Letter

Data-driven discovery of the governing equations for transport in heterogeneous media by symbolic regression and stochastic optimization

Jinwoo Im,[1] Felipe P. J. de Barros,[1] Sami Masri,[1] Muhammad Sahimi,[2,*] and Robert M. Ziff[3]

[1]*Sonny Astani Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, California 90089, USA*

[2]*Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, California 90089-1211, USA*

[3]*Michigan Center for Theoretical Physics and Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA*

(Received 13 September 2022; accepted 8 January 2023; published 25 January 2023)

With advances in instrumentation and the tremendous increase in computational power, vast amounts of data are becoming available for many complex phenomena in macroscopically heterogeneous media, particularly those that involve flow and transport processes, which are problems of fundamental interest that occur in a wide variety of physical systems. The absence of a length scale beyond which such systems can be considered as homogeneous implies that the traditional volume or ensemble averaging of the equations of continuum mechanics over the heterogeneity is no longer valid and, therefore, the issue of discovering the governing equations for flow and transport processes is an open question. We propose a data-driven approach that uses stochastic optimization and symbolic regression to discover the governing equations for flow and transport processes in heterogeneous media. The data could be experimental or obtained by microscopic simulation. As an example, we discover the governing equation for anomalous diffusion on the critical percolation cluster at the percolation threshold, which is in the form of a fractional partial differential equation, and agrees with what has been proposed previously.

DOI: 10.1103/PhysRevE.107.L013301

Heterogeneous media and materials, both natural and engineered, are ubiquitous [1,2]. They are often multiscale systems in which the heterogeneity is relevant over multiple and disparate length scales and contain long-range correlations. They vary anywhere from tissues and other biological materials to composite solids, membranes, such large-scale porous media as aquifers, and a vast number of other systems. Many phenomena occur in heterogeneous media that are of fundamental and practical interest, and include flow, transport, reaction, deformation, and other physical processes.

A most important question regarding heterogeneous materials and media is the governing equations for the physical phenomena that occur in them. To address this question, we first divide them into two groups. In one group are those that are microscopically disordered but macroscopically homogeneous. Thus, provided that the size of such media is larger than the representative elementary volume (REV)—the minimum size for macroscopic homogeneity—the phenomena of interest are governed by the classical equations of continuum mechanics [3], averaged over the REV, such as the Navier-Stokes equations for fluid flow, the convective-diffusion equation (CDE) for heat and mass transfer, and equations of linear elasticity. The transport coefficients that appear in such equations represent averaged values, with the averaging taken over the distribution of the heterogeneities, and must be measured by experiments or predicted based on a model of the media.

In the second group are materials and media that are macroscopically heterogeneous [4], implying that the REV is either larger than their size or does not exist. This implies that volume or ensemble averaging of the equations of continuum mechanics is no longer appropriate. A review of a broad class of heterogeneous materials and media indicates that macroscopic heterogeneity is more of a rule than an exception, as they are encountered in astrophysics [5], oceanography [6,7], large-scale porous media [8,9], spatial patterns of environmental pollution [10], and biological tissues and organs [11]. In addition, any statistically self-similar fractal structure, such as the critical percolation cluster (CPC) at the percolation threshold p_c, is also macroscopically heterogeneous up to the length scale over which it is self-similar.

Even if one attempts to carry out a *large-scale* averaging [12–14] over multiple scales, the result is a highly complex equation with many terms, such that direct computer simulation of the phenomena and averaging the numerical results over the relevant length scales are more straightforward than solving the equations that result from large-scale averaging. It is also known that averaging over strong heterogeneity gives rise to the memory effect [15,16], hence complicating the task of deriving the governing equations. To include the memory effects, approaches based on continuous-time random walks [17], miltirate mass transfer equations [18,19], and fractional advective equation [20,21] have been developed. Such approaches are, however, mostly phenomenological.

