Statistical Inference Over Persistent Homology Predicts Fluid Flow in Porous Media

Chul Moon[1], Scott A. Mitchell[2], Jason E. Heath[2], and Matthew Andrew[3]

[1]Department of Statistical Science, Southern Methodist University, Dallas, Texas, USA.
[2]Sandia National Laboratories, Albuquerque, New Mexico, USA.
[3]Carl Zeiss X-ray Microscopy Inc., Dublin, California, USA.

Key Points:

- Statistical inference of persistent homology over 3D rock images predicts constitutive behaviors.
- Principal component analysis and a penalized regression model computes structural characteristics of sample volumes.
- Output is consistent with established sREV sizes, and fluid flow and transport properties, but requires low computational cost.

Corresponding author: Chul Moon, `chulm@smu.edu`

Abstract

We statistically infer fluid flow and transport properties of porous materials based on their geometry and connectivity, without the need for detailed material properties and expensive physical simulations. Our predictions are consistent with traditional approaches. We summarize structure by persistent homology, then determines the similarity of structures using image analysis and statistics. Longer term, this may enable quick and automated categorization of rocks into known archetypes. We first compute persistent homology of binarized 3D images of material subvolume samples. The persistence parameter is the signed Euclidean distance from inferred material interfaces, which captures the distribution of sizes of pores and grains. Each persistence diagram is converted into an image vector. We infer structural similarity by calculating image similarity. For each image vector, we compute principal components to extract features. We fit statistical models to features estimates material permeability, tortuosity, and anisotropy. We develop a Structural SIMilarity index (SSIM) to determine Statistical Representative Elementary Volumes (sREV).

1 Introduction

Algebraic topology offers powerful mathematical tools for describing the connectivity of a space, and how the connectivity varies in response to parameter changes. It converts local shapes to global connectivity properties in a concrete and measurable way. The hope is that we can predict porous material properties from an algebraic topology description of pore geometry and topology. Materials have been analyzed before using the Euler characteristics χ (Scholz et al., 2012; Andrew, 2018a). Persistent homology analysis (Robins et al., 2016) may be more accurate, because it contains more information. Homology is described by Betti numbers, β_i, one per dimension. Betti numbers are equal to the numbers of connected components, loops, shells, etc. In contrast, the Euler characteristic is merely a check-sum on homology, the alternating sum of Betti numbers: $\chi = \beta_0 - \beta_1 + \beta_2 - \beta_3 + \cdots$.

Persistent homology goes further and summarizes homology dynamics as a function of the chosen *persistence parameter* (metric) (Hatcher, 2002; Edelsbrunner & Harer, 2008; Ghrist, 2008). As the parameter is increased, data become more connected, and features appear and disappear. One has the freedom to choose the metric to illuminate the properties of interest. It may be abstract, and need not be an actual time-varying physical parameter such as pressure. Indeed, it is usually chosen to depend on the geometric distance between data; in our case, we use the signed distance to a material interface. This captures pore and grain sizes. Varying this parameter is analogous to growing or shrinking the material. At larger values, pore throats close and pores disappear. At smaller values pores connect and grains disappear. The dynamics of a topological feature is quantified by the parameter values at which the feature first occurs and finally disappears: the interval between its "birth" and "death."

Persistent homology has been used to quantify topology changes and predict the physical behavior of porous materials (Robins et al., 2011). Herein we go further by combining this approach with classical Statistical/Machine Learning (SML) methods to make even better predictions. However, we need the additional step of transforming persistent data into a form that SML can use. Persistence data depends on dataset size; this dependence is poorly understood. Traditional methods use the concept of a statistical Representative Elementary Volume (REV) (Bear, 1972). The REV is the smallest size of a material sample above which selected property calculations are stable. For porous media, REVs for porosity, permeability, and tortuosity provides the subsample scale at which continuum methods accurately predict bulk fluid flow and transport. The REV enables the complex pore microstructure of rock to be replaced by a fictitious continuum, so partial differential equations can calculate properties. In an analogous way, we develop a

Statistical REV (sREV), the minimum subsample size for which our statistical machine learning methods produce stable and accurate predictions.

1.1 Rock Image Input

We have three types of data; see Table 1. First, Focused-Ion-Beam Scanning-Electron-Microscopy (FIB-SEM) examined the Selma Chalk, a microporous carbonate rock. Second, X-Ray Microscopy (XRM) examined the intergranular pore structures in a series of sandstones: Bentheimer, Doddington, and Rotleigend (Moon & Andrew, 2019b). Third, simulation of elastic deformation of the Bentheimer sandstone created a series of compressed networks (Moon & Andrew, 2019a).

1. Selma Chalk data comes from the binarized images of Yoon and Dewers (2013). We also use their porosity, permeability, tortuosity and anisotropy values by subvolume size.
2. Sandstones were imaged by a ZEISS Versa X-ray Microscope (XRM) at ZEISS X-Ray Microscopy, Pleasanton, CA. We acquired a series of 1600 2D projections at regular angular increments over a 360° rotation. These reconstructed a series of 3D volumes: cylinders which were then cropped to the maximally inscribed cubes. 2D projections were 1024^2 pixels, and 3D volumes are 1024^3 voxels. Images were segmented by ZEISS Zen Intellesis machine learning (Andrew, 2018b).
3. The simulated Bentheimer network was deformed using the algorithm of Rutka and Wiegmann (2006), using target strains of 2%, 4%, 6%, 8%, 10%, 20%, 30%, 40% and 50%.

Table 1. Porous material data.

Rock Type	Modality	Resolution (m)	Sample	Subsamples
Selma Chalk	FIB-SEM	15.6×10^{-9}	$930 \times 520 \times 962$	150^3, 300^3, 400^3, 500^3, $600 \times 520 \times 600$, $765 \times 520 \times 962$
Bentheimer	XRM	8.9×10^{-6}	1024^3	$700 \times 600 \times 700$
Doddington	XRM	5.4×10^{-6}	1024^3	700^3
Rotleigend	XRM	2.0×10^{-6}	1024^3	$700 \times 700 \times 980$
Bentheimer	simulation	8.9×10^{-6}	1024^3	$700 \times 600 \times 700$

2 Analysis Pipeline

1. Data acquisition and preprocessing
   - original images → binary images
2. Persistent homology and transformation for SML
   - binary images → grayscale images → cubical complexes → persistence diagrams → vectorized persistence diagrams
3. sREV estimation for single dataset
   - Select subvolumes of various sizes. For each perform step 2.
   - vectorized persistence diagrams → similarity metric → determine sREV
4. Property estimation and dataset comparison
   - Feature extraction:

vectorized persistence diagrams → principal component analysis → weights

- weights → fit a penalized regression model → property prediction
- weights → fit SML models → clustering/classification

2.1 Persistent Homology Computation

We follow the persistent homology framework of Robins et al. (2011, 2016). The persistence metric is the Signed Euclidean Distance Transform (SEDT). It assigns a numeric value to each point: negative for pore and positive for grain, with magnitude equal to the Euclidean distance to the closest point of opposite sign. A large negative value indicates a large pore size, and a large positive value indicates a large grain size.

