Engineering with Computers (2023) 39:3895–3926
https://doi.org/10.1007/s00366-023-01841-8

ORIGINAL ARTICLE

A data-driven framework for permeability prediction of natural porous rocks via microstructural characterization and pore-scale simulation

Jinlong Fu[1] · Min Wang[2] · Bin Chen[3] · Jinsheng Wang[4] · Dunhui Xiao[5] · Min Luo[6] · Ben Evans[1]

Received: 11 October 2022 / Accepted: 3 May 2023 / Published online: 19 May 2023

Abstract

Understanding the microstructure–property relationships of porous media is of great practical significance, based on which macroscopic physical properties can be directly derived from measurable microstructural informatics. However, establishing reliable microstructure–property mappings in an explicit manner is difficult, due to the intricacy, stochasticity, and heterogeneity of porous microstructures. In this paper, a data-driven computational framework is presented to investigate the inherent microstructure–permeability linkage for natural porous rocks, where multiple techniques are integrated together, including microscopy imaging, stochastic reconstruction, microstructural characterization, pore-scale simulation, feature selection, and data-driven modeling. A large number of 3D digital rocks with a wide porosity range are acquired from microscopy imaging and stochastic reconstruction techniques. A broad variety of morphological descriptors are used to quantitatively characterize pore microstructures from different perspectives, and they compose the raw feature pool for feature selection. High-fidelity lattice Boltzmann simulations are conducted to resolve fluid flow passing through porous media, from which reliable permeability references are obtained. The optimal feature set that best represents permeability is identified through a performance-oriented feature selection process, upon which a cost-effective surrogate model is rapidly fitted to approximate the microstructure-permeability mapping via data-driven modeling. This surrogate model exhibits great advantages over empirical/analytical formulas in terms of prediction accuracy and generalization capacity, which can predict reliable permeability values spanning four orders of magnitude. Besides, feature selection also greatly enhances the interpretability of the data-driven prediction model, from which new insights into the mechanism of how microstructural characteristics determine intrinsic permeability are obtained.

Keywords Porous rocks · Permeability prediction · Microstructural characterization · Lattice Boltzmann simulation · Feature selection · Data-driven modeling

✉ Min Wang
sacewangmin@gmail.com

1 Zienkiewicz Institute for Modelling, Data and AI, Faculty of Science and Engineering, Swansea University, Swansea SA1 8EN, UK

2 Fluid Dynamics and Solid Mechanics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

3 College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China

4 School of Civil Engineering, Southwest Jiaotong University, Chengdu 610031, China

5 School of Mathematical Sciences, Tongji University, Shanghai 200092, China

6 Ocean College, Zhejiang University, Zhoushan 316021, Zhejiang, China

1 Introduction

Permeability quantifies the ability of a porous medium to transmit fluid and serves as a fundamental characteristic for the transport behavior of fluid flow inside porous media [1, 8]. It plays a critical role in such geological applications as oil and gas recovery, geothermal energy exploitation, CO_2 underground storage, radioactive waste disposal, and contaminant hydrogeology. The permeable pore spaces in different geologic materials are often highly distinctive, leading to an extremely broad range of permeability values that vary up to 13 orders of magnitude [73]. The macroscopic physical properties of porous media [10, 35, 41, 47, 74, 75] strongly depend on the microstructural characteristics, such that the hydraulic, mechanical, electrical, and thermal properties of porous media can be evaluated/estimated from

the measurable microstructural informatics, at least in principle. Indeed, the microstructure–property relationship is one of the most fundamental challenges in porous media research. However, the intricacy, stochasticity, and heterogeneity inherent in natural porous rocks make it difficult to accurately and rapidly evaluate permeability, especially for tight rocks with low porosity. Therefore, a deep insight into the microstructure–permeability mapping is desirable and has attracted extensive research efforts, to develop a reliable and efficient method for permeability prediction [13, 34, 80, 83, 100].

Laboratory measurement is the routine way to determine permeability, where fluid flow is driven by a constant pressure difference to pass through a rock core and permeability is then evaluated according to Darcy's law when the fluid flow reaches the steady state [8]. In practical applications, long waiting time and high cost are the main limitations of experimental measurement, especially for tight rocks. In addition to the experimental measurement, analytical and empirical models have also been developed to predict permeability of porous media, such as the well-known Kozeny–Carman relation [21, 35] and many variants derived from it [10, 11, 28]. Generally, these models rely on specific microstructural characteristics of porous media, such as porosity, specific surface area, tortuosity, characteristic length, pore size, constriction factor, and fractal dimension, among others. Despite the simplicity and convenience in practical applications, analytical models are often overly idealized and empirical models usually contain adjustable parameters to accommodate uncertainty. As a result, they are restricted to some specific pore microstructures and for natural rocks with complicated pore networks, their prediction accuracy drops significantly (and can become unacceptable due to excessive errors).

In recent years, the digital rock physics (DRP) technique has progressed rapidly, offering an alternative to laboratory measurement and analytical/empirical models for permeability evaluation. The DRP approach uses advanced microscopy imaging techniques [4, 17], such as X-ray micro-computed tomography (micro-CT) and focused ion beam scanning electron microscopy (FIB-SEM), to obtain 3D geometries of pore microstructures, on which high-fidelity numerical simulations are performed to evaluate various transport properties and investigate specific physical phenomena [3, 16]. The DRP approach is convenient and promising for microstructural characterization and petrophysical property evaluation [3, 34].

Pore-network modeling (PNM) and direct numerical simulation (DNS) are the two primary pore-scale computing approaches to mimicking transport processes occurring inside porous rocks. According to some specific criteria [16, 113], PNM simplifies the complicated pore space into a topologically representative network of pore bodies interconnected by pore throats with ideal shapes (such as sphere and cylinder). The transport behaviors within each network element are described by semi-analytical laws (such as Hagen–Poiseuille law), which greatly reduces the computational cost and enables multi-scale modeling to incorporate strong heterogeneity in large volumes. PNM is widely used for capillary-controlled transport processes. However, due to the simplification of complicated pore space, PNM may produce inaccurate estimations. To date, it remains a major challenge to correctly identify the key microstructural features that are critical for effective PNM estimation and the less important ones that can be safely ignored to reduce computational complexity [113].

By contrast, DNS directly discretizes the raw pore space into computing elements by preserving pore geometry (voxels can be used as the computing elements), and then transport equations (such as Navier–Stokes or Laplace equations) are numerically solved or approximated on the computational meshes [3, 17, 37, 107]. The lattice Boltzmann method (LBM), finite-element method (FEM), and finite volume method (FVM) are commonly used to approximate or solve transport equations at the pore scale. Generally, DNS can provide direct insight into the impact of pore microstructure on transport properties, but it has severe limitations in computational intensity. The 3D digital microstructure with large representative size and high resolution usually contains hundreds of millions (or even billions) of computational elements (or voxels). As a result, massively parallel programming, long computing time, high-performance computing (HPC) platform, and large data storage are usually required to run such large-scale numerical simulations [71, 91]. The computation-intensive nature of DNS makes it difficult to accommodate all details of pore microstructures and involve all relevant transport physics.

As discussed above, both PNM and DNS approaches have their own limitations, which have been long recognized by the DRP community. More recently, many attempts have been made to develop surrogate microstructure–property models through artificial intelligence, to rapidly and accurately predict macroscopic properties from the measurable microstructural informatics. Due to the powerful capacities in massive data analysis and hidden rule exploration, machine/deep learning algorithms are becoming increasingly popular in this field, especially the convolutional neural network (CNN). CNN [2] is capable of automatically extracting task-related features from spatial data such as images through its convolution layers, avoiding the manual feature selection procedure, and it has achieved tremendous success in the computer vision field. Therefore, many similar studies have been conducted to construct CNN-based surrogate models for permeability prediction, where the 2D or 3D digital images of porous microstructures are directly used as the input data [48, 55, 96, 97, 99, 100, 110]. Besides,

CNN has also been applied to establish the linkages between microstructures and other macroscopic properties/behaviors for various heterogeneous materials, including effective thermal conductivity [108], effective elastic moduli [22, 68], effective diffusivity [109], P/S-wave velocity [56], formation factor [85], and fluid velocity filed [90].

However, despite the rapid growth in publications, the CNN-based surrogate modeling strategy is not without limitations, at least in its current forms. (1) Shortage of training images: A reliable predictive CNN model usually requires a large number of training images to feed it, but acquirement of high-quality 3D digital rocks is quite expensive, which can explain why most of the previous studies only use 2D images or sphere packing to investigate the microstructure–property relations of porous media. (2) Heavy computational burden: High computational intensity and excessive memory requirement are the inherent challenges of the 3D CNN algorithm, which strictly limit both the quality and quantity of 3D training images. Using 3D representative elementary volumes (REVs) of digital rocks to fit a CNN model usually demands an HPC platform. (3) Feature extraction problem: Kernels (convolving windows) are applied across the input image to extract local features, but the internal connection of components, as well as the relative spatial relationships, are not captured by the convolution layers of CNN. It means that the global features of porous media (such as long-distance connectivity and topological information) that are crucial to transport properties are rarely considered. (4) Rotation dependence: The internal representation of a pore microstructure in CNN is not independent of the view angle, which means that rotation of the input image can potentially affect the prediction result. This issue can be solved through data augmentation, but the computational cost of CNN model training will be dramatically increased. (5) Over-fitting problem: The CNN model is prone to overfitting due to a large database for training. (6) Low-level interpretability: The complicated CNN architecture, formed by a deep stack of distinct layers, is often referred to as a "black box", because it is difficult to understand the underlying mechanics and no inherent way exists to interpret how features influence a particular prediction. (7) Inflexibility: Once a CNN model is fitted for physical property prediction, both the size and resolution of input images are fixed, which is very inflexible for the common cases where adjustments of image size or resolution have to be made without losing information.

As discussed above, the poor explainability of CNN does not contribute to a good understanding of the microstructure–property linkages. In contrast, results from simple regression algorithms, such as linear regression, decision tree, random forest, support vector machine, and shallow neural network, are much easier to be interpreted, which helps to reveal the underlying mechanisms of the microstructure–property relationships. In addition, CNN is free from manual feature-extraction, but in porous media research, this feature does not constitute a comparative advantage over other regression algorithms that require predefined feature variables. This is because various morphological descriptors that quantitatively characterize pore microstructures have already been properly designed (as listed in Table 1). These descriptors can provide measurable microstructural informatics from multiple perspectives to predict macroscopic physical properties. Compared with the unreadable features extracted by CNN, morphological descriptors characterize porous microstructures from multiple perspectives with clear physical indications, and they form the feature pool that can be readily used to develop the surrogate microstructure–property relationships through simpler regression algorithms. Specifically, feature selection can be conducted to identify the morphological descriptors that are significant to permeability and remove the abundant and irrelevant ones, through which the microstructural complexity is reduced to a limited number of descriptive parameters related to permeability, and then, a high-fidelity data-driven prediction model can be achieved (as illustrated in Fig. 1). More importantly, the dependence of permeability on microstructural characteristics can also be well interpreted through the feature selection process, providing a deep insight into the microstructure–permeability relation.

Although simple regression algorithms have been adopted to model physical properties of porous media in the previous studies [32, 87, 101, 104, 106], they provide little insightful understanding of the linkages between macroscopic physical properties and microstructural characteristics, and the corresponding pore-scale behaviors are still poorly understood. This study distinguishes itself from previous studies in the following five aspects (as graphically illustrated in Fig. 1). (1) Plenty of 3D digital rocks with diverse morphologies are acquired from micro-CT scanners at high resolution, which are used to construct the predictive model with strong generalization capacity. (2) A large number of 3D microstructure samples are stochastically reconstructed by preserving statistical equivalence, morphological similarity, and transport properties, which are used as the raw data to capture the stochasticity in permeability modeling. (3) A wide variety of morphological descriptors are collected through an extensive literature study, aiming to provide comprehensive characterization of porous microstructures. (4) High-fidelity simulations of pore-scale fluid flow passing through the REVs of digital microstructures are performed to obtain reliable permeability values. (5) Feature selection is conducted to identify the optimal feature set that best represents permeability, based on which a data-driven model with excellent prediction performance can be constructed, and the model interpretability can be also enhanced.

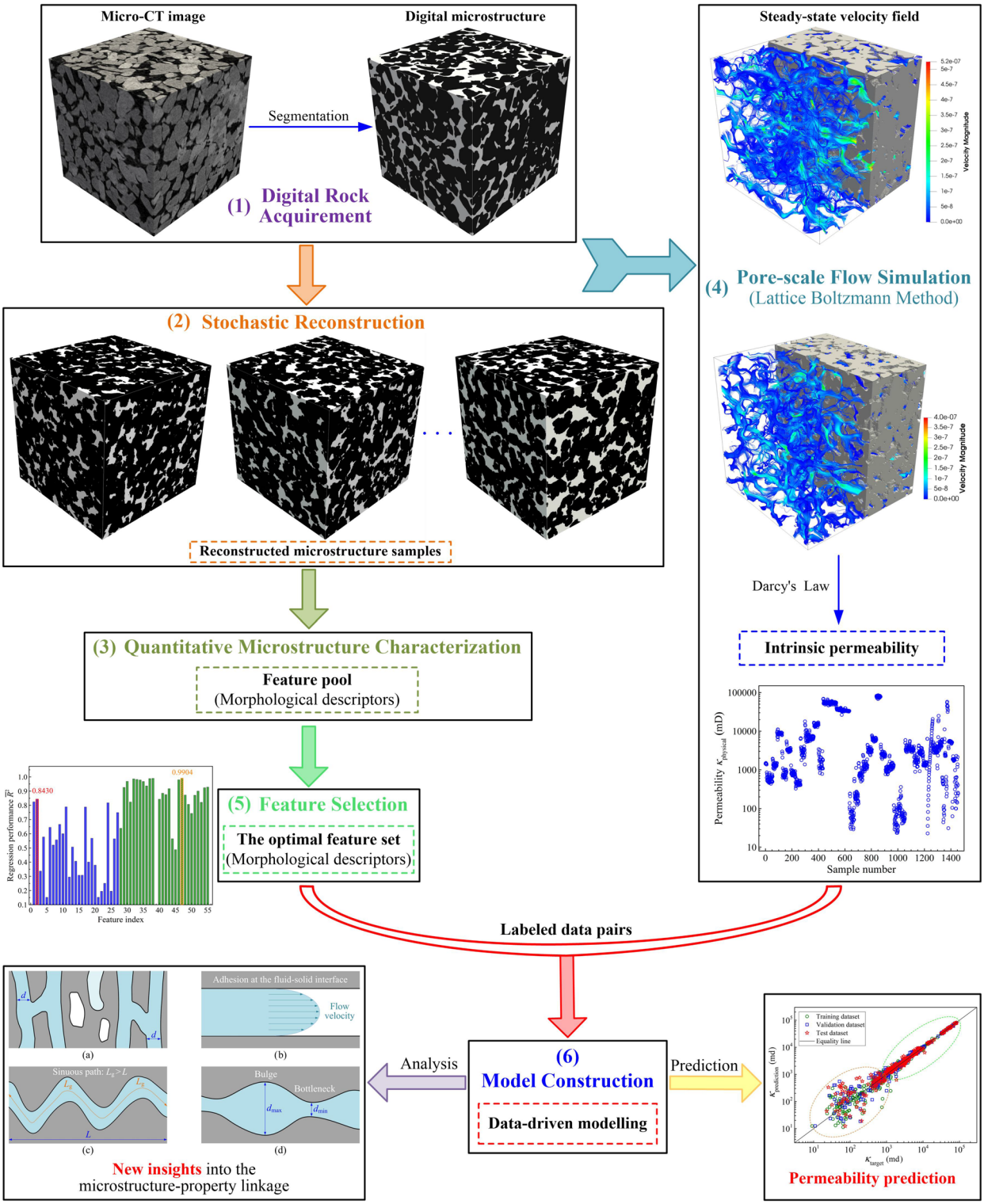


Fig. 1 The data-driven framework to investigate the microstructure–permeability relation for natural porous rocks, and it contains six functional modules: (1) Digital rock acquirement, (2) Stochastic microstructure reconstruction, (3) Quantitative microstructure characterization, (4) Pore-scale flow simulation, (5) Feature selection, and (6) Data-driven modeling (the paired observations are comprised of the morphological descriptors selected from the 5th module and the permeability values evaluated from the 4th module, based on which data-driven machine learning models can be trained to construct the nonlinear microstructure–permeability mapping)

In summary, this work proposes a data-driven framework to investigate the dependence of macroscopic physical properties on microstructural characteristics for porous media. Here, we focus on the intrinsic permeability of porous rocks, but this framework is generally applicable to study other physical properties, such as hydraulic, thermal, electrical, diffusional, and mechanical behaviors. The remainder of this paper is organized as follows. The methodology of the data-driven framework is explained in detail in Sect. 2, and the raw datasets are also prepared and organized, including microstructure sample generation, morphological descriptor extraction, and permeability evaluation. In Sect. 3, different types of feature selection are tested to identify the morphological descriptors that are significant to permeability. In Sect. 4, the optimal feature sets identified by the wrapper methods used to construct data-driven prediction models, and regression performances are also deeply analyzed. The data-driven models are compared with two popular empirical/analytical formulas in terms of prediction accuracy and generalization performance in Sect. 5. Finally, the key findings and relevant thinking are discussed, and the main contributions of this work are also summarized in Sect. 6.

2 Methodology and data preparation

As illustrated in Fig. 1, the proposed data-driven framework comprises six functional modules: digital rock acquirement, stochastic microstructure reconstruction, quantitative microstructure characterization, pore-scale flow simulation, feature selection, and data-driven modeling. These modules are integrated together to form a data-driven approach to investigating the microstructure–permeability relation of natural porous rocks. The first four modules are briefly explained in this section, while the last two modules will be detailed in Sect. 3 and Sect. 4, respectively.

2.1 Digital rock acquirement

To investigate how microstructural characteristics determine intrinsic permeability, a variety of porous media (mainly porous rocks) are used in this study, including sandstones, carbonate rocks, sand packs, and synthetic silicas among others. The pore network systems inside these porous media are rather different in terms of their geometry, topology, fractal property, and statistical attribute, which assures that the resulting prediction model holds for a diverse range of porous media. For sedimentary rocks in hydrocarbon reservoirs, the porosity values generally vary from 10% to 40% in sandstones and from 5% to 25% in carbonates. As illustrated in Fig. 2, the porous media samples used in this study have a wide porosity range varying from 6.85% to 50.73%. Besides, the permeability values also broadly vary across 4 orders of magnitude, as shown in Fig. 8.