*moe@usc.edu

In this Letter, we propose an approach that uses a set of data for a transport process in a heterogeneous medium, obtained by either experiments or computer simulation, together with a stochastic optimization method and symbolic regression (see below), to discover the governing equation for the process. An approach for discovering the governing equation for data sets that represent a nonstationary time series has already been developed [22,23]. In addition, the Mori-Zwanzig approach [24,25] provides a procedure for developing reduced-order models for high-dimensional systems and data, which are constructed based on projection operators, although determining the precise form of the kernel in their approach remains difficult. Our goal in the present Letter is to develop an approach for flow and transport in two- or three-dimensional heterogeneous media.

Suppose that $\mathcal{T}$ is the transport process in a heterogeneous medium for which we have an extensive set of data describing the spatiotemporal evolution of a quantity $q(\mathbf{r}, t)$, where $\mathbf{r}$ is the position vector at time t. According to the equations describing conservation of mass, momentum, and energy, the searched-for model $\mathcal{M}$ for the transport process is described by partial differential equations (PDEs). Large-scale averaging methods [12–14] tell us that the spatial variability is expressed by the PDEs that contain, first- and second-order, and possibly third-order, spatial partial derivatives of q, while we also know that averaging over the spatial heterogeneity induces long-term memory [15,16]. Thus, the time evolution of $q(\mathbf{r}, t)$ might be described by its fractional derivative, defined by [26]

$$\partial_{t^\alpha} q \equiv \frac{\partial^\alpha q}{\partial t^\alpha} = \frac{1}{\Gamma(1-\alpha)} \frac{\partial}{\partial t} \int_0^t d\tau \frac{q(\mathbf{r}, \tau)}{(t-\tau)^\alpha} , \qquad (1)$$

where $\Gamma(x)$ is the gamma function. Thus, the goal is to identify a model $\mathcal{M}$ that minimizes the difference between its predictions q_p and the given data q_d, i.e., it minimizes the loss function $\mathcal{L} = \sigma^2 + np$, where σ^2 is the normalized error σ^2, defined by

$$\sigma^2 = \frac{\sum_i \sum_j [q_p(\mathbf{r}_i, t_j) - q_d(\mathbf{r}_i, t_j)]^2}{\sum_i \sum_j [q_d(\mathbf{r}_i, t_j)]^2} , \qquad (2)$$

with n being the number of the nodes in the binary expression tree converted from the PDE, and p is a *complexity penalty coefficient* (see the Supplemental Material (SM) [27], as well as Refs. [15,16,28–32]). Minimizing $\mathcal{L}$ is, of course, a nonlinear optimization problem for which many approaches have been developed [33], such as simulated annealing [34] and the genetic algorithm (GA) [35].

To describe the method concretely, we consider diffusion on 2D CPC at p_c, which has a fractal dimension $D_f = 91/48 \simeq 1.9$ at all length scales. Diffusion in the CPC is anomalous [36], i.e., the mean-squared displacement of a diffusant grows with time as $\langle R^2(t) \rangle \propto t^\alpha$, where $\alpha = 2/D_w$. Here, $D_w = 2 + (\mu - \beta)/\nu$ is the fractal dimension of the walk, with μ, β, and ν being, respectively, the scaling exponents of the conductivity, order parameter, and correlation length of percolation, so with $\mu \simeq 1.3$, $\beta = 5/36$, and $\nu = 4/3$ in 2D, one obtains $D_w \simeq 2.87$. An important, and for quite some time controversial, issue was the governing equation for $q = P(\mathbf{r}, t)$, the average probability that a diffusant is at position $\mathbf{r}$ at time t, for which various equations [32,37,38] were suggested. It now appears that the equation derived by Metzler *et al.* [32] is the generally accepted correct equation (see below).

We generated the CPC on the square lattice at its site percolation threshold, $p_c \simeq 0.5927$, using the Leath algorithm [39], with periodic boundary conditions in both directions. The size of the cluster was 4096×4096, and we averaged the computed $P(\mathbf{r}, t)$ over 500 realizations of the clusters. Diffusion was simulated by an unbiased random walk (RW) on the CPS by the so-called blind ant method [40] using a highly efficient RW simulator, which is an open-source GPU-accelerated [41] algorithm, hence allowing us to use 30 000 particles and simulate 10^6 time steps.