The reader can gain some intuition about persistent homology by starting with nothing, then adding (groups of) voxels in order of increasing SEDT, at each stage computing the homology of the selected voxels. An example of connectivity changes is shown in Figure 1. The voxels at the middle of the large pore appear first. As the SEDT threshold value becomes positive, voxels in the grain phase are added.

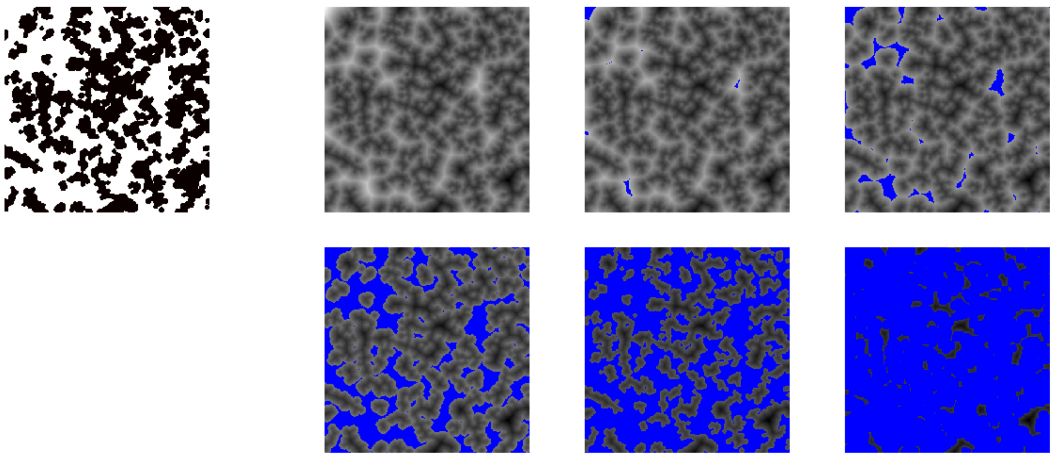

Figure 1. Left, a binary image of pores (white) and grains (black). Right, a grayscale SEDT for a pore network and a sequence of five cell complexes. White and black represent the smallest and the largest SEDT values. The complex at the current persistent homology filtration value consists of the blue pixels.

However, we do not compute persistent homology over the voxels directly, but define a cubical cell complex. The voxels are added to the cell complex as vertices (0-cells) when the filtration reaches the corresponding SEDT value. When adjacent voxels are added, then it constructs higher dimensional cells such as edges (1-cells), faces (2-cells), and cubes (3-cells).We add *cells* in order of increasing value, where the value of a *cell* is the maximum SEDT of its vertices.

Further detail is that it is sufficient to track critical points of the discrete Morse function. This reduces the number of cells we must consider and significantly reduces the computational burden, since the complexity of the standard algorithm is cubic in the number of cells. Critical points include local minima at some 0-cells, local maxima at some 3-cells, and saddle points at some 1-cells and 2-cells.

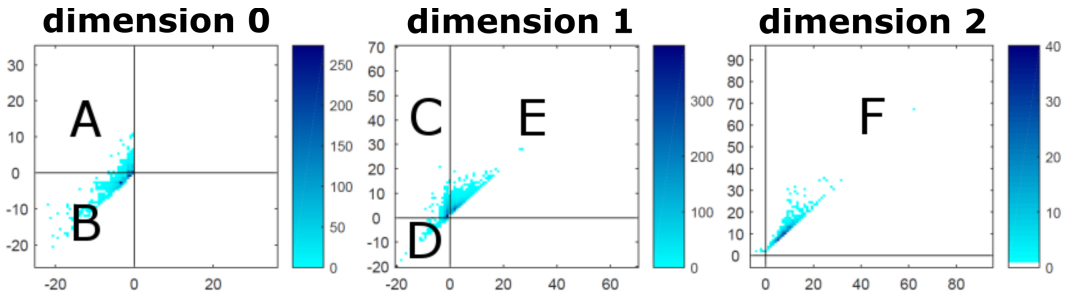

Figure 2. Dimension 0, 1, and 2 persistence diagrams of the Selma Chalk data. Axes are Euclidean distance, with the conventional scale of a voxel side = 1 unit, here 15.6×10^{-9} m.

Persistent homology is computed by updating the homology as cells are added, rather than recomputing from scratch. We use the software "Diamorse" (Delgado-Friedrichs, 2015).

Results for each dimension are plotted separately. Figure 2.1 shows persistence diagrams for the Selma Chalk and Table 2 interprets the lettered quadrants. For each dimension's persistence diagram, the different quadrants reveal the distributions of different material aspects (Robins et al., 2011, 2016). Visual examples of these structures are given in Figure S1 in the Supplementary Material.

Table 2. Interpretation of persistence diagrams. Pore (grain) size is defined as the largest radius of a sphere that can fit inside the pore (grain). Pore throat radius is the maximum radius of a sphere that can move between two pores. Tube radius is the largest radius of a sphere that can be placed within the tube, an extrema of the SEDT within the pore (grain).

Dim.	Quad	Structure	Value of X (Birth)	Value of Y (Death)
0	A	disconnected pore	size of pore	narrowest grain contact
	B	connected pore	size of pore	pore throat radius
1	C	contact grain	pore tube radius	grain contact radius
	D	non-convex pore	pore tube radius	non-convex pore throat radius
	E	non-convex grain	grain tube radius	non-convex grain throat radius
2	F	grain	grain-contact radius	size of grain

2.2 Vectorized Persistence Diagram for Statistical Learning

Persistence homology results are sets of [Birth, Death] intervals. They are typically summarized in persistence diagrams; see Figure 3. Statistical learning methods cannot be applied directly to the intervals or either type of graphic; the data must first be transformed into canonical form. The main issues are that intervals are a non-classical data type, and the number of intervals generated varies from dataset to dataset. The SML methods we choose rely on feature-vectors having identical numbers of features.

