Statistical Inference for Porous Materials using Persistent Homology

CHUL MOON*, JASON E. HEATH †, AND SCOTT A. MITCHELL ‡

Abstract. We propose a porous materials analysis pipeline using persistent homology. We first compute persistent homology of binarized 3D images of sampled material subvolumes. For each image we compute sets of homology intervals, which are represented as summary graphics called *persistence diagrams*. We convert persistence diagrams into image vectors in order to analyze the similarity of the homology of the material images using the mature tools for image analysis. Each image is treated as a vector and we compute its principal components to extract features. We fit a statistical model using the loadings of principal components to estimate material porosity, permeability, anisotropy, and tortuosity. We also propose an adaptive version of the structural similarity index (SSIM), a similarity metric for images, as a measure to determine the statistical representative elementary volumes (sREV) for persistence homology. Thus we provide a capability for making a statistical inference of the fluid flow and transport properties of porous materials based on their geometry and connectivity.

1. Introduction. Algebraic topology offers powerful mathematical tools for describing the connectivity of a space, and how the connectivity varies. It connects the local shapes in a dataset to several global connectivity properties, in a concrete, measurable way. In particular, persistent homology analyzes the dynamics of topological features (connected components, loops, shells, etc.) as a function of a metric. As the metric is increased, the data become more connected, and features appear and disappear. One has the freedom to choose the metric to illuminate the properties of interest. The dynamics of a topological feature is quantified as the interval between its "birth" and "death," which are the values of the metric at which the feature first occurs and finally disappears. Birth and death intervals are called persistence data. A goal of applied persistent homology is to predict physical behavior of a system from persistence data. Classical statistical learning methods, however, cannot be applied directly to make predictions from persistence data, due to several challenges. The number of intervals of persistence data generated can vary from dataset to dataset, which makes it difficult to develop methods for predicting behaviors for a given system. Also, persistence data depends on the size of the dataset, or subsample of the dataset, being considered. But how it depends is poorly understood, and may change unpredictably as the subsample size is increased. Traditional statistical methods use the concept of a Representative Elementary Volume, REV [1]. The REV is the smallest size of a material for which statistical stable properties can be defined and measured. For porous media, such as rock, REVs for key parameters such as porosity, permeability, and tortuosity enable one to apply continuum methods to predict fluid flow and transport. The REV enables the complex pore microstructure of rock to be replaced by a fictitious continuum and thereby allow the use of partial differential equations. REVs are thus useful, but discard pore-scale information.

Algebraic topology offers powerful mathematical tools for describing local-to-global structures of shapes. Persistent homology, in particular, analyzes the dynamics

*Department of Statistics, University of Georgia, chulmoon@uga.edu
†Sandia National Laboratories, jeheath@sandia.gov
‡Sandia National Laboratories, samitch@sandia.gov

NSF CCR Center for Computing Research Sandia National Laboratories NNSA National Nuclear Security Administration U.S. DEPARTMENT OF ENERGY

of topological features and summarizes it by numeric values. The dynamics of topological features are recorded as an interval between each feature's birth and death. However, the classical statistical learning methods cannot be directly applied; the type of data is an interval, and the number of intervals generated varies from dataset to dataset.

1.1. Summary of contribution. We present a persistence homology analysis framework of 3D image sets. We demonstrate it over a focused ion beam scanning electron microscopy dataset of the Selma Chalk. We compute and extract structural characteristics of sampling volumes via persistent homology and principal component analysis. We fit a statistical model using the summarized values to estimate porosity, permeability, anisotropy, and tortuosity. The Lattice Boltzmann methods for single phase flow modeling are used to obtain the relationships. The suggested framework efficiently predicts fluid flow and transport properties based on geometry and connectivity.

2. Data. We analyze a FIB-SEM dataset of the Selma Chalk, based on previously binarized images used in the study by Yoon and Dewers [10]. The original dataset includes grayscale images that are 930 x 520 x 962 voxels in size. We use previously calculated parameters, which were calculated as a function of an increasingly-larger subvolume. These include porosity, permeability, tortuosity, and anisotropy.

