Petroleum Science 23 (2026) 5819–5835

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

Classification of shale oil micro-occurrence types and flow behavior: Combining QSGSM, pore network, machine learning, and LBM

Yi-Fan Yin [a,c], Zhi-Xue Sun [b,c,d,*], Jing Wang [e], Kai-Jun Tong [f], Yong-Fei Yang [b,c,d]

[a] Department of Chemical Engineering, The University of Manchester, Manchester, M13 9PL, UK
[b] State Key Laboratory of Deep Oil and Gas, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[c] School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[d] Key Laboratory of Unconventional Oil and Gas Development (China University of Petroleum (East China)), Ministry of Education, Qingdao, 266580, Shandong, China
[e] PipeChina Engineering Technology Innovation Co., Ltd., Tianjin, 300450, China
[f] CNOOC International Limited, Beijing, 100028, China

ARTICLE INFO

Article history:
Received 10 February 2025
Received in revised form
2 June 2026
Accepted 3 June 2026
Available online 8 June 2026

Edited by Yan-Hua Sun

Keywords:
Shale oil
Digital core
Quartet structure generation set
Lattice Boltzmann method
Machine learning

ABSTRACT

Shale oil occurs in varying forms within porous media, with diverse micro-occurrences exhibiting intricate topological characteristics within the pores. The impact of shale oil presence on fluid flow patterns in these complex environments remains uncertain. This study investigates how different shale oil micro-occurrence types and saturation levels influence pore network characteristics and fluid flow behavior in shale reservoirs. Firstly, the shale digital core is established using the focused ion beam scanning electron microscope (FIB-SEM) method, and the pore skeleton is segmented via the watershed method. Utilizing the quartet structure generation set method, we constructed 12 digital cores representing floating, cementing, and coating shale oil micro-occurrence types at different oil saturations based on the real digital core model. The pore network analysis, conducted using the maximum ball method, revealed that increasing oil saturation generally leads to a reduction in pore radius and an increase in pore quantity due to pore subdivision, with the cementing type showing the most significant effect. Additionally, machine learning techniques, including random forest, support vector machine (SVM), and K-nearest neighbors (KNN), were employed to classify micro-occurrence types of shale oil based on topological features such as volume ratio, cross-sectional area ratio, three-dimensional shape factor, and Euler number. The support vector machine (SVM) method achieved the highest predictive accuracy (93%), outperforming the random forest (91%) and KNN (75%) methods, indicating its effectiveness in discerning subtle differences between micro-occurrence types. Lattice Boltzmann simulations were employed to study the impact of shale oil micro-occurrence type and saturation on fluid flow properties within the pore networks. Results showed that coating shale oil exhibited the highest permeability, followed by floating, whereas cementing shale oil caused a notable reduction in permeability due to pore blockage. This study enhances the understanding of the intricate relationship between shale oil geometric micro-occurrence patterns, pore structure, and fluid flow behavior within a digital-core framework, providing valuable insights into pore-scale fluid dynamics in shale reservoirs.

1. Introduction

Unconventional oil and gas resources are playing an increasingly dominant role in the global petroleum industry, surpassing conventional resources, and have become a new focus of oil and gas exploration and development worldwide (Wang et al., 2016; Xu et al., 2022; Mcmahon et al., 2024; Lin et al., 2022b, 2025). Compared to conventional oil and gas reservoirs, shale reservoirs exhibit complex, multi-scale pore structures with diverse pore types (Lei et al., 2021). From a microscopic perspective, the pore

* Corresponding author.
E-mail address: upcszx@upc.edu.cn (Z.-X. Sun).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2026.06.007

structure of shale reservoirs includes various types of pores at the nanometer and micrometer scales, as well as fractures, displaying diverse morphologies and spatial distributions (Li et al., 2020; Zhang et al., 2022b).

The state of shale oil occurrence is highly complex, encompassing adsorbed and free states, which significantly impact the permeability characteristics of shale reservoirs and the release of oil and gas (Cao et al., 2021; Li et al., 2022). Traditional methods for characterizing shale pore structures include gas adsorption (Rowsell et al., 2005) and mercury intrusion porosimetry (MIP) (Abell et al., 1999). Gas adsorption effectively measures specific surface area, porosity, and pore size distribution, making it suitable for mesoporous and microporous analysis. However, it provides only statistical pore size data without 3D structural details. MIP is a fast and versatile technique applicable to various materials, capable of measuring small pores and total pore volume. Nevertheless, it cannot capture the geometric complexity or image-based distribution of pores, and it fails to detect isolated pores.

With the advancement of experimental techniques and further research, various methods have been developed for characterizing rock pore structures, including scanning electron microscopy (SEM), X-ray microtomography (micro-CT), and small-angle scattering techniques such as small-angle X-ray scattering (SAXS) and small-angle neutron scattering (SANS) (Neil et al., 2020; Lin et al., 2022a, 2023). These techniques collectively offer comprehensive insights into the pore structures of rocks, each contributing unique and complementary information. Therefore, digital core has gradually become an irreplaceable tool in rock analysis (Wang et al., 2021; Sun et al., 2023; Yin et al., 2023). First, it can scan limited rock samples without causing any damage, which is especially important when dealing with rare samples that are difficult to obtain or replicate (Ponomarev et al., 2022). Second, digital cores can delve into the rock's micro structural characteristics, pore connectivity, permeability, and other key features in great detail (Hu et al., 2024). These insights are often difficult to obtain through traditional laboratory-based techniques, which makes digital cores indispensable in improving our understanding of rock properties and fluid flow behavior. Among these methods, the most widely used in characterizing rock pore structures is focused ion beam-scanning electron microscopy (FIB-SEM) (Goral et al., 2020; Garum et al., 2020). This technique combines focused ion beam technology with scanning electron microscopy imaging, enabling both micro-nanometer-scale sectioning and high-resolution imaging of rock samples. FIB-SEM offers exceptionally high spatial resolution, typically reaching the nanometer scale, which surpasses the resolution of CT. This enables FIB-SEM to observe the microstructure of shale pores in detail, including small fractures, pores, mineral particles, and the specific morphology of the pore network.

Once core images are acquired, segmentation is required to distinguish the rock matrix from pore spaces for accurate analysis of porosity, permeability, and microstructure. This process often employs specialized image processing algorithms such as thresholding, watershed, or deep learning-based methods. The watershed algorithm, based on image gradients, automatically detects region boundaries by computing gradient values, making it well-suited for segmenting complex and fine structures (Bleau and Leon, 2000). Compared to threshold method, which struggles with similar grayscale values between pores and the matrix, the watershed method can effectively capture subtle features like small fractures, pores, and mineral grains in high-resolution images. While deep learning offers high accuracy, it demands extensive labeled data and significant computational resources. In contrast, the watershed algorithm strikes a balance between accuracy and efficiency, making it a robust choice for segmenting rock matrix and pore structures. Its effectiveness has been demonstrated in multiple recent studies (Silvoster et al., 2021; Guo et al., 2022; Fotos et al., 2023).

The occurrence states in shale, as well as the pore sizes in which they occur can influence the mobility and fluid flow behavior of shale oil. A previous study has shown that oil in pores exists in dissolved, adsorbed, and free states (Zhang et al., 2022a). The proportion of dissolved shale oil is typically low, as shale oil primarily comprises large hydrocarbons with low water solubility. Due to the strong solid–liquid interactions, a significant amount of oil in the shale pore spaces is more likely to exist in an adsorbed state rather than a dissolved state (Cui and Cheng, 2017). Therefore, in shale pores, there are mainly adsorbed and free states of shale oil. The shale oil is “solid-like” adsorbed on the surface of organic matter and mineral particles with poor mobility. The free shale oil has less adsorption by kerogen and mineral particles, making it relatively more mobile (Yang et al., 2020). NMR and micro-CT studies have identified six microscopic occurrence forms: film-like, clustered, throated, emulsified, granular, and isolated, with emulsified and film-like forms being dominant (about 70%) (Wang et al., 2015). Additional analysis using EDS and ESEM confirmed that tight oil mainly exists as films and droplets within pores (Gong et al., 2015). Previous studies have primarily classified shale oil's micro-occurrence types based on physical properties, but have largely overlooked its geometric and topological characteristics within pore spaces. Compared to physical-state-based classifications (e.g., adsorbed vs. free), geometric and topological descriptors such as volume ratio, cross-sectional area ratio, shape factor, and Euler number offer three distinct advantages: they are quantitatively computable from digital core images, reproducible across different samples, and directly applicable as input features for machine learning classifiers. Moreover, it should be noted that current evidence for shale oil micro-occurrence states at the nanoscale is largely indirect, relying on theoretical models, numerical simulations, or post-processing of NMR/micro-CT signals, rather than direct in-situ observation. This further underscores the need for quantitative, image-based classification approaches. This limitation hinders the development of precise classification criteria, as subtle topological differences are difficult to quantify using simple mathematical expressions. Consequently, it remains challenging to fully understand the spatial distribution and flow behavior of shale oil at the microscale. To overcome these challenges, the rapid development of machine learning, particularly deep learning and neural networks, provides new tools for extracting complex, nonlinear features from data (Tariq et al., 2021; Sircar et al., 2021). These methods have been widely applied in the oil and gas sector for reservoir prediction, exploration, and production forecasting. For example, the application of a deep neural network (DNN) combined with particle swarm optimization (PSO) can effectively improve production forecasts and optimize fracturing parameters (Lu et al., 2022). However, most of these applications focus on macroscale analysis, and using machine learning to classify microscale shale oil occurrence states remains underexplored. Nonetheless, machine learning excels at recognizing patterns in large datasets (Lawal et al., 2024). For instance, convolutional neural networks (CNNs) can automatically extract features from images for classification; support vector machines (SVM) handle small datasets effectively; and ensemble methods like random forests and gradient boosting machines enhance accuracy and robustness. Recognizing shale oil microscopic occurrence types involves processing a large amount of high-dimensional data and differentiating subtle differences between types. Therefore, it necessitates the use of machine learning methods that can effectively handle complex data and capture subtle feature differences. Common machine learning methods include:

(1) Supervised learning: Techniques such as logistic regression, support vector machines , decision trees, and random forests are suitable for datasets with known class labels. These methods learn the relationship between data features and classes to predict the category of unknown samples.
(2) Unsupervised learning: Methods like clustering analysis are suitable for datasets without class labels. These methods automatically group data based on features, revealing hidden data structures.

Therefore, by extracting descriptors of shale oil's topological features within pores and training machine learning models, it is feasible to achieve automated, high-precision classification of its microscopic occurrence states.

The shale pores exhibit a complex multiscale structure, and the pore geometry, throat size, and connectivity play a crucial role in the fluid flow behavior within the rock. Therefore, investigating the impact of different microscopic occurrence types of shale oil on flow characteristics is also of significant research value. There are several methods available for studying the flow of shale oil, including pore network models (Blunt, 2001), lattice Boltzmann methods (Kuznik et al., 2010), and molecular dynamics simulations (Alipour et al., 2019). Pore network models are widely used for characterizing the flow behavior in porous media, including shale. These models simplify the complex pore structure into a network of interconnected nodes and channels, allowing for efficient computation of fluid flow. However, they often rely on assumptions such as the homogeneity and isotropy of the pore system, which may limit their accuracy in capturing the true flow behavior in shale formations. In addition, the proportions of shale oil occurrence states (dissolved, adsorbed, and free) in shale, as well as the pore sizes in which they occur, also influence the mobility and fluid flow behavior of shale oil.

By summarizing previous research results (Yang et al., 2019; Zhang et al., 2022c; Yin and Sun, 2024), based on the topological characteristics and geometric relationships between shale oil and pores, the micro-occurrence types of shale oil in the pores can be classified into three categories:

(1) floating shale oil droplets in the pores, referred to as floating type;
(2) thin film oil layers coating the surfaces of matrix grains, known as coating type;
(3) clusters of shale oil blocking the pore throats, known as cementing type.

Among them, tight oil of the floating type mainly exists in larger intergranular pores and corroded pores; coating appears in thin film shape, often on mineral grain surfaces; cementing predominantly exists in clusters at pore connections and throats, blocking flow channels in a cemented shape. However, in real core samples from formations, the complex pore structure results in various complex micro-occurrence types, making the study of the topological characteristics and fluid flow laws of different shale oil micro scale occurrence types particularly challenging. Therefore, by establishing shale oil digital core models with specific micro scale occurrence types, topological structure analysis and micro scale flow simulations can be conducted to reveal the topological characteristics of different micro scale occurrence types of shale oil and their impact on fluid flow.