For each realization i, the probability $P_i(r, t)$ $(r = |\mathbf{r}|)$ of a diffusant being within a hyperspherical shell between $(r - \Delta r/2)$ and $(r + \Delta r/2)$ at time t (counted as the number of RW steps) was computed (we used $\Delta r = 4$ in units of the bonds' length). The probability distribution function of $P_i(r, t)$ was then computed by normalizing the numerical results, i.e., by setting, $\int_0^\infty r^{D_f - 1} P_i(r, t) dr = 1$, and then averaging over all the realizations.

Though any optimization algorithm can be used, we utilized the genetic programming for system identification (GPSI) [28]. The complete details are given in the SM [27] (see also Ref. [28]). Briefly, one first specifies the mathematical expressions that will be tried by the GPSI. We included $\partial^n P/\partial r^n$ ($n = 0,\ 1$, and 2), $\partial P/\partial t$, and $\partial_{t^\alpha} P(r, t)$ $(0 < \alpha \leqslant 1)$, together with the boundary conditions, and used the fourth-order Runge-Kutta method when the time derivative was simply $\partial P/\partial t$, and the predictor-corrector method suggested by Diethelm *et al.* [29] when the trial PDEs involved fractional derivatives. The GPSI generates a PDE at random, solves it numerically to compute $P(r, t)$, and calculates the loss function σ^2. If σ^2 is larger than a threshold ϵ, the algorithm continues generating the trial PDEs through the evolutionary process of the GA—the crossover and mutation—until $\sigma^2 < \epsilon$. This generates a few plausible solutions, most of which can be eliminated by imposing other physical constraints, such as $0 < P(r, t) \leqslant 1$. Such a procedure amounts to symbolic regression [42,43], since one tries to fit certain expressions in terms of spatial and temporal derivatives to a given set of data.

As a test, we first carried out RW simulations on the fully connected square lattice at $p = 1$. The algorithm easily identified the spherically symmetric diffusion equation as the only viable governing equation. The simulations on the CPC at p_c yield $D_w \simeq 2.875 \pm 0.003$, in agreement with the theoretical expectation. We used the data for the final 40% of time steps and, therefore, the predictions of the equation to be discovered for the initial 60% of the total time is a stringent test of its accuracy. The algorithm rejected all the PDEs with integer-order time derivatives of $P(r, t)$. Only three possible solutions with fractional derivatives were deemed viable. Of the three, one given by

$$\frac{\partial^{0.718} P}{\partial t^{0.718}} = -0.288 P^2 + 0.202 \frac{\partial^2 P}{\partial r^2} \qquad (3)$$

was rejected, even though its predictions for $P(r, t)$ were accurate, because it violates mass balance. A second