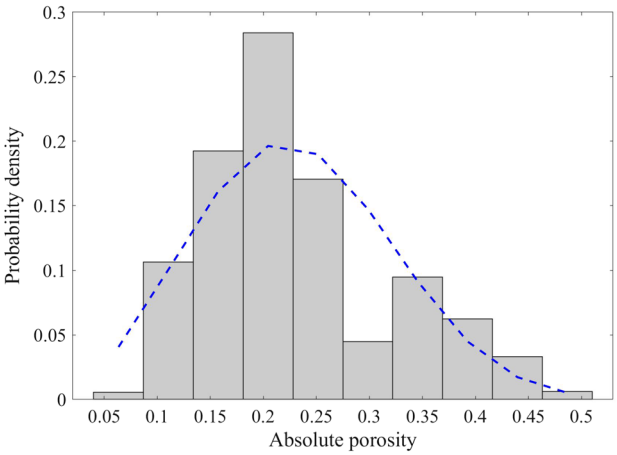


Fig. 2 The porosity distribution of the porous media samples used in this study

Modern microscopy imaging techniques can be used to characterize the internal geometries of opaque porous rocks at the micro-scale. Here, 3D digital rock samples are acquired from micro-CT scanning, and they can be used for subsequent studies including microstructural analyses and pore-scale numerical simulations. The raw micro-CT image is usually in grayscale, as shown in Fig. 3a. It is necessary to convert the raw grayscale image to a segmented format that permits quantitative characterization of the porous microstructure and pore-scale simulation of fluid flow. As illustrated in Fig. 3, the raw micro-CT image of a Mt. Simon sandstone sample [59] is denoised and segmented. The binary segmentation is often referred to as the digital microstructure, where the pore space is separated from the solid matrix. The digital microstructure provides a computational mesh for quantitative characterization of the pore network system and numerical simulation of pore-scale flow. More details on image processing and segmentation can be found in relevant references [50, 93].

It is noted that the micro-CT images used here are collected from several open-access databases, and these images are processed and segmented using *ImageJ* [51], a popular image processing tool in the DRP community. Due to the high costs of rock core drilling and microscopy imaging, there are only a limited number of 3D digital rock samples available. A total of 185 micro-CT images are used in this study, covering 37 types of porous media with distinct morphological features (the representatives of them are shown in Figs. 7 and 9).

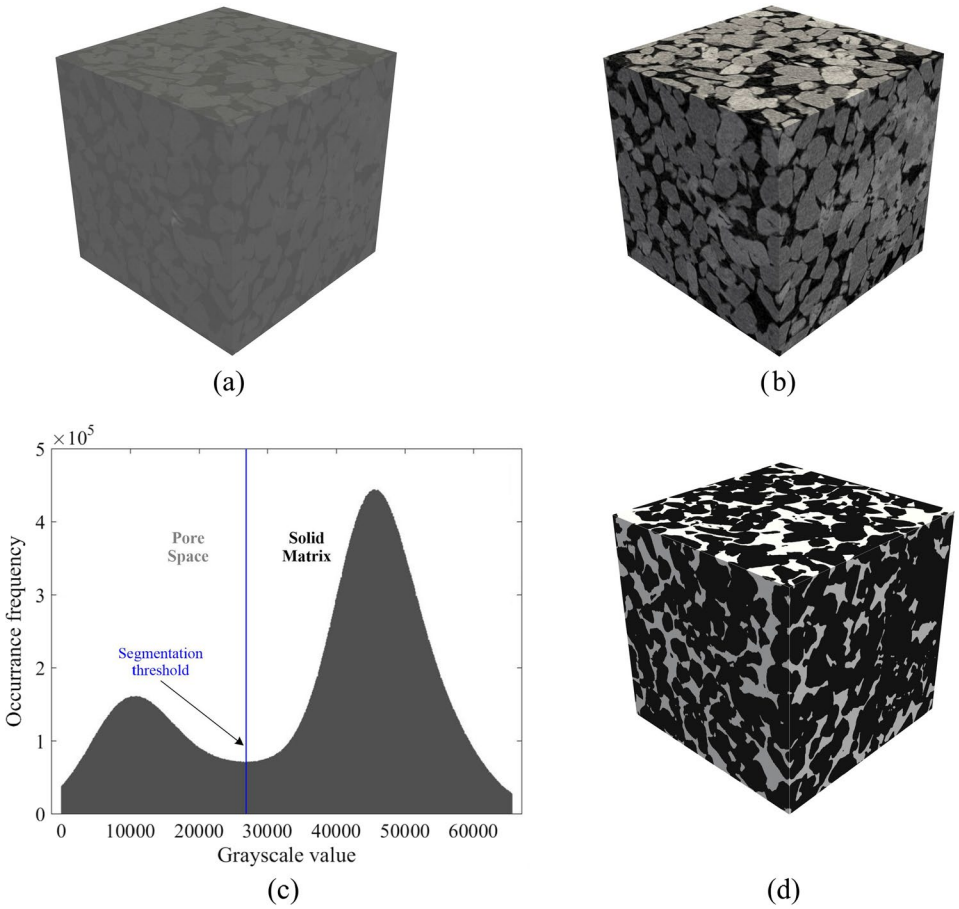


Fig. 3 Illustration of image processing and segmentation: (**a**) The raw micro-CT image of a Mt. Simon sandstone sample (resolution is 2.80 μm, and image size is 480^3 voxels); (**b**) the grayscale image after denoising and enhancement; (**c**) the histogram of voxel grayscale value; and (**d**) the binary image segmented by a global thresholding method (pore space is shown in white, and solid matrix is shown in black)

2.2 Stochastic microstructure reconstruction

The transport properties of porous media usually exhibit strong uncertainty, due to the random distribution of pore bodies. As a result, the limited number of digital rock samples obtained from micro-CT scans is far from sufficient to cover all possible morphology configurations of pore microstructures. In general, the complete computational dataset [33, 38] is an ensemble of representative/statistical volume elements that cover all morphological possibilities and share the same averaged characteristics, based on which a generalized prediction model can be achieved with high reliability.

Stochastic microstructure reconstruction [18, 33, 81] is an effective and economical approach to generating statistically equivalent samples of porous media, and the numerous reconstructed samples can be used to investigate the microstructure–property correlations when the availability of real porous media samples is limited. In this work, a high-fidelity reconstruction method developed in our previous study [33, 36, 38] is adopted to generate 3D pore microstructure samples. This method first characterizes the morphology patterns of the real 3D microstructures by fitting statistics-informed neural networks, based on which virtual 3D microstructure samples can then be generated via probability sampling. These virtual samples have been proven to preserve statistical equivalence, morphological similarity, long-distance connectivity, and transport properties of the real ones, and more details can be found in relevant references [33, 36, 38]. As shown in Fig. 4, Fontainebleau sandstone samples with different porosities are taken as examples to illustrate stochastic microstructure reconstruction using this new method. Guided by the morphological information extracted from the real digital microstructures, a total of 1270 virtual microstructure samples are reconstructed, with the image size varying from 200^3 to 320^3 voxels. The scanned digital rocks together with the reconstructed microstructure samples compose the raw dataset of 1455 samples in total for subsequent analyses.

2.3 Quantitative microstructure characterization

Quantitative characterization of porous microstructures in an explicit expression is the essential prerequisite to exploring the microstructure–property linkage of porous media. The pore space inside natural porous rocks usually exhibits great disorder and strong randomness, which needs to be quantified in statistical terms. Through quantitative characterization [103, 114], the microstructural complexity of a

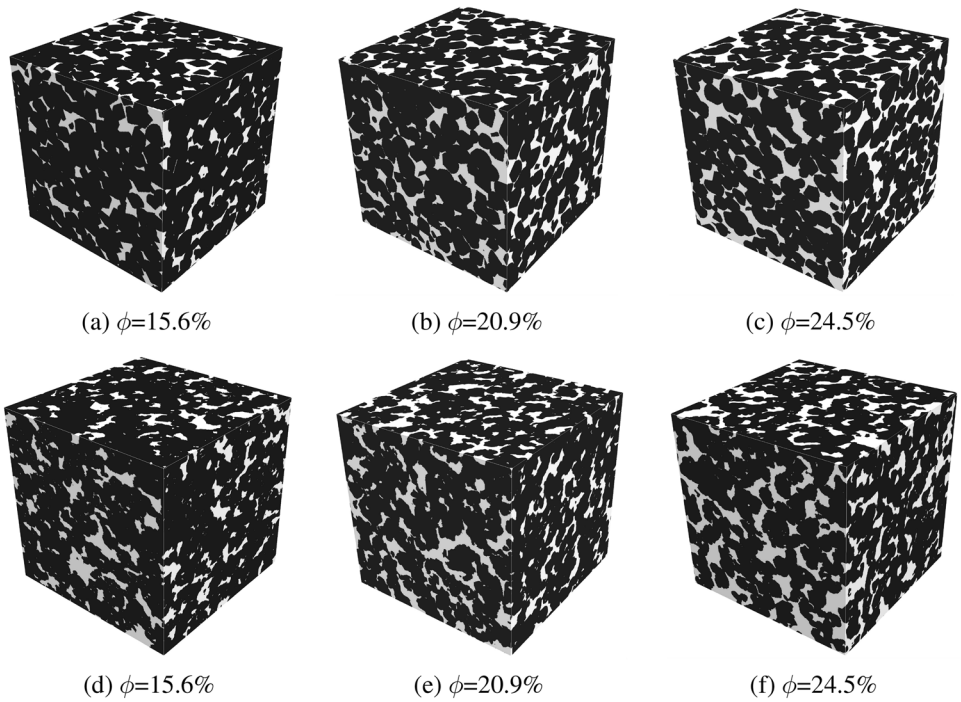


Fig. 4 Stochastic microstructure reconstruction (image size: 320^3 voxels): (**a**–**c**) are the scanned microstructures of Fontainebleau sandstones with different porosities ϕ; (**d**–**f**) are the representative reconstructed microstructures

porous medium can be reduced to a small set of morphological descriptors related to the macroscopic physical property of interest. A broad range of microstructure characterization approaches have been developed for porous media, such as statistical characterization, geometrical measurement, topological representation, and fractal analysis. As listed in Table 1, the commonly used morphological descriptors are collected through an extensive literature study, and they will be used as the microstructural features for permeability prediction. Many of these descriptors have been used to investigate the microstructure–property relations of porous media in the previous studies.

These morphological descriptors characterize porous microstructures from different perspectives, and they can be roughly grouped into four levels. Porosity and specific surface area are the typical descriptors at the first level to simply represent the global/mean properties of porous microstructures via single numbers, but they ignore the detailed morphological features of pore networks that may have significant effects on transport processes. As to the second level, local or size-dependent features are measured by such morphological descriptors as local porosity distribution, coarseness, local percolation probabilities, and lacunarity. When it comes to the third level, geometric attributes of porous media are quantified from various aspects such as pore shape, pore size, and surface roughness. The frequently used descriptors includes pore/throat size distribution, mean curvature, chord length distribution, lineal path function, and spatial correlations functions. The fourth level focuses on the topological characteristics of pore microstructures, which is related to long-distance connectivity and percolation of pore networks. Total curvature (Euler characteristic), two-point cluster correlation function, pair connectivity function, total fraction of percolating cells, and succolarity are commonly used indicators of connectivity. Pore coordination number represents the number of adjacent pore bodies connected to a specific pore. Besides, geometrical tortuosity characterizes the sinuosity and complexity of percolation paths inside porous media, while constriction factor quantitatively represents cross-sectional variation along pore channels.

All morphological descriptors in Table 1 are extracted from the microstructure dataset with 1455 samples, and they serve as the possible predictors to construct data-driven models for permeability prediction, as illustrated in Fig. 1. It is noted that some descriptors have multiple definitions and therefore different evaluation methods, and they are all used in this study to achieve a microstructure characterization as comprehensive as possible. As shown in Fig. 5, the results of average pore size, geometrical tortuosity, and construction factor are computed using different methods. Generally, different evaluation methods yield inconsistent values of morphological descriptors, but the results show similar changing trends and are highly correlated as well. There is no general standard to judge the rationality of the descriptor result calculated from a specific evaluation method, so feature selection could be an effective mean to choose an appropriate evaluation of a morphological descriptor. Besides, the evaluation results of another 12 representative descriptors are provided in Fig. 6.

Table 1 The collected morphological descriptors for quantitative characterization of porous microstructures

Index	Morphological descriptor (Evaluation method)	Denotation	Data dimension	Representative references
D1	Absolute porosity	ϕ	1	Carman [21], Adler [1]
D2	Effective porosity	ϕ_e	1	Géraud [39], Fu et al. [34]
D3	Specific surface area	S	1	Liang et al. [69], Cui et al. [29]
D4	Integral of mean curvature	m_2	1	Lehmann et al. [66]
D5	Integral of total curvature	m_3	1	Vogel et al. [105]
D6	Geometrical tortuosity (Direct shortest path search method)	τ_g	1	Koponen et al. [61], Cecen et al. [23]
D7	Geometrical tortuosity (Skeleton shortest path search method)	τ_g	1	Sevostianova et al. [95], Fu et al. [35]
D8	Constriction factor (Mercury intrusion porosimetry simulation)	β	1	Holzer et al. [47], Berg [10]
D9	Constriction factor (Morphological opening method)	β	1	Berg [10], Dong et al. [31]
D10	Mean chord length	$\langle z \rangle$	1	Coker et al. [26], Bertei et al. [12]
D11	Average pore coordination number	η	1	Hormann et al. [49]
D12	Average pore size (Continuous method)	d	1	Münch and Holzer [78]
D13	Average pore size (Discrete method)	d	1	Holzer et al. [46]
D14	Average pore size (Morphological opening method)	d	1	Paterson [82], Dong et al. [31]
D15	Average pore size (Random point method)	d	1	Torquato [102]
D16	Average pore size (Skeleton method)	d	1	Delerue et al. [30]
D17	Average pore throat size	d_t	1	Liang et al. [69]
D18	Effective pore size	d_e	1	Blair et al. [15]
D19	Hydraulic pore diameter	d_h	1	Bear [8]
D20	Characteristic length I	l_a	1	Coker et al. [26]
D21	Characteristic length II	l_b	1	Coker et al. [26]
D22	Characteristic length III	l_c	1	Coker et al. [26]
D23	Characteristic length IV	l_d	1	Ioannidis et al. [52]
D24	Average connectivity distance	l_e	1	Knudby and Carrera [58]
D25	Characteristic length V	$L^\star$	1	Hilfer [45]
D26	Fractal dimension	α	1	Yu and Cheng [118]
D27	Succolarity	ψ	1	Xia et al. [112]
D28	Lacunarity	$\delta(\varepsilon)$	10	N'Diaye et al. [79]
D29	Chord length distribution	$\rho(z)$	70	Muche and Stoyan [77], Cui et al. [29]
D30	Lineal path function	$L(z)$	50	Hilfer [45], Cui et al. [29]
D31	Spherical contact distribution function	$H_S(d)$	16	Lehmann et al. [66]
D32	1st Minkowski function	$m_0(d)$	16	Vogel et al. [105]
D33	2nd Minkowski function	$m_1(d)$	16	Armstrong et al. [5]
D34	3rd Minkowski function	$m_2(d)$	16	Vogel et al. [105]
D35	4th Minkowski function	$m_3(d)$	16	Armstrong et al. [5]
D36	Two-point correlation function	$S_2(r)$	50	Blair et al. [15], Fu et al. [33]
D37	Two-point cluster correlation function	$C_2(r)$	50	Jiao et al. [53], Cui et al. [29]
D38	Normalized auto-covariance function	$R(r)$	50	Bentz and Martys [9]
D39	Pair connectivity function	$H(r)$	50	Knudby and Carrera [58], Fu et al. [38]
D40	Surface-surface correlation function	$F_{SS}(r)$	20	Rubinstein and Torquato [88]
D41	Surface-void correlation function	$F_{SV}(r)$	50	Rubinstein and Torquato [89]
D42	Local porosity distribution	$\mu(\widetilde{\phi}, L = 50)$	100	Biswal et al. [14], Fu et al. [38]
D43	Local porosity distribution	$\mu(\widetilde{\phi}, L = L^\star)$	100	Hilfer [45], Fu et al. [38]
D44	Local percolation probabilities	$\lambda(\widetilde{\phi}, L = 50)$	50	Cosenza et al. [27], Fu et al. [38]

Table 1 (continued)

Index	Morphological descriptor (Evaluation method)	Denotation	Data dimension	Representative references
D45	Local percolation probabilities	$\lambda(\widetilde{\phi}, L = L^{\star})$	50	Cosenza et al. [27]
D46	Total fraction of percolating cells	$T_1(L)$	75	Latief et al. [65], Fu et al. [38]
D47	Total fraction of percolating cells	$T_3(L)$	75	Hilfer [45], Fu et al. [33]
D48	Pore coordination number distribution	$O(\eta)$	20	Hormann et al. [49]
D49	Pore size distribution (Continuous method)	$p(d)$	50	Münch and Holzer [78]
D50	Pore size distribution (Discrete method)	$p(d)$	30	Holzer et al. [46]
D51	Pore size distribution (Morphological opening method)	$p(d)$	20	Dong et al. [31]
D52	Pore size distribution (Random point method)	$p(d)$	30	Torquato [102]
D53	Pore size distribution (Skeleton method)	$p(d)$	30	Delerue et al. [30]
D54	Pore throat size distribution	$p(d_t)$	15	Lindquist et al. [70]
D55	Coarseness	$C(L)$	100	Quintanilla and Torquato [84]

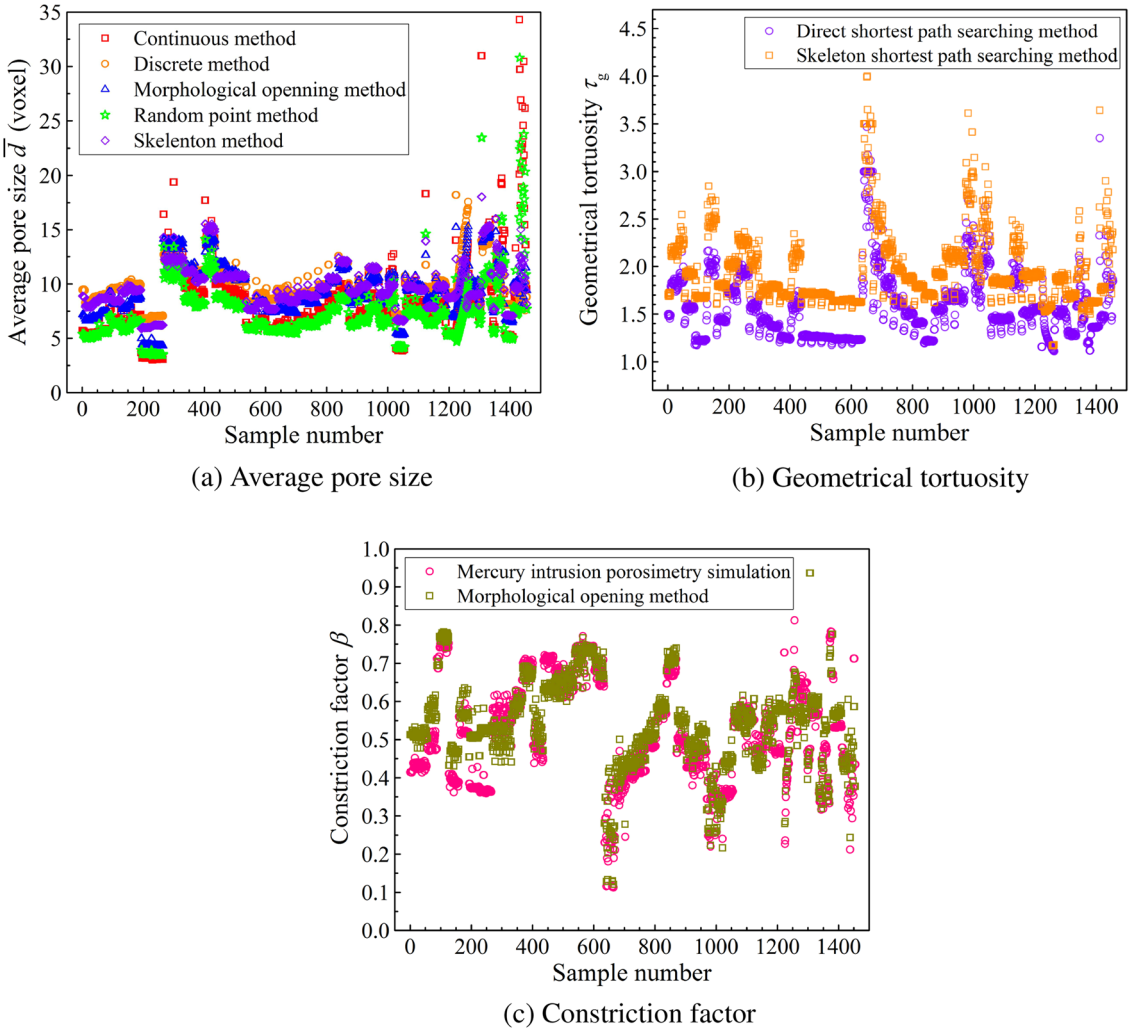


Fig. 5 The morphological descriptors with multiple definitions extracted from the digital microstructure dataset with 1455 samples

