



This is the accepted manuscript made available via CHORUS, the article has been published as:

# Cross-Correlation Function for Accurate Reconstruction of Heterogeneous Media

Pejman Tahmasebi and Muhammad Sahimi

Phys. Rev. Lett. **110**, 078002 — Published 14 February 2013

DOI: [10.1103/PhysRevLett.110.078002](https://doi.org/10.1103/PhysRevLett.110.078002)

# Cross-Correlation Function for Accurate Reconstruction of Heterogeneous Media

Pejman Tahmasebi<sup>‡</sup> and Muhammad Sahimi<sup>†,¶</sup>

<sup>‡</sup>*Department of Energy Resources Engineering, Stanford University, Stanford, California  
94305-2220, U.S.A.*

<sup>†</sup>*Mork Family Department of Chemical Engineering & Materials Science, University of  
Southern California, Los Angeles, California 90089-1211, U.S.A.*

Porous media, heterogeneous materials, and biological tissues are examples of ubiquitous disordered systems, understanding of which and any physical phenomenon in them entails having an accurate model. We show that a new reconstruction method based on a cross-correlation function (CCF) and a one-dimensional (1D) raster path provides accurate description of a wide variety of such materials and media. The reconstruction uses a *single* 2D slice of data to reconstruct an *entire 3D medium*. Seventeen examples are reconstructed accurately, as indicated by two connectivity functions that we compute for them. The reconstruction method may be used for both *unconditioned* and *conditioned* problems, and is highly efficient computationally.

PACS number(s): 47.46.+r, 81.05.Rm, 61.43.Bn, 05.20.-y, 02.70.-c, 89.90.+n

<sup>¶</sup>E-mail address: moe@usc.edu

Disordered multiphase media, ranging from porous materials to biological tissues, are ubiquitous and their understanding is of fundamental importance [1,2]. Hence, their modeling has been a long-standing problem. In principle, one needs an infinite set of  $n$ -point correlation functions in order to obtain a complete statistical characterization of such disordered media. Chief among such functions is [1,2] the standard  $n$ -point correlation function [1-3]  $S_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ , the probability of locating  $n$  points at positions  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ , all in one phase. But, measuring  $S_n$  for  $n \geq 5$  is impractical, while it is a formidable challenge for  $n = 3$  and 4, although [5]  $S_3$  and  $S_4$  do not provide much more information beyond what  $S_2$  does. Describing a disordered medium based solely on  $S_2$  is not accurate [4,5]. Thus, a long-standing problem has been the development of correlation functions that not only can be measured, but also provide accurate description of disordered media.

In this Letter we introduce a cross-correlation function (CCF) that is shown to provide accurate description of disordered media. To demonstrate this we use an inverse method, usually referred to as the *reconstruction technique* [1,2,6-10]: given some data for a target system (TS), one tries to construct a model for it that closely matches the data. Complete reconstruction is impractical. Thus, one generates an ensemble of realizations to compute the pertinent properties. We incorporate the CCF in a new reconstruction algorithm that is different from those of the past, and produces accurate realizations of the TS, even if it is highly complex. The reconstruction does not use computationally intensive methods, such as simulated annealing [11], which have been used in the past reconstruction methods. As such, it is highly efficient computationally.

Let  $\mathbf{G}$  represent the computational grid used in the reconstruction. The template (blocks) of  $\mathbf{G}$  by which the TS is reconstructed is denoted by  $\mathbf{T}$ , while  $\mathbf{D}_T(\mathbf{u})$  is the *data event* at position  $\mathbf{u}$  in  $\mathbf{T}$ . We use *data event* because during the reconstruction  $\mathbf{D}_T$  changes. Finally,  $\mathbf{O}$  represents the overlap regions between neighboring templates. We describe the CCF and the reconstruction method for two-dimensional (2D) media; the extension to 3D media is explained in the Supplementary Information (SI) [12]. Let  $\text{TS}(x, y)$  represent the datum at point  $(x, y)$  of a TS of size  $L_x \times L_y$ , with  $x \in \{0, \dots, L_x - 1\}$  and  $y \in \{0, \dots, L_y - 1\}$ . Examining the TS, one focuses on a portion  $\mathbf{D}_T(\mathbf{u})$  of size  $\ell_x \times \ell_y$  and regenerates it based on the data, such that it matches the corresponding portion in the TS. We introduce a CCF to quantify the similarity

between the TS and  $\mathbf{D}_T$ ,

$$\mathcal{C}(i, j) = \sum_{x=0}^{\ell_x-1} \sum_{y=0}^{\ell_y-1} \text{TS}(x+i, y+j) \mathbf{D}_T(x, y), \quad (1)$$

with  $0 \leq i < L_x + \ell_x - 1$  and  $0 \leq j < L_y + \ell_y - 1$ . Equation (1) indicates that the desired position of  $(i, j)$  - the best match with the TS - is one that maximizes  $\mathcal{C}(i, j)$ . If the size of the TS is not too large, all the computations are carried out in the spatial domain. When the TS is very large, calculations in the Fourier space result in significant savings in the computations [13].