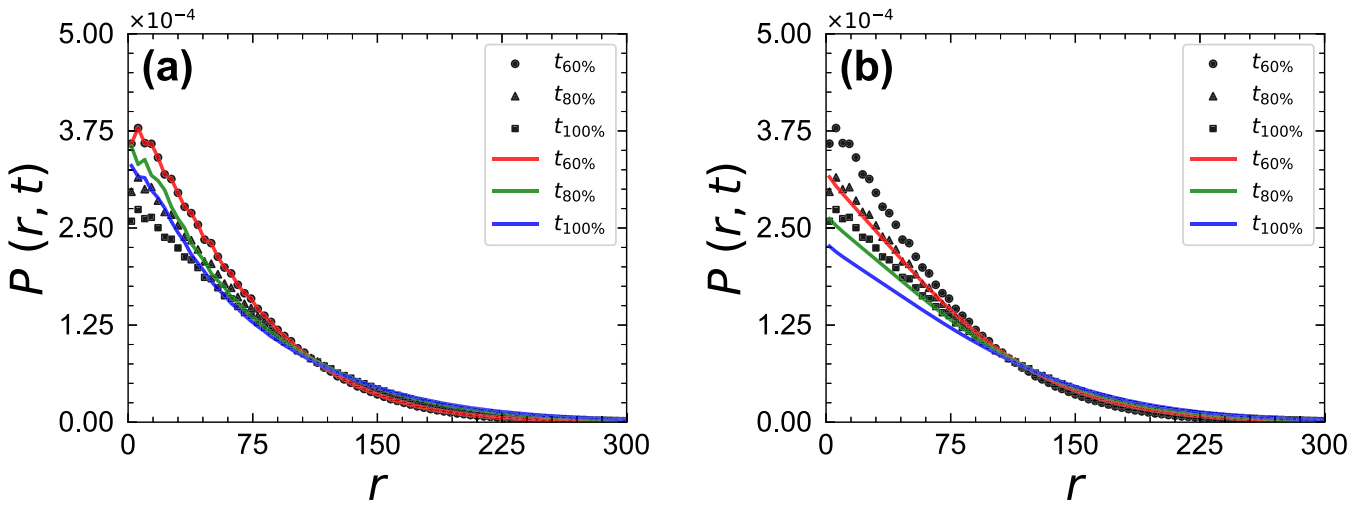

FIG. 1. Comparison of the predictions of the discovered fractional diffusion equation (curves) with the results of numerical simulations of diffusion on the critical percolation cluster at the percolation threshold: (a) Eq. (5) and (b) Eq. (7). To discover the equation, the last 40% of the data at the longest times were used; circle. Triangles and squares represent the numerical results, while red, green, and blue show the model's predictions.

solution,

$$\frac{\partial^{0.645}P}{\partial t^{0.645}} = \frac{0.555}{r}\frac{\partial P}{\partial r} + 0.640\frac{\partial^2 P}{\partial r^2}\,, \tag{4}$$

which is still accurate, was also rejected because it implies anisotropic diffusion. Thus, the final governing equation identified by the approach is given by

$$\frac{\partial^{0.614}P}{\partial t^{0.614}} = \frac{0.849}{r}\frac{\partial P}{\partial r} + \frac{\partial^2 P}{\partial r^2}\,. \tag{5}$$

Note that the factor $1/r$ in the first term of the right side of Eq. (5) was identified by the algorithm, and was not included in the set of trial searches.

On the other hand, the governing equation for $P(r,t)$, derived by Metzler *et al.* [32], is given by

$$\frac{\partial^{\alpha}P}{\partial t^{\alpha}} = \frac{1}{r^{d_s-1}}\frac{\partial}{\partial r}\left[r^{d_s-1}\frac{\partial P(r,t)}{\partial r}\right] = \frac{d_s-1}{r}\frac{\partial P}{\partial r} + \frac{\partial^2 P}{\partial r^2}\,, \tag{6}$$

where $d_s = 2D_f/D_w \simeq 1.321$ is the spectral dimension [44]. Thus, substituting for d_s and $\alpha = 2/D_w \simeq 0.696$, Eq. (6) becomes

$$\frac{\partial^{0.669}P}{\partial t^{0.669}} = \frac{0.321}{r}\frac{\partial P}{\partial r} + \frac{\partial^2 P}{\partial r^2}\,, \tag{7}$$

which is practically identical with what the proposed approach identified. In Fig. 1, we compare the predictions of Eqs. (5) and (7); the agreement is excellent. Since only the last 40% of the data was used in the stochastic optimization, we compare in Fig. 2 the predictions of the discovered equation for