It should be emphasized that the micro-occurrence classification proposed in this study is a geometry/topology-based classification within a digital-core framework, aimed at distinguishing the spatial occupation patterns of oil within pore spaces, rather than a complete physical classification of in-situ shale oil occurrence states (e.g., adsorbed, dissolved, and free states).

In this study, a high-resolution digital core model of shale is established using FIB-SEM, and the watershed method is used to segment the pore skeleton. Quartet structure generation set method (QSGSM) is used to construct digital core models containing specific micro-occurrence of shale oil. Pore network models of these cores are extracted by the maximum ball method, and the influence of different micro-occurrence types of shale oil on the inner pore topological characteristics is analyzed. Furthermore, machine learning techniques, including random forest, SVM, and KNN, are applied to classify micro-occurrence types of shale oil based on their topological features such as volume ratio, cross-sectional area ratio, three-dimensional shape factor, and Euler number. Additionally, for the aforementioned core, the lattice Boltzmann method is utilized for flow simulation to reveal the impact of different micro-occurrence types and saturations of shale oil on fluid flow laws.

2. Digital core construction using FIB-SEM experiment

The experiments were carried out at the China University of Petroleum (East China) using Leica EM RES 102 and Zeiss Crossbeam 540 instruments. A total of 400 images were obtained by the experimental scan, which have a pixel resolution of 400 $\times$ 400 pixels and a physical size of 1000 nm $\times$ 1000 nm.

Core samples require drying prior to experimentation. The cores were heated at 50 °C for 48 h to remove residual water and oil. This prevents steam formation during scanning, which would degrade image quality. Core mass was monitored periodically during heating; weight stabilization confirmed complete removal of pore fluids.

The detailed sample preparation procedure, including mechanical polishing and two-step Pt deposition, is provided in Appendix A.

Then, the rock samples are put into the FIB-SEM system for imaging, which comprises two beams: an electron beam and an ion beam. The focused gallium ion beam, operating at a voltage of 30 kV, is utilized to continuously mill the samples by sputtering away the shale material, as shown in Fig. 1(a). Simultaneously, the SEM captures in-situ images of the newly milled shale surface, with a resolution of 2.5 nm at an accelerating voltage of 2 kV and a working distance of 4 mm. This milling and imaging procedure generates a sequence of SEM images, which collectively form a 3D image dataset depicting the internal microstructure of the shale. The image we obtained directly through the experiment is a grayscale image with pixel values between 0 and 255.

Following the completion of the FIB-SEM experiment, the watershed algorithm (Watson et al., 1992) based on topological theory, was employed for image segmentation, as shown in Fig. 1(b). The basic principle of the watershed method involves generating a gradient map based on the grayscale information in the image. This gradient map is used to calculate the height values of each pixel in the image. Pixels with higher grayscale values are considered high areas, while pixels with lower grayscale values are considered as low areas. Then, a series of basins are constructed at the boundaries between high and low areas. These basins are filled during the water flow simulation, forming the segmentation boundaries. After segmentation by the watershed method, the original core image was divided into rock skeletons and pores. To facilitate the subsequent calculation, we set the part of the rock skeleton as 1 and the pore part as 0, so that the binary core image after image processing can be obtained, as shown in Fig. 1(c).

However, it must be acknowledged that the watershed algorithm is fundamentally an image-based segmentation tool. It

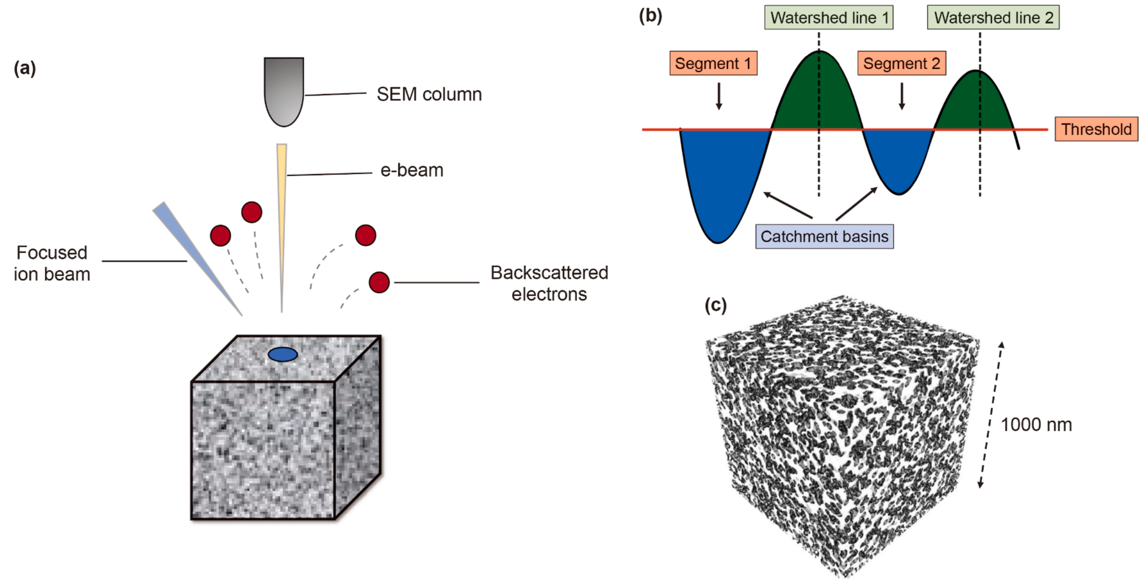


Fig. 1. **(a)** Schematic diagram of the FIB-SEM experimental principle. **(b)** Image segmentation principle of the watershed method. **(c)** Schematic diagram of the digital core model obtained from the experiment.

cannot capture pore morphology and topology changes induced by temperature–pressure evolution, desorption, or strong polar fluid invasion. Therefore, the segmented digital core represents only the static geometric framework of the imaged sample under laboratory conditions, and does not by itself reflect dynamic pore-structure evolution under reservoir conditions.

3. Generation of digital core models for shale oil and extraction of topological structure parameters

While the FIB-SEM digital core obtained in the previous section accurately represents the rock framework, it fails to capture the in-situ distribution of shale oil. The evaporation of pore fluids is unavoidable during sample preparation (drying at 50 °C for 48 h under atmospheric conditions), and the high-vacuum environment of FIB-SEM imaging further precludes fluid preservation. Furthermore, even if fluid retention were possible, the laboratory setting differs substantially from the subsurface reservoir environment regarding temperature, pressure, and stress states. Consequently, the FIB-SEM digital core should be regarded strictly as a realistic geometric template of the rock matrix. Within this framework, the specific micro-occurrence patterns of shale oil must be numerically reconstructed as synthetic models, thus necessitating the adoption of the QSGSM described below.

The QSGSM utilizes the true rock framework information obtained in Section 2, taking into account the micro-occurrence types of shale oil and its interactions with the rock matrix. By simulating the distribution of shale oil with various occurrence types, the model generates a digital core model that possesses prescribed geometric micro-occurrence patterns of shale oil. QSGSM not only allows for the precise insertion of oil droplets with different flow properties into the digital core framework but also effectively simulates the distribution state and occurrence forms of shale oil, such as floating, cementing, and coating types. In this way, we can synthetically generate realistic geometric occupation patterns of shale oil within the digital core, providing more accurate numerical simulations and theoretical support for understanding pore-scale fluid dynamics in shale reservoirs.

The generation mechanism of the QSGS (Wang et al., 2007; Yang et al., 2024) conceptually mirrors the physical diffusion of actual porous media, driving structural growth outward from designated seed cores. By adjusting four parameters (specified seed distribution S_d, porosity ϕ, growth probability of inter-component interactions $I_i^{m,n}$, and constant growth probability G_i), the topological characteristics of porous media can be controlled. This method has practical physical significance and is simple and efficient in algorithm. In this method, the porous media model we used is constructed in Section 2 (using FIB-SEM and watershed method), and then continue to generate shale oil within this porous media. The detailed QSGS generation procedure, including seed initialization, directional growth, and iterative filling, is described in Appendix B.

In actual core samples from reservoir formations, the heterogeneous pore structures give rise to a complex mixture of various micro-occurrence types of shale oil. This structural complexity significantly complicates the study of the topological characteristics and fluid flow behavior of different shale oil occurrence types. The challenge lies in isolating the individual contributions of each occurrence type within the porous network. Therefore, in this phase of the study, we systematically generated three distinct micro-occurrence types of shale oil (coating, floating, and cementing) within the core. This approach enables a more precise and focused examination of the spatial distribution, interaction patterns, and flow characteristics of different shale oil types, thereby facilitating a deeper understanding of their respective contributions to fluid dynamics in the reservoir.

The digital cores containing three different types of shale oil with floating, cementing, and coating configurations were constructed using the above method, as shown in Fig. 2(b). For each type, digital core models were built with shale oil concentrations of 10%, 20%, 30%, and 40%, totaling 12 pieces, as shown in Fig. 3.

Then, to investigate the different microscopic occurrence types of shale oil and the structural relationships between pores, as well as their impact on flow patterns, it is first necessary to characterize the topological structure of the pores within the shale oil core. Thus, it is essential to extract the pore network model and the shale oil network model from the core. Through the pore network

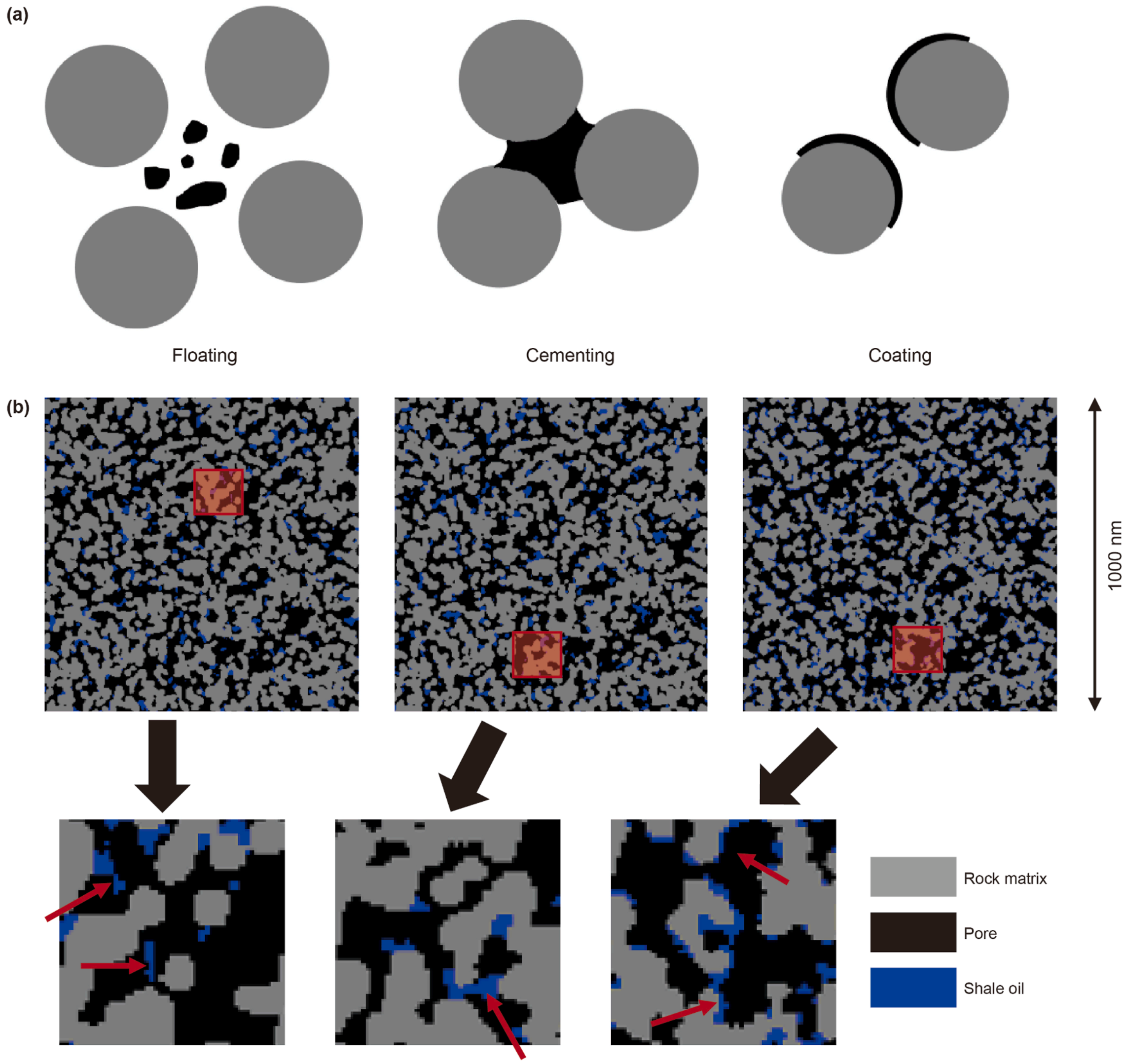


Fig. 2. Two-dimensional slice diagram of three micro-occurrence types of shale oil generated by QSGS method.

model, parameters such as pore radius, coordination number, and shape factor can be calculated. Similarly, the shale oil network model can provide information on the radius and shape factor of the shale oil. By combining both models, it is possible to analyze the segmentation of the original pores before and after the presence of shale oil. In this study, the extraction of the pore network model is achieved using the maximal ball method (Al-Kharusi and Blunt, 2007).