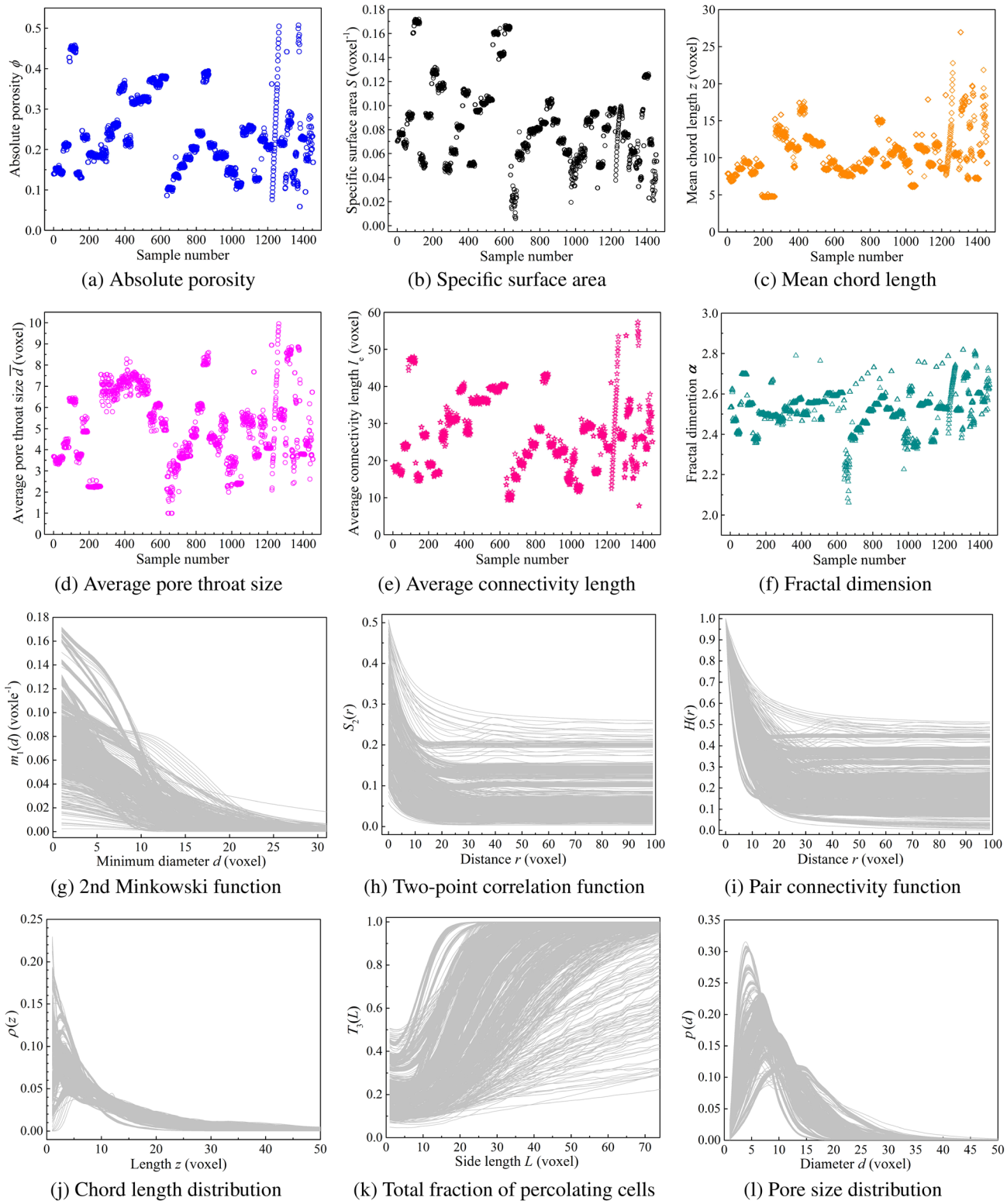


Fig. 6 The results of representative morphological descriptors extracted from the digital microstructure dataset with 1455 samples

As mentioned above, the digital rocks are collected from several open-access databases, so the image resolutions (voxel sizes) of them are slightly different, which are all around 5 μm. It means that microstructural analyses are conducted in voxel domains with different length scales to compute morphological descriptors. For the dimensionless descriptors, such as porosity, geometrical tortuosity, constriction factor, and poor coordination number, no additional data processing is required. As to the descriptors with length dimension, such as specific surface area, mean curvature, average pore size, and characteristic length, they are all quantified using voxel as the basic length unit, instead of converting them into the physical length scale. This treatment enables the seamless combination between the morphological descriptors in the voxel length unit and the LBM permeability in the lattice length unit, just by setting the lattice length equal to the voxel size for each porous media sample.

2.4 Pore-scale flow simulation

For pore-scale simulation of fluid flow, lattice Boltzmann method (LBM) [62] is more mathematically rigorous than pore network modeling (PNM), and the former can also provide more reliable permeability evaluations for porous media with complicated geometries [113]. Besides, the LBM simulation is directly performed on voxel domain of digital microstructures without any simplification, and the computed permeability values in the lattice unit can be directly linked to the morphological descriptors in the voxel unit, avoiding additional data conversion/processing. Therefore, LBM is adopted in this work to evaluate the intrinsic permeability values of these digital microstructure samples.

2.4.1 Basic theory of LBM

LBM [62, 111, 119] models the fluid flow through a time-dependent distribution of fluid particles propagating on a regular lattice. In DRP research, pore voxels in digital rock images serve as the regular lattice for LBM to simulate pore-scale fluid flow, and each lattice node is located in the center of corresponding pore voxel. The numerical grid of lattice Boltzmann simulation completely coincides with the image voxel grid in this study. The particle distribution function $f_i(\mathbf{x}, t)$ represents the probability of finding a fluid particle with the lattice velocity $\mathbf{c}_i$ in the location $\mathbf{x}$ and at the time t. Beginning with an initial state, $f_i(\mathbf{x}, t)$ moves from one lattice node to its neighboring nodes at each time step, and evolves itself locally subject to both mass and momentum conservation.

The conventional LBM scheme with the D3Q19 lattice arrangement and the Bhatnagar–Gross–Krook (BGK) collision operator [24] are adopted in this study. The evolution of $f_i(\mathbf{x}, t)$ along the direction of $\mathbf{c}_i$ from the time t to $t + \Delta t$ can be expressed as

$$f_i(\mathbf{x} + \mathbf{c}_i\Delta t, t + \Delta t) - f_i(\mathbf{x}, t) = -\frac{1}{\tau}\left[f_i(\mathbf{x}, t) - f_i^{\mathrm{eq}}(\mathbf{x}, t)\right], \quad (1)$$

where τ is the single-relaxation time, $f_i^{\mathrm{eq}}(\mathbf{x}, t)$ is the equilibrium distribution function, and the subscript i indicates the direction of lattice velocity around the lattice node. The relaxation time τ is a function of kinematic lattice viscosity ν of simulated fluid, i.e., $\nu = c_s^2\Delta t(\tau - 0.5)$, where c_s is the lattice speed of sound and it is assigned with the dimensionless value of $\sqrt{1/3}$.

The equilibrium distribution function $f_i^{\mathrm{eq}}(\mathbf{x}, t)$ corresponds to an ideal state where the particle distributions tend to a specific macroscopic state, to recover the macroscopic Navier–stokes equation. For the D3Q19 lattice arrangement with BGK collision operator, $f_i^{\mathrm{eq}}(\mathbf{x}, t)$ is expressed as [24]

$$f_i^{\mathrm{eq}}(\mathbf{x}, t) = w_i\rho\left[1 + 3(\mathbf{c}_i \cdot \mathbf{u}) + \frac{9(\mathbf{c}_i \cdot \mathbf{u})^2}{2} - \frac{3(\mathbf{u} \cdot \mathbf{u})}{2}\right], \quad (2)$$

where w_i is the weight factor of D3Q19 lattice structure, ρ is the fluid density, and $\mathbf{u}$ is the macroscopic fluid velocity. For the D3Q19 lattice model, the weight factors w_i are equal to $\frac{12}{36}$, $\frac{2}{36}$, and $\frac{1}{36}$ for the velocity directions of the central lattice node, face-connected neighbors, and edge-connected neighbors, respectively.

At the end of each time step, the macroscopic properties of fluid flow, including density ρ and velocity $\mathbf{u}$, can be approximated from $f_i(\mathbf{x}, t)$ through the following equations, and these macroscopic properties will be used for the LBM computation at the next time step

$$\rho = \sum_{i=1}^{n} f_i(\mathbf{x}, t) \quad (3)$$

$$\mathbf{u} = \frac{\sum_{i=1}^{n} f_i(\mathbf{x}, t)\mathbf{c}_i}{\rho}, \quad (4)$$

where n is the number of lattice directions ($n = 19$ in D3Q19 lattice structure used in this study).

2.4.2 Permeability evaluation

Driven by a constant pressure difference between the inlet and outlet faces, LBM is performed on the cubic digital rock sample to simulate a single-phase fluid flow with low Reynolds number ($Re \ll 1$) passing through it. In this study, small pressure gradients are applied to 3D porous media samples, to ensure permeability results are evaluated from laminar fluid flows. As shown in Fig. 7, the Mt. Simon sandstone sample in Fig. 3 is taken as the example to illustrate lattice Boltzmann simulation of pore-scale

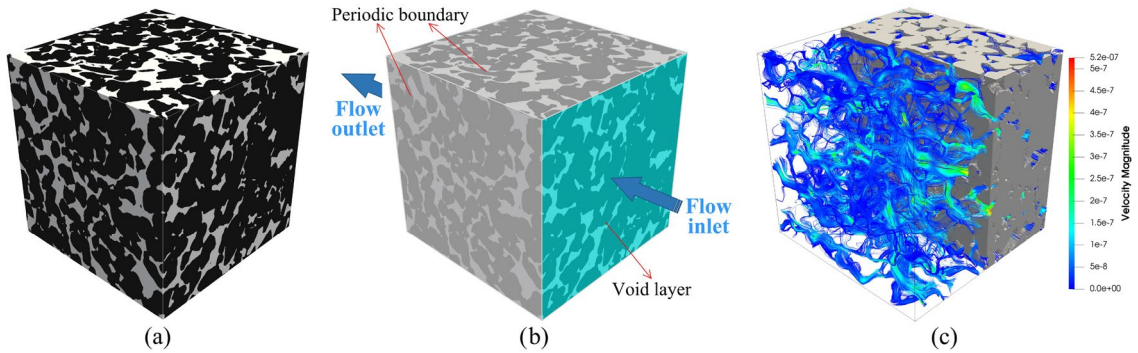


Fig. 7 Evaluation of intrinsic permeability through lattice Boltzmann simulation: (**a**) the 3D digital microstructure of a Mt. Simon sandstone sample; (**b**) the boundary conditions; and (**c**) the steady-state fluid velocity field inside the porous medium

fluid flow. When the fluid flow reaches a steady state, it can be described by Darcy's law, and the intrinsic permeability κ of this sample is quantified by the following equation:

$$\kappa = -\frac{\mu}{\nabla p}\langle \mathbf{u} \rangle, \tag{5}$$

where ∇p denotes the pressure gradient along the direction of macroscopic fluid flow, μ is the dynamic viscosity, and $\langle \mathbf{u} \rangle$ denotes the volume averaged fluid velocity across the entire simulation domain.

As the initial condition of lattice Boltzmann simulation is less important for steady-state flows and corresponding long-term behaviors, we simply assign the initial velocity $\mathbf{u} = 0$ and initial density $\rho = 1$ to all lattice nodes in the simulation domain [64]. Three types of boundary conditions are adopted for the pore-scale simulation: the no-slip boundary condition on the pore-solid surface, fixed pressure boundary condition at the inlet and outlet faces, and periodic boundary condition applied to the surfaces that are parallel to the main flow direction. The complete bounce-back scheme is implemented to simulate the no-slip boundary condition, where a fluid particle bounces back to the node it comes from with no relaxation when it meets a solid node. To apply the constant pressure difference, two void layers are added to both inlet and outlet faces [34, 54], and the pressure difference between the inlet and outlet is expressed by fluid density difference. The lattice Boltzmann simulation runs continuously until reaching the user-prescribed convergence criterion. In this case, the fluid flow is assumed to be stable when the standard deviation of average kinetic energy falls below 10^{-6} (the maximum number of iterations is 60,000). As shown in Figs. 7 and 9, lattice Boltzmann simulations are performed on eight representative digital rock samples to achieve the steady-state flow velocity fields for permeability evaluation.

The permeability value computed from lattice Boltzmann simulation is in dimensionless lattice unit, and it can be converted to the physical unit via the following equation [98]:

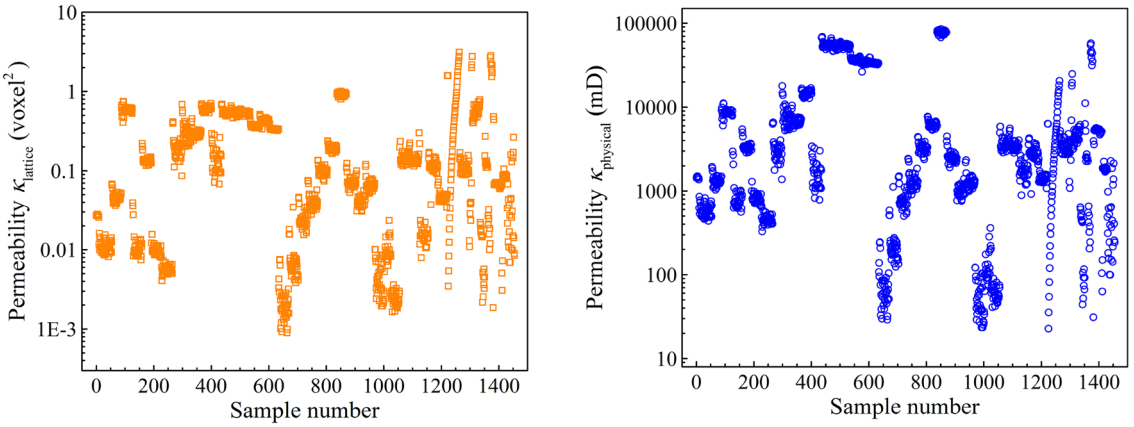
$$\kappa_{\text{physical}} = \kappa_{\text{lattice}} \left(\frac{L_{\text{physical}}}{L_{\text{lattice}}} \right)^2, \tag{6}$$

where κ_{physical} and κ_{lattice} are the permeability values in the physical and lattice unit, respectively; and L_{physical} and L_{lattice} are the lengths of any identical feature in the physical sample and the LBM domain, respectively. As the numerical grid of LBM coincides with the voxel grid of the digital microstructure in this study, the value of $\frac{L_{\text{physical}}}{L_{\text{lattice}}}$ is equal to the image resolution (voxel size).

For each porous media sample, lattice Boltzmann simulations of fluid flow are conducted along three-axial directions, and the average value of three directional permeabilities is used to investigate the microstructure–permeability relationship. As recorded in Fig. 8, the permeability values of the digital microstructure dataset with 1455 samples are plotted both in lattice unit and physical unit. From the above figures, one can see that the permeability values span in a broad range over 4 orders of magnitude. To avoid extra data processing, the permeability values in lattice unit will be used to explore the microstructure–permeability relation via feature selection and data-driven modeling, as illustrated in Fig. 1.

2.5 Feature selection

As listed in Table 1, a variety of morphological descriptors that quantitatively characterize the internal microstructures of porous media are collected. However, these descriptors are not equally important for permeability, and some of them represent overlapping features. In addition, the inconsistent results of morphological descriptors obtained from different evaluation methods can also negatively affect investigation of the microstructure–permeability linkage. Hence, a brute-force regression model based on all available morphological



(a) Permeability values in lattice unit (b) Permeability values in physical unit

Fig. 8 The permeability results evaluated from lattice Boltzmann simulations for the digital microstructure dataset with 1455 samples

descriptors cannot provide accurate and reliable permeability prediction, due to the noise from irrelevant (or less important) descriptors and the conflicts between overlapping descriptors. Besides, the unnecessary involvement of irrelevant and abundant features can increase the model complexity and make it harder to interpret. Therefore, feature selection is an indispensable step for predictive model construction, where the most relevant and significant features are to be identified from a large set of morphological descriptors in Table 1.

The objectives of feature selection in this work include: (1) enhancing interpretability of the implicit regression model to obtain deep insights into the underlying dependence of permeability on microstructural characteristics; (2) reducing the computational complexity and avoiding overfitting to built a cost-effective predictor using the selected features; (3) achieving a generalized and rational model with the optimal performance in permeability prediction.

For a dataset of m observations $\{\mathbf{x}_i, \kappa_i\}$ $(i = 1, 2, ..., m)$ consisting of n input feature variables $x_{i,j}$ $(j = 1, ... n)$ and an output permeability value κ_i, various methods can be applied to select the feature variables $x_{i,j}$ that are important to the response κ_i. Generally, feature selection techniques [42, 67] can be divided into three categories: filter, embedded, and wrapper methods. Among them, filter type feature selection is independent of learning algorithms, while wrapper and embedded methods interact with a particular learning process. All these methods are tried and tested in this study to identify the most suitable feature selection for the morphological descriptors that best represent permeability of porous media.

2.5.1 Filter type feature selection

Filter type feature selection [42, 67] assesses feature importance according to certain data characteristics, so it is unrelated to any learning algorithms. Typically, a filter method consists of two steps: feature importance ranking and feature filtering. Different feature evaluation criteria have been proposed to rank feature importance, such as feature correlation, mutual information, the feature discriminative ability, the feature ability to maintain the data manifold, and the feature capacity to reconstruct the raw data. Four representative criteria of feature importance evaluation are covered in this study, which are Pearson's correlation coefficient $R\left[\widetilde{\mathbf{x}}_j\right]$, RReliefF importance weight $W\left[\widetilde{\mathbf{x}}_j\right]$ [86], F-test importance score F [6], and nearest-neighbor-based feature weight $f(\boldsymbol{w})$ [116]. More algorithm details can be found in relevant references.