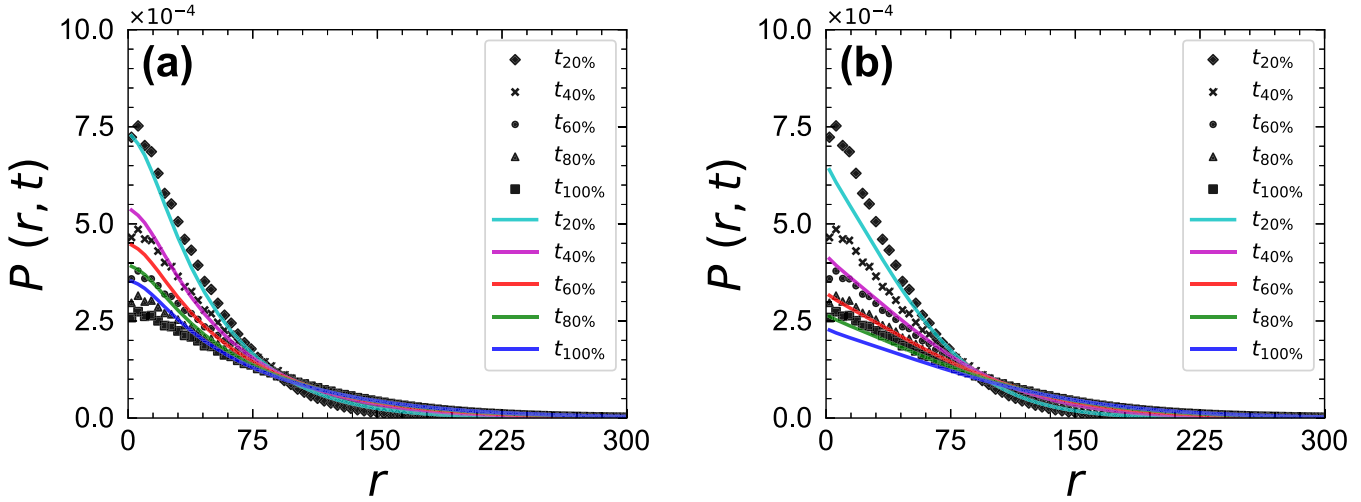

FIG. 2. Comparison of the predictions of the discovered fractional diffusion equation (curves) with the results of numerical simulations of diffusion on the critical percolation cluster at the percolation threshold over the entire simulation time: (a) Eq. (5) and (b) Eq. (7). Diamond, cross, circle, triangle, and square represent the numerical results while cyan, magenta, red, green, and blue show the predictions.

$P(r,t)$ over the entire time that was simulated. It is clear that Eq. (5) provides accurate predictions even for the initial 60% of the data. Note that since our estimate of $\alpha = 2/D_w$ is in agreement with the theoretical expectation, the reason for the difference between the value of $d_s \simeq 1.849$ that Eq. (5) identified and the theoretical prediction, $d_s \simeq 1.321$, is due to the finite-size effect that influences the value of the fractal dimension D_f of the CPC.

Let us point out that, as He *et al.* [45] showed, the dynamics of transport processes in heterogeneous media that are described by a fractional diffusion equation is not self-averaging, in that time and ensemble averages of the observables, such as the mean-squared displacements, do not converge to each other. This is consistent with what is known for diffusion on the CPC at the percolation threshold [46,47], for which the distribution of the displacements of the diffusing particle does not exhibit self-averaging. Our discovery of a fractional diffusion equation for diffusion on the CPC at the percolation threshold is fully consistent with this picture and indicates the accuracy of the approach.

As a further test of the method, we used experimental data of Scheidegger [48] for dispersion of a solute in the flow of a solvent through a heterogeneous porous medium, which have been subject to debate for decades because the data cannot be accurately described by the standard 1D CDE,

$$\frac{\partial C}{\partial t} + v\frac{\partial C}{\partial z} = D_L \frac{\partial^2 C}{\partial z^2} \ , \tag{8}$$

where C is the solute concentration, v is the mean flow velocity, and D_L is the dispersion (effective diffusion) coefficient. Our preliminary computations based on the method proposed here indicate that the data can be accurately described by a fractional CDE of the following form:

$$\frac{\partial^\alpha C}{\partial t^\alpha} + v\frac{\partial C}{\partial z} = D_L \frac{\partial^\beta C}{\partial z^\beta} \ , \tag{9}$$

where $\alpha < 1$ and $1 < \beta < 2$, hence shedding light on decades-old experimental data. The details will be reported elsewhere [49].

Summarizing, with advances in instrumentation and the tremendous increase in computational power, vast amounts of data are becoming available for various phenomena in macroscopically heterogeneous media. To understand and analyze such data and make predictions for future states of the phenomena, one must be able to represent them by accurate governing equation(s). We proposed a data-driven approach, based on stochastic optimization and symbolic regression, which provides an effective solution for this unsolved problem and opens the way to many applications of the method for a wide variety of complex phenomena in heterogeneous media.

M.S. was supported by the National Science Foundation Grant No. CBET #2000966.

