Big Data for Microstructure-Property Relationships: A Case Study of Predicting Effective Conductivities

Ole Stenzel, Omar Pecho, and Lorenz Holzer
Institute of Computational Physics, ZHAW, CH-8400 Winterthur, Switzerland

Matthias Neumann [ORCID] **and Volker Schmidt**
Institute of Stochastics, Ulm University, D-89069 Ulm, Germany

The analysis of big data is changing industries, businesses and research as large amounts of data are available nowadays. In the area of microstructures, acquisition of (3-D tomographic image) data is difficult and time-consuming. It is shown that large amounts of data representing the geometry of virtual, but realistic 3-D microstructures can be generated using stochastic microstructure modeling. Combining the model output with physical simulations and data mining techniques, microstructure-property relationships can be quantitatively characterized. Exemplarily, we aim to predict effective conductivities given the microstructure characteristics volume fraction, mean geodesic tortuosity, and constrictivity. Therefore, we analyze 8119 microstructures generated by two different stochastic 3-D microstructure models. This is—to the best of our knowledge—by far the largest set of microstructures that has ever been analyzed. Fitting artificial neural networks, random forests and classical equations, the prediction of effective conductivities based on geometric microstructure characteristics is possible. © 2017 American Institute of Chemical Engineers AIChE J, 00: 000–000, 2017
Keywords: big data, effective conductivity, geodesic tortuosity, microstructure characteristics, predictive simulation, stochastic microstructure modeling

Introduction

Data is the new oil. The analysis of big data is changing industries, businesses, and research. Big data is also used to advance materials research[1] aiming an accelerated systematic design of functional materials[2] like (organic) solar cells, fuel cells and batteries. This includes the identification of new chemical compositions with desired properties as well as the optimization of microstructures or nanostructures, that is, the spatial arrangement of materials components, which have a large influence on the functional properties of these materials.[3]

To optimize microstructures in functional materials, the relationship between microstructure characteristics and functional properties has to be understood quantitatively, which is often not the case or just for some special types of simple structures.[3] Here, it is important to note that many functional properties cannot be predicted only by the volume fraction of the microstructure.[4,5] Thus, further microstructure characteristics have to be considered for a better understanding of the relationship between microstructure characteristics and functionality.

The progress of 3-D imaging during the last decades enables the computation of well-defined microstructure characteristics from real data, which can be compared to effective properties that are either measured experimentally or simulated with numerical models.[6–9] Although this approach allows a direct investigation of the relationship between microstructure and effective properties, it is limited due to the high costs of 3-D imaging.

Thus, virtual materials testing (VMT), that is, the combination of stochastic microstructure models (SMM) with numerical simulations of physical processes, was used to investigate the quantitative relationship between microstructure characteristics and effective conductivity in porous materials.[10,11] The use of SMM allows us to generate virtual microstructures in short time, where certain microstructure characteristics can be varied systematically. The virtual microstructures are used as an input for finite element modeling (FEM) where the corresponding effective conductivities are simulated. The generation of virtual microstructures leads to big data and thus, the microstructure-property relationships can be considered as a statistical learning problem.

It was shown that effective conductivity σ_{eff} of porous microstructures can be approximately predicted by three microstructure characteristics,[11] which are volume fraction ε of the solid phase, its mean geodesic tortuosity τ_{geod} and a certain constriction factor β, using the equation

$$\sigma_{\text{eff}} = \sigma_0 \frac{\varepsilon^{1.15} \beta^{0.37}}{\tau_{\text{geod}}^{4.39}}, \qquad (1)$$

where σ_0 denotes the intrinsic conductivity of the bulk material without microstructure limitation. This empirical
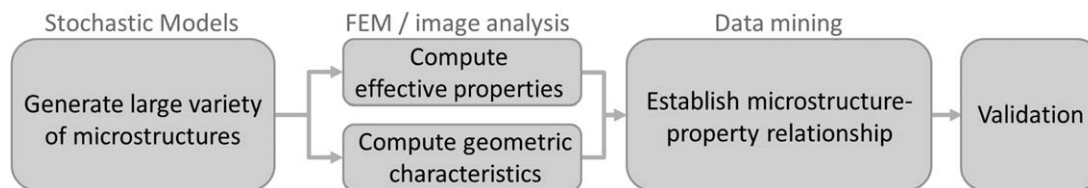
Figure 1. Combination of VMT with statistical learning to analyze relationships between microstructure characteristics and functional properties using big data.

relationship has been established on the basis of 43 virtual microstructures, where the corresponding effective conductivities have been computed with the software GeoDict.[12] In this article, we consider 8119 virtual microstructures and use neural networks and random forests to predict the effective conductivity given the structural properties. Although, it is more difficult to interpret these new prediction formulas in comparison with Eq. 1, they increase accuracy of the prediction. The effective conductivity can now be predicted with a prediction error of less than 9% instead of 13.6%, which is the prediction error when applying Eq. 1 to the 8119 virtual microstructures. This shows that combining stochastic microstructure modeling, physical computations and data mining is a powerful and helpful approach to establish quantitative microstructure-property relationships. This concept, outlined in Figure 1, is not restricted to conductive transport processes, but can—in principle—be applied to establish all kinds of microstructure-property relationships.

The concept, however, has the disadvantage that the effective conductivities of the virtual microstructures cannot be compared with experimental measurements. For a proper validation, one needs to prepare real samples, measure the effective conductivities experimentally, do 3-D imaging of the samples and then compare experimentally measured with predicted conductivities. Such a comparison is expensive in costs and time and can only be performed for a small number of samples. This was done to validate the VMT approach,[10] where a reasonably good agreement between predicted and measured conductivities was found. Validation of simulating effective conductivity by GeoDict can be found in the literature.[13]

Moreover, it has to be emphasized that we do not want to replace experimental 3-D imaging by the in-silico approach of VMT. It can be understood as an additional tool which makes 3-D imaging more powerful in tailoring new microstructures with specific properties. To create virtual, but realistic microstructures, a certain SMM is fitted to experimental microstructures such that the model creates statistically equivalent microstructures. Then, model parameters of the SMM can be correlated to production parameters of the microstructures, to suggest production parameters that lead to a certain type of microstructure.[14] In many materials, for example, in fuel cells, various different transport processes take place simultaneously which makes microstructure optimization difficult: each type of transport process may prefer a different microstructure. Thus, for a successful microstructure optimization, quantitative microstructure-property relationships must be established not only for conductive transport but also for other kinds of transport processes, for example, for effective permeability or mechanical stress-strain curves.

