R-VINE COPULAS FOR DATA-DRIVEN QUANTIFICATION OF DESCRIPTOR RELATIONSHIPS IN POROUS MATERIALS

MATTHIAS NEUMANN[1,⋆], PHILLIP GRÄFENSTEINER[1,⋆], EDUARDO MACHADO CHARRY[2,3], ULRICH HIRN[3,4], ANDRÉ HILGER[5], INGO MANKE[5], ROBERT SCHENNACH[2,3], VOLKER SCHMIDT[1], KARIN ZOJER[2,3]

[1]*Institute of Stochastics, Ulm University, Helmholtzstraße 18, 89069 Ulm, Germany*

[2]*Institute of Solid State Physics, NAWI Graz, Graz University of Technology, Petersgasse 16/II, 8010 Graz, Austria*

[3]*Christian Doppler Laboratory for Mass Transport through Paper, Graz University of Technology, Petersgasse 16/II, 8010 Graz, Austria*

[4]*Institute of Bioproducts and Paper Technology, Graz University of Technology, Inffeldgasse 23, 8010 Graz, Austria*

[5]*Institute of Applied Materials, Helmholtz-Zentrum Berlin für Materialien und Energie, Hahn-Meitner-Platz 1, 14109 Berlin, Germany*

⋆*M.N. and P.G. contributed equally to this work*

Abstract. Local variations in the 3D microstructure can control the macroscopic behavior of heterogeneous porous materials. For example, the permittivity through porous sheets or membranes is governed by local high-volume pathways or bottlenecks. Due to local variations, unfeasibly large amounts of microstructure data would be needed to reliably predict such material properties directly from image data. Here it is demonstrated that a vine copula approach provides parametric models for local microstructure descriptors that compactly capture the 3D microstructure including its local variations and efficiently probe it with respect to selected, measurable properties. In contrast to common methods of complexity reduction, the proposed approach creates parametric models for the multivariate probability distribution of high-dimensional descriptor vectors that inherently contain the complex, nonlinear dependencies between these descriptors. Therein, material properties are offered in physically motivated distributions of microstructure descriptors rather than as normally distributed data. Applied to porous fiber networks (paper) before and after unidirectional compression, it is shown that the copula-based models reveal material-characteristic relationships between two and more microstructure descriptors. In this way, the presented modeling approach can provide deeper insight into the microscopic origin of effective macroscopic properties of heterogeneous porous materials.

Key words and phrases. porous material, local heterogeneity, vine copula, 3D microstructure, multivariate statistical data analysis, complexity reduction.

1. Introduction

Although heterogeneous porous materials surround us in daily life, be them membranes, fiber-reinforced materials, wood, concrete, or paper, it is still a challenge to predict their macroscopic behavior from the microscopic structure. In contrast to ordered or homogeneous materials, such materials exhibit pronounced local variations in their microstructure, so that structure-driven material properties depend on the actual location in the material. Consequently, also any relationship between these properties depends on their local spatial correlation. Reliable predictions are therefore only possible if models are informed about the limits in which the underlying local microstructure descriptors vary over the entire sample and how likely certain local realizations of relevant descriptors, so-called configurations, occur. If it is known how often, *i.e.*, how likely, each permissible configuration occurs in the material, it is possible to quantify how these descriptors vary across the material due to the heterogeneous structure. In mathematical terms, the likelihood of encountering a given configuration of microstructure descriptors can be computed from the joint multivariate probability distribution of these descriptors.

When focusing on the 3D morphology of the microstructure, multiple geometric descriptors are needed to capture the complex microstructure and to distinguish it from the structure of other materials. If the microstructure is known, *e.g.*, from tomographic imaging of the material, possible configurations and, consequently, the spatial variations and relationships between the geometric descriptors are contained in the 3D image data. However, since such data sets are typically very large, the key challenge is to cast the information contained in the microstructure into a much more compact form, *i.e.*, to provide and compactly store the parameters of high-dimensional multivariate distributions.

In this article, we demonstrate the use of R-vine copulas to build and apply a compact, parametric model for the multivariate distribution of microstructure descriptors. Such vine copulas represent the distribution contained in the 3D image data more flexibly and more precisely than conventional multivariate distribution models. Unlike the latter, vine copulas reliably predict highly likely as well as rare configurations [1] and inherently describe nonlinear relations between various descriptors of the 3D microstructure. This flexibility of R-vine copulas is rooted in their construction: One assembles simple and easy-to-interpret building blocks. The first building blocks are the univariate distributions of each descriptor. Then, the univariate distributions are coupled in pairs with one or two additional model parameters using bivariate copula functions. These model parameters directly provide the relationships between pairs of microstructure descriptors. Since R-vine copulas are not routinely used for designing multivariate distribution models to capture heterogeneous materials, we illustrate the conceptual steps necessary to construct such a distribution from 3D image data considered in the present article.

With this compact model of the multivariate distribution at our disposal, we will (i) showcase how to extract relationships between pairs or even triples of microstructure descriptors, (ii) show that we are capable of finding relations that are not revealed with other methods, and (iii) show that these relationships are characteristic for our heterogeneous materials.

Paper sheets serve here as a heterogeneous porous model system to instructively explain how to construct and use such multivariate probabilistic models. The pore space in paper sheets, that is accessible to 3D imaging techniques such as microcomputed X-ray tomography [2, 3, 4] or FIB-SEM,[5, 6] is associated to pores formed between fibers. When a paper sheet is formed, the number and diameters of the locally involved fibers may profoundly vary [7]. This variation

results in a pronounced heterogeneity of the pore space, especially in the plane of the sheets. Intriguing microstructure descriptors of the pore space with foreseeable local variations are the porosity ε (*i.e.,* the volume fraction of the pore space), the sheet thickness δ, and the specific surface area S_V (*i.e.,* the surface area per unit volume) [4, 8, 9, 10, 11]. For the investigation of transport phenomena, also the lengths of transportation pathways through the sheet are crucial. To provide such lengths, we consider here τ_0, the mean geodesic tortuosity of all pathways, and τ_3, the mean geodesic tortuosity of pathways with a minimum radius of 3 μm, to account for high-volume pathways [12].

Thanks to two deliberately chosen paper grades, markedly different realizations of the microstructure descriptors can be offered for modelling (Figure 1): A new, second paper grade was formed from an original paper grade by compressing the sheets unidirectionally in the thickness direction without allowing the fibers to change their in-plane orientation [13]. In the original paper, there are pronounced variations in the local thickness, as the number of stacked fibers laterally varies from place to place (*uncompressed*, Figure 1a,c). During the compression process, the original paper was converted to a paper with only slight local thickness variations, but marked lateral variations in the space between stacked fibers (*compressed*, Figure 1b,d). While compression predominantly reduces the space between stacked fibers in previously thicker regions, previously thinner regions remain practically unchanged. Further structural changes, such as significantly deformed fiber cross sections or even densification of the fiber walls, are not expected. Since the two paper grades still share the same fiber type and in-plane distribution of fibers (i.e., same basis weight), it possible to interpret model-predicted changes in relations between descriptor pairs as a result of compression. The microstructure dataset associated to each paper grade stems from X-ray computed microtomography scans is large enough to ensure that local variations in the descriptors are fully captured [13].

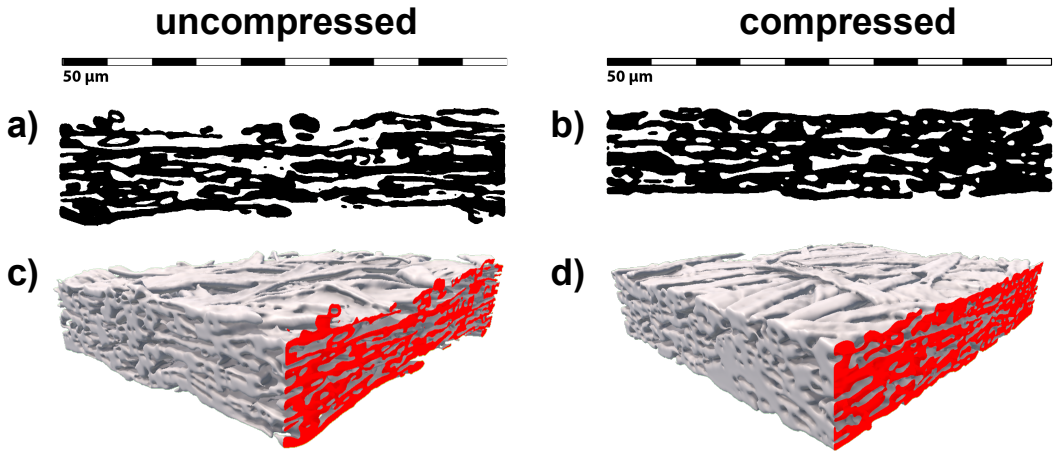

FIGURE 1. Microstructure of a local cutout from uncompressed (a,c), and compressed paper (b,d) obtained with X-ray computed microtomography [13]. Shown is the fiber material in a cross section of the cutout (a,b) and within the 3D-rendered volume (c,d). This visualization only shows a small region of the available image data. Courtesy of E. Baikova.

In a first use case, our model formulated in terms of these five microstructure descriptors directly reveals how strongly descriptor pairs are cross-related in each paper grade. As a second instructive use case, we will combine our knowledge on porosity ε and specific surface area S_V

to unravel the relation of the latter to the more routinely quoted surface area per unit mass. As a final model application involving three or even four microstructure descriptors, we show how the conventional Bruggemann-relation between porosity and tortuosity can be material-specifically modified to reliably predict mean geodesic tortuosities τ_0 and τ_3 from knowing local porosity and local thickness. Such a relation is a highly desirable to ease the prediction of structure-determined transport properties such as local air permeances [12].

