Machine learning framework for analysis of transport through complex networks in porous, granular media: A focus on permeability

Joost H. van der Linden, Guillermo A. Narsilio, and Antoinette Tordesillas

Phys. Rev. E **94**, 022904 — Published 17 August 2016

DOI: 10.1103/PhysRevE.94.022904

A machine learning framework for analysis of transport through complex networks in porous, granular media: a focus on permeability

Joost H. van der Linden and Guillermo A. Narsilio
Department of Infrastructure Engineering
The University of Melbourne, Australia

Antoinette Tordesillas*
School of Mathematics and Statistics
School of Earth Sciences
The University of Melbourne, Australia

(Dated: 06/07/2016)

We present a data-driven framework to study the relationship between fluid flow at the macroscale and the internal pore structure, across the micro- and meso-scales, in porous, granular media. Sphere packings with varying particle size distribution and confining pressure are generated using the discrete element method. For each sample, a finite element analysis of the fluid flow is performed to compute the permeability. We construct a pore network and a particle contact network to quantify the connectivity of the pores and particles across the mesoscopic spatial scales. Machine learning techniques for feature selection are employed to identify sets of microstructural properties and multiscale complex network features that optimally *characterize* permeability. We find a linear correlation (in log-log scale) between permeability and the average closeness centrality of the weighted pore network. With the pore network links weighted by the local conductance, the average closeness centrality represents a multiscale measure of efficiency of flow through the pore network in terms of the mean geodesic distance (or shortest path) between all pore bodies in the pore network. Specifically, this study objectively quantifies a hypothesized link between high permeability and efficient shortest paths that thread through relatively large pore bodies connected to each other by high conductance pore throats, embodying connectivity and pore structure.

I. INTRODUCTION

Transport through porous, granular systems is of central importance in a wide range of technological and engineering applications, including: ceramics [1], pervious concrete [2], hydrocarbon recovery [3], hydraulic fracking [4], geosequestration of CO_2 [5], exploitation of geothermal energy as a renewable energy source [6] and geologic disposal of radioactive waste [7]. In the energy resource sector alone, the economic cost of many processes, including ground exploration (i.e., site investigations), construction and maintenance of associated infrastructure, to risk monitoring and mitigation, runs into billions of dollars [8, 9]. Ultimately, the design and management of these processes use estimations of the hydraulic, thermal, and mechanical properties of porous, granular media (e.g., ground, concrete etc) at the macroscopic scale. In turn, robust predictions of such properties rely on fundamental knowledge of the material's internal grain and pore structure and of its influence on the efficiencies of transmission pathways for interstitial fluid flow, heat transfer, electrical flow, stress transfer, etc. [10–13].

This effort focuses on quantifying the relationship between the internal pore structure, across the micro- and meso-scales, and permeability at the macroscopic scale. Before proceeding, it is instructive to place this study in the context of the state-of-the-art, especially given the immense research attention that has been paid to the characterization and modeling of transport properties in granular media. Despite significant past efforts, many aspects that are fundamental to engineering scale transport in these materials remain poorly understood. A long standing impediment to progress has been the limited access to the internal structure of a material under load. Recent advances in nondestructive, high resolution 3D and 4D imaging, however, have rapidly overcome this limitation and are now able to deliver unprecedented detail at the scale of individual grains and pores [14–17]. Advances in post-processing techniques, such as data-constrained modeling, can now also infer submicron porosity and compositional information [18]. Such developments, coupled with data generated from high performance computing and discrete element models [19–21], have prompted a pressing need for new data-driven concepts and tools that can embrace the information embodied in these rich microstructural data sets, and uncover patterns that facilitate an understanding of how the underlying physics at the microscopic and mesoscopic scales (the cause) relate to transport phenomena at the macrocopic scale (the effect). In particular, of crucial importance to transport phenomena are patterns in the connectivity of the solid grain phase and of the interstitial pore space across the mesoscale, since these provide vital clues on the relative efficiencies of transmission pathways for different granular materials [2, 13, 22–25].

Different strategies have been employed to capitalize on the rich data sets from nondestructive, high resolu-

* atordesi@unimelb.edu.au

tion imaging techniques. One strategy has been to extract hidden patterns in the data. For example, novel spatio-temporal patterns uncovered in studies of force transmission have shed light on the underlying mechanisms of various phenomena at the macroscale, including: shear jamming [26], aging [27] and strain localization [28]. Emergent linear and cyclic mesoscale structures, which form the structural building blocks of self-organization, have been characterized (e.g., force chains, cycles) and introduced into continuum models that can capture the defining dynamics inside shear bands (e.g., [29–31]). In the engineering literature on pore fluid transport, a large body of knowledge has been gained from use of standard statistical methods to analyze microstructural data; this strategy has led to important insights on the relationships between macroscopic transport properties and structural characteristics of porous media such as: porosity and path tortuosity [32], grain size [33], particle shape [34], local pore space connectivity, e.g. through coordination number, [2, 13, 25, 35], and pore geometry [36]. What is still missing, however, are robust multiscale descriptors of pore connectivity and associated transmission pathways, and their relationship to macroscopic transport [13, 37–42]. Although techniques employing local percolation probabilities [22], the Euler characteristic [43] and n-point correlation functions [12, 44–46] have helped to fill this knowledge gap, studies continue to highlight a critical need for explicit, higher-order, three-dimensional topology and connectivity descriptors to be incorporated in predictions of permeability and thermal conductivity[13, 37–42].

A perennial challenge for characterization and modeling of phenomena involving granular media is that the internal connectivity of, and interactions between, the pores and the particles exhibit hallmarks of complexity: multiscale and nonlinear interactions that lead to patterns of self-organization at the mesoscale [47, 48]. In this study, we take the first steps in a new line of investigation which fuses modern advances in statistics (i.e., machine learning) and complex systems (i.e. complex networks) to develop a data-driven framework that is particularly suited for multiscale and nonlinear phenomena germane to complex systems. Although this study focuses solely on unraveling the details of permeability, our approach is general and applicable to studies of other transport (thermal, electrical) phenomena in porous, granular media. With respect to studies of permeability, our approach distinguishes itself from past efforts in two fronts. First, our approach exploits emerging developments in big data analytics, high resolution imaging and high performance computing — by combining discrete element methods, finite element methods, complex networks, machine learning and computerized tomography in a single data-driven platform. Second, in a first of its kind, we fuse machine learning with complex networks to establish an objective method for identifying metrics that parsimoniously characterize the internal connectivity and concomitant efficiency of transmission pathways across multiple spatial scales.

The rest of this paper is organized as follows. Because our proposed data-driven framework combines techniques and concepts from separate research disciplines, we first provide an overview of this framework along with a brief review of relevant extant literature at the start of section II, before discussing the implementation of the different components of this framework. In section III, we focus on the machine learning analysis and present a new relationship between permeability and a complex network descriptor of pore connectivity. We summarize our key findings and identify future research directions in section IV.

II. METHODS

Our framework weaves together multiple techniques into one platform. Thus, to aid understanding, we begin with an overview of the main components, placing these in the context of relevant past work and the most pressing research needs, before providing details of the implementation each component in subsequent subsections.

A. Proposed framework

The framework is divided into three components (Figure 1). The first component delivers the complete data to be analyzed, i.e., the feature set, comprising the 'input variables' and the 'output variable' (steps 1-4, Figure 1). The input variables consist of two groups of data. The first group consists of the raw high resolution data at the level of individual pores and particles; the second group consists of multiscale complex network metrics that include connectivity descriptors. Those in the second group utilize information from the raw data in the first group. Although the generation of high resolution imaging data sets in the first group is still prohibitively expensive for many real materials, there is a clear trend towards these data sets becoming increasingly accessible and routine [14, 15]. High resolution imaging may guide the DEM simulations, as, for example, shown by Delaney et al. [49], or potentially replace these altogether. In anticipation of such data capability and assets, we thus envisage that this framework may ultimately be applied to microstructural data of real porous, granular media samples using microstructural data gathered *directly* from high resolution imaging techniques (e.g. X-ray micro computed tomography). In the present study, however, in order to perform the requisite machine learning analysis, $\mathcal{O}(100)$ samples are needed. Repeated use of (high resolution) imaging equipment and post-processing software for large quantities of samples is presently costly.

Consequently, for this study, we resort to artificially generated samples of porous, granular media to gather data spanning micro-, meso- and macro- scales: see

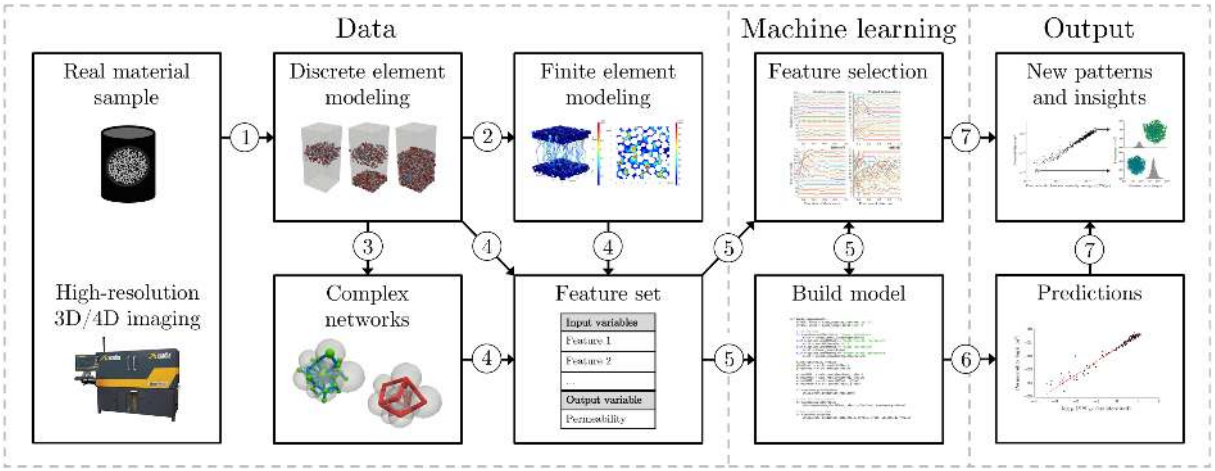

FIG. 1. (Color online) Real sample parameters, partially obtained from high resolution imaging, are used in the DEM simulation to generate realistic artificial samples (step 1). Fluid flow is simulated with a finite element method to compute the permeability (step 2). A pore network and contact network are constructed to compute multiscale complex network variables (step 3). Micro-meso-macro data comprise the physical properties at the pore and grain scale, the network variables and permeability (step 4). The resulting feature set is used for feature selection and model construction (step 5) to generate predictions (step 6) and new insights (step 7).

steps 1-3, Figure 1. The microstructure of samples can be generated using stochastic reconstruction methods [45, 50, 51], (non-)ballistic procedures [52–54], process-based reconstruction methods [36] and discrete-element methods (DEM) [55]. Each technique has its own strengths and limitations. For example, a proper stochastic reconstruction requires high computational effort (for the simulated annealing) [56] and does not capture the dynamic processes that precede the creation of porous, granular media [39]. For this work, we choose DEM for its simplicity, reproducibility, broad acceptance and extensibility (step 1, Figure 1). In its simplest form, spherical particles may represent old alluvial deposits (porous, granular systems of typically high sphericity and low angularity). Assemblies of different grain shapes and degrees of (interparticle) cementation may be modeled within DEM with clusters of spheres that more realistically represent non-spherical grains [19]. In addition, DEM may be used to capture samples with evolving fracture patterns (e.g., [24, 57]). As a reference case for our DEM samples, we include a single monodisperse sphere packing for which the centroids and radii have been determined using X-ray computed tomography [49, 58–60].