This article is organized as follows. At first, we present the SMM used for the generation of virtual microstructures, their geometric characteristics (volume fraction, constrictivity,

mean geodesic tortuosity), the considered transport processes as well as the predictive models from statistical learning. The results are presented and discussed, where the proposed microstructure-property relationships are validated by experimental image data, too.

Data and Methods

Stochastic microstructure modeling

With increasing availability of highly resolved image data stochastic microstructure modeling becomes a frequently used tool in materials science.[3] During the last years a number of stochastic microstructure and nanostructure models has been created for specific types of microstructures in organic solar cells, Li-ion batteries, and fuel cells.[15]

In general, an SMM uses tools from stochastic geometry[16] to generate virtual, random microstructures whose properties can be adjusted by the model parameters. To develop an SMM, a purposive combination of random variables is used to model spatial data, like point configurations, spatial networks, or random sets. The generation of a virtual microstructure typically requires little computational effort and therefore many different microstructures can be simulated in short time.

A simple example for an SMM is the Boolean model with spherical grains,[17] where possibly overlapping spheres are distributed completely at random in space (2-D or 3-D) with a predefined distribution of radii. The influence of model parameters on transport properties has been recently investigated for Boolean models with more general grains.[18]

For our case study, we consider two SMM that generate different types of microstructures: the stochastic spatial graph model[10] (SSGM) and a simplified version of the multiscale sphere model (MSM).[14,19] By means of the SSGM microstructures within a wide range of different values for volume fraction, mean geodesic tortuosity, and constrictivity can be generated. We additionally incorporate the MSM into our investigation as it was fitted to image data of real microstructures. Moreover, considering two models instead of one reduces the errors introduced by the model type.

Stochastic spatial graph model

The stochastic spatial graph model (SSGM) is based on a random spatial graph that is randomly dilated,[10] see Figure 2. The model has a large flexibility to generate microstructures with different volume fractions, mean geodesic tortuosities, and constrictivities. All microstructures realized by the SSGM are completely connected by definition. Using the SSGM, 3900 microstructures with different structural characteristics have been generated for this study.

Multiscale sphere model

The second SMM is the multiscale sphere model[14,19] (MSM). It follows a completely different approach in
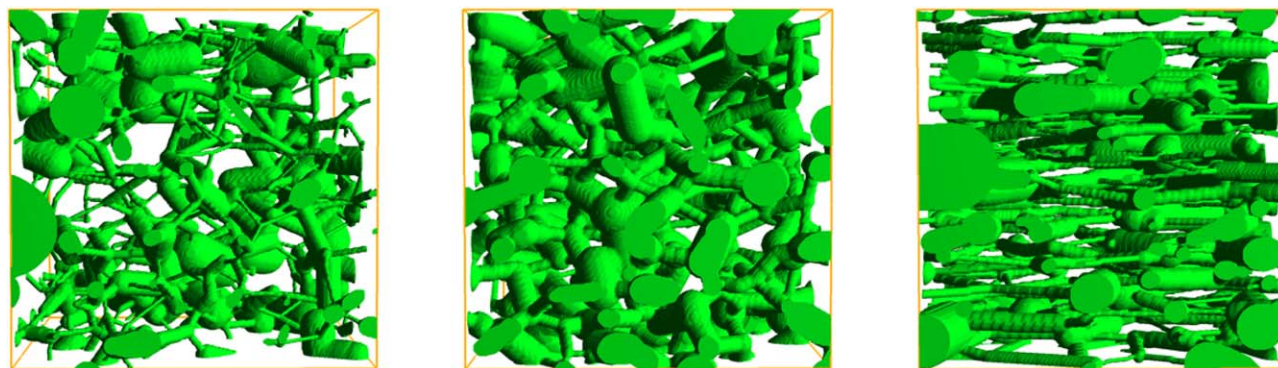
Figure 2. Virtual microstructures generated by the SSGM.

[Color figure can be viewed at wileyonlinelibrary.com]

comparison to the SSGM described in the previous section. It is based on a random, anisotropic arrangement of spheres. The midpoints of the spheres follow a Markov-chain of 2-D point processes. The model has two components: a macroscale component and a microscale component that adds structural complexity. In this article, we only use the macroscale component of the MSM. Examples of realizations are displayed in Figure 3. In total, we consider 2131 microstructures where the sphere system is the transport phase and 2088 microstructures where transport takes place in the complement of the sphere system. As only the connected (nonisolated) part of the considered material phase contributes to transport, a postprocessing is applied where all material is removed that is not connected with both, inlet and outlet plane.

