Science Arts & Métiers (SAM)

is an open access repository that collects the work of Arts et Métiers Institute of Technology researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: https://sam.ensam.eu
Handle ID: .http://hdl.handle.net/10985/18955

To cite this version :

Minyoung YUN, Clara ARGERICH, Elias CUETO, Jean Louis DUVAL, Francisco CHINESTA SORIA - Nonlinear Regression Operating on Microstructures Described from Topological Data Analysis for the Real-Time Prediction of Effective Properties - Materials - Vol. 13, n°10, p.1-12 - 2020

Any correspondence concerning this service should be sent to the repository Administrator : scienceouverte@ensam.eu

Arts et Métiers Institute of Technology

Nonlinear Regression Operating on Microstructures Described from Topological Data Analysis for the Real-Time Prediction of Effective Properties

Minyoung Yun [1], Clara Argerich [1], Elias Cueto [2], Jean Louis Duval [3] and Francisco Chinesta [1,3,*]

[1] PIMM Laboratory & ESI Group Chair, Arts et Métiers Institute of Technology, CNRS, Cnam, HESAM Université, 151 boulevard de l'Hôpital, 75013 Paris, France; minyoung.yun@ensam.eu (M.Y.); clara.argerich_martin@ensam.eu (C.A.)
[2] Aragon Institute of Engineering Research, Universidad de Zaragoza, 50009 Zaragoza, Spain; ecueto@unizar.es
[3] ESI Group, Bâtiment Seville, 3bis rue Saarinen, 50468 Rungis, France; Jean-Louis.Duval@esi-group.com
* Correspondence: Francisco.Chinesta@ensam.eu

Abstract: Real-time decision making needs evaluating quantities of interest (QoI) in almost real time. When these QoI are related to models based on physics, the use of Model Order Reduction techniques allows speeding-up calculations, enabling fast and accurate evaluations. To accommodate real-time constraints, a valuable route consists of computing parametric solutions—the so-called computational vademecums—that constructed off-line, can be inspected on-line. However, when dealing with shapes and topologies (complex or rich microstructures) their parametric description constitutes a major difficulty. In this paper, we propose using Topological Data Analysis for describing those rich topologies and morphologies in a concise way, and then using the associated topological descriptions for generating accurate supervised classification and nonlinear regression, enabling an almost real-time evaluation of QoI and the associated decision making.

Keywords: machine learning; data-driven mechanics; TDA; *Code2Vect*; nonlinear regression; effective properties; microstructures

1. Introduction

Recently, industry is experiencing a new revolution. In the past, product design, as well as their associated manufacturing processes, were based on the use of nominal models, nominal loadings (in their broadest sense), and a small amount of data for calibrating those models, with the product performance as a design target.

Very recently, predictions enabling real-time decision-making targeting zero defects in processing and zero unexpected faults in operation, were needed everywhere within the Internet of Things (IoT) paradigm, on the work-floor (smart processes), in the city (autonomous systems and smart-city), at the nation level (e.g., smart nation), etc., i.e., anywhere where engineering designs operate.

In those circumstances, the use of traditional simulation-based engineering (SBE) that was the major protagonist of 20th century engineering, is not anymore a valuable option due to three main reasons: (i) models become sometimes crude approximations of the observed reality; (ii) assimilating data enabling the continuous calibration of the models in operation remains difficult to perform under the stringent real-time constraint; and (iii) the real-time simulation of those extremely complex mathematical models needs alternative techniques to those commonly employed in traditional SBE.

It was at the beginning of the XXI century that two new revolutions in the domain of digital engineering emerged.

1.1. Model Order Reduction

Advances in applied mathematics, computer science (high-performance computing) and computational mechanics met to give rise to a diversity of Model Order Reduction (MOR) techniques [1]. These techniques do not reduce or modify the model, they simply reduce the complexity of its resolution and thus transform a complex and time-consuming calculation, into a real-time response while maintaining precision. These new techniques have completely altered traditional approaches of simulation, optimization, inverse analysis, control and uncertainty propagation, all them operating under the stringent real-time constraint.

In a few words, when approximating the solution $u(\mathbf{x}, t)$ of a given Partial Differential Equation (PDE), the multipurpose finite element method assumes an approximation

$$u(\mathbf{x}, t) = \sum_{i=1}^{\mathtt{N}} U_i(t) N_i(\mathbf{x}), \tag{1}$$

where U_i represents the value of the unknown field at node i and $N_i(\mathbf{x})$ is the associated shape function. When $\mathtt{N}$ (the number of nodes) increases the solution process becomes cumbersome.

POD-based model order reduction learns offline the most adequate (in a given sense) reduced approximation basis $\{\phi_1(\mathbf{x}), \cdots, \phi_{\mathtt{R}}(\mathbf{x})\}$, and project the solution in it

$$u(\mathbf{x}, t) \approx \sum_{i=1}^{\mathtt{R}} \xi_i(t) \phi_i(\mathbf{x}), \tag{2}$$

where now, the complexity scales with $\mathtt{R}$ instead of $\mathtt{N}$, with $\mathtt{R} \ll \mathtt{N}$ in general.

The so-called Proper Generalized Decomposition (PGD from now on) goes a step forward and assume a general approximation

$$u(\mathbf{x}, t) \approx \sum_{i=1}^{\mathtt{M}} T_i(t) X_i(\mathbf{x}), \tag{3}$$

where now both the space and time functions, $X_i(\mathbf{x})$ and $T_i(t)$ respectively, are computed during the solution process.