[1] S. Torquato, *Random Heterogeneous Materials* (Springer, New York, 2002).
[2] M. Sahimi, *Heterogeneous Materials I & II* (Springer, New York, 2003).
[3] S. Whitaker, *The Method of Volume Averaging* (Springer, New York, 2013).
[4] P. Tahmasebi and M. Sahimi, Reconstruction of nonstationary disordered materials and media: Watershed transform and cross-correlation function, Phys. Rev. E **91**, 032401 (2015).
[5] R. A. Frazin, M. D. Butala, A. Kemball, and F. Kamalabadi, Time-dependent reconstruction of non-stationary objects with tomographic or interferometric measurements, Astrophys. J. **635**, L197 (2005).
[6] A. Seppänen, M. Vauhkonenl, P. J. Vauhkonenl, E. Somersalo, and J. P. Kaipio, State estimation with fluid dynamical evolution models in process tomography—an application to impedance tomography, Inverse Probl. **17**, 467 (2001).
[7] R. T. Lemos and B. Sansó, A spatio-temporal model for mean, anomaly, and trend fields of north Atlantic sea surface temperature, J. Am. Statist. Asso. **104**, 5 (2009).
[8] M. Honarkhah and J. Caers, Stochastic simulation of patterns using distance-based pattern modeling, Math. Geosci. **42**, 487 (2010).
[9] M. Sahimi, *Flow and Transport in Porous Media and Fractured Rock*, 2nd ed. (Wiley-VCH, Weinheim, 2011).
[10] J. L. Mennis and L. Jordan, The distribution of environmental equity: Exploring spatial nonstationarity in multivariate models of air toxic releases, Ann. Asso. Am. Geog. **95**, 249 (2005).
[11] T. McInerney and D. Terzopoulos, A dynamic finite element surface model for segmentation and tracking in multidimensional medical images with application to cardiac 4D image analysis, Comput. Med. Graph. **19**, 69 (1995).
[12] M. Quintard and S. Whitaker, Two-phase flow in heterogeneous porous media: The method of large-scale averaging, Transp. Porous Media **3**, 357 (1988).
[13] L. W. Gelhar, A. L. Gutjahr, and R. J. Naff, Stochastic analysis of macrodispersion in stratified aquifers, Water Resour. Res. **15**, 1387 (1979).
[14] L. W. Gelhar and C. L. Axness, Three-dimensional stochastic analysis of macrodispersion in aquifers, Water Resour. Res. **19**, 161 (1983).
[15] J. Klafter and R. Silbey, Derivation of the Continuous-Time Random-Walk Equation, Phys. Rev. Lett. **44**, 55 (1980).
[16] M. Sahimi, B. D. Hughes, L. E. Scriven, and H. T. Davis, Stochastic transport in disordered systems, J. Chem. Phys. **78**, 6849 (1983).
[17] B. Berkowitz, A. Cortis, M. Dentz, and H. Scher, Modeling non-Fickian transport in geological formations as a continuous time random walk, Rev. Geophys. **44**, RG2003 (2006).
[18] M. Dentz and B. Berkowitz, Transport behavior of a passive solute in continuous time random walks and multirate mass transfer, Water Resour. Res. **39**, 1111 (2003).
[19] O. Silva, J. Carrera, M. Dentz, S. Kumar, A. Alcolea, and M. Willmann, A general real-time formulation for multi-rate mass transfer problems, Hydrol. Earth Syst. Sci. **13**, 1399 (2009).

[20] Y. Zhang, D. A. Benson, M. M. Meerschaert, and E. M. LaBolle, Space-fractional advection-dispersion equations with variable parameters: Diverse formulas, numerical solutions, and application to the Macrodispersion Experiment site data, Water Resour. Res. **43**, W05439 (2007).

[21] S. P. Neuman and D. M. Tartakovsky, Perspective on theories of non-Fickian transport in heterogeneous media, Adv. Water Resour. **32**, 670 (2009).

[22] R. Friedlich, J. Peinke, M. Sahimi, and M. R. Rahimi Tabar, Approaching complexity by stochastic methods: From biological systems to turbulence, Phys. Rep. **506**, 87 (2011).

[23] F. Nikakhtar, L. Parkavosi, M. R. Rahimi Tabar, M. Sahimi, K. Lehnertz, and U. Feudel, Data-Driven Reconstruction of Stochastic Dynamical Equations based on Statistical Moments Phys. Rev. Lett. (to be published).