Consider, first, unconditional reconstruction - one in which the algorithm must not honor *exactly* specific hard (quantitative) data (HD) in the TS.  $\mathbf{G}$  is partitioned into blocks of size  $L_x \times L_y$ . Between every two neighboring blocks is an  $\mathbf{O}$  region of size  $\ell_x \times \ell_y$  representing  $\mathbf{D}_T$ , where  $\ell_x (\ell_y) = L_x (L_y)$  and  $\ell_y (\ell_x) \ll L_y (L_x)$ , if the  $\mathbf{O}$  region is between two templates that are neighbors in the  $x$  ( $y$ ) direction. Suppose, for example, that the 1D raster path is along the vertical ( $y$ ) direction; see the SI [12] for an illustration. The algorithm begins at the path's origin in  $\mathbf{G}$  - the leftmost bottom template in  $\mathbf{G}$  - and moves along it. A realization of the disorder in the TS, equal in size to  $\mathbf{T}$ , is generated and inserted into the first block, part of which represents the  $\mathbf{O}$  region with the next block. The purpose of the  $\mathbf{O}$  region is to preserve the continuity near the common boundary between two blocks. Thus, one generates the best realization of the disorder for the next neighboring block along the path whose bottom  $\mathbf{O}$  section makes a seamless transition between the two. If several realizations of the disorder have the same degree of similarity with  $\mathbf{O}$ , one of them is selected randomly. The procedure continues until reconstruction of the first vertical column of the raster path is finished. The algorithm then moves to the bottom block of the next vertical column of  $\mathbf{G}$  and uses the same procedure, taking into account its  $\mathbf{O}$  region with the neighboring block on the left. The next block is more complex as it has two  $\mathbf{O}$  regions, at the bottom and on the left. Thus, the reconstruction proceeds as follows

(i) Since we do not know a priori the location of maximum of the CCF, we use  $1/\mathcal{C}$  and set a threshold  $0 \leq \delta \leq 1$ . If  $\delta = 0$ , the similarity between the  $\mathbf{O}$  regions and the corresponding portions in the TS is perfect, whereas  $\delta > 0$  generates an ensemble of realizations that do not match the TS *exactly*. After some preliminary simulations we used,  $\delta = 0.2$ .

(ii) Generate for the first block of  $\mathbf{G}$  along the raster path a realization of the disorder in

the TS, based on and constrained by the statistics of the TS (volume fractions of the phases), or by sampling TS.

(iii) Compute the CCF between the  $\mathbf{O}$  and the *entire* TS.

(iv) If  $1/\mathcal{C} < \delta$ , the realization is accepted. If not, other realizations are made, or  $\delta$  is increased somewhat, until  $1/\mathcal{C} < \delta$ . Typically, one generates 30-50 realizations in order to satisfy the similarity criterion.

(v) Fill up the next block with a realization of the disorder, identify the new  $\mathbf{O}$  with the next block to be reconstructed, and repeat (ii) - (iv). The entire procedure continues until  $\mathbf{G}$  is completely reconstructed.

In conditional reconstruction  $\mathbf{G}$  contains some HD that must be honored *exactly*, due to which the data event  $\mathbf{D}_T$  is the entire block, not just the  $\mathbf{O}$  regions. Thus, conditional reconstruction is a two-step process. First, one identifies the realizations that honor the HD, and then determines the matching  $\mathbf{D}_T$ . The algorithm first computes  $\mathcal{C}$  and checks whether  $1/\mathcal{C} < \delta$ . Those that honor the HD are identified, one of which is selected randomly and inserted in the  $\mathbf{T}$ .

A problem may arise if  $\mathbf{T}$  is very large, or if a realization does not honor the HD, leading to discontinuities and failure of the reconstruction. In such a rare case one may increase the threshold  $\delta$  to obtain new realizations that, although may have significant differences with the  $\mathbf{O}$  regions, they are still acceptable.  $\delta$  is increased until a proper data event  $\mathbf{D}_T$  is identified. But, doing so also inserts incorrect structures in the ensemble. To overcome the problem, we use an adaptive algorithm described below.