To capture pore connectivity, i.e. higher-order, three-dimensional topology and geometry [13, 39], we opt for a new class of multiscale connectivity descriptors from complex networks [61]. These will generate meso-scale data (step 3, Figure 1) – in addition to the aforementioned physical properties of the constituent grains and pores that form our initial input data. Complex network theory opens a new avenue for multiscale characterization of fluid flow phenomena in porous, granular media. Using a shortest paths analysis of the pore network, for example, a region of efficient transport in the shear bands of deforming, dense granular media was identified [57]. More recently, Russell et al. [24] uncovered optimized flow pathways that are driven by complex jamming-unjamming dynamics unique to shear bands, giving explicit structural insights into causes of enhanced flow and permeability in fractured media. They also provide a template for the abstract representation of the pore space in a three-dimensional granular assembly using concepts from dual graphs. In this work, however, we construct a pore network that more closely represents the physical pore domains in a manner similar to those adopted in past network models of porous, granular media [23, 62–64]. Finally, at the macroscopic level, we compute the permeability of each of our samples by performing a finite element simulation of the fluid flow through the pore space, using a model that has been validated against physical experiments [11, 65–67] (step 2, Figure 1).

Having collated all the data on material properties spanning micro-, meso- and macro- scales (step 4, Figure 1), we next employ machine learning techniques to establish *objectively* a parsimonious relationship between permeability (i.e., the output variable) and the internal structure of our granular samples (i.e., the input variables): see step 5, Figure 1. Distinct from traditional curve fitting of data (e.g. Hazen formula, Archie's law) which seeks to establish unknown parameters based on a known model function derived from theory and/or experiments, here we use machine learning to establish the model function itself from the data. Machine learning provides a rigorous statistical framework for analysis of complex data sets, such as noisy, high dimensional data, through: feature selection (i.e., finding a subset of relevant and non-redundant input variables that can best

predict a given output variable), model construction and error and uncertainty quantification [68, 69]. The use of machine learning techniques has precedence in studies of materials and transport phenomena. Ma et al. [70] presented a machine learning framework to classify and predict fluid flow properties of stochastically reconstructed rocks, studying the relationships with geometry, topology and statistical correlation functions. Xu et al. [71] predict the damping parameters of polymer nanocomposites, using correlation functions, particle shape descriptors and pore size descriptors. Feature selection for materials science has been explored by Ghiringhelli et al. [72]. Khandelwal [73] used machine learning to predict the thermal conductivity of rocks, based on the uniaxial compressive strength, density, porosity and P-wave velocity. Machine learning has also been used extensively with (macro-scale) soil survey data for pedo-transfer functions [74, 75] and in geotechnical applications such as slope stability and liquefaction [76].

In our proposed framework, we use machine learning to establish a model of, and new insights on internal structural features that define, permeability (steps 6 and 7, Figure 1). Although we predict the permeability at the end, our main objective is to characterize permeability through feature selection. We proceed in two phases. In the first phase, we choose feature selection methods that are appropriate for the dataset at hand, to identify a non-redundant subset of the most relevant properties, including novel descriptors of connectivity, to characterize permeability. Using more than one feature selection algorithm allows us to investigate feature 'importance' from various angles: that is, we can rule out, and thus ensure our conclusions are robust to, algorithmic artifacts and assumptions. In addition, we assess our results against well established microstructural properties known to influence permeability (e.g. void ratio). In the second phase, we use the selected features in a predictive model and employ techniques, such as cross-validation, to quantify the uncertainty in our predictions. We discuss the methods from Figure 1 in detail in the upcoming subsections, in the order illustrated in Figure 1. The terms 'variable' and 'feature' are used interchangeably.

B. Discrete Element Modeling

We develop a model for Ottawa sand and sandstone, comprising rounded quartz particles [10], for which DEM can provide an reasonable approximation of real geomechanical behavior [20, 67]. The simulation is implemented in `Yade` [77]. A total of 536 packings are generated using the simulation parameters summarized in Table I. The friction angle is drawn from a uniform distribution in an expanded range similar to the range used by Garcia et al. [78] (between approximately 5.7° and 24.2°), encapsulating the commonly used quartz friction angles reported by Procter and Barton [79]. For simplicity, a uniform grain size distribution with mean radius 0.5 mm is used, varying the extrema of the distribution between the mono-dispersed packing $U(0.5, 0.5)$ mm and the poly-dispersed packing $U(0.2, 0.8)$ mm. The Young's modulus is set to an artificially low value to generate a wide void ratio distribution. Samples under low confining pressure (small particle overlap) approximate Ottawa sand, while samples under higher (e.g. $> 10^6$) confining pressure (larger particle overlap) act as a simple proxy for sandstone.

TABLE I. Simulation parameters used in DEM.

Number of grains	4000
Grain shape	spherical
Density [kg/m^3]	2650
Young's modulus [Pa]	10^8
Poisson's ratio	0.2
Friction angle [deg]	$\theta \in U(5.7°, 31°)$
Grain radius [mm]	$U(0.5-\alpha, 0.5+\alpha)$, $\alpha \in U(0.0, 0.3)$
Confining pressure [Pa]	10^n, $n \in U(5, 7)$

In each DEM simulation, a rectangular box with base dimensions 15 by 15 mm is created. Periodic boundary conditions are imposed on the four vertical plane boundaries. After drawing the friction angle, grain size range and confining pressure from their respective distributions shown in Table I, the DEM simulation proceeds in two stages. In the first stage, shown in Figure 2(a), 400 grains (10%) are placed randomly, without overlap, slightly above the box floor. After placement, the grains fall and settle under gravity. When the unbalanced force (ratio of the average contact force and average per-body force) reaches a small threshold value, a new batch of 400 particles is introduced. Care is taken to introduce each batch at equal height from the top of the settled packing. This process is repeated ten times, until all 4000 particles are settled. Repeatedly settling batches of grains effectively simulates the air pluviation method that is used in the preparation of laboratory soil samples [80]. Grains are not frozen during the simulation. In the second stage, shown in Figure 2(b), the packing from the first stage is subjected to an isotropic, confining pressure, effectively reducing the porosity. `Yade` approximates a quasi-static equilibrium condition of the packing by reducing the loading velocity while approaching the goal confining pressure. The simulation is terminated when the unbalanced force reaches a small threshold value.

In addition to the 536 packings generated with our DEM simulation, we include a reference case of a real sphere packing. Using X-ray computed tomography, Aste et al. [58, 59] developed a technique to extract sphere centroids and raddii from real packings of glass beads. We use one of the stationary samples (FB18) described in [49, 60]. In a settlement process somewhat similar to our gravity deposition, approximately 1.5×10^5 monodisperse particles in a fluidized column are subjected to a flow pulse from the bottom, after which the particles settle in a mechanically stable configuration. We scale the

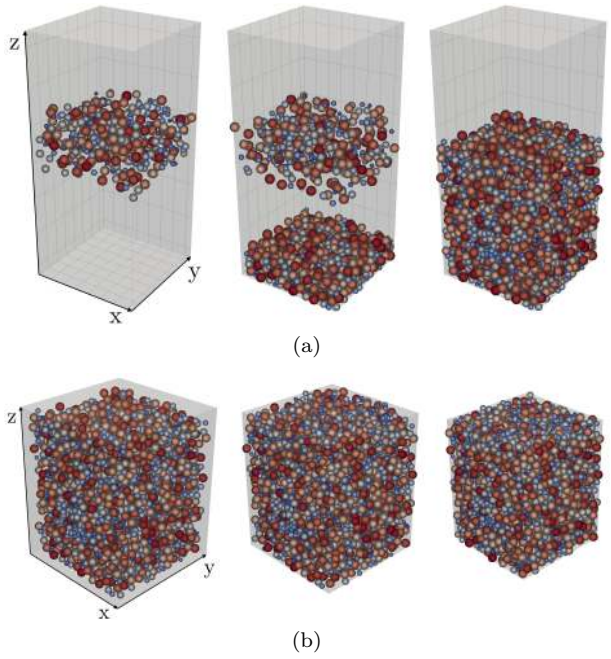

FIG. 2. (Color online) (a) Gravity deposition. Batches of grains are sequentially added to a rectangular box and settle under gravity. (b) Triaxial compression. A confining pressure is applied along the x, y and z axes.

packing to have the same mean particle radius as our DEM samples (0.5 mm) and extract a small cubic REV (with length 13% of the original x length) from the center of the packing.

C. Finite Element Modeling

The permeability computation consists of three stages, summarized in Table II. Firstly, to control computational costs, a representative element volume (REV) is subsampled from the center of each DEM sample. Because the sample limits vary, depending on the amount of confining pressure applied in the triaxial compression, the REV length, height and width are taken proportionally to the original sample limits. The fraction (45%) is determined from a mesh-convergence study, in which we increased the REV size until the permeability and porosity converged. Secondly, we mesh the 536 subsamples using finite elements in `Simpleware ScanIP` [81]. Lastly, the fluid flow simulation is performed in `COMSOL Multiphysics` [82]. The simulation solves the governing Navier-Stokes equations, assuming the flow is incompressible and isotropic, assuming the fluid (water) is Newtonian, and assuming a no-slip boundary condition on the solid surfaces. The permeability is obtained by modifying Darcy's law to [66]

TABLE II. Preprocessing, meshing and simulation settings.

Preprocessing	REV %	45%
ScanIP	Mesh algorithm	`+FE Grid`
	Mesh type	Smoothed
	Elements	Linear
	Minimum quality	0.1
COMSOL	Side BC	Symmetric
	Top pressure	10 Pa
	Top BC	Inlet
	Bottom pressure	9 Pa
	Bottom BC	Outlet
	Fluid dynamic viscosity	0.001002 Pa s
	Fluid unit weight	9.789×10^3 N m^{-3}
	Linear solver	Direct (LU)

$$k = \frac{\eta}{\gamma}\frac{\bar{v}}{\bar{i}} = \frac{\frac{\eta n}{A_V}\int_{A_V} v_z \, dA_V}{(\Delta p/L)} \tag{1}$$

where k [m^2] is the numerically computed permeability and n is the porosity of the sample. The fluid properties are the dynamic viscosity η [Pa s^{-1}] and the unit weight γ [N/m^3]. A pressure difference Δp [Pa] is imposed over the sample length L [m] along the z-axis and the vertical velocity v_z [m/s^{-1}] is averaged over an x-y plane with void area A_V [m^2]. The variables $\bar{v}$ and $\bar{i}$ are the averaged Darcy velocity and hydraulic gradient, respectively. The value of k is computed for both the inlet- and outlet-plane, and subsequently averaged, similar to work by Narsilio et al. [66]. Because our permeability data spans three orders of magnitude and we wish to predict each magnitude equally well, we consider the natural logarithm of the permeability in the upcoming analysis.