A particularly appealing extension of the just introduced space-time separated representation consists of the space-time-parameter separated representation leading to the a so-called *computational vademecum* that expresses the solution of a parametrized PDE from [2,3]

$$u(\mathbf{x}, t, \mu_1, \ldots, \mu_{\mathtt{Q}}) \approx \sum_{i=1}^{\mathtt{M}} X_i(\mathbf{x}) T_i(t) \prod_{j=1}^{\mathtt{Q}} M_i^j(\mu_j), \tag{4}$$

where μ_j, $j = 1, \ldots, \mathtt{Q}$, represent the model parameters. Once constructed off-line that parametric solution (4), it offers under very stringent real-time constraints—in the order of milliseconds—simulation, optimization, inverse analysis, uncertainty propagation and simulation-based control, to cite a few. Thus, at the beginning of the third millennium a real-time dialogue with physics no longer seemed to be the domain of the impossible.

PGD-based techniques have been widely considered for the real-time simulation and decision-making in a variety of problems of industrial relevance. However, prior to use it, one must extract the parameters to be included as extra-coordinates in the problem statement, and then included in the parametric representation of its solution. In the case of morphological and topological descriptions, as considered later in the present work, the extraction of the adequate parametrization

represents the most difficult task. Some attempts of combing PGD-based MOR and manifold learning [4] were addressed in [5–8].

1.2. Engineered Artificial Intelligence

Data bursts within engineering disciplines. For years, data was used in other areas where models were less developed or remained quite inaccurate. Data collected massively was successfully classified, cured, distilled, ... using artificial intelligence (AI) techniques. Thus, correlations between data ca be removed, proving that a certain simplicity remains hidden behind a rather apparent complexity. Data-driven modeling developed exponentially and advanced artificial intelligence techniques were developed, covering six major domains: (i) Multidimensional data visualization [9]; (ii) Data classification and clustering [10,11]; (iii) Learning models from input/output pairs of data, with adequate techniques enabling real-time learning and able to operate in the low-data limit (e.g., sPGD [12], *Code2Vect* [13], iDMD [14–16], NN [17], ThemodynML [5,18], ...); (iv) Knowledge extraction in order to identifying combined parameters and model richness/complexity, discovering hidden parameters, discarding useless parameters or even to extract governing equations; (v) Explaining for certifying; and (vi) Hybridizing physics and data for defining advanced and powerful Dynamic Data-Driven Application Systems, DDDAS [19].

However, these data-driven models, when used in engineering and industry, were quickly confronted with three major and recurrent difficulties: (i) the need for a huge amount of data to make predictions accurate and reliable, knowing that data is synonymous with cost (acquisition and processing costs); (ii) the difficulty of explaining and interpreting predictions obtained by artificial intelligence; and (iii) related to the the latter, the difficulty of certifying engineering products.

1.3. Towards Real-Time Decision Making

In summary, on one side models based on physics can be solved fast but, as discussed, in many engineering areas they remain poor approximations of the real components and systems. On the other hand, when approaching the problem from the data perspective, impressive amounts of data are sometimes needed (with the associated cost and technological difficulty of collecting them), to be processed in real time and then explained in order to certify both the designs and the decisions.

A possible winning option consists of merging both concepts and methodologies. The *hybrid paradigm* was born [19,20], associating in it two type of models: the first based on physics; the second being a completely new type of model, more pragmatic and phenomenological, based on data.

Real-time decision making in engineering design, manufacturing and predictive and operational maintenance, needs the evaluation of quantities of interest in almost real-time. The present work aims at proposing a technique able to determine under the stringent real-time constraints, effective properties of a complex microstructure by assimilating an image of it.

To conciliate accuracy and real-time constraints, the hybrid paradigm is retained: (i) the prediction engine will be trained offline from data coming form physics; then (ii) a non-linear regression, acting on some topological descriptors extracted from those images, will ensure a real-time evaluation of the effective properties (in the present case the homogenized thermal conductivity).

As previously discussed, dealing with shapes and topologies, the parametric description requires performant techniques able to express them in a compact and concise way. In this paper, we propose using Topological Data Analysis, TDA [21], for representing these rich topologies and morphologies, and then using the associated topological descriptors for generating accurate supervised classification and nonlinear regressions, enabling an almost real-time evaluation of the quantities of interest.

In the next section we will present the main methodologies used in the present study, that will be considered later for the training and then for the real-time evaluation of effective properties (homogenized thermal conductivity) of rich microstructures.

2. Methods

This section revisits the main methodologies that will be considered later for real-time classification and prediction of the effective thermal properties from collected images. For that purpose we will consider a rich enough training stage that consists of generating several microstructures whose effective thermal conductivity will be evaluated by using a standard linear homogenization technique, revisited in Section 2.1.

In order to associate the resulting homogenized conductivity tensor to each microstructure, the last must be described in a compact and concise way. For that purpose Topological Data Analysis and Principal Component Analysis (PCA) will be employed. Both are revisited in Sections 2.2 and 2.3, respectively.

The last step aims at performing a nonlinear regression to link the parameters extracted by the TDA to the thermal conductivity. The technique retained in our study is the so-called *Code2Vect* nonlinear regression, revisited in Section 2.4.