2.5.2 Embedded type feature selection

Embedded methods [42, 67] conduct feature selection during the learning processes, which are deeply embedded in specific learning algorithms. For example, during the training process of a decision tree [72], feature importance is evaluated from the sum of changes in the mean squared error due to splits on each feature and the number of branch nodes. For a random forest [19], feature importance can be evaluated by permutation to measure the influence degree of a feature variable in predicting the response. As to Gaussian process regression [94], feature importance can be evaluated from corresponding separate length scales of the kernel function. In this study, the above three learning algorithms are tested to assess the importance of morphological descriptors to permeability modeling, and more algorithm details can be found in relevant references.

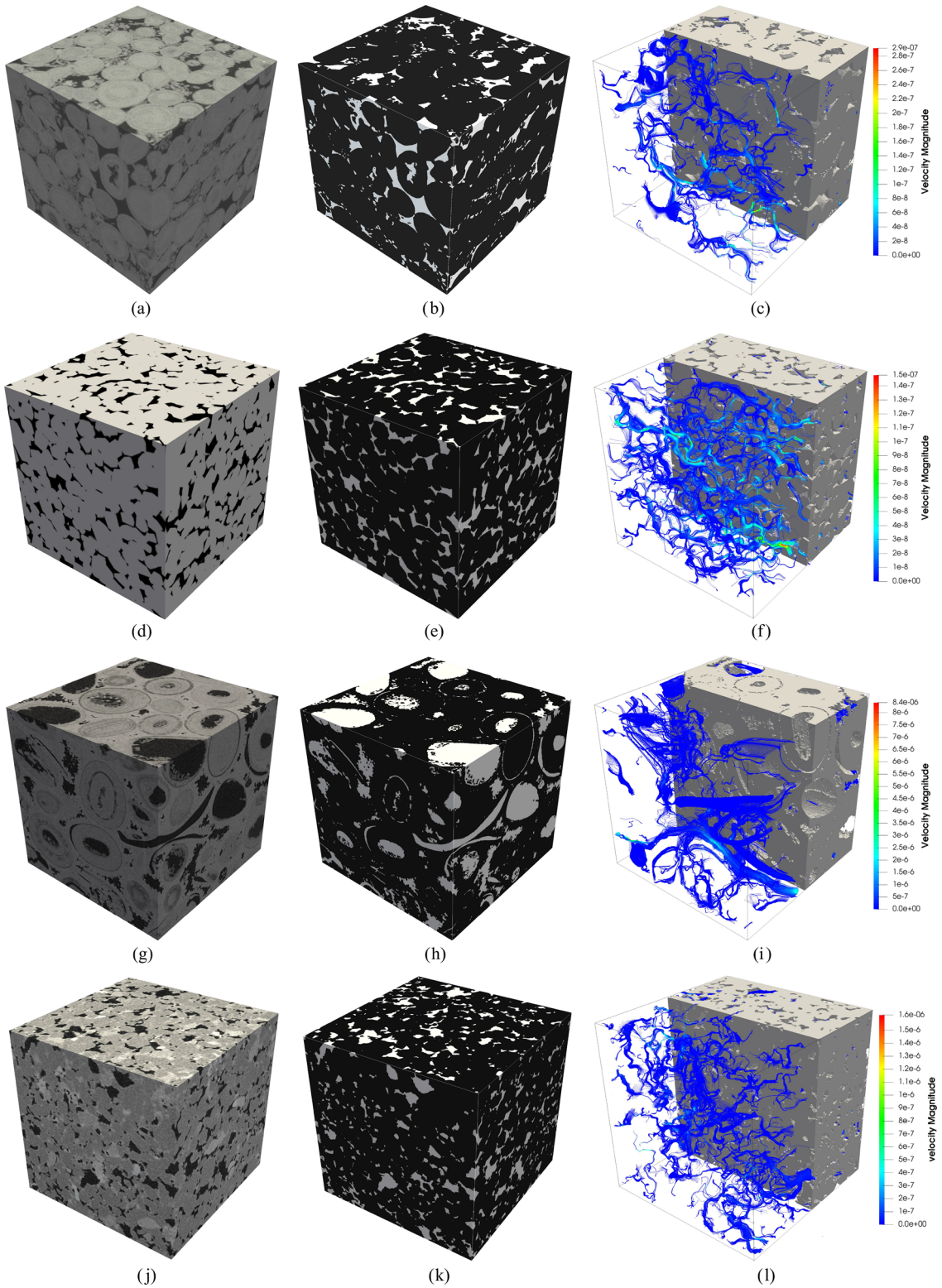


Fig. 9 Digital rock samples, image segmentation, and lattice Boltzmann simulations: The micro-CT scanning images of (**a**) Ketton carbonate [92], (**d**) Fontainebleau sandstone [65], (**g**) Savonnières carbonate [20], and (**j**) Leopard sandstone [44]; (**b**), (**e**), (**h**), and (**k**) are the segmented images; (**c**), (**f**), (**i**), and (**l**) are the steady-state flow velocity fields inside porous microstructures

2.5.3 Wrapper type feature selection

Wrapper type feature selection is more applicable to heterogeneous features, compared to the filter and embedded methods. Considering the dimension differences between morphological descriptors listed in Table 1, wrapper type feature selection is attractive. Wrapper methods [43, 60] assess the quality of feature selection according to the prediction performance of the predefined learning algorithms. It searches the optimal feature subset through greedily evaluating the possible combinations of features based on a certain evaluation criterion. For regression problems, the coefficient of determination R^2, the mean-squared error (MSE), and the mean-absolute error (MAE) can be used as the metrics to evaluate the model performance, which are mathematically expressed by the following equations, respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(\widetilde{\kappa}_i - \widehat{\widetilde{\kappa}}_i\right)^2}{\sum_{i=1}^{m} \left(\widetilde{\kappa}_i - \overline{\widetilde{\kappa}}\right)^2} \tag{7}$$

$$\mathrm{MSE} = \frac{1}{m} \sum_{i=1}^{m} \left(\widetilde{\kappa}_i - \widehat{\widetilde{\kappa}}_i\right)^2 \tag{8}$$

$$\mathrm{MAE} = \frac{1}{m} \sum_{i=1}^{m} \left|\widetilde{\kappa}_i - \widehat{\widetilde{\kappa}}_i\right|, \tag{9}$$

where $\widetilde{\kappa}_i$ and $\widehat{\widetilde{\kappa}}_i$ are the target and predicted permeability respectively corresponding to the i-th porous media sample, and $\overline{\widetilde{\kappa}}$ is the average value of the target permeability. Among them, R^2 quantifies the degree to which the feature variables explain the variation of the response, and its value ranges from 0 to 1, where a larger value indicates a better model performance.

Exhaustive search is a "brute-force" strategy in wrapper type feature selection, which usually requires enormous amounts of computation, especially when the number of feature variables is large. By contrast, greedy search strategies are of lower computation cost, which can be further divided into two categories: forward selection and backward elimination. Here, the wrapper method with sequential forward adding strategy [40] is adopted, as explained in Fig. 10.

Starting with a null model, each morphological descriptor in Table 1 is used individually to construct a predictive surrogate model, and the descriptor that achieves the best predictive performance (the maximum R^2 or the minimum RMSE value) is picked out as the first selected feature. A new predictive model with two features is then constructed by sequentially combining the previously selected feature with one of the remaining descriptors, and the descriptor resulting in the largest R^2 or the smallest RMSE is selected as the second feature. The above procedure is repeated iteratively until no improvement of prediction performance or reaching the desired number of included features, and a subset of features are consequently selected through this performance-orientated process.

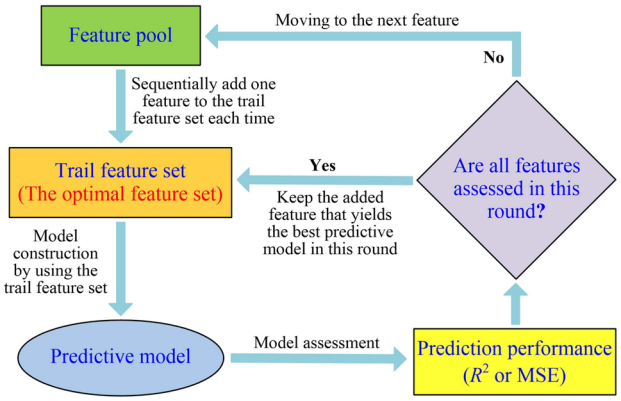


Fig. 10 The flowchart of wrapper type feature selection through a sequential forward adding strategy (it should be noted that this flowchart is only for one round of feature selection, and the remaining rounds just repeat this procedure)

3 Feature selection results

As listed in Table 1, the first 27 morphological descriptors are in the format of a single number, while the remaining 28 descriptors are in the form of distributions with different data dimensions. Generally, the filter and embedded methods are not applicable to feature selection with heterogeneous data, while the wrapper type feature selection possesses good versatility. Considering the above situation, feature selection is first performed on the first 27 morphological descriptors using different methods, and later the wrapper method is applied to select features from the entire feature pool with all 55 descriptors.

3.1 Data normalization

After microstructural characterization and pore-scale simulation, the paired data with m observations $\{\mathbf{x}_i, \kappa_i\}$ can be obtained to study the microstructure–permeability relation. For both feature selection and data-driven modeling, the scale of labeled data can greatly affect the results, and thus, data normalization is required to deal with this issue. As the jth feature variable, $\mathbf{x}_j$ within a range of interest can be scaled using the minimum and maximum values, given by

$$\widetilde{\mathbf{x}}_j = \frac{\mathbf{x}_j - \frac{\mathbf{x}_j^{(max)} + \mathbf{x}_j^{(min)}}{2}}{\frac{\mathbf{x}_j^{(max)} - \mathbf{x}_j^{(min)}}{2}} = \frac{2\mathbf{x}_j - \left(\mathbf{x}_j^{(max)} + \mathbf{x}_j^{(min)}\right)}{\mathbf{x}_j^{(max)} - \mathbf{x}_j^{(min)}}, \tag{10}$$

where $\mathbf{x}_j^{(max)}$ and $\mathbf{x}_j^{(min)}$ are the maximum and minimum values, respectively, of the jth feature variable. As to the output variable, the range of permeability value κ_i is not known for unseen data, so it is statistically normalized as follows:

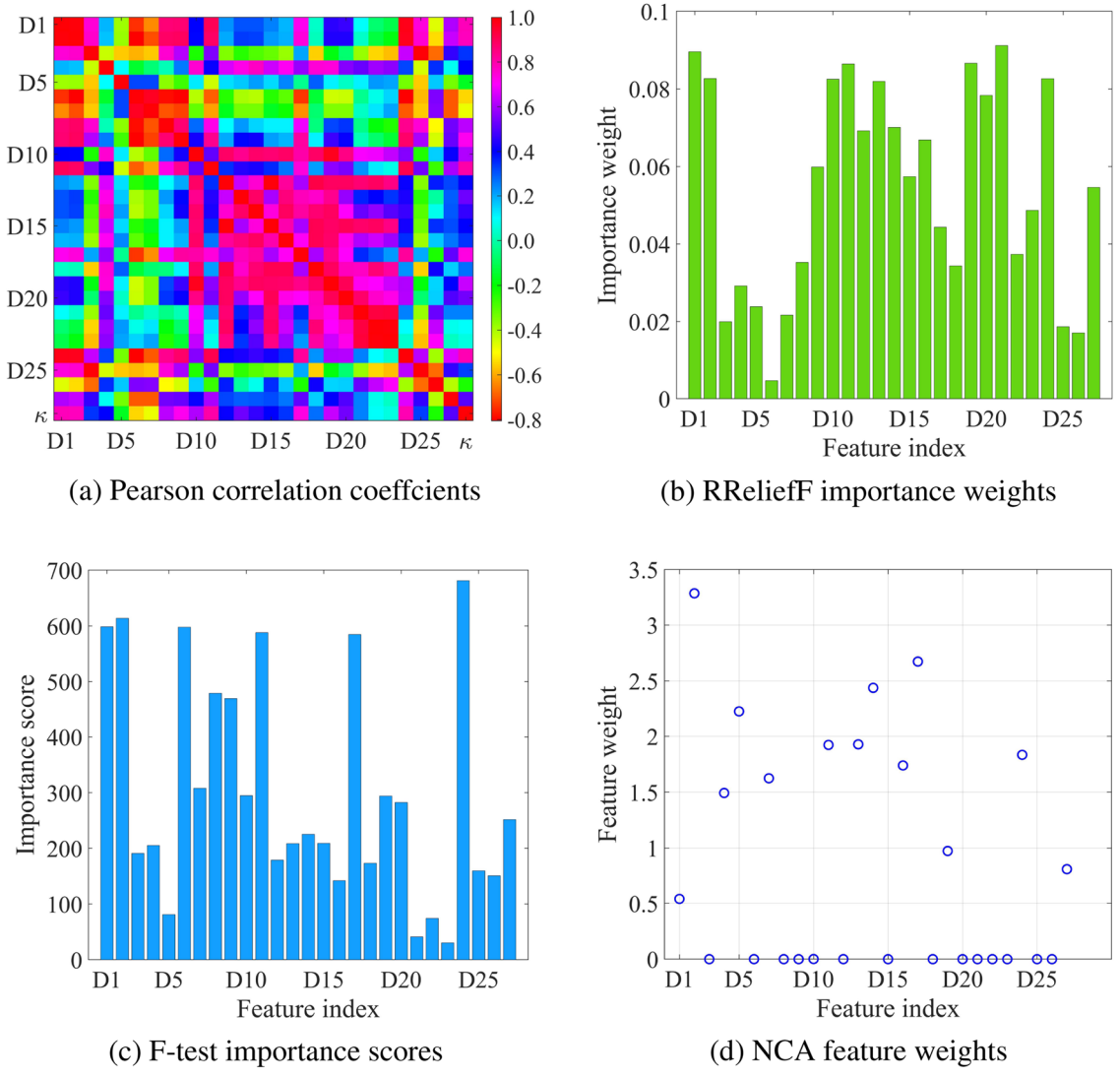
$$\widetilde{\kappa_i} = \frac{\kappa_i - \overline{\kappa}}{\sigma_\kappa}, \tag{11}$$

where $\overline{\kappa}$ and σ_κ are the mean value and standard deviation, respectively, of permeability data.

3.2 Filter type feature selection results

As plotted in Fig. 11, the results of feature importance ranking are estimated for the first 27 descriptors using four different filter methods. Due to the evaluation criterion difference of feature selection, the importance ranking results estimated from these methods are not completely consistent with each other, but the overall assessment results are similar. In all four filter methods, effective porosity (D2) and average connectivity distance (D24) are identified as the influential microstructure characteristics to permeability (κ), which agrees with the common knowledge of porous media [1, 21, 34].

However, specific surface area (D3) is evaluated to be an insignificant/irrelevant feature, which is contrary to the general consensus that the specific surface area is critical



(a) Pearson correlation coeffcients

(b) RReliefF importance weights

(c) F-test importance scores

(d) NCA feature weights

Fig. 11 The feature importance ranking results estimated from four filter methods

to permeability of porous media [1, 21, 34]. Filter methods evaluate the importance of feature variables individually, but a feature variable that is recognized to be unimportant by itself can be significant to the response when used with the other features [42]. Basically, filter methods are unable to detect the joint importance of multi-variable features, which is one of their main drawbacks.

3.3 Embedded type feature selection results

Three embedded methods are also used to assess the importances of the first 27 morphological descriptors to permeability, and corresponding results are plotted in Fig. 12. Because different learning algorithms are embedded in these three feature selection processes, the importance rankings of morphological descriptors are not completely consistent. Similar to the results of filter methods, effective porosity (D2) and average connectivity distance (D24) are also selected as important microstructure features by embedded methods, but specific surface area (D3) is assigned with low importance scores, especially in the regression tree and the GPR model. Besides, embedded type feature selection is associated with specific learning algorithms, which is inflexible for prediction model construction.

In summary, the intended purpose of feature selection has not been achieved using the filter or embedded method. Feature importance has been missed for some specific descriptors that are known to be critical to permeability of porous rocks, while only the scaler-valued morphological descriptors in Table 1 are covered by these two methods. This task will be continued with the wrapper type feature selection in the following part.

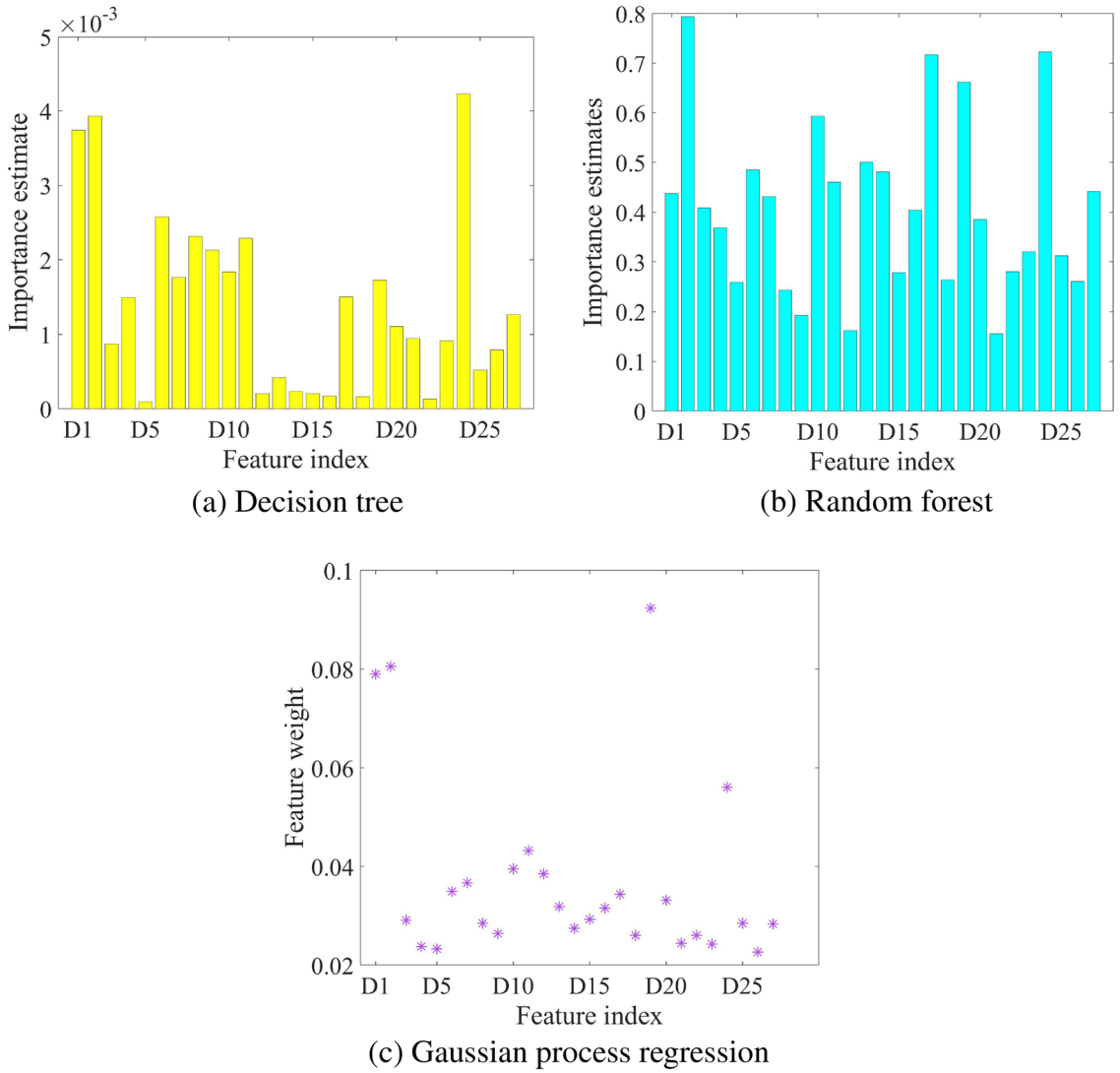


Fig. 12 The results of feature importance ranking evaluated from three embedded methods

3.4 Wrapper type feature selection results

As explained in Sect. 2.5.3, wrapper type feature selection is highly interrelated to the learning algorithm. Therefore, it is crucial to choose an appropriate learning algorithm for both wrapper-based feature selection and data-driven modeling. On the one hand, the chosen learning algorithm should possess strong learning capacity to deal with high-dimensional data; on the other hand, the model response should also be sensitive to influential feature variables to capture the feature significance. After conducting the comparison between linear regression, decision tree, random forest, support vector machine, and feed-forward neural network (FNN), the FNN with a shallow architecture is found to be the most appropriate learning algorithm for this study.