An example of the resulting mesh is shown in Figure 3, along with several fluid flow streamlines and a vertical slice of the velocity field. Red colors indicate higher velocities.

D. Complex networks

For each sample, a weighted *contact network* and weighted *pore network* are constructed. In order to relate complex network features to the permeability, it is crucial that the network weighting is physically representative. Ideally, the weighting is, by itself, strongly related to the permeability. Our choices for the representation and weighting are summarized in Table III. The conductance weighting is outlined in Appendix A. In the next two sections, the network construction and derived variables are discussed in detail.

Network construction

The contact network is constructed by assigning a node to each grain and an edge if the corresponding grains touch. The edges in the contact network are weighted with the contact area. To construct the pore network, nodes are assigned to pores, connected by an edge if the corresponding pores share a throat.

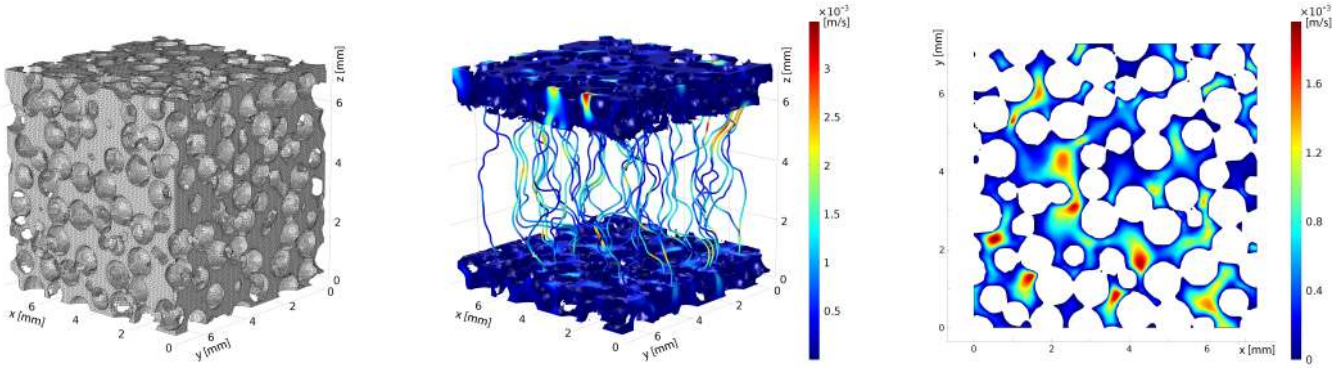

FIG. 3. (Color online) The finite element mesh, an illustration of the fluid flow paths from top (high pressure) to bottom (low pressure) and a cross-section of the velocity field.

TABLE III. Network construction. Edges in the pore network are weighted by local conductance, while edges in the contact network are weighted by contact area.

	Pore network	Contact network
Node representation	Pores	Particles
Node features	Pore void ratio	Grain size
	Surface area	Surface area
Edge representation	Throats	Particle contacts
Edge features	Conductance	Contact area
	Throat void ratio	

What constitutes a 'pore' and 'throat' remains ambiguous [83]. We opt to use the modified Delaunay tessellation approach by Al-Raoush and Willson [64], with several adaptations [84]. Similar to their approach, we assume pore bodies are encapsulated by (merged collections of) tetrahedra. Each tetrahedron consists of four triangular faces. Pore throats are found on shared faces of tetrahedra. Our approach to constructing the pore network proceeds as follows:

1. A Delaunay tessellation is constructed first, using the centroids of the grains.

2. Rather than using a non-linear optimization procedure with inscribed spheres, a conceptually simpler approach is introduced. A pair of tetrahedra in the Delaunay tessellation is merged if the areal porosity of the shared face is higher than a certain threshold. The merging procedure is illustrated in Figure 4 for a simple setup of six spheres and three tetrahedra.

3. Nodes are assigned to (merged collections of) tetrahedra, representing the pore bodies. Next, each pair of tetrahedra (collections) is connected with an edge if they share a face, representing pore throats.

4. The boundaries of the pore volume are the surface of the grains and the throats. Rather than defining pore volume by an inscribed sphere, we record the pore void ratio as the fraction of void and solid volume in the (merged collection of) tetrahedra. The pore surface area is computed as the area of grains exposed to void space inside the (merged collection of) tetrahedra. In addition, we compute the throat void ratio on the faces of the (merged collections of) tetrahedra (see n in Figure 4). Lastly, to construct a network weighting related to the permeability, we compute the local conductance in a tube model of adjacent pores and throats. For more details on the conductance computation, we refer to Appendix A.

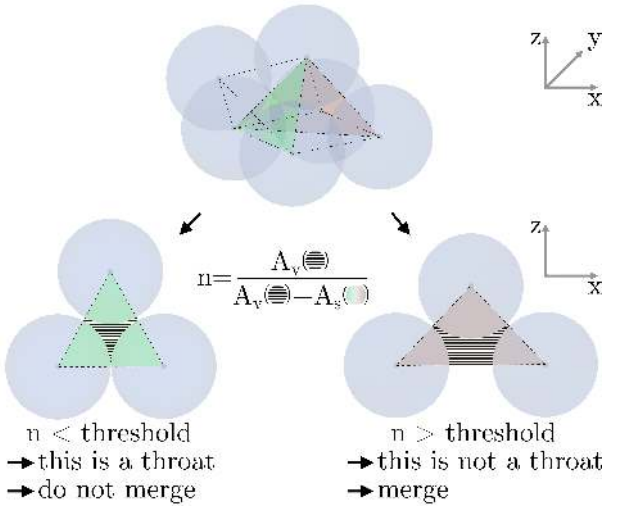

FIG. 4. (Color online) Depiction of the construction of the pore network. Two faces are isolated and the void area and solid area (in 2D) on the faces is determined. Then, the areal porosity of each face is compared to the threshold ϵ. In this example, only the adjacent tetrahedra for the red (right) face are merged.

The only hyperparameter in the pore network construction, the porosity threshold ϵ, is set to 0.4. Based on our experience, this choice results in an reasonable

distinction between pores and pore throats. Our implementation (using the computational geometry library `CGAL` [85]) avoids voxelation in most cases, except for the pore volume calculations. An example of the contact network and pore network is shown in Figure 5. By

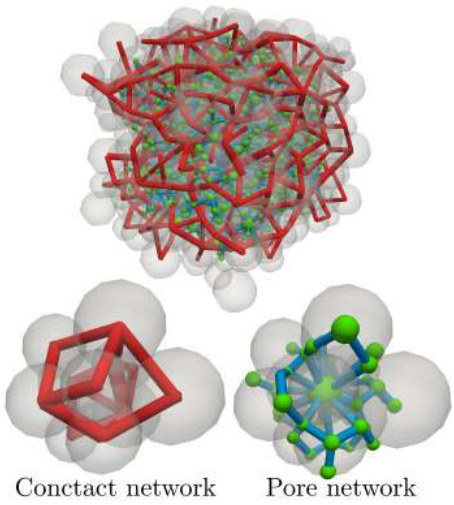

FIG. 5. (Color online) Example of the contact network and pore network. Nodes in the pore network are visualized as spheres with (scaled) equivalent volume to the corresponding pore void volume. The pore network construction correctly identified both the particle connections on the bottom left and the large pore body in the center of the excerpt on the bottom right.

avoiding inscribed spheres or medial axes and adopting the aforementioned merging criterion, we gain a useful, physically representative distinction between pores and throats based on simple surface areas (throats) and volumes (pores).

Network features

Having established the network construction procedure, we discuss the complex network features next. Denote (V, E) as the set of vertices and edges in the network, respectively. The length of a weighted shortest path between nodes $i, j \in V$, denoted $d(i, j)$, is the path from i to j that minimizes the summed weights of the traversed edges. Because higher conductance generally corresponds to more flow, but shortest paths are computed through minimization of the edge weighting, edges in the pore network are assigned the reciprocal of the conductance as a weighting. Similarly, in preparation for future applications to heat flow, edges in the contact network are assigned the reciprocal of the contact area as a weighting. A total of eight network properties are computed, both for the contact network and the pore network.

For a full review of complex network theory, developments and applications, refer to the work by Newman [61]. We compute both the *degree*, i.e. the number of edges adjacent to a node, and the *weighted degree*, which is the sum of the edge weights adjacent to a node. In the granular media research community, the degree is often referred to as the coordination number. In the pore network, degree represents the number of throats for a particular pore, while in the contact network, degree is the number of particle contacts for a particular particle. We also compute the *network density*, not to be confused with the packing density, as the ratio $2|E|/(|V|(|V|-1))$ of potential edges over actual edges in the network. The *network diameter* is calculated as $\max_{i,j\in V} d(i,j)/(|V|-1)$, finding the length of the 'longest' shortest path between all pairs of nodes in the network. The *betweenness centrality* quantifies the fraction of shortest paths passing through a particular node $i \in V$ [86],

$$C^B_{node}(i) = \beta \sum_{j,k\in V} \frac{\sigma(j,k|i)}{\sigma(j,k)},$$

where $\sigma(j,k)$ is the total number of shortest paths between node j and k, $\sigma(j,k|i)$ is the number of shortest paths between j and k that pass through i and $\beta = 1/((|V|-1)(|V|-2)/2)$ is a normalization term equal to the number of pairs of nodes excluding i. The edge betweenness centrality $C^B_{edge}(e)$ for edge $e \in E$ is computed by computing $\sigma(j,k|e)$ as the number of shortest paths passing through edge e and setting $\beta = 2/(|V|(|V|-1))$. For the pore network, high values should indicate that the corresponding pore (*node betweenness*) or pore throat (*edge betweenness*) is 'important' for the fluid flow. Finally, we compute the *closeness centrality* for each node $i \in V$ as the reciprocal of the summed shortest path distances to all other nodes $j \in V$ [87],

$$C^C(i) = \beta \left[\sum_{j=1}^{|V|-1} d(i,j) \right]^{-1},$$

where $\beta = |V| - 1$ is the normalization term. High closeness centrality indicates a 'central' pore and, again, hints towards a strong contribution to the fluid flow. The degree, weighted degree, betweenness centrality and closeness centrality can be averaged over all nodes/edges to obtain a global network feature. For example, high average closeness centrality may indicate relatively 'short' shortest paths throughout the network, hinting towards a more permeable sample with many large throats.

E. Feature Set

Referring back to step 4 in the overview Figure 1, the next step is to extract relevant physical features from the DEM packing and connectivity features from the network representations. The physical features include the global void ratio, local (pore, throat) void ratio, pore surface area, specific surface area and the coefficients of uniformity and curvature. We also compute the throat to pore volume ratio, where the throat volume is equal to the

volume of a sphere with radius equal to the radius of the throat area represented as a circle, and record the confining pressure, friction angle and the grain radius distribution range α (see Table I) for each sample. In terms of network descriptors, we compute all fatures listed in the previous section.