First, a voxel point is arbitrarily selected within the pore space, and its surrounding 26 directions are searched to identify contacting particles, referred to as skeleton voxels. Subsequently, based on the formation range of this voxel, a contraction algorithm is employed to locate the maximal ball and determine its upper and lower radius limits, thereby establishing the size and position of this maximal ball (Fig. 4(a)). This procedure is repeated iteratively to identify subsequent maximal balls, continuing until the final configuration of all maximal balls is achieved (Fig. 4(b) and (c)). During this process, redundant balls may emerge. Redundant balls are defined as those completely encapsulated within another maximal ball, rendering them ineffective for the characterization of the pore body. If ball B is entirely enclosed within ball A, the condition is

$$Dist(p_A, p_B) \leq R_A - R_B \tag{1}$$

where $Dist(p_A, p_B)$ represents the distance between the centers of balls A and B; R_A and R_B are the lower radius limits of the two balls. At this point, ball B can be removed. Repeat the process to remove all redundant balls.

To elucidate the differential impact of various microscopic occurrence types of shale oil on the topological structure of pore space, this study conceptualizes the shale oil phase as the occupying phase within the pore system, while the remaining non-oil-bearing regions are defined as the effective pore space for analysis. By characterizing this distinction, it becomes possible to investigate the relationship between different microscopic shale oil occurrence modes and the topological morphology of the pore network.

The results of shale oil network model construction are shown in Fig. 5. The image reveals that at lower saturation levels of shale oil, the pores within the core exhibit larger radius and are less abundant, appearing as red-colored balls and sticks in the pore network model with a relatively sparse distribution. As shale oil saturation increases, the pore radius decreases while the number

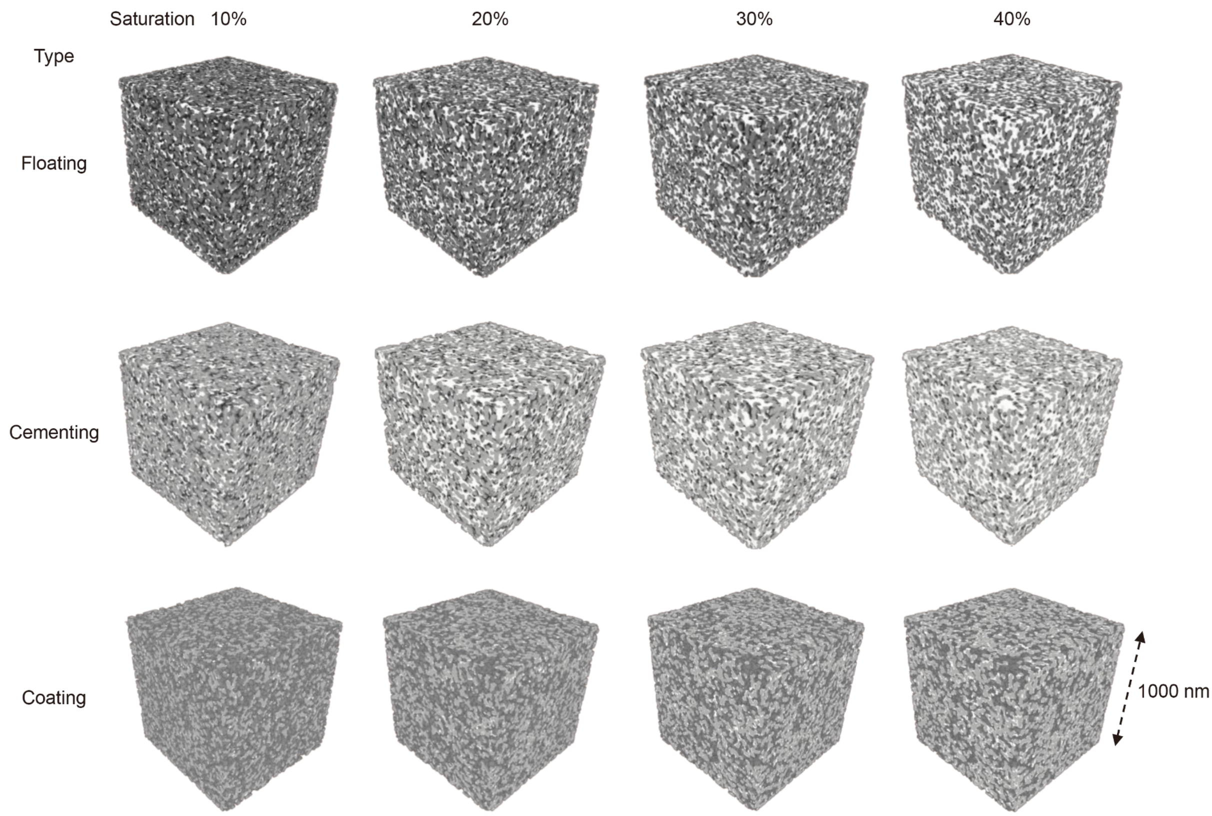


Fig. 3. Shale oil cores constructed by QSGS method with three different micro-occurrence types: floating, cementing, and coating.

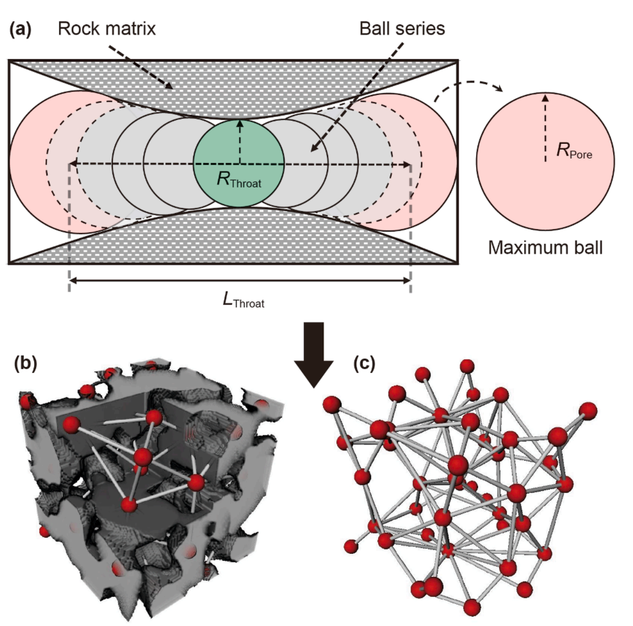


Fig. 4. (**a**) Schematic diagram of the maximal ball method for extracting the pore network model. (**b**) Schematic diagram of digital core with pore network. (**c**) Schematic diagram of the pore network model.

of pores rises, resulting in a proliferation of blue-colored balls and a denser pore network. This phenomenon is driven by the increase in shale oil saturation, causing the original pores within the core to be subdivided into smaller pores, leading to decreased radii and increased pore counts. Additionally, different shale oil occurrence states affect the extent of pore subdivision, with cementing shale oil showing a more pronounced division effect on pores compared to floating shale oil, which primarily reduces pore radius without significant pore subdivision.

In the process of constructing a pore network model, the number, radius, and coordination number of pores can be obtained. The coordination number represents the number of pores directly connected to a single pore, and the larger the coordination number, the better the connectivity of the pore. Specifically, the number of pores represents the number of balls in the pore network model, the pore radius is derived from the radius of the largest ball, and the coordination number of pores is the number of sticks connected to an individual ball.

As shown in Fig. 6, at a shale oil saturation of 10%, the number of effective pores in the floating-type and cementing-type shale oil configurations is nearly identical, indicating that both types exhibit similar characteristics in segmenting the original pore space at low saturation levels. As the saturation increases, the impact of the floating-type shale oil on the topological structure of the pore space becomes more pronounced. The oil phase progressively subdivides the pore space, resulting in an increase in the number of effective pores (from 5.2×10^3 to 9.5×10^3). In contrast, the coating-type shale oil primarily exists in association with the surfaces of mineral particles and does not directly subdivide the pore space. Consequently, the number of effective pores increases only marginally (by 1.7×10^3), especially at higher saturation levels, where the increase is significantly smaller compared to the other two types.

With increasing saturation, the average equivalent radius of oil-free pore spaces in all three occurrence types shows a general decreasing trend. The coating-type shale oil exhibits a larger equivalent radius, but also the greatest reduction

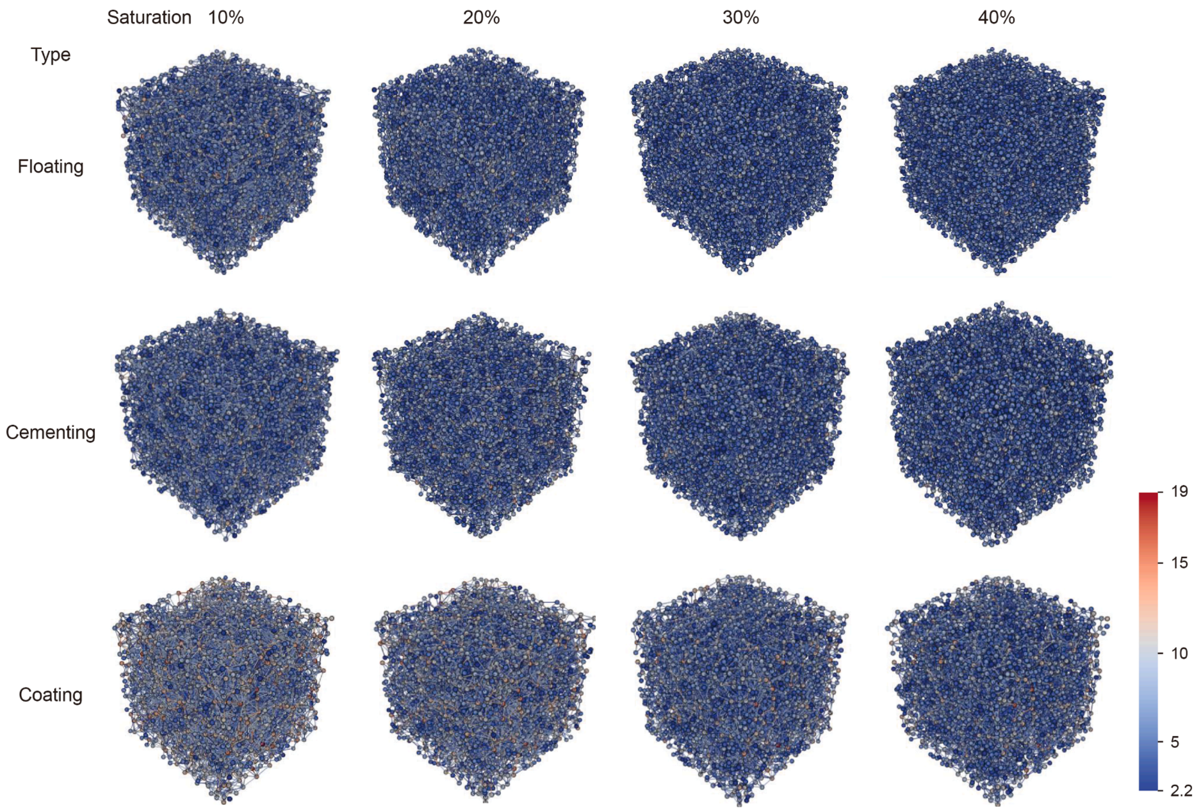


Fig. 5. Construction results of shale oil network models under different micro-occurrence types and saturation.