3.4.1 Feed-forward neural network

Artificial neural networks [117] are function approximators to map the inputs to the output through many interconnected computation elements called neurons. Each elementary neuron possesses a certain degree of approximation capacity, and a powerful learning performance can be achieved by cohesively combining many neurons. It has been proved that a fairly simple neural network is capable of fitting many practical functions [63]. The feed-forward neural network with a shallow architecture is adopted to construct the implicit microstructure–permeability model in this study.

As illustrated in Fig. 13, morphological descriptors are used as the feature variables to feed an FNN model with one or two hidden layer(s), and the final output is a permeability prediction. The predicted permeability $\widehat{\widetilde{\kappa}}$ is computed through a series of forward-propagation equations that occur at particular layers, given by

$$\begin{cases} \text{Input layer}: \ \boldsymbol{y}_0 = \widetilde{\mathbf{x}} \\ \text{Hidden layer I}: \ \boldsymbol{y}_1 = \tanh\left(\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{y}_0 + b_1\right) \\ \text{Hidden layer II}: \ \boldsymbol{y}_2 = \tanh\left(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{y}_1 + b_2\right) \\ \text{Output layer}: \ \widehat{\widetilde{\kappa}} = \boldsymbol{y}_3 = \boldsymbol{W}_3^{\mathrm{T}}\boldsymbol{y}_2 + b_3, \end{cases} \tag{12}$$

where $\widetilde{\mathbf{x}}$ denotes the input features; $\boldsymbol{y}_k$ denotes the output of the kth layer; $\boldsymbol{W}_i$ and b_i are the weight matrix and bias of the ith layer, respectively; $\tanh(\cdot)$ denotes the activation function (hyperbolic tangent function is adopted here).

Essentially, the above FNN model is a vector-valued network surrogate to approximate the input–output relation of the $\widetilde{\mathbf{x}}$-$\widetilde{\kappa}$ mapping, and the approximation function can be mathematically expressed as follows:

$$\mathcal{FNN}\left(\widetilde{\mathbf{x}};\mathbf{W},\boldsymbol{b}\right) : \ \widetilde{\mathbf{x}} \in \mathbb{R}^{d_{\mathbf{x}}} \rightarrow \widetilde{\kappa} \in \mathbb{R}^{d_{\kappa}}, \tag{13}$$

where $\mathcal{FNN}(\cdot)$ denotes the approximation function of the FNN model; $d_{\mathbf{x}}$ and d_{κ} are the data dimensions of the input and the output, respectively.

The next key issue here is to optimally adjust the weight matrices $\mathbf{W}$ and bias vector $\boldsymbol{b}$ of the neural network by making full use of the available labeled data. Basically, data-driven training is to optimize $\mathbf{W}$ and $\boldsymbol{b}$ of the neural network by minimizing of the discrepancy between the targets $\widetilde{\kappa}$ and the outputs $\widehat{\widetilde{\kappa}}$ for the observational data. This optimization problem can be mathematically expressed as follows:

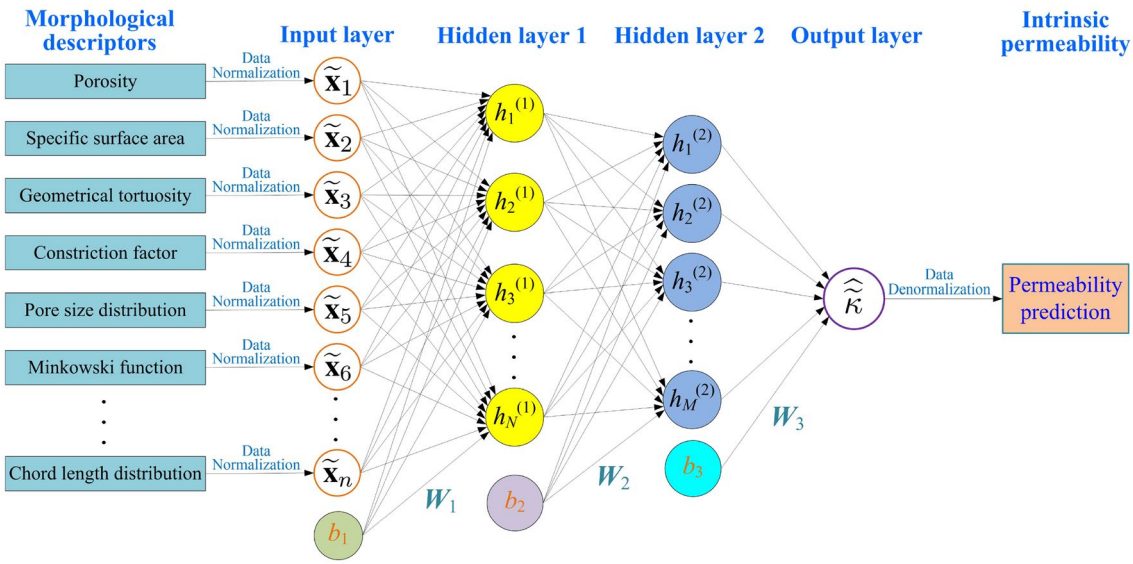


Fig. 13 The graphic illustration of an FNN model with two hidden layers for permeability prediction

$$\arg\min_{\mathbf{W},\boldsymbol{b}} \ \mathcal{L}\left(\widetilde{\mathbf{x}},\widetilde{\kappa};\mathbf{W},\boldsymbol{b}\right) = \frac{1}{m}\sum_{i=1}^{m}\left\|\widetilde{\kappa}_i - \widehat{\widetilde{\kappa}}_i\right\|_2^2 + \gamma\left\|\mathbf{W}\right\|_F^2$$
$$= \frac{1}{m}\sum_{i=1}^{m}\left\|\widetilde{\kappa}_i - \mathcal{FNN}\left(\widetilde{\mathbf{x}}_i;\mathbf{W},\boldsymbol{b}\right)\right\|_2^2 + \gamma\left\|\mathbf{W}\right\|_F^2, \tag{14}$$

where $\mathcal{L}\left(\widetilde{\mathbf{x}},\widetilde{\kappa};\mathbf{W},\boldsymbol{b}\right)$ denotes the loss function, and $\gamma \geq 0$ is the weight regulation constant. The first term of the loss function is the mean-squared error (MSE) to represent the discrepancy between the targets $\widetilde{\kappa}$ and the predictions $\widehat{\widetilde{\kappa}}$. The second term is the L_2 weight regulation term, also called weight decay, which can force the network response to be smoother and thus to reduce over-fitting.

The above minimization problem can be solved by many optimization algorithms, such as stochastic gradient descent methods. Generally, Levenberg–Marquardt algorithm (LMA) [76] is considered to be the most cost-effective method to train moderate-sized feed-forward neural networks (up to several hundred weights) with high accuracy, especially for regression problems. Therefore, LMA is adopted here to obtain the optimal weights and biases of feed-forward neural networks for the purpose of function approximation. Besides, cross-validation is usually performed to avoid over-fitting, thereby improving the generalized predictive ability for new observations. More details about parameter optimization and cross-validation of artificial neural networks can be found in relevant references [63, 117].

3.4.2 Feature importance indicator

After data normalization (as explained in 3.1), the entire dataset is randomly split into three subsets: training (50%), validation (25%), and test (25%). The training dataset is used to fit the neuron network, where network parameters are optimally adjusted to minimize regression error. The validation dataset is used to measure network generalization, and the training process is halted when generalization stops improving, so as to avoid over-fitting. The testing dataset is used to provide an independent measure of the model performance on unseen data.

Due to the randomness of initial configurations (such as random sampling of initial network parameters, and random division of the entire dataset for training, validation, and testing), training the neural network multiple times usually generates different results (as illustrated in Fig. 14a). Here, the feed-forward neural network is trained for 100 times for each case of input features, and the average value of R^2 over these 100 trials is used as the indicator to represent feature importance for feature selection using the wrapper method. The hyper-parameters of feed-forward neuron networks are summarized in Table 2.

In Fig. 14b, the average values of R^2 over 100 trials are used as the indicators to represent feature importance for all 55 morphological descriptors in Table 1, where each descriptor is used individually to train the neural network. The first 27 morphological descriptors in Table 1 are in the format of a single number, which are represented by the blue bars, and we call them *blue descriptors* for convenience. As to the remaining 28 morphological descriptors, they are in the format of a distribution, which are called *green descriptors* here.

3.4.3 Feature selection strategy I

In this part, feature selection is restricted to the *blue descriptors* (the first 27 descriptors in Table 1) using the wrapper method, aiming to built a cost-effective surrogate model with fewer input variables. The methodology of wrapper type feature selection is graphically illustrated in Fig. 10. In the first round of feature selection, the blue descriptors are used individually to fit the neural network one by one, and the regression performances are recorded in Fig. 14b. Effective porosity (D2) is selected as the most important feature in this round, because it yields a regression model with the best predictive performance ($\overline{R^2} = 0.8430$) among all the blue descriptors. This feature selection result is consistent with that of the filter and embedded methods, as shown in Figs. 11 and 12.

Table 2 The hyper-parameters of neural networks trained by using Levenberg–Marquardt algorithm

Data-driven modeling	Model hyper-parameters		Algorithm hyper-parameters						
	Number of hidden layers	Neurons in per hidden layer	Initial damping factor θ	Decrease factor for θ	Increase factor for θ	Minimum value for θ	Maximum value for θ	Regulation constant γ	Maximum validation failures
Predictive model I (Selection strategy I)	2	15	10^{-3}	0.1	10	10^{-7}	10^{10}	5×10^{-3}	20
Predictive model II (Selection strategy II)	2	30	10^{-3}	0.1	10	10^{-7}	10^{10}	5×10^{-3}	50

Note: θ is the combination coefficient in Levenberg–Marquardt algorithm, and its meaning can be found in the appendix section of this paper

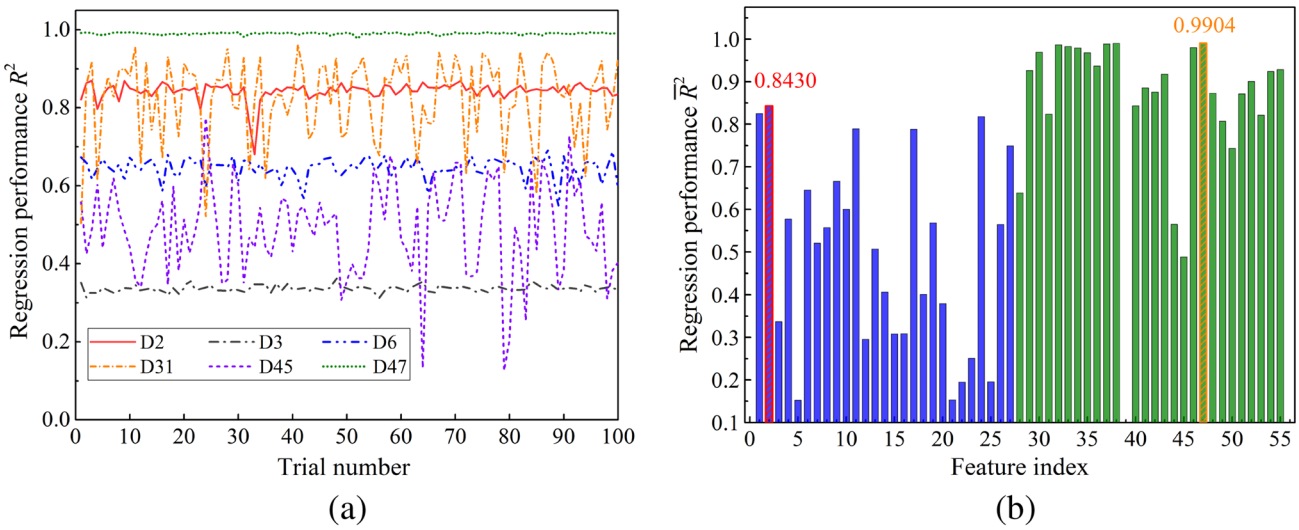


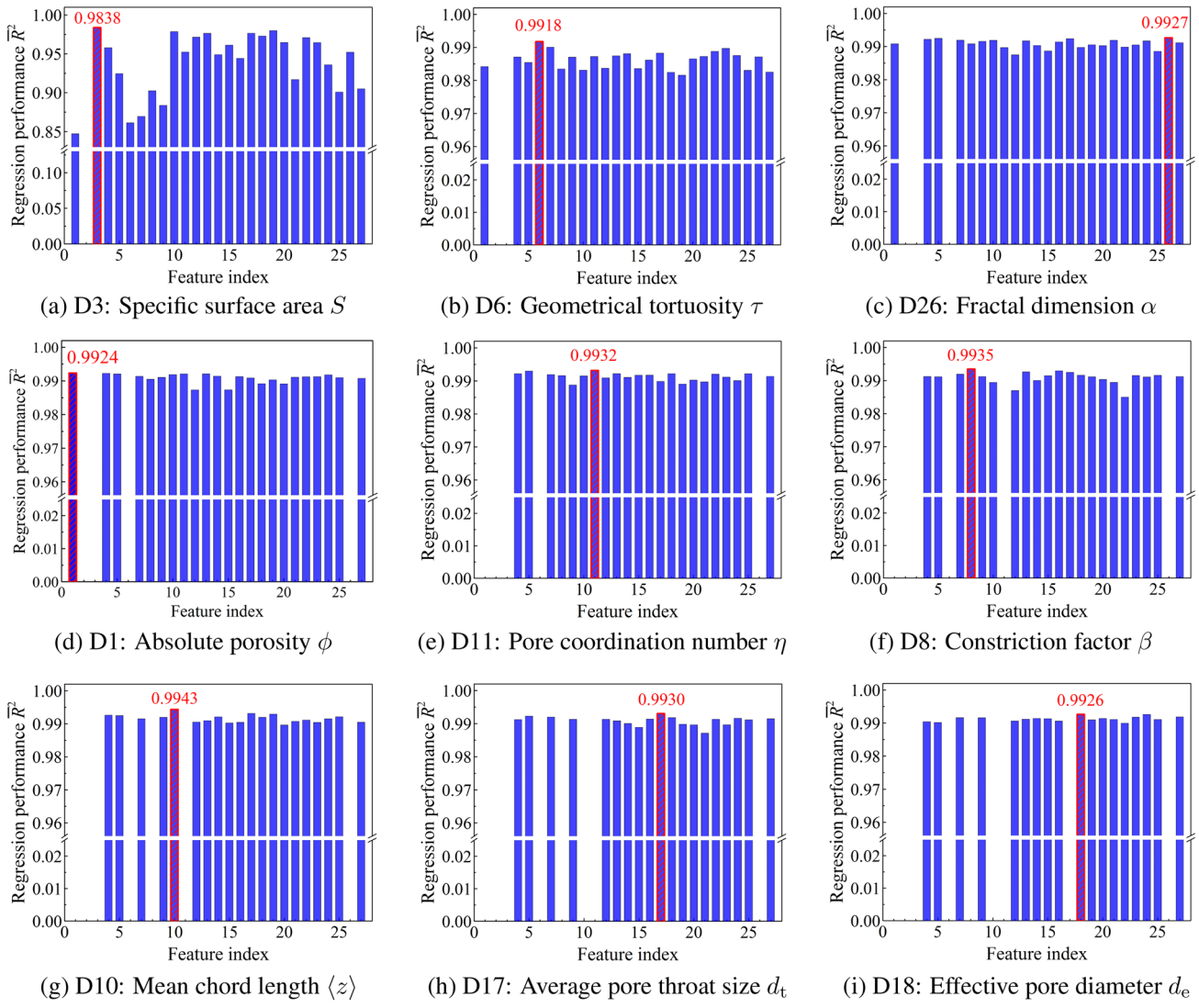
Fig. 14 Regression performances of the FNN models trained by individual features: (**a**) The varying values of R^2 for different trails; (**b**) The average regression performance $\overline{R^2}$ is used as the indicator to represent feature importance for all 55 morphological descriptors in Table 1

In the second round, D2 is combined with the remaining descriptors one by one to jointly fit the neural network, and the regression performances are remarkably improved, as shown in Fig. 15a. The maximum value of $\overline{R^2}$ is 0.9838, and the corresponding descriptor D3 (specific surface area) is selected as the second feature. In contrast to the filter and embedded methods, the significance of specific surface area to permeability can be well recognized by the wrapper type feature selection. Repeating the above procedures, morphological descriptors D6, D26, D1, D11, D8, D10, D17, and D18 are then successively identified, as illustrated in Fig. 15b–i.

As illustrated in Fig. 16a, the regression performance $\overline{R^2}$ increases continuously when more selected morphological descriptors are included in the FNN model. However, the regression performance reaches the peak ($\overline{R^2} = 0.9943$) when the first eight selected descriptors are used to train the FNN model, which is highlighted by the red star in Fig. 16a. Continuing to add more selected descriptors for model training, the regression performance starts to decline. Therefore, the optimal feature set obtained from the wrapper method contains eight morphological descriptors, which are: effective porosity (D2), specific surface area (D3), geometrical tortuosity (D6), fractal dimension (D26), absolute porosity (D1), pore coordination number (D11), constriction factor (D8), and mean chord length (D10).

The above feature selection result agrees with the existing knowledge of porous media, and the selected morphological descriptors have all been directly used to build analytical/empirical formulas for permeability evaluation in the previous studies. For instance, the first three selected descriptors (effective porosity, specific surface area, and geometrical tortuosity) are used in the well-known Kozeny–Carmon relation [21, 25, 34]. Different from the filter and embedded methods that only analyze the relationship between an individual descriptor and permeability, the wrapper method selects features in a multivariable analytical manner.