Vectorized persistence diagrams provide the needed transformation. Besides enabling image analysis and statistical learning methods, vectorization allows us to compute mean persistence diagrams. A disadvantage is that persistence comparisons will not be exact, unlike using Wasserstein or bottleneck distances, for example.

There are several known ways to vectorize persistence diagrams, including binning (Bendich et al., 2016), persistence images (Adams et al., 2017), and kernel methods (Reininghaus et al., 2015; Kusano et al., 2016). However, none of the prior approaches are perfect for our context for the following two reasons. First, some vectorizations transform [birth, death] to (birth, death−birth). This emphasizes features far from the 45-degree line, whereas we seek a uniform assessment of the distribution. Worse, it maps quadrants into differently shaped regions, obscuring the geological interpretation in Table 2. Second, some vectorizations smooth the persistence diagram, e.g., Adams et al. (2017) creates a smooth surface over the data before transforming it into a vector. In their context, this improves robustness, but in our context it blurs the geological distinction between quadrants. As an example, in Figure 2.1, smoothing can make non-convex *pores* in region D affect non-convex *grains* in region E. Smoothing persistence for porous materials is unstudied, as are the effects of noise in the original images and the binarization filters.

Thus we use a new binning without transformation or smoothing. We use bins that are one unit wide, which start and end at integer values. This preserves quadrants. We use a square range, $m \times m$ bins. The range is scaled to the data: the first bin starts at *floor*(min data point), and the last bin ends at *ceiling*(max data point), where the min and max are taken over all barcodes we wish to compare. The value of a bin is the number of barcode data points in that bin. (The lower-diagonal bins are always empty and may be discarded.) We convert the array of values into a vector in column-major order: first column top to bottom, second column top to bottom, etc. We found this sufficient, but, for other contexts, alternatives that better represent proximity might be advantageous; e.g., scan in order of a space-filling curve (Bader, 2012). Figure 3 illustrates the vectorization process.

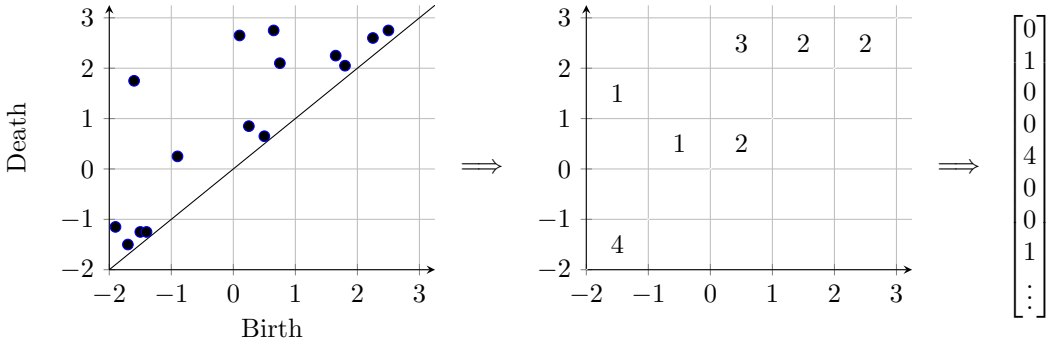

Figure 3. A small illustrative example of vectorization. Left, persistence diagram divided into 5×5 bins. Middle, numeric bins. Right, numeric vector, from column-major order of bins.

2.3 Determining sREV using Persistence Homology

The three-dimensional material images are expensive data to obtain. For accuracy, we need a sufficiently large subsample; but for efficiency, we do not want to use a larger subsample than necessary. Persistence diagrams depend on subsample size, but how they depend is poorly understood, and barcodes may change unpredictably as size increases. Traditional statistical methods use the concept of a representative elementary volume (REV) (Bear, 1972). The REV is the scale at which smaller-scale fluctuations dampen out, and statistically stable properties can be defined. This scale depends on the medium, the property, and the calculation technique. For porous media, we have REVs for porosity, permeability, and tortuosity from partial differential equation calculations.

The statistical REV, sREV, represents a smaller scale: property means are constant, and variations are small. We have sREV's for microstructures of a variety of materials, including single-phase flow in sandstones (Zhang et al., 2000), mechanics of fiber-reinforced composites (Trias Mansilla, 2005), and transport in fuel cells (Wargo et al., 2012). Until now, the sREV has not been evaluated for nanopore FIB-SEM data from geo-materials: e.g., carbonate rocks and shale mudstones.

Computations to determine the sREV are traditionally expensive; e.g., lattice Boltzmann simulations (Zhang et al., 2000). In contrast, persistence is less expensive. Hence we propose developing an sREV based on it. Our hypothesis is that above a certain scale persistence diagrams are stable, and materials with similar persistence diagrams have similar structural properties.

First, we need a way to measure the similarity of persistence diagrams. A similarity between persistence diagrams can be measured by various pairwise distances such as bottleneck and Wasserstein distances (Cohen-Steiner et al., 2007, 2010). Various comparison methods have been developed including clustering (Marchese et al., 2017) and hypothesis test (Robinson & Turner, 2017). However, a computational cost is high for computing such distances when there are a large number of topological features. In our porous material applications, persistence diagrams include thousands of features, which lead to a long computational time.

We have similarity measures for images that are robust to shift, scale, and noise (Wang et al., 2004; Li & Lu, 2009). However, we want a persistence measure that is sensitive to shift and scale. It will naturally be robust to perturbations because persistence diagrams themselves are invariant to rotating or orienting the data.

Mean squared error (MSE) and persistence landscape (Bubenik, 2015) measure the relative distance between diagrams. However, they depend on image scale, but we need to compare subsamples of different sizes to determine the sREV. Hence we suggest using the Structural SIMilarity (SSIM) of Wang et al. (2004). The range is $[-1, 1]$, where 1 indicates identical images. The SSIM index is the product of three components: luminance $l(x, y)$, contrast $c(x, y)$, and structure $s(x, y)$.

$$\mathrm{SSIM}(x, y) = l(x, y)^{\alpha} * c(x, y)^{\beta} * s(x, y)^{\gamma},$$

where

$$\begin{aligned} l(x, y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \\ c(x, y) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \\ s(x, y) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \end{aligned}$$

Here μ is the average intensity of an image and σ is its standard deviation. We use the default setting $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$ as Wang et al. (2004) so that

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}.$$

Instead of computing SSIM for the whole image, Wang et al. (2004) suggest computing SSIM for multiple local blocks of the image. The Mean SSIM (MSSIM) is the average of the SSIM values of blocks:

$$\mathrm{MSSIM}(x, y) = \frac{1}{M}\sum_{i=1}^{M} \mathrm{SSIM}(x_i, y_i),$$

where x_i and y_i are the ith block of images x and y.