2. Capturing local variations and correlations with R-vine copulas

The five local microstructure descriptors considered in the present study are modeled by a random vector $W = (W_1, \ldots, W_5) = (\varepsilon, \delta, S_V, \tau_0, \tau_3)$: A realization $w = (w_1, \ldots, w_5) \in \mathbb{R}^5$ of W can be interpreted as the vector of descriptors evaluated for a predefined local inspection region. In principle, the multivariate distribution of the random vector W is readily sampled by collecting the configurations of $(\varepsilon, \delta, S_V, \tau_0, \tau_3)$ in non-overlapping inspection regions of a given size. This yields a sampling-dependent point cloud in the five-dimensional descriptor space. Based on such a point cloud, we calibrate the probabilistic model. In other words, we seek for a multivariate probability density function $f\colon \mathbb{R}^5 \to [0, \infty)$ of $(\varepsilon, \delta, S_V, \tau_0, \tau_3)$ which describes the observed data appropriately and, in particular, reflects interdependencies between pairs of descriptors accurately. While it is possible to model multivariate probability density functions directly, for example by kernel density estimation, such an approach does not allow for adjusting the final model and makes inter- or extrapolation to different scenarios impossible.

We therefore choose to employ a parametric approach. By using so-called R-vine copulas, parametric univariate and bivariate distributions can be combined to construct the multivariate probability density function of a random vector in arbitrary dimension, where no limiting assumptions on the shape of the marginal distributions are necessary.

In this section, we explain the theory underlying this approach at the example of the five local descriptors of paper sheets mentioned above, and demonstrate the increased flexibility and accuracy that this type of modeling approach brings compared to classical techniques, such as multivariate normal distributions or product densities.

2.1. **Univariate distributions.**

As a first step, we analyze the univariate distribution of each local descriptor individually. That is, we are interested in modeling the univariate probability density functions $f_i\colon \mathbb{R} \to [0, \infty)$, $i = 1, \ldots, 5$, which are the density functions of the components W_i of $W = (W_1, \ldots, W_5)$. Each of these density functions is determined by the frequency of encountering values of the corresponding descriptor across the sample. From these observed frequencies, an empirical density function is determined by kernel density estimation [14]. In order to obtain a parametric model, these empirical densities are approximated by an appropriate parametric probability density function with correspondingly adjusted parameters. A candidate probability density function is found appropriate if (i) it approximates the empirical density well and (ii) the function accounts for the descriptor distributions regardless of the size of the inspection region [15]. Furthermore, it is desirable, albeit not necessary, that (iii) the support of the distribution is consistent with the nature of the descriptor. The last aspect often excludes the classical Gaussian density function of the normal distribution, as its support is the whole real line, and it cannot accurately represent distributions that are skewed to one side. By employing criteria (i)–(iii) explained above, we follow a data-driven approach in modeling the univariate distributions of the five local descriptors. To find the best fitting model density

functions, we did not impose physically motivated density functions. For a general overview on candidate density functions the reader is referred to [16].

Figure 2 shows that the empirical density functions (filled symbols) are accurately approximated with their respective fitted parametric density functions (open symbols) in all cases. Depending on the descriptor, we either use beta, generalized gamma, Weibull, or Rician distributions. Their parameters are fitted via maximum likelihood estimation [17], where the procedure for beta and shifted gamma distributions follows [15] and the one for the Rician distribution follows [18].

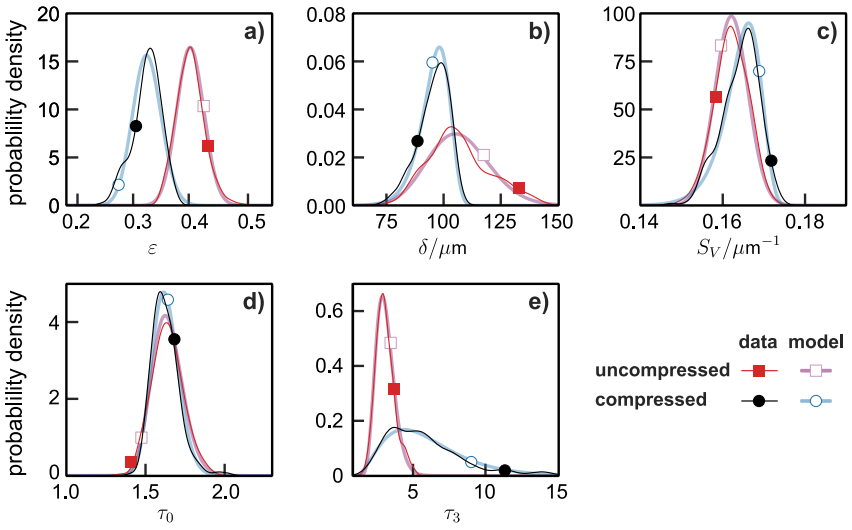

FIGURE 2. Univariate density functions of microstructure descriptors porosity ε (a), sheet thickness δ (b), specific surface area S_V (c), as well as the mean geodesic tortuosities τ_0 (d), and τ_3 (e) for the uncompressed and the compressed paper sample. The densities of microstructure descriptors computed from 3D image data via kernel density estimation (filled symbols) are compared with the corresponding parametric model densities (open symbols).

For the local porosity ε, beta distributions are highly suitable (Figure 2a), because their support is $[0, 1]$ and thus matches exactly the range of possible porosity values. Path-length-related descriptors τ_0 and τ_3 are modelled with shifted generalized gamma distributions (Figure 2d,e). Shifting the gamma distribution ensures that its support does not exceed the range of possible tortuosity values, because $\tau_0, \tau_3 \geqslant 1$. Moreover, gamma distributions readily account for skewed distributions, be that for the tortuosities τ_0, τ_3 (Figure 2d,e) or for the local thickness δ in the uncompressed paper (Figure 2b). The Weibull distribution models the local thickness δ (Figure 2b) and the local specific surface area S_V (Figure 2c) of compressed paper sheets. Note that the Weibull distribution nicely reproduces the negative skewness, that compression induces in the distributions of δ and S_V. As already discussed in [13] for local thicknesses, a negative skewness indicates that there are less outliers towards high values of δ and S_V. The specific surface area S_V of uncompressed paper sheets (Figure 2c) is modeled by a Rician distribution. Though a fit to a normal distribution has a comparable quality, a Rician distribution is preferred, since its support is the positive half axis, *i.e.,* only positive S_V-values are permitted. The expressions for the probability density functions used in this article, their adjustable parameters, and their support are provided in the Supporting Material.

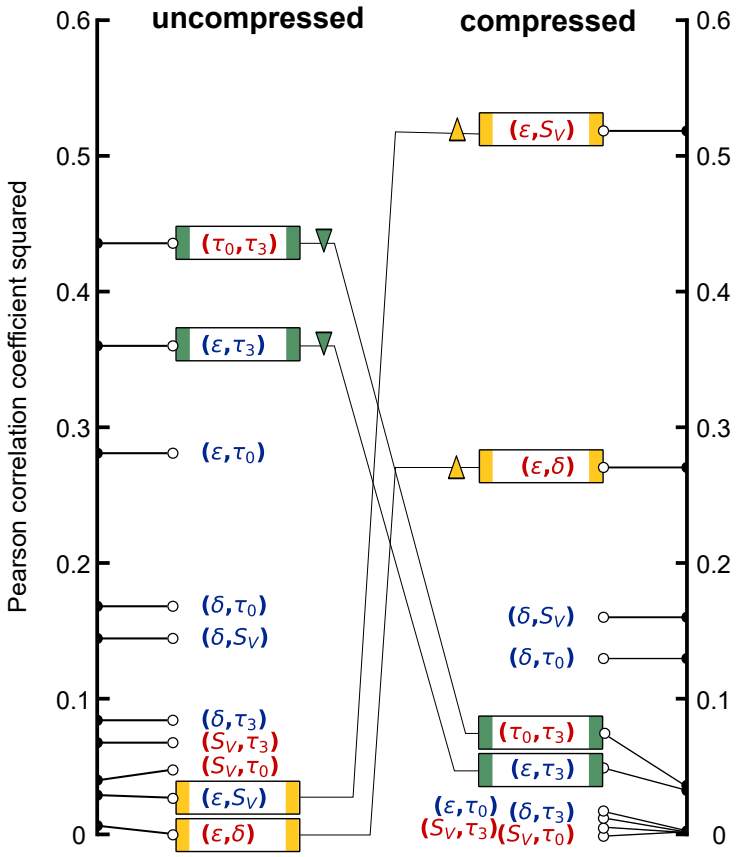

FIGURE 3. Pairwise Pearson correlation coefficients squared for porosity (ε), sheet thickness (δ), specific surface area (S_V), and the two path-length related descriptors τ_0 and τ_3 for the uncompressed (left) and compressed (right) paper sheets. Labels in red indicate positive correlations, labels in blue negative correlations. To guide the eye through changes in correlations upon compression, some property pairs are highlighted with a box. Green bars (down triangles) indicate pairs, whose relations weaken after compression; yellow bars (up triangles) indicate pairs with increased dependence after compression.