Essentially, the optimal feature set consist of eight morphological descriptors quantitatively characterize pore network systems inside porous media from seven different perspectives, based on which the dependence of permeability on microstructural characteristics can be interpreted as follows: (1) Absolute and connected porosity represent the entire pore space and the permeable portion permitting fluid to flow through, respectively; (2) Specific surface area approximately reflects the area of fluid–solid interface that provides adhesive friction to fluid flow; (3) Geometrical tortuosity measures the sinuosity of percolating pore paths that extends the average length of flow streamlines; (4) Fractal dimension is a measurement of scaling irregularity and complexity of porous microstructures in which transport phenomenon of fluid flow occurs; (5) Average pore coordination number characterizes the topology properties of porous media, which represents the number of adjacent pore bodies connected to a specific pore body; (6) Constriction factor quantifies the degrees of cross-section variation along pore channels, which can converge and diverge fluid streamlines and thus hinder transport flow; (7) Mean chord length measures the spatial distances between opposite walls of pore channels allowing fluid to pass through.



(a) D3: Specific surface area S (b) D6: Geometrical tortuosity τ (c) D26: Fractal dimension α

(d) D1: Absolute porosity ϕ (e) D11: Pore coordination number η (f) D8: Constriction factor β

(g) D10: Mean chord length $\langle z \rangle$ (h) D17: Average pore throat size d_t (i) D18: Effective pore diameter d_e

Fig. 15 The selected descriptor at each round via the wrapper type feature selection

3.4.4 Feature selection strategy II

Despite the encouraging results, the *green descriptors* have not been included in the above discussion. In this part, the wrapper type feature selection is performed on all morphological descriptors listed in Table 1, expecting to identify a set of descriptors that comprehensively characterizes porous microstructures, from which a deep insight into the microstructure–permeability relation can be obtained.

In the first round of feature selection, D47 (total fraction of percolating cells) is picked out, because it yields the neural network model with the best performance ($\overline{R^2} = 0.9904$), as shown in Fig. 14b. In the second round, D47 is combined with the remaining descriptors one by one to jointly fit the neural network, and the regression performances are shown in Fig. 17a. The maximum value of $\overline{R^2}$ is 0.9921, and the corresponding descriptor D6 (geometrical tortuosity) is selected as the second feature. Repeating the above procedures, morphological descriptors D3, D2, D27, D25, D1, D11, D15, and D26 are then successively picked out, as demonstrated in Fig. 17b–i.

As illustrated in Fig. 16b, the regression performance of the FNN model reaches its peak (the red start) with $\overline{R^2} = 0.9945$, when the first nine selected descriptors are used as the input feature variables, Therefore, the optimal feature set contains nine morphological descriptors, which are: total fractional of percolating cells (D47), geometrical tortuosity (D6), specific surface area (D3), effective porosity (D2), succolarity (D27), characteristic length (D25), absolute porosity (D1), average coordination number (D11), and average pore size (D15). Only one

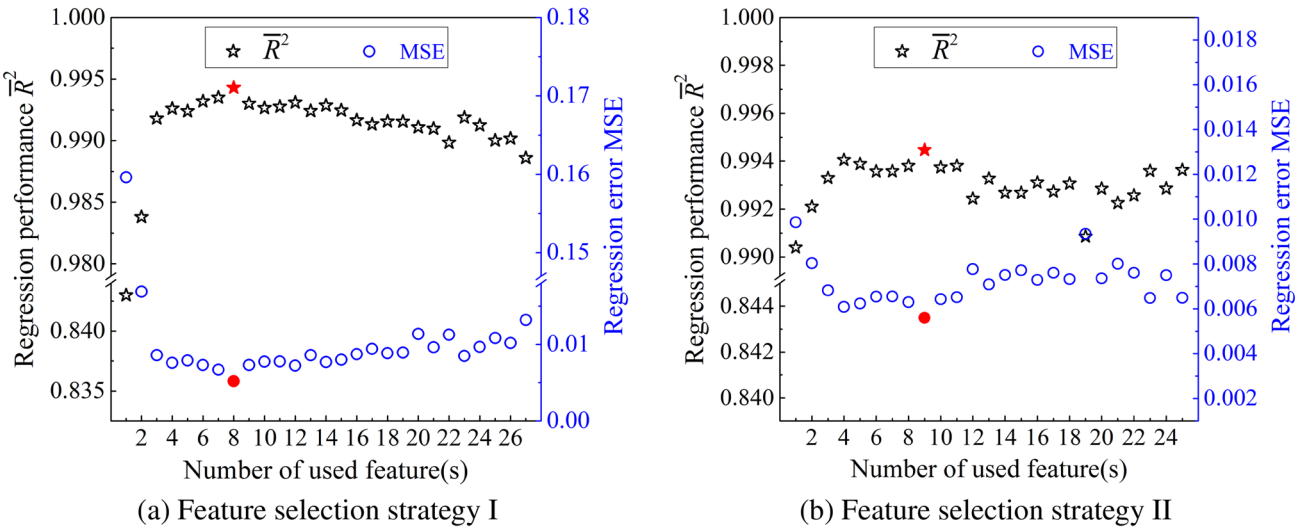


Fig. 16 The regression performance of FNN models varying with the number of selected descriptors

green descriptor is contained in the optimal feature set, and the remainders are blue descriptors.

Comparing the feature selection results in Sect. 3.4.3 and Sect. 3.4.4, there are five morphological descriptors in common, which are D6, D3, D2, D1, and D11. Besides, the newly selected descriptor D15 quantitatively characterizes the spatial size of pore channels allowing fluid percolation, which is conceptually similar to the formerly selected descriptor D10 (mean chord length) in Sect. 3.4.3. The remaining selected descriptors still contain D47 (total fractional of percolating cells), D27 (succolarity), and D25 (characteristic length), and they provide a new perspective to understand the microstructure–permeability relation. In essence, D47, D27, and D25 are quantitative indicators to characterize the percolation degree and long-distance connectivity of the pore networks that allow fluid flow to pass through.

4 Data-driven modeling

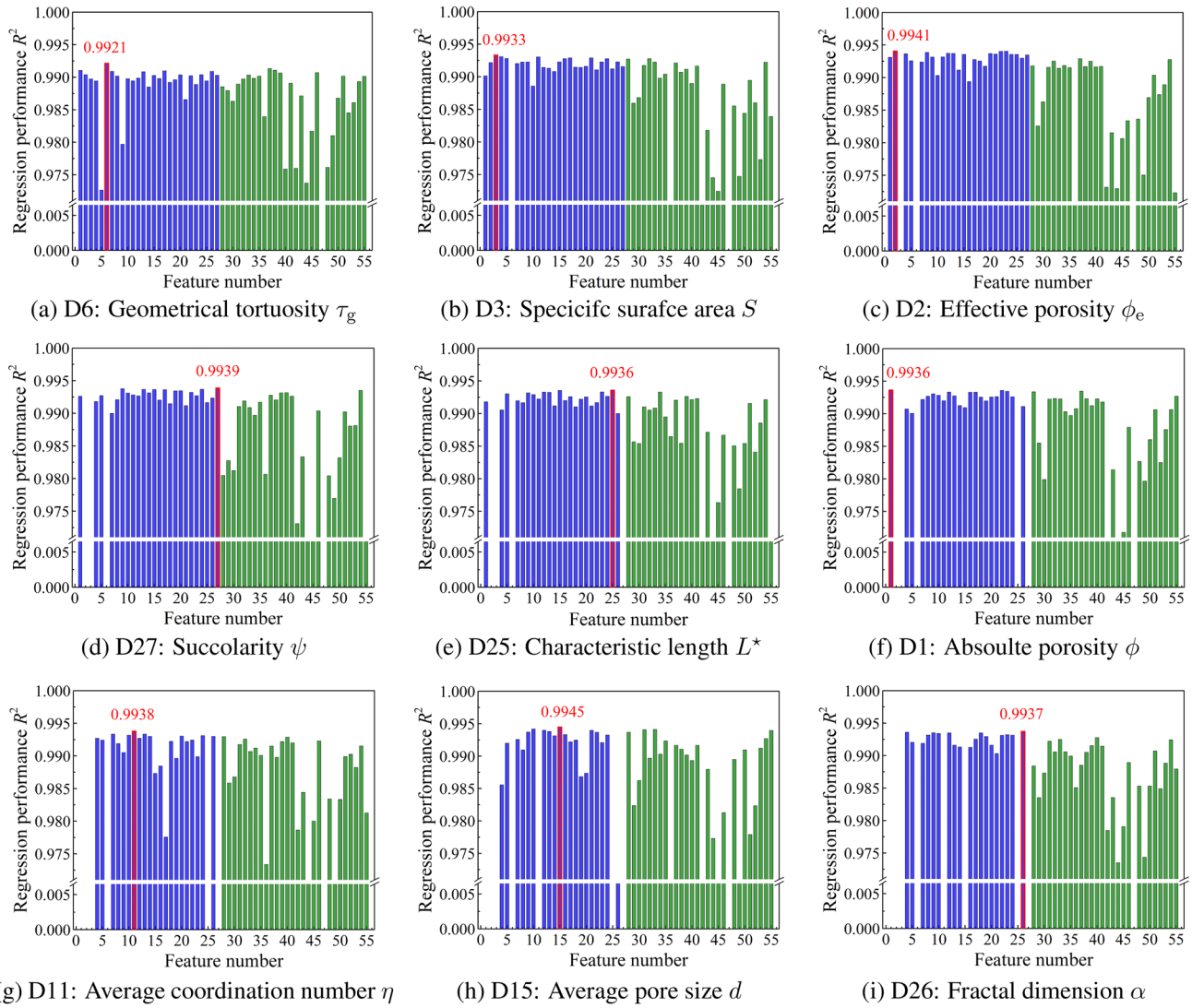
In this section, the feature selection results in Sect. 3.4.3 and Sect. 3.4.4 are separately applied to construct two surrogate models to approximate the microstructure–permeability mapping through data-driven modeling. As shown in Table 3, the morphological descriptors used to construct predictive models are clearly listed. The feed-forward neural networks used for function approximation are completely same to the one used for feature selection, and the hyper-parameters are summarized in Table 2.

4.1 Data-driven model I

As explained in Sect. 3.4.3, the optimal feature set containing eight blue descriptors has been obtained through the wrapper type feature selection, and they are listed in Table 3. These eight morphological descriptors are used as the feature variables to fit an FNN for permeability prediction. In fact, such a predictive model has already been obtained together with the feature selection result in Sect. 3.4.3. The critical issue is how well the microstructure–permeability relation is represented by the surrogate model, which should be further analyzed.

As shown in Fig. 18a, the loss function of the neural network is quickly minimized, and the best training state is determined by the validation performance. Once the neural network is properly trained, it can be used for permeability prediction, and the results are plotted in Fig. 18b. By overall comparison, the permeability prediction results agree well with the lattice Boltzmann simulation results (the targets), especially for large permeability values. However, a clear trend can be seen from Fig. 18b is that the prediction–target discrepancy increases as the permeability value declines.

To make a thorough analysis of the prediction results, the entire observational data are divided into two subsets by a permeability threshold equal to 1,000 millidarcy (md). As shown in Fig. 18b, the data points in the green ellipse correspond to high-permeable rock samples, and the remaining points in the orange ellipse represent low-permeable rock samples. Obviously, the fitted FNN model is of high accuracy for permeability evaluation of high-permeable rock samples, and the evaluation errors are summarized in



(a) D6: Geometrical tortuosity τ_g (b) D3: Specicifc surafce area S (c) D2: Effective porosity ϕ_e

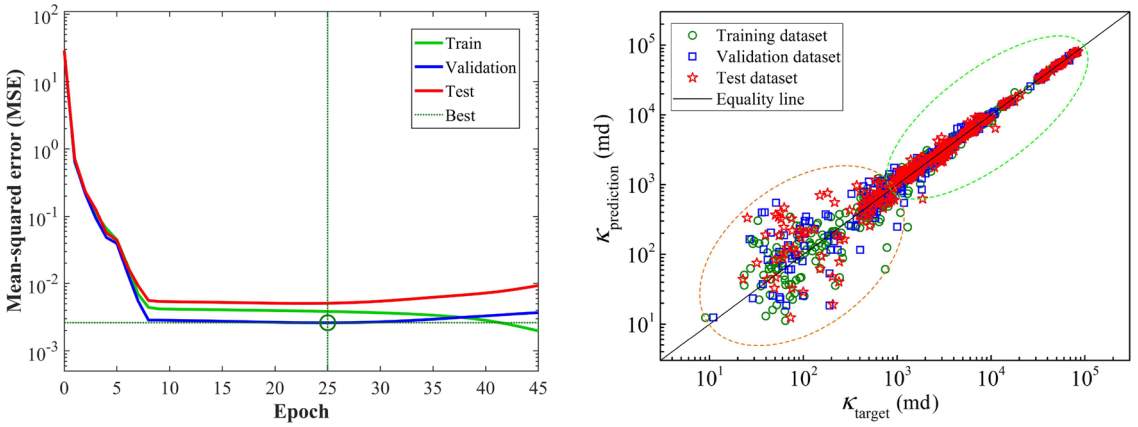
(d) D27: Succolarity ψ (e) D25: Characteristic length $L^\star$ (f) D1: Absoulte porosity ϕ

(g) D11: Average coordination number η (h) D15: Average pore size d (i) D26: Fractal dimension α

Fig. 17 The selected descriptor at each round via the wrapper type feature selection

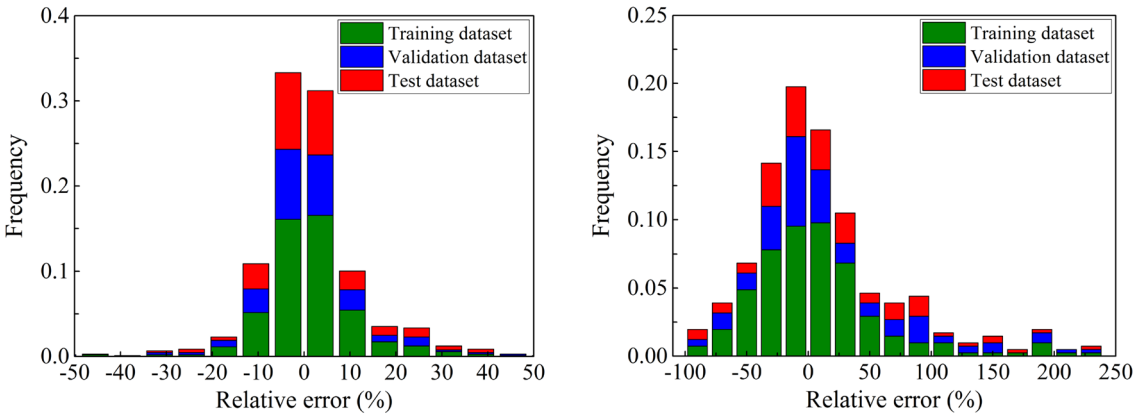
Table 3 Comparisons between different predictive models in terms of permeability evaluation accuracy

Predictive model	Involved morphological descriptors	High-permeable rocks ($\kappa \geq 1,000$ md)		Low-permeable rocks (10 md $< \kappa <$ 1,000 md)	
		Range of relative evaluation errors	Average error magnitude (%)	Range of relative evaluation errors	Average error magnitude (%)
Data-driven model I	D2, D3, D6, D26, D1, D11, D8 and D10	[−49.11%, 47.58%]	7.70	[−90.75%, 287.03%]	53.30
Data-driven model II	D47, D6, D3, D2, D27, D25, D1, D11 and D15	[−50.66%, 46.79%]	8.80	[−85.82%, 272.81%]	49.73
Kozeny–Carman relation	D2, D3 and D6	[−28.56%, 277.81%]	48.58	[−5.14%, 2393.81%]	334.58
Berg's relation	D2, D6, D8 and D15	[−66.04%, 93.66%]	32.06	[−79.10%, 1839.49%]	102.11



(a) Convergence of the loss function (b) Comparison between targets and predictions

Fig. 18 The data-driven FNN model I for permeability prediction



(a) High-permeable rock samples (b) Low-permeable rock samples

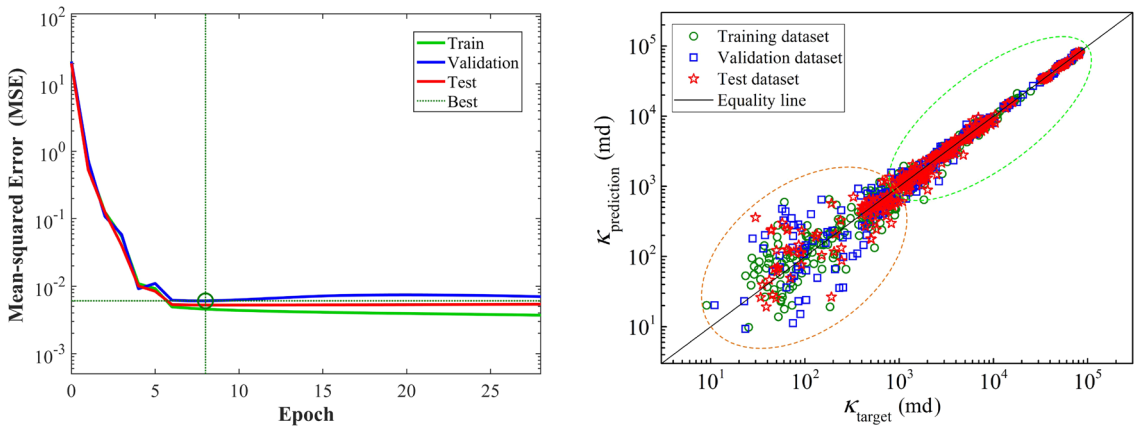
Fig. 19 Relative error distributions of the permeability values evaluated from the data-driven model I

Table 3. As illustrated in Fig. 19a, the relative prediction error varies from −49.11% to 47.58%, with the average error magnitude as low as 7.70%. Besides, for 85.46% of high-permeable rock samples, this predictive model is able to provide permeability values with very low relative errors within ±10%.

However, the fitted FNN model becomes less accurate, when it comes to low-permeable porous media samples. As illustrated in Fig. 19b, the relative prediction errors are distributed in a wider range from −90.75% to 287.03%, with the average error magnitude equal to 53.30%. This regression error is mainly caused by the inadequacy of microstructure characterization for low-permeable rock samples, instead of under-fitting, over-fitting, or other training problems. Compared to the popular PNM [7, 113], which usually provides permeability evaluations with relative errors around ±40% for low-permeable rocks, the accuracy of the fitted neural network model is acceptable. On the other hand, this machine learning-based surrogate model also possesses an excellent generalization performance to predict permeability spanning four orders of magnitude for natural reservoir rocks.