We compute the MSSIM between vectorized persistence diagrams and their mean image for each subvolume. However, the MSSIM is close to one because most pixels are zero. Therefore, we only sum over non-zero blocks:

$$\text{MSSIM}_{PH}(x, \mu) = \frac{1}{|\{k|\mu_k \neq 0\}|} \sum_{i \in \{k|\mu_k \neq 0\}} \text{SSIM}(x_i, \mu_i).$$

There is no prior standard for deciding sREV using persistent homology; we suggest thresholds of 0.9 (weak) and 0.95 (strict) for MSSIM_{PH}.

2.4 Extracting Features with Principal Component Analysis

Images $\longrightarrow$ Vectorized persistence diagrams $\longrightarrow$ Principal Component Analysis $\longrightarrow$ Weights

We extract features of the vectorized persistence diagrams as M. A. Turk and Pentland (1991); M. Turk and Pentland (1991). First, we standardize persistence diagrams by subtracting the mean persistence diagram from all other diagrams. The covariance matrix of the standardized persistence diagrams shows how much variabilities. We reduce the number of eigenvectors of the covariance matrix, which can be up to m^2, using principal component analysis. We compute n principal components of the covariance matrix that correspond to the n largest eigenvalues.

The computed principal components form a basis: the vectorized persistence diagram are approximated by a linear combination of the principal components. Let k be the diagram dimension, $D_{k,i}$ the ith k-dimensional vectorized persistence diagram, $D_{k,\mu}$ the average of n diagrams, PC_{kj} the jth principal component of the covariance matrix of standardized persistence diagrams for $i,\ j \in \{1, \cdots, n\}$. Then the vectorized persistence diagrams can be represented as a linear combination of n principal components.

$$D_{k,i} - D_{k,\mu} \approx c_{ik1} \times PC_{k1} + c_{ik2} \times PC_{k2} + \cdots + c_{ikn} \times PC_{kn}. \tag{1}$$

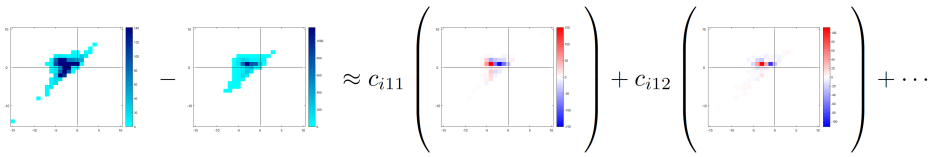

Figure 4. Illustration of a vectorized diagram represented as a linear combination of principal components.

We call the coefficients c_{ikj} the *weights*. Vector $v_i = \{c_{i01}, c_{i02}, \cdots c_{ikn}\}$ summarizes the i^{th} porous material. Figure 4 illustrates Equation (1) for the first and second principal components of the covariance matrix. We can now apply classical statistical approaches to make an inference. For example, the weight vector values can be used as explanatory variables for classification or regression.

2.5 Prediction of Fluid Flow and Transport by Penalized Regression

We seek to fit a model that explains geometric properties (y variables) using principal component analysis weights (x variables). For n subvolumes, we will obtain n weights

for each dimension (i.e., for each Betti number index). As a result, the number of weights obtained over all dimensions is $3n$, and is larger than the number of samples n. Even after the dimension reduction, the number of variables could be larger than the number of samples, which causes over-fitting. This is commonly called the "small n large p" case. One solution is to use an embedded feature selection method. A penalized regression model fits the same linear regression but gives a penalty to the coefficients. The Least Absolute Shrinkage and Selection Operator (LASSO) uses the L_1 penalty (Tibshirani, 1996). LASSO fits a regression and selects variables simultaneously, and can be obtained by solving

$$\min_{\beta}\{||y - X\beta||_2^2 + \delta||\beta||_1\}.$$

The advantage of this approach is that we can see which principal components play a role in predicting fluid flow and transport properties. The LASSO model over the weights estimates the geophysical variables as

$$y = \text{geophysical variables} \sim f_{LASSO}(\text{PCA weights}).$$

3 Results

3.1 Barcodes Describe Deformation

GeoDict ("GeoDict User Guide", 2017) finite volume simulations created a series of deformations of the Bentheimer Sandstone. The rock was subjected to strain levels of 2%, 4%, 6%, 8%, 10%, 20%, 30%, 40%, and 50%. Simulation generated rock images, and their persistence diagrams are given in Figure S2, S3, and S4 in the Supplementary Material. These persistence diagrams capture the structural changes as the stress increases in the following three ways. First, zero-dimensional barcodes reflect decreasing pore size. The negative birth and death values approach zero as pressure goes to 50%. Second, one-dimensional barcodes describe structural changes in the pores and grains. Figure 5 shows the fraction of each barcode quadrant. Non-convex pores (region D in Figure 2.1) make up less than 2.5% initially, and decrease moderately as pressure increases. In contrast, non-convex structures in the grain phase (region E in Figure 2.1) increase significantly with pressure. At 50% strain, non-convex grains comprise 90.3% of the one-dimensional barcodes. This reflects non-convex pores and contact grains gluing together to form non-convex grains as pressure increases. Third, two-dimensional barcodes imply that grain sizes increase with pressure. The positive pixels in the persistence diagrams shift to the upper-right.

3.2 Persistent sREV Determination

We determined the sREV of the Selma chalk (Yoon & Dewers, 2013), the Bentheimer sandstone, and the Doddington sandstone. Figure 6 show a single slice from the grayscale tomographic data of the three datasets. The 2D slice images may not be representative of the corresponding rocks.

3.2.1 Selma Group Chalk sREV is Consistent with Alternatives

Yoon and Dewers (2013) determine sREV by considering the variation of five geophysical characteristics: porosity, permeability, tortuosity, anisotropy, and specific surface area. They define the sREV to be when the subsample is large enough that the coefficient of variation (the standard deviation divided by the mean) is less than 15% for some properties. Their *weak* condition considers porosity, tortuosity, and specific surface area. Their *strict* condition considers all five properties. They conclude the sREV is between 400^3 voxels (weak) and $600 \times 520 \times 600$ voxels (strict).