2.2. **Multivariate distributions and copulas.** Recall that we consider the five descriptors ε, δ, S_V, τ_0, and τ_3 as components of a random vector $W = (W_1, \dots, W_5)$. As we ultimately want to model the joint distribution of the random vector W, we are interested in the multivariate probability density function $f\colon \mathbb{R}^5 \to [0, \infty)$ of W. Since we already have a parametric representation of the univariate density functions f_i, $i = 1, \dots, 5$, the simplest approach would be to model f as the density of the product measure, *i.e.*,

$$f(w_1, \dots, w_5) = f_1(w_1) \dots f_5(w_5), \qquad \text{for all } w_1, \dots, w_5 \in \mathbb{R}. \tag{1}$$

However, as this approach ignores all interactions between the descriptors, this model of f only leads to reasonable results if all five descriptors are independent. In Figure 3, the values of the squared empirical Pearson correlation coefficient [19] are visualized for all descriptor pairs for the compressed and uncompressed paper sheets. As these values are non-zero and most of them significantly differ from zero, the five local descriptors are not independent and therefore influence each other. These dependencies, which we model separately, have to be taken into account in order to obtain an accurate model for the joint distribution of the random vector

W. For a more detailed discussion on the effect of compression on the correlation between descriptor pairs, see Section 3.1.

In the fortunate case of a multivariate normal distribution, it is enough to incorporate these pair-wise correlation coefficients in the covariance matrix in order to capture the dependencies between the individual descriptors [9, 20]. However, as shown in Section 2.1, the marginal distributions of the descriptors cannot be approximated well by normal distributions. We therefore require a more advanced approach for modeling the interdependencies between the individual descriptors, which is precisely what so-called copulas can be used for.

Formally, for any fixed integer $n \geqslant 2$, where $n = 5$ in our case, a function $C\colon [0,1]^n \to [0,1]$ is called a copula if it is the multivariate probability distribution function of a random vector $U = (U_1, \ldots, U_n)$ for which the marginal distributions of all components U_i, $i = 1, \ldots, n$, are uniform on $[0, 1]$.

The merit of this concept is that it allows to decompose the information carried by the multivariate distribution of any n-dimensional random vector $W = (W_1, \ldots, W_n)$ in two separate parts. One part is the information about the univariate marginal distributions, which is carried by the univariate probability density functions or, equivalently, the corresponding cumulative distribution functions of $W_1, \ldots, W_n$. The second part is the information about the correlation of the random variables $W_1, \ldots, W_n$. In the context of multivariate normal distributions, such a decomposition is available by combining univariate normal distributions with a suitable covariance matrix. This strategy becomes available for all types of marginal distributions and correlation structures when copulas are used. This idea is formalized through Sklar's theorem (see, *e.g.*, Theorem 1.1 in [21]), which states that for any cumulative distribution function $F\colon \mathbb{R}^n \to [0,1]$ of an n-dimensional random vector there exists a copula C such that F can be written in terms of its marginal distribution functions F_i, $i = 1, \ldots, n$, and C as

$$F(w_1, \ldots, w_n) = C\big(F_1(w_1), \ldots, F_n(w_n)\big), \qquad \text{for all } w_1, \ldots, w_n \in \mathbb{R}. \tag{2}$$

In this representation, the copula C models the correlation structure between the components $W_1, \ldots, W_n$ of W, while information about the marginal distributions themselves is captured in the respective univariate distribution functions F_i, $i = 1, \ldots, n$. This splits the problem of modeling a multivariate distribution into the two subproblems of modeling the distributions of the individual components, and modeling their correlation structure separately. Moreover, if the distribution of the random vector W is absolutely continuous and, in particular, if its multivariate distribution function F is differentiable, the multivariate density function f corresponding to F can be expressed via the density function $c\colon [0,1]^n \to [0, \infty)$ of the copula C, together with the marginal densities f_i, of the components W_i, $i = 1, \ldots, n$, by

$$f(w_1, \ldots, w_n) = c\big(F_1(w_1), \ldots, F_n(w_n)\big) \prod_{i=1}^{n} f_i(w_i), \qquad \text{for all } w_1, \ldots, w_n \in \mathbb{R}. \tag{3}$$

In general, this is the product of an n-variate density function c and n univariate density functions $f_1, \ldots, f_n$. However, the density c in Equation (3) can be decomposed further such that the multivariate density f is represented by a product of univariate and bivariate density functions only.

In brief, the necessary steps involve (i) applying the chain rule for conditional density functions and (ii) recursively expressing each conditional density by some (conditional) bivariate densities $\widetilde{f}_{ij}\colon \mathbb{R}^2 \to [0, \infty)$ for $i, j \in \{1, \ldots, n\}$ with $i \neq j$. To avoid complicated notation, the

tilde symbol is used to emphasize that these probability densities and distribution functions $\widetilde{f}_{ij}$ can be either unconditional bivariate ones or conditional ones with respect to one or more other components of $W = (W_1, \ldots, W_n)$.. A detailed explanation using more involved notation be found in the Supplementary Material.

Step (iii) decomposes each bivariate density $\widetilde{f}_{ij}$ (regardless whether $\widetilde{f}_{ij}$ is conditional or not) using a bivariate copula density $\widetilde{c}_{ij}$ as in Equation (3) by

$$\widetilde{f}_{ij}(w_i, w_j) = \widetilde{c}_{ij}\big(\widetilde{F}_i(w_i), \widetilde{F}_j(w_j)\big)\ \widetilde{f}_i(w_i)\ \widetilde{f}_j(w_j), \qquad \text{for all } w_i, w_j \in \mathbb{R}, \tag{4}$$

where $\widetilde{f}_i$ is some (conditional) univariate density with the corresponding cumulative distribution function $\widetilde{F}_i$, $i \in \{1, \ldots, n\}$, and the copula density $\widetilde{c}_{ij}$ is the partial derivative of some (conditional) bivariate copula function $\widetilde{C}_{ij} \colon \mathbb{R}^2 \to [0, 1]$, given by

$$\widetilde{c}_{ij}(u_i, u_j) = \frac{\partial^2 \widetilde{C}_{ij}(u_i, u_j)}{\partial u_i \partial u_j} \qquad \text{for all } u_i, u_j \in [0, 1]. \tag{5}$$

Ultimately, the multivariate density f can be decomposed into factors of the form given in Equation (4), which implies that we only need to model (conditional) joint distributions of descriptor pairs (W_i, W_j) to capture the full correlation structure as in Equation (3). For this, many parametric families of bivariate copula densities are readily available in order to achieve a good fit to the bivariate densities, see the Supplementary Material for further details.

2.3. **R-vine tree representations.** To find all univariate and bivariate densities that ultimately enter the decomposition in Equation (3), it is not necessary to step through the formal decomposition (being provided in the Supporting Material). Instead, an appropriate structure for the decomposition can be determined using a graph representation with trees. So-called regular vine (R-vine) trees give rise to a pair-copula decomposition of the multivariate density [22]. A specific choice of bivariate copulas for every pair in this decomposition is then called an R-vine copula.

We illustrate the graph representation for the case of three descriptors (*i.e.,* considering $W = (W_1, W_2, W_3)$) added to the R-vine tree one by one, see Figure 4. If we only consider one descriptor W_1, then f is equal to f_1, see Figure 4a. In Figure 4b, two descriptors W_1 and W_2 are considered, which adds the univariate density f_2 (circle) and a third factor $c_{1,2}$ (square, connected by a dashed line to the edge of the graph above). The latter factor $c_{1,2}$ is a bivariate copula density and accounts for the correlation between W_1 and W_2 in the spirit of Equation (4). Figure 4c extends this scheme to incorporate a third descriptor W_3. The new factors are the univariate density f_3 (circle), the bivariate copula density $c_{2,3}$ (square), and the conditional copula density $c_{1,3;2}$ (diamond), which models the dependency between descriptors W_1 and W_3 given that $W_2 = w_2$ for some $w_2 \in \mathbb{R}$, see Equation (3.23) in [22]. In the spirit of $c_{1,3;2}$, we will list all descriptors, to whose values bivariate copula densities are conditioned, as indices after the semicolon in $c_{\cdot;\cdot}$. In general, this conditional copula density $c_{1,3;2}$ depends on the value $w_2 \in \mathbb{R}$ of W_2.

In the following we assume the so-called simplifying assumption, *i.e.*, that copulas of conditional distributions do not depend on the value we are conditioning on. This is a commonly used simplification step that enormously reduces the complexity of the model. Without this step, the number of parameters for the final five-dimensional density would become computationally infeasible. Under this assumption, the copula density $c_{13;2}(\,\cdot\,, \cdot\,; w_2)$ does not depend

on the particular choice of w_2. This allows us to model the joint distribution of W_1, W_2, W_3 by means of three bivariate copulas and the three marginal distributions. We will see later that the distribution of our data is modeled reasonably well under this assumption; for a more detailed discussion on the simplifying assumption, see [23].

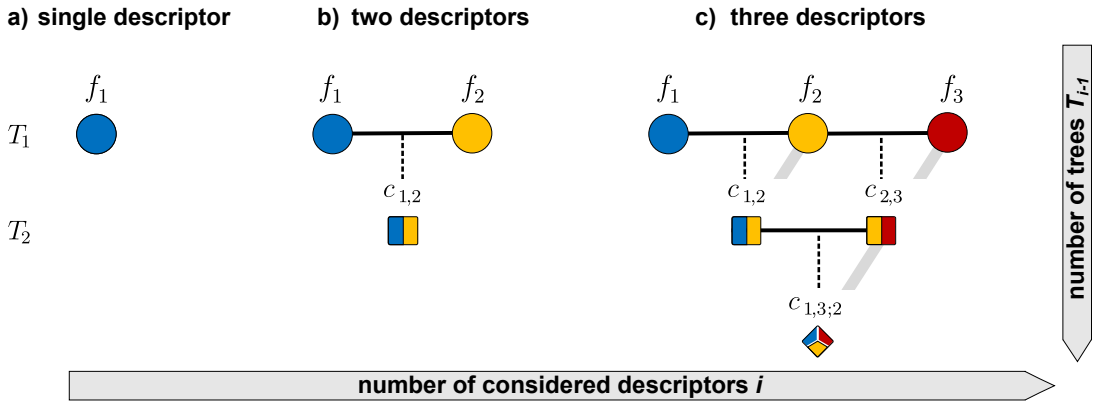

FIGURE 4. Pair copula decomposition of f represented by graphs for the example of one (a), two (b) and three correlated descriptors (c). Each vertex in a graph corresponds to a density function that enters the vine copula of f as a factor. Edges indicate a dependence between two descriptors. As a guide for the eye, all factors associated to a new descriptor added are connected with a diagonal, grey bar. Each new descriptor i extends the scheme to the right via factors (vertices) aligned along such a diagonal and adds a new tree T_{i-1} if $i \geqslant 2$.