The algorithm's parameters are the templates' size and the threshold  $\delta$ . Our study indicates that the templates' size is the most important factor, which depends on the heterogeneity of the TS: for a relatively homogeneous TS a coarse grid suffices, whereas highly disordered media require grids with small blocks. We have developed a method for determining the optimal block size based on the Shannon entropy [14]  $S$ . The TS is partitioned into large coarse blocks, for each of which  $S$  is computed where,  $S = -\sum_{i=1}^n p_i \ln p_i$ , with  $p_i$  being the probability of having states  $i$  in the TS, and  $n$  the number of pixels in the block.  $p_i$  is given by,  $p_i = \text{histogram of sample } i / \text{length of the sample}$ . The entropies are normalized by their maximum value, and denoted by  $\hat{S}$ . A quadtree-based partitioning based on  $\hat{S}$  is then used. If  $\hat{S} \ll 1$ , the coarse block is homogeneous enough that need not be partitioned any further. But, if  $\hat{S}$  is close to

1, the coarse block is divided into 4 (8 in 3D) blocks, the procedure is repeated for the new smaller blocks and continued until the optimal block size is determined. The optimal block size for the binary images of the TS analyzed here was  $20 \times 20$  with an **O** region of size  $20 \times 4$ , while the corresponding sizes for the continuous TS were  $25 \times 25$  and  $25 \times 5$ .

To quantify the accuracy of the reconstruction, we compute two connectivity functions for the TS and its reconstruction. One [5,15] is the two-point cluster (TPC) function  $C_2(r)$ , the probability of finding two points separated by a distance  $r$  in the same cluster of the phase of interest. The second measure is the multiplepoint connectivity (MPC) function [16], the probability  $p(\mathbf{r}; m)$  of having a *sequence* of  $m$  points in a phase in a direction  $\mathbf{r}$ . If an indicator function  $I^{(i)}(\mathbf{u})$  is defined by,  $I^{(i)}(\mathbf{u}) = 1$ , if  $\mathbf{u} \in$  phase  $i$ , and  $I^{(i)} = 0$ , otherwise, then,

$$p(\mathbf{r}; m) = \text{Prob}\{I^{(i)}(\mathbf{u}) = 1, I^{(i)}(\mathbf{u} + \mathbf{r}) = 1, \dots, I^{(i)}(\mathbf{u} + m\mathbf{r}) = 1\}. \quad (2)$$

Reproducing  $p(\mathbf{r}; m)$  represents a most stringent test of the method's accuracy. We used  $m = 100$  and, for convenience, delete  $m$  and denote the MPC function by  $p(\mathbf{r})$ . We note that the previous methods that used optimization techniques based on, for example, simulated annealing, become computationally intractable if the number of data points is very large, whereas the computational cost of the present algorithm is very low, even for a very large 3D system.

A most difficult problem in reconstruction is using *a single 2D section of an entire 3D medium in order to reconstruct it*. Such 2D sections are obtained by, for example, x-ray computed tomography, whereas 3D images are difficult and costly to produce. Thus, an algorithm that utilizes 2D data to reconstruct a 3D medium is a powerful tool for studying disordered materials and media. We show that our reconstruction method is such an algorithm, hence demonstrating that the amount of the data used in the algorithm is not the crucial factor for its accuracy.

As the first example we reconstruct a sample of Berea sandstone of porosity 0.22, whose complete 3D digitized image is available [17]. One 2D slice is used for reconstructing the next layer and adding it to the medium under reconstruction. Since the first layer represents actual hard data (HD) taken from the 3D image, the reconstruction is conditional. The newly reconstructed layer acts as a source of the HD for the next layer to be reconstructed. The important questions are, (i) how should one select the HD in each layer? (ii) How many HD points from the 2D section should one select for reconstructing the next 2D section? If too much

information is used, the next layer will be unrealistically similar to the current one, whereas if the selected HD is too little, the connectivity will be lost. (iii) Where in the 2D slice should one pick the HD? To address these we proceed as follows.

We first use the quadtree partitioning and the Shannon entropy to superimpose a nonuniform grid on the 2D slice. The smallest cell in the 2D grid is of size  $3 \times 3$ , from which a single data point at the center is used, while from all other larger cells 2-5 percent of the data are selected at random. Because the HD must be honored *exactly*, the connectivity is preserved automatically. Every time a 2D slice is reconstructed and added to the system, it is used in the same way for reconstructing the next slice.