2.1. Linear Homogenization Procedure

Due to the microscopic nature of heterogeneity, a procedure is required for extracting the effective thermal conductivity. In what follows we proceed in the linear case, as was also the case in [22], in a representative volume element Ω with a microstructure perfectly defined at that scale. The microscopic conductivity $\mathbf{k}(\mathbf{x})$ is known at every point $\mathbf{x} \in \Omega$.

The macroscopic temperature gradient $\mathbf{G}$ is defined from the space average

$$\mathbf{G} \;=\; \langle \mathbf{g}(\mathbf{x}) \rangle \equiv \frac{1}{|\Omega|} \int_{\Omega} \mathbf{g}(\mathbf{x})\, d\mathbf{x}, \tag{5}$$

where $\mathbf{g}(\mathbf{x})$ represents the microscopic temperature gradient, i.e., $\mathbf{g}(\mathbf{x}) = \nabla T(\mathbf{x})$.

We define the localization tensor $\mathbf{L}(\mathbf{x})$ such that

$$\mathbf{g}(\mathbf{x}) = \mathbf{L}(\mathbf{x})\; \mathbf{G}. \tag{6}$$

The microscopic heat flux $\mathbf{q}(\mathbf{x})$ follows the Fourier law

$$\mathbf{q}(\mathbf{x}) \;=\; -\mathbf{k}(\mathbf{x})\; \mathbf{g}(\mathbf{x}), \tag{7}$$

and its macroscopic counterpart $\mathbf{Q}$ reads

$$\mathbf{Q} \;=\; \langle \mathbf{q}(\mathbf{x}) \rangle \;=\; \langle -\mathbf{k}(\mathbf{x})\; \mathbf{g}(\mathbf{x}) \rangle \;=\; \langle -\mathbf{k}(\mathbf{x})\; \mathbf{L}(\mathbf{x}) \rangle\; \mathbf{G}, \tag{8}$$

from which the homogenized thermal conductivity reads

$$\mathbf{K} \;=\; \langle -\mathbf{k}(\mathbf{x})\; \mathbf{L}(\mathbf{x}) \rangle. \tag{9}$$

Thus, the calculation of the homogenized thermal conductivity tensor only needs the computation of the tensor $\mathbf{L}(\mathbf{x})$. The present work considers the simplest procedure that in the 2D case consists of solving two steady state thermal problems in Ω

$$\begin{cases} \nabla \cdot \left(\mathbf{k}(\mathbf{x})\; \nabla T^1(\mathbf{x}) \right) \;=\; 0 \\ T^1(\mathbf{x} \in \partial\Omega) = x, \end{cases}, \tag{10}$$

and

$$\begin{cases} \nabla \cdot \left(\mathbf{k}(\mathbf{x})\; \nabla T^2(\mathbf{x}) \right) \;=\; 0 \\ T^2(\mathbf{x} \in \partial\Omega) = y \end{cases}, \tag{11}$$

whose solutions verify by construction

$$\left\{ \begin{array}{rcl} \mathbf{G}^1 & = & \langle \nabla T^1(\mathbf{x}) \rangle^T = (1,0) \\ \mathbf{G}^2 & = & \langle \nabla T^2(\mathbf{x}) \rangle^T = (0,1) \end{array} \right. , \tag{12}$$

and whose gradients define the localization tensor columns

$$\mathbf{L}(\mathbf{x}) = \left(\nabla T^1(\mathbf{x}) \quad \nabla T^2(\mathbf{x}) \right), \tag{13}$$

that allows calculating the effective thermal conductivity.

Remark 1. *In the present work, since we are only interested in the effective thermal conduction along the y-direction, a single problem, problem (11), suffices for calculating the only component of interest, component* K_{22}.

2.2. Topological Data Analysis

Topological data analysis, TDA [21], is one of the most promising techniques in high-dimensional data analysis. In essence, TDA is a powerful tool to find the topology of data: if there are clusters, a manifold structure or even noise that is not relevant for the analysis.

For an intuitive description of the method consider the set of points depicted in Figure 1. In general, these points will live in high dimensional spaces, such that their intrinsic topology will not be visible at first glance. We then equip the set with a distance parameter r. By making r grow, different k-simplexes will appear. Remember that a 0-simplex is a point, a 1-simplex is an edge, a 2-simplex is a triangle, on so on.

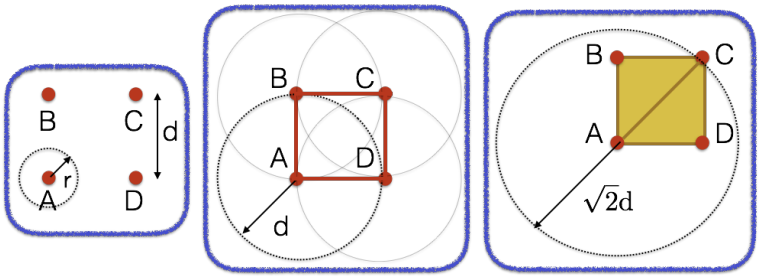

Figure 1. Illustrating TDA: (**left**) For $r < d$ the four points (A, B, C and D) remain disconnected; (**center**) At $r = d$ the hole ABCD appear from the four edges AB, BC, CD and DA; (**right**) The just created hole persist until $r = \sqrt{2}d$, value at which A connects with C and the two resulting triangles ABC and ACD cover the initial hole that disappears consequently.

As r grows, holes appear (as the one defined by the edges between points A, B, C and D in Figure 1, for instance when $r = d$), and disappear for higher values of r (when $r = \sqrt{2}d$, the initial hole is covered by triangles ABC and ACD). Which is important in this discussion is that the overall structure of data is the one that *persists* for longer r values. Holes defined by noisy data are rapidly eliminated from the simple complex.