Geometric characteristics

VMT has shown that three microstructure characteristics of the conducting phase carry significant information with respect to σ_{eff}.[10,11] These microstructure characteristics are volume fraction ε, mean geodesic tortuosity τ_{geod} and constrictivity β. A detailed discussion about the importance of these three characteristics for conduction processes is given in a previous publication,[20] where, additionally, the concept of constrictivity is transferred from simplified geometries[21] to complex microstructures.
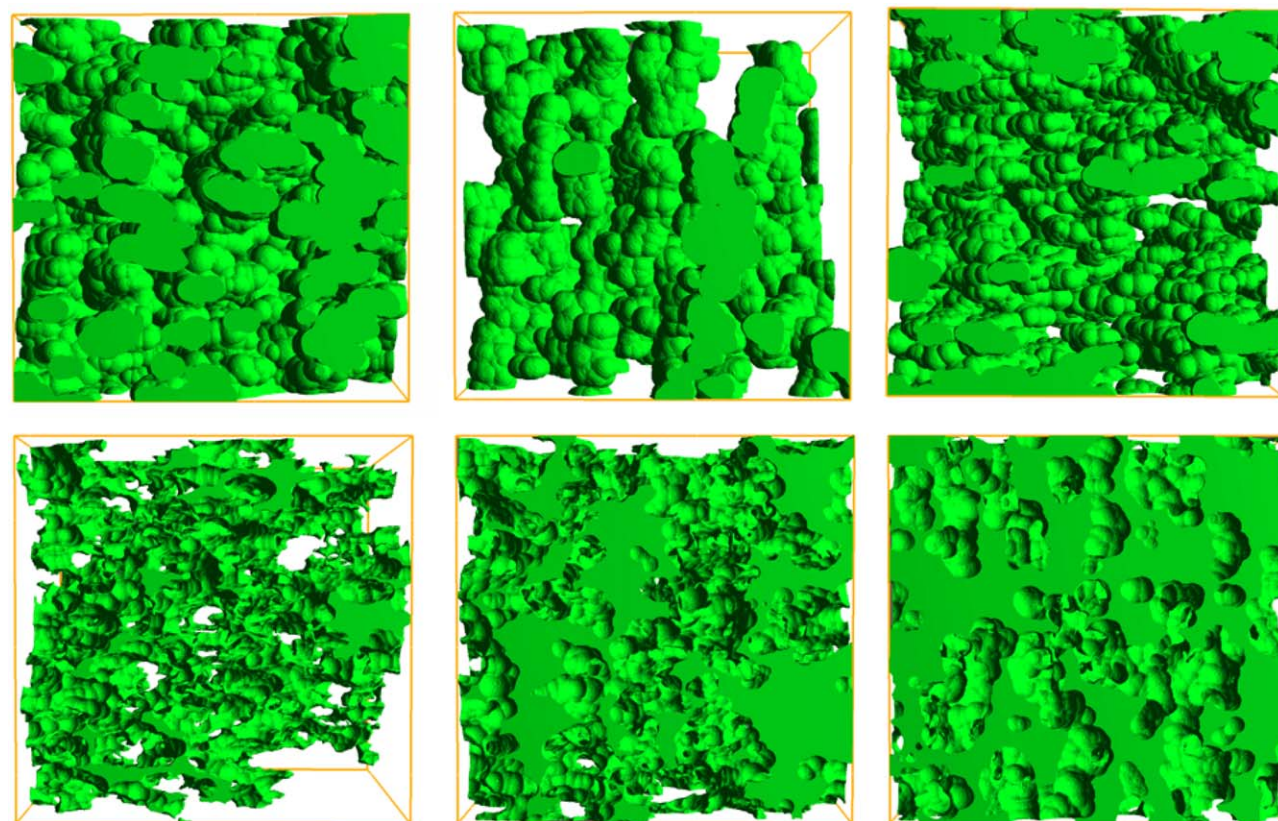

Figure 3. Virtual microstructures generated by a simplified version of the MSM.

Top row: microstructures generated by a sphere system. Bottom row: microstructures generated as complementary phase of a sphere system. [Color figure can be viewed at wileyonlinelibrary.com]
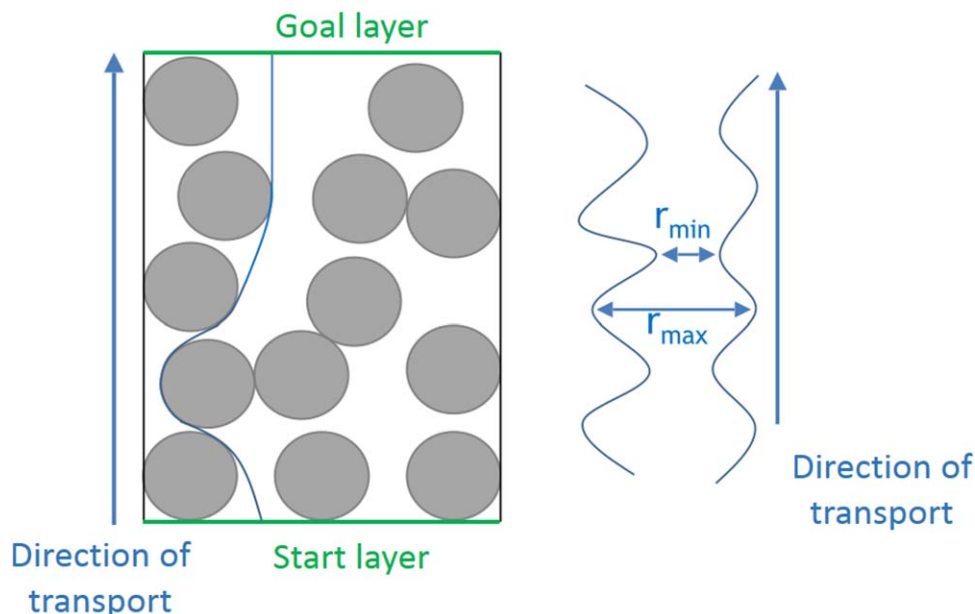
Figure 4. Concept of geodesic tortuosity τ_{geod} (left) and concept of constrictivity $\beta = (r_{\min}/r_{\max})^2$ (right) (Reprinted from Figures 3 and 4 in Ref. 11 with permission from [Wiley]).

[Color figure can be viewed at wileyonlinelibrary.com]

Note that the considered micostructure characteristics ε, τ_{geod}, and β can be defined by means of expectations with respect to the underlying stochastic model[16] or they can be estimated from a given microstructure. In this article, we use the latter way, because given a certain microstructure we are interested in the influence of ε, τ_{geod}, and β on σ_{eff}.

