140968495.
[22] Burrage K, Lenane I, Lythe G. Numerical methods for second-order stochastic differential equations. *SIAM Journal on Scientific Computing* 2007; **29**(1):245–264.