All factors occurring in this decomposition are arranged in graphs T_m for $m = 1, 2$. Vertices in graph T_1 visualize factors depending on one descriptor and correspond to univariate densities. Edges (solid lines in Figure 4a and 4b) mark the dependencies between the variables of the connected vertices and indicate a bivariate copula (dashed line). The densities $c_{1,2}$ and $c_{2,3}$ of these bivariate copulas enter graph T_2 as a vertex if W_1 and W_3 are connected by an edge in T_2 (Figure 4c).

Collecting all factors generated with the graphs T_1 and T_2, the density f of $W = (W_1, W_2, W_3)$ is given by

$$\begin{aligned} f(w_1, w_2, w_3) = {} & c_{1,3;2}\big(F_{1;2}(w_1; w_2), F_{3;2}(w_3; w_2)\big) \\ & \times c_{2,3}(F_2(w_2), F_3(w_3))\, c_{1,2}\big(F_1(w_1), F_2(w_2)\big)\, f_3(w_3) f_2(w_2) f_1(w_1), \end{aligned} \tag{6}$$

for all $w_1, w_2, w_3 \in \mathbb{R}$, where $c_{ij} \colon [0,1]^2 \to \mathbb{R}$ denotes the bivariate copula of W_i and W_j, $1 \leqslant i < j \leqslant 3$, $c_{1,3;2}$ is the conditional copula density of W_1 and W_3 given the value of W_2 under the simplifying assumption, and $F_{i;2}(\cdot; w_2)$ is the cumulative distribution function of W_i given that $W_2 = w_2$ for $i = 1, 3$, see also Example 4.1 in [22]. In the general case of n random descriptors for some fixed integer $n \geqslant 2$, a particular decomposition is represented as a sequence of $n - 1$ trees $T_1, \ldots, T_{n-1}$.

2.4. **Resulting R-vine copula structure.** The representation of the joint density of a random vector through R-vine trees is not unique. For instance, the descriptor indices $1, 2$, and 3 can be permuted in Equation (6). Commonly, the variables in the vine trees are ordered based on the strength of the pair-wise correlations. Performing such an ordering for the pair copula decomposition gives trees for uncompressed (Figure 5a) and compressed paper sheets (Figure 5b),

in which the descriptors arrange in tree T_1 depending on whether the paper has undergone compression or not. We will interpret in which detail the trees of uncompressed and compressed paper sheets differ later in Section 3.1. Figure 5a,b shows that order and arrangement of the descriptors (in tree T_1) can differ from the ordering in the initially introduced random vector $W = (\varepsilon, \delta, S_V, \tau_0, \tau_3)$ and from the example tree in Figure 4. Hence it is convenient (i) to index the copula densities by the descriptors rather than their corresponding indices, for example $c_{1,5}$ is denoted by c_{ε,τ_3}, and, (ii), to indicate descriptors to be conditioned on after the semicolon in the copula indices (in analogy to the notation used in Equation (6)); e.g., $c_{\varepsilon,\delta;\tau_0,\tau_3}$ is the conditional bivariate copula density of (ε, δ) given the values of τ_0 and τ_3 under the simplifying assumption.

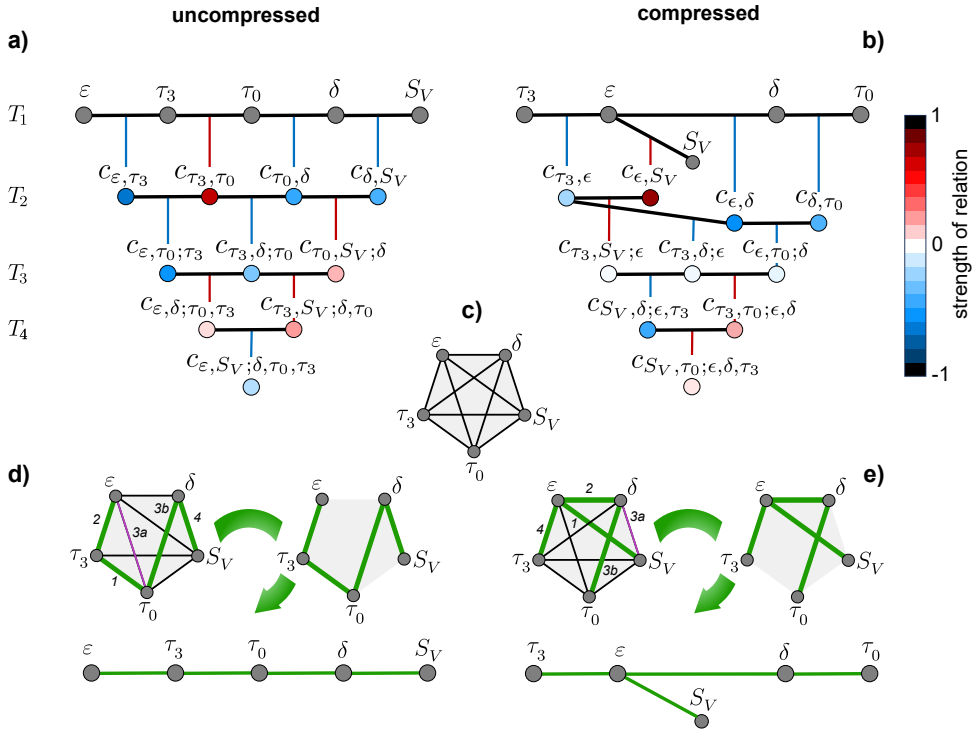

FIGURE 5. (a,b) Tree representation of pair copula decomposition obtained by fitting R-vine copulas to the multivariate distribution of the local descriptors porosity ε, sheet thickness δ, surface area per unit volume S_V and path-length related descriptors τ_0 and τ_3. Tree representations are shown for uncompressed (a) and compressed (b) paper sheets. Colored vertices of trees T_2, T_3 and T_4 indicate the size and sign of Kendall's tau τ_K. (c-e) Determination of the material-specific order of descriptors in tree T_1 using maximum spanning trees for the uncompressed (d) and compressed paper sample (e). (c) Graph in which descriptors are vertices arranged as in the random vector $W = (\varepsilon, \delta, S_v, \tau_0, \tau_3)$ and all pairwise connections represented as edges. (d,e) Shows how the maximum spanning tree for the graph (c) is determined by connecting descriptors with strongest pair-relations in descending order.

The arrangement of descriptors in Figure 5a,b strictly results from an ordering algorithm applied to the paper-specific data. This algorithm is implemented in the statistical software package `R` by the function `RVineStructureSelect` in the package `VineCopula`, see p. 134 in [24]. In brief, the structure of the trees is determined using maximum spanning trees with respect to Kendall's tau [22], which is a rank correlation coefficient used to quantify the ordinal association

between two measured quantities. In the first step we consider a complete graph in which every vertex represents a descriptor, see Figure 5c). Then, every edge represents a descriptor pair and receives a weighting based on the value of Kendall's tau. The so-called maximum spanning tree of this graph is determined according to Kruskal's algorithm, as illustrated in Figure 5d,e). Edges are marked (green lines) according to the absolute value of Kendall's tau, $|\tau_K|$, that is associated with the pair correlation of the vertices connected by the respective edge. The marking starts with the edge of largest $|\tau_K|$ (indexed with 1) and continues in descending order of $|\tau_K|$ (indices 2,3, and 4). If an edge is marked that connects vertices that are already connected via previously found edges, the edge (marked as purple line) is not considered in the tree and one proceeds with the edge of next lower $|\tau_K|$, for example discarding edge 3a in favor of edge 3b in the uncompressed case. The marking stops as soon as all five vertices are connected, i.e., four edges are identified. The resulting tree contains the descriptor *pairs* (e.g., (ε, τ_3) and (τ_3, τ_0) in T_1) as edges whose joint distribution is to be modeled by a bivariate copula function.

As the descriptors are arranged in T_1 according to their position in the maximum spanning tree (cf. Figure 5d,e)), the most strongly related pair (with the highest absolute value of Kendall's tau) is not necessarily placed in left-most position. In the next step, the *edges* of the maximum spanning tree become the vertices of a new graph (vertices in T_2). The set of edges of this new graph is constructed according to the so-called proximity condition, see p. 98 in [22], to ensure that the formal decomposition of the resulting vine into bivariate copulas is well-defined. The proximity condition requires that the corresponding edges of two connected vertices (e.g., c_{τ_3,τ_0} and $c_{\tau_0,\delta}$ in T_2, Figure 5a) of the new graph share a common vertex in the previous tree (τ_0 in T_1). This common vertex represents the common descriptors on which subsequent bivariate densities are conditioned on ($c_{\tau_3,\delta;\tau_0}$ in T_3). Due to the simplifying assumption, these conditional descriptors are ignored and Kendall's tau is again determined on the remaining descriptor pair. A maximum spanning tree is computed and the procedure is repeated recursively until only one vertex remains in the final tree.