The volume fraction ε is estimated by the ratio of the volume of the transporting phase divided by the total volume of the 3-D image. The influence of winded transport paths of the conducting phase is described by the mean geodesic tortuosity τ_{geod}, which is defined as the ratio of the expected shortest path lengths from inlet- to outlet-plane over the material thickness. Thereby, the shortest path lengths (in terms of geodesic distance)[22] in transport direction from inlet- to outlet-planes are computed within the voxel space that represents the transporting phase (see the left-hand side of Figure 4). To determine τ_{geod}, we consider an average of geodesic tortuosities computed for all voxels of the transporting phase in the inlet-plane. Obviously, it holds that $\tau_{\text{geod}} \geq 1$ and higher values of τ_{geod} indicate more winded pathways.

Besides the windedness of transport paths through the material, narrow constrictions of the conducting phase, quantified by the so-called constrictivity β, have a strong influence on σ_{eff}. Constrictivity is defined as

$$\beta = \left(\frac{r_{\min}}{r_{\max}}\right)^2, \tag{2}$$

where, heuristically speaking, $r_{\min}$ indicates the radius of the characteristic bottleneck and $r_{\max}$ indicates the radius of the characteristic bulge, see the right-hand side of Figure 4. More precisely, $r_{\max}$ is the 50% quantile of the continuous pore size distribution (c-PSD) and $r_{\min}$ is the 50% quantile of the MIP pore size distribution, which is based on a geometrical simulation of mercury intrusion porosimetry[23] (MIP). Constrictivity takes values between 0 and 1, where values close to 0 indicate strong bottleneck effects while values close to 1 indicate that there are no bottlenecks at all.

For details regarding these structural characteristics and their estimation from 3-D image data, the reader is referred to previous work.[11] The 8119 microstructures generated by the aid of SSGM and MSM cover a wide range of values for the characteristics ε, τ_{geod}, and β, see the Section 'Characteristics of simulated virtual 3D microstructures'.

Conductive transport

We consider conductive transport processes within composite materials, where only one phase is conducting.[10,11] The electric charge transport is described by Ohm's law

$$J = -\sigma \frac{dU}{dx} \tag{3}$$

and

$$\frac{dU}{dt} = \sigma \frac{d^2U}{dx^2}, \tag{4}$$

where J is the current density, σ is the conductivity, U is the electric potential, and t is time. Assuming constant boundary conditions, such systems converge to an equilibrium which is described by the Laplace equation

$$\frac{d^2U}{dx^2} + \frac{d^2U}{dy^2} + \frac{d^2U}{dz^2} = 0, \tag{5}$$

where x, y, and z denote the coordinates in the 3-D Euclidean space.

As transport only takes place in one phase, the geometry of the microstructure reduces the intrinsic conductivity σ_0 of the material to the effective conductivity σ_{eff}, that is

$$\sigma_{\text{eff}} = \sigma_0 M \tag{6}$$

for some $0 \leq M \leq 1$. The influence of the microstructure on the effective conductivity is described by the factor M. Our goal is to validate the prediction of the M-factor based on the geometric characteristics ε, τ_{geod}, and β, which has already been derived.[11] Moreover, we improve the prediction formula

using methods from statistical learning, that is, by neural networks and random forests. For each of the 8119 synthetic microstructures, the effective conductivity and the associated M-factor are determined by numerical simulation using the software GeoDict.[12]

Statistical learning

Neural networks and random forests are two methods from statistical learning that can be used for nonlinear regression.[24] Both methods are used to predict the M-factor of virtual 3-D microstructures by the corresponding values of ε, τ_{geod}, and β. We give a short description of neural networks and random forests. In both cases, an output variable $Y \in \mathbb{R}$ is predicted by an input vector $X \in \mathbb{R}^p$ consisting of p features, where $p \in \mathbb{N}$. In our case, we have $X = (\varepsilon, \tau_{\mathrm{geod}}, \beta)$ and $Y = \log_2(M)$. As the computed M-factors vary over several orders of magnitude, a better fit is obtained by putting the M-factors on a $\log_2$-scale.

Neural networks are two-stage regression models. Here, we use a single hidden layer network. For prediction of Y, the vector X is mapped to the hidden layer, which is a vector $Z \in \mathbb{R}^L, L \in \mathbb{N}$, where for each $l \in \{1, \ldots, L\}$ we have $Z_l = \sigma \left(\alpha_{0,l} + \sum_{i=1}^p \alpha_{i,l} X_i \right)$ for a parameter matrix $\alpha = (\alpha_{i,j}) \in \mathbb{R}^{(p+1) \times L}$ and some function $\sigma : \mathbb{R} \to \mathbb{R}$. Here, we choose σ as the sigmoid function, that is, $\sigma(t) = (1 + e^{-t})^{-1}$ for each $t \in \mathbb{R}$. The predictor $\hat{Y}$ of Y is finally constructed by a linear combination of the entries of Z, to be more precise $\hat{Y} = \min \{ \max \{ \hat{Y}^*, 0 \}, 1 \}$ with

$$\hat{Y}^* = \theta_0 + \sum_{i=1}^L \theta_i Z_i \tag{7}$$

for some parameter vector $\theta \in \mathbb{R}^{L+1}$. To fit the parameters α and θ, we minimize the mean squared error (MSE) between $\hat{Y}^*$ and Y by the Matlab implementation[25] of the Levenberg-Marquardt backpropagation algorithm,[26] where the initial values are determined by the Nguyen-Widrow algorithm.[27] During the fitting procedure, data is divided completely at random into training data (70%), validation data (15%), and test data (15%). Training data is directly used to fit α and θ, whereas validation data is used to define a stopping criterium for the Levenberg-Marquardt algorithm.[24] The dimension L of the hidden layer is chosen such that the mean absolute percentage error (MAPE) of test data is minimized. For this purpose, we average over 200 random subdivisions, where data is divided into training data, validation data, and test data. Altogether, the described division of data avoids overfitting by the neural network.