[24] H. Mori, Transport, collective motion, and Brownian motion, Prog. Theor. Phys. **33**, 423 (1965).

[25] R. Zwanzig, Nonlinear generalized Langevin equations, J. Stat. Phys. **9**, 215 (1973).

[26] *Applications of Fractional Calculus in Physics*, edited by R. Hilfer (World Scientific, Singapore, 2000).

[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.107.L013301 for for details of optimization and numerical procedures.

[28] J. Im, C. B. Rizzo, F. P. J. de Barros, and S. F. Masri, Application of genetic programming for model-free identification of nonlinear multi-physics systems, Nonlinear Dyn. **104**, 1781 (2021).

[29] K. Diethelm, N. J. Ford, and A. D. Freed, A predictor-corrector approach for the numerical solution of fractional differential equations, Nonlinear Dyn. **29**, 3 (2002).

[30] S. Havlin and D. Ben-Avraham, Diffusion in disordered media, Adv. Phys. **36**, 695 (1987).

[31] M. Sahimi and A. O. Imdakm, The effect of morphological disorder on hydrodynamic dispersion in flow through porous media, J. Phys. A **21**, 3833 (1988).

[32] R. Metzler, W. G. Glöckle, and T. F. Nonnenmacher, Fractional model equation for anomalous diffusion, Physica A **211**, 13 (1994).

[33] M. Sahimi and P. Tahmasebi, Reconstruction, optimization, and design of heterogeneous materials and media: Basic principles, computational algorithms, and applications, Phys. Rep. **939**, 1 (2021).

[34] S. Kirkpatrick Jr., C. D. Gelatt, Jr., and M. P. Vecchi, Optimization by simulated annealing, Science **220**, 671 (1983).

[35] M. Mitchell, *An Introduction to Genetic Algorithms* (MIT Press, Cambridge, 1996).

[36] Y. Gefen, A. Aharony, and S. Alexander, Anomalous Diffusion on Percolating Clusters, Phys. Rev. Lett. **50**, 77 (1983).

[37] B. O'Shaughnessy and I. Procaccia, Analytical Solutions for Diffusion on Fractal Objects, Phys. Rev. Lett. **54**, 455 (1985).

[38] M. Giona and H. E. Roman, Fractional diffusion equation for transport phenomena in random media, Physica A **185**, 87 (1992).

[39] P. L. Leath, Cluster size and boundary distribution near percolation threshold, Phys. Rev. B **14**, 5046 (1976).

[40] M. Sahimi, *Applications of Percolation Theory*, 2nd ed. (Springer, New York, 2023).

[41] C. B. Rizzo, A. Nakano, and F. P. J. de Barros, PAR^2: Parallel random walk particle tracking method for solute transport in porous media, Comput. Phys. Commun. **239**, 265 (2019).

[42] J. Bongard and H. Lipson, Automated reverse engineering of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA **104**, 9943 (2007).

[43] M. Schmidt and H. Lipson, Distilling free-form natural laws from experimental data, Science **324**, 81 (2009).

[44] R. Rammal and G. Toulouse, Random walks on fractal structures and percolation clusters, J. Phys. Lett. **44**, L-13 (1983).

[45] Y. He, S. Burov, R. Metzler, and E. Barkai, Random Time-Scale Invariant Diffusion and Transport Coefficients, Phys. Rev. Lett. **101**, 058101 (2008).

[46] A. Bunde and J. Dräger, Localization in disordered structures: Breakdown of the self-averaging hypothesis, Phys. Rev. E **52**, 53 (1995).

[47] A. Pacheco-Pozo and I. M. Sokolov, Universal fluctuations and ergodicity of generalized diffusivity on critical percolation clusters, J. Phys. A **55**, 345001 (2022).

[48] A. E. Scheidegger, An evaluation of the accuracy of the diffusivity equation for describing miscible displacement in porous media, in *Proceedings of the Theory of Flow in Porous Media Conference* (University of Oklahoma, Norman, Oklahoma, 1959), p. 103.

[49] J. Im, F. P. J. de Barros, S. Masri, and M. Sahimi (unpublished).