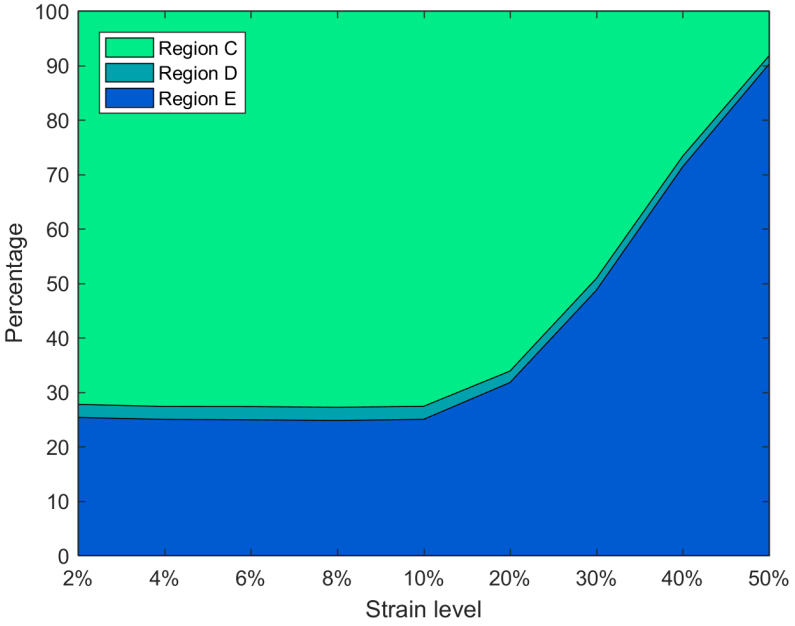

Figure 5. Fraction of each quadrant in one-dimensional persistence diagrams of Bentheimer Sandstone, as strain increases. See Figure 2.1 for the definitions of C, D and E.

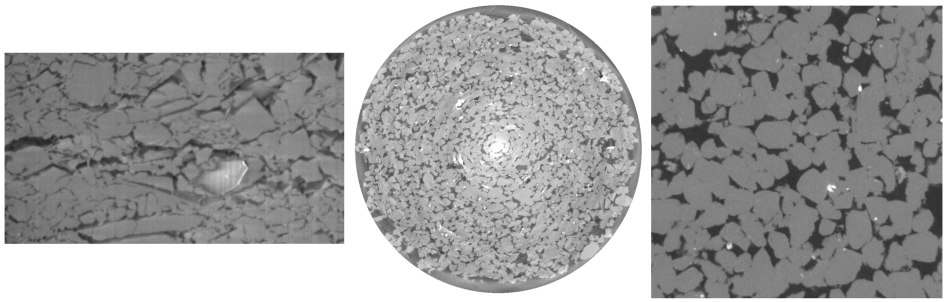

Figure 6. 2D slices from the grayscale tomographic data of Selma chalk (left), Bentheimer (center), and Doddington (right).

We determine the sREV for the Selma group chalk data using the method of Subsection 2.3. Figure 7 Figure 7 is changed to deliver more information in a compact way shows the Mean SSIM ($\text{MSSIM}_{PH}(x, \mu)$) for six subvolumes. We measure the similarities of persistence diagrams for all three dimensions: 0, 1, and 2. For all dimensions, as the size of subvolumes increase, structures become more similar, and the average MSSIM_{PH} values increase. The two-dimensional persistence diagrams show the biggest variation. This implies that the greatest variability comes from the irregular-sized grains in Selma chalk. Our criteria for the sREV is for the average MSSIM_{PH} to exceed 0.9 (weak) and 0.95 (strict). This occurs at the 400^3 and 500^3 datapoints, or around 330^3 and 450^3 voxels when linearly interpolating the data.

Thus our sREV is in $[330^3, 500^3]$, vs. $[400^3, 600 \times 520 \times 600]$ for Yoon and Dewers (2013). Our sREV range is slightly smaller, but the right order of magnitude and the ranges overlap. However, our thresholds are categorically different from Yoon and Dewers (2013)'s. We consider them consistent to the extent that they are comparable.

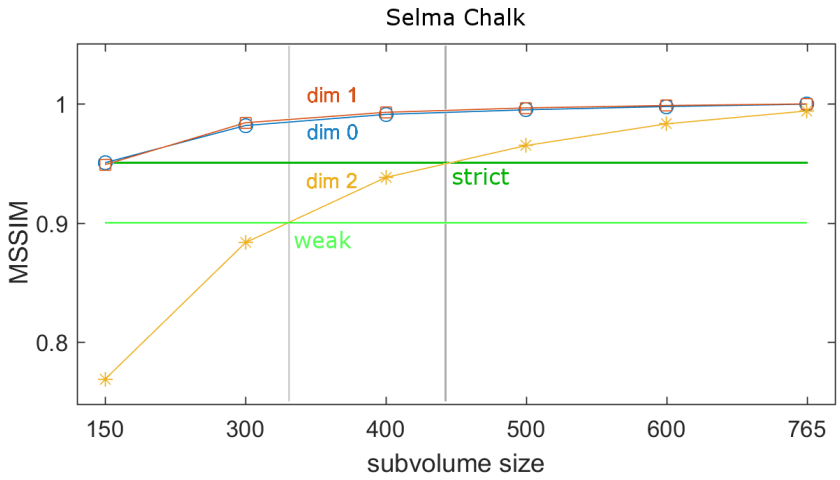

Figure 7. Mean SSIM (MSSIM_{PH}) and sREV thresholds for Selma group chalk.

3.2.2 Bentheimer and Doddington Sandstones

For the Bentheimer and Doddington sandstones we generated four sizes of subvolumes: 100^3, 150^3, $200 \times 200 \times 150$, and $300 \times 300 \times 200$ voxels. The number of subvolumes is 100, 64, 36 and 12. For each size, the location of subvolumes are chosen so that they overlap as little as possible.

Figure 8 shows the Mean SSIM (MSSIM_{PH}) of the sandstone subvolumes. The sandstones' trends are different than the Selma Chalk, and different from each other. The largest dissimilarity (lowest MSSIM) for Bentheimer occurs in the zero-dimensional persistence diagrams, whereas Doddington's occurs in the one-dimensional. This implies that the largest structural variability of Bentheimer comes from the differences in pore sizes. In contrast, Doddington has more variability in non-convex pores and non-convex grains. The [weak,strict] sREV range of Benthheimer is $[150^3, 200 \times 200 \times 150^3]$ (or $[110^3, 160^3]$ interpolated) and Doddington is $[150^3, 300 \times 300 \times 200^3]$ ($[150^3, 230^3]$). We do not have sREV's calculated by other means to compare against.