Random forests[28] are further regression models from statistical learning, which are based on so-called regression trees. The predictor $\hat{Y}$ of Y obtained by a single regression tree is a linear combination of indicators, that is

$$\hat{Y} = \sum_{m=1}^M c_m \mathbb{1}_{x \in R_m} \tag{8}$$

for an appropriate partition $\mathcal{R} = \{R_1, \ldots, R_M\}$ of $\mathbb{R}^p$, where $\mathbb{1}_{x \in R} = 1$ if $x \in R$ and $\mathbb{1}_{x \in R} = 0$ otherwise for each $R \subset \mathbb{R}^p$. Beginning with $\mathcal{R} = \{\mathbb{R}^p\}$ the partition $\mathcal{R}$ is refined iteratively. In each iteration, all regions are split into two half-spaces such that by an optimal choice of coefficients c_m the MSE can be minimized. The refinement is stopped when each region contains a predefined minimum number of observations of X. For our purpose this minimum number is set to five as recommended in the literature.[24] In random forests averaging over randomized regression trees improves the prediction. Randomization takes place in two different ways. To fit the individual regression trees, different random subsets of the input vector are chosen. Moreover, $k < p$ features of X, denoted by $i_1, \ldots, i_k$, are chosen at random for each splitting of a region. Then, splitting is only possible along one of the axes $i_1, \ldots, i_k$. Usually $k = \sqrt{p}$ is used. This procedure allows a variance reduction of the predictor $\hat{Y}$, caused by averaging of single regression trees as well as by the described randomization. For prediction of the M-factor, we choose $k = 2$. Similar to neural networks we divide data into training data (70%) and test data (30%) completely at random. The number of trees used for averaging is chosen such that the MSE of test data does not decrease significantly for a larger number of trees. As in the case of neural networks, we consider 200 random subdivisions to determine the number of trees. In order to fit and simulate random forests, we use the random Forest-package[29] of the statistical software R.[30]

Results and Discussion

Simulations, which are based on the stochastic models presented in the Section 'Stochastic microstructure modeling' provide 8119 virtual 3-D microstructures. For each of these virtual 3-D microstructures, we compute the geometric microstructure characteristics ε, τ_{geod}, and β as well as the corresponding M-factor.

Characteristics of simulated virtual 3-D microstructures

Figure 5 shows that the generated virtual 3-D microstructures cover a wide range of constellations for ε, τ_{geod}, and β. For small values of ε, many microstructures are generated, the transport paths of which are more than 1.5 longer than the
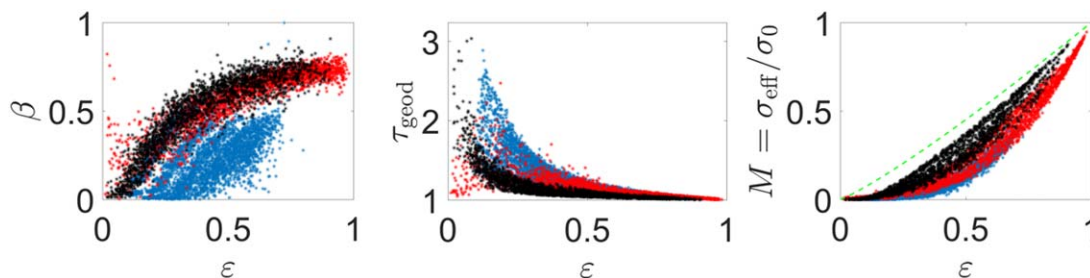

Figure 5. Characteristics of the 8119 virtual 3-D microstructures generated by the SSGM (blue), the MSM (red) and the complement of the MSM (black).

The plots show mean geodesic tortuosity τ_{geod} (left), constrictivity β (center) and M-factor versus volume fraction ε. [Color figure can be viewed at wileyonlinelibrary.com]
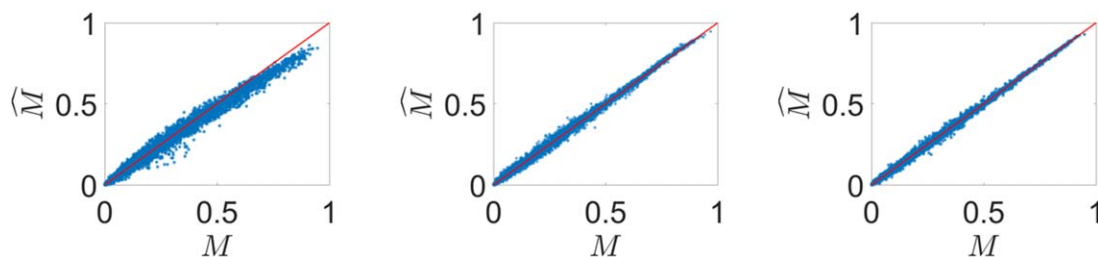
Figure 6. Scatter plots of computed and predicted *M*-factors using the prediction formula from Eq. 1 (left), neural networks (center) and random forests (right), where the identity function is added in each plot (red lines).

[Color figure can be viewed at wileyonlinelibrary.com]

materials thickness, that is $\tau_{\mathrm{geod}} \geq 1.5$. The unflexibility of the models regarding τ_{geod} for large values of ε is not surprising, as—excluding pathological counterexamples—the mean length of transport paths through the material decreases strongly with increasing volume fractions.

Most values of β are in the interval $[0, 0.8]$ and for virtual microstructures with $\varepsilon \in [0.4, 0.7]$ the corresponding constrictivities take nearly all values between 0.05 and 0.7. While higher values of constrictivity are observed in virtual microstructures generated by the MSM (blue and red dots in Figure 5), the correlation between ε and β is less strong in the SSGM (black dots in Figure 5). The SSGM was especially developed for varying the considered microstructure characteristics as independently as possible.[10]

The right-hand side of Figure 5 shows that the simulated *M*-factors of the virtual 3-D microstructures cover the whole range between 0 and 1. All *M*-factors are below the upper bound $M \leq \varepsilon^{1.15}$ (green dashed line) resulting from the empirically derived prediction formula, see Eq. 1. Note that a rigorous upper bound for M is given[3] by $M \leq \varepsilon$.