(approximately 2 nm). This suggests that at lower saturation, due to strong adhesion of the coating-type oil on pore surfaces, the remaining oil-free pore space is relatively large; however, with increasing oil saturation, the equivalent radius significantly diminishes. In comparison, the average equivalent radius for the floating and cementing types is smaller but exhibits a more gradual decline, decreasing from around 4 nm to approximately 3 nm, indicating that their effects on oil-free pore space are relatively consistent.

Coordination number is a critical metric for evaluating pore connectivity. In this study, the coordination number of the pore space after segmentation by the oil phase is defined as the "equivalent coordination number" to capture its relationship with the pore-structure-modifying effect of shale oil. At low saturation (10%), the coating-type shale oil exhibits the highest average equivalent coordination number (greater than 9), suggesting that its oil-free pore network remains relatively dense and well-connected. However, as saturation increases, this advantage diminishes, and the connectivity of the coating-type network becomes comparable to that of the floating and cementing types, resulting in reduced differences in average equivalent coordination numbers. The floating and cementing types maintain consistently smaller differences in coordination number across saturation levels, reflecting their similarly moderate impact on pore connectivity.

It is important to note that the shale pore system exhibits not only a geometric hierarchy (nanopores–micropores–fractures) but also an occurrence-state hierarchy, including adsorption, miscibility/dissolution, and free-fluid distribution. The present study focuses exclusively on the geometric hierarchy and does not address the physicochemical transitions between different occurrence states, which are governed by intermolecular interactions and thermodynamic conditions at the pore scale.

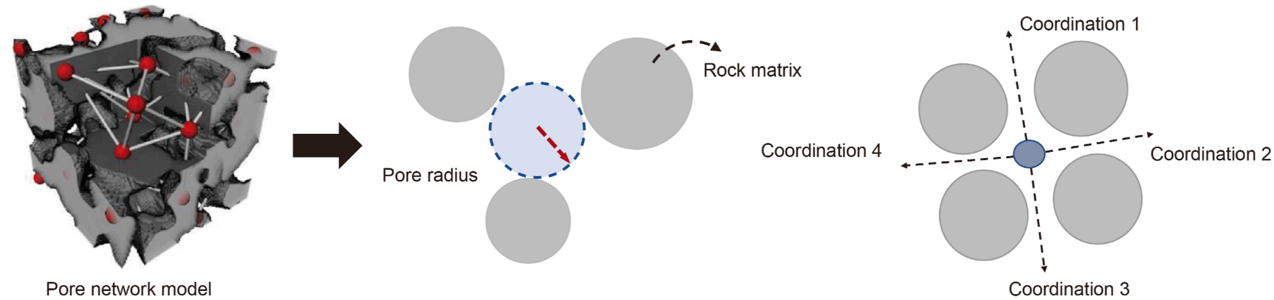
4. Micro-occurrence type classification based on machine learning

In previous studies, the focus at the microscopic scale has primarily been on the qualitative description of the topological features of different occurrence types of shale oil. However, a clear distinction standard between various micro-occurrence types of shale oil has yet to be established. Therefore, machine learning techniques are employed in this section to train data corresponding to the different micro-occurrence types of shale oil discussed in the previous sections. The resulting model can then be utilized to classify the occurrence type of unknown shale oil.

4.1. Data collection and preprocessing

In Section 3, pore network models of digital cores containing three types of shale oil, namely floating, coating, and cementing, are extracted. These shale oil-bearing cores are all generated from the same digital core using the QSGS method (Fig. 2). As a result, the shale oil portions can be further extracted from these models and their topological structure parameters can be analyzed. The resulting parameter sets can then be used for machine learning applications.

In the machine learning approach, the selected dataset parameters include the volume ratio of shale oil to the pore it resides in (V_r), the cross-sectional area ratio of shale oil to the pore (S_r), the three-dimensional shape factor (G), and the Euler number (Eu). The reason for choosing these parameters is that different microscopic occurrence types of shale oil are mainly characterized by differences in the topological structure and shape of shale oil itself as well as the topological relationship between shale oil and the pore space it is located in. Therefore, the volume ratio of shale oil to the pore, the cross-sectional area ratio of shale oil to the pore, and the



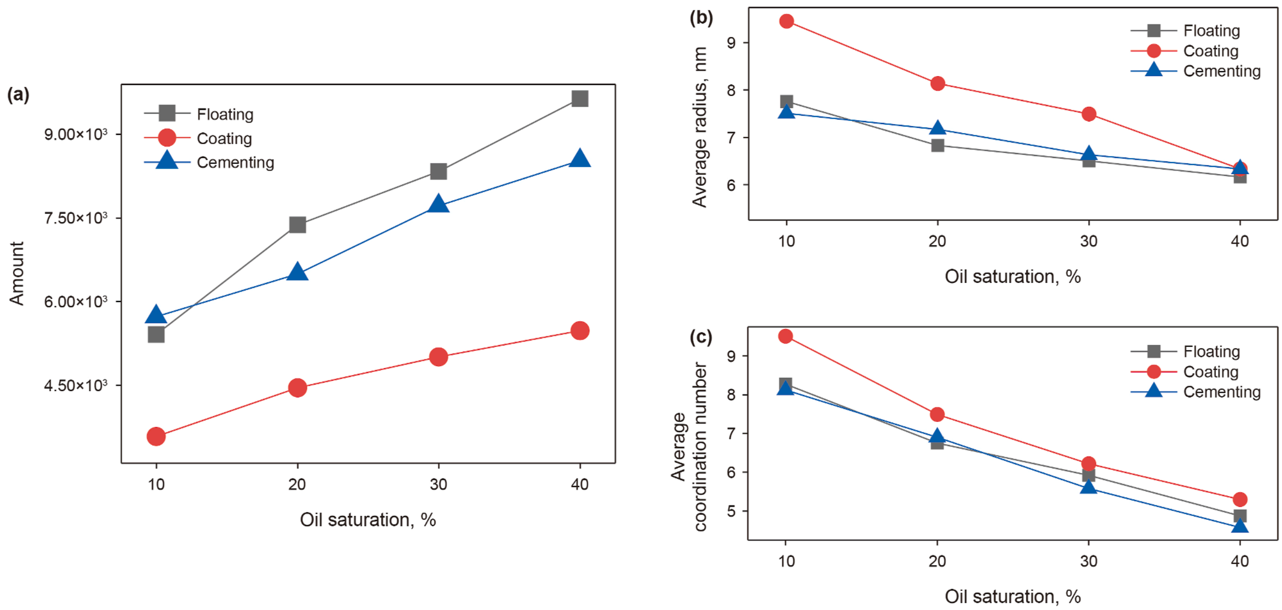


Fig. 6. Values and variation curves of the three parameters for different micro-occurrence types of shale oil under different shale oil saturation levels: (**a**) amount; (**b**) average radius; (**c**) average coordination number.

Euler number can reflect the topological relationship between shale oil and the pore space. The three-dimensional shape factor can reflect the topological structure and shape of shale oil itself. Thus, combining these four parameters for machine learning can effectively distinguish different microscopic occurrence types of shale oil.

Among them, the calculation of the volume ratio (V_r) of shale oil to the pore it is in and the cross-sectional area ratio (S_r) of shale oil to the pore is based on the pore network model. The formulas are as follows:

$$V_r = \frac{V_o}{V_p} \tag{2}$$

$$S_r = \frac{S_o}{S_p} \tag{3}$$

where V_o and S_o represent the volume and area of shale oil in the pore; V_p and S_p represent the volume and area of the pore. The three-dimensional shape factor G is obtained from the above two parameters. The calculation formula is as follows:

$$G = \frac{V_r}{S_r} \tag{4}$$

In three-dimensional images, the Euler number Eu is calculated by the formula:

$$Eu = V - E_0 + F - C \tag{5}$$

where V represents the number of vertices; E_0 represents the number of edges; F represents the number of faces; and C represents the number of connected components.

The data set formed by the calculation results is shown in Table 1.

4.2. Machine learning processing

Given the intricate nature and diverse range of shale oil microscopic occurrence types, and the availability of labeled datasets, we opt for supervised learning methods to classify these occurrences. Among various approaches, we have selected three methods: random forest, SVM, and KNN, each offering unique strengths for this specific task. The schematic diagram of the three machine learning methods is shown in Fig. 7.

The random forest algorithm addresses the challenge of shale oil microscopic occurrence type recognition by constructing

Table 1
Data set of micro-occurrence types of shale oil divided by machine learning.

ID	V_r	S_r	Eu	G	Micro-occurrence
1	0.03827	0.08412	2	0.455	Cementing
2	0.05654	0.12434	5	0.454	Cementing
			......		
6565	0.27173	0.39535	0	0.647	Coating
6566	0.29192	0.42698	1	0.643	Coating
			......		
11636	0.19626	0.32368	4	0.606	Floating
11637	0.15714	0.25954	5	0.605	Floating
			......		

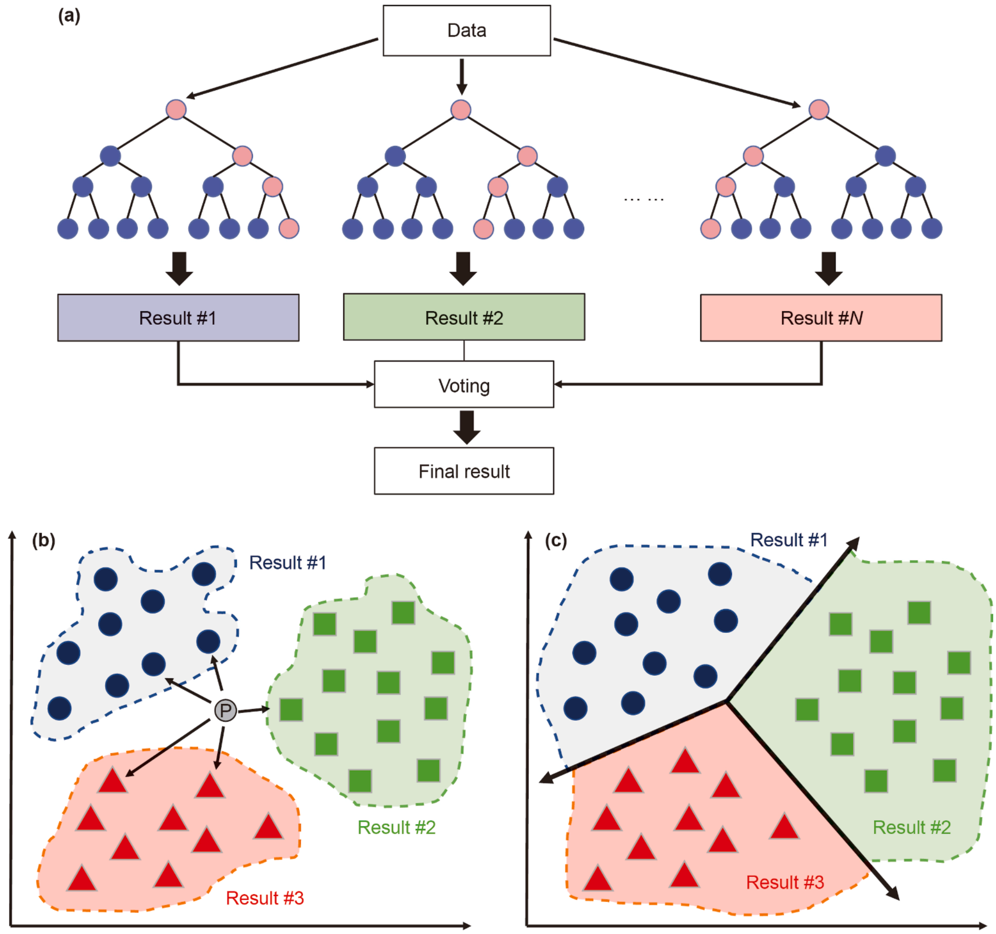


Fig. 7. Schematic diagrams of machine learning methods: (**a**) random forest method; (**b**) KNN method; (**c**) SVM method.

multiple decision trees. Each decision tree is trained on a randomly sampled subset of data points and features from the original dataset, effectively mitigating overfitting and enhancing model generalization. Ultimately, the predictions of all decision trees are aggregated through voting or averaging, as shown in Fig. 7(a).