3. Analysis.

3.1. Analysis pipeline.

1. Computation
   - Original Images → Binary images → Transformed grayscale images → Cubical complexes → Persistence diagrams
2. Data preparation
   - Persistence diagrams → Vectorized persistence diagrams
3. Sampling
   - Use sub-volumes of increasing sizes to determine the scale at which persistent homology predicts behavior, i.e. a REV.
   - Vectorized persistence diagrams → Similarity metric → Determine REV
4. Feature extraction
   - Vectorized persistence diagrams → Principal component analysis → Loadings
5. Modelling
   - Prediction: Fit a statistical model → Prediction
   - Classification

3.2. Persistent homology computation. We use the persistent homology computation framework used by Robins et al. [6, 5]. In Robins et al. [6], the geometric characteristics of the binary image are defined by the Signed Euclidean Distance Transform (SEDT). The SEDT assigns a numeric value to each pixel: negative for pore and positive for grain. Its magnitude represents the Euclidean distance between the pixel and the closest opposite status pixel; a large negative value indicates a large pore size, and a large positive value indicates a large grain size. Then, a cubical cell complex is defined based on the discrete Morse function. The SEDT value of a cell is the maximum value of all of its vertices. The cubical cell complex is an appropriate topological space for the images. The components of cubical cell complexes are 0-cell

(vertex), 1-cell (edge), 2-cell (patch) and 3-cell (solid). Figure 3.1 shows the four components of a cubical cell complex.

Fig. 3.1: Components of a cubical cell complex

As we change the filtration value, k-cell components are added to the cubical cell complex. That is, we add cells one by one in order of increasing SEDT value, thus a cell is added when the current filtration value reaches the maximum value of all of its vertices. By Morse theory, it is sufficient to track critical points: local minimum (0-cell), local maximum (3-cell), saddle points (1-cell and 2-cell). By tracking the homology of the sequence of cubical complexes, we can compute the persistent homology. We use the persistent homology computation code called Diamorse [4].

The barcodes for each dimension are reported separately: the zero, one, and two dimensional homology groups. For each dimension's barcode, the different quadrants of its image reveal different aspects of the rock [6, 5]. For the dimension zero image, the death < 0 quadrants imply the sizes of pore throats, and the death > 0 quadrants describe the disconnected pore space components. Figure 3.2 illustrates the interpretations of the dimension zero persistence diagram.

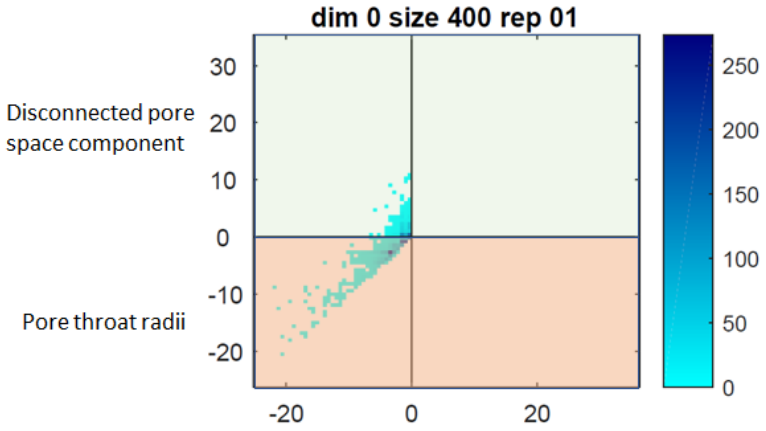

Fig. 3.2: Dimension zero persistence diagram

The dimension one image describes non-convex pores, non-convex grain structures and the narrowest throat radius along the cycle. Figure 3.3 illustrates the interpreta-

tions of the dimension zero persistence diagram.