The full feature set is presented in Table IV as our educated, physically-inspired 'initial guess' of features that could be relevant in characterizing and predicting the permeability. The physical features (#1-13) include pore, throat and grain geometry, as well as several packing features and DEM input features. None of these features, however, address the connectivity in the packing, which is highly-relevant aspect for the permeability [13]. To this end, we include the complex network features (#14-27) in our feature set, which are inherently multiscale and are able to succinctly describe connectivity of pores (pore network) and grains (contact network). Note that, in Section III, quantities are made dimensionless in the plots, using (depending on the units) the mean particle diameter (d_{50}) and the dynamic viscosity η ($d_{50} = 0.001$ m, refer to Table I, and $\eta = 0.001002$ Pa s, refer to Table II). We experimented with various distribution indicators, such as the mean, variance, skewness, kurtosis and percentiles. For simplicity, we include only the mean μ for each distribution, averaging over the corresponding entity (throats, pores, particles, nodes or edges). We found that, although the other distribution indicators do reveal some interesting relationships, the main conclusions of this work hold when only the mean is used. Given the small length-scale under consideration, and the homogeneity of our samples, averaging is a valid and straightforward method to reduce each distribution to a single parameter. The features $p, \mu[T]_K, \mu[B]_{A_c}, \mu[G^c]_{\kappa_w}, \mu[G^c]_{C^C}, \mu[G^p]_{\kappa_w}$ and $\mu[G^p]_{C^C}$ are found to span multiple orders of magnitude. In order to weigh different magnitudes *within* these features equally, we compute the natural logarithm. Finally, in order to weigh different magnitudes *between* different features equally, we standardize each feature by subtracting the mean and dividing by the standard deviation, as is standard practice in machine learning:

$$\tilde{x}_i = \frac{x_i - \mu(X)}{\sigma(X)}$$

where $X = (x_1, \ldots, x_N)$ is a feature vector with N values (for N packings) and μ and σ are the mean and standard deviation, respectively. Figure axis used in Section III are shown in original scales, however, for a more physically meaningful discussion.

TABLE IV. Feature notation. Note that we use $[X]_a$ to denote a distribution of parameter a of an entity X to emphasize the difference between scalars (no brackets) and distributions.

#	Notation	Entity	Attribute	Units
1	e	packing	void ratio	
2	p	packing	confining pressure	[Pa]
3	ssa	packing	specific surface area	[m^{-1}]
4	$[T]_V/[P]_V$	throat/pore	throat/pore volume ratio	
5	$[T]_K$	throat	conductance	[$\mathrm{m}^3\mathrm{Pa}^{-1}\mathrm{s}^{-1}$]
6	$[T]_e$	throat	void ratio	
7	$[P]_e$	pore	void ratio	
8	$[P]_{A_s}$	pore	surface area	[m^2]
9	$[B]_{A_c}$	particle	contact area	[m^2]
10	α	particle	grain size range	[m]
11	c_u	particle	coefficient of uniformity	
12	c_c	particle	coefficient of curvature	
13	θ	particle	friction angle	[deg]
14	G^p_ρ	pore net.	network density	
15	G^p_D	pore net.	network diameter	[m^{-3}Pa s]
16	$[G^p]_\kappa$	pore net.	degree	
17	$[G^p]_{\kappa_w}$	pore net.	weighted degree	[m^{-3}Pa s]
18	$[G^p]_{C^B_{edge}}$	pore net.	edge betwnness centrality	
19	$[G^p]_{C^B_{node}}$	pore net.	node betwnness centrality	
20	$[G^p]_{C^C}$	pore net.	closeness centrality	[$\mathrm{m}^3\mathrm{Pa}^{-1}\mathrm{s}^{-1}$]
21	G^c_ρ	contact net.	network density	
22	G^c_D	contact net.	network diameter	[m^{-2}]
23	$[G^c]_\kappa$	contact net.	degree	
24	$[G^c]_{\kappa_w}$	contact net.	weighted degree	[m^{-2}]
25	$[G^c]_{C^B_{edge}}$	contact net.	edge betwnness centrality	
26	$[G^c]_{C^B_{node}}$	contact net.	node betwnness centrality	
27	$[G^c]_{C^C}$	contact net.	closeness centrality	[m^2]

F. Feature Selection

In the presence of a large feature set, such as the one presented in the previous section, we aim to uncover the most 'important' features for the permeability in an objective manner using feature selection. Formally, given N *instances* of M features $\boldsymbol{F} = \{X_i; i = 1, \ldots, M\}$, the objective of feature selection is to find a subset $\boldsymbol{S} \subseteq \boldsymbol{F}$ with m features that 'optimally' characterizes a *target* variable Y [88]. Each instance corresponds to a packing, and the target variable is, in our case, the natural logarithm of the permeability. We present four feature selection algorithms of increasing complexity. The first three (Kendall correlation, mutual information and mRMR) are *myopic*, i.e. conditional dependencies between features are ignored. The fourth algorithm (RReliefF) is *non-myopic*.

Kendall rank correlation

In contrast with the Pearson correlation coefficient, the Kendall rank correlation coefficient (Kendall's τ) can measure non-linear dependence between variables [89]. Let $X = (x_1, \ldots, x_N) \in \boldsymbol{F}$ be the values of a certain feature X and $Y = (y_1, \ldots, y_N)$ the target variable values, and define N_c as the number of concordant pairs ($x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$) and N_d as the number of discordant pairs ($x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$

and $y_i > y_j$). Then, $-1 \leq \tau \leq 1$ is defined as

$$\tau = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}$$

Values close to 1 or -1 indicate a good agreement between rankings, generally indicating a strong relationship between the feature X and target variable Y. We employ Kendall rank correlation because of its initial simplicity and its ability to identify non-linear relationships.

Mutual information

Let X and Y be two random variables with joint probability $p(X,Y)$ and marginal probabilities $p(X)$ and $p(Y)$, then the mutual information $I(X;Y)$ is defined as

$$I(X;Y) = \int_Y \int_X p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} dxdy$$

Computing the integrals is often difficult with a limited number of instances [88]. As a solution, continuous variables can be discretized and the mutual information is computed as

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}$$

Mutual information measures the degree of mutual dependence between X and Y. For the discretization, we experiment with a range of different techniques, including equal-sized bins, percentile bins and mean-based splittings using $\mu \pm a\sigma$, where μ is the mean of a feature, σ is the standard deviation and a is a tuning parameter. We find that k equal-sized bins between the minimum and maximum of each feature deliver robust results, particularly when the results from this discretization are averaged for $k = 5, 6, \ldots, 150$. The permeability is discretized using 50 equal-sized bins between the minimum and maximum. Although some information is lost by discretizing continuous variables, mutual information is included in our study because it can identify non-linear, non-monotonic relationships.

minimum-Redundancy, Maximum Relevance

The mRMR method [88] is essentially an optimization procedure with two objectives:

1. To *maximize dependency*, that is, find features that are strongly correlated with the target variable,

$$\max D(\boldsymbol{S}, Y),\ D = \frac{1}{|\boldsymbol{S}|} \sum\nolimits_{X_i \in \boldsymbol{S}} I(X_i;Y)$$

2. To *minimize redundancy*, that is, avoid features in $\boldsymbol{S}$ that are highly correlated among themselves,

$$\min R(\boldsymbol{S}),\ R = \frac{1}{|\boldsymbol{S}|^2} \sum\nolimits_{X_i, X_j \in \boldsymbol{S}} I(X_i;X_j)$$

An exhaustive search of all possible subsets $\boldsymbol{S} \subseteq \boldsymbol{F}$ is often computationally unfeasible. Hence, in practice, an incremental search method is used. First, set $\boldsymbol{S}_0 = \{X_{i_0}\}$ where $X_{i_0} = \arg\max_{X_i \in \boldsymbol{F}} I(X_i;Y)$. We then incrementally add a feature to the current subset $\boldsymbol{S}_{k-1}, k \geq 1$ with the criterion

$$\max_{X_i \in \boldsymbol{F} \setminus \boldsymbol{S}_{k-1}} \left[I(X_i;Y) - \frac{1}{m-1} \sum\nolimits_{X_j \in \boldsymbol{S}_{k-1}} I(X_i;X_j) \right] \quad (2)$$

Note that the mRMR ranking should be interpreted collectively. That is, for a ranking of three features, the combination of ranked features 1 and 2 may be better than feature 3, with respect to characterizing the permeability, but this does not mean that feature 3 is less relevant (individually) than feature 1 or 2. Instead, the feature score in the incremental optimization procedure is based on the relevance to the permeability *and* the degree of redundancy with the already selected features in $\boldsymbol{S}_{k-1}$. We discretize all features and the permeability using 50 equal-sized bins between the minimum and maximum. We use mRMR because the method combines the strengths of mutual information (non-linear, non-monotonic relationships) with the ability to maximize dependency and minimize redundancy.

RReliefF

The last feature selection algorithm under investigation is the RReliefF method [90–92]. The family of Relief methods estimate a feature's importance $W[X]$ based on its ability to separate values of the target variable, approximating the following difference of probabilities [91]

$$\begin{aligned} W[X] = &P(dissimilar X \mid dissimilar Y) \\ &- P(dissimilar X \mid similar Y) \end{aligned}$$

In words, a feature X is rewarded for separating dissimilar values of Y and penalized for separating similar values of Y. Given two instances I_1 and I_2, 'similar' and 'dissimilar' for either the feature X or target Y is defined using the distance function

$$\text{diff}(Z, I_1, I_2) = \frac{|\text{value}(Z, I_1) - \text{value}(Z, I_2)|}{\max(Z) - \min(Z)}$$

for continuous features, where $\text{value}(Z, I)$ is the value of $Z \in \{X, Y\}$ in instance I. For a detailed discussion of the implementation of RReliefF, which involves the use of an exponential correction to the distance function and k-nearest neighbors to improve robustness, we refer to the overview by Robnik-Šikonja and Kononenko [93]. In our implementation, we use the parameters $k = 70$ nearest neighbors and $\sigma = 20$ in the exponential distance function, as recommended by the authors. We select RReliefF because it can detect conditional dependencies between features given the target variable values, a highly desirable property in feature sets with strong dependencies, because two features that appear useless individually may be useful together [94]. A drawback is that RReliefF, in contrast with mRMR, does not detect feature redundancy.

G. Prediction

Characterizing the permeability through feature selection is the main objective of this work. The final stage of prediction, though, is included for completeness and as an experiment to quantify the predictive capability of the chosen features. Traditionally, in *supervised* machine learning, we *train* (or *fit*) a predictive model using both the input (values for selected features) and output (permeability values) of the model, and subsequently ask the model to predict (or *test*) the permeability given a collection of unseen instances. In this context, machine learning mitigates limited understanding of the fundamental, underlying governing equations of a system by performing data-driven predictions [76]. Even though we deliberately made no assumptions regarding the linearity of the relationships between features and the permeability, we restrict our use of predictive methods to linear regression for simplicity. As will become clear in the results section, many high-scoring features have, in fact, a linear relationship with the permeability, for which linear regression suffices. For linear regression theory, we refer to [68, 69]. In this section, we discuss the validation methods used to quantify the error and uncertainty in our predictions of the permeability.