For types with minimal feature differences, the KNN algorithm excels. KNN operates based on the principle of similarity-based classification. It assigns a new data point to the class that is most prevalent among its k nearest neighbors in the feature space, as shown in Fig. 7(b). The key parameter k determines the number of nearest neighbors considered during classification. KNN is a non-parametric and lazy learning algorithm, requiring no explicit training phase. This method effectively captures local features based on data similarity, achieving high classification accuracy.

When distinct classes exhibit significant feature differences, the SVM algorithm proves effective. SVM aims to find an optimal hyperplane that maximally separates different types of microscopic occurrences. By mapping data points into a high-dimensional feature space, SVM seeks to construct a decision boundary that maximizes the margin between distinct classes, achieving robust classification, as shown in Fig. 7(c).

After preparing the data set, features and corresponding classification results are first set in the algorithm. The first four columns in this data set are features, and the fifth column contains the classification labels, which are extracted as feature matrix $\boldsymbol{X}$ and classification result vector $\boldsymbol{y}$, respectively. After that, the data set is divided into the training set and the test set, where the test set accounts for 10% The random forest model, constructed with 100 decision trees and a random seed of 42 for reproducibility, was trained on the training data. SVM employed a kernel function to map the data into a high-dimensional feature space, enabling the identification of a hyperplane that maximized the margin between classes. This process determined support vectors and hyperplane parameters. KNN, in contrast, simply stores the training data for subsequent prediction. To classify a new data point, KNN calculates its distances to all stored data points, selects the k nearest neighbors, and assigns the new data point to the class with the highest frequency among those neighbors. The performance of each model was then assessed on the testing set, providing insights into the suitability of each algorithm for the specific task.

To evaluate the classification performance of the models, we employed feature importance values and confusion matrices to analyze the accuracy of the prediction results.

In random forest, each decision tree selects the feature and split point that maximizes the reduction in impurity after splitting a node. Upon completing the random forest construction, the average impurity reduction value for each feature across all decision trees is calculated. This value serves as a measure of the feature's importance. Features with higher importance values contribute more significantly to the random forest classification process.

A confusion matrix compares the actual classification results from the test set with the model's predictions, providing a count of correct and incorrect classifications for each category. By tabulating correct and incorrect classifications for distinct category labels, a matrix is formed. The rows represent the actual categories, and the columns represent the predicted categories. In a three-class problem, for example, the matrix might be a 3×3 matrix where the diagonal elements represent the number of correct classifications, and off-diagonal elements represent incorrect classifications. This allows for a visual understanding of the model's performance across different classes.

Furthermore, we utilized learning curves to assess the performance changes of the model across varying training set sizes. By dividing the training set size into ten equal parts from 0.1 to 1.0 and applying five-fold cross-validation, we observed the training scores and cross-validation scores at different training data volumes. Learning curves illustrate the trend of model performance as training data volume increases, enabling us to determine whether the model is overfitting or underfitting and whether more training data is required.

4.3. Result analysis

Based on the confusion matrices of the three machine learning methods (Fig. 8(a)–(c)), it is evident that the KNN method exhibits the lowest predictive accuracy at only 75%, while the random forest method achieves a higher predictive accuracy of 91%. The SVM method demonstrates the highest predictive accuracy, reaching up to 93%. Analyzing the learning curves, we observe that the accuracy of the KNN method decreases as the training set size increases (Fig. 8(a)). This suggests that, with a smaller training set, the KNN method may be able to fit the data effectively, as it can memorize the relationships between the features and labels of the training samples. However, as the training set expands, the model may become overly focused on noise and specific details within the training data, neglecting the overall patterns, which can lead to overfitting. The cross-validation scores gradually rise to approximately 0.6 and then stabilize, indicating that the model performs generally on different data splits, resulting in moderate predictive accuracy on unseen data. (Cross-validation is an evaluation method that assesses model performance by partitioning the dataset into multiple subsets, training and validating on different combinations to simulate the model's performance on new data.) For the random forest method (Fig. 8(b)), the learning curve shows that the training score remains at 1, while the maximum cross-validation score only reaches around 0.73, suggesting some degree of overfitting; the method performs well on the training set but not as effectively on new, unseen data. The SVM method (Fig. 8(c)), exhibiting the highest predictive accuracy, effectively learns the patterns and features of the data, demonstrating a high degree of fit to the training data. Moreover, the cross-validation curve yields favorable results, with the cross-validation scores rising as the training set size increases, indicating an improving generalization capability of the model. This also reflects the model's stability and reliability, as it maintains high accuracy across different data splits.

Fig. 9 illustrates the decision boundaries for the three machine learning methods. The random forest method generally succeeds in clearly separating different classes; however, there are some regions where the decision boundaries exhibit alternate fluctuations, suggesting potential overfitting and indicating a relatively poor generalizability of the model. In the KNN method, the decision boundaries appear quite coarse, signifying that the model has over-fitted to the training data and is excessively sensitive to noise. In contrast, the SVM method displays smoother decision boundaries compared to the other two methods, suggesting that the model's fit to the data is more stable and less influenced by noise and outliers. Additionally, this method effectively distinguishes between different classes, as evidenced by the distinct distribution of sample points on either side of the decision boundary, which indicates a favorable classification performance of the model. Therefore, a comprehensive comparison of the results suggests that the SVM method holds a significant advantage in addressing the classification problem of micro-occurrence types in shale oil.

5. LBM flow simulation for different shale oil occurrences in digital cores

Building on the machine learning classification results, we employ LBM to conduct in-depth studies of shale oil cores with diverse micro-occurrence types. Through simulation analysis, we aim to uncover the intrinsic mechanisms by which these micro-occurrence types influence core parameters such as fluid flow patterns and permeability.

The LBM is utilized to investigate the influence of different saturation and types of shale oil occurrences on fluid flow. LBM incorporates the geometric properties of the porous medium directly into its framework, allowing for realistic simulation of complex pore structures and fluid interactions. For the simulations, we used the D3Q19 LBM discretization with 19 discrete molecular velocities, denoted by e_i ($i = 0, 1, 2, \ldots, 18$). To increase numerical stability, we adopted the multi-relaxation-time (MRT) model. The governing equation is shown in Eq. (6).

$$n_i(\boldsymbol{x} + e_i\Delta t, t + \Delta t) = n_i(\boldsymbol{x}, t) - M^{-1}\widehat{S}\left[m_i(\boldsymbol{x}, t) - m_i^{\mathrm{eq}}(\boldsymbol{x}, t)\right] \tag{6}$$

where vector $\boldsymbol{x}$ represents lattice location; t is time; subscript i is the predefined directions of the molecular motion; $n_i(x, t)$ denotes the particle population; $n_i(x + e_i\Delta t, t + \Delta t)$ is the particle population after the evolution; $\widehat{S}$ is the collision matrix; $m_i(x, t)$ represents an equal number of moments; M is an orthogonal transformation, and the moments satisfy $m_i(\boldsymbol{x}, t) = Mn_i(\boldsymbol{x}, t)$. The streaming step incorporates a local solid ratio $P(\boldsymbol{x}) = V_{\mathrm{solid}}(\boldsymbol{x})/V(\boldsymbol{x})$ to account for partially solid nodes. The detailed collision-streaming formulation is standard and can be found in the cited reference.

This relaxation process effectively mimics the effect of inter-particle collisions, driving the system towards thermodynamic equilibrium. During the streaming step, particles propagate to neighboring lattice nodes along their respective discrete velocity directions. As multiple particles from different directions arrive at the same node, they undergo collision, redistributing their velocities according to the prescribed collision rules, thus mimicking the momentum exchange in a real fluid. This combination of local collision and streaming steps effectively simulates the macroscopic flow behavior. During the flow process, Darcy's law is employed to compute the absolute permeability of the porous medium, this law utilizes parameters such as the average fluid velocity, pressure gradient, and dynamic viscosity of the fluid. Due to the typically high viscosity of shale oil and its relatively poor mobility compared to water in rock pores, shale oil is often trapped within the pore spaces of the rock during actual flow processes, making it difficult to mix or flow with water. In this study, a simplified model is employed in which the oil phase is treated as a stationary phase, to simulate the flow of water within shale oil cores containing different micro-occurrence types. For the above 12 cores, the LBM method was applied to conduct flow simulations. Both the shale oil and rock skeleton were treated as stationary phases. A constant pressure difference was applied at the inlet and outlet boundaries as the driving force. The specific parameters are listed in Table 2. The above calculations were performed on a Dell tower workstation with an Intel Xeon E5 CPU and 64.0 GB RAM.

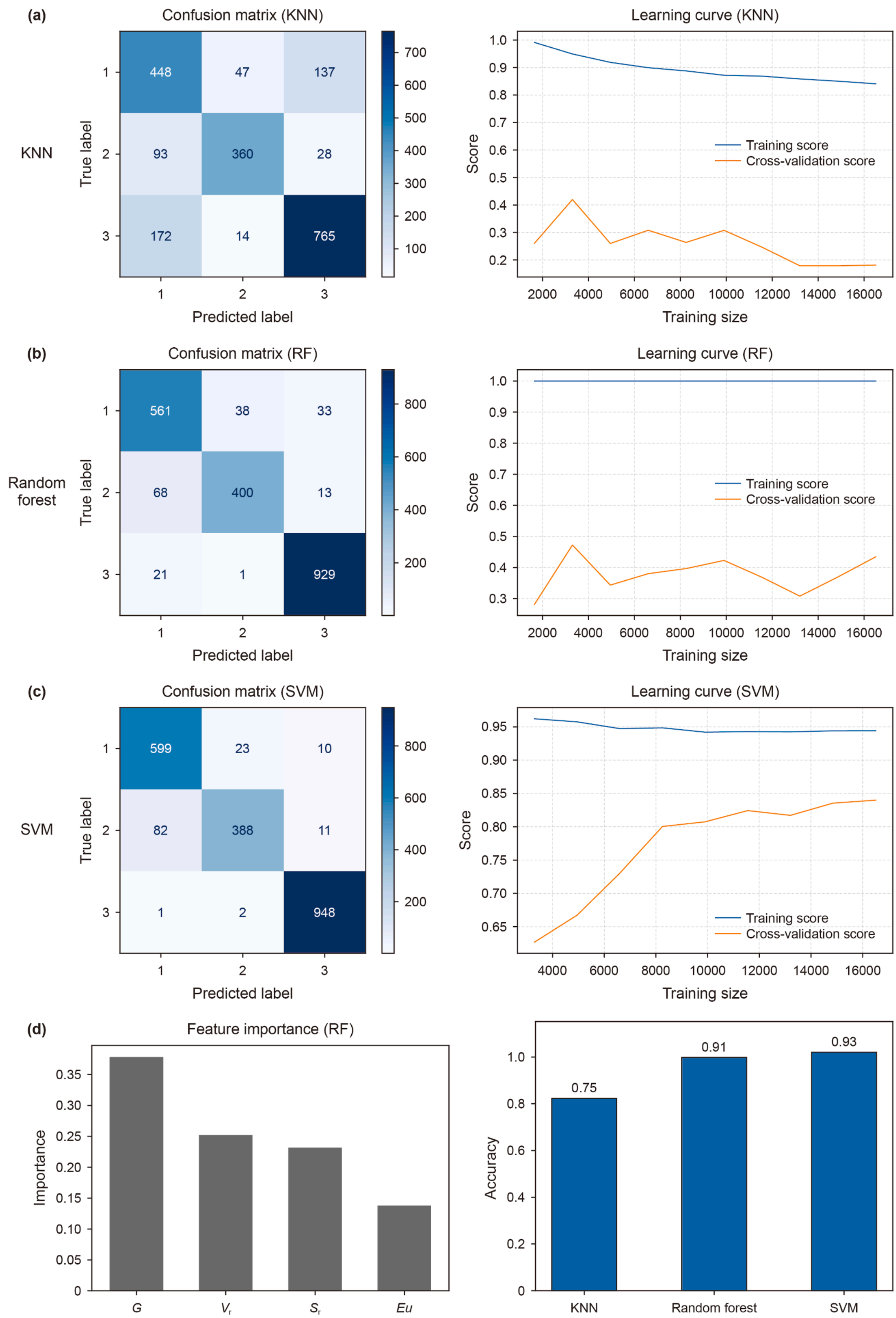


Fig. 8. Confusion matrix and learning curve for the KNN method (**a**), random forest method (**b**), and the SVM method (**c**). (**d**) Predictive accuracy results for different machine learning methods.