The value of r at which a hole appears, and then the one at which it disappears, defines a bar joining both, which characterizes the hole persistence. When collecting all the bars associated with all the holes appearing and then disappearing when r grows, the so-called persistence barcode results, the last representing compactly a given morphology.

An alternative consists of using a 2D representation, the so-called persistence diagram (PD), reporting in the x_1-axis the value of r at which a hole appears, and on the x_2-axis the value at which it disappears. Obviously, with the hole birth preceding its death, all the point are place on the upper domain defined by the bisector $x_2 = x_1$, and any point (x_1, x_2) remaining close to that bisector represents noise, a small scale, with the associated hole death following immediately its birth. Points far from the bisector represent the topology that persists.

The persistence barcode and the persistence diagram are two representations with a high physical content; however both representations can not be used for comparison purposes, because they are defined in a non-metric space where the calculation of distances for concluding on proximity has not sense.

To move to a more appropriate space making possible the calculation of distances, we first transform the persistence diagram according to $(x_1, x_2) \rightarrow (y_1 = x_1, y_2 = x_2 - x_1)$ and then apply on the last a convolution (usually with a Gaussian kernel) leading to the so-called persistence image (PI) $\mathbf{y}$, the last defined in a vector space, $\mathbf{y} \in \mathbb{R}^D$, that allows applying most of AI algorithms [23].

2.3. Principal Component Analysis

TDA is able to analyze a complex microstructure through its associated image, and to extract its relevant topological features in form of a persistence image, that can be viewed a matrix. However, using these matrix components is not the most compact and concise way of representing the microstructure, because it contains too many components that makes difficult using it for constructing regressions. Thus, in practice, a linear dimensionality reduction such as principal component analysis (PCA) can be applied for extracting the most representative modes of the persistence images and then to represent in a compact and concise way the microstructures by using the weight associated with the most important modes extracted.

Let us consider a vector $\mathbf{y} \in \mathbb{R}^D$ containing the different components of a persistence image. When considering a set of $\mathtt{P}$ microstructures, the associated PIs lead to $\mathbf{y}_i$, $i = 1, \ldots, \mathtt{P}$. If they are somehow correlated, there will be a linear transformation $\mathbf{W}$ defining the vector $\boldsymbol{\xi} \in \mathbb{R}^d$, with $d < D$, which contains the still unknown *latent variables*, such that [4]

$$\mathbf{y} = \mathbf{W}\boldsymbol{\xi}. \tag{14}$$

The transformation matrix $\mathbf{W}$, $D \times d$, satisfies the orthogonality condition $\mathbf{W}^T\mathbf{W} = \mathbf{I}_d$, where $\mathbf{I}_d$ represents the $d \times d$ identity matrix.

PCA proceeds by guaranteeing maximal preserved variance and de-correlation in the latent variable set $\boldsymbol{\xi}$. Thus, the covariance matrix of $\boldsymbol{\xi}$,

$$\mathbf{C}_{\xi\xi} = \mathrm{E}\{\boldsymbol{\Xi}\boldsymbol{\Xi}^T\}, \tag{15}$$

will be diagonal. PCA will then extract the d uncorrelated latent variables from

$$\mathbf{C}_{yy} = \mathrm{E}\{\mathbf{Y}\mathbf{Y}^T\} = \mathrm{E}\{\mathbf{W}\boldsymbol{\Xi}\boldsymbol{\Xi}^T\mathbf{W}^T\} = \mathbf{W}\mathrm{E}\{\boldsymbol{\Xi}\boldsymbol{\Xi}^T\}\mathbf{W}^T = \mathbf{W}\mathbf{C}_{\xi\xi}\mathbf{W}^T, \tag{16}$$

that pre- and post-multiplying by $\mathbf{W}^T$ and $\mathbf{W}$, respectively, reads

$$\mathbf{C}_{\xi\xi} = \mathbf{W}^T\mathbf{C}_{yy}\mathbf{W}. \tag{17}$$

By factorizing the covariance matrix $\mathbf{C}_{yy}$, applying the singular value decomposition, SVD,

$$\mathbf{C}_{yy} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T, \tag{18}$$

and taking into account Equation (17), it results

$$\mathbf{C}_{\xi\xi} = \mathbf{W}^T\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\mathbf{W}, \tag{19}$$

that holds when the d columns of $\mathbf{W}$ are taken collinear with d columns of $\mathbf{V}$, i.e.,

$$\mathbf{W} = \mathbf{V}\mathbf{I}_{D\times d}. \tag{20}$$

2.4. Code2Vect

Code2Vect [13] maps data into a vector space where the distance between points is proportional to the difference of the QoI associated with those points, as sketched in Figure 2.