2.5. **Bivariate distributions.** With the R-vine decomposition shown in Figure 5 at hand, the only remaining step is to fit a bivariate copula function to every descriptor pair in the R-vine trees. The procedure for this is implemented by the `R` function `BiCopSelect` in the package `VineCopula`, where a wide range of parametric copula families is available for fitting, see p. 69 in [24] for a full list. First, the parameters of all available copulas are determined through maximum likelihood estimation. The selection of the parametric family is then performed according to the Akaike information criterion [19].

The resulting fits for the joint bivariate densities are shown in Figure 6 for the examples of (i) the porosity ε and specific surface area S_V (Figure 6a and 6b) and (ii) for the porosity ε and thickness δ (Figure 6c and 6d). For the descriptor pairs appearing in the first tree of the R-vine decomposition (T_1 in Figure 5), this is a direct fit of a bivariate copula function to the joint bivariate densities. For the conditional distributions of descriptor pairs appearing in the subsequent trees, it is a fit under the simplifying assumption. All empirical and copula-modeled joint bivariate densities as well as the chosen copula families and their associated copula parameters are provided in the Supporting Material.

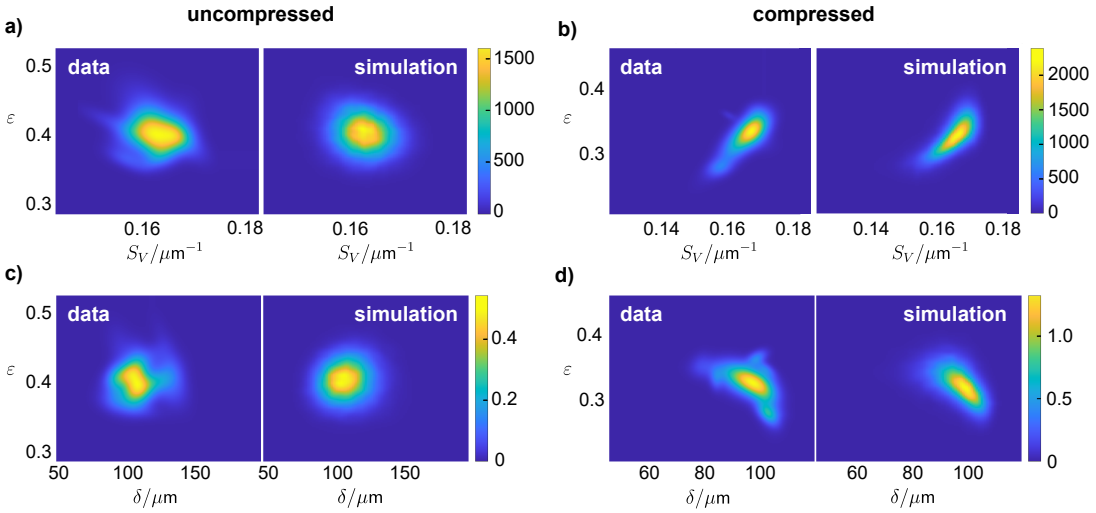

FIGURE 6. Bivariate probability densities of porosity ε and specific surface area S_V (a,b), and of porosity ε and sheet thickness δ (c,d), each shown for the uncompressed (a,c) and compressed (c,d) paper sheets. For each case, the bivariate density obtained via kernel density estimation is shown for measured image data (left) and simulated data drawn from the copula-based model (right).

3. Interdependence relations for microstructure descriptors

With the parametric model described above, the multivariate probability distribution of $W = (\varepsilon, \delta, S_v, \tau_0, \tau_3)$ can be analyzed with respect to different aspects. In particular, we use the R-vine copula model in order to quantitatively assess the impact of compressing the paper sheets. This compression acts only in the thickness direction, compacting the paper mainly in thicker areas with many stacked fibers, without the possibility of changing the in-plane orientation of the fibers, see Section 2 in [13].

In this way, a sample with a roughly uniform mass density and marked thickness variations (uncompressed) was transformed into a sample of marked mass density variations with only small thickness variations (compressed).

The copula-based model, the tree representations of which are shown in Figure 5, allow us to quantify compression-induced differences in multiple different ways.

3.1. Interdependence relations distinguish paper grades.

The tree representations of the parametric R-vine copula models inform on dependence structures between the descriptors for the two paper grades (Figure 5a,b). Each set of trees reveals the strongest, most-needed relationships with descriptors based on Kendall's tau in tree T_1. For uncompressed paper, for example, the three pairs that can be formed between the porosity ε and the two path length descriptors τ_0 and τ_3 relate most strongly (cf. Figure 3), but only two pairs appear in the tree T_1, (ε, τ_3) and (τ_3, τ_0) (Figure 5a). The pairwise relationship (ε, τ_0) is quite strong, but is not considered in T_1 because it is the least strong and therefore least needed to quantify the joint variations in ε, τ_0, and τ_3. Compression leads to a significant change in the dependency structure; it rearranges the descriptors within the first tree T_1 (Figure 5b) to emphasize the strong interdependencies between ε, δ with S_V. In contrast to the uncompressed sample, τ_0 and τ_3 interact only weakly with the other three descriptors.

Instead of comparing the strongly condensed information stored in the copula-based dependence structures, we can directly monitor how compression alters the correlation coefficients of each descriptor pair before and after compression. The values of these coefficients are visualized in Figure 3 for both samples using their squares, varying within the interval $[0, 1]$. The uncompressed paper has, as mentioned above, two particularly strong relations, between τ_3 and τ_0, and between ε and τ_3 (Figure 3, left green boxes), and concomitantly between ε and τ_0. Note that the correlation coefficient of the pairs (ε, τ_3) and (ε, τ_0) are negative, as indicated by their blue labels in Figure 3. That means that longer paths form the more likely, the smaller the local porosity is, regardless whether all possible local paths or local high-volume paths, i.e., paths with a minimum diameter of 3 μm, are considered. Upon compression, this intuitive relation between ε and τ_0 as well as between ε and τ_3, respectively, is practically lost. The strongly reduced squared values of the correlation coefficients of ε and τ_3, and τ_3 and τ_0 (Figure 3, right green boxes), respectively, suggest that the lengths of high-volume paths neither dependent on the available local porosity nor on the lengths of all conceivable local paths. As argued in [13], the low correlation between ε and τ_0, together with a reduced porosity and a practically unchanged distribution of all pathlengths, τ_0, suggests that the topology of the pathways is nearly unchanged. However, high volume paths get much longer upon compression (Figure 2b and 6c); some of them exceed the local thickness by a factor of 10 and more. Such pathways may start in a local environment of a given ε, but certainly leave this spot and run through denser or more open regions. Thus, whether such paths can form is not determined by the local porosity of their starting position anymore.

This illustrative discussion demonstrates that the set of correlation coefficients (Figure 3) and the copula-based dependence structure (Figure 5a,b) of $W = (\varepsilon, \delta, S_v, \tau_0, \tau_3)$ can serve as finger prints to compare and to distinguish differences between materials, here between differently treated (compressible) papers.

3.2. **Revealing relations between different kinds of specific surface areas.** Our parametric copula-based model enables us to unravel relationships to descriptors that are associated with, but not directly incorporated into, the parametric model. To demonstrate the use of such relationships, we will now focus on the surface area per unit mass, denoted by S_M, which is routinely measured, for example, by gas sorption, mercury intrusion porosimetry, or inverse gas chromatography. Thus, S_M is commonly used to compare and distinguish materials in terms of their internal surface area rather than S_V, the surface area per unit volume. However, S_M cannot simply replace S_V to convey information about the available internal surface area per unit volume, i.e., the amount of available surface sites. Using the models for the two paper samples considered in this study, we show that S_M includes not only the locally available surface area, but also the local porosity and its variation; correspondingly, samples can have the same value for S_V, but very different ones for S_M, and vice versa.

The relation between S_M and S_V is established via the local mass density ρ. Recall that, in a certain inspection region of the sample with volume V and internal surface area A, the surface per unit volume S_V is defined as

$$S_V = \frac{A}{V} \,. \tag{7}$$

Its counterpart S_M requires, besides the internal surface area A, also the mass M of the inspection region, *i.e.*,

$$S_M = \frac{A}{M} \ . \tag{8}$$

Thus,

$$S_M = \frac{1}{\rho} S_V, \tag{9}$$

where $\rho = M/V$. The mass density ρ can be estimated using mass densities ρ_s in the solid and ρ_p in the pore space, respectively, provided that ρ_s and ρ_p are constant. In our samples, ρ_s corresponds to the mass density of the fiber walls, which is expected to be invariant in each paper grade and unaffected by compression. Assuming that $\rho_p \ll \rho_s$, as the density ρ_p of air in the pores is three orders of magnitude smaller than the density of the fiber walls, we obtain

$$\rho = (1 - \varepsilon)\rho_s + \varepsilon\rho_p \approx (1 - \varepsilon)\rho_s \ . \tag{10}$$

In this approximation, all possible variations in the local density ρ of the porous material are caused by variations in the local porosity ε. Inserting the approximated relation given in Equation (10) into Equation (9) yields

$$S_M = \frac{S_V}{\rho_s(1 - \varepsilon)} \ . \tag{11}$$

Consider the surface area per unit solid volume S_{V_s} given by

$$S_{V_s} = \frac{S_V}{1 - \varepsilon}, \tag{12}$$

and note that S_M only differs from S_{V_s} by the constant factor ρ_s, since

$$S_M = \frac{S_{V_s}}{\rho_s}. \tag{13}$$

By means of S_{V_s} we can therefore predict the statistical behavior of S_M with the parametric copula-based model, even without knowing the value of the constant factor ρ_s.