Cross-validation

A common problem of machine learning in scarce data settings is that the test set might not be sufficiently large to provide a robust estimate of the generalization performance. The most widely used method to remedy this issue is K-fold cross-validation, for which the data is split into K parts of equal size [69]. We run an iterative procedure for $k = 1, \ldots, K$, where, in iteration k, the model is trained using $k-1$ parts and tested on the single, remaining part. By averaging the resulting K test set scores, we obtain a more robust estimate of the generalization performance. Feature selection may be combined with cross-validation by running the feature selection algorithm on the $k-1$ parts before training the model [69, Sec. 7.10.2]. Note that because the feature selection is repeated K times, rankings may differ for different folds. We expect the differences to be minimal, however, if the feature selection method is robust and sufficient data is available. Cross-validation should be understood as a method to evaluate the process of fitting a model, rather than evaluating the model itself [95].

Assessment

To study the generalization performance of a model, we use two commonly-used performance indicators. The root-mean-squared error (RMSE) is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{3}$$

where $\hat{y}_i$ is the predicted value and y_i is the measured value for each of the N test set instances. We also report the coefficient of determination R^2, defined as,

$$\mathrm{R}^2 = 1 - \frac{u}{v},\ u = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2,\ v = \sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{4}$$

where $\bar{y}$ is the mean of the measured values.

III. RESULTS AND DISCUSSION

Feature selection

Before we highlight the key insights from our analysis, we present the broad results first. The feature selection scores for Kendall rank correlation, mutual information and RReliefF, applied to the feature set from Section II E and permeability values from Section II C, are summarized in Table V. Key parameters, discussed in the subsequent analysis, are highlighted. We make a number of observations. Firstly, we observe a similar top ten for Kendall correlation and mutual information, and a slightly different top ten for RReliefF. We attribute this distinction to the ability of RReliefF to detect conditional dependencies, whereas Kendall correlations and mutual information are purely myopic methods. The best scoring feature in RReliefF (throat conductance $[T]_K$) can be thought of as having both a strong individual dependency and conditional dependency (combined with other features) on the permeability. A difference between RReliefF and the other two rankings is the confining pressure p. Further inspection of this feature reveals that p is heavily penalized for not having a clear relationship with the permeability for low confining pressures ($p < 10^6$ Pa). For the mutual information scores, we observe that the standard deviation over various binnings is relatively small compared to the mean scores, indicating a robust ranking.

Secondly, in terms of network parameters, the pore network closeness centrality $[G^p]_{C^C}$ receives high scores in all three methods, indicating the importance of this feature for the permeability. Not surprisingly, the edge weighting for the pore network, $[T]_K$, scores high in all three methods as well. Other important network features appear to be the degree $[G^c]_\kappa$ in the contact network and the weighted degree $[G^p]_{\kappa_w}$ in the pore network. We observe that the betweenness centrality receives medium scores and the network diameter and network density receive relatively low scores in all three feature selection methods. Based on these observations, the closeness centrality in the pore network appears to be the most promising network feature to predict the permeability.

Thirdly, in terms of physical features, the local pore void ratio $[P]_e$, global void ratio e and local throat void ratio $[T]_e$ are given relatively high scores in all three methods, confirming the well-known importance of the void ratio. The friction angle θ and parameters related to the grain size distribution (c_u, c_c and α) receive low scores in all three algorithms. We attribute this result to the fact that these parameters are only suitable to predict the permeability if all other parameters of the porous

TABLE V. Ranked scores assigned to each feature for three feature selection algorithms. For the notation, refer to Table IV.

	Kendall Correlation				Mutual Information[a]				RReliefF			
1	ln	μ	$[G^p]_{C^C}$	.878	ln	μ	$[G^p]_{C^C}$	.412 ± .017	ln	μ	$[T]_K$	.313
2		μ	$[P]_e$	.848	ln	μ	$[T]_K$	.374 ± .013	ln	μ	$[G^p]_{C^C}$	.290
3		μ	$[G^c]_\kappa$	−.848		μ	$[P]_e$	.365 ± .017			e	.237
4	ln	μ	$[T]_K$	.847			e	.361 ± .014		μ	$[T]_e$	.233
5			e	.842		μ	$[G^c]_\kappa$	.358 ± .016	ln	μ	$[G^p]_{\kappa_w}$	.223
6		μ	$[G^c]_{C^B_{edge}}$	.828		μ	$[T]_e$	.347 ± .017		μ	$[P]_e$	.175
7		μ	$[T]_e$	.816	ln	μ	$[G^p]_{\kappa_w}$	.336 ± .023		μ	$[G^c]_\kappa$	.160
8	ln	μ	$[G^p]_{\kappa_w}$	−.787	ln		p	.323 ± .033			ssa	.140
9	ln		p	−.775		μ	$[G^c]_{C^B_{edge}}$	.318 ± .024			G^c_ρ	.123
10			G^c_ρ	−.732	ln	μ	$[B]_{A_c}$	.296 ± .041		μ	$[G^p]_{C^B_{edge}}$	.096
11	ln	μ	$[B]_{A_c}$	−.723			G^c_ρ	.291 ± .034		μ	$[G^p]_{C^B_{node}}$	.093
12		μ	$[G^c]_{C^B_{node}}$	.723	ln	μ	$[G^c]_{C^C}$	.280 ± .036		μ	$[T]_V/[P]_V$	.081
13	ln	μ	$[G^c]_{C^C}$	−.704			ssa	.271 ± .046		μ	$[P]_{A_s}$	.052
14			G^c_D	.623		μ	$[G^c]_{C^B_{node}}$	.270 ± .039			G^p_ρ	.049
15	ln	μ	$[G^c]_{\kappa_w}$	.607		μ	$[G^p]_{C^B_{node}}$	.252 ± .048			θ	.045
16			ssa	.594		μ	$[G^p]_{C^B_{edge}}$	.246 ± .049		μ	$[G^c]_{C^B_{edge}}$	.036
17		μ	$[T]_V/[P]_V$	.572	ln	μ	$[G^c]_{\kappa_w}$	.243 ± .049		μ	$[G^c]_{C^B_{node}}$	.027
18		μ	$[G^p]_{C^B_{node}}$	−.547		μ	$[P]_{A_s}$	.232 ± .059			c_u	.024
19		μ	$[G^p]_{C^B_{edge}}$	−.513			G^c_D	.227 ± .045		μ	$[G^p]_\kappa$	.004
20		μ	$[P]_{A_s}$	.467			G^p_ρ	.219 ± .066			α	.002
21		μ	$[G^p]_\kappa$	−.465		μ	$[T]_V/[P]_V$	.213 ± .033			c_c	−.001
22			G^p_ρ	.265		μ	$[G^p]_\kappa$	.211 ± .055			G^p_D	−.010
23			θ	.192			G^p_D	.201 ± .065	ln	μ	$[G^c]_{\kappa_w}$	−.028
24			G^p_D	−.106			α	.200 ± .051			G^c_D	−.052
25			c_u	.087			θ	.186 ± .032	ln		p	−.064
26			α	.087			c_u	.148 ± .025	ln	μ	$[G^c]_{C^C}$	−.094
27			c_c	−.000			c_c	.133 ± .023	ln	μ	$[B]_{A_c}$	−.115

[a] mean ± standard deviation of scores for a set of different binnings, as explained in Section II F.

medium are held constant, or, equivalently, if the conditional dependencies between the friction angle or grain size distribution with other features are utilized. Kendall correlation and mutual information, being myopic, do not account for such feature interactions. In the case of RReliefF, we believe that the conditional dependencies of θ, c_u, c_c and α are relatively weak, resulting in low scores. These observations are consistent with geotechnical literature and are revealed even in the absence of disciplinary knowledge.

Having performed a quantitative analysis of the feature set, we further investigate a number of features, inspired by the ranking in Table V. We hypothesized a strong relationship between the permeability and the average closeness centrality of the pore network. Indeed, the pore network closeness centrality consistently ranks high (1st for Kendall correlation, 1st for Mutual Information and 2nd for RReliefF) in our feature selection algorithms. Figure 6 depicts this key result in the form of a scatter plot of the data, along with the two pore networks corresponding to the permeability extrema and the Aste et al. reference packing. Recall that the pore network is weighted using the conductance, as outlined in Appendix A. We observe an approximately linear relationship between the logarithm of the average closeness centrality and the logarithm of the permeability. The variance in closeness centrality decreases slightly, as the permeability increases. The reference case shows good resemblence with the observed trend in the DEM data, exhibiting a high permeability due to the fact that Aste et al. [60] did not subject the sample to compression. As may be expected, the maximum dimensionless permeability (1.8×10^{-3}) is found in a sample subjected to a low confining pressure (1.7×10^5 Pa) resulting in a high overall void ratio (0.68). The minimum permeability (1.3×10^{-5}) corresponds to a low void ratio (0.17) stemming, in turn, from a high confining pressure (9.9×10^6 Pa). The shortest paths distributions on the right of Figure 6 show that a high permeability and high average closeness centrality corresponds to a pore network with relatively fewer, shorter shortest paths. In contrast, a low permeability and low average closeness centrality corresponds to relatively many, longer shortest paths. We can explain the difference using the geometry and connectivity in the corresponding pore networks: the highly permeable sample at the top contains large pores and throats, whereas the pores and throats in the bottom network are much smaller. In summary, the average closeness centrality is able to capture the interplay between fluid flow, shortest paths, pore sizes and throat sizes in a single scalar.

Figure 7 contains a selection of other relationships between features and the permeability. In Figure 7(a), a power-law-like dependency is observed between the permeability and the various void ratio parameters, consistent with past results [25, 96–99]. The global void ratio and averaged pore void ratio data closely resembles $k \propto e^3$ while, interestingly, the averaged throat void ratio

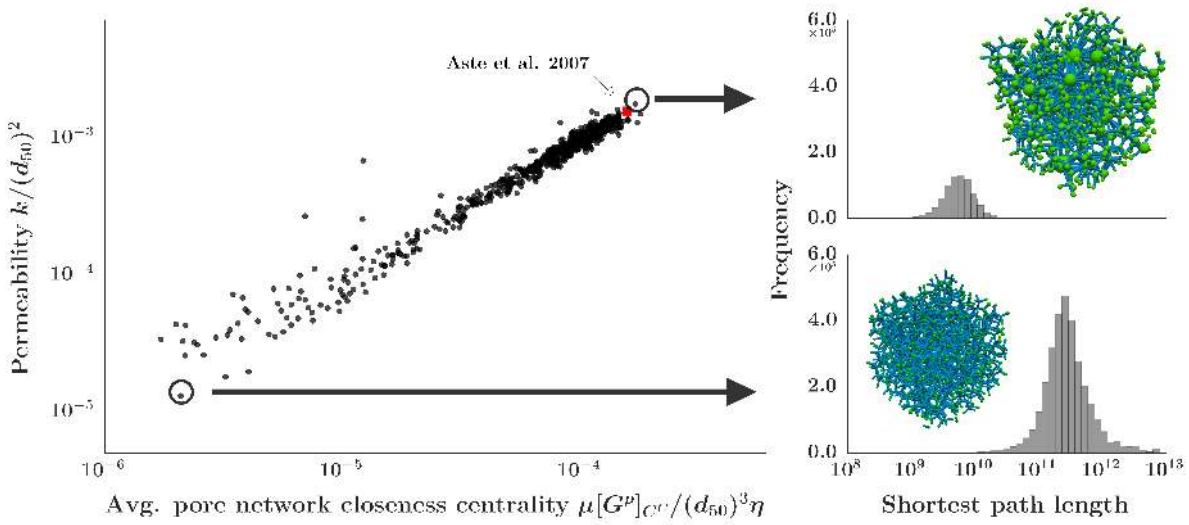

FIG. 6. (Color online) Scatter plot of the average closeness centrality in the pore network and the permeability, along with the pore networks at the extrema and the corresponding distributions of the shortest path lengths. The edge widths in the pore network are scaled by the conductance.

fits better with $k \propto e^4$. Moreover, while the volumetric void ratio parameters (e and $\mu[P]_e$) vary between 0.17 and 0.74, the averaged throat (areal) void ratio is much smaller, varying between 0.10 and 0.30. Showing similar variance compared to the void ratio, the average contact network degree (i.e. coordination number) in Figure 7(b) also has a defined relationship with the permeability. For higher confining pressures, the average degree in the contact network increases and the void ratio decreases. Consequently, the pore connectivity decreases, and, similar to Figure 7(a), the permeability reduction accelerates for larger values of the average contact network degree. Similar, accelerated permeability reduction is observed by Fredrich and co-authors [13, 25]. With an average contact network degree of 1.95, the Aste et al. reference case is showing a similar trend but much lower value compared to our DEM data. We attribute this result to (1) the fact that the reference case is not subjected to triaxial compression, and (2) the imaging resolution, which, as noted in [49], may not always be sufficient to identify particle contacts.