3.3 Prediction of Fluid Flow and Transport in Selma Chalk is Consistent with Alternatives

We fit a LASSO model to the Selma group chalk data of Yoon and Dewers (2013). We use 42, 23, and 23 subvolumes of sizes 150^3, 300^3, and 400^3. For larger sizes, not enough subsamples are available for our analysis. We train the models for 100 sets of training and test data. The R package glmnet (Friedman et al., 2010) is used to fit the model. The regularization parameter δ is determined by 20 repeated 4-fold cross-validations. The fractions of training, validation, and test sets are 60%, 20%, and 20%, respectively.

Our model predicts three fluid flow and transport properties: permeability k, anisotropy λ, and tortuosity τ. We restate their definitions from Yoon and Dewers (2013) for completeness. Permeability measures how readily a fluid flows. It is measured in $x, y,$ and z directions since they often differ. We define the representative permeability as the geometric mean $k = (k_x k_y k_z)^{1/3}$. Anisotropy measures structural differences between $x, y,$ and z directions. Tortuosity quantifies twisting of pore paths. It is measured in each of the same three directions, and net tortuosity is the arithmetic mean $\tau = (\tau_x + \tau_y + \tau_z)/3$.

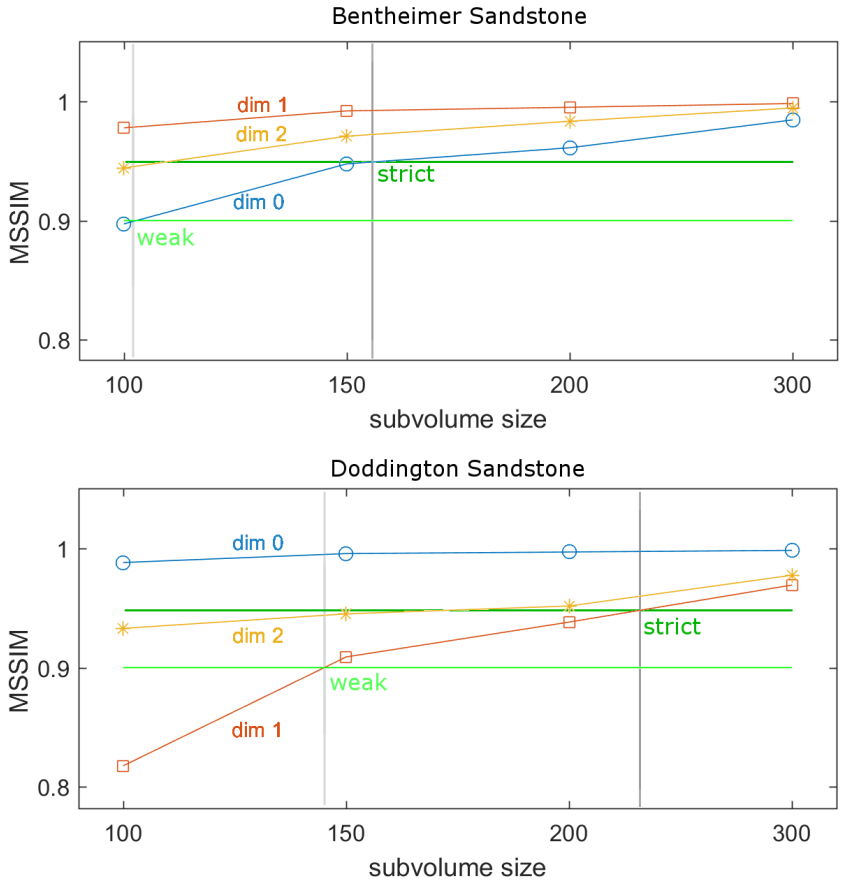

Figure 8. Mean SSIM (MSSIM_{PH}) of Bentheimer (top) and Doddington (bottom).

We compare the prediction results of the proposed model with three models: intercept, linear regression, and PCA model. First, the intercept model estimates the response variable y only by an intercept (average of y). Second, for the linear regression model, we use two x variables, porosity ϕ, and specific surface area S_S. The two x variables are the properties that are not obtained by the lattice Boltzmann simulation. Lastly, the PCA model is a linear regression model using nine weights that correspond to the principal components that have the largest eigenvalues. We select three weights from each dimension.