On the other hand, Equation (11) readily shows that the spatial variation in S_M is not only governed by the variation in S_V, but also by variations in local mass density $\rho \approx \rho_s(1 - \varepsilon)$ and hence by variations in local porosity ε. Even though our paper samples possess only slightly different distributions of S_V, they represent two contrasting scenarios in terms of mass density variations and, thus, give quantitatively and qualitatively different values of S_M and S_{V_s}. Figure 7 tracks how the relation between S_{V_s} and S_V changes when passing from uncompressed to compressed paper.

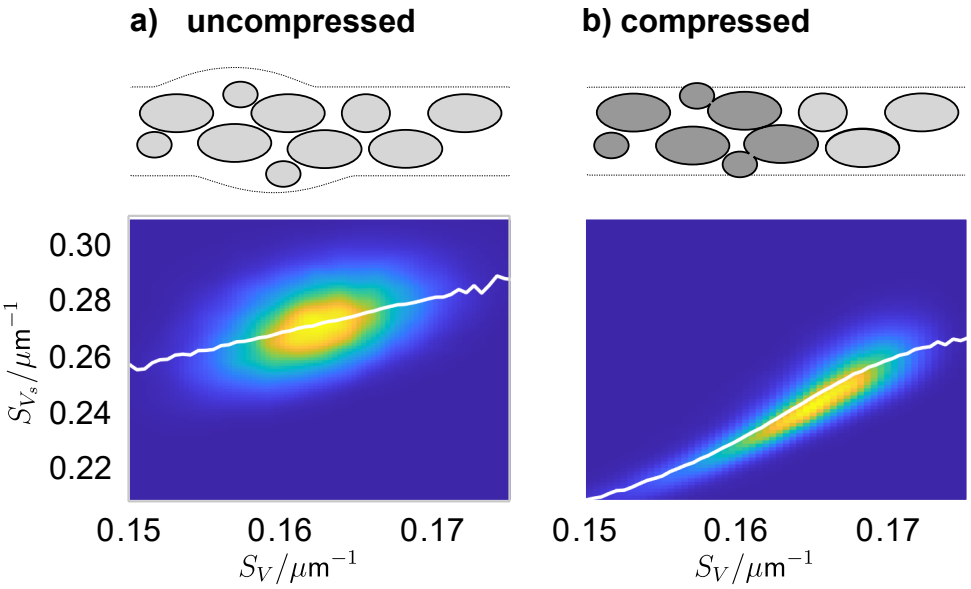

FIGURE 7. Comparison of uncompressed (a) and compressed (b) paper sheets to reveal the impact of compression on the relation between local surface area per unit volume, S_V, and local surface area per solid volume, S_{V_s}. For both paper sheets, the arrangement of fibers (gray) in a cross section of a sheet is schematically visualized. The bold contour of the cross sections of fibers indicate their contributions to the local surface area. In the compressed paper (b), compacted regions are indicated with fiber cross sections shaded in dark-gray. The joint probability density of S_{V_s} and S_V is shown as heatmap. The white curves indicate how the (conditional) mean value of S_{V_s} changes with the value of with S_V.

The uncompressed paper is composed of virgin, long softwood fibers that arrange layer-by-layer into stacks of fiber mats as shown in a schematic cross section in Figure 7a. This fiber arrangement remains largely unaffected during the production process. Hence, local accumulations of fibers lead to thickness variations [13] and the local surface area A (and, hence S_V) is expected to relate to the local porosity [25, 26]. The more surprising find is that the local specific surface area S_V does not correlate with the local porosity ε, as seen from the joint density of ε and S_V in Figure 6a. Even though we cannot explain the absence of any relation between S_V and ε in this paper sheet, theoretical models that aim to connect S_V to ε in packed beds (such as soils) [27, 28] or multiphase battery electrodes [29] predict indeed that a porosity range ($0 \ll \varepsilon \ll 1$) exists in which a change in local porosity does not trigger a change in S_V. As S_V is unaffected by ε, we expect S_{V_s} to link linearly to S_V according to Equation (12). The fitted joint density of S_{V_s} and S_V is visualized in Figure 7a by a heat map. Therein, associated mean values of S_{V_s} relate practically linearly to S_V (white line) as expected. Note that the underlying heat map gives additional insight into the predictive power of the white mean line, as it carries information about the local amount and spread of the available data-pairs of S_{V_s} and S_V.

Compression of the paper changes the relation between S_V and S_{V_s} as a result of *local* compaction. Compression, which is intended to smooth surfaces [13], acts exclusively in the thickness direction. It compresses thicker regions while thinner regions remain unchanged, as shown schematically in the cross section of Figure 7b. This process strongly alters the

relationships between descriptors of microstructure (see Section 3.1), albeit the distribution of S_V changes only slightly (see Figure 2d). The apparent increase in the mean value of S_V compared to the uncompressed case is likely due to lesser contributions of specific surface areas $\lessapprox 0.16\ \mu\mathrm{m}^{-1}$. The descriptor relations originate from the previously thicker, now compacted regions. Compaction reduces the vertical space between fibers and thus the local porosity ε. Wherever the vertical space is completely eliminated, the fibers are in direct contact, which reduces their contribution to the total surface area, so that S_V shrinks [25, 26]. Even when considering the whole sample, this strong dependence of S_V on ε is clearly reflected in their joint bivariate density (Figure 6b) and in the strong positive correlation (Figure 3). How S_V grows with ε determines how the relationship (12) between S_{V_s} and S_V accounts for local porosity: In the heat map of Figure 7b, we observe values of S_{V_s} that are much smaller than in the uncompressed case and the mean value curve of S_{V_s} has a much steeper slope. The steeper slope is not only due to lower local porosity ε, but also reflects the changed dependence of S_{V_s} on ε and, in parallel, the unchanged behavior of the uncompressed regions.

Note that S_V and S_M are not simply linearly related, as the proportionality factor $1/\rho_s(1-\varepsilon)$ in Equation (11) depends on the sample-specific local porosity ε, so that S_V and S_M cannot just be used interchangeably. While the white curve in Figure 7a might suggest a nearly linear dependence between S_V and S_{V_s} (and therefore also between S_V and S_M due to Equation (13)), this is only true for the conditional mean value of S_{V_s} (or S_M) given the value of S_V. However, the relation between S_V and S_{V_s} (or S_M) is random and cannot be captured by a simple linear relation.

Thus, the use of S_M alone to compare specific surface areas in a series of samples is not sufficient, because it cannot be excluded that samples differ in their local porosity.

3.3. **Bruggeman-type relations for the prediction of tortuosity.** The copula-based model presented in this study also allows us to justify formulas that relate various descriptors of porous materials with each other. We illustrate this for a common wisdom associated to porous materials: Higher porosities promote shorter pathlengths and, hence, smaller tortuosities [30, 31]. This expectation is often cast into the form of Bruggeman-type relations for the so-called tortuosity factor τ, which characterizes the length of effective (not necessarily shortest) transportation paths. In particular, when the transport is obstructed by cylindrical or spherical objects, literature considers the formula

$$\tau = \frac{1}{\varepsilon^a} \tag{14}$$

for random arrangements of solid cylinders or spheres, in which the exponent a in Equation (14) is $1/2$ or unity, respectively [32].

In the following we show how the relation between τ and ε given in Equation (14) can be extended for the mean geodesic tortuosities τ_0 and τ_3 considered in the present study. With the parametric copula-based model delivering ε and τ_r, with r being either 0 or 3, we can readily determine the quotient

$$a = -\frac{\log \tau_r}{\log \varepsilon}, \tag{15}$$

where a is not longer a deterministic constant as in Equation (14), but a random variable.

Since we learned that the local porosity ε is either strongly correlated with the tortuosities τ_0, τ_3 (uncompressed sample) or the local thickness δ (compressed sample), see Figure 3, it is worthwhile to check whether the random Bruggeman exponent a in Equation (15) depends

on the thickness δ. Figure 8 shows how the exponent a in Equation (15) predicted from the

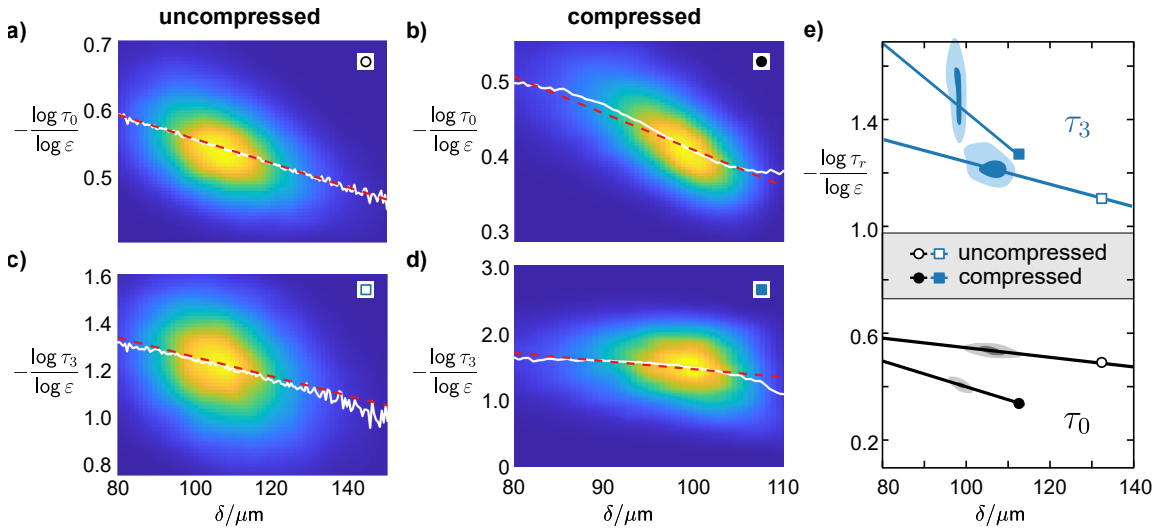

FIGURE 8. Dependency of the exponent $a = -\log \tau_r / \log \varepsilon$ on the local thickness δ, shown for uncompressed (a,c) and compressed paper sheets (b,d) by the bivariate probability density of (a, δ) (heatmap), together with the conditional mean values of a for given values of δ (white curves) and its linear fit (dashed lines). Left panels (a,c) concern the tortuosity τ_0 (i.e., $r = 0$) with respect to all paths, while the middle panels (b,d) refer to tortuosity τ_3 with respect to high volume paths ($r = 3$). The right panel (e) shows a comparison of the linear relation between a and δ for all four cases on equal length scales. As a guide to the eye, the regions with high probability in (a-d) are indicated.

parametric model varies with thickness δ when either the tortuosity τ_0 of all pathways (Figure 8a and 8b) or τ_3 of high volume pathways is considered (Figure 8c and 8d). Each panel shows (i) the bivariate density of the joint probability distribution of a and δ (heatmap), and (ii) the conditional mean value of a given δ (white solid curves).