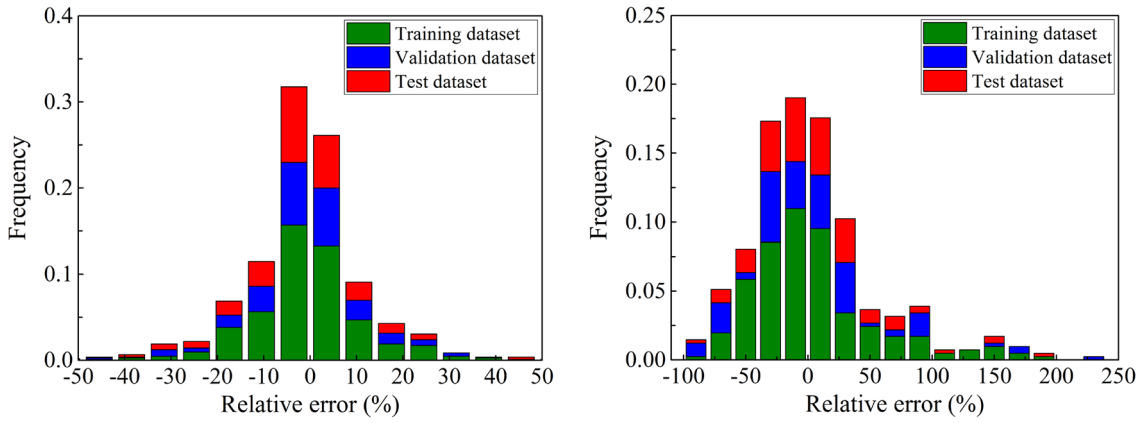
4.2 Data-driven model II

In this part, the optimal feature set obtained in Sect. 3.4.4 is used as predictor variables to fit another data-driven model for permeability evaluation. This feature set contains nine morphological descriptors, as listed in Table 3. An FNN model is trained by minimizing its loss function, and its best training state can be reflected by the validation performance, as illustrated in Fig. 20a. Once the training process is completed, the FNN model is able to predict permeability for new observations, and the results are plotted in Fig. 20b.



(a) Convergence of the loss function (b) Comparison between targets and predictions

Fig. 20 The data-driven FNN model II for permeability prediction



(a) High-permeable rock samples (b) Low-permeable rock sampples

Fig. 21 Relative error distributions of the permeability values evaluated from the data-driven model II

It seems that this regression model is comparable to data-driven model I in terms of prediction accuracy.

The relative prediction errors of data-driven model II are summarized in Table 3, and corresponding error distributions are plotted in Fig. 21. Compared to data-driven model I, data-driven model II possesses a slightly better prediction performance for low-permeable rock samples, but its prediction accuracy for high-permeable rock samples drops slightly. It may imply that blue descriptors are more accurate for microstructure characterization of high-permeable rocks, while green descriptors are more powerful to capture microstructural complexities of low-permeable rocks. Specifically, the relative prediction error varies from −50.66% to 46.79% for high-permeable rocks, with a average error magnitude equal to 8.80%. As to low-permeable rock samples, the range of relative evaluation error is from −85.82% to 272.81%, with the average error magnitude up to 49.73%. Although the prediction performances of data-driven model I and II are rather similar, the former (8 predictor variables) has much less predictor variables than the latter (83 predictor variables), and thus, data-driven model I can be considered to be a more cost-effective surrogate.

Generally, the pore network systems inside low-permeable rocks exhibit strong randomness, complexity, and heterogeneity, which makes it extremely difficult to achieve accurate and complete characterization of the internal microstructures. For high-permeable rock samples, the selected morphological descriptors can well represent the microstructural complexities for permeability evaluation. However, for low-permeable rock samples, more powerful morphological descriptors may be required to completely capture the microstructural characteristics related to permeability.

5 Comparisons

To examine the proposed data-driven framework in exploring the microstructure–permeability relation for porous media, the predictive models constructed through microstructural characterization and pore-scale simulation are compared with two popular empirical/analytical formulas in this section. Table 3 summarizes the performances of different predictive models in permeability evaluation. Generally, these two explicit formulas are significantly inferior to the data-driven models in terms of evaluation accuracy and generalization capacity.

5.1 Kozeny–Carman relation

The semi-empirical Kozeny–Carman relation [21, 25, 35] is one of the best-known formulas to estimate permeability, given by

$$\kappa = \frac{\phi^3}{cS^2\tau^2}, \tag{15}$$

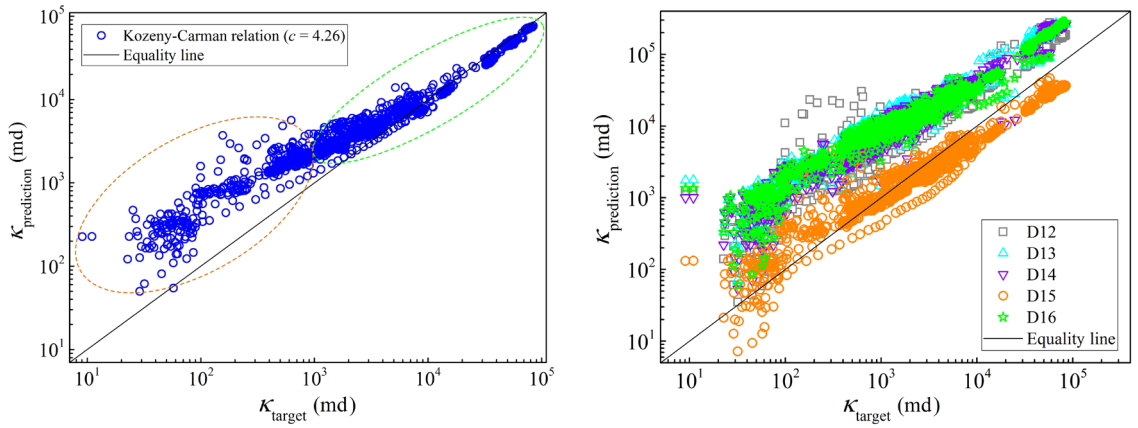
where ϕ, S, and τ are the porosity, specific surface area, and tortuosity of a porous medium, respectively; and c is a dimensionless coefficient called Kozeny's constant.

Kozeny's constant c is an unknown coefficient, and its value can significantly vary with the microstructural characteristics of porous media. Without a universal value, c is usually estimated by empirically fitting numerical or experimental data for specific types of porous media [115]. For beds packed with spherical particles, c is around 2.50 [57]. Due to microstructural complexities, c should be larger than 2.50 for natural porous rocks. Besides, the three predictor variables ϕ, S, and τ involved in Kozeny–Carman equation are also contained in the feature selection results in Sect. 3.4.3 and Sect. 3.4.4. Here, the selected descriptors D2 (effective porosity), D3 (specific surface area), and D6 (geometrical tortuosity) are substituted into Eq. (15) to take the place of ϕ, S, and τ, respectively, and then, c is determined to be 4.26 by fitting the permeability results evaluated from lattice Boltzmann simulations.

As shown in Fig. 22a, Kozeny–Carman relation is unable to provide accurate predictions for a wide range of permeability values. Roughly, its prediction accuracy is acceptable when the permeability value is larger than 1,000 md, as illustrated by the data points in the green ellipse in Fig. 22a. The prediction error varies from −28.56% to 277.81%, with the average error magnitude up to 48.58%. However, Kozeny–Carman relation becomes much less reliable when it comes to lower permeable samples (κ 1,000 md). Permeability values are systematically overestimated, as illustrated by the data points in the orange ellipse in Fig. 22a. The prediction error varies widely from −5.14% to 2393.70%, with the average error as high as 334.58%, which can be seen in Fig. 23b. In general, Kozeny–Carman relation possesses high level of uncertainty due to empirical selection of the adjustable coefficient c. Besides, the intrinsic microstructure–permeability mapping is also not fully represented by Kozeny–Carman relation, especially for low-permeable rocks, because only three simple morphological descriptors (namely, ϕ, S, and τ) are used, which is far from sufficient to capture the microstructure complexities of natural porous rocks.

5.2 Berg's relation

Inspired by Kozeny–Carman equation, Berg [10] derived a physical relation from the measurable microstructural



(a) Kozeny-Carman relation (b) Berg's relation

Fig. 22 Permeability predictions obtained from (**a**) Kozeny–Carman relation, and (**b**) Berg's relation (D12, D13, D14, D15, and D16 are the average pore size d results computed from different methods, and they are all used to replace L_h in Eq. (16) for permeability evaluation)

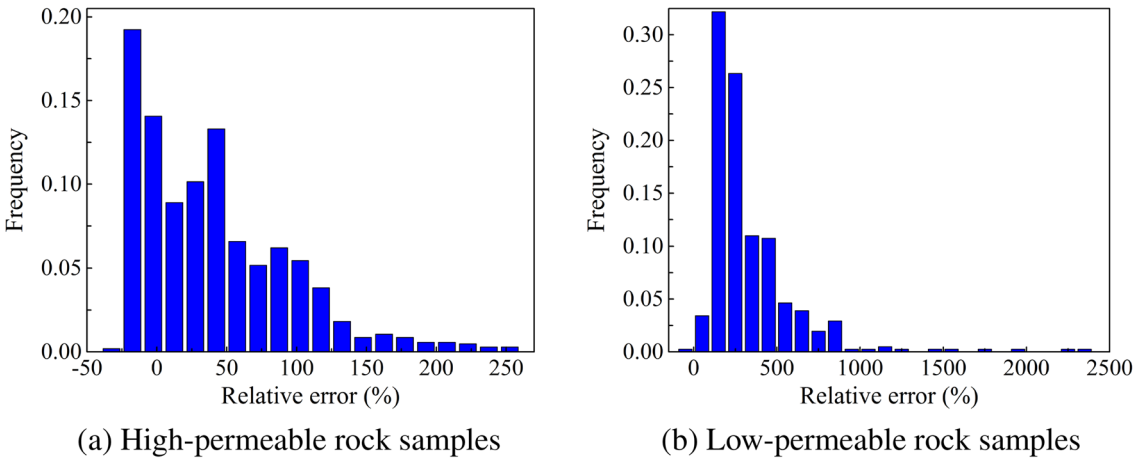


Fig. 23 Relative error distributions of the permeability values estimated from Kozeny–Carman relation

descriptors of porous media, without introducing any tuning parameters or free constants. Berg's relation reproduces Darcy's law for idealized pipe flow, but it is also used to evaluate permeability for natural porous rocks, which is mathematically expressed as follows:

$$\kappa = \phi_s \frac{L_h^2 \beta_s}{8\tau_s^2}, \tag{16}$$

where ϕ_s is the effective porosity to describe the fractional volume conducting flow, L_h denotes a characteristic length related to hydraulic pore radius, β_s is the constriction factor to represent the fluctuation in local hydraulic pore radii, and τ_s denotes the tortuosity to quantify the effective length of streamlines.

Here, the selected morphological descriptors D2 (effective porosity), D6 (geometrical tortuosity), and D8 (constriction factor) are substituted into Eq. (16) to replace ϕ_s, τ_s, and β_s respectively. Average pore size d (such as D12, D13, D14, D15, and D16) is used to approximate the characteristic length L_h, and permeability results are then estimated from Berg's relation, as shown in Fig. 22b. Five different methods are used to determine average pore sizes for the porous media samples used in this work, and the results from them are not consistent with each other, as illustrated in Fig. 5a. Obviously, the average pore size (D15) determined from the random point method is the most suitable estimation of L_h for permeability evaluation via Berg's relation, as can be seen from Fig. 22b. It should be emphasized that D15 is also one of the selected descriptors contained in the feature selection results in Sect. 3.4.4, which further substantiates the rationality of the feature selection results.

Here, the permeability results (the orange dots in Fig. 24) obtained by substituting D2, D6, D8, and D15 to into Eq. (16) are used to assess the performance of Berg's relation. As shown in Fig. 24, the relative error distributions of

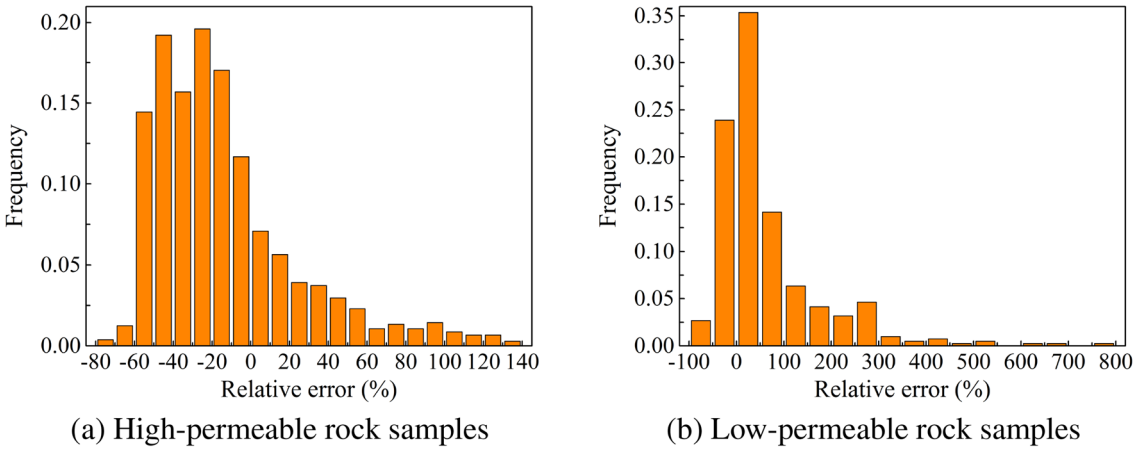


Fig. 24 Relative error distributions of the permeability values estimated from Berg's relation

the permeability values estimated from Berg's relation are plotted. For high-permeable rock samples, Berg's relation exhibits an acceptable accuracy, and the relative evaluation error varies from −66.04% to 93.66%, with the average error magnitude of 32.06%. However, the relative evaluation error increases significantly when it comes to low-permeable rock samples, as can be seen from Fig. 24b. The relative evaluation error varies from −79.10% to 1839.49%, with the average error magnitude up to 102.11%. Compared with Kozeny–Carman relation, Berg's model exhibits a better prediction performance, and its greatest success is the exclusion of empirical parameter by introducing constriction factor β_s. However, the inherent microstructure–permeability mapping is still not fully described by Berg's relation, especially for low-permeable porous rocks.

6 Discussion and conclusions

6.1 Discussion

As demonstrated in Sect. 3 and Sect. 4, it is an effective route to investigate the microstructure–property relationships through feature selection and data-driven modeling. Using the optimal feature set as the predictor variables for data-driven regression, a highly cost-effective model can be obtained with excellent prediction performance, and new insights into the microstructure–property linkage can also be gained from the feature selection results. However, achieving the above objectives is on the condition that the available feature pool can provide a comprehensive characterization of porous microstructures in an explicit expression, from which the optimal feature set that best represents the macroscopic physical property can be picked out through feature selection.

Generally, the data-driven surrogate models are of high accuracy and reliability to represent the microstructure–permeability mapping for high-permeable rocks. When it comes to low-permeable rocks, which usually possess more complicated internal pore network systems, neither the data-driven models nor explicit formulas can guarantee a high accuracy of permeability evaluation. The primary reason for such disparate performance is that the selected morphological descriptors capture the major microstructural characteristics that are important to permeability but also neglect some microstructural details, and such informatics loss can lead to increased uncertainty in permeability evolution when porous microstructures become more complicated.

Although there are many well-designed morphological descriptors (as listed in Table 1), extremely complicated microstructures could be beyond their capacity scope for accurate quantitative characterization. To cope with this inadequate characterization problem, effective descriptors should be specially developed from new perspectives, and microstructural details should also be preserved as much as possible by the new descriptors, which all put forward higher requests for quantitative microstructure analysis. Also, piecewise analysis can be adopted to establish microstructure–permeability mappings for different permeability ranges, because global predictive models are usually less accurate for low-permeable rock samples (as illustrated in Table 3) and the requirement of quantitative characterization also becomes higher as microstructural complexity increases.

Finally, the average computational costs of different predictive models to evaluate the permeability values of porous media samples used in this study are recorded in Table 4. It can be seen that the proposed data-driven models are able to provide instant predictions, which is 10^7 times faster that the lattice Boltzmann simulation of pore-scale fluid flow.

6.2 Conclusions

The main contribution of this work is to present a novel data-driven computational framework to fundamentally investigate the microstructure–property relationships of porous media through feature selection and data-driven regression. This framework can not only construct cost-effective surrogate models with high prediction accuracy and strong generalization capacity, but also provide new insights into the mechanisms of how microstructural characteristics determine microscopic behaviors.

This study especially focuses on the microstructure–permeability mapping of natural porous rocks. A large number of 3D digital microstructure samples with a wide porosity range are acquired from microscopy imaging and stochastic reconstruction. Pore-scale fluid flow passing through porous media is numerically simulated using high-fidelity lattice Boltzmann models, to provide reliable references of permeability values. A broad variety of morphological descriptors are collected from an extensive literature survey, and they compose the feature pool that quantitatively characterizes

Table 4 The average computational costs of different predictive models for evaluating permeability of porous media samples used in this study

Predictive model	Lattice Boltzmann simulation	Data-driven predictive models	Kozeny–Carman relation	Berg's relation
Computational time (s)	30860	1.4×10^{-2}	6.5×10^{-3}	8.0×10^{-3}

porous microstructures from global, local, geometrical, and topological perspectives. A performance-oriented feature selection is conducted to identify and pick out the microstructural characteristics that are significant to permeability. Based on the optimal feature sets, data-driven models are rapidly fitted to approximate the microstructure–permeability mapping, and these surrogate models can reliably predict permeability value spanning four orders of magnitudes, which are greatly superior to commonly used empirical/analytical formulas in terms of evaluation accuracy and generalization ability.

In addition to constructing cost-effective models, feature selection is also greatly beneficial to understanding the microstructure–permeability relation. By comparing the three categories of feature selection techniques (including filter, embedded, and wrapper methods), we found that the wrapper method is more applicable to exploring the microstructure–permeability linkage, because it is not only capable of identifying the joint importance of multiple features, but also effective for heterogeneous feature selection problems. According to the selected morphological descriptors, intrinsic permeability of porous media primarily depends on the microstructural characteristics in the following aspects: permeable pore volume, pore-solid interface, pore channel sinuosity, pore fractal dimension, pore coordination number, pore channel constriction, pore size, and percolation/connectivity degree. Besides, the proposed data-driven framework can be straightforwardly applied to analyze other physical properties (such as effective diffusivity, thermal conductivity, formation factor, and effective elastic moduli) of porous media by linking them to relevant microstructural informatics of importance.

Acknowledgements The authors would like to acknowledge the support of EPSRC grant: PURIFY (*EP/V*000756/1), Swansea University (FSE Impact Fund), Higher Education Funding Council for Wales (COVID-19 Higher Education Student Support Fund), and Great Britain China Centre (Chinese Students Award).

Declarations

Conflict of interest The authors declare that they have no conflict of interest in this paper.