We investigate two lower-scoring features in Figure 7(c). It can be observed that taken together, the specific surface area *ssa* and coefficient of uniformity c_u are able to explain a reasonable fraction of the variance in the permeability. This relationship is expected to be even stronger when considering finer grained porous media than the Ottawa sand-like medium studied in this work. The conditional dependency explains why *ssa* ranks higher in the RReliefF scores, compared to the Kendall and mutual information scores. In Figure 7(c), we also compare least-square fits to the simulation data and the Kozeny-Carman (KC) estimate (data itself not shown). We use the KC equation as presented by Carrier III [100], setting the empirical coefficient at 5. Although the KC estimate slightly underpredicts the permeability compared to our simulation values, we observe a reasonable agreement between the two lines and the Aste et al. reference packing. Lastly, we include the pore network diameter in Figure 7(d) as an example of a feature with little predictive value, showing no clear relationship with the permeability. The lack of predictive value is confirmed by the low feature selection scores in Table V.

Redundancy reduction

None of the methods from Table V take a critical aspect of the feature set into account: redundancy. To quantify this phenomenon, we compute the *inter*-feature correlation values using the (absolute) Kendall correlation score between each pair of features. The average inter-correlation values between features 1-9 and 17-27 (see Table IV) is relatively high (0.64), while the remaining features (10-16) exhibit low inter-feature correlations (0.30). We conclude that the pore network features and contact network features are correlated with the packing features, throat features and pore features, highlighting the overlap between the traditional physical features and the network-based features.

Having observed redundancy in the feature set, we can employ the minimum-Redundancy, Maximum Relevance (mRMR) method. The result of applying the mRMR method to the full feature set is shown in Table VI for the top-10 features. We reiterate that for a particular feature, the mRMR score should be interpreted as the sum of a bonus for the relevance to the permeability and a penalty for redundancy with higher-ranked features. Hence, $[G^p]_{CC}$ (ranked first) has the highest relevance,

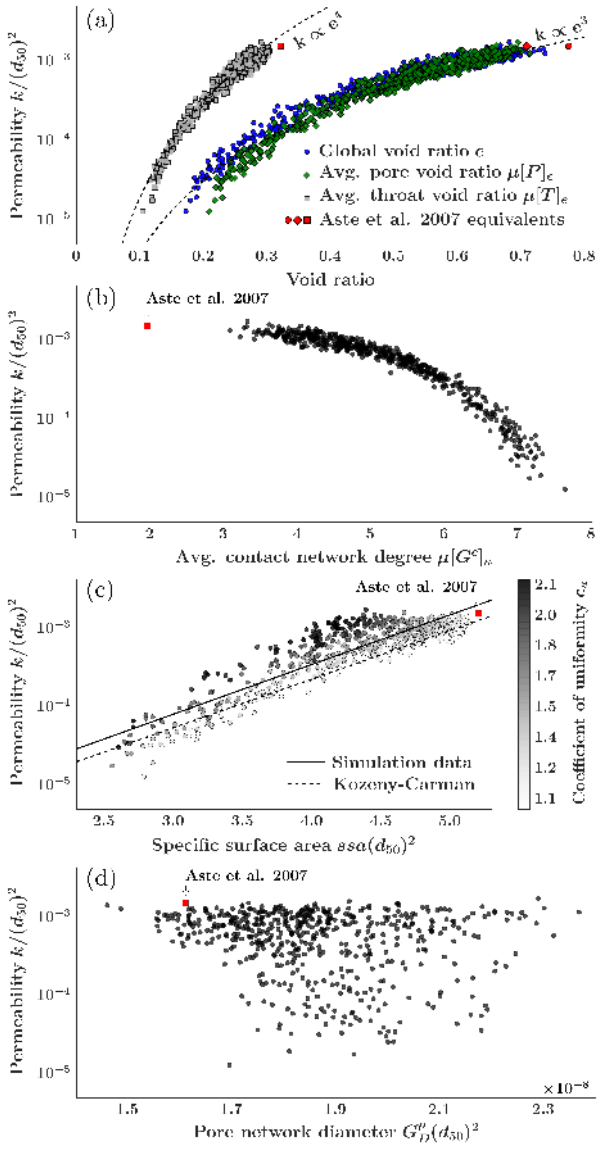

FIG. 7. (Color online) (a) Global and averaged local (pore/throat) void ratio. (b) Averaged contact network degree. (c) Specific surface area, colored by coefficient of uniformity. Also includes a least-squares fit of the simulation data and the Kozeny-Carman estimate. (d) Pore network diameter.

in terms of mutual information, to the permeability. The feature ranked second, $[G^p]_{\kappa_w}$, maximizes Equation (2) by simultaneously having minimal redundancy with $[G^p]_{C^C}$ and maximal relevance to the permeability. Note that the mRMR score drops significantly after the first feature, indicating either low relevance with the permeability or high redundancy with the first feature. Based on the methods from Table V, for which $[G^p]_{\kappa_w}$ achieves high relevancy scores, we conclude that the drop in mRMR scores can be attributed to high redundancy. In conclusion, the pore network closeness centrality is

TABLE VI. Top-10 feature selection scores of the mRMR method.

1	ln	μ	$[G^p]_{C^C}$	.422
2	ln	μ	$[G^p]_{\kappa_w}$	.042
3		μ	$[G^c]_\kappa$	.042
4			e	.029
5		μ	$[P]_e$	.024
6		μ	$[G^c]_{C^B_{edge}}$	.022
7		μ	$[T]_e$	.020
8		μ	$[T]_V/[P]_V$	.018
9	ln	μ	$[T]_K$	.015
10	ln		p	.015

able to capture a large fraction of the available microstructural information in the sample, resulting in any other features being mostly redundant. Equivalently, none of the other features are able to explain much of the remaining variance in the permeability for a particular value of the pore network closeness centrality.

Stability

We analyze the stability of each feature selection algorithm by re-computing the feature selection scores and corresponding ranking using an increasingly large fraction of the 536 samples. Unless the algorithm is unstable or none of the feature correlate with the permeability, we expect the scores to converge as more data becomes available. Figure 8 depicts the result of this analysis. Kendall correlation produces the most stable

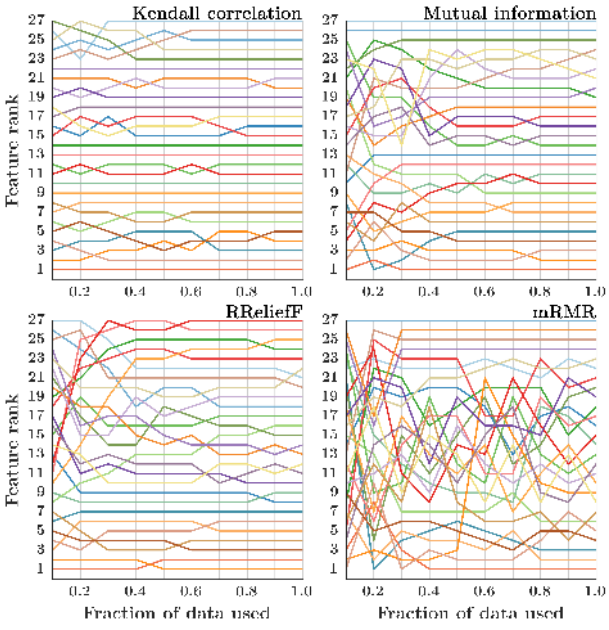

FIG. 8. (Color online) Convergence of the feature selection rankings as more data becomes available. When 100% of the data is used, the ranking corresponds to Tables V and VI.

results, consistently ranking the pore network closeness centrality as the best feature. The features that do change in ranking correspond, in fact, to scores that are relatively close. Mutual information and RReliefF show less stable rankings, compared to Kendall correlation, although the top ten only shows minor changes when at least 50% of the data is used. The mRMR method is clearly the least stable, showing large variations in the ranking for all but the highest-ranked feature. We attribute this result to the optimization procedure in mRMR, which struggles to identify the most relevant and least redundant features after the first feature (pore network closeness centrality) has been chosen. We conclude that the top ten ranking for Kendall correlation, mutual information and RReliefF are reasonably reliable in terms of stability. The mRMR ranking, beyond the highest-ranking feature, is less reliable.

Prediction

Having discussed the characterization of permeability using feature selection, the remaining step in our framework (Figure 1) is the prediction of the permeability. We randomly split the data in 80% (428 packings) and 20% (108 packings) for training and testing purposes, respectively, and run Kendall correlation feature selection on the training set. Assuming that the top features in the resulting ranking are strongly correlated with the permeability, we pick the top two features ($\ln \mu[G^p]_{C^C}$ and $\mu[P]_e$) as the independent variables in the linear regression model. We take the natural logarithm of the average pore void ratio $\mu[P]_e$ because the permeability k and $\mu[P]_e$ approximately follow a power-law relation (see Figure 7(a)) for which a log-log plot is linear. Note that taking the logarithm of a feature does not change its ranking, because Kendall correlation is invariant to monotone transformations. Figure 9 shows the prediction plane and Table VII lists the root-mean-square error (RMSE) and R^2. For the single 80/20 split, the RMSE over the training set is lower than the RMSE over the test set, which we attribute to the inclusion of some of the outliers (shown in Figure 6) in the test data. Indeed, when run with 10-fold cross-validation, which should average out the effect of outliers, the RMSE of the test set and training set are approximately equal at 0.14. The R^2 scores in Table VII suggest that 98% of the variance in the permeability (or, equivalently 86% of the standard deviation) is explained by the pore void ratio and pore network closeness centrality. Not shown here are the cross-validation scores when only a single feature is used, i.e. either $\mu[G^p]_{C^C}$ or $\mu[P]_e$, which are worse than the result above (test set RMSE of 0.19 ± 0.04 and 0.17 ± 0.02, respectively). Hence, despite the fact that the mRMR method (see Table VI) identifies most features as redundant, combining the closeness centrality with a geometry feature does (slightly) improve the prediction. Observe that the regression coefficients of $\mu[G^p]_{C^C}$ (0.48) and $\mu[P]_e$ (1.77) deviate from the observed coefficients of approximately 1 (see Figure 6) and 3 (see Figure 7(a)), respectively. The discrepancy appears due to a trade-off in fitting two independent variables simulateneously.