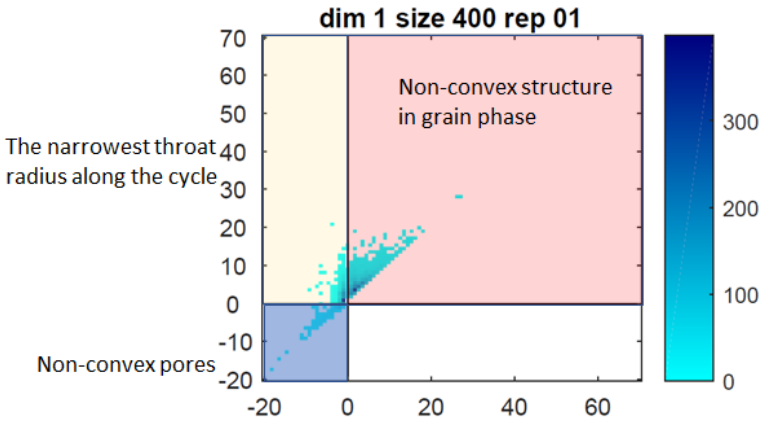

Fig. 3.3: Dimension one persistence diagram

The dimension two barcodes describe the radius of a sphere that can fit within a grain.

3.3. Vectorization of persistence diagram. Persistence homology results are given as an interval of [Birth, Death]. They are summarized in persistence diagrams or barcode plots. The summary graphics are give in Figure 3.4.

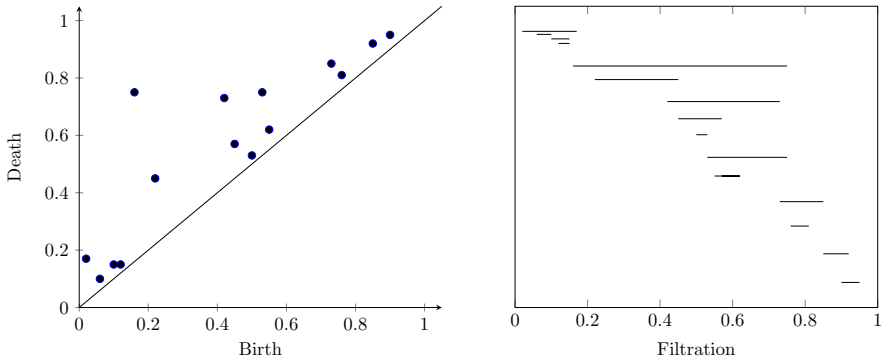

Fig. 3.4: Summary graphics of persistent homology: persistence diagram (left) and barcode plot (right)

The classical statistical learning methods cannot be directly applied to the persistence diagrams. The barcodes are an interval data, a non-classical data type and the number of barcodes varies from dataset to dataset. Instead, we convert persistence diagrams to image vectors. The vectorization process has two advantages: 1) we may apply image analysis technique 2) we may compute the mean persistence diagram. A disadvantage is that the comparison between vectorized persistence diagrams will not be exact, compared to using a Wasserstein distance for example.

We bin the elements of the persistence diagram into $m \times m$ bins; each bin is an output pixel. We count the number of dots that correspond to barcodes in each bin, and assign it as the output pixel intensity. We convert the array of pixels into a vector by scanning by column: visit the bins in the first column from top to bottom, then the bins in the second column, etc. We found this sufficient, but an alternative would be to scan the image in order of a space filling curve. Figure 3.5 illustrates the vectorization process.

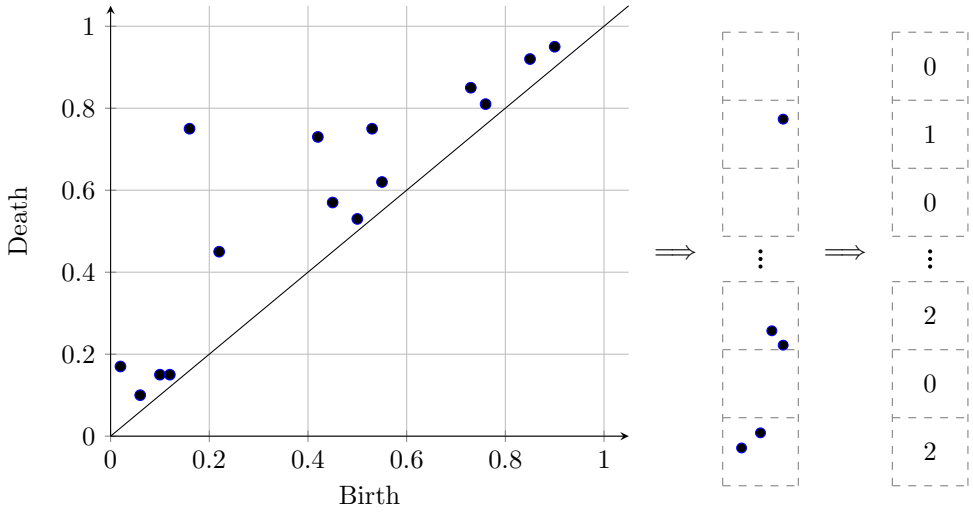

Fig. 3.5: Persistence diagram divided by 5×5 pixels (left), concatenated persistence diagram (middle) and persistence vector (right)