Prediction of M-factor by geometric microstructure characteristics

On the basis of the simulated microstructures we validate the prediction formula given in Eq. 1. Furthermore, we present the predictions obtained by neural networks and random forests, which are fitted to simulated data as it is described in the Section 'Statistical learning'. Figure 6 shows scatter plots of computed and predicted *M*-factors, while the MAPE as well as the coefficient of determination R^2 are listed in Table 1.

For the prediction formula given in Eq. 1, the MAPE corresponding to the 8119 virtual microstructures is 13.6%, while the MAPE was 19.6% for the virtual microstructures, which have been used to derive Eq. 1. The reason for this smaller value of MAPE is that the microstructures analyzed in this article are less extreme, that is, they have a larger average *M*-factor. Altogether, the formula given in Eq. 1 offers a good

prediction of the *M*-factor, see Figure 6 (left), which is also indicated by a high coefficient of determination R^2. However, the formula seems to systematically underestimate the *M*-factor for values above 0.7, that is, for materials with a high-volume fraction. Note that fitting the exponents in Eq. 1 to all 8119 cannot solve this underestimation. Just a slight improvement leading to a MAPE of 12.7% can be obtained and the formula would change to

$$\sigma_{\mathrm{eff}} = \sigma_0 \, \frac{\varepsilon^{1.22} \, \beta^{0.31}}{\tau_{\mathrm{geod}}^{4.7}}. \tag{9}$$

Due to this small change, which occurs, when using all 8119 microstructures for fitting the exponents, there is no need to replace Eq. 1 by Eq. 9 from our point of view.

Using neural networks and random forests the prediction of the *M*-factor can be improved. Fitting a single hidden layer neural network leads to a hidden layer of size $L = 20$. The MAPE for the test data is 8.94%, while $R^2 = 0.997$. For prediction by a random forest we average over 500 trees and obtain a MAPE of 8.47%, which is slightly better than prediction by neural networks. Also $R^2 = 0.999$ shows a better prediction by random forests. Note that random forests, in contrast to neural networks, have a much smaller MAPE for training data than test data, see Table 1. Matlab- and R-code is provided to use the trained neural network and the trained random forest for prediction of the *M*-factor, see the Section 'Supplementary information'.

Random forests and neural networks offer a much lower prediction error than the formula given by Eq. 1, see Table 1. Thus, for prediction purposes, random forests, or neural networks should be used from our point of view. Both methods are equivalent in terms of their prediction accuracy. However, random forests and neural networks are extremely difficult if not impossible to interpret. Thus, it is difficult to explain why a microstructure has a certain *M*-factor. The big advantage of the prediction formula from Eq. 1 is that it allows us to explain how ε, τ_{geod}, and β influence the *M*-factor. In short, we propose to use neural networks and random forests for prediction and Eq. 1 for explanation.

Using neural networks or random forests, the MAPE of test data is smaller than 9%. This means that the considered volume-averaged characteristics ε, τ_{geod}, and β carry significant information about effective conductivity, but certainly not all information. One possibility to further improve the prediction accuracy would be to consider the active volume fraction instead of the connected volume fraction. Imagine a microstructure that is completely connected yet has many dead-ends which are not used for transport. Then considering active volume

Table 1. MAPE and Coefficient of Determination R^2 of the Different Prediction Models

Model	MAPE (Training Data)	MAPE (Test Data)	R^2
Prediction formula	–	13.6%	0.984
Neural network	8.20%	8.94%	0.997
Random forest	3.99%	8.47%	0.999

Note that the prediction formula from Eq. 1 was not fitted to the data simulated in this study. Thus, the complete data can be considered as test data in this case.
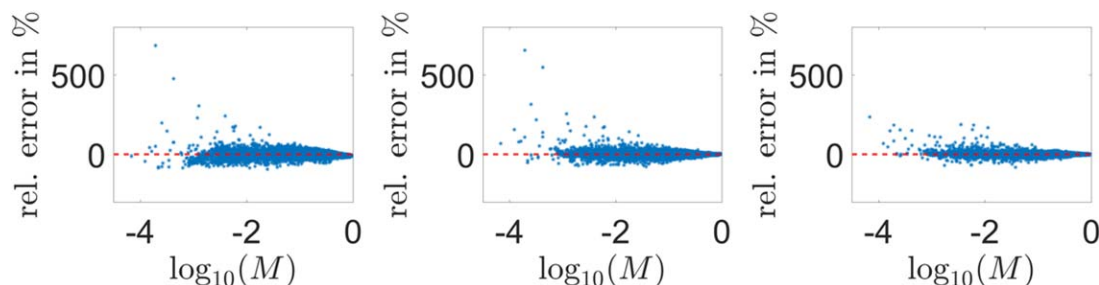
Figure 7. Computed *M*-factors on a log₁₀-scale versus relative prediction errors. Predictions are obtained by the prediction formula from Eq. 1 (left), neural networks (center) and random forests (right).

[Color figure can be viewed at wileyonlinelibrary.com]

(connected volume minus "dead-end"-volume) instead of connected volume should further increase prediction accuracy. The precise mathematical definition and computation of active volume, however, is challenging and subject of current research. We are not aware of any definition for active volume in the literature, which is based on the geometry of the underlying microstructure.

Considering Figure 6, it seems that all three prediction models work well for all microstructures without any exceptions. However, for all three methods, the prediction error increases for decreasing *M*-factors and extreme errors occur for very small *M*-factors (below 10^{-2}), see Figure 7. Note that the errors are less extreme when random forests are used for prediction of the *M*-factor.

Interestingly, all extreme errors overestimate the *M*-factor, that is, the corresponding microstructures have a smaller *M*-factor than predicted. These extreme deviations are caused by microstructures, which are close to their percolation threshold, that is, eroding the microstructure a little bit would eliminate connectivity. The microstructures have a low connectivity and much of the volume is not used for transport ("dead-end" volume). Measuring active volume instead of connected volume could lead to a better prediction of the *M*-factor.