We experimented with more advanced, non-linear regression method (e.g. support-vector regression, random forest regression) but encountered only a small reduction in RMSE and, more importantly, a larger degree of overfitting. We attribute this result to two factors: (1) the relationship between either of our chosen features and the permeability is linear in a log-log scale, as shown in Figures 6 and 7(a), for which a linear model is most appropriate, and (2) the combination of the pore network closeness centrality and pore void ratio already captures most of the variance in the permeability, so adding another feature (or using a higher-order model) is not going to significantly reduce the prediction error. Essentially, we find that the combination of a connectivity feature (closeness centrality) and a geometry feature (pore void ratio) performs well in characterizing (e.g. explaining the variance in) permeability.

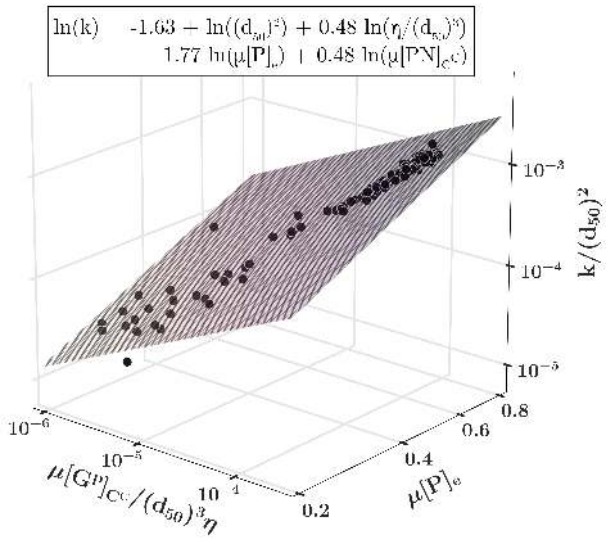

FIG. 9. Linear regression using the logarithm of the average pore void ratio and the logarithm of the average pore network closeness centrality. Only the 108 test set instances are shown. Terms used to make the permeability and closeness centrality dimensionless have been moved to the intercept term of the regression formula to indicate scale-dependence.

TABLE VII. (Color online) RMSE and R^2 values are computed using equations (3) and (4), respectively. For cross-validation, we report the mean and variance over the different folds.

	80/20 split		10-fold Cross-validation	
Data	RMSE	R^2	RMSE ($\mu \pm \sigma$)	R^2 ($\mu \pm \sigma$)
Train	0.14	0.98	0.14 ± 0.002	0.98 ± 0.001
Test	0.17	0.98	0.14 ± 0.023	0.98 ± 0.013

IV. CONCLUSION

We developed a general data-driven framework for modeling transport in porous, granular media from high resolution microstructural data. We quantitatively analyzed a large feature set, spanning micro- to meso-scales, to optimally 'characterize' permeability. By employing multiple feature selection algorithms, we gather objective evidence that certain features are important in predicting permeability and others are not. In particular, the weighted pore network closeness centrality consistently outperforms all other features across all the methods used. The weighted pore network closeness centrality parsimoniously characterizes the internal connectivity and concomitant efficiency of transmission pathways across multiple spatial scales. Specifically, a sample with a high permeability has an internal pore structure encompassing many efficient shortest paths that run through relatively large pore bodies connected to each other by high conductance pore throats. This closeness centrality metric renders most other features redundant in explaining variance in the permeability. Analysis of the corresponding shortest paths and pore network data reveals the interplay between shortest paths, pore- and throat-geometry and fluid flow captured by the closeness centrality. As an example of utilizing the feature selection results, we fit a linear model to the pore void ratio and pore network closeness centrality, which is able to explain approximately 86% of the standard deviation in the permeability.

The framework presented here can be applied to investigate the relationship between permeability (or any other transport property at the macroscopic, engineering scale) and a given feature set, where the latter contains any number of measurable internal properties that span the micro- to meso-scales, for a porous granular material. We demonstrate that feature selection methods are a useful, quantitative approach to extract key parameters from a large dataset. Caution must be exercised however, since the methods are subject to algorithmic subtleties (e.g. myopic versus non-myopic) that influence the results. Therefore, the feature selection score of a variable, a measure of the extent to which it characterizes the permeability, should always be interpreted in light of the assumptions of the particular algorithm used. The use of multiple feature selection methods and the stability test shown in Figure 8, which we adopt here to rule out algorithmic artifacts, are useful checks for the robustness of the results.

In applying the framework to relatively simple sphere packings, we take the first step in applying complex network theory to pore networks. Ongoing work is focused on (1) developing, implementing and validating stable feature selection methods to ensure robustness of features to variations in the training data for accurate characterization and prediction of permeability, (2) *characterizing* other transport properties (e.g. thermal conductivity) within the proposed framework, and (3) *predicting* local pore phenomena (e.g. clogging, filtration) with macroscale implications, using complex networks.

ACKNOWLEDGMENTS

The authors acknowledge the support of the Melbourne Energy Institute, the Australian Research Council (FT140100227, DP120104759), the US Air Force (AFOSR 15IOA059) and the US Army Research Office (W911NF-11-1-0175). We thank Dr. Benjamin Rubinstein and Prof. Stephan Matthai for the helpful discussions, and Dr. Tomaso Aste for sharing the reference sample data.

Appendix A: Conductance computation

In this appendix, we briefly detail our approach to computing the local conductance, used as the edge weighting in the pore network. Define (p_1, p_2) as the pair of pores, connected by pore throats t_i, $i = 1, \ldots, N_t$. Assume the pores can be represented as cylinders with lengths (L_{p_1}, L_{p_2}), radii (r_{p_1}, r_{p_2}) and volumes (V_{p_1}, V_{p_2}) equal to the original pore volumes. Furthermore, assume the throats can be represented by cylinders with lengths L_{t_i}, radii r_{t_i} and top/bottom areas A_{t_i} equal to the original pore throat areas. We then compute the conductance weighting C [m^3 Pa^{-1} s^{-1}] as the harmonic mean

$$C = \frac{L_{p_1} + L_{t_{\text{eqv}}} + L_{p_2}}{\frac{L_{p_1}}{C_{p_1}} + \frac{L_{t_{\text{eqv}}}}{C_{t_{\text{eqv}}}} + \frac{L_{p_2}}{C_{p_2}}} \tag{A1}$$

where $L_{t_{\text{eqv}}}$ and $C_{t_{\text{eqv}}}$ are the arithmetic means

$$L_{t_{\text{eqv}}} = \frac{\sum_{i=1}^{N_t} L_{t_i} A_{t_i}}{\sum_{i=1}^{N_t} A_{t_i}}, \qquad C_{t_{\text{eqv}}} = \frac{\sum_{i=1}^{N_t} C_{t_i} A_{t_i}}{\sum_{i=1}^{N_t} A_{t_i}}$$

Equation (A1) is also illustrated in Figure 10. The

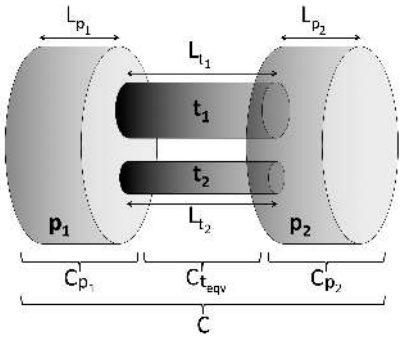

FIG. 10. Illustration of Equation (A1) for $N_t = 2$ throats connecting a pair of pore bodies (p_1, p_2)

conductances C_{p_1}, C_{p_2} and C_{t_i} are computed from the

Hagen-Poiseuille equation

$$Q = C_{p_k}\Delta p = \frac{\pi r_{p_k}^4}{8\eta L_{p_k}}\Delta p, \qquad k = 1, 2$$

$$Q = C_{t_i}\Delta p = \frac{\pi r_{t_i}^4}{8\eta L_{t_i}}\Delta p, \qquad i = 1, \ldots, N_t$$

where Q is the fluid discharge [$\mathrm{m^3 s^{-1}}$], η is the dynamic viscosity [Pa s] and Δp is the pressure drop [Pa]. Note that in our case, 'pores' are (merged collections of) tetrahedra and 'throats' are shared faces, but the theory applies to any pore network construction algorithm (e.g. inscribed spheres) as long as the pores and throat representations can be translated to equivalent cylinders.

[1] R. W. Rice, *Porosity of Ceramics: Properties and Applications* (CRC Press, Coca Raton, FL, USA, 1998).

[2] N. Neithalath, M. S. Sumanasooriya, and O. Deo, Materials Characterization **61**, 802 (2010).

[3] D. J. Hartmann and E. A. Beaumont, in *Exploring for Oil and Gas Traps: AAPG Treatise of Petroleum Geology, Handbook of Petroleum Geology*, edited by E. A. Beaumont and N. H. Foster (American Association of Petroleum Geologists, Tulsa, OK, USA, 1999) Chap. 9, pp. 9.1–9.154.

[4] M. Hubbert and D. Willis, Petroleum Transactions, AIME **210**, 239 (1972).

[5] F. Jiang and T. Tsuji, Physical Review E **90**, 053306 (2014).

[6] H. Brandl, Géotechnique **56**, 81 (2006).

[7] L. Katz, D. Humphrey, P. Jankauskas, and F. Demascio, Hazardous Waste and Hazardous Materials **13**, 283 (1996).

[8] MIT, *The Future of Geothermal Energy*, Tech. Rep. (Massachusetts Institute of Technology (MIT), Idaho Falls, ID, 2006).

[9] OPEC, *World Oil Outlook*, Tech. Rep. (Organization of the Petroleum Exporting Countries (OPEC), Vienna, Austria, 2015).

[10] J. C. Santamarina, K. A. Klein, and M. A. Fam, (2001).

[11] Y. Zaretskiy, S. Geiger, K. Sorbie, and M. Förster, Advances in Water Resources **33**, 1508 (2010).

[12] S. Blair, P. Berge, and J. Berryman, Journal of Geophysical Research **101**, 20359 (1996).

[13] J. T. Fredrich and W. B. Lindquist, International journal of rock mechanics and mining sciences & geomechanics abstracts **34**, 368 (1997).

[14] V. Cnudde, B. Masschaele, M. Dierick, J. Vlassenbroeck, L. V. Hoorebeke, and P. Jacobs, Applied Geochemistry **21**, 826 (2006).

[15] I. Vlahinić, E. Andò, G. Viggiani, and J. E. Andrade, Granular Matter **16**, 9 (2013).

[16] K. Alshibli and A. Reed, eds., *Advances in Computed Tomography for Geomaterials* (Wiley, New Orleans, LA, USA, 2010).