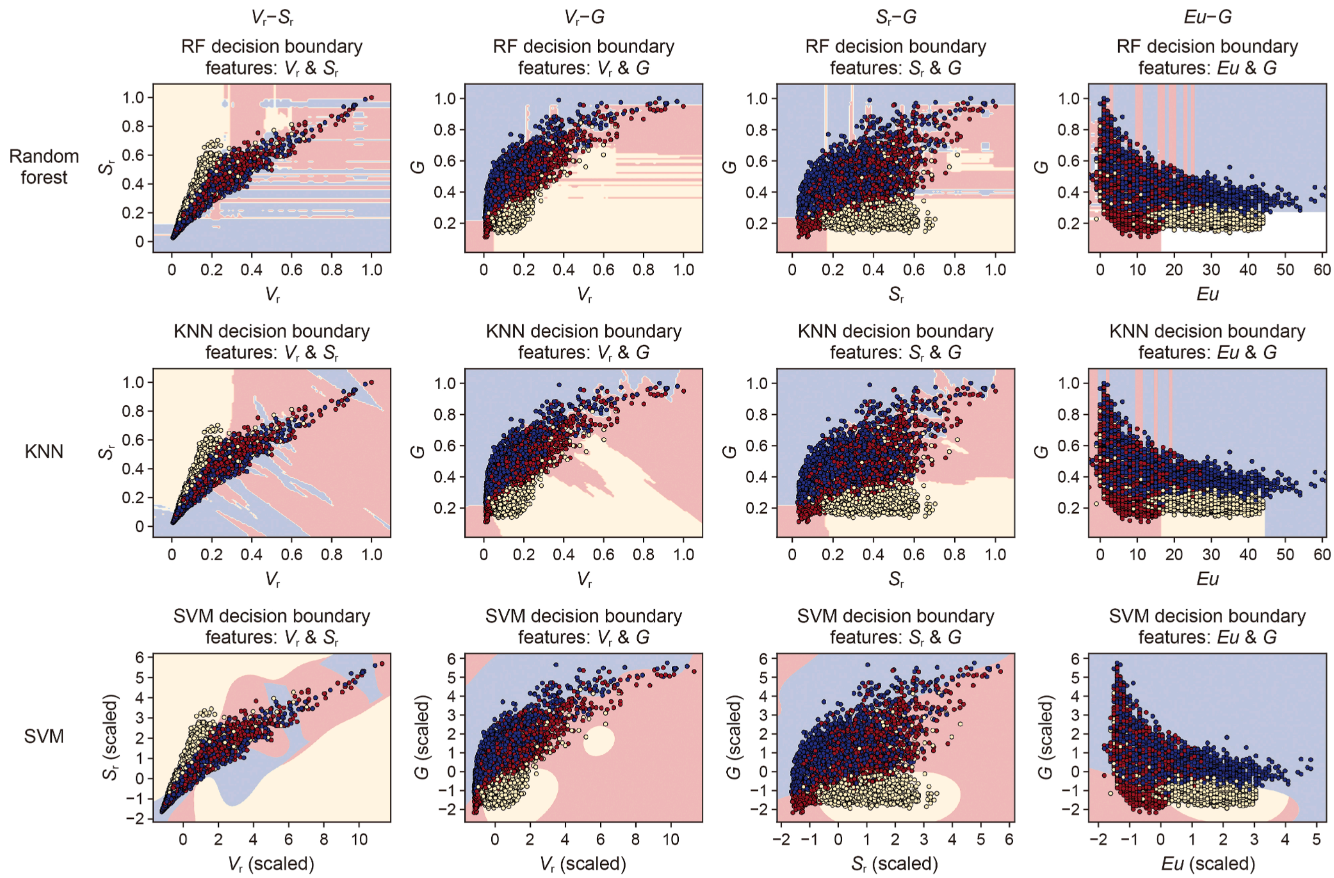


Fig. 9. Decision boundaries for classification using random forest, KNN, and SVM methods.

The streamline diagrams from LBM simulations are shown in Fig. 10. Under conditions of low oil saturation (10%), the obstructive effects of all three occurrence modes on fluid flow are relatively weak. The streamline distribution within the core exhibits dense and continuous characteristics. Coating occurs as a thin layer uniformly attached to the surface of the rock framework, only slightly reducing the effective pore diameter and minimally disturbing the main flow paths. Consequently, cores with coating show the highest streamline density and the lowest flow resistance. Floating is dispersed as isolated droplets in the pore centers. Its flow resistance primarily arises from the bypassing flow around the droplets, leading to localized velocity variations, although the overall flow paths remain relatively unimpeded. In contrast, cementing, though partially filling cemented regions between mineral grains, does not fully block pore connectivity due to its limited extent under low saturation. As a result, its streamline density is comparable to that of floating. At this stage, fluid flow is mainly governed by pore geometry, and the influence of occurrence modes on overall permeability is not yet significant; flow behavior is still primarily determined by the structural features of the pore network.

With the increase in shale oil saturation, the different effects of shale oil in various micro-occurrence on fluid flow become more prominent, as shown in Fig. 10. As oil saturation increases to 30%, the obstruction to fluid flow from different occurrence modes diverges significantly. In coating, the thickened oil layer further narrows pore throats without completely sealing the channels. Flow resistance here stems primarily from viscous boundary layer effects, causing notable velocity reduction near the walls, while the mainstream retains relatively high streamline density. In floating, the number of suspended droplets increases with saturation and gradually coalesces into larger oil clusters. These clusters force the fluid to bypass them, greatly increasing the tortuosity of the flow paths and reducing permeability. On the microscale, the bypassing process dissipates local kinetic energy, manifested as streamlines splitting and rejoining around the droplets. In cementing, the cemented region expands with rising saturation, directly blocking the connecting throats between pores. This forces fluid to seep only through the remaining uncemented spaces, sharply reducing the number of streamlines and creating a fragmented distribution.

The permeability and flow rate calculated by LBM are shown in Fig. 11. From Fig. 11(a)–(c), it can be observed that the coating-type shale oil exhibits the highest core permeability, followed by the floating-type. In contrast, the presence of cementing-type shale oil significantly impairs permeability, resulting in the largest reduction among the three occurrence types. Specifically, for the cementing type, permeability drops steeply from 0.453 mD at 10% saturation to nearly zero at 40% saturation.

From the normalized flow and permeability curves (Fig. 11(d) and (e)), several key observations emerge: the marked decreases in permeability and flow rate are primarily attributable to the cementing type, which effectively blocks pore channels, as evidenced by the near-zero flow rate at 30% and 40% saturation. Floating-type shale oil also impedes fluid flow to some extent by

Table 2
Parameters in LBM simulation.

Parameter	Value
Resolution, m	2.5×10^{-9}
Kinematic viscosity, m^2/s	10^{-6}
Density, kg/m^3	10^3
Pressure difference, Pa	10^6

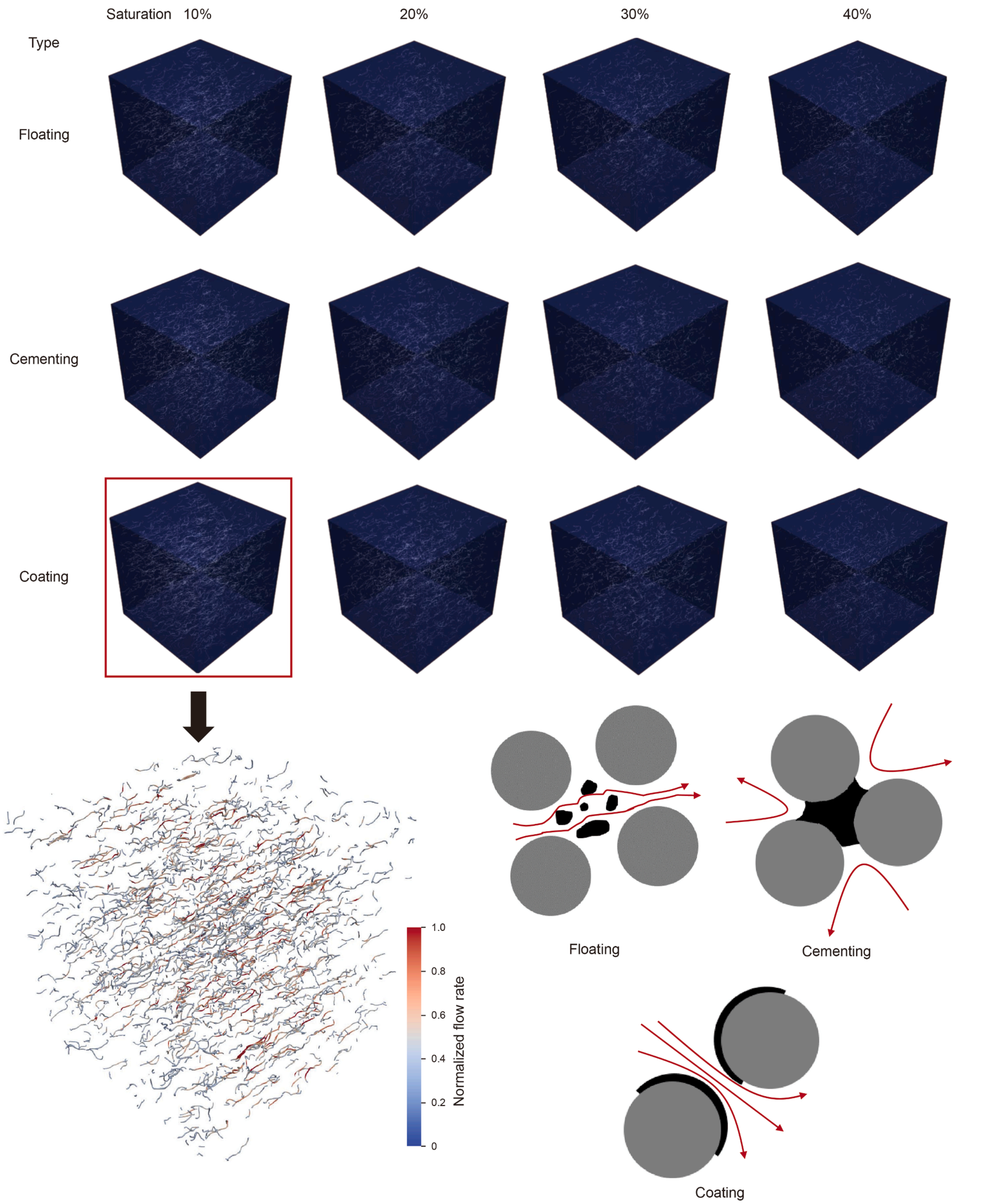


Fig. 10. Streamline diagrams of fluid flow inside cores at different saturation and micro-occurrence types obtained from LBM simulation.

occupying the pore space as suspended droplets, forcing fluids to detour around it and increasing the tortuosity of flow paths. This is reflected in the more gradual decline of flow rate from 0.703 to 0.134 as saturation increases from 10% to 30% for the floating type. In contrast, coating-type shale oil predominantly adheres to rock particle surfaces, minimally hindering flow paths. However, at higher saturations (e.g., 30% and 40%), coating-type shale oil begins to encroach into pore centers and interconnected throats, causing noticeable reductions in permeability and flow rate. Specifically, the flow rate for the coating type drops from 0.835 at 20% saturation to 0.294 and 0.113 at 30% and 40% saturation, respectively.