Our use of vectorization in this way is new. The most similar approach was suggested by Bendich et al. [2]. Like us, they count the number of barcodes within the pixel. But they transform the mapping of (birth, death) to (birth, death?birth) in persistence diagram to highlight the distance of the points from the 45-degree line. However, this method cannot be applied directly in our setting because the different quadrants of our persistence diagrams (e.g. for dimension 1, quadrant birth<0 and death<0) imply different meanings, and Bendich et al.'s transformation makes it difficult to identify these sections.

3.4. Determining sREV using persistence homology. The rock 3D scan images are expensive data to obtain. For accuracy, we need a sufficiently large subsample; but for efficiency, we do not want to use a much larger subsample than necessary. The smallest sample size that is sufficiently large is formalized by the Representative Elementary Volume, REV. It is the smallest volume for which a measurement is representative of the whole; see Bear [1] for a better discussion). The REV is the scale at which smaller-scale fluctuations dampen out and statistically stable properties can be defined.

The statistical REV, sREV, represents a scale smaller than that of the REV. The sREV is a scale where the means of properties are constant and their variations are small. The concept of sREV has been used for quantifying microstructures of various materials including single phase flow in sandstones at the microscale [11], mechanical properties of fiber-reinforced composites [7], and transport properties of

fuel cell materials [9]. However, until now, the sREV has not been evaluated for quantitative analysis of FIB-SEM data with nanopore structures observed in geomaterials: e.g., carbonate rocks and shale mudstones. The sREV is closely related to defining the sampling unit from the rock images, and the "right scale" for rock analysis.

sREV determination. We suggest the following method to determine sREV for persistence homology. If the persistence diagrams of sampled subvolumes are similar to each other, then their structural properties would be consistent. We can determine the resemblance by comparing persistence diagrams: the similarity measures for images can be used.

The similarity measures for images have been developed to be robust to differences in shift/scale/noise. However, appropriate measure for the persistence diagrams should be sensitive to the shift and scale differences while robust to perturbations because there are no rotational or directional changes in persistence diagrams. We considered using measures of mean squared error (MSE) and persistence landscape [3]. However, these measure the relative distance between two persistence diagrams. Also, these distances depends on image scales, and so cannot be immediately applied when comparing two subsamples of different sizes. Hence we concluded that these two measures are not the best way to determine sREV for persistent homology.

We instead suggest using the structural similarity (SSIM) [8]. The SSIM index is defined as a multiplication of three components: the luminance $l(x, y)$, contrast $c(x, y)$, and structure $s(x, y)$. The SSIM varies from zero to one, where one indicates two images x and y are identical.

$$\mathrm{SSIM}(x, y) = l(x, y)^{\alpha} * c(x, y)^{\beta} * s(x, y)^{\gamma} \tag{3.1}$$

where

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

We use the same setting $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$ as Wang et al. [8] so that

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{3.2}$$

Instead of computing SSIM for the whole image, Wang et al. [8] suggests using a local block. The mean SSIM (MSSIM) is the average of the SSIM values of blocks:

$$\mathrm{MSSIM}(x, y) = \frac{1}{M}\sum_{i=1}^{M} \mathrm{SSIM}(x_i, y_i), \tag{3.3}$$

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, for the persistence diagrams, the MSSIM is

very high because most of the image pixels are zero, e.g. all lower diagonal pixels. Therefore, we only consider the local block of mean images that have a non-zero element:

$$\text{MSSIM}_{PH}(x, \mu) = \sum_{i \in \{k | \mu_k \neq \mathbf{0}\}} \text{SSIM}(x_i, \mu_i). \tag{3.4}$$