Validation with experimental microstructures

To validate our method, we compare *M*-factors predicted by the three different methods with computed *M*-factors (using
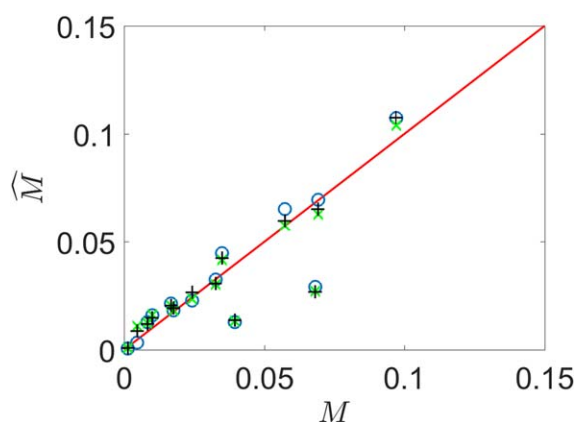

Figure 8. Computed *M*-factors and the corresponding predictions $\widehat{M}$ for experimental image data. Predictions have been performed by Eq. 1 (blue circles), by neural networks (green crosses) and random forests (black plus signs).

[Color figure can be viewed at wileyonlinelibrary.com]

GeoDict) for different 3-D image data obtained by FIB-SEM tomography. For this purpose, the same datasets are considered, which have also been used for validation[11] of Eq. 1. In total we have 10 images, where six of them representing anodes in solid oxide fuel cells (SOFC) consisting of pores, nickel (Ni) and yttrium-stabilized zirconia (YSZ)[31] and four of them represent porous membranes used as liquid junctions in pH-Sensors.[32] In the SOFC anodes electric conduction takes place in the Ni phase and ionic conduction in the YSZ phase, while liquid electrolyte diffusion (classical Fickian diffusion) occurs in the pores of the membranes of pH-Sensors. More information about the experimental data can be found in a previous publication.[11]

As the classical Fickian diffusion converges also to an equilibrium described by the Laplace equation, given in Eq. 5, effective diffusivity is simulated in the same way as effective conductivity. Thus, the definition of the *M*-factor for diffusion processes is analogous to the *M*-factor for conduction processes in this application. In Figure 8, the *M*-factors computed by numerical simulation on the image datasets are compared to the predictions by Eq. 1, neural networks and random forests. In general, the prediction fits the simulated *M*-factors nicely, where the results obtained from statistical learning are slightly worse than those obtained by the prediction formula. The MAPE is 28.0% for the prediction formula, 33.8% for the neural network and 30.3% for the random forest. However, note that only 16 values of the *M*-factors are considered. Thus, there is no need to withdraw the conclusion from the Section 'Prediction of *M*-factor by geometric microstructure characteristics' based on more than 8000 virtual microstructures, which is that methods from statistical learning improve the prediction of the *M*-factor by ε, τ_{geod}, and β.

Figure 8 shows two outliers which can be explained as follows: The two data points represent electric conductivity in the Ni phase of SOFC anodes that were exposed to harsh conditions, which led to strong microstructure alteration (i.e., Ni-agglomeration). It was shown that due to the strong alteration, the representative volume is much larger than the observation window that can be obtained by FIB-tomography.[11,31] Therefore, the analyses based on these two 3-D datasets suffer from a high uncertainty. For all other data points the predictions are reasonably well. From the validation with experimental microstructures, we can conclude that the stochastic models are realistic enough to use them to derive predictors for effective conductivity.

Conclusion

In this article, we investigate microstructure-property relationships for conductive transport processes using 8119 virtual microstructures generated by SMM. Effective conductivity is

predicted by the three microstructure characteristics volume fraction ε, mean geodesic tortuosity τ_{geod} and constrictivity β. The interpretable prediction formula, given in Eq. 1, yields a prediction error of 13.6%, which can be considered as a further validation of this prediction formula since only 43 virtual microstructures have been used to derive it. Random forests and neural networks which are difficult to interpret yield smaller prediction errors of less than 9%, where in all cases the prediction becomes unstable for microstructures at their percolation threshold.

Validation with experimental microstructures shows that the generated virtual microstructures are sufficiently realistic to derive prediction models for effective conductivity. Overall, this article points out that the combination of stochastic microstructure modeling with physical computations and data mining techniques is a powerful tool to establish quantitative microstructure-property relationships. These relationships enable the identification of improved microstructures with respect to effective conductivity.

Due to physical arguments, this article is based on the assumption that ε, τ_{geod}, and β are the most relevant transport characteristics. This assumption can be tested using the large dataset of 8119 virtual microstructures. Computing many microstructure characteristics and ranking them according to their impact on effective conductivity, as it was done in a recent study on transport through complex networks,[33] one can find out if ε, τ_{geod}, and β are the most important microstructure characteristics. Moreover, one can even think of extracting the most important microstructure characteristics from 3-D image data using statistical learning for feature selection.

The method of VMT itself is not restricted to conduction processes and can also be used to investigate relationships between microstructure characteristics and other functional properties, for example, effective permeability or mechanical stress-strain curves.

Supplementary material

For supplementary information, see http://onlinelibrary.wiley.com/doi/10.1002/aic.15757/full. The fitted neural network (aic15757-sup-0004-suppinfo4.mat) as well as the fitted random forest (aic15757-sup-0002-suppinfo2.RData) are provided as supplementary material. Additional code is provided, which can be used to predict the M-factor for given volume fraction, mean geodesic tortuosity and constrictivity by the neural network (aic15757-sup-0003-suppinfo3.m) and the random forest (aic15757-sup-0001-suppinfo1.R).