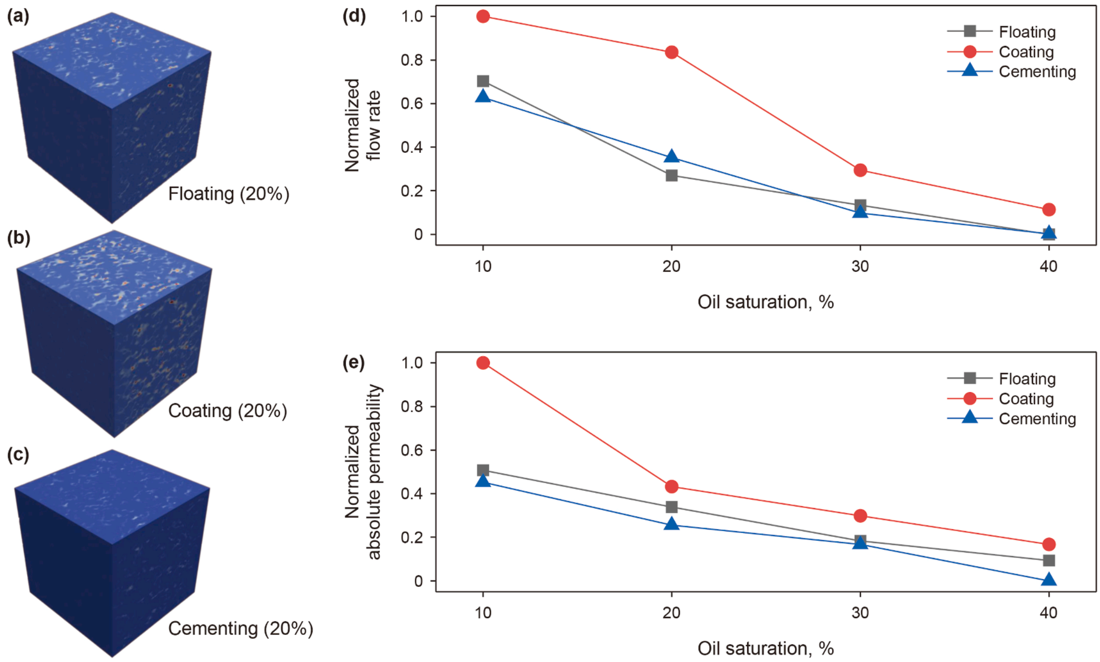


Fig. 11. (**a**–**c**) Permeability fields of three micro-occurrence types of shale oil cores under 20% shale oil saturation. (**d**) Normalized flow rate curve. (**e**) Normalized absolute permeability curves.

Further insights from LBM simulation results reveal that different occurrence modes exert distinctly different control over pore-scale seepage behavior. Coating-type shale oil forms a nanoscale film along the rock surfaces, reducing effective pore size without disrupting overall channel continuity. Mainstream regions retain relatively high velocity at low saturation (normalized velocity of 0.835 at 20% saturation). When saturation exceeds 30%, however, the coating-type oil extends toward pore centers and forms a double-layer coating structure, leading to geometric blockage of flow paths. This structural transition results in a sharp decline in permeability, with flow velocity decreasing by approximately 70% from 20% to 40% saturation, suggesting a shift from surface adsorption effects at low saturation to volumetric blockage at higher saturation levels.

Floating-type shale oil occupies the central pore space as discrete droplets, with its primary hindrance arising from increased path tortuosity rather than direct blockage. As saturation rises from 10% to 40%, suspended droplets gradually coalesce into pore-scale clusters, forcing flow to take longer, more tortuous paths and consuming more energy. Streamline simulations indicate a significant elongation of average flow paths and elevated flow resistance, consistent with the observed normalized velocity decline from 0.703 to 0.134. Notably, the resistance introduced by floating-type shale oil remains partially reversible when subjected to increased displacement pressure, as the suspended droplets may be mobilized.

In contrast, cementing-type shale oil dominates through pore throat blockage, which intensifies as the cemented regions expand. At 30% saturation, over 80% of the primary flow paths are obstructed. This causes a dramatic deterioration in the pore network's topological connectivity: coordination numbers drop sharply, pore interconnections decrease markedly, and isolated pore clusters begin to form. As the fraction of disconnected pore volume crosses a critical threshold, previously continuous flow pathways fragment into residual micro-paths, leading to core-scale permeability collapse (from 0.453 to near-zero). This phenomenon reflects a network transition from global connectivity to localized isolation of pore spaces.

To provide a more systematic comparison, the flow behavior and permeability variations associated with the three micro-occurrence types are further summarized as follows:

The coating-type shale oil exerts the weakest influence on the overall flow pattern within the pore system. At low saturation levels (10% to 20%), it mainly exists as a thin nanoscale film uniformly coating the surfaces of the rock framework. This configuration slightly reduces the effective pore radius and generates a local viscous boundary layer, but it maintains the geometric continuity and connectivity of the principal flow channels. As a result, this occurrence type shows the lowest flow resistance and the highest initial permeability and streamline density among the three. With increasing saturation (30% to 40%), the coating film thickens and gradually extends toward the pore centers, leading to a more noticeable reduction in permeability. However, the main mechanism remains the narrowing of existing pore throats rather than complete blockage, which distinguishes this type from the more obstructive behaviors of the other modes.

The floating-type shale oil mainly impedes flow by increasing the tortuosity of the flow paths. At low saturation, it occurs as discrete droplets suspended within pore centers, forcing the fluid to bypass these obstacles by splitting and rejoining streamlines. This bypassing increases the effective flow distance and results in higher flow resistance. As saturation rises, the droplets coalesce into larger oil clusters, forming more irregular flow channels. This process significantly enhances the tortuosity effect, consuming additional pressure energy to maintain flow. Consequently, the floating type exhibits an intermediate permeability reduction. Although it does not destroy pore connectivity as severely as the cementing type, it still causes considerable resistance due to the complex flow paths within the pore network.

The cementing-type shale oil has the most detrimental effect on fluid flow due to its mechanism of direct pore-throat blockage.

Unlike the other two modes, the cementing type preferentially occupies the narrow pore throats that connect larger pores, which severely reduces pore connectivity even at low saturation. As saturation increases beyond 30%, cemented regions expand rapidly, blocking the major flow channels. This results in fragmentation of the originally connected pore network into isolated clusters and a significant decline in topological connectivity. Simulation results clearly demonstrate this effect, showing a sharp decrease in streamline density and permeability. This phenomenon reflects a transition from a connected, percolating network to a disconnected, non-percolating one caused by pore-throat blockage.

6. Conclusions

(1) A shale digital core model was constructed using the FIB-SEM method. Based on the analysis of the topological morphological characteristics of each micro-occurrence type of shale oil, the application of the quartet structure generation set method and the selection of appropriate growth parameters (such as growth probability and inter-component interactions), a total of 12 shale oil cores of three different micro-occurrence types with different saturation levels were constructed, which are as follows: floating, cementing, and coating.

(2) The pore network model of the core is extracted by the maximum ball method, and the corresponding pore radius and coordination number are calculated. The effects of different microscopic types of shale oil and saturation on pore size and connectivity are revealed. As shale oil saturation increases, original pores are subdivided into smaller ones, resulting in decreased radii and increased pore counts. Additionally, different shale oil occurrence states affect the extent of pore subdivision, with cementing shale oil displaying a more pronounced division effect on pores compared to floating shale oil, which primarily reduces pore radius without significant pore subdivision.

(3) Machine learning was employed to investigate the classification of micro-occurrence types of shale oil based on their topological features. Three supervised learning methods, random forest, support vector machine (SVM), and K-nearest neighbors (KNN), were employed to analyze a dataset of shale oil samples characterized by their volume ratio, cross-sectional area ratio, three-dimensional shape factor, and Euler number. Results showed that the SVM method achieved the highest predictive accuracy (93%), outperforming the random forest (91%) and KNN (75%) methods. The SVM model demonstrated effective learning of data patterns and features, exhibiting smooth decision boundaries and maintaining high accuracy across various data splits. This indicates that SVM is particularly well-suited for classifying subtle differences between micro-occurrence types of shale oil. Analysis of learning curves provided insights into the generalizability and overfitting tendencies of each model. These results highlight the efficacy of machine learning as a powerful tool for classifying complex microscopic structures in shale oil.

(4) By using lattice Boltzmann method, the fluid flow laws of shale oil cores under different micro-occurrence types and saturation conditions are studied, and the permeability and velocity of fluid in the cores are calculated. Coating-type shale oil exhibits the highest core permeability, followed by floating type. However, the cementing-type shale oil significantly affects the permeability, resulting in the greatest decrease among the three types. The significant decreases in permeability and flow rate are primarily attributed to the cementing type, which blocks the channels for fluid flow. Floating-type shale oil also impedes fluid flow by increasing the tortuosity of flow paths. Coating-type shale oil predominantly adheres to the surfaces of rock particles. However, at higher saturations (30% and 40%), it begins to expand towards the central regions of pores and through interconnected throats, thereby exerting some influence on permeability and fluid flow rates. Coating type impedes flow through viscous boundary-layer effects at low saturation and volumetric blockage at high saturation; floating type increases flow resistance by enhancing flow-path tortuosity; cementing type causes the most severe impact through direct pore-throat blockage, leading to a percolation transition from a connected to a fragmented pore network.

CRediT authorship contribution statement

Yi-Fan Yin: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation. **Zhi-Xue Sun:** Project administration, Funding acquisition. **Jing Wang:** Investigation, Funding acquisition, Data curation. **Kai-Jun Tong:** Supervision, Resources. **Yong-Fei Yang:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos: 51774317, U23A20595). First author Yi-Fan Yin is grateful for the financial support by the China Scholarship Council (No. 202506450054).

Appendix A. FIB-SEM sample preparation details

The sample was initially polished using dry sandpaper, starting with a coarse P600 grit to remove macroscopic irregularities, followed by a P800 grit to achieve a smoother surface. Subsequently, two distinct Pt deposition steps were performed. First, electron-beam-induced deposition (EBID) was utilized to deposit a thin protective Pt layer over the pre-polished surface, safeguarding the region of interest from damage during subsequent ion beam milling. Next, U-shaped trenches were milled into the surface, and a second Pt layer was deposited within them via ion-beam-induced deposition (IBID). This secondary layer was applied to mitigate charge accumulation under the high-vacuum SEM environment, thereby enhancing imaging contrast and quality.

Appendix B. QSGS generation procedure

The specific construction steps of the QSGS method are as follows.

1. Firstly, random initial nodes for shale oil growth are generated within the pore regions. Let the shale oil be the growth phase, denoted by 1, and the porous media be the non-growth phase, denoted by 0. Initially, all phases are set to 0. Determine the solid phase nucleus growth probability S_d. Generate random numbers within the [0, 1] interval for all nodes in the initial

phase. When the random number of a node is less than or equal to S_d, the node becomes part of the shale oil and is referred to as an oil phase nucleus. The oil phase nucleus growth probability S_d cannot be greater than the porosity ϕ.

2. Determine the growth probability G_i in the i direction. With the nodes that have already become part of the porous medium as cores, if the neighborhood of a node is in the non-growth phase, generate random numbers within the [0, 1] interval for these nodes. When the random number for a node is less than or equal to D_i, that neighborhood node becomes part of the oil. This step represents the process in which the shale oil grows from the nucleus towards the surrounding direction. The constructed porous medium in this study is three-dimensional, hence the growth direction is three-dimensional, with the range of i values being from 0 to 18.
3. Repeat the above steps until the set porosity ϕ is satisfied in the calculation domain.
4. Subsequently, the inter-component interaction $I_i^{m,n}$ is adjusted to control the topological shape of the shale oil, in order to generate shale oil with different micro structural types. This parameter is used to control the degree of interaction between different components. During the shale oil growth process, the interactions between components determine the expansion and aggregation of the oil phase within the pore space. Therefore, the interaction probability between components can regulate the geometric and topological distribution characteristics of the oil phase.

Specifically, the physical connection between QSGS parameters and the resulting micro-occurrence morphologies can be summarized as follows. A higher seed distribution probability S_d produces more dispersed nucleation sites throughout the pore space, favoring the formation of isolated droplets characteristic of the floating type. The directional growth probability G_i controls the anisotropy of oil-phase growth: isotropic growth tends to produce spherical or irregular clusters (floating and cementing), while preferential growth along pore-wall directions generates thin-film structures typical of the coating type. The inter-component interaction $I_i^{m,n}$ governs the affinity between the oil phase and the rock matrix: high affinity drives oil growth along rock surfaces (coating), low affinity allows independent oil aggregation in pore centers (floating), and intermediate affinity promotes oil accumulation at pore-throat junctions (cementing).

References

Abell, A., Willis, K., Lange, D., 1999. Mercury intrusion porosimetry and image analysis of cement-based materials. J. Colloid Interface Sci. 211, 39–44. https://doi.org/10.1006/jcis.1998.5986.