3.5. Feature extraction: principal component analysis. We extract features of the vectorized persistence diagrams using principal component analysis. First, we subtract the mean (vectorized) persistence diagram from all the persistence diagram vectors. We compute the principal components of the covariance matrix. The principal components form a basis to explain the vectorized persistence diagrams. We use the singular vector decomposition to find the principal components to reduce the computational load. The vectorized persistence diagram can be represented as a linear combination of the principal components.

$$\begin{aligned} &i^{th} \text{ persistence diagram of dimension } k - \text{mean persistence diagram} \\ &= c_{ik1} * PC_{k1} + c_{ik2} * PC_{k2} + \cdots + c_{ikn} * PC_{kn} \end{aligned}$$

After computing principal components, we select a subset of principal components that explain at least 85% of total variabilities, and discard the rest to achieve a reduction in dimension. Figure 3.6 illustrates the equation above with the first, second, and third principal components of the dimension one persistence diagram.

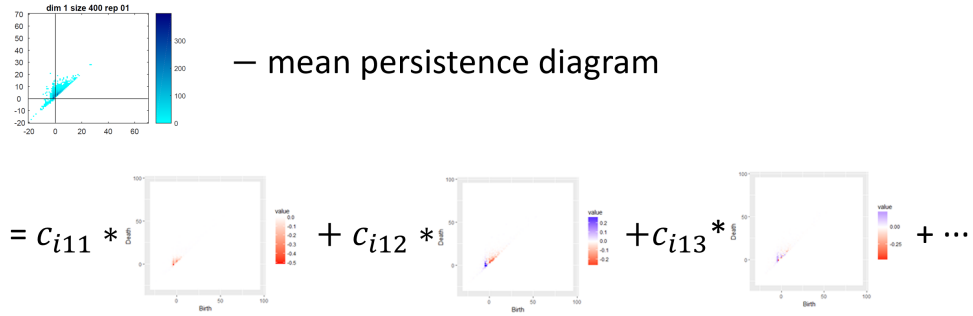

Fig. 3.6: Dimension zero persistence diagram

We call the coefficients of the principal components the *loadings*. We can use the set of loadings to summarize persistence diagrams. The Euclidean vector $v = \{c_{i01}, c_{i02}, \cdots c_{ikn}\}$ summarizes the i^{th} porous material.

Rock Images ⟶ Persistence Diagrams ⟶ Euclidean Vector

Once we convert data into the Euclidean vector, we can then apply classical statistical approaches to make an inference. For example, the numeric values can be used as an explanatory variable for a classification or regression

3.6. Statistical Inference. We aim to make a statistical inference using the converted Euclidean vector, to predict fluid flow and transport properties of already known rocks. As future work, we propose classifying rock types.

3.6.1. Prediction of fluid flow and transport characteristics: penalized regression. We would like to fit a model that explains the geometric properties (y variables) using loadings obtained from the principal component analysis (x variables). If there are n subvolumes, then we will obtain n principal components for each dimension. As a result, the number of loadings obtained for 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. We call this the small n large p case, which causes over-fitting.

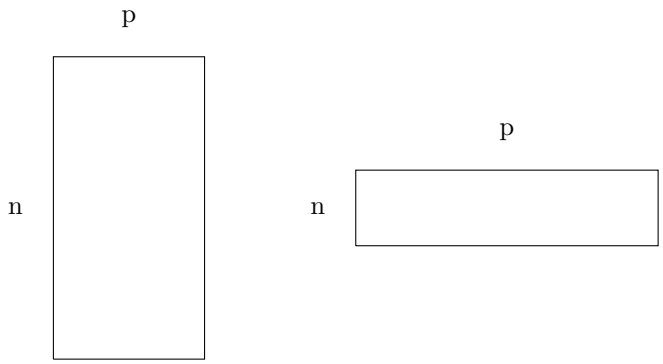

Fig. 3.7: $n > p$ vs. $n < p$ data

One of the solutions to the overfitting problem is to use a penalized regression. We fit the same regression but give a penalty to the coefficients. LASSO is a penalized regression model using a L_1 penalty. The result of LASSO can be obtained by solving

$$\min_{\beta}\{||y - X\beta||_2^2 + \lambda||\beta||_1\}. \tag{3.5}$$

It fits a regression and does variable selection at the same time. The advantage of the LASSO model is that we can see which principal components play a role in predicting fluid flow and transport properties.