The exponent a, best seen from the mean value curve, shows a marked dependence on the thickness δ regardless of the (uncompressed or compressed) sample or the chosen variant of the tortuosities τ_0, τ_3. Moreover, for each considered case, the evolution of the mean value curve of a within the range of encountered thicknesses δ can be well quantified by a linear function (dashed lines in Figure 8a-d), *i.e.*, the exponent a is approximated by

$$a = b\delta + c\,, \tag{16}$$

with $b, c \in \mathbb{R}$ being adjustable parameters. The fitted values of b and c as well as the coefficient of determination R^2 are given in Table 1. Note that for the compressed sample, the goodness of fit could be further improved if the fit only considers thicknesses near the mean thickness of ca. 100 μm, see Figure 2e. These linear approximations of a, collected in Figure 8e, exhibit significant differences that depend on the sample and variant of tortuosity considered. Thus, the porosity- and thickness-dependence of the mean geodesic tortuosity τ_r (r being either 0 or 3), is captured with the parameterized formula

$$\tau_r = \frac{1}{\varepsilon^{b\delta+c}}\,, \tag{17}$$

where the values of $b, c \in \mathbb{R}$ are chosen separately for each of the four cases uncompressed paper, compressed paper.

Having in mind that the Bruggeman-type relation in Equation (14) was easily extended to include the dependence on the local thickness δ, the question arises whether the prediction of τ_r via Equation (17) could be further improved if the exponent a also took a dependence on the remaining descriptor, the surface area per unit volume S_V, into account. Therefore, an extended linear model of the form

$$a = b\delta + cS_V + d\,, \tag{18}$$

where $b, c, d \in \mathbb{R}$ are some parameters, was fitted to realizations of $(\varepsilon, \delta, S_V, \tau_0, \tau_3)$ drawn from the copula-based model. However, it turned out that the fit using Equation (18) improves the prediction of τ_r only marginally compared to Equation (16), see the last two columns of Table 1. Hence, there is no significant benefit in increasing the complexity of the model beyond Equation (17), see also the Supporting Material.

Sample	Tortuosity	b of Eq. (16)	c of Eq. (16)	R^2 of Eq. (16)	R^2 of Eq. (18)
compressed	τ_0	$-4.8 \cdot 10^{-3}$	0.89	0.365	0.392
compressed	τ_3	$-12.9 \cdot 10^{-3}$	2.72	0.045	0.052
uncompressed	τ_0	$-1.8 \cdot 10^{-3}$	0.73	0.186	0.195
uncompressed	τ_3	$-4.2 \cdot 10^{-3}$	1.66	0.091	0.097

TABLE 1. Parameters b, c and coefficients of determination of the linear model given in Equation (16), as well as coefficients of determination of the model in Equation (18). Note that the model in Equation (16) only considers local thickness, while the model in Equation (18) considers local thickness and local surface area per unit volume.

4. Discussion

Being equipped with the general concept of R-vine copulas, two additional technical considerations promise to substantiate the model and further predictions to be derived from it: (1) As a sanity check for the distilled cross-descriptor relations it is fair to ask whether an R-vine copula keeps its shape upon considering other local cutout sizes. As the model relies on local descriptors obtained for a certain, fixed cutout size, it inherits the descriptor variation that is characteristic for the cutout size [33]. The larger the cutout size, the smaller the variation [10, 13, 34]. So it is intriguing to establish, whether the size-dependent R-vine copula models preserve (i) the correlation ranking of the descriptors (*i.e.*, the tree shape of the R-vine copula) and (ii) the same bivariate copula density functions. (2) The results from Section 3.2 and 3.3 include information about the predictive power of considered models by indicating the local amount and spread of available data in the heat maps of the probability densities. This predicted validity range aids to suggest which further data ranges ought to be explored. Additional data could be supplied either by probing further regions in the sample or by inspecting deliberately altered samples, provided that the initial data set is already representative of the sample [13].

The R-vine-based parametric models allow the complexity of highly varying microstructures to be compactly captured by sample-specific multivariate distributions of five microstructure

descriptors. Each model distribution consists of a set of parameterized univariate and bivariate distributions ordered and linked according to their pairwise dependence expressed by the Kendall-Tau rank coefficient. With this formal act of complexity reduction, our models provide valuable structure to the five-dimensional space of descriptors itself. The very composition of the model, and thus the dependence structure determined, is highly material-specific and capable of distinguishing materials: Pairwise dependencies between descriptors are quantified as an inherent feature of the model. Beyond the pairwise dependencies, probability distributions with three or more descriptors can be extracted and used to confirm or find previously unknown analytical formulas for the relationship between descriptors, as shown for the surface area per unit mass, S_M, and for the Bruggemann relationship between τ_r and ε. A final and very important, but easily overlooked, aspect is that the choice of descriptors allows us to interpret the relationships between each descriptor in terms of microstructure and its changes. Rather than relying on descriptors whose interrelationships are completely unclear and often come from unrelated measurements or are simply readily available, we have taken care to select descriptors that are known to contribute collectively, yet in a structure-specific manner, to transport through disordered materials. Considering the above, parametric models based on vine copulas are expected to reach their full potential when coupled with subsequent searches for structural properties or optimization of related properties, e.g., permittivity.

Most of the conceivable benefits exemplified in this study are not limited to unordered materials. Even before the actual matching procedure, the identified dependencies between descriptors may suggest additional dimensionality reduction [35]; for strongly related pairs of descriptors, one descriptor could be completely replaced by another [20]. Our model is also useful in selecting or generating configurations: (i) A parametric model, such as the one presented in this paper, inherently contains the physically consistent range of descriptor values. Since the descriptors do not need to be normally distributed, no additional constraints are required for the generation of configurations to ensure physically consistent values. (ii) Thanks to the parametric formulation of the multivariate probability density, interpolation between known configurations to generate unseen configurations is fast and robust. In particular, the prediction of rare and extreme configurations directly benefits from the ability of copulas to accurately describe the likelihood of configurations that are far from the most likely. (iii) The latter implies that material-specific conditions can be imposed on algorithms to generate statistically equivalent morphologies that reproduce the dependency structure between descriptors. (iv) Our parametric multivariate model can even be made compatible with frameworks that use Gaussian approaches, although these approaches necessarily require multidimensional normal distributions. The idea is to iteratively transform the probability distribution model into multivariate normal distributions and vice versa, using normalizing fluxes [36, 37].

Taken together, vine copula-based parametric models characterize vectors of microstructure descriptors for various kinds of materials and their relationships in a compact, structured, and easily interpretable manner. They provide an excellent relation-wise and physically informed starting point for fast and robust sampling of higher-dimensional descriptor spaces of materials, in particular of disordered heterogeneous materials. Therefore, these models are a good tool to generate a data-based understanding of local, structural relationships in materials.

5. Methods

Material. Two types of paper are considered. The first, uncompressed one serves as our reference paper. It is a commercial product consisting of unbleached softwood virgin pulp fibers, that does not contain fillers and has not passed any mechanical post-treatment. The paper sheets exhibit low variability in local mass density, while its local thickness varies appreciably across the sheet. To obtain the second type, *i.e.*, the compressed paper, the reference paper was subjected to a hard-nip, steel-steel calendering with a line load of 90 N/m. To determine whether this compression had the desired impact on local mass density and thickness, each paper type underwent a characterization in terms of basic paper-specific quantifiers; the detailed values are given in [13]. The compression preserves the basis weight of 100 g/m^2 [38], but markedly reduces the sheet thickness, determined via the caliper-based thickness (the apparent thickness associated with the most protruding regions of the paper sheets [39]). Calendering smoothes both surfaces as indicated by an enhanced Bekk smoothness parameter (informing on the time (in seconds) for a fixed volume of air to leak between the surfaces of a paper sample and a smooth glass) [40]. In parallel, markedly enhanced air retention times [41] demonstrates a compression-induced reduction in air permeance.

Microstructure acquisition. The microstructures of the two considered paper grades were acquired by imaging the papers with X-ray microcomputed tomography followed by a segmentation of the 3D images. The voxel size in the microstructure data sets is 1.3 μm. To capture the variations in the microstructure, 150 volumes of the uncompressed sample with a field of view of 1.7×1.4 mm^2 were acquired that covered a totally scanned area of approximately 2.9 cm^2. For the compressed sample less volumes sufficed as the field of view was with 3.29×2.46 mm^2 larger. Details related to sample preparation and structure acquisition, as well as post-treatment and binarization of the 3D image data is provided in [13].