Al-Kharusi, A.S., Blunt, M.J., 2007. Network extraction from sandstone and carbonate pore space images. J. Pet. Sci. Eng. 56, 219–231. https://doi.org/10.1016/j.petrol.2006.09.003.

Alipour, P., Toghraie, D., Karimipour, A., Hajian, M., 2019. Molecular dynamics simulation of fluid flow passing through a nanochannel: Effects of geometric shape of roughnesses. J. Mol. Liq. 275, 192–203. https://doi.org/10.1016/j.molliq.2018.11.057.

Bleau, A., Leon, L., 2000. Watershed-based segmentation and region merging. Comput. Vis. Image Understand. 77, 317–370. https://doi.org/10.1006/cviu.1999.0822.

Blunt, M.J., 2001. Flow in porous media—Pore-network models and multiphase flow. Curr. Opin. Colloid Interface Sci. 6, 197–207. https://doi.org/10.1016/S1359-0294(01)00084-X.

Cao, Z., Jiang, H., Zeng, J., Saibi, H., Lu, T., Xie, X., Zhang, Y., Zhou, G., Wu, K., Guo, J., 2021. Nanoscale liquid hydrocarbon adsorption on clay minerals: A molecular dynamics simulation of shale oils. Chem. Eng. J. 420, 127578. https://doi.org/10.1016/j.cej.2020.127578.

Cui, J., Cheng, L., 2017. A theoretical study of the occurrence state of shale oil based on the pore sizes of mixed gaussian distribution. Fuel 206, 564–571. https://doi.org/10.1016/j.fuel.2017.06.047.

Fotos, G., Campbell, A., Murray, P., Yakushina, E., 2023. Deep learning enhanced watershed for microstructural analysis using a boundary class semantic segmentation. J. Mater. Sci. 58, 14390–14410. https://doi.org/10.1007/s10853-023-08901-w.

Garum, M., Glover, P.W., Lorinczi, P., Drummond-Brydson, R., Hassanpour, A., 2020. Micro- and nano-scale pore structure in gas shale using Xμ-CT and FIB-SEM techniques. Energy Fuels 34, 12340–12353. https://doi.org/10.1021/acs.energyfuels.0c02025.

Gong, Y., Liu, S., Zhu, R., Liu, K., Tang, Z., Jiang, L., 2015. Micro-occurrence of Cretaceous tight oil in southern Songliao Basin, NE China. Petrol. Explor. Dev. 42, 323–328. https://doi.org/10.1016/s1876-3804(15)30022-7.

Goral, J., Andrew, M., Olson, T., Deo, M., 2020. Correlative core- to pore-scale imaging of shales. Mar. Petrol. Geol. 111, 886–904. https://doi.org/10.1016/j.marpetgeo.2019.08.009.

Guo, Q., Wang, Y., Yang, S., Xiang, Z., 2022. A method of blasted rock image segmentation based on improved watershed algorithm. Sci. Rep. 12. https://doi.org/10.1038/s41598-022-11351-0.

Hu, C., Zhao, Z., Gao, S., Liu, C., Wu, K., Pang, P., 2024. Characterization of pore systems in fine-grained carbonate rocks using digital core technology. Adv. Geo-Energy Res. 12, 77–80. https://doi.org/10.46690/ager.2024.04.07.

Kuznik, F., Obrecht, C., Rusaouen, G., Roux, J.J., 2010. LBM based flow simulation using GPU computing processor. Comput. Math. Appl. 59, 2380–2392. https://doi.org/10.1016/j.camwa.2009.08.052.

Lawal, A., Yang, Y., He, H., Baisa, N.L., 2024. Machine learning in oil and gas exploration: A review. IEEE Access 12, 19035–19058. https://doi.org/10.1109/ACCESS.2023.3349216.

Lei, J., Pan, B., Guo, Y., Fan, Y., Xue, L., Deng, S., Zhang, L., Ruhan, A., 2021. A comprehensive analysis of the pyrolysis effects on oil shale pore structures at multiscale using different measurement methods. Energy 227, 120359. https://doi.org/10.1016/j.energy.2021.120359.

Li, J., Wang, M., Jiang, C., Lu, S., Li, Z., 2022. Sorption model of lacustrine shale oil: Insights from the contribution of organic matter and clay minerals. Energy 260, 125011. https://doi.org/10.1016/j.energy.2022.125011.

Li, X., Cai, J., Liu, H., Zhu, X., Li, Z., Liu, J., 2020. Characterization of shale pore structure by successive pretreatments and its significance. Fuel 269, 117412. https://doi.org/10.1016/j.fuel.2020.117412.

Lin, S., Hou, L., Luo, X., 2022a. Shale mineralogy analysis method: Quantitative correction of minerals using QEMSCAN based on MAPS technology. Appl. Sci. 12. https://doi.org/10.3390/app12105013.

Lin, S., Hou, L., Luo, X., Wu, Y., 2022b. A millimeter-scale insight into formation mechanism of lacustrine black shale in tephra deposition background. Sci. Rep. 12. https://doi.org/10.1038/s41598-022-15715-4.

Lin, S., Li, H., Luo, X., Zhang, Y., Zhang, L., 2023. An experimental investigation into the elastic and thermal properties of oil shales: Yanchang Formation, Ordos Basin. Mar. Petrol. Geol. 158, 106530. https://doi.org/10.1016/j.marpetgeo.2023.106530.

Lin, S., Luo, X., Zhao, Z., Hou, L., Li, S., Li, H., Pang, Z., Zhang, L., 2025. Grading evaluation methods for lacustrine shale oil resources developed by in-situ heating conversion technology. Front. Earth Sci. 12. https://doi.org/10.3389/feart.2024.1518236.

Lu, C., Jiang, H., Yang, J., Wang, Z., Zhang, M., Li, J., 2022. Shale oil production prediction and fracturing optimization based on machine learning. J. Pet. Sci. Eng. 217, 110900. https://doi.org/10.1016/j.petrol.2022.110900.

Mcmahon, T.P., Larson, T.E., Zhang, T., Shuster, M., 2024. Geologic characteristics, exploration and production progress of shale oil and gas in the United States: An overview. Petrol. Explor. Dev. 51, 925–948. https://doi.org/10.1016/s1876-3804(24)60516-1.

Neil, C.W., Hjelm, R.P., Hawley, M.E., Watkins, E.B., Cockreham, C., Wu, D., Mao, Y., Fischer, T.B., Stokes, M.R., Xu, H., 2020. Small-angle neutron scattering (SANS) characterization of clay- and carbonate-rich shale at elevated pressures. Energy Fuels 34, 8178–8185. https://doi.org/10.1021/acs.energyfuels.0c01009.

Ponomarev, A., Kadyrov, M., Tugushev, O., Drugov, D., Vaganov, Y.V., Leontiev, D., Zavatsky, M., 2022. Digital core reconstruction research: Challenges and prospects. Geol. Ecol. Landsc. 8, 49–56. https://doi.org/10.1080/24749508.2022.2086201.

Rowsell, J.L.C., Spencer, E.C., Eckert, J., Howard, J.A.K., Yaghi, O.M., 2005. Gas adsorption sites in a large-pore metal-organic framework. Science 309, 1350–1354. https://doi.org/10.1126/science.1113247.

Silvoster, M.L., Mathusoothana, R., Kumar, S., 2021. Watershed based algorithms for the segmentation of spine MRI. Int. J. Inf. Technol. 14, 1343–1353. https://doi.org/10.1007/s41870-021-00644-8.

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., Oza, H., 2021. Application of machine learning and artificial intelligence in oil and gas industry. Pet. Res. 6, 379–391. https://doi.org/10.1016/j.ptlrs.2021.05.009.

Sun, Z., Yin, Y., Wu, Y., Sun, Z., Zhu, L., Zhan, Y., Niasar, V., An, S., 2023. Morphological and hydrodynamic properties of hydrates during dissociation in sediment. Fuel 353, 129032. https://doi.org/10.1016/j.fuel.2023.129032.

Tariq, Z., Aljawad, M.S., Hasan, A., Murtaza, M., Mohammed, E., El-Husseiny, A., Alarifi, S.A., Mahmoud, M., Abdulraheem, A., 2021. A systematic review of data science and machine learning applications to the oil and gas industry. J. Pet. Explor. Prod. Technol. 11, 4339–4374. https://doi.org/10.1007/s13202-021-01302-2.

Wang, H., Ma, F., Tong, X., Liu, Z., Zhang, X., Wu, Z., Li, D., Wang, B., Xie, Y., Yang, L., 2016. Assessment of global unconventional oil and gas resources. Petrol. Explor. Dev. 43, 925–940. https://doi.org/10.1016/S1876-3804(16)30111-2.

Wang, M., Wang, J., Pan, N., Chen, S., 2007. Mesoscopic predictions of the effective thermal conductivity for microscale random porous media. Phys. Rev. 75, 036702. https://doi.org/10.1103/physreve.75.036702.

Wang, M., Zhang, S., Zhang, F., Liu, Y., Guan, H., Li, J., Shao, L., Yang, S., She, Y., 2015. Quantitative research on tight oil microscopic state of Chang 7 Member of Triassic Yanchang Formation in Ordos Basin, NW China. Petrol. Explor. Dev. 42, 827–832. https://doi.org/10.1016/s1876-3804(15)30079-3.

Wang, Y., Blunt, M.J., Armstrong, R.T., Mostaghimi, P., 2021. Deep learning in pore scale imaging and modeling. Earth Sci. Rev. 215, 103555. https://doi.org/10.1016/j.earscirev.2021.103555.

Watson, A., Vaughan, R., Powell, M., 1992. Classification using the watershed method. Int. J. Rem. Sens. 13, 1881–1890. https://doi.org/10.1080/01431169208904237.

Xu, Y., Lun, Z., Pan, Z., Wang, H., Zhou, X., Zhao, C., Zhang, D., 2022. Occurrence space and state of shale oil: A review. J. Pet. Sci. Eng. 211, 110183. https://doi.org/10.1016/j.petrol.2022.110183.

Yang, G., Liu, T., Lu, X., Wang, M., 2024. Fast-QSGS: A GPU accelerated program for structure generation of granular disordered media. Comput. Phys. Commun., 109241. https://doi.org/10.1016/j.cpc.2024.109241.

Yang, Y., Wang, K., Zhang, L., Sun, H., Zhang, K., Ma, J., 2019. Pore-scale simulation of shale oil flow based on pore network model. Fuel 251, 683–692. https://doi.org/10.1016/j.fuel.2019.03.083.

Yang, Y., Liu, J., Yao, J., Kou, J., Li, Z., Wu, T., Zhang, K., Zhang, L., Sun, H., 2020. Adsorption behaviors of shale oil in kerogen slit by molecular simulation. Chem. Eng. J. 387, 124054. https://doi.org/10.1016/j.cej.2020.124054.

Yin, Y., Sun, Z., 2024. A multiscale approach for assessing shale oil availability: Digital core, molecular simulation, and machine learning analysis. SPE J. 29 (6), 3033–3042. https://doi.org/10.2118/219475-PA.

Yin, Y., Sun, Z., Song, W., Sun, Z., Ren, Z., 2023. Investigation of natural gas hydrate microscopic occurrence types and pore-scale flow simulation based on digital cores. Energy Fuels 37, 16434–16445. https://doi.org/10.1021/acs.energyfuels.3c01500.

Zhang, H., Ahmed, M., Zhan, J.H., 2022a. Recent advances in molecular simulation of oil shale kerogen. Fuel 316, 123392. https://doi.org/10.1016/j.fuel.2022.123392.

Zhang, P., Yin, Y., Lu, S., Li, J., Chang, X., Zhang, J., Pang, Y., Chen, G., Liu, Y., Li, Z., 2022b. Insights into pore structures and multifractal characteristics of shale oil reservoirs: A case study from Dongying Sag, Bohai Bay Basin, China. Energy Fuel. 36, 8224–8237. https://doi.org/10.1021/acs.energyfuels.2c01763.

Zhang, W., Feng, Q., Wang, S., Zhang, J., Jin, Z., Xia, T., Xing, X., Lv, P., 2022c. Pore network modeling of oil and water transport in nanoporous shale with mixed wettability. J. Pet. Sci. Eng. 209, 109884. https://doi.org/10.1016/j.petrol.2021.109884.