3.6.2. Classification extension. After extracting features using principal component analysis, we summarized them with loadings. Given the vectors of loadings, we predicted fluid flow and transport characteristics, using LASSO. For future work, if we had data on multiple rock types, we could attempt to classify them. We propose using the vectors of loadings as input objects to the classification algorithm. In principle, one could use any existing classification method, such as a random forest.

4. Results.

4.1. Determination of sREV. We determine the sREV for the Selma group chalk data of Yoon and Dewers [10]. We compute persistent homology of subvolume images and vectorize the computed persistence diagrams as in Section 3.2 and 3.3. We compute SSIM index between the mean persistence diagram and persistence diagrams and report the average SSIM.

4.1.1. Selma group chalk data. We use the binary image of Yoon and Dewers [10]. We compute persistent homology for six differently-sized subvolumes. There is no standard for deciding REV using SSIM. We set the threshold to be 0.9 because

Yoon and Deweres [10] finds that the subvolume size 400^3 is the sREV. Figure 4.1 shows the average SSIM values for six subvolumes.

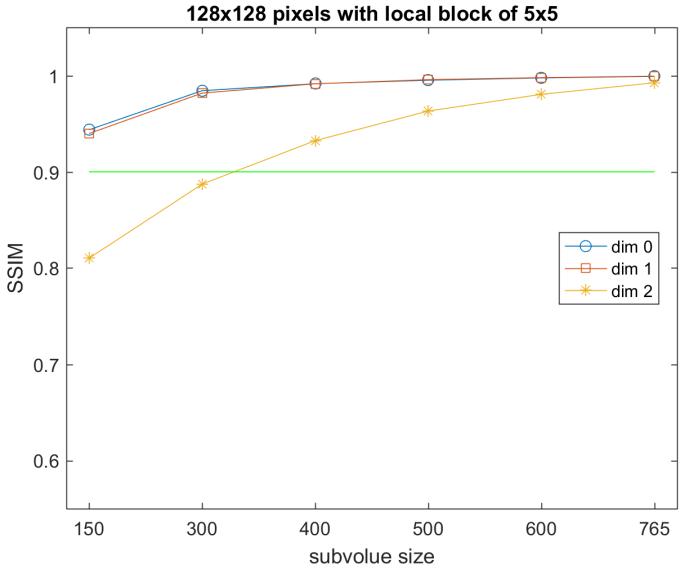

Fig. 4.1: Average SSIM of Selma group chalk

4.2. Prediction of fluid flow and transport properties. We use the Selma group chalk data of Yoon and Dewers [10]. There are in total six different sizes of subvolumes: 150^3, 300^3, 400^3, 500^3, $600 \times 520 \times 600$, and $765 \times 520 \times 765$. We fit a model only for three smallest sizes of subvolumes (150^3, 300^3, and 400^3) because for the other sizes the number of subvolumes is insufficient. We have 42, 23, and 23 subvolumes for the sizes 150^3, 300^3, and 400^3.

We fit a LASSO model to predict four fluid flow and transport properties: porosity ϕ, permeability k, anisotropy λ, and tortuosity τ. We restate their definitions from Yoon and Dewers [10] for completeness. Porosity is the ratio of the volume of the pore space over the total volume. Permeability measures how readily a fluid or gas flows through a material. Permeability is measured in x, y, and z directions. We define the representative permeability as a geometric mean $(k_x * k_y * k_z)^{1/3}$. Anisotropy measures structural differences along the directions. Tortuosity quantifies how much pore paths are twisted. Tortuosity is also measured in three directions. We define the representative tortuosity as an arithmetic mean $(\tau_x + \tau_y + \tau_z)/3$. To decide the λ in the LASSO model, we train the model with 5000 repetitions. The ratio of training, validation, and test sets is 62%, 18%, and 10%, respectively.

We summarize the prediction results of four properties as plots of actual vs. fitted values in Appendix 5. The black dots represent data points of training and validation, whereas the red dots are the test sets. At the sREV, the subvolume size of 400, we found that the models explained all four fluid flow and transport variables. Also, as we increase the subvolume size, the predictions tend to be more accurate.