Acknowledgments

References

1. White A. Big data are shaping the future of materials science. *MRS Bull.* 2013;38:594–595.
2. Kalidindi SR, De Graef M. Materials data science: current status and future outlook. *Ann Rev Mater Res.* 2015;45:171–193.
3. Torquato S. *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*. New York: Springer, 2013.
4. Skibindi J, Cwieka K, Kowalkowski T, Wysocki B, Wejrzanowski T, Kurzydlowski KJ. The influence of pore size variation on the pressure drop in open-cell foams. *Mater Design.* 2015;87:650–655.
5. Wejrzanowski T, Grybczuk M, Chmielewski M, Pietrzak K, Kurzydlowski KJ, Strojny-Nedza A. Thermal conductivity of metal-graphene composites. *Mater Design.* 2016;99:163–173.
6. Doyle M, Fuller TF, Newman J. Modeling of galvanostatic charge and discharge of the lithium/polymer/insertion cell. *J Electrochem Soc.* 1993;140(6):1526–1533.
7. Holzer L, Wiedenmann D, Münch B, Keller L, Prestat M, Gasser P, Robertson I, Grobéty B. The influence of constrictivity on the effective transport properties of porous layers in electrolysis and fuel cells. *J Mater Sci.* 2013;48:2934–2952.
8. Shikazono N, Kanno D, Matsuzaki K, Teshima H, Sumino S, Kasagi N. Numerical assessment of SOFC anode polarization based on three-dimensional model microstructure reconstructed from FIB-SEM images. *J Electrochem Soc.* 2010;157(5):B665–B672.
9. Tippmann S, Walper D, Balboa L, Spier B, Bessler WG. Low-temperature charging of lithium-ion cells part I: electrochemical modeling and experimental investigation of degradation behavior. *J Power Sources.* 2014;252:305–316.
10. Gaiselmann G, Neumann M, Pecho OM, Hocker T, Schmidt V, Holzer L. Quantitative relationships between microstructure and effective transport properties based on virtual materials testing. *AIChE J.* 2014;60(6):1983–1999.
11. Stenzel O, Pecho OM, Holzer L, Neumann M, Schmidt V. Predicting effective conductivities based on geometric microstructure characteristics. *AIChE J.* 2016;62:1834–1843.
12. Becker J, Glatt E, Wiegmann E. GeoDict. Available at: www.geodict.com. 2012.
13. Becker J, Flückiger R, Reum M, Büchi FN, Marone F, Stampanoni M. Determination of material properties of gas diffusion layers: experiments and simulations using phase contrast tomographic microscopy. *J Electrochem Soc.* 2009;156(10):B1175–B1181.
14. Stenzel O, Koster L, Thiedmann R, Oosterhout SD, Janssen RAJ, Schmidt V. A new approach to model-based simulation of disordered polymer blend solar cells. *Adv Func Mater.* 2012;22(6):1236–1244.
15. Neumann M, Schmidt V. Stochastic 3D modeling of amorphous microstructures - a powerful tool for virtual materials testing. *Proceedings of the VII European Congress on Computational Methods in Applied Sciences and Engineering.* Paper-ID 8172, Crete. 2016.
16. Chiu SN, Stoyan D, Kendall WS, Mecke J. *Stochastic Geometry and its Applications*, 3rd ed. Chichester: Wiley, 2013.
17. Molchanov I. *Statistics of the Boolean Model for Practitioners and Mathematicians*. Chichester: Wiley, 1997.
18. Scholz C, Wirner F, Klatt MA, Hirneise D, Schröder-Turk GE, Mecke K, Bechinger C. Direct relations between morphology and transport in Boolean models. *Phys Rev E.* 2015;92(4):043023.
19. Stenzel O, Hassfeld H, Thiedmann R, Koster LJA, Oosterhout SD, van Bavel SS, Wienk MM, Loos J, Janssen RAJ, Schmidt V. Spatial modeling of the 3D morphology of hybrid polymer-ZnO solar cells, based on electron tomography data. *Ann Appl Stat.* 2011;5:1920–1947.
20. Holzer L, Cantoni M. Review of FIB-tomography. In: Utke I, Moshkalev S, Russel P, editors. *Nanofabrication Using Focused Ion and Electron Beams: Principles and Applications*, New York: Oxford University Press, 2012:410–435.
21. Petersen EE. Diffusion in a pore of varying cross section. *AIChE J.* 1958;4(3):343–345.
22. Soille P. *Morphological Image Analysis: Principles and Applications*. New York: Springer, 2003.
23. Münch B, Holzer L. Contradicting geometrical concepts in pore size analysis attained with electron microscopy and mercury intrusion. *J Am Ceramic Soc.* 2008;91(12):4059–4067.
24. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2008.
25. MATLAB 2015b, The MathWorks. Available at: www.mathworks.com, 2015.
26. Hagan MT, Menhaj MB. Training feedforward networks with the Marquardt algorithm. *IEEE Trans Neural Netw.* 1994;5(6):989–993.
27. Nguyen D, Widrow B. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In: *IJCNN International Joint Conference on Neural Networks*. IEEE, San Diego. 1990:21–26.
28. Breiman L. *Random Forests. Machine Learning.* 2001;45(1):5–32.
29. Liaw A, Wiener M. Classification and regression by randomForest. *R News.* 2002;2(3):18–22.
30. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. 2015.

31. Pecho OM, Stenzel O, Gasser P, Neumann M, Schmidt V, Hocker T, Flatt RJ, Holzer L. 3D microstructure effects in Ni-YSZ anodes: prediction of effective transport properties and optimization of redox-stability. *Materials.* 2015;8(9):5554–5585.

32. Holzer L, Stenzel O, Pecho OM, Ott T, Boiger G, Gorbar M, De Hazan Y, Penner D, Schneider I, Cervera R, Gasser P. Fundamental relationships between 3D pore topology, electrolyte conduction and flow properties: towards knowledge-based design of ceramic diaphragms for sensor applications. *Mater Design.* 2016;99:314–327.

33. Van der Linden JH, Narsilio GA, Tordesillas A. Machine learning framework for analysis of transport through complex networks in porous, granular media: a focus on permeability. *Phys Rev E.* 2016; 94(2):022904.