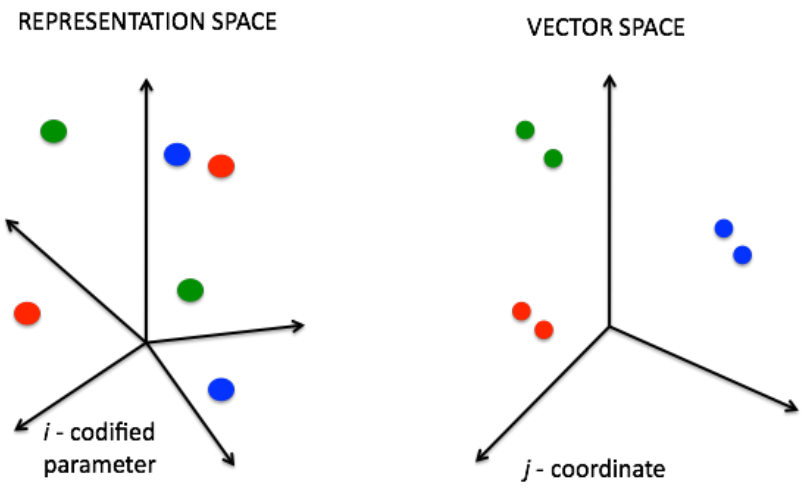

Figure 2. Input space $\boldsymbol{\xi}$ (**left**) and target vector space $\mathbf{z}$ (**right**).

We assume the available data consisting of $\mathtt{P}$ d-dimensional arrays, $\boldsymbol{\xi}_i \in \mathbb{R}^d$, with a QoI $\mathcal{O}_i$ associated with each datum. The images, $\mathbf{z}_i \in \mathbb{R}^q$ ($q = 2$ in our numerical implementation for the sake of visualization clarity), results from

$$\mathbf{z}_i = \mathbf{W}\boldsymbol{\xi}_i, \quad i = 1, \ldots, \mathtt{P}, \tag{21}$$

that preserves the quantity of interest associated with is origin point $\boldsymbol{\xi}_i$, denoted by $\mathcal{O}_i$.

In order to place points such that distances scales with their QoI differences we enforce

$$(\mathbf{W}(\boldsymbol{\xi}_i - \boldsymbol{\xi}_j)) \cdot \mathbf{W}(\boldsymbol{\xi}_i - \boldsymbol{\xi}_j)) = \|\mathbf{z}_i - \mathbf{z}_j\|^2 = |\mathcal{O}_i - \mathcal{O}_j|. \tag{22}$$

Thus, there are $\frac{\mathtt{P}^2}{2} - \mathtt{P}$ relations to determine the $q \times d + \mathtt{P} \times q$ unknowns. Linear mappings are limited and do not allow proceeding in nonlinear settings. Thus, a better choice consists of a nonlinear mapping $\mathbf{W}(\boldsymbol{\xi})$, expressible as a general polynomial form.

3. Results

3.1. Model Training

Several microstructures, based on a population of holes (from now on called pores) with different sizes, shapes, location, and number of pores, distributed in the 2D square domain Ω, are created. Four of these microstructures are shown in Figure 3. They are equipped with a mesh on which finite element calculations will be done for computing the reference effective (homogenized) thermal conductivity, in particular the component K_{22} of the homogenized conductivity tensor.

These meshes also serve to apply the TDA in order to obtain the persistence diagram (PD) and its associated persistence image (PI). As previously indicated, the last consists of a convolution applied on the former. Each persistence image defines a 20×20 matrix, or its vector counterpart $\mathbf{y}_i \in \mathbb{R}^{400}$.

Thus, TDA is able to analyze a complex microstructure through its image, and extract its relevant topological features in form of a persistence image, that can be viewed as a matrix. However, this matrix still contains too much information (its number of components, here 20×20) to perform classification or regression when not too much data is available (scarce-data limit). Obviously, large amounts of synthetic data can be produced by solving numerically thousands or even millions of thermal

problems. However, in engineering cheap solutions are usually preferred, and in particular smart-data is preferred to its big counterpart. Efficiency seems a better option that brute force, and for this reason, here we prefer keeping the amount of data as reduced as possible, and compensate its absence by enhancing the amount of information that data contains.

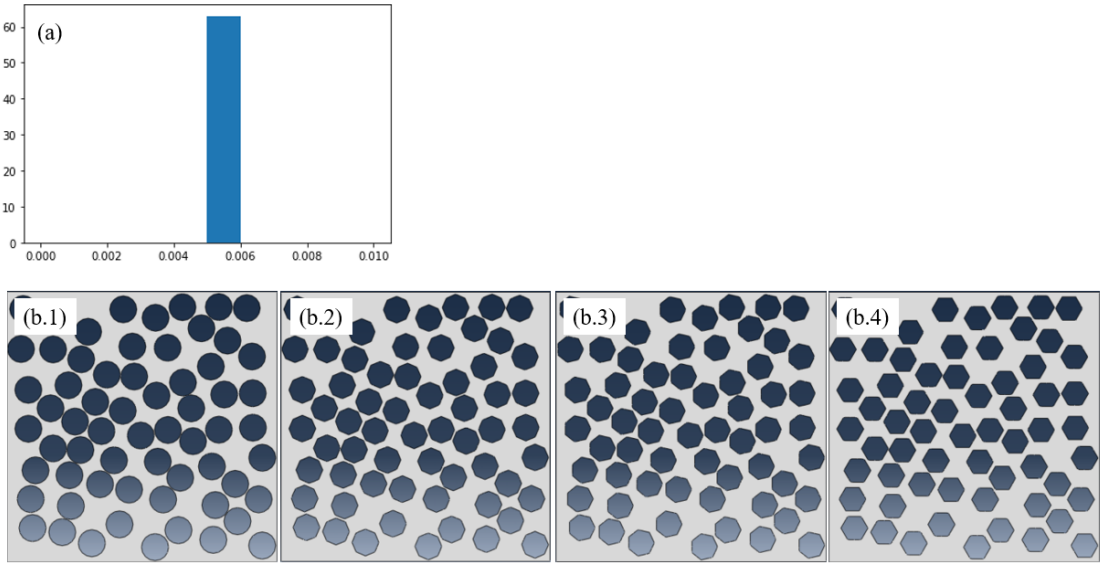

Figure 3. (**a**) Histogram of pores radius; (**b**) Pore shapes: Circle, Octagon, Heptagon and Hexagon.