5. Prediction results. See Figures 5.1—5.4.

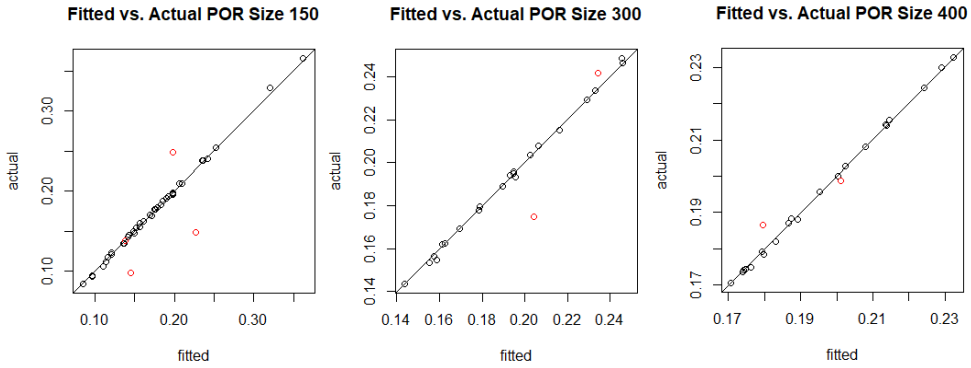

Fig. 5.1: Porosity prediction result. Note the predictions are fairly accurate at all scales, from 150–400.

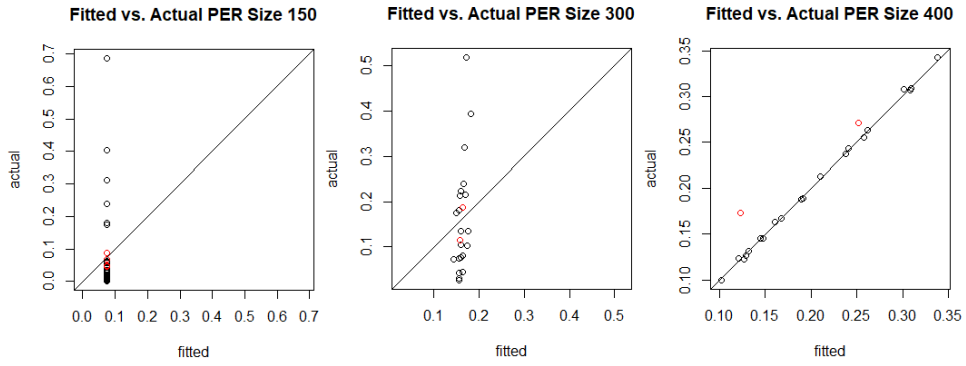

Fig. 5.2: Permeability prediction result. Note the sudden increase in predictive accuracy at size 400.

6. Conclusion. We present a framework for rock analysis using persistent homology, spanning from data preparation to sampling to inference. In future work, we plan to implement prediction models that can make further statistical inferences, such as producing confidence intervals. Also, we aim to develop an adaptive SSIM which can be applied to every type of rock in a sampling process. It would be worthwhile to further investigate how to best apply classical statistical learning methods to the persistence diagrams, expanding our normalization and scaling in a rigorous way.

Acknowledgement. We thank Nickolas Callor, Hongkyu Yoon, Thomas A. Dewers, and Matthew Andrew for helpful discussions, images, and data. This material is based upon work supported by the U.S. Department of Energy, Office of Science: both the Basic Energy Sciences Program under contract DE-SC0006883, and the Office of Advanced Scientific Computing Research (ASCR), Applied Mathematics Program. This research was supported in part by an appointment with the NSF Mathematical Sciences Summer Internship Program sponsored by the National