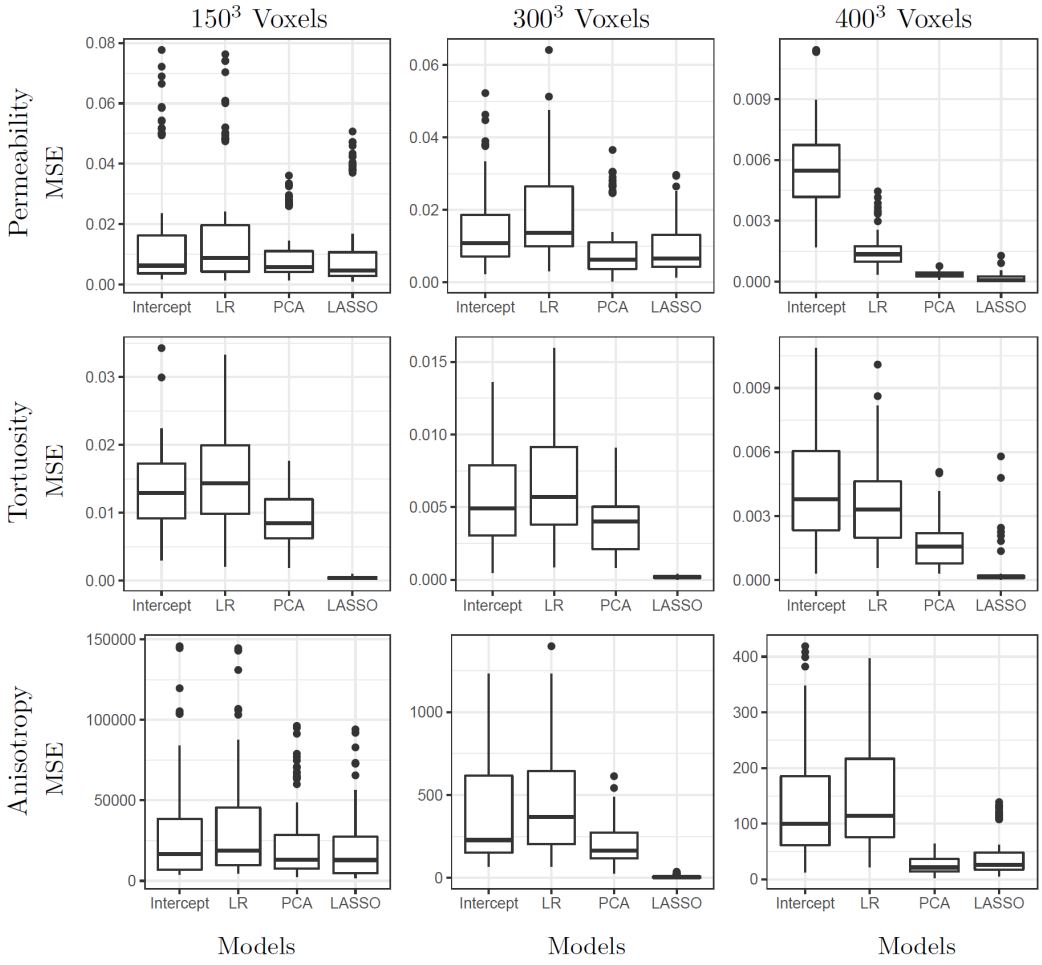

Figure 9. Boxplots of mean squared errors of the intercept, linear regression, PCA model, and LASSO models of permeability, tortuosity, and anisotropy for Selma chalk. Subsample sizes are 150^3, 300^3, and 400^3. Small mean squared errors imply more accurate predictions.

Figure 9 presents a boxplot of the 100 mean squared error (MSE) results for each model and subvolume. Outlier mean squared error values are drawn as dots. The closer MSEs are to zero, the more accurate the prediction. For most cases, the linear regression model does not perform well, even compared to the intercept model.

The suggested LASSO model performs well compared to other models. Permeability has a sudden increase in predictive accuracy at size 400. Tortuosity is relatively predictive for all of the sample scales. Anisotropy is predictive at size 300 but becomes less

accurate at size 400. The PCA model shows similar or better prediction peformance when the suggested LASSO model drops all the variables in cross-validations (See Table 3).

Also, the LASSO model is able to predict the properties for the smaller-sized subsamples than sREV. This comes from the differences between two approaches. The sREV is determined when the topological features become consistent whereas the predictive model is applicable to relatively inconsistent geophysical values.

Table 3. Percentages of existence of LASSO models that dropped all the variables in repeated cross-validations.

		Subsamples		
		150^3	300^3	400^3
Properties	Permeability	88%	0%	0%
	Tortuosity	0%	1%	0%
	Anisotropy	29%	0%	81%

Table 3 presents the percentages when there exist the LASSO models that drop all the weights (x variables) in cross-validations. LASSO drops all the x variables because they do not contain enough information to predict y variable. In this case, the LASSO model only has an intercept, which is the same as the intercept model. These indicate that the weights might not be useful to predict the geophysical properties. Please note that even though LASSO drops all the variables for one of the cross-validation folds, the final model might not be the same as the intercept model. The final LASSO model could be selected from the other cross-validation folds that achieve better predictions than the intercept model. We can expect that higher percentages of LASSO models dropping all the variables in cross-validations tend to have higher MSEs. We can check this tendency in Figure 9.

4 Discussion

Very small subsample sizes can lead to a false conclusion that the sREV has been reached. Tiny subvolumes contain little topological structure of any kind, and only a few barcodes are generated. Since there are few data points, the MSSIM_{PH} differences may be small. We encountered this phenomenon for the two sandstones Bentheimer and Doddington at subvolume size 50^3. There, the MSSIM_{PH} values are large, and higher than at the next-larger sizes. Thus it is important to compute MSSIM_{PH} over a broad range of sizes.

We encountered a similar problem when determining the sREV of Rotleigend sandstone; see Figure S5 in the Supplementary Material. We used the same five subvolumes sizes as the other sandstones. However, some 50^3 subvolumes generated *no* zero-dimensional barcodes; the subsample was enclosed by a single grain! Although the Rotleigend image has more voxels ($700 \times 700 \times 980$) than the other sandstones', it covers a smaller physical extent because it was acquired at a higher resolution. All MSSIM_{PH} values are greater than 0.95, leading to the (false) conclusion that the sREV is at most 100^3. We expect the MSSIM_{PH} graph to be qualitatively v-shaped: high MSSIM_{PH} values for tiny and huge subvolumes, and low MSSIM_{PH} values for some range in between. The sREV should be the size such that *it and all larger sizes* have $\text{MSSIM}_{PH} \geq 0.95$. When selecting subvolume sizes, one should consider the nominal grain and pore sizes, not just the number of image voxels.

5 Conclusion

We determine the sREV for several rock types and properties using topology and statistics. We believe the methodology extends to additional rocks and properties. Whereas Yoon and Dewers (2013) determine the sREV by comparing the stability of geophysical properties, our approach directly compares structural similarities. Our sREV by persistent homology requires less computation since it is based on algebra rather than PDE simulations.

We suggest a statistical model to predict geophycial properties of rocks using their topological summaries. The suggested model performs well to predict permeability and tortuosity even with the small sized subsamples.

Future work includes prediction models capable of further statistical inferences, such as producing confidence intervals. An adaptive SSIM, which adjusts to arbitrary rock type and adjusts sample size, would be valuable. Applying other classical statistical learning methods to persistence diagrams may rigorously expand normalization and scaling. Our similarity measures and sREV thresholds produced results consistent with other approaches. However, more study is required to determine which thresholds to use, or even that the MSSIM is the best measure. One should consider additional data transformations, such as alternative vectorizations. The SEDT has proven useful to many research teams, yet alternatives that consider flow directionality and other physics-inspired features may provide more accurate predictions, or predictions for other quantities such as strength.

Acknowledgments

The three sandstones data (Bentheimer, Doddington, and Rotleigend) are available the digital rock portal (`https://www.digitalrocksportal.org/projects/222`). The Bentheimer network data are available on the digital rock portal (`https://www.digitalrocksportal.org/projects/223`).

The author CM was supported by an appointment with the NSF Mathematical Sciences Summer Internship Program sponsored by the National Science Foundation, Division of Mathematical Sciences (DMS). This program is administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and NSF. ORISE is managed by ORAU under DOE contract number DE-SC0014664. CM performed the majority of his work while visiting the Computer Science Research Institute at Sandia National Laboratories, and CM developed the statistical inference concepts.

SAM and JEH developed the objectives and scientific research questions and guided CM. They were supported by the Exploratory Express Laboratory Directed Research and Development program at Sandia National Laboratories. SAM was funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences Program under contract DE-SC0006883 for supporting interpretation of the persistent homology results, and by the Office of Advanced Scientific Computing Research (ASCR), Applied Mathematics Program.