Thus, persistence images are still not the most compact way of representing the topological and morphological features of the analyzed microstructures. For improving the representation we apply a linear dimensionality reduction, the principal component analysis, for extracting the most representative modes of the persistence images. Thus, the weights of those PCA modes will constitute the compact and concise way to represent those microstructures.

From a practical viewpoint PCA allowed reducing from $400 = (20 \times 20)$ the dimension of PI resulting from TDA, to 3 dimensions. Thus, each analyzed microstructure is concisely represented by 3 coordinates (the weights of the first three most relevant PCA modes) and each one has attached a QoI, the effective thermal conductivity K_{22} obtained from a finite element simulation following the rationale described in Section 2.1. Now, the nonlinear regression relating the output, the QoI (the effective thermal conductivity in our case), with the parameters describing the microstructure, the three PCA weights, is performed by applying the *Code2Vect* nonlinear regression, summarized in Section 2.4.

As soon as the regression is constructed at the present training stage, it could be used online for predicting the conductivity of new microstructures.

3.2. Inferring Effective Properties

We prepared 5 samples, four of them were used in the training stage, represented in Figure 3, in which the pores volume fraction was kept constant ($\phi = 0.5$) and the spatial distribution almost uniform.

The constructed nonlinear regression (based on the use of *Code2Vect*) described in the previous section, is now applied to the sample shown in Figure 4 where while keeping the same almost uniform pore distribution and the same volume fraction, hexagons and heptagons were randomly mixed. In this same figure, the solution of the thermal problem at the microscopic scale for obtaining the effective thermal conductivity that will serve as reference value, is also included. Finally, it also shows both the PD and the PI.

The PI, $\mathbf{y} \in \mathbb{R}^D$, is then projected into the three retained orthonormal PCA modes to give the thee weights that constitute the data $\boldsymbol{\xi} \in \mathbb{R}^3$ ($d = 3$) to be processed by the nonlinear regression (based on

the *Code2Vect*) that produces vector $\mathbf{z} \in \mathbb{R}^2$ (we enforce a 2D representation, $q = 2$, for the sake of clarity in the data visualization)

$$\mathbf{z} = \mathbf{W}(\boldsymbol{\xi})\, \boldsymbol{\xi}, \tag{23}$$

and then identify the set $\mathcal{S}(\mathbf{z})$ of data $\mathbf{z}_i$ closest to $\mathbf{z}$, from which the QoI, the effective thermal conductivity, is interpolated

$$\mathcal{O} = \sum_{i \in \mathcal{S}(\mathbf{z})} \mathcal{F}(\mathbf{z}, \mathbf{z}_i)\, \mathcal{O}_i, \tag{24}$$

with in the present case $\mathcal{O} \equiv K_{22}$ and with radial bases as interpolation functions $\mathcal{F}(\mathbf{z}, \mathbf{z}_i)$.

Figure 5 places $\mathbf{z}$ with respect to its neighbors, where color scales with the target quantity, that is, with K_{22}. The inferred value of the effective thermal conductivity K_{22} using Equation (24) for the micorstructure depcited in Figure 4 results $K_{22}(\mathbf{z}) = 73.4$ W/mK, very close to the reference value computed numerically from the temperature distribution shown also in Figure 4, of $K_{22,\mathrm{REF}} = 74$ W/mK.

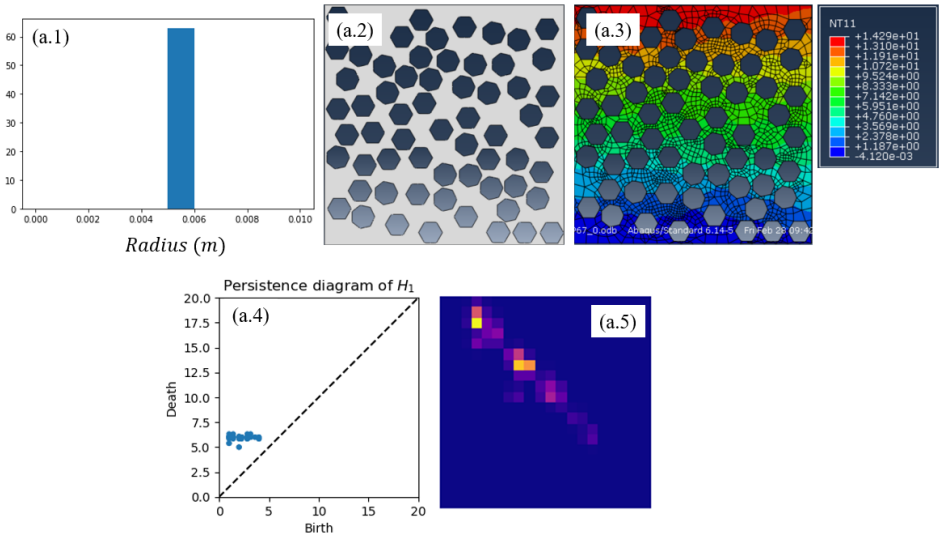

Figure 4. (**a.1**) Histogram of the pores radius; (**a.2**) considered microstructure; (**a.3**) temperature field used for computing the effective thermal conductivity that will serve as reference for evaluating the regression performance; (**a.4**) persistence diagram; and (**a.5**) persistence image.