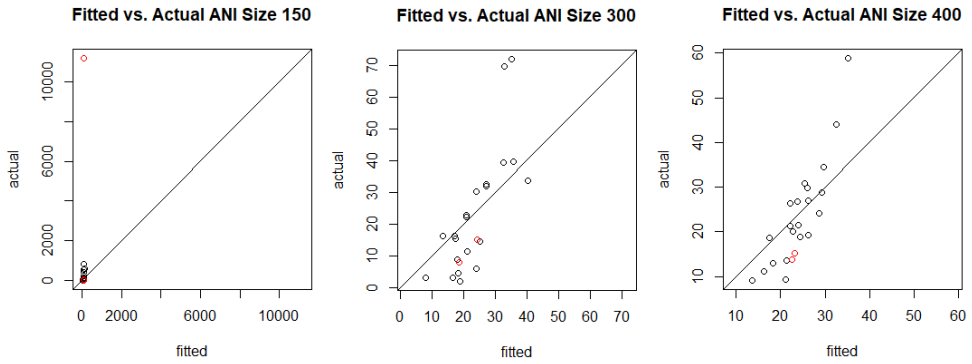

Fig. 5.3: Anisotropy prediction result. Note the gradual increase in predictive accuracy between sizes 300 and 400. The accuracy for anisotropy is not as good as for permeability.

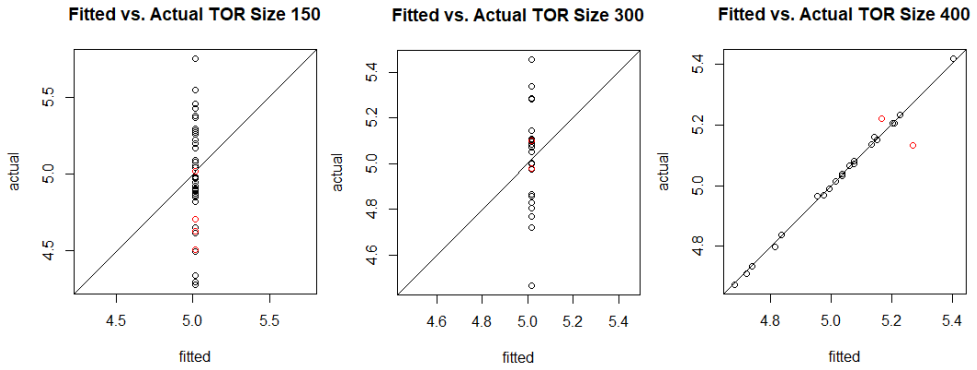

Fig. 5.4: Tortuosity prediction result. As with permeability, there is a sudden increase in predictive accuracy at size 400, supporting the notion of an sREV at that scale.

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. Sandia National Laboratories is a multi-mission 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 Energy's National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

[1] J. Bear, *Dynamics of Fluids in Porous Media*, Dover Civil and Mechanical Engineering Series, Dover, 1972.
[2] P. Bendich, S. P. Chin, J. Clark, J. Desena, J. Harer, E. Munch, A. Newman, D. Porter,

D. Rouse, N. Strawn, and A. Watkins, *Topological and statistical behavior classifiers for tracking applications*, IEEE Transactions on Aerospace and Electronic Systems, 52 (2016), pp. 2644–2661.

[3] P. Bubenik, *Statistical topological data analysis using persistence landscapes*, Journal of Machine Learning Research, 16 (2015), pp. 77–102.

[4] O. Delgado-Friedrichs, *Diamorse: Digital Image Analysis Using Discrete Morse Theory and Persistent Homology*, 2015. `https://github.com/AppliedMathematicsANU/diamorse`.

[5] V. Robins, M. Saadatfar, O. Delgado-Friedrichs, and A. P. Sheppard, *Percolating length scales from topological persistence analysis of micro-CT images of porous materials*, Water Resources Research, 52 (2016), pp. 315–329.

[6] V. Robins, P. J. Wood, and A. P. Sheppard, *Theory and algorithms for constructing discrete Morse complexes from grayscale digital images*, IEEE Transactions on pattern analysis and machine intelligence, 33 (2011), pp. 1646–1658.

[7] D. Trias Mansilla, *Analysis and Simulation of Transverse Random Fracture of Long Fibre Reinforced Composites*, University of Girona, Spain, 2005.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, *Image quality assessment: From error visibility to structural similarity*, IEEE transactions on image processing, 13 (2004), pp. 600–612.

[9] E. Wargo, A. Hanna, A. een, S. Kalidindi, and E. Kumbur, *Selection of representative volume elements for pore-scale analysis of transport in fuel cell materials*, Journal of Power Sources, 197 (2012), pp. 168 – 179.

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

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