Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR038379

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

A Novel Explainable Hybrid Model for Permeability Prediction of Tight Sandstone Using PI-DeepFM Machine Learning Algorithm

Lina Huo[1,2], Fujie Jiang[1,2], Liu Cao[1,2], Di Chen[1,2], Renda Huang[3], Xiaowei Zheng[1,2], Guangjie Zhao[1,2], Weiye Chen[1,2], Zezhang Song[1,2], Libin Zhao[4], and Yuanyuan He[4]

[1]National Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum (Beijing), Beijing, China, [2]College of Geosciences, China University of Petroleum (Beijing), Beijing, China, [3]Institute of Energy, Peking University, Beijing, China, [4]Taxinan Exploration and Development Company, PetroChina Tarim Oilfield Company, Kolar, China

Abstract Permeability is essential in controlling hydrocarbon fluid distribution and production in tight sandstone reservoirs. However, most conventional permeability prediction models are low-accuracy and cannot evaluate the importance of factors quantitatively because of the complex interplay among them, stemming from the complex pore structure of tight sandstone. Although machine learning (ML) constitutes an advanced solution, existing ML models often neglect the importance of explainable feature selection and interactions, thereby significantly restricting accuracy and reliability. Here, we proposed an explainable hybrid ML model, namely, PI-DeepFM, for permeability prediction. In this model, the Permutation Importance (PI) algorithm for feature selection is incorporated with Deep-Learning Neural Network and Factorization Machine (DeepFM) algorithm for feature interaction and prediction. To establish a more accurate feature set for prediction, we conducted various experiments, identified Shen & Li's fractal model as the best method for characterizing tight sandstone pore structure, and analyzed the factors influencing permeability through geological and correlation mathematical analysis methods. The quantitative feature evaluation results of the PI-DeepFM revealed that the fractal dimension, derived from Shen & Li's fractal model, influenced permeability the most, followed by *skewness*, *Clay*, *porosity*, *Carbonate* and *sorting coefficient*. This indicates that fluid seepage is dominated by pore structure complexity and connectivity. The PI-DeepFM model demonstrated a remarkable reduction of 25.76% in the root mean square error compared with that of conventional methods.

Key Points:

- The fractal dimension calculated using Shen & Li's model optimally quantifies the pore structure of tight sandstone
- The fractal dimension is the most important factor for permeability prediction
- The explainable hybrid model outperforms traditional models, providing insights into complex and hidden feature interaction predictions

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

F. Jiang,
jfjhtb@163.com

Citation:

Huo, L., Jiang, F., Cao, L., Chen, D., Huang, R., Zheng, X., et al. (2025). A novel explainable hybrid model for permeability prediction of tight sandstone using PI-DeepFM machine learning algorithm. *Water Resources Research*, *61*, e2024WR038379. https://doi.org/10.1029/2024WR038379

Received 21 AUG 2024
Accepted 21 JUN 2025

Author Contributions:

Conceptualization: Lina Huo
Data curation: Liu Cao
Funding acquisition: Fujie Jiang, Di Chen
Investigation: Weiye Chen
Methodology: Lina Huo, Fujie Jiang
Project administration: Fujie Jiang, Zezhang Song

1. Introduction

Tight sandstone gas, a clean and green unconventional fossil fuel, accounts for approximately 75% of the total unconventional natural gas resources, playing a critical role in meeting the growing global energy demands (Jia et al., 2022; Qiao et al., 2020; Sun et al., 2013; Worden et al., 2018). Permeability is a macroscopic parameter that describes the pore structure of porous media and governs fluid distribution and production in tight sandstone reservoirs (Behrang et al., 2016; X. Chen & Yao, 2017; Golab et al., 2010; X. Zhou et al., 2024). Permeability prediction based on its relationship with other petrophysical properties or pore structure parameters has emerged as a key focus in reservoir evaluation research, particularly for tight sandstone (R. Guo et al., 2020; Rezaee et al., 2012). Nevertheless, the heterogeneous and complex pore structure of tight sandstone, which results in various factors affecting permeability and intricate interactions among them, presents challenges in establishing an accurate permeability prediction model (Qiao et al., 2022).

Traditional permeability prediction models based on Poiseuille's equation or Darcy's law, such as the Winland and Pittman models, are usually employed to predict permeability and are commonly applied in conventional carbonate and sandstone reservoirs (Y. Guo et al., 2022). Tight sandstone exhibits a complicated pore structure and high heterogeneity, with its permeability resulting from the comprehensive influence of multiple geological factors. Therefore, a single or several feature parameters in empirical models may not fully reflect the seepage characteristics or might not be the critical factors influencing permeability. These will lead to inapplicability and large errors in prediction (Al-Kharra'a et al., 2023; R. Guo et al., 2020; Liu et al., 2018; Qiao et al., 2020). Previous studies have revealed that the main influencing factors include (a) macro petrophysical parameters like porosity (Dou et al., 2021; Z. Li et al., 2020; F. Wang & Wang, 2022); (b) experimental quantitative characterization parameters of pore structure, including high pressure mercury intrusion (HPMI) parameters and fractal dimension

Resources: Libin Zhao, Yuanyuan He
Software: Liu Cao
Supervision: Fujie Jiang, Di Chen, Renda Huang, Xiaowei Zheng, Guangjie Zhao, Weiye Chen, Zezhang Song, Libin Zhao, Yuanyuan He
Validation: Lina Huo, Liu Cao
Visualization: Lina Huo, Liu Cao
Writing – original draft: Lina Huo
Writing – review & editing: Lina Huo, Di Chen, Renda Huang, Xiaowei Zheng, Guangjie Zhao

(*D*) (Dai et al., 2021; Lai et al., 2018; F. Wang et al., 2019); and (c) the mineral composition, of which different mineral types and genetic mechanisms affect the pore size, shape and distribution, in turn influencing permeability (Qiao et al., 2022; Q. Zhang et al., 2022; F. Zhu et al., 2018). Although mercury intrusion porosity (MIP)—based models have been extensively used for pore structure characterization, their optimal selection remains understudied, particularly in the Bozi-Dabei area of the Tarim Basin, which is critical for precise pore structure characterization (Dou et al., 2021; K. Li, 2010; Z. Zhang & Weller, 2014). To construct a more accurate feature set for prediction, it is essential to characterize the pore structure better, analyze the factors affecting permeability at greater depths, and quantitatively assess their importance.