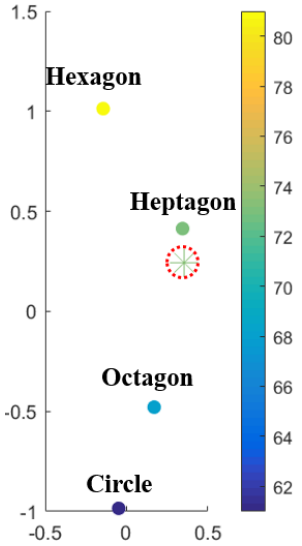

Figure 5. Interpolation space $\mathbf{z}$ with color scaling with the values of the effective thermal conductivity K_{22}.

3.3. Microstructures with Varying Shapes and Size Distribution

These first preliminary successful results were pushed forward by considering quite more complex microstructures. Thus, a total of 35 samples were generated, while varying other parameters, in particular pore size (following uniform and Gamma distributions) and pore shape (circular or 5 to 8 side polygons, randomly chosen). The volume fraction was kept constant ($\phi = 0.5$). 34 samples were used in the training, keeping one, the one shown in Figure 6, for inferring the thermal conductivity and concluding on the ability of the proposed technique to infer accurately it. Figure 7 places the considered microstructure in the **z**-space where the thermal conductivity is interpolated, to infer the value of $K_{22} = 81$ W/mK, for a reference value of $K_{22,\mathrm{REF}} = 78$ W/mK.

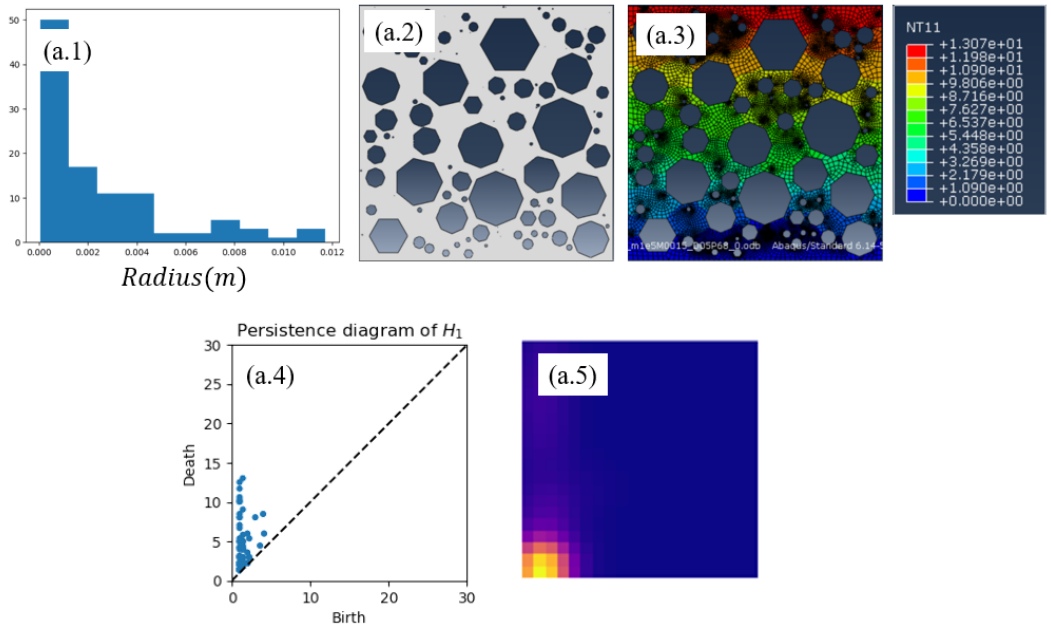

Figure 6. (**a.1**) Histogram of pores radius; (**a.2**) testing microstructure; (**a.3**) temperature field used for calculating the reference effective thermal conductivity; (**a.4**) persistence diagram; and (**a.5**) persistence image.

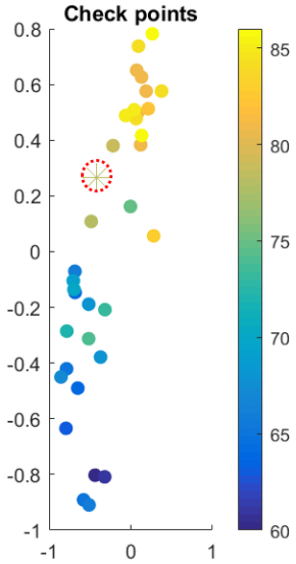

Figure 7. Interpolation space **z** with color scaling with the values of the effective thermal conductivity K_{22}.

To check the prediction improvement with the sampling richness, the effective thermal conductivity in the microstructure shown in Figure 6 while considering different samplings in the training stage, from 13 to 35 microstructures, with the relative errors reported in Table 1.

Table 1. Relative error in the effective conductivity prediction depending on the number of samples considered in the regression (training stage).

Number of Samples	Relative Error
13	0.076
16	0.056
19	0.046
35	0.037

4. Conclusions

The present study proves that effective properties can be associated with microstructures with complex morphological and topological features. For this purpose, those features are extracted by using TDA, post-compressed by using linear dimensionality reduction (PCA) which output represents the parameters employed by the nonlinear *Code2Vect* regression that finally assign a effective property (here the effective thermal conductivity) to a given microstructure.