The single 2D section of the 3D image of Berea sandstone that we used had a size of  $200 \times 200$  pixels, representing 2% of the total data for the 3D sample. A total of 199 2D sections were reconstructed and added to the initially selected 2D section. Figure 1 presents the connectivity functions  $C_2(r)$  and  $p(r)$ , computed for both the Berea sandstone and its reconstruction. The agreement is excellent. The total computation time for the reconstruction was about 500 CPU seconds.

Next, we generated a synthetic 3D heterogeneous porous medium on a  $200^3$  grid using geostatistical techniques [18]; see Fig. 2. The local permeabilities vary by three orders of magnitude. We used one  $200 \times 200$  slice at the grid's center and reconstructed the 3D porous medium. The results are shown in Fig. 2. The agreement is again excellent.

The next 2D-to-3D reconstruction example that we present is a granular open-cell copper foam that contains about 800 cells with a porosity of 0.33. The size of its binary image is  $128^3$  pixels, and we used a 2D slice of size  $128 \times 128$  to reconstruct the entire image. Figure 3 presents the results. Once again, the agreement is excellent.

It is useful to make a comparison between the performance of the present reconstruction and that of an accurate method developed previously. Thus, we reconstructed a 2D piece of concrete [19] with nanometer-sized pores and centimeter-sized aggregates, using a binarized image of a cross section of the material with a linear size of 170 pixels. The results are shown in Fig. 4. We also computed the TPC and MPC functions for another reconstruction of the same material due to Jiao *et al.* [5]. Whereas the CCF reproduces  $p(r)$  very accurately, Jiao *et al.*'s does not, because theirs is based on  $C_2(r)$ , a radially-averaged TPC function that contains much less microstructural information than the CCF. [A recently-proposed [20] dilation/erosion procedure

using a generalized  $C_2(r)$  yields a better reconstruction of the same sample.] Depending on the template size, the 2D CCF-based reconstruction might use more data than the Jiao *et al.*'s method and, thus, one might argue that it must produce more accurate results. Even if true, as we demonstrate for the 2D-to-3D reconstructions, it is not the primary factor for the method's accuracy. The amount of the data used is based on the templates' size. If the template size is much smaller than the TS (as is the case here), then, the amount of data used is also small.

The last example is a digitized realization of an equilibrium distribution of equal-sized hard spheres [5], generated by the Metropolis Monte Carlo method, with particle volume fraction of 0.446 and a linear size of 100 pixels. Figure 5 presents the results and compares them with the TPC and MPC functions that we also computed for another reconstruction of the sample, due to Jiao *et al.* [5]. While both reconstructions produce the  $C_2(r)$  function so accurately that it is nearly impossible to distinguish them, the CCF-based reconstructed sample also reproduces  $p(r)$  very accurately.

We emphasize a few important points. (i) The reconstruction of a continuous TS based on  $C_2(r)$  [5] is not only a function of the separation distance  $r$ , but also of the values of the pixels, whereas in the CCF-based method the computations for a continuous TS and binary images remain the same. (ii) The method's aim is *not* reproducing explicitly all the statistical properties of a TS, but reproducing *multiscale structures* stochastically, such that they reconstruct the *multipoint statistics*. It is this feature of the method that generates accurate reconstruction. (iii) The method is particularly accurate for highly heterogeneous media.

In addition to the results presented here, twelve other heterogeneous materials and media were studied. The results have been deposited as the Supplementary Information [12].

Summarizing, the advantages of the approach, applicable to any type of disordered material and media, are, (i) low computational cost, but high accuracy; (ii) the ability to include hard data in the computations; (iii) the ability to do the computations in parallel mode; (iv) reconstructing a 3D medium based on a *single 2D section*, and (v) the applicability of the method to *nonstationary* media, which are most prevalent in practice, but have been very difficult to reconstruct. This aspect will be reported in the near future.

Work at the USC was supported in part by the Department of Energy. We thank M. Saadatfar for the data of Figure 3, and Yang Jiao and Sal Torquato for the data of Figure 5.



- [1] S. Torquato, *Random Heterogeneous Materials* (Springer, New York, 2002).
- [2] M. Sahimi, *Heterogeneous Materials*, volumes I & II (Springer, New York, 2003).
- [3] S. Torquato and G. Stell, J. Chem. Phys. **78**, 3262 (1983).
- [4] P. Debye and A. M. Bueche, J. Appl. Phys. **20**, 518 (1948).
- [5] Y. Jiao, F. H. Stillinger, and S. Torquato, Proc. Natl. Acad. Sci. USA **106**, 17634 (2009).
- [6] N. F. Berk, Phys. Rev. Lett. **58**, 2718 (1987); P. Spanne, J.-F. Thovert, C. J. Jacquin, W. B. Lindquist, K. W. Jones, and P. M. Adler, *ibid.* **73**, 2001 (1994); A. P. Roberts and M. Teubner, Phys. Rev. E **51**, 4141 (1995); B. Biswal, P.-E. Øren, R. J. Held, S. Bakke, and R. Hilfer, *ibid.* **75**, 061303 (2007).
- [7] C. L. Y. Yeong and S. Torquato, Phys. Rev. E **57**, 495 (1998); *ibid.* **58**, 224 (1998).
- [8] H. Kumar, C. L. Briant, and W. A. Curtin, Mech. Mater. **38**, 818 (2006).
- [9] Y. Jiao, F. H. Stillinger, and S. Torquato, Phys. Rev. E **76**, 031110 (2007); *ibid.* **77**, 031135 (2008).
- [10] H. Hamzeshpour and M. Sahimi, Phys. Rev. E. **74**, 026308 (2006); H. Hamzeshpour, M. R. Rasaei, and M. Sahimi, *ibid.* **75**, 056311 (2007).
- [11] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
- [12] Link for the Supplementary Information should be given here.
- [13] P. Tehmasebi and M. Sahimi, Computers & Geosciences (to be published).
- [14] C. E. Shannon, Bell Syst. Technol. J. **27**, 379 (1948).
- [15] S. Torquato, J. D. Beasley, and Y. C. Chiew, J. Chem. Phys. **88**, 6540 (1988).
- [16] S. Krishna and A. G. Journel, Math. Geol. **35**, 915 (2003).
- [17] <http://www3.imperial.ac.uk/earthscienceandengineering/research/perm/porescalemodeling>

- [18] M. Sahimi, *Flow and Transport in Porous Media and Fractured Rock*, 2nd ed. (Wiley-VCH, Weinheim, 2011), chapter 5.
- [19] E. J. Garboczi and D. P. Bentz, J. Adv. Cement-Based Mater. **8**, 77 (1998).
- [20] C. E. Zachary and S. Torquato, Phys. Rev. E **84**, 056102 (2011).

## Captions

Fig. 1 The target system (TS) is a 3D sample of Berea sandstone. Shown is the comparison between the connectivity functions  $C_2(r)$  and  $p(r)$  for both TS and the reconstructed medium.

Fig. 2. The TS is a 3D heterogeneous porous medium. The color bar indicates the local porosities.

Fig. 3. The TS is a 3D sample of copper foam.

Fig. 4. The TS is the cross section of a binarized image of concrete microstructure. Comparison is made between the computed connectivity functions for the CCF-based reconstruction and that developed by Jiao *et al.* [5].

Fig. 5. The TS is a digitized realization of a hard-sphere packing with porosity of 0.446.

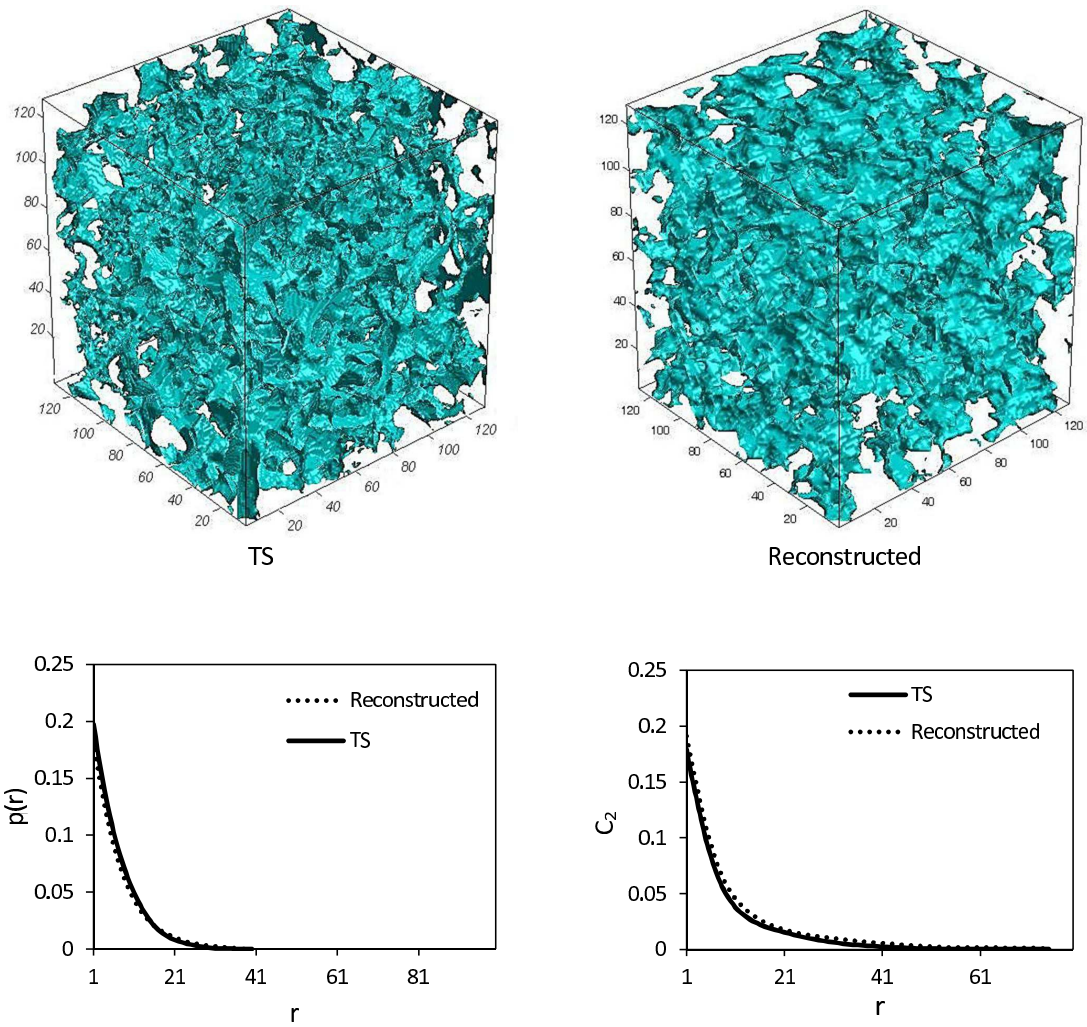


Fig. 1

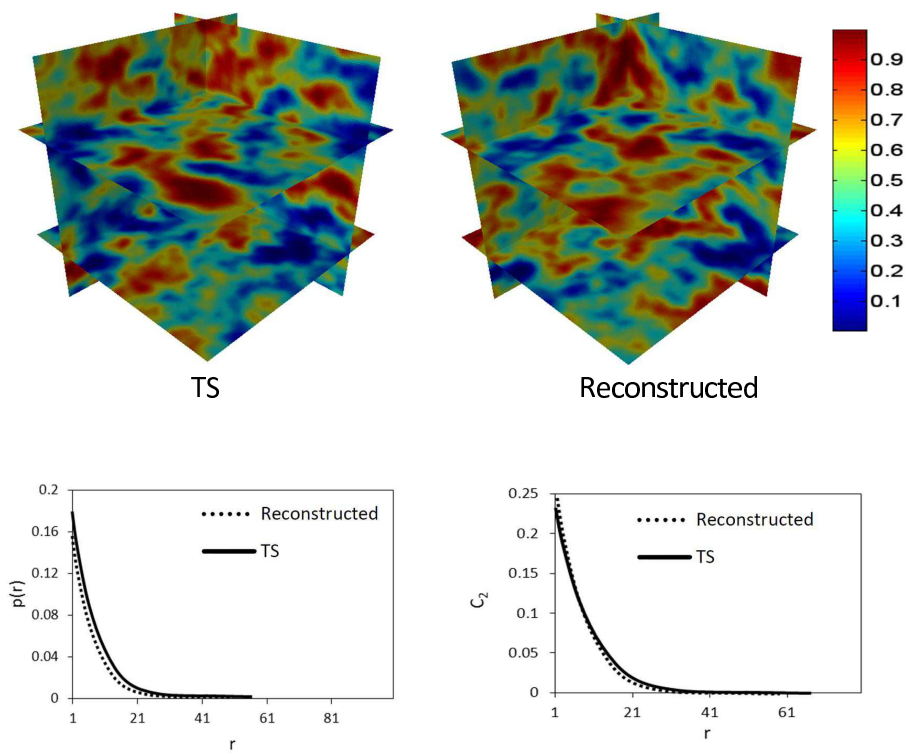


Fig. 2

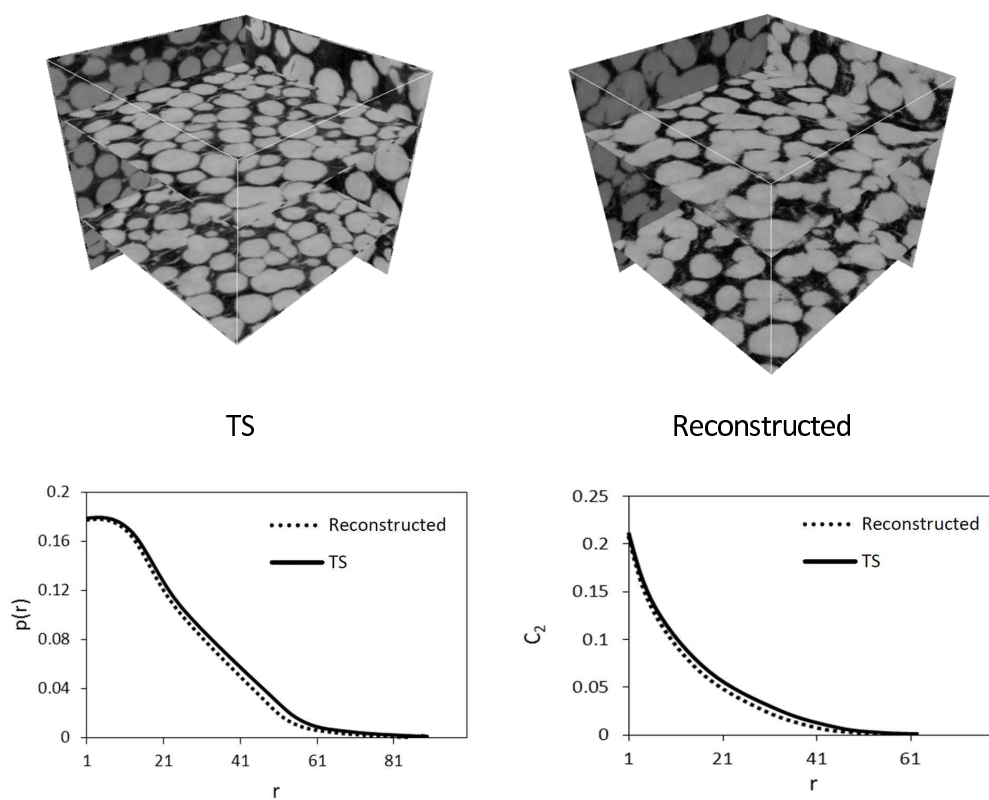


Fig. 3

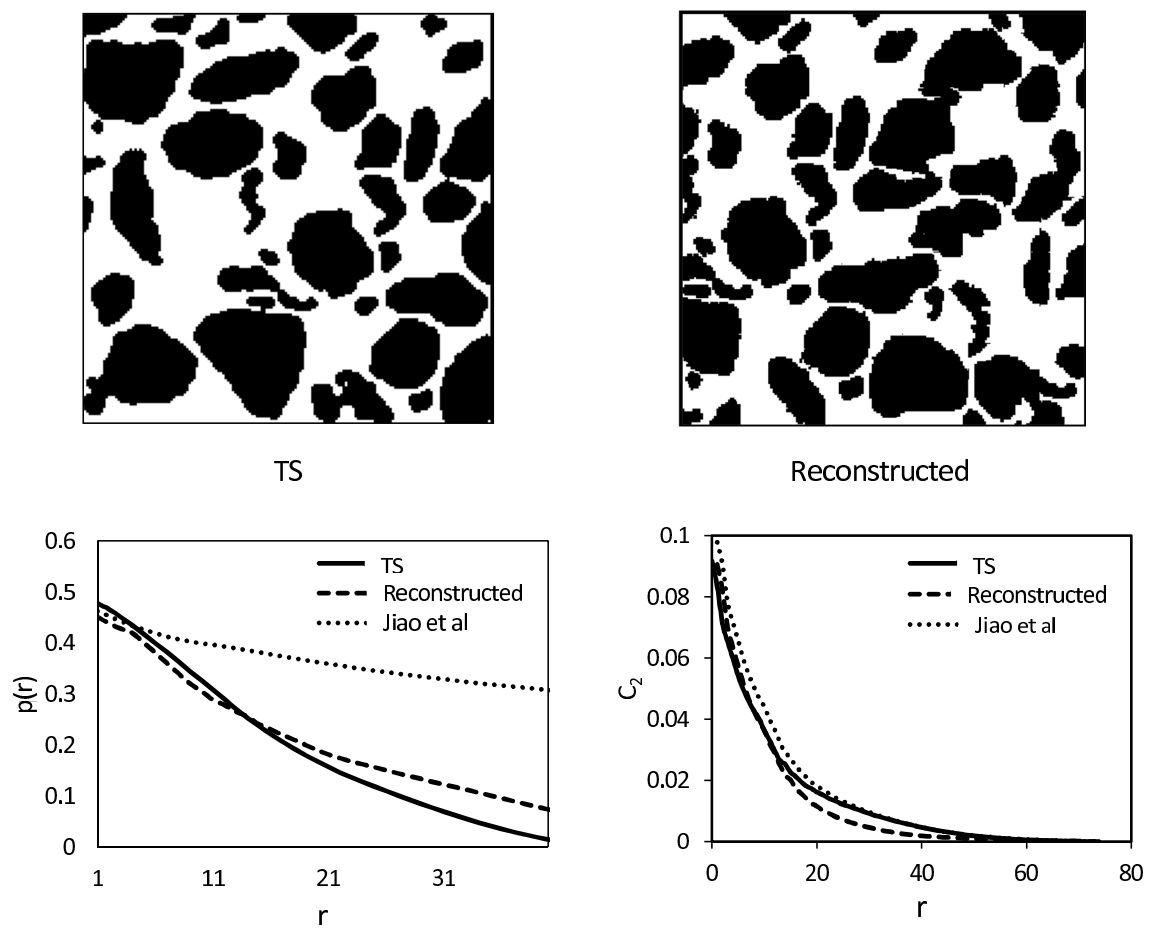


Fig. 4

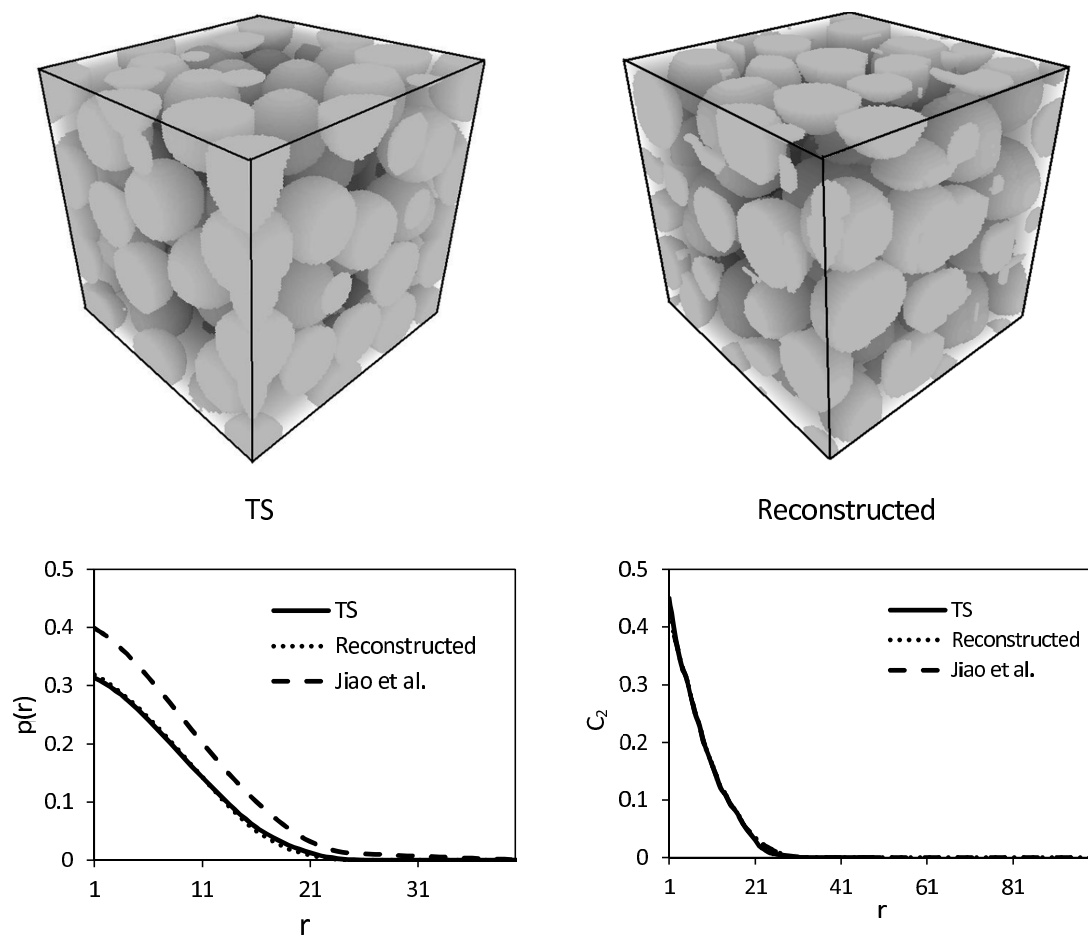


Fig. 5