[17] S. A. Hall, J. Desrues, G. Viggiani, P. Bésuelle, and E. Andò, Procedia IUTAM **4**, 54 (2012).

[18] Y. S. Yang, A. M. Tulloh, T. Muster, A. Trinchi, S. C. Mayo, and S. W. Wilkins, in *Proc. SPIE 7804, Developments in X-Ray Tomography VII*, edited by S. R. Stock (San Diego, CA, USA, 2010).

[19] X. Garcia, J. Xiang, J.-P. Latham, and J. Harrison, Géotechnique **59**, 779 (2009).

[20] H. Zhu, Z. Zhou, R. Yang, and a.B. Yu, Chemical Engineering Science **63**, 5728 (2008).

[21] G.-F. Zhao, *High Performance Computing and the Discrete Element Model* (ISTE Press Ltd - Elsevier Inc, London, UK, 2015).

[22] R. Hilfer, Physical Review B **44**, 60 (1991).

[23] J. Jang, G. Narsilio, and J. Santamarina, Geophysical journal International **184**, 1167 (2011).

[24] S. Russell, D. Walker, and A. Tordesillas, Journal of the Mechanics and Physics of Solids **88**, 227 (2016).

[25] J. T. Fredrich, A. A. DiGiovanni, and D. R. Noble, Journal of Geophysical Research **111**, 1 (2006).

[26] D. Bi, J. Zhang, B. Chakraborty, and R. P. Behringer, Nature **480**, 355 (2011).

[27] D. M. Walker, A. Tordesillas, J. Ren, J. A. Dijksman, and R. P. Behringer, Europhysics Letters **107**, 18005 (2014).

[28] A. Tordesillas, S. Pucilowski, S. Tobin, M. R. Kuhn, E. Andò, G. Viggiani, A. Druckrey, and K. Alshibli, Europhysics Letters **110**, 58005 (2015).

[29] A. Tordesillas and M. Muthuswamy, Acta Geotechnica **3**, 225 (2008).

[30] A. Tordesillas and M. Muthuswamy, Journal of the Mechanics and Physics of Solids **57**, 706 (2009).

[31] A. Tordesillas, D. M. Walker, and Q. Lin, Physical Review E **81**, 011302 (2010).

[32] J. Bear, *Dynamics of fluids in porous media* (Dover, New York, NY, 1988).

[33] D. Coelho, J.-F. Thovert, and P. M. Adler, Physical Review E **55**, 1959 (1997).

[34] J. Santamarina and G. Cho, in *Advances in Geotechnical Engineering. Proceedings of the Skempton Conference* (Thomas Telford, London, UK, 2004) pp. 604–617.

[35] R. Hilfer, Advances in Chemical Physics **XCII**, 299 (1996).

[36] S. Bakke and P. Øren, SPE Journal **2**, 136 (1997).

[37] B. Biswal, C. Manwart, R. Hilfer, S. Bakke, and P. Oren, Physica A **273**, 452 (1999).

[38] C. Manwart, S. Torquato, and R. Hilfer, Physical review E **62**, 893 (2000).

[39] R. Hilfer and C. Manwart, Physical review E **64**, 021304 (2001).

[40] C. Manwart, U. Aaltosalmi, A. Koponen, R. Hilfer, and J. Timonen, Physical Review E **66**, 016702 (2002).

[41] H. Okabe and M. J. Blunt, Physical Review E **70**, 66135 (2004).

[42] T. Yun, T. Han, S. Chung, and G. Narsilio, KSCE Journal of Civil Engineering **18**, 132 (2014).

[43] C. Scholz, F. Wirner, J. Götz, U. Rüde, G. E. Schröder-Turk, K. Mecke, and C. Bechinger, Physical Review Letters **109**, 264504 (2012).

[44] J. G. Berryman and S. C. Blair, Journal of Applied Physics **60**, 1930 (1986).

[45] P. Adler, C. Jacquin, and J. Quiblier, International Journal of Multiphase Flow **16**, 691 (1990).

[46] D. Coker and S. Torquato, Journal of Applied Physics **77**, 6087 (1995).

[47] G. H. Ristow, *Pattern formation in granular materials* (Springer, Berlin, Germany, 2000).

[48] N. F. Johnson, *Simply Complexity: A Clear Guide to Complexity Theory* (Oneworld, London, UK, 2009).

[49] G. W. Delaney, T. Di Matteo, and T. Aste, Soft Matter **6**, 2992 (2010).

[50] J. Quiblier, Journal of Colloid and Interface Science **98**, 84 (1984).

[51] C. Yeong and S. Torquato, Physical Review E **57**, 495 (1998).

[52] M. Vold, The Journal of Physical Chemistry **64**, 1616 (1960).

[53] W. M. Visscher and M. Bolsterli, Nature **239**, 504 (1972).

[54] R. Jullien and P. Meakin, Europhysics Letters **4**, 1385 (1987).

[55] P. Cundall and O. Strack, Géotechnique **29**, 47 (1979).

[56] T. Yun, T. Han, S. Chung, and G. Narsilio, KSCE Journal of Civil Engineering **18**, 132 (2013).

[57] D. Walker, K. Vo, and A. Tordesillas, International Journal of Bifurcation and Chaos **23**, 1330034 (2013).

[58] T. Aste, M. Saadatfar, A. Sakellariou, and T. J. Senden, Physica A: Statistical Mechanics and its Applications **339**, 16 (2004).

[59] T. Aste, M. Saadatfar, and T. Senden, Physical Review E **71**, 061302 (2005).

[60] T. Aste, T. D. Matteo, M. Saadatfar, T. J. Senden, M. Schröter, and H. L. Swinney, Europhysics Letters **79**, 24003 (2007).

[61] M. E. J. Newman, *Networks: an introduction* (Oxford University Press, New York, NY, 2010).

[62] I. Fatt, Petroleum Transactions, AIME **207**, 144 (1956).

[63] W. Lindquist, Contemporary Mathematics **295**, 355 (2002).

[64] R. Al-Raoush, K. Thompson, and C. S. Willson, Soil Science Society of America Journal **67**, 1687 (2003).

[65] W. Fourie, R. Said, P. Young, and D. Barnes, in *Comsol conference* (Newton, MA, 2007) pp. 2–7.

[66] G. Narsilio, O. Buzzi, S. Fityus, T. Yun, and D. Smith, Computers and Geotechnics **36**, 1200 (2009).

[67] G. Narsilio, J. Kress, and T. Yun, Computers and Geotechnics **37**, 828 (2010).

[68] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag New York, Inc., Secaucus, NJ, 2006).

[69] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. (Springer-Verlag, New York, NY, 2009).

[70] J. Ma, Z. Jiang, K. Wu, Q. Tian, and G. D. Couples, in *13th European Conference on the Mathematics of Oil Recovery* (European Association of Geoscientists & Engineers, Biarritz, France, 2012) pp. 1637–1643.

[71] H. Xu, R. Liu, A. Choudhary, and W. Chen, Journal of Mechanical Design **137**, 051403 (2015).

[72] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Physical Review Letters **114**, 105503 (2015).

[73] M. Khandelwal, Neural Computing and Applications **21**, 1341 (2012).

[74] L. J. Briggs and H. L. Shantz, Botanical Gazette **53**, 20 (1912).

[75] J. Bouma and H. van Lanen, in *Quantified Land Evaluation*, edited by K. Beek, P. Burrough, and D. McCormack (Enschede, The Netherlands, 1987) pp. 106–110.

[76] M. Shahin, in *Metaheuristics in Water, Geotechnical and Transport Engineering*, edited by X.-S. Yang, A. H. Gandomi, S. Talatahari, and A. H. Alavi (Elsevier, 2013) Chap. 8, pp. 169–204.

[77] V. Šmilauer, E. Catalano, B. Chareyre, S. Dorofeenko, J. Duriez, A. Gladky, J. Kozicki, C. Modenese, L. Scholtès, L. Sibille, J. Stránský, and K. Thoeni, in *Yade Documentation*, edited by V. Šmilauer (The Yade Project, 2010) 1st ed.

[78] X. Garcia, L. Akanji, M. Blunt, S. Matthai, and J. Latham, Physical Review E **80**, 021304 (2009).

[79] D. C. Procter and R. R. Barton, Géotechnique **24**, 581 (1974).

[80] F. Tatsuoka, K. Ochi, S. Fujii, and M. Okamoto, Soils and Foundations **26**, 23 (1986).

[81] Simpleware Ltd., "Simpleware ScanIP," `https://www.simpleware.com/software/scanip/` (2015).

[82] Comsol AB, "Comsol multiphysics v. 5.0," `http://www.comsol.com` (2015).

[83] J.-W. Kim, D. Kim, and W. B. Lindquist, Water Resources Research **49**, 7615 (2013).

[84] J. van der Linden, A. Sufian, G. Narsilio, A. Russell, and A. Tordesillas, "Pore network construction for sphere packings," (2016), unpublished manuscript.

[85] H. Brönnimann, A. Fabri, G.-J. Giezeman, S. Hert, M. Hoffmann, L. Kettner, S. Pion, and S. Schirra, in *CGAL User and Reference Manual* (CGAL Editorial Board, 2016) 4.8 ed.

[86] L. C. Freeman, Sociometry **40**, 35 (1977).

[87] L. C. Freeman, Social Networks **1**, 215 (1978).

[88] H. C. Peng, F. H. Long, and C. Ding, IEEE Transactions on Pattern Analysis and Machine Intelligence **27**, 1226 (2005).

[89] M. G. Kendall, Biometrika **30**, 81 (1938).

[90] K. Kira and L. A. Rendell, in *Proceedings of the ninth international workshop on Machine learning*, edited by D. Sleeman and P. Edwards (Morgan Kaufmann, Aberdeen, 1992) pp. 249–256.

[91] I. Kononenko, in *European Conference on Machine Learning*, Vol. 784, edited by F. Bergadano and L. de Raedt (Srpinger-Verlag, Catania, 1994) pp. 171–182.

[92] M. Robnik-Šikonja and I. Kononenko, in *Proceedings of the Fourteenth International Conference on Machine Learning*, Vol. 5, edited by D. H. Fisher (Morgan Kaufmann, San Francisco, CA, 1997) pp. 296–304.

[93] M. Robnik-Šikonja and I. Kononenko, Machine Learning **53**, 23 (2003).

[94] I. Guyon and A. Elisseeff, The Journal of Machine Learning Research **3**, 1157 (2003).

[95] G. C. Cawley and N. L. C. Talbot, Journal of Machine Learning Research **11**, 20792107 (2010).

[96] T. Bourbie and B. Zinszner, Journal of Geophysical Research **90**, 11524 (1985).

[97] J. Dvorkin, N. Derzhi, E. Diaz, and Q. Fang, Geophysics **76**, E141 (2011).
[98] W. Zhu, C. David, and T.-f. Wong, Journal of Geophysical Research: Solid Earth **100**, 451 (1995).
[99] U. Mok, Y. Bernabé, and B. Evans, Journal of Geophysical Research: Solid Earth **107**, ECV 4 (2002).
[100] W. Carrier III, Journal of Geotechnical and Geoenvironmental Engineering **129**, 1054 (2003).