MA performed the imaging, image processing and segmentation of X-ray tomography volumes, performed geomechanical deformation simulations on the same and assisted in the analysis and interpretation.

The authors thank Honkgyu Yoon and Thomas Dewers for providing the Selma Chalk FIB-SEM data set. The authors thank Nickoas Callor, Hongkyu Yoon, and Thomas Dewers for discussing the relationships between geology-controlled pore structure and constitutive behaviors such as fluid flow.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energys National Nuclear Security Administration under contract DE-NA-0003525.This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., ... Ziegelmeier, L. (2017). Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, *18*, 1-35.

Andrew, M. (2018a). Comparing organic-hosted and intergranular pore networks: Topography and topology in grains, gaps and bubbles. *Geological Society, London, Special Publications*, *484*. doi: 10.1144/SP484.4

Andrew, M. (2018b). A quantified study of segmentation techniques on synthetic geological XRM and FIB-SEM images. *Computational Geosciences*, *22*, 1503–1512. doi: 10.1007/s10596-018-9768-y

Bader, M. (2012). *Space-filling curves: An introduction with applications in scientific computing*. Springer Publishing Company, Incorporated.

Bear, J. (1972). *Dynamics of fluids in porous media*. Dover. Retrieved from `https://books.google.com/books?id=lurrmlFGhTEC`

Bendich, P., Chin, S. P., Clark, J., Desena, J., Harer, J., Munch, E., ... Watkins, A. (2016). Topological and statistical behavior classifiers for tracking applications. *IEEE Transactions on Aerospace and Electronic Systems*, *52*, 2644-2661. doi: 10.1109/TAES.2016.160405

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, *16*(1), 77-102.

Cohen-Steiner, D., Edelsbrunner, H., & Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, *37*, 103-120.

Cohen-Steiner, D., Edelsbrunner, H., Harer, J., & Mileyko, Y. (2010). Lipschitz functions have L_p-stable persistence. *Foundations of Computational Mathematics*, *10*, 127-139. doi: 10.1007/s10208-010-9060-6

Delgado-Friedrichs, O. (2015). Diamorse: Digital image analysis using discrete Morse theory and persistent homology [Computer software manual]. (`https://github.com/AppliedMathematicsANU/diamorse`)

Edelsbrunner, H., & Harer, J. (2008). Persistent homology — a survey. *Surveys on Discrete and Computational Geometry: Twenty Years Later*, *453*, 257-282.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from `http://www.jstatsoft.org/v33/i01/`

GeoDict user guide [Computer software manual]. (2017). Kaiserslautern, Germany: Math2Market GmbH. Retrieved from `https://www.math2market.com/Support/UserGuide.php`

Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, *45*, 61-75.

Hatcher, A. (2002). *Algebraic topology*. Cambridge University Press.

Kusano, G., Fukumizu, K., & Hiraoka, Y. (2016). Persistence weighted Gaussian kernel for topological data analysis. In *Proceedings of the 33rd international conference on machine learning* (pp. 2004–2013). (Available as ArXiv e-print at `http://adsabs.harvard.edu/abs/2016arXiv160101741K`.)

Li, J., & Lu, B.-L. (2009). An adaptive image Euclidean distance. *Pattern Recognition*, *42*, 349–357. doi: 10.1016/j.patcog.2008.07.017

Marchese, A., Maroulas, V., & Mike, J. (2017). K-means clustering on the space of persistence diagrams. In (Vol. 10394). Retrieved from https://doi.org/10.1117/12.2273067 doi: 10.1117/12.2273067

Moon, C., & Andrew, M. (2019a). *Bentheimer networks.* http://www.digitalrocksportal.org/projects/223. Digital Rocks Portal. doi: 10.17612/1a36-rn45

Moon, C., & Andrew, M. (2019b). *Intergranular pore structures in sandstones.* http://www.digitalrocksportal.org/projects/222. Digital Rocks Portal. doi: 10.17612/ze8a-1z13

Reininghaus, J., Huber, S., Bauer, U., & Kwitt, R. (2015). A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4741–4748). doi: 10.1109/CVPR.2015.7299106

Robins, V., Saadatfar, M., Delgado-Friedrichs, O., & Sheppard, A. P. (2016). Percolating length scales from topological persistence analysis of micro-CT images of porous materials. *Water Resources Research*, *52*, 315-329. doi: 10.1002/2015WR017937

Robins, V., Wood, P. J., & Sheppard, A. P. (2011). Theory and algorithms for constructing discrete Morse complexes from grayscale digital images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*, 1646-1658.

Robinson, A., & Turner, K. (2017). Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology*, *1*, 241–261.

Rutka, V., & Wiegmann, A. (2006). Explicit jump immersed interface method for virtual material design of the effective elastic moduli of composite materials. *Numerical Algorithms*, *43*, 309-330.

Scholz, C., Wirner, F., Götz, J., Rüde, U., Schröder-Turk, G. E., Mecke, K., & Bechinger, C. (2012). Permeability of porous materials determined from the Euler characteristic. *Physical Review Letters*, *109*, 264504. doi: 10.1103/PhysRevLett.109.264504

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.

Trias Mansilla, D. (2005). *Analysis and simulation of transverse random fracture of long fibre reinforced composites* (Doctoral dissertation, University of Girona, Spain). Retrieved from http://hdl.handle.net/10803/7762

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71-86. Retrieved from https://doi.org/10.1162/jocn.1991.3.1.71 (PMID: 23964806) doi: 10.1162/jocn.1991.3.1.71

Turk, M. A., & Pentland, A. P. (1991). Face recognition using eigenfaces [Journal Article]. In *Proceedings IEEE computer society conference on computer vision & pattern recognition* (p. 586-591). Retrieved from http://proxy-remote.galib.uga.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=92216393&site=eds-live

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*, 600-612.

Wargo, E., Hanna, A., Çeçen, A., Kalidindi, S., & Kumbur, E. (2012). Selection of representative volume elements for pore-scale analysis of transport in fuel cell materials. *Journal of Power Sources*, *197*(Supplement C), 168-179. doi: 10.1016/j.jpowsour.2011.09.035

Yoon, H., & Dewers, T. A. (2013). Nanopore structures, statistically representative elementary volumes, and transport properties of chalk. *Geophysical Research Letters*, *40*, 4294-4298.

Zhang, D., Zhang, R., Chen, S., & Soll, W. E. (2000). Pore scale study of flow in porous media: Scale dependency, REV, and statistical REV. *Geophysical Research Letters*, *27*, 1195-1198.