Preprocessing methods providing descriptor data. To compute the local microstructure descriptors, we partitioned all binarized data sets into non-overlapping, local inspection regions. These inspection regions are square-shaped in lateral direction and contain the entire thickness of the paper sheets. For each inspection region, the values of the five descriptors were determined and collected in 5-tuples of $(\varepsilon, \delta, S_V, \tau_0, \tau_3)$. Though the descriptor values depend on the size of the inspection regions [13], we illustrate the construction of the parametric probabilistic model and the interdependence relations for regions with a side length of 330 μm. A total of 200 non-overlapping regions were considered for each paper grade. When combining these regions, the associated total area 200 × (330 μm)^2 = 21.78 mm^2 ensures a comprehensive representation of the variations in each paper grade, i.e., any further region added would not change the univariate distributions within a Kolmogorov distance of 0.05. [13]

Acknowledgments

The authors gratefully acknowledge financial support from the Christian Doppler Research Association, Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology, and Development, Austria. M.N. acknowledges funding by the German Research Foundation (DFG) under Project ID 390874152 (POLiS Cluster of Excellence, EXC 2154). K.Z. thanks the "TU Graz Lead Project LP-03: Porous Materials @ Work for Sustainability" for inspiration.

Author contributions

E.M.C., R.S., A.H., I.M. acquired the micro-CT scans. M.N., E.M.C., and A.H. segmented the image data. U.H. suggested use cases and supported the interpretation of the model predictions. M.N. and P.G. performed the statistical analysis of the microstructure data. M.N., P.G., and K.Z. conceptualized and wrote the manuscript. M.N., P.G., R.S., I.M., U.H., V.S. and K.Z. were involved in reading & editing of the manuscript. All authors have approved the final version of the manuscript.

Competing interests

There are no competing interests related to this work.

References

[1] E. Bevacqua, L. Suarez-Gutierrez, A. Jézéquel, F. Lehner, M. Vrac, P. Yiou, and J. Zscheischler. Advancing research on compound weather and climate events via large ensemble model simulations. *Nature Communications*, 14:2145, 2023.

[2] S. Rolland du Roscoat, J. F. Bloch, and X. Thibault. Synchrotron radiation microtomography applied to investigation of paper. *Journal of Physics D: Applied Physics*, 38:A78, 2005.

[3] R. Holmstad, A. Goel, S. Ramaswamy, and O. W. Gregersen. Visualization and characterization of high resolution 3D images of paper samples. *Appita Journal: Journal of the Technical Association of the Australian and New Zealand Pulp and Paper Industry*, 59:370, 2006.

[4] G. Chinga-Carrasco, M. Axelsson, O. Eriksen, and S. Svensson. Structural characteristics of pore networks affecting print-through. *Journal of Pulp and Paper Science*, 34:13–22, 2008.

[5] G. Chinga-Carrasco. Exploring the multi-scale structure of printing paper - A review of modern technology. *Journal of Microscopy*, 234:211–242, 2009.

[6] H. Aslannejad, S. M. Hassanizadeh, A. Raoof, D. A. M. de Winter, N. Tomozeiu, and M. T. van Genuchten. Characterizing the hydraulic properties of paper coating layer using FIB-SEM tomography and 3D pore-scale modeling. *Chemical Engineering Science*, 160:275–280, 2017.

[7] C. T. J. Dodson, Y. Oba, and W. W. Sampson. On the distributions of mass, thickness and density of paper. *Appita Journal: Journal of the Technical Association of the Australian and New Zealand Pulp and Paper Industry*, 54:385–389, 2001.

[8] C. T. J. Dodson and W. W. Sampson. Spatial statistics of stochastic fiber networks. *Journal of Statistical Physics*, 96:447–458, 1999.

[9] C. T. J. Dodson, Y. Oba, and W. W. Sampson. Bivariate normal thickness-density structure in real near-planar stochastic fiber networks. *Journal of Statistical Physics*, 102:345–353, 2001.

[10] S. Rolland du Roscoat, M. Decain, X. Thibault, C. Geindreau, and J.-F. Bloch. Estimation of microstructural properties from synchrotron X-ray microtomography and determination of the REV in paper materials. *Acta Materialia*, 55:2841–2850, 2007.

[11] D. S. Keller, D. L. Branca, and O. Kwon. Characterization of nonwoven structures by spatial partitioning of local thickness and mass density. *Journal of Materials Science*, 47:208–226, 2012.

[12] P. Leitl, E. Machado Charry, E. Baikova, M. Neumann, U. Hirn, V. Schmidt, and K. Zojer. Joint distributions of local pore space properties quantitatively explain simulated air flow variations in paper. *Transport in Porous Media*, 148:627–648, 2023.

[13] M. Neumann, E. Machado-Charry, E. Baikova, A. Hilger, U. Hirn, R. Schennach, I. Manke, V. Schmidt, and K. Zojer. Capturing centimeter-scale local variations in paper pore space via μ-CT: A benchmark study using calendered paper. *Microscopy and Microanalysis*, 27:1305–1315, 2021.

[14] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38:2916–2957, 2010.

[15] M. Neumann, E. Machado Charry, K. Zojer, and V. Schmidt. On variability and interdependence of local porosity and local tortuosity in porous materials: A case study for sack paper. *Methodology and Computing in Applied Probability*, 23:613–627, 2021.

[16] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. J. Wiley & Sons, 1995.

[17] A. C. Cohen. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. *Technometrics*, 7:579–588, 1965.

[18] K. K. Talukdar and W. D. Lawing. Estimation of the parameters of the Rice distribution. *The Journal of the Acoustical Society of America*, 89:1193–1197, 1991.

[19] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 1995.

[20] C. T. J. Dodson, M. Mettänen, and W. W. Sampson. Dimensionality reduction for information geometric characterization of surface topographies. In F. Nielsen, F. Critchley, and C. T. J. Dodson, editors, *Computational Information Geometry: For Image and Signal Processing*, pages 133–147. Springer International Publishing, 2017.

[21] H. Joe. *Dependence Modeling with Copulas*. CRC Press, 2014.

[22] C. Czado. *Analyzing Dependent Data with Vine Copulas*. Springer, 2019.

[23] M. Killiches, D. Kraus, and C. Czado. Examination and visualisation of the simplifying assumption for vine copulas in three dimensions. *Australian & New Zealand Journal of Statistics*, 59:95–117, 2017.

[24] T. Nagler, U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, and T. Erhardt. *VineCopula: Statistical Inference of Vine Copulas*, 2023. R package version 2.4.5.

[25] S. Eichhorn and W. Sampson. Statistical geometry of pores and statistics of porous nanofibrous assemblies. *Journal of the Royal Society Interface*, 2:309–318, 2005.

[26] S. Eichhorn and W. Sampson. Relationships between specific surface area and pore size in electrospun polymer fibre networks. *Journal of the Royal Society Interface*, 7:641–649, 2009.

[27] S. Torquato. *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*. Springer, 2002.

[28] F. Wadsworth, J. Vasseur, E. Llewellin, R. Brown, H. Tuffen, J. Gardner, J. Kendrick, Y. Lavallée, K. Dobson, M. J. Heap, D. Dingwell, K.-U. Hess, J. Schauroth, F. Von Aulock, A. R. L. Kushnir, and F. Marone. A model for permeability evolution during volcanic welding. *Journal of Volcanology and Geothermal Research*, 409:107118, 2021.

[29] H. Moussaoui, R. K. Sharma, J. Debayle, Y. Gavet, G. Delette, and J. Laurencin. Microstructural correlations for specific surface area and triple phase boundary length for composite electrodes of solid oxide cells. *Journal of Power Sources*, 412:736–748, 2019.

[30] P. U. Foscolo, L. G. Gibilaro, and S. P. Waldram. A unified model for particulate expansion of fluidised beds and flow in fixed porous media. *Chemical Engineering Science*, 38:1251–1260, 1983.

[31] L. Pisani. A geometrical study of the tortuosity of anisotropic porous media. *Transport in Porous Media*, 114:201–211, 2016.

[32] B. Tjaden, S. J. Cooper, D. J. L. Brett, D. Kramer, and P. R. Shearing. On the origin and application of the Bruggeman correlation for analysing transport phenomena in electrochemical systems. *Current Opinion in Chemical Engineering*, 12:44–51, 2016.

[33] W. W. Sampson. The structural characterisation of fibre networks in papermaking processes - A review. In C. F. Baker, editor, *The Science of Papermaking*, Transactions of the 12th Fundamental Research Symposium, pages 1205–1288, Oxford, 2001. FRC.

[34] T. Kanit, S. Forest, I. Galliet, V. Mounoury, and D. Jeulin. Determination of the size of the representative volume element for random composites: statistical and numerical approach. *International Journal of Solids and Structures*, 40:3647–3679, 2003.

[35] E. Boattini, M. Dijkstra, and L. Filion. Unsupervised learning for local structure detection in colloidal systems. *The Journal of Chemical Physics*, 151:154901, 2019.

[36] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

[37] I. Kobyzev, S. D. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43:3964–3979, 2021.

[38] ISO 536:2019. Paper and board: Determination of grammage. Standard, International Organization for Standardization, Geneva, CH, 2019.

[39] ISO 534:2011(E). Paper and board: Determination of thickness, density and specific volume. Standard, International Organization for Standardization, Geneva, CH, 2011.

[40] ISO 5627:1995. Paper and board: Determination of smoothness (Bekk method). Standard, International Organization for Standardization, Geneva, CH, 1995.

[41] ISO 5636-5:2013. Paper and board: Determination of air permeance (Medium range) – Part 5: Gurley method. Standard, Geneva: Standard, International Organization for Standardization, Geneva, CH, 2013.