References

1. Adler PM (1992) Porous media: geometry and transports. Butterworth-Heinemann, Boston
2. Agrawal A, Choudhary A (2019) Deep materials informatics: Applications of deep learning in materials science. MRS Commun 9(3):779–792
3. Andrä H, Combaret N, Dvorkin J, Glatt E, Han J, Kabel M, Keehm Y, Krzikalla F, Lee M, Madonna C et al (2013) Digital rock physics benchmarks-part ii: computing effective properties. Comput Geosci 50:33–43
4. Anovitz LM, Cole DR (2015) Characterization and analysis of porosity and pore structures. Rev Mineral Geochem 80(1):61–164
5. Armstrong RT, McClure JE, Robins V, Liu Z, Arns CH, Schlüter S, Berg S (2019) Porous media characterization using minkowski functionals: theories, applications and future directions. Transp Porous Media 130(1):305–335
6. Bache K, Lichman M (2013) UCI machine learning repository
7. Baychev TG, Jivkov AP, Rabbani A, Raeini AQ, Xiong Q, Lowe T, Withers PJ (2019) Reliability of algorithms interpreting topological and geometric properties of porous media for pore network modelling. Transp Porous Media 128(1):271–301
8. Bear J (2013) Dynamics of fluids in porous media. Courier Corporation
9. Bentz DP, Martys NS (1994) Hydraulic radius and transport in reconstructed model three-dimensional porous media. Transp Porous Media 17(3):221–238
10. Berg CF (2014) Permeability description by characteristic length, tortuosity, constriction and porosity. Transp Porous Media 103(3):381–400
11. Berryman JG, Blair SC (1987) Kozeny-carman relations and image processing methods for estimating darcy's constant. J Appl Phys 62(6):2221–2228
12. Bertei A, Nucci B, Nicolella C (2013) Effective transport properties in random packings of spheres and agglomerates
13. Bignonnet F (2020) Efficient fft-based upscaling of the permeability of porous media discretized on uniform grids with estimation of rve size. Comput Methods Appl Mech Eng 369:113237
14. Biswal B, Manwart C, Hilfer R (1998) Three-dimensional local porosity analysis of porous media. Phys A 255(3–4):221–241
15. Blair SC, Berge PA, Berryman JG (1996) Using two-point correlation functions to characterize microgeometry and estimate permeabilities of sandstones and porous glass. J Geophys Res Solid Earth 101(B9):20359–20375
16. Blunt MJ (2017) Multiphase flow in permeable media: A pore-scale perspective. Cambridge University Press, Cambridge
17. Blunt MJ, Bijeljic B, Dong H, Gharbi O, Iglauer S, Mostaghimi P, Paluszny A, Pentland C (2013) Pore-scale imaging and modelling. Adv Water Resour 51:197–216
18. Bostanabad R, Zhang Y, Li X, Kearney T, Brinson LC, Apley DW, Liu WK, Chen W (2018) Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques. Prog Mater Sci 95:1–41
19. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
20. Bultreys T (2016) Savonnières carbonate. http://www.digitalrocksportal.org/projects/72
21. Carman PC (1937) Fluid flow through granular beds. Trans Inst Chem Eng 50:150–166
22. Cecen A, Dai H, Yabansu YC, Kalidindi SR, Song L (2018) Material structure-property linkages using three-dimensional convolutional neural networks. Acta Mater 146:76–84
23. Cecen A, Wargo E, Hanna A, Turner D, Kalidindi S, Kumbur E (2012) 3-d microstructure analysis of fuel cell materials: spatial distributions of tortuosity, void size and diffusivity. J Electrochem Soc 159(3):B299
24. Chen H, Chen S, Matthaeus WH (1992) Recovery of the navier-stokes equations using a lattice-gas boltzmann method. Phys Rev A 45(8):R5339
25. Clennell MB (1997) Tortuosity: a guide through the maze. Geol Soc Lond Spec Publ 122(1):299–344
26. Coker DA, Torquato S, Dunsmuir JH (1996) Morphology and physical properties of fontainebleau sandstone via a tomographic analysis. J Geophys Res Solid Earth 101(B8):17497–17506

27. Cosenza P, Prêt D, Zamora M (2015) Effect of the local clay distribution on the effective electrical conductivity of clay rocks. J Geophys Res Solid Earth 120(1):145–168
28. Costa A (2006) Permeability-porosity relationship: a reexamination of the kozeny-carman equation based on a fractal pore-space geometry assumption. Geophys Res Lett 33(2):5
29. Cui S, Fu J, Cen S, Thomas HR, Li C (2021) The correlation between statistical descriptors of heterogeneous materials. Comput Methods Appl Mech Eng 384:113948
30. Delerue J, Perrier E, Yu Z, Velde B (1999) New algorithms in 3d image analysis and their application to the measurement of a spatialized pore size distribution in soils. Phys Chem Earth Part A 24(7):639–644
31. Dong H, Gao P, Ye G (2017) Characterization and comparison of capillary pore structures of digital cement pastes. Mater Struct 50(2):154
32. Erofeev A, Orlov D, Ryzhov A, Koroteev D (2019) Prediction of porosity and permeability alteration based on machine learning algorithms. Transp Porous Media 128(2):677–700
33. Fu J, Cui S, Cen S, Li C (2021) Statistical characterization and reconstruction of heterogeneous microstructures using deep neural network. Comput Methods Appl Mech Eng 373:113516
34. Fu J, Dong J, Wang Y, Ju Y, Owen DRJ, Li C (2020) Resolution effect: An error correction model for intrinsic permeability of porous media estimated from lattice boltzmann method. Transp Porous Media 132(3):627–656
35. Fu J, Thomas HR, Li C (2020) Tortuosity of porous media: Image analysis and physical simulation. Earth-Sci Rev 2:103439
36. Fu J, Wang M, Xiao D, Zhong S, Ge X, Ben E (2023) Hierarchical reconstruction of 3d well-connected porous media from 2d exemplars using statistics-informed neural network. Comput Methods Appl Mech Eng
37. Fu J, Xiao D, Fu R, Li C, Zhu C, Arcucci R, Navon IM (2023) Physics-data combined machine learning for parametric reduced-order modelling of nonlinear dynamical systems in small-data regimes. Comput Methods Appl Mech Eng 404:115771
38. Fu J, Xiao D, Li D, Thomas HR, Li C (2022) Stochastic reconstruction of 3d microstructures from 2d cross-sectional images using machine learning-based characterization. Comput Methods Appl Mech Eng 390:114532
39. Géraud Y (1994) Variations of connected porosity and inferred permeability in a thermally cracked granite. Geophys Res Lett 21(11):979–982
40. Goodarzi M, Dejaegher B, Heyden YV (2012) Feature selection methods in qsar studies. J AOAC Int 95(3):636–651
41. Guest JK, Prévost JH (2007) Design of maximum permeability material structures. Comput Methods Appl Mech Eng 196(4–6):1006–1017
42. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
43. Guyon I, Gunn S, Nikravesh M, Zadeh LA (2008) Feature extraction: foundations and applications, vol 207. Springer, Berlin
44. Herring A, Sheppard A, Turner M, Beeching L (2018) Multiphase flows in sandstones. http://www.digitalrocksportal.org/projects/135
45. Hilfer R (2002) Review on scale dependent characterization of the microstructure of porous media. Transp Porous Media 46(2–3):373–390
46. Holzer L, Iwanschitz B, Hocker T, Münch B, Prestat M, Wiedenmann D, Vogt U, Holtappels P, Sfeir J, Mai A et al (2011) Microstructure degradation of cermet anodes for solid oxide fuel cells: Quantification of nickel grain growth in dry and in humid atmospheres. J Power Sources 196(3):1279–1294
47. Holzer L, Wiedenmann D, Münch B, Keller L, Prestat M, Gasser P, Robertson I, Grobéty B (2013) The influence of constrictivity on the effective transport properties of porous layers in electrolysis and fuel cells. J Mater Sci 48(7):2934–2952
48. Hong J, Liu J (2020) Rapid estimation of permeability from digital rock using 3d convolutional neural network. Comput Geosci 24:1523–1539
49. Hormann K, Baranau V, Hlushkou D, Höltzel A, Tallarek U (2016) Topological analysis of non-granular, disordered porous media: determination of pore connectivity, pore coordination, and geometric tortuosity in physically reconstructed silica monoliths. New J Chem 40(5):4187–4199
50. Iassonov P, Gebrenegus T, Tuller M (2009) Segmentation of x-ray computed tomography images of porous materials: A crucial step for characterization and quantitative analysis of pore structures. Water Resour Res 45(9):6
51. ImageJ (2016) Website. https://imagej.net/Welcome
52. Ioannidis M, Kwiecien M, Chatzis I (1996) Statistical analysis of the porous microstructure as a method for estimating reservoir permeability. J Petrol Sci Eng 16(4):251–261
53. Jiao Y, Stillinger F, Torquato S (2009) A superior descriptor of random textures and its predictive capacity. Proc Natl Acad Sci 106(42):17634–17639
54. Jin G, Patzek T, Silin D (2004) Direct prediction of the absolute permeability of unconsolidated and consolidated reservoir rock. spe 90084. In (2003) SPE Annual Technical Conference and Exhibition (Houston. Texas, USA), SPE
55. Kamrava S, Tahmasebi P, Sahimi M (2020) Linking morphology of porous media to their macroscopic permeability by deep learning. Transp Porous Media 131(2):427–448
56. Karimpouli S, Tahmasebi P (2019) Image-based velocity estimation of rock using convolutional neural networks. Neural Netw 111:89–97
57. Kaviany M (2012) Principles of heat transfer in porous media. Springer Science & Business Media, Berlin
58. Knudby C, Carrera J (2005) On the relationship between indicators of geostatistical, flow and transport connectivity. Adv Water Resour 28(4):405–421
59. Kohanpur AH, Valocchi A, Crandall D (2019) Micro-ct images of a heterogeneous mt. simon sandstone sample. http://www.digitalrocksportal.org/projects/247
60. Kohavi R, John GH et al (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
61. Koponen A, Kataja M, Timonen J (1997) Permeability and effective porosity of porous media. Phys Rev E 56(3):3319
62. Krüger T, Kusumaatmaja H, Kuzmin A, Shardt O, Silva G, Viggen EM (2016) The lattice boltzmann method: principles and practice. Springer, Berlin
63. Kuhn M, Johnson K et al (2013) Applied predictive modeling, vol 26. Springer, Berlin
64. Kutay ME, Aydilek AH, Masad E (2006) Laboratory validation of lattice boltzmann method for modeling pore-scale flow in granular materials. Comput Geotech 33(8):381–395
65. Latief F, Biswal B, Fauzi U, Hilfer R (2010) Continuum reconstruction of the pore scale microstructure for fontainebleau sandstone. Phys A 389(8):1607–1618
66. Lehmann P, Berchtold M, Ahrenholz B, Tölke J, Kaestner A, Krafczyk M, Flühler H, Künsch H (2008) Impact of geometrical properties on permeability and fluid phase distribution in porous media. Adv Water Resour 31(9):1188–1204
67. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2017) Feature selection: a data perspective. ACM Comput Surv (CSUR) 50(6):1–45
68. Li X, Liu Z, Cui S, Luo C, Li C, Zhuang Z (2019) Predicting the effective mechanical property of heterogeneous materials by image based modeling and deep learning. Comput Methods Appl Mech Eng 347:735–753

69. Liang Z, Ioannidis M, Chatzis I (2000) Permeability and electrical conductivity of porous media from 3d stochastic replicas of the microstructure. Chem Eng Sci 55(22):5247–5262
70. Lindquist WB, Venkatarangan A, Dunsmuir J, Wong T-F (2000) Pore and throat size distributions measured from synchrotron x-ray tomographic images of fontainebleau sandstones. J Geophys Res Solid Earth 105(B9):21509–21527
71. Liu J, Pereira GG, Liu Q, Regenauer-Lieb K (2016) Computational challenges in the analyses of petrophysics using microtomography and upscaling: a review. Comput Geosci 89:107–117
72. Loh W-Y (2002) Regression tress with unbiased variable selection and interaction detection. Stat Sin 2:361–386
73. Luhmann AJ, Tutolo BM, Bagley BC, Mildner DF, Seyfried WE Jr, Saar MO (2017) Permeability, porosity, and mineral surface area changes in basalt cores induced by reactive transport of co 2-rich brine. Water Resour Res 53(3):1908–1927
74. Łydżba D, Różański A, Sevostianov I, Stefaniuk D (2021) A new methodology for evaluation of thermal or electrical conductivity of the skeleton of a porous material. Int J Eng Sci 158:103397
75. Moctezuma-Berthier A, Vizika O, Adler PM (2002) Macroscopic conductivity of vugular porous media. Transp Porous Media 49(3):313–332
76. Moré JJ (1978) The levenberg-marquardt algorithm: implementation and theory. Numerical analysis. Springer, Berlin, pp 105–116
77. Muche L, Stoyan D (1992) Contact and chord length distributions of the poisson voronoi tessellation. J Appl Probab 2:467–471
78. Münch B, Holzer L (2008) Contradicting geometrical concepts in pore size analysis attained with electron microscopy and mercury intrusion. J Am Ceram Soc 91(12):4059–4067
79. N'Diaye M, Degeratu C, Bouler J-M, Chappard D (2013) Biomaterial porosity determined by fractal dimensions, succolarity and lacunarity on microcomputed tomographic images. Mater Sci Eng, C 33(4):2025–2030
80. Nordlund M, Penha DJL, Stolz S, Kuczaj A, Winkelmann C, Geurts BJ (2013) A new analytical model for the permeability of anisotropic structured porous media. Int J Eng Sci 68:38–60
81. Okabe H, Blunt MJ (2004) Prediction of permeability for porous media reconstructed using multiple-point statistics. Phys Rev E 70(6):066135
82. Paterson M (1983) The equivalent channel model for permeability and resistivity in fluid-saturated rock-a re-appraisal. Mech Mater 2(4):345–352
83. Pia G, Sanna U (2014) An intermingled fractal units model and method to predict permeability in porous rock. Int J Eng Sci 75:31–39
84. Quintanilla J, Torquato S (1997) Local volume fraction fluctuations in random media. J Chem Phys 106(7):2741–2751
85. Rabbani A, Babaei M, Shams R, Da Wang Y, Chung T (2020) Deepore: a deep learning workflow for rapid and comprehensive characterization of porous materials. Adv Water Resour 2:103787
86. Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of relieff and rrelieff. Mach Learn 53(1–2):23–69
87. Röding M, Ma Z, Torquato S (2020) Predicting permeability via statistical learning on higher-order microstructural information. Sci Rep 10(1):1–17
88. Rubinstein J, Torquato S (1988) Diffusion-controlled reactions: Mathematical formulation, variational principles, and rigorous bounds. J Chem Phys 88(10):6372–6380
89. Rubinstein J, Torquato S (1989) Flow in random porous media: mathematical formulation, variational principles, and rigorous bounds. J Fluid Mech 206:25–46
90. Santos JE, Xu D, Jo H, Landry CJ, Prodanović M, Pyrcz MJ (2020) Poreflow-net: A 3d convolutional neural network to predict fluid flow through porous media. Adv Water Resour 138:103539
91. Saxena N, Hofmann R, Alpak FO, Berg S, Dietderich J, Agarwal U, Tandon K, Hunter S, Freeman J, Wilson OB (2017) References and benchmarks for pore-scale flow simulated using micro-ct images of porous media and digital rocks. Adv Water Resour 109:211–235
92. Scanziani A, Singh K, Blunt M (2018) Water-wet three-phase flow micro-ct tomograms. http://www.digitalrocksportal.org/projects/167
93. Schlüter S, Sheppard A, Brown K, Wildenschild D (2014) Image processing of multiphase images obtained via x-ray microtomography: a review. Water Resour Res 50(4):3615–3639
94. Schulz E, Speekenbrink M, Krause A (2018) A tutorial on gaussian process regression: modelling, exploring, and exploiting functions. J Math Psychol 85:1–16
95. Sevostianova E, Leinauer B, Sevostianov I (2010) Quantitative characterization of the microstructure of a porous material in the context of tortuosity. Int J Eng Sci 48(12):1693–1701
96. Srisutthiyakorn N (2016) Deep-learning methods for predicting permeability from 2d/3d binary-segmented images. In: SEG technical program expanded abstracts 2016, pages 3042–3046. Society of Exploration Geophysicists
97. Sudakov O, Burnaev E, Koroteev D (2019) Driving digital rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks. Comput Geosci 127:91–98
98. Sukop MC, Huang H, Alvarez PF, Variano EA, Cunningham KJ (2013) Evaluation of permeability and non-darcy flow in vuggy macroporous limestone aquifer samples with lattice boltzmann methods. Water Resour Res 49(1):216–230
99. Tembely M, AlSumaiti AM, Alameri W (2020) A deep learning perspective on predicting permeability in porous media from network modeling to direct simulation. Comput Geosci 24:1541–1556
100. Tian J, Qi C, Sun Y, Yaseen ZM (2020) Surrogate permeability modelling of low-permeable rocks using convolutional neural networks. Comput Methods Appl Mech Eng 366:103–113
101. Tian J, Qi C, Sun Y, Yaseen ZM, Pham BT (2020) Permeability prediction of porous media using a combination of computational fluid dynamics and hybrid machine learning methods. Eng Comput 2:1–17
102. Torquato S (2002) Statistical description of microstructures. Annu Rev Mater Res 32(1):77–111
103. Torquato S (2013) Random heterogeneous materials: microstructure and macroscopic properties, vol 16. Springer Science & Business Media, Berlin
104. van der Linden JH, Narsilio GA, Tordesillas A (2016) Machine learning framework for analysis of transport through complex networks in porous, granular media: a focus on permeability. Phys Rev E 94(2):022904
105. Vogel H-J, Weller U, Schlüter S (2010) Quantification of soil structure based on minkowski functions. Comput Geosci 36(10):1236–1245
106. Wang J, Li Z, Yan S, Yu X, Ma Y, Ma L (2019) Modifying the microstructure of algae-based active carbon and modelling supercapacitors using artificial neural networks. RSC Adv 9(26):14797–14808
107. Wang M, Pan N (2008) Modeling and prediction of the effective thermal conductivity of random open-cell porous foams. Int J Heat Mass Transf 51(5–6):1325–1331
108. Wei H, Zhao S, Rong Q, Bao H (2018) Predicting the effective thermal conductivities of composite materials and porous media by machine learning methods. Int J Heat Mass Transf 127:908–916

109. Wu H, Fang W-Z, Kang Q, Tao W-Q, Qiao R (2019) Predicting effective diffusivity of porous media from images by deep learning. Sci Rep 9(1):1–12
110. Wu J, Yin X, Xiao H (2018) Seeing permeability from images: fast prediction with convolutional neural networks. Sci Bull 63(18):1215–1222
111. Xia M, Fu J, Feng Y, Gong F, Jin Y (2023) A particle-resolved heat-particle-fluid coupling model by dem-imb-lbm. J Rock Mech Geotech Eng 2:2
112. Xia Y, Cai J, Perfect E, Wei W, Zhang Q, Meng Q (2019) Fractal dimension, lacunarity and succolarity analyses on ct images of reservoir rocks for permeability prediction. J Hydrol 579:124198
113. Xiong Q, Baychev TG, Jivkov AP (2016) Review of pore network modelling of porous media: experimental characterisations, network constructions and applications to reactive transport. J Contam Hydrol 192:101–117
114. Xu H, Dikin DA, Burkhart C, Chen W (2014) Descriptor-based methodology for statistical characterization and 3d reconstruction of microstructural materials. Comput Mater Sci 85:206–216
115. Xu P, Yu B (2008) Developing a new form of permeability and Kozeny–Carman constant for homogeneous porous media by means of fractal geometry. Adv Water Resour 31(1):74–81
116. Yang W, Wang K, Zuo W (2012) Neighborhood component feature selection for high-dimensional data. JCP 7(1):161–168
117. Yegnanarayana B (2009) Artificial neural networks. PHI Learning Pvt Ltd., Delhi
118. Yu B, Cheng P (2002) A fractal permeability model for bi-dispersed porous media. Int J Heat Mass Transf 45(14):2983–2993
119. Zeng Z, Fu J, Feng Y, Wang M (2023) Revisiting the empirical particle-fluid coupling model used in dem-cfd by high-resolution dem-lbm-imb simulations: a 2d perspective. Int J Numer Anal Methods Geomech 2:2