The procedure demonstrated its robustness and performance in the low-data limit, as well as its capacity to provide better predictions when considering larger training sets. It successfully combines physics-based data for learning purposes, with almost real-time inference based on the topological analysis of images.

Author Contributions: Conceptualization, software and validation, M.Y. and C.A.; methodology, J.L.D., E.C. and F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors acknowledge the ESI Chairs at Arts et Métiers Institute of Technology and the University of Zaragoza, as well as the French ANR through the DataBEST project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chinesta, F.; Huerta, A.; Rozza, G.; Willcox, K. *Model Order Reduction Chapter in the Encyclopedia of Computational Mechanics*, 2nd ed.; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2015.
2. Chinesta, F.; Leygue, A.; Bordeu, F.; Aguado, J.V.; Cueto, E.; González, D.; Alfaro, I.; Ammar, A.; Huerta, A. Parametric PGD based computational vademecum for efficient design, optimization and control. *Arch. Comput. Methods Eng.* **2013**, *20*, 31–59. [CrossRef]
3. Chinesta, F.; Keunings, R.; Leygue, A. *The Proper Generalized Decomposition for Advanced Numerical Simulations*; Springerbriefs, Springer: Berlin, Germany, 2014.
4. Lee, J.A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer: New York, NY, USA, 2007.
5. González, D.; Chinesta, F.; Cueto, E. Thermodynamically consistent data-driven computational mechanics. Continuum Mech. *Thermodynamics* **2018**, *31*, 239–253. [CrossRef]
6. González, D.; Cueto, E.; Chinesta, F. Computational patient avatars for surgery planning. *Ann. Biomed. Eng.* **2016**, *44*, 35–45. [CrossRef] [PubMed]
7. Lopez, E.; Gonzalez, D.; Aguado, J.V.; Abisset-Chavanne, E.; Cueto, E.; Binetruy, C.; Chinesta, F. A manifold learning approach for integrated computational materials engineering. *Arch. Comput. Methods Eng.* **2016**, *25*, 59–68. [CrossRef]
8. Gonzalez, D.; Aguado, J.V.; Cueto, E.; Abisset-Chavanne, E.; Chinesta, F. kPCA-based Parametric Solutions within the PGD Framework. *Arch. Comput. Methods Eng.* **2018**, *25*, 69–86. [CrossRef]
9. Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

10. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines: And other Kernel-Based Learning Methods*; Cambridge University Press: New York, NY, USA, 2000.
11. Forgy, E. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics* **1965**, *21*, 768–769.
12. Ibanez, R.; Abisset-Chavanne, E.; Ammar, A.; Gonzalez, D.; Cueto, E.; Huerta, A.; Duval, J.L.; Chinesta, F. A multi-dimensional data-driven sparse identification technique: The sparse Proper Generalized Decomposition. *Complexity* **2018**, 5608286. [CrossRef]
13. Argerich, C.; Ibanez, R.; Barasinski, A.; Chinesta, F. Code2vect: An efficient heterogenous data classifier and nonlinear regression technique. *Comptes Rendus Mécanique* **2019**, *347*, 754–761. [CrossRef]
14. Reille, A.; Hascoet, N.; Ghnatios, C.; Ammar, A.; Cueto, E.; Duval, J.L.; Chinesta, F.; Keunings, R. Incremental dynamic mode decomposition: A reduced-model learner operating at the low-data limit. *Comptes Rendus Mécanique* **2019**, *347*, 780–792. [CrossRef]
15. Schmid, P.J. Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.* **2010**, *656*, 5–28. [CrossRef]
16. Williams, M.O.; Kevrekidis, G.; Rowley, C.W. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *J. Nonlinear Sci.* **2015**, *25*, 1307–1346. [CrossRef]
17. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, UK, 2016
18. Moya, B.; Gonzalez, D.; Alfaro, I.; Chinesta, F.; Cueto, E. Learning slosh dynamics by means of data. *Comput. Mech.* **2019**, *64*, 511–523. [CrossRef]
19. Chinesta, F.; Cueto, E.; Abisset-Chavanne, E.; Duval, J.L.; El Khaldi, F. Virtual, Digital and Hybrid Twins: A New Paradigm in Data-Based Engineering and Engineered Data. *Arch. Comput. Methods Eng.* **2020**, *27*, 105–134. [CrossRef]
20. González, D.; Chinesta, F.; Cueto, E. Learning corrections for hyperelastic models from data. *Front. Mater. Comput. Mater. Sci.* **2019**, *6*. [CrossRef]
21. Wasserman, L. Topological data analysis. *Ann. Rev. Stat. Appl.* **2018**, *5*, 501–532. [CrossRef]
22. Lamari, H.; Ammar, A.; Cartraud, P.; Legrain, G.; Jacquemin, F.; Chinesta, F. Routes for Efficient Computational Homogenization of Non-Linear Materials Using the Proper Generalized Decomposition. *Arch. Comput. Methods Eng.* **2010**, *17*, 373–391. [CrossRef]
23. Adams, H.; Emerson, T.; Kirby, M.; Neville, R.; Peterson, C.; Shipman, P.; Chepushtanova, S.; Hanson, E.; Motta, F.; Ziegelmeier, L. Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.* **2017**, *18*, 218–252.