Machine learning (ML) has been extensively applied in geological research due to process data efficiently and its ability to capture non-linear relationships (Elmorsy et al., 2022; Jiang et al., 2023; Ma et al., 2022; L. Wang et al., 2020, 2021). ML provides a promising solution to quantitative feature importance and prediction problems with complex feature interactions. Given the diversity and complexity of influencing factors, effective feature selection to eliminate redundancy and irrelevance is essential before predicting. The traditional feature selection methods, such as Pearson correlation (Pearson, 1901) and Principal Component Analysis (PCA) (Jolliffe, 2002), are more suitable for linear relationships, whereas the Spearman correlation (Spearman, 1904) is better suited for monotonic relationships among features and is robust to outliers. With the continuous development of ML, linear regression algorithms, such as Lasso regression (Hastie et al., 2015), achieve sparse feature selection through regularization, but they are still limited to linear modeling. Tree-based algorithms like Decision Trees (DT) (L. Breiman et al., 1984), Random Forests (RF) (L. Breiman, 2001; Strobl et al., 2008), Extreme Gradient Boost (XGBoost) (T. Chen & Guestrin, 2016), and LightGBM (Ke et al., 2017) can more effectively capture nonlinear relationships between features and target variables (Kuhn & Johnson, 2018; Strobl et al., 2008). However, these models have varying approaches and sensitivities to feature handling, necessitating the selection of feature selection methods that conform to the predictive model (Guyon et al., 2002; Kuhn & Johnson, 2018). And the regular prediction algorithms face specific challenges, such as the need for pre-training in Factorized Neural Network (FNN), limitations in capturing low-order feature interactions in Product-based Neural Network (PNN), the necessity for manual feature engineering in Wide & Deep models, and the risk of overfitting in Random Forest (RF), especially with an excessive number of trees, which can limit their ability to capture complex feature interactions effectively. In conclusion, existing ML-based permeability prediction models exhibit problems including a lack of effective feature selection, reliance on certain prior knowledge, limited interpretability, and deficiencies in mining feature interaction information, which affect their trustworthiness and accuracy (L. Breiman, 2001; Cortes & Vapnik, 1995; Mahdaviara et al., 2021; Rasmussen & Williams, 2005; Wold et al., 1987; W. Zhang et al., 2021). There is a pressing need for advanced, explainable ML algorithms that can quantify the importance of factors to eliminate redundant and irrelevant features and effectively mine intricate interactions.

Here, we propose an explainable hybrid ML model (PI-DeepFM) for permeability prediction, which incorporates the Permutation Importance (PI) algorithm with the Deep-Learning Neural Network & Factorization Machine (DeepFM) algorithm. The PI algorithm, which is well-known for its robust interpretability and suitability for structured small sample data and any model, enables feature selection via the computation of the importance score of each feature (Altmann et al., 2010; Fisher et al., 2018). The DeepFM algorithm can be used to simultaneously mine linear, pairwise, and high-order feature interactions without manual feature engineering (H. Guo et al., 2017). It represents the cutting-edge in recommendation systems and has demonstrated exceptional performance across domains such as information retrieval, agriculture, and meteorology (Khaki et al., 2020; Yi et al., 2022). To our knowledge, this study presents the first application of this algorithm in the geological field. In this study, we applied the hybrid ML model to the Cretaceous Bashijiqike Formation (BS Fm.) in the Bozi-Dabei area, Tarim Basin. We conducted varied experimentsto characterize the pore structure, identified the optimal model from six fractal models and analyzed the factors influencing the tight sandstone permeability by correlation analysis with geological knowledge to construct a more accurate feature set. The feature set was input into the hybrid ML model to analyze the importance of each feature quantitatively to implement feature screening and feature interaction for prediction. This study provides new insights and has reference significance for geological feature prediction problems in the geological exploration and development process with various features, hidden feature interaction, difficult prior identification, and tough modeling. We believe that it is necessary and promising to introduce more advanced explainable feature selection and feature interactive ML algorithms into geological exploration and development.

2. Fractal Models

Fractal theory is applicable to the study of irregular and complex natural phenomena. According to fractal geometry, the number of pores $N(r)$ exceeding a certain pore radius r follows a power-law dependence on r (Mandelbrot, 1982):

$$N(>r) = \int_{r}^{r_{\max}} P(r)\,dr \propto r^{-D} \tag{1}$$

where $r_{\max}$ represents the maximum pore radius, and $P(r)$ denotes the distribution density function of pore radius.

Both the volume fractal dimension (D_v) and surface fractal dimension (D_s) can quantitatively characterize the microscopic pore structure characteristics. Specifically, D_v quantitatively characterizes the space filling characteristics of the pore, while D_s quantifies the roughness of the interface between the solid and pore spaces (Dou et al., 2021; F. Wang et al., 2019; L. Wang et al., 2021). In this study, we employed six widely used fractal models, covering both classes, as shown in Table S1 in Supporting Information S1. The D of rock porous media typically ranges from 2 to 3. The D nearer to 3 corresponds to rougher pore surface, more irregular pore shape, and a higher abundance of smaller pores (Cui et al., 2022; Su et al., 2018; N. Zhou et al., 2020; H. Zhu et al., 2019).

2.1. Model I

According to Equation 1, the pore structure distribution can be expressed as:

$$\frac{dN(>r)}{dr} = ar^{-D-1} \tag{2}$$

where a is a constant.

The cumulative volume with radius below r may be calculated by:

$$V(<r) = \int_{r_{\min}}^{r} ar^{-D-1}Fr^3dr = b\left(r^{3-D} - r_{\min}^{3-D}\right) \tag{3}$$

where F and b are constant, the shape factor F, which equals 1 for cubic pores, quantifies pore geometry $r_{\min}$ and $r_{\max}$ represent the smallest and largest pore radius, respectively.

Similarly, for pores with radius above r, the total volume is given by:

$$V(>r) = \int_{r}^{r_{\max}} ar^{-D-1}Fr^3dr = b\left(r_{\max}^{3-D} - r^{3-D}\right) \tag{4}$$

For the wetting phase cumulative saturation (gS_{1-S_H}) with r can be calculate by:

$$S_{1-S_{Hg}} = \frac{V(<r)}{V} = \frac{r^{3-D} - r_{\min}^{3-D}}{r_{\max}^{3-D} - r_{\min}^{3-D}} \tag{5}$$

When $r_{\min} \ll r_{\max}$, it can be simplified as

$$S_{1-S_{Hg}} = \frac{r^{3-D}}{r_{\max}^{3-D}} \tag{6}$$

According to Equation 17, it can be transformed into:

$$S_{1-S_{Hg}} = \frac{P_c^{D-3}}{P_{cd}^{D-3}} \tag{7}$$

where P_{cd} is the displacement pressure, and P_c is the external pressure.

The D is determined from the slope of logarithmic plot of $S_{1-S_{Hg}}$ versus P_c (He & Hua, 1998), by $D = 3 + \text{Slop}$. D calculated by this model is denoted as $D1$.

2.2. Model II

The relationship between S_{Hg} and V_{Hg} for MIP can be calculated as follows:

$$S_{Hg} = \frac{V_{Hg}}{V_{\text{pore}}} \tag{8}$$

where V_{pore} represents the total pore volume of the porous medium; V_{Hg} represents the cumulative mercury volume.

Substituting Equation 8 into Equations 15 and 1, we obtain (K. Li, 2010; Shen et al., 1995):

$$S_{Hg} \propto P_c^{D-2} \tag{9}$$

Therefore, the D is calculated by the slope of $\log(S_{Hg})$ versus $\log(P_c)$, by $D = 2 + \text{Slop}$. D computed using this model is named as $D2$.

2.3. Model III

D_s can be obtained from the pore volume and pore radius according to (Friesen & Mikula, 1987):

$$\text{Log}\left(\frac{-dS_{Hg}}{dr}\right) \sim (2-D)\log(r) \tag{10}$$

Combining Equation 1, the inverse relationship between r and Pc, Equation 10 can be simplified as:

$$\log\left(\frac{dS_{Hg}}{dP_c}\right) \sim (D-4)\log(P_c) \tag{11}$$

The D of the porous medium surface is computed from the slope of $\log(dS_{Hg}/dP_c)$ versus $\log(P)_c$ of which $D = 4 + \text{Slop}$. D calculated by this model is denoted as $D3$.

2.4. Model IV

B. Zhang and Li (1995) developed a fractal model grounded in the energy balance between surface energy increase during nonwetting phase displacement and external work required for mercury intrusion. The surface dimension can be determined by:

$$\ln\left(\frac{W_n}{r_n^2}\right) = C + D\ln\left(\frac{V_n^{\frac{1}{3}}}{r_n}\right) \tag{12}$$

where C is a constant; Wn is the intruded work, calculated by $W_n = \sum_{i-1}^{n} P_i \Delta V_i$ with P_i and V_i correspond to the external pressure and intruded mercury volume at each step i, respectively; V_n represents the cumulative mercury volume intruded into pores of radius r. D derived from the slope of $\ln\left(W_n/r_n^2\right)$ versus $\ln\left(V_n^{\frac{1}{3}}/r_n\right)$, by $D = \text{Slop}\, D$ calculated by this model is denoted as $D4$.

2.5. Model V

Neimark (1990) introduced a method to calculated D_s using the thermodynamic relationship between interfacial surface area and adsorbed mass, which does not use any adsorption model on fractal surfaces. We can calculate

the solid/liquid interface area S as a function of Pc from the Rootare-Prenzlow equation (Rootare & Prenzlow, 1967). By combining with Equation 17, the D calculation proceeds as Equation 14:

$$S(P_c) = \left(\frac{1}{\gamma \cos\theta}\right) \int_0^V P_c dV \tag{13}$$

$$D = 2 + \frac{d[\log S(P_c)]}{d \log P_c} \tag{14}$$

where γ represents the liquid surface tension; θ denotes the wetting angle, and V corresponds to the cumulative intruded mercury volume. The D is determined from the slope of $\log S(P_c)$ versus $\log P_c$, and $D = 2 + \text{Slop}$. D calculated by this model is denoted as *D5*.

2.6. Model VI

Assuming pores as a capillary bundle model (cylindrical pores with length l and radius r), $N(r)$ is calculated using Equation 15 (Shen & Li, 1994):

$$N(r) = \frac{V_{Hg}(r)}{\pi r^2 l} \tag{15}$$

where $V_{Hg}(r)$ denotes the cumulative mercury volume with pore radius r; $N(r)$ is the number of pores modeled as cylindrical capillaries with radius r, following the distribution described by Equation 1. The D is calculated by the incline of $\log(N(r))$ versus $\log(r)$ by $D = -\text{Slop}$. This model-derived dimension is designated as *D6*.

2.7. Multifractal

Multifractal can quantitatively describe the distribution of a physical quantity within an irregular fractal space/set and represents a heterogeneous collection of subsets with varying fractal behaviors (Paz Ferreiro & Vidal Vázquez, 2010). The strong heterogeneity of rock porous media may lead to differences of D in different pore size segments. Using the segmental regression method based on pore size distribution (PSD) to calculate D_i can better reflect the partial feature (P. Li et al., 2017; J. Wu et al., 2019). And the total D of the whole PSD can be calculated using Equation 16:

$$D = \sum_{i=1}^{n} \frac{\varphi_i}{\varphi_{\text{total}}} D_i \tag{16}$$

here, n is the number of fractal intervals divided according to the PSD; φ_i denotes the porosity (%) of various pore size ranges; D_i defines the fractal dimension of the corresponding interval, ranging from 2 to 3, as discussed above; φ_{total} is the total porosity of the rock. In this study, this method is specifically employed for samples exhibiting multiple fractals. We calculated the total fractal dimension and discussed its relation to the petrophysical properties and pore structure parameters of tight sandstone. This approach ensures completeness and accuracy in calculating D.

3. Materials and Methods

3.1. Sampling

The Kuqa Depression, a Mesozoic-Cenozoic superimposed foreland basin, is located in the northern Tarim Basin near the South Tianshan foothills (X. Guo et al., 2016). The Kelasu structural belt is the first row of thrust belts in the South Tianshan piedmont. To the west lies the Bozi-Dabei area, an imbricated structural belt formed under the strong thrust and nappe, covering an area of 2,070 km^2 (Tian et al., 2020), with relatively limited exploration so far. Achieving breakthroughs in natural gas exploration and development in this area is essential for ensuring long-term stable gas supply for the China West-East Gas Pipeline Project. All samples were collected from the BS

Fm. in Bozi-Dabei area. The BS Fm. is the main producing formation, lying at a significant depth of 5,000–8,000 m (details in Figure S1 in Supporting Information S1).

3.2. Experiments

For all test samples, chloroform and distilled water were used to remove residual bitumen and salt and finally dried in a drying cabinet at 105°C for 24 hr. The irregular cutting blocks, excised from both ends of the core, were utilized for microscopic observation and X-ray diffraction analysis (XRD). The regular plunger was used for physical property analysis, and subsequently, a 1 cm section of the regular plunger excised for the HPMI experiment.

For most solid interfaces, mercury exhibits a non-wetting behavior. Thus, *Pc* (external pressure) is required to penetrate hydrophilic rock pores. The attraction of mercury to the surface of porous media is equal to the circumference (supposing a circular pore). The equilibrium, when external force matches resistance, can be described as Washburn (1921):

$$-\pi \mathrm{D}\gamma \cos\theta = \frac{\pi \mathrm{D}^2 P_c}{4} \tag{17}$$

Simplified as:

$$\mathrm{W} = -\frac{4\gamma \cos\theta}{P_c} \tag{18}$$

where γ is 485 dyne cm^{-1} for air/mercury system; θ is 140° for mercury/rock system; W is the pore diameter. The instrument operates at a maximum mercury injection pressure of 200 MPa, enabling measurement of pore radius down to 0.004 μm. Due to the low compressible of tight sandstone, which contains more rigid components compared to shale and coal, this study omits compressibility corrections.

Pcd is defined as the mercury pressure at the intersection of the tangent to the inflection point of the mercury curve and the *y*-axis. The volume of mercury injection into pores larger than the pore size corresponding to *Pcd* is regarded as rock surface volume and can be ignored. Pittman (1992) employed the hyperbolic graphical method, utilizing the vertex as an effective indicator, while Winland (Kolodzie, 1980) defined the pore radius *R35*, at 35% mercury saturation, to evaluate reservoir physical properties. Detailed illustrations are provided in Figure S2 in Supporting Information S1.

3.3. Algorithms

3.3.1. Permutation Importance

There are many feature parameters affecting permeability, so convincing feature selection to obtain more important features is crucial to enhance the prediction model (Chandrashekar & Sahin, 2014; Guyon & Elisseeff, 2003; Kuhn & Johnson, 2018). In recent years, the PI algorithm has garnered significant attention as a model-agnostic feature selection approach, outlined in Equation 19 (Altmann et al., 2010; Fisher et al., 2018). This algorithm evaluates feature importance through randomly permuting feature values and monitoring the impact on model performance. It boasts strong interpretability and generous applicability, making it suitable for structured, small-sample data and any model. Specifically, for each feature, the algorithm generates a randomly permuted version and replaces the original feature with this permuted one, observing changes in model performance. A significant decrease in model performance indicates the importance of that feature for prediction. Compared to other commonly used feature selection algorithms, the PI algorithm solely relies on validation results and requires no model modifications, offering greater flexibility and convenience. The adoption of the PI algorithm for feature selection enables a more comprehensive and precise evaluation of feature importance, providing valuable support for subsequent model training and prediction.

$$i_j = s - 1/K \sum_{k=1}^{K} s_{k,j} \tag{19}$$

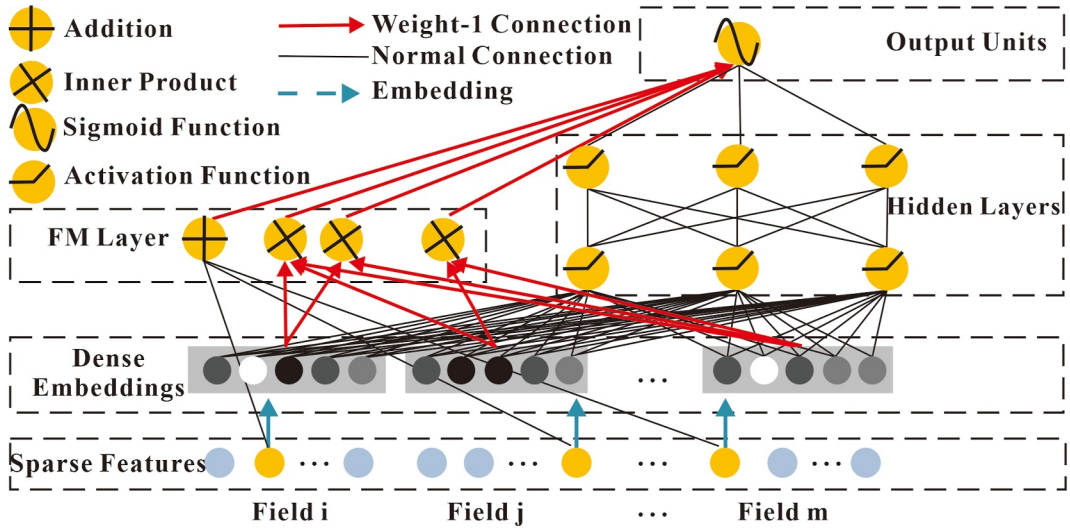

Figure 1. Wide & deep architecture of DeepFM. Both components share identical raw feature inputs, enabling simultaneous learning of low- and high-order feature interaction information directly from the original features (H. Guo et al., 2017).

where j is the feature index in the feature tabular data set A. The importance of feature j, denotes as i_j, is computed, and s represents the compute reference score of the validation model m on data A. For each repetition k in 1… K, column j of data set A is randomly shuffled to create a corrupted version, designated as $\tilde{D}_{k,j}$, and $s_{k,j}$ is the compute score of model m on corrupted data $\tilde{D}_{k,j}$.

3.3.2. DeepFM

The DeepFM model is a neural network framework that integrates deep neural networks with factorization machines (H. Guo et al., 2017). It's built upon the Wide & Deep structure, using FM to replace the LR of the Wide module, avoiding the complex feature engineering of manual construction. FM extracts features by the inner product of hidden variables of each dimension feature, having strong interpretability. The FM component in DeepFM simultaneously models linear (order-1) and pairwise (order-2) feature interactions (Rendle, 2010). The deep component, a feed-forward neural network, is designed for mining high-order feature interaction information. Notably, the FM and Deep modules share a common feature embedding part, enabling mutual optimization of low- and high-order feature representations through backpropagation (Figure 1 and Table 1). This joint training strategy enhances representation learning precision by synergizing both interaction types. And the structures of FM, Deep and embedding layer show in Figures S3, S4, and S5 in Supporting Information S1, respectively. Compared to other deep models like FNN, PNN and Wide & Deep, DeepFM can more effectively capture low- and high-order feature interactions simultaneously without the requirement for pre-training or manual feature engineering (Cheng et al., 2016; H. Guo et al., 2017; Rendle, 2010). The prediction function of DeepFM is:

$$\hat{y} = \text{sigmoid}\left(y_{\text{FM}} + y_{\text{DNN}}\right) \tag{20}$$

Table 1
Comparison of the Models for Prediction

Model	Linear (order-1)	Pairwise (order-2)	High-order	No manual feature engineering
Classical models	√	×	×	×
FM	√	√	×	×
DNN	×	×	√	×
DeepFM	√	√	√	√

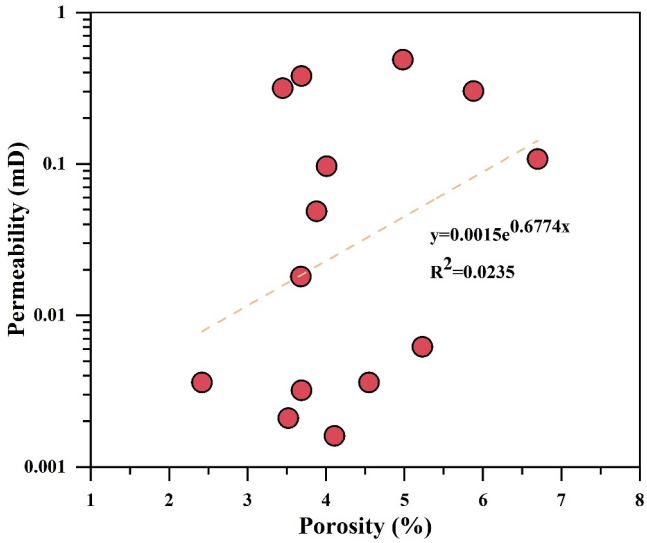

Figure 2. Porosity shows a poor correlation with permeability.

where $\hat{y}$ is the prediction permeability, y_{FM} is the output of the FM component, and y_{DNN} is the output of the deep component, with a sigmoid function applied for prediction.

FM Model Equation: The equation for a factorization machine with a degree of $d = 2$ is defined as:

$$y_{FM} = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=j+1}^{n} <v_i, v_j> x_i x_j \quad (21)$$

The parameters of the model that must be assessed are:

$$w_0 \in R, w \in R, V \in R^{n \times k} \quad (22)$$

And $<\cdot,\cdot>$ denotes the dot product of two vectors of size k:

$$<v_i, v_j> = \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f} \quad (23)$$

A row v_i within V describes the ith variable with k factors. $k \in N_0^+$ is a hyperparameter that determines the dimensionality of the factorization. w_0 represents the global bias; w_i models the strength of the ith variable. $<v_i, v_j>$ represents the inner product of eigenvector v_i and eigenvector v_j; Vi is the hidden factor vector of x_i feature; v_j is the hidden factor vector of x_j feature. The relationship between features is represented by the dimension of the latent factor vector, which is essentially an embedding representation of the features. According to Cholesky decomposition, for any symmetric positive semi-definite matrix W, as long as k is large enough, there must be a matrix V such that $W = V \cdot V^T$ can approximate any second-order cross feature in this way. Thus, FM more effectively learns feature interactions that are rare or absent in the training data and captures order-2 feature interaction information better than former methods.

4. Results

4.1. Petrophysical Properties

The petrophysical parameters indicate that the BS Fm. is a typical tight sandstone reservoir, with porosity (φ) and permeability (K) values ranging from 2.42% to 6.70% (average of 4.27%) and 0.0016 mD to 0.486 mD (average of 0.13 mD), respectively. The porosity and permeability are weakly correlated, with the permeability varying by two orders of magnitude among core samples with similar porosities (Figure 2), indicating that the porosity alone impacts the permeability to a limited extent (Cai et al., 2019).

The reservoir quality index (RQI) and the flow zone indicator (FZI) were computed via Equations 24 and 25, respectively (Abbaszadeh et al., 1996; Amaefule et al., 1993), with detailed results in Table S2 in Supporting Information S1. The RQI and FZI are key parameters in hydraulic flow unit theory and can serve to classify highly heterogeneous reservoirs. The larger their values are, the better the reservoir quality (P. Zhao et al., 2017).

$$RQI = 0.0314 \times \sqrt{\frac{K}{\varphi}} \quad (24)$$

$$FZI = RQI \times \frac{100 - \varphi}{\varphi} \quad (25)$$

4.2. Pore Space Characteristics

4.2.1. Pore Types

The casting thin section and SEM results (as shown in Figure S6 in Supporting Information S1) indicate that the tight sandstone contained various types and multi-scale pores, such as residual primary intergranular pores,

dissolution pores, intercrystalline pores and microfractures. The dominant pore types are residual intergranular and intergranular dissolution pores. The residual intergranular pores exhibited simple geometric shapes, typically triangular, polygonal or long strip, with radius generally exceeding 10 μm. They develop between skeleton grains like quartz and feldspar grains and are often filled with secondary authigenic quartz, authigenic feldspar and I/S mixed layer, resulting in relatively isolated pores with poor connectivity. Intragranular dissolution pores, formed by partial dissolution in feldspar and debris grains, are usually irregular polygons and relatively small in size, whereas the mold pores formed by complete dissolution can reach tens of microns. Intergranular dissolution pores, resulting from the dissolution of feldspar and debris, exhibit complex, irregular and varied geometries. Their size depends on the dissolution intensity, and the connectivity of the dissolution pores is relatively good. Intercrystalline pores, occurring within authigenic clay minerals such as illite and I/S mixed layers, range from several to hundreds of nanometers in size, with geometric shape determined by the mineral type, distribution and arrangement. In general, the connectivity between the pores on the plane is poor, and connected by sheet-shaped and curved-sheet-shaped throats. Various and multistage microfractures are present in some samples, including structural fractures formed by tectonic stress, grain breaking cracks and grain edge dissolution cracks. Although they contribute little to pore space, unfilled microfractures can significantly enhance permeability. Owing to the strong heterogeneity of tight sandstone, thin-section analysis alone provides limited insight into pore types and space characteristics, but cannot fully capture three-dimensional pore characteristics and real PSDs. Additionally, the pore types and distributions observed on the plane under the microscope also demonstrate strong heterogeneity. Thus, while thin-section analysis offers valuable reference, it cannot serve as the sole standard to judge the impact of pore type and structure on physical properties.

4.2.2. Pore Structure

Table S2 in Supporting Information S1 shows the HPMI experimental results. The *Pcd* shows relatively high values varying between 0.14 and 5.51 MPa (average of 1.43 MPa). The median radius (*R50*), defined as the radius at 50% mercury injection saturation, varies from 0.006 to 0.67 μm (average 0.22 μm). Sorting coefficient (*Sp*) ranges from 1.76 to 3.23 (average 2.63). In addition, the samples have relatively high maximum mercury saturation (*Smax*) (40.41%–96.96%/70.92%), but low mercury removal efficiency (*We*) (28.05%–58.54%/42.13%), and a large amount of mercury remained. These indicate poor sorting, distinct pore-throat distribution, and strongly heterogeneous and complex pore structure. The pore-throat distribution curve of tight sandstone obtained by HPMI analysis is shown in Figure S7 in Supporting Information S1. There are two distribution types, bimodal (Figure S7b in Supporting Information S1) and bias to unimodal (Figure S7c in Supporting Information S1), with maximum pore size below 2 μm. Furthermore, the macroscopic petrophysical properties like porosity, permeability, RQI and FZI show poor correlation with the parameters derived from HPMI (Figure 3). Therefore, a more effective approach for establishing tight reservoir classification and evaluation criteria involves the integrated analysis of multiple factors, including PSD, pore type and pore geometry characteristics.

Type I reservoir exhibits permeability greater than 0.01 mD, RQI greater than 0.02 (average 0.06), and FZI greater than 0.55 (average 1.43), indicating relatively good macroscopic physical properties. Most sample platform areas are narrow with relatively large slopes. It has relatively high *Pcd* (average 2.12 MPa), relatively low *Smax* (average 59.61%), and small *R50* and average pore radius (*Rave*) (averages of 0.038 and 0.18 μm, respectively). The PSD curve is bimodal, with peaks between 0.003 and 0.008 μm and 0.08 and 0.7 μm. The development of dissolution pores, including intragranular and intergranular types with complex geometric shapes, as well as grain edge dissolution fractures, observed under microscope. Some samples develop fractures running through thin section, including open, half-open or closed fractures.

Type II reservoir has permeability less than 0.01 mD, RQI below 0.012 (average 0.009), and FZI under 0.49 (average 0.24), suggesting poor macroscopic physical properties. Most sample platform areas are relatively wider with gentler slopes. It features relatively lower *Pcd* (average 0.52 MPa), higher *Smax* (average 86.01%), and larger *R50* and *Rave* (averages of 0.34 and 0.47 μm, respectively). The PSD curve bias to unimodal with peak distribution ranging from 0.2 to 1 μm. Under the microscope, intergranular pores with regular geometric shapes are relatively developed, and most of them exist relatively isolated on the plane or are connected by sheet-like throats.

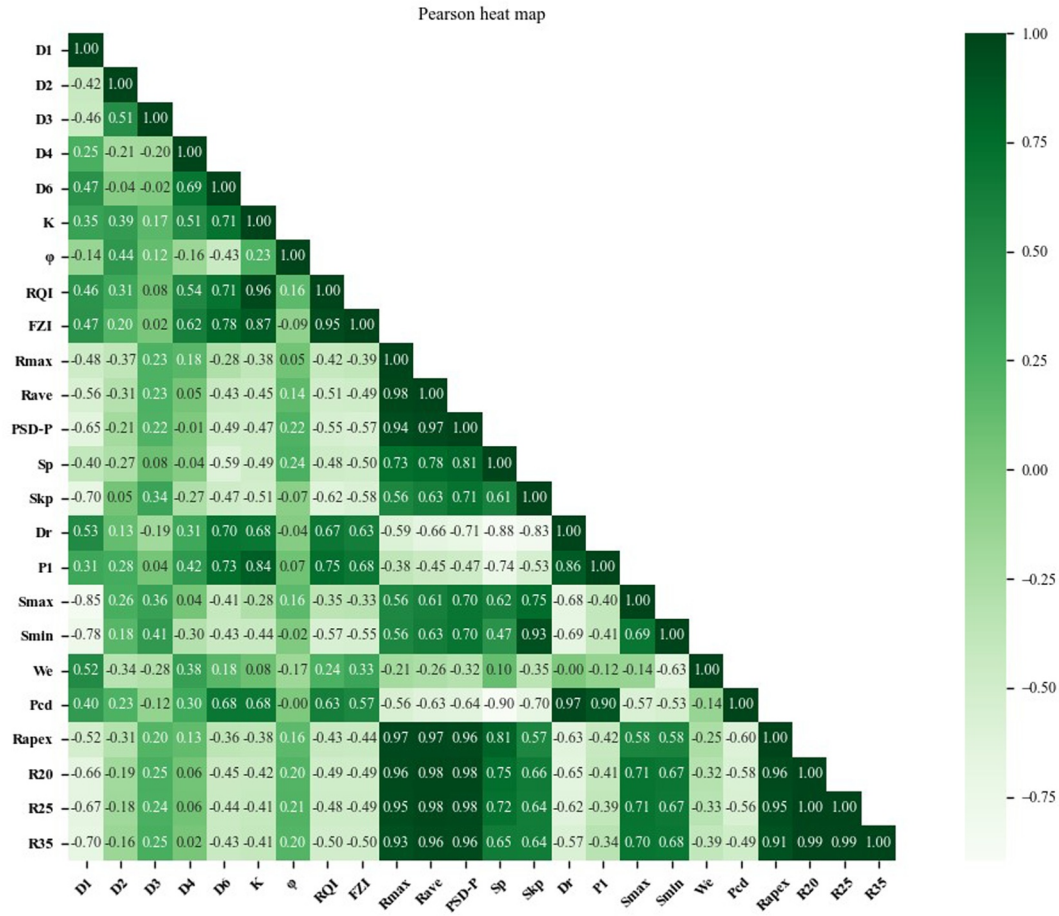

Figure 3. Heatmap reflecting the relationship among fractal dimension, HPMI parameters and macro petrophysical parameters. Note that *PSD-P* is the radius of pore throat distribution peak; *Dr* represents the relative sorting coefficient; *P1* is the characteristic structure coefficient; *Smin* represents the final remaining mercury saturation. *Notes.* The Pearson correlation coefficient (|R|) are categorized as: no ($|R| < 0.01$), negligible ($0.01 \leq |R| < 0.2$), weak ($0.2 \leq |R| < 0.3$), moderate ($0.3 \leq |R| < 0.4$), strong ($0.4 \leq |R| < 0.7$), and very strong ($0.7 \leq |R| \leq 1$) association (L. Wang & Fu, 2018).

4.3. Fractal Dimension

The D calculations utilize data points above P_{cd}. Table S3 in Supporting Information S1 and Figures S8, S9, and S10 in Supporting Information S1 detail the results from the above six fractal models.

4.3.1. Two-Regions Fractal Models

For most samples, two fractal regions can be identified via model I. Region I, corresponding to large pores, exhibits D ranging from 2.59 to 2.93. Region II, corresponding to small pores, shows D between 1.15 and 2.95, with some not having physical meaning. The total fractal dimension *D1* distribution spans 2.58 to 2.89.

Model II also reveals two distinct regions. For region I, the calculated D ranges from 2.39 to 3.92, with some not having physical meaning. For region II, the calculated D ranges from 2.09 to 2.29. The total fractal dimension *D2* is varied between 2.12 and 2.38.

Model III also yields two fractal regions. For region I, the calculated D is between 1.70 and 2.94, whereas for region II, D ranges from 2.64 to 4.13. Some samples, such as S6 and S2, have no fractal geometric meaning for the large or small pores. The total fractal dimension *D3* ranges between 2.01 and 2.91.

4.3.2. Single-Region Fractal Models

Fractal models IV, V and VI are all satisfied in the whole range of mercury injection characterization apertures, and all samples get only one fractal dimension, *D4* and *D6* ranges from 2.55 to 2.84, and 2.13 to 2.38, respectively, and have real physical meaning. However, the *D5* of all samples are greater than 3 (ranging from 3.06 to 3.20), thus, the fractal results have no physical meaning.

5. Discussions

5.1. Differences Among Fractal Models

All *D5* values are greater than 3, indicating that the fractal results have no real physical meaning, so it is deemed inapplicable and excluded from further discussion. Both models IV and VI recognize only one fractal region, *D4* and *D6* can represent the entire pore size range measured by HPMI. And they are significantly positively correlated (Figure 3), indicating their consistency in characterizing reservoir pore structure heterogeneity. Furthermore, basically all samples exhibit obvious inflection points, making fractal models I, II, and III effective for distinguishing different fractal regions, which is consistent with former research (Cui et al., 2022; Dou et al., 2021). We calculate *D* of Regions I and II, distinguished through segmental regression applied to PSD and obtain the total value using Equation 16. As shown in Figure 3, correlation analysis reveals that *D3* is positively correlated with *D2*, whereas *D1* is negatively correlated with *D2*. This indicates that using *D* derived from model I, model II or III to analyze rock pore structure and physical properties may lead to opposite conclusions. A higher *D6* value signifies a complex pore structure, lower porosity, higher permeability, and elevated FZI and RQI values; in addition, the correlation between *D6* and physical properties is the best. In contrast, *D1*, *D2*, and *D3* show no significant or only weak correlations with the macroscopic physical parameters. In summary, *D6* can effectively describe the fractal characteristics across the entire PSD and shows the best correlation with reservoir physical properties. Therefore, it is the optimal choice for evaluating tight sandstone reservoirs physical properties.

5.2. Factors Influencing Permeability

Permeability is the comprehension reflect of multiple geological factors. It is necessary to deeply analyze the factors influencing permeability before modeling (R. Guo et al., 2020; Oluwadebi et al., 2019). The results of correlation analysis of the HPMI parameters, fractal dimension, main mineral types, and permeability and porosity, are shown in Figures 3 and 4: (a) Permeability is correlated with certain HPMI experimental parameters. For example, *Rave* and *Skp* are negatively correlated with permeability, indicating that the more the PSD is biased toward smaller pore sizes, the higher the permeability; while *P1* is positively correlated with permeability, reflecting the better the degree of pore-throat sorting and pore-throat connectivity, the better the pore structure and the higher the permeability, that is, permeability is affected by pore size and distribution (Cai et al., 2019; F. Wang et al., 2019). (b) *D6* is positively correlated with permeability for type I reservoir and negatively correlated for type II reservoir, with a similar pattern observed for porosity. For type I reservoir, the presence of dissolution pores with complex geometric but relatively better connectivity and fractures leads to a higher *D* and consequently higher permeability, although the porosity inclines to decrease. Conversely, for type II reservoirs, which are dominated by intergranular pores with relatively poor connectivity, a higher *D* indicates a more complex pore structure, resulting in lower permeability and porosity (Lai et al., 2016). (c) The type and genetic mechanism of minerals affect pore characteristics (e.g., type, geometry and size distribution), thereby affecting reservoir physical properties. Researchers have investigated the relationships among mineral composition, microscopic pore structure and fractal dimension (Al-Kharra'a et al., 2023; Qiao et al., 2022; S. Zhao et al., 2022), demonstrating that distinct mineral compositions exert varying controls on these parameters, so we will not elaborate here. Minerals and physical properties exhibit complex and diverse relationships, even with opposing trends observed in different reservoirs types (Figure 4). In Type I reservoirs, an increase in carbonate mineral indicates enhanced dissolution and the potential presence of microfractures filled with carbonate mineral, leading to increased permeability and porosity. In contrast, higher quartz mineral and clay mineral contents suggest relatively weaker dissolution and poorer pore connectivity, resulting in lower permeability and a tendency for the porosity to decrease. For Type II reservoirs dominated by intergranular pores, carbonate mineral primarily blocks pores and throats, belong to destructive diagenesis. Feldspar and quartz, which support intergranular pores

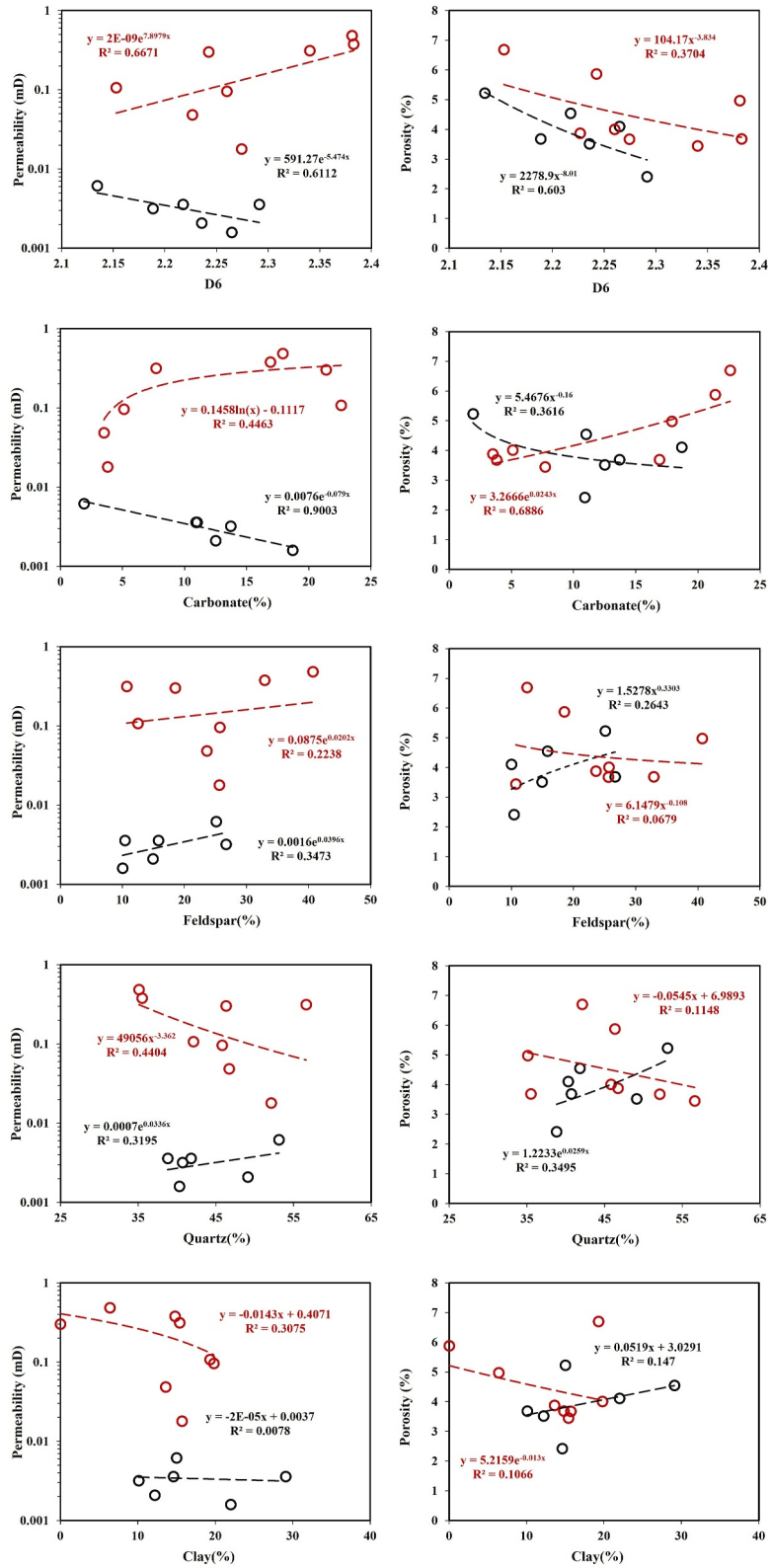

Figure 4. Correlation analysis of permeability, porosity with *D*6 and different minerals contents.

Table 2
The Results and Errors of Classic Permeability Prediction Models

Model	Equation	MSE (mD^2)	RMSE (mD)
Pittman	logK = 3.4699logφ—1.3524logRapex—4.3405	0.425	0.652
Winland(*R10*)	logK = 3.7789logφ—1.3399logR10–4.1792	0.375	0.612
Winland(*R15*)	logK = 3.9089logφ—1.3444logR15–4.3974	0.347	0.589
Winland(*R20*)	logK = 4.0795logφ—1.3208logR20–4.6299	0.326	0.571
Winland(*R25*)	logK = 4.2632logφ—1.2164logR25–4.8432	0.318	0.564
Winland(*R30*)	logK = 4.5027logφ—1.0021logR30–5.0482	0.320	0.566
Winland (*R35*)	logK = 5.0402logφ—0.8584logR35–5.4809	**0.312**	**0.559**
Winland(*R40*)	logK = 5.1349logφ—0.6969logR40–5.5627	0.338	0.581

Note. The bold values represent the results of the optimal model.

development, positively impact permeability. Consequently, as the carbonate mineral content decreases and the quartz and feldspar contents increase, permeability becomes higher and porosity tends to increase.

In summary, the factors affecting permeability are highly diverse and complexly interactive, making it challenging to characterize permeability fully and well through one or a few fixed parameters. Even though the interplay between factors influencing permeability has been meticulously analyzed and discussed, manual modeling remains challenging. Therefore, it is essential to employ feature interaction algorithms based on ML to automatically mine hidden geological information for more precise and intelligent permeability prediction modeling.

5.3. Winland and Pittman Permeability Prediction Models

Percolation theory establishes characteristic length as the dominant control on seepage properties (Katz & Thompson, 1986, 1987). It can be calculated through HPMI experiment (Kolodzie, 1980; Pittman, 1992). Pittman (1992) employed mathematical methods to confirm *Rapex* as the characteristic length controlling permeability, while Kolodzie (1980) proposed that the characteristic length matches the pore radius (*R35*) at 35% mercury saturation and provided empirical formulas. We obtained the pore radius at various mercury saturation levels of 10%, 15%, 20%, 25%, 30%, 35%, and 40% and denoted these values as *R10*, *R15*, *R20*, *R25*, *R30*, *R35*, and *R40*, respectively. On the basis of classic empirical formulas, the permeability is predicted through the multiple regression method based on the least square algorithm. The accuracy of each model is evaluated using the mean squared error (MSE) and the root mean squared error (RMSE) (Table 2). However, since the dimension of the RMSE is consistent with that of the predicted value, making it more intuitively interpretable and widely applicable in practical scenarios.

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (K_{\mathrm{exp}} - K_{\mathrm{pred}})^2}{n} \tag{26}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (K_{\mathrm{exp}} - K_{\mathrm{pred}})^2}{n}} \tag{27}$$

where K_{exp} is the experimentally measured permeability, and K_{pred} is the permeability predicted by the model.

The MSE is in a range from 0.312 to 0.425, the RMSE ranges from 0.559 to 0.652, and the Winland (*R35*) model has the smallest error among them. Possible explanations for the unsatisfactory prediction performance of the Pittman and Winland models are as follows: (a) As an important parameter of these models, the relationship between porosity and permeability is not clear in the study area. (b) A single characteristic length parameter cannot fully characterize the pore structure characteristics, and more parameters need to be considered for permeability prediction.

(a) Weight	Feature
0.4005 ± 0.0986	D6
0.1699 ± 0.2906	P1
0.0524 ± 0.0076	φ
0.0355 ± 0.0797	Dr
0.0175 ± 0.0146	Carbonate
0.0156 ± 0.0178	Clay
0.0074 ± 0.0139	Feldspar
0.0071 ± 0.0138	Pcd
0.0031 ± 0.0045	Sp
0.0028 ± 0.0343	R25
0.0019 ± 0.0032	Rapex
0.0014 ± 0.0085	R20
0.0002 ± 0.0039	Rave
0.0001 ± 0.0061	Quartz
-0.0127 ± 0.0767	Skp
-0.0209 ± 0.0318	PSD-P
-0.0217 ± 0.0726	Smin
-0.0228 ± 0.0255	Rmax

(b) Weight	Feature
2.1698 ± 1.2837	D6
0.3111 ± 0.3549	Carbonate
0.0449 ± 0.0449	Dr
0 ± 0.0000	R20
0 ± 0.0000	φ
0 ± 0.0000	Rmax
0 ± 0.0000	Rave
0 ± 0.0000	PSD-P
0 ± 0.0000	Skp
0 ± 0.0000	R25
0 ± 0.0000	Smin
0 ± 0.0000	Pcd
0 ± 0.0000	Quartz
0 ± 0.0000	Clay
0 ± 0.0000	Feldspar
0 ± 0.0000	Rapex
0 ± 0.0000	P1
-0.0251 ± 0.0615	Sp

(c) Weight	Feature
0.6672 ± 1.6621	D6
0.4944 ± 1.9970	Skp
0.3455 ± 0.4146	Clay
0.1559 ± 0.0416	φ
0.1266 ± 0.0960	Carbonate
0.0532 ± 0.0653	Sp
0 ± 0.0000	Smin
0 ± 0.0000	R20
0 ± 0.0000	R25
0 ± 0.0000	Pcd
0 ± 0.0000	P1
-0.0001 ± 0.0002	PSD-P
-0.0040 ± 0.0068	Dr
-0.0065 ± 0.0131	Rave
-0.0165 ± 0.0231	Quartz
-0.0174 ± 0.0454	Rmax
-0.0209 ± 0.0446	Feldspar
-0.0647 ± 0.1387	Rapex

(d) Weight	Feature
2.0396 ± 0.7039	D6
0.0038 ± 0.0006	φ
0 ± 0.0000	Smin
0 ± 0.0000	Rave
0 ± 0.0000	PSD-P
0 ± 0.0000	Sp
0 ± 0.0000	Dr
0 ± 0.0000	R20
0 ± 0.0000	R25
0 ± 0.0000	Pcd
0 ± 0.0000	Quartz
0 ± 0.0000	Clay
0 ± 0.0000	Carbonate
0 ± 0.0000	Feldspar
-0.0015 ± 0.0021	Rapex
-0.0030 ± 0.0047	Rmax
-0.0249 ± 0.0802	P1
-0.0336 ± 0.1813	Skp

Figure 5. The results of PI using different validation models (a) RF; (b) DT; (c) LightGBM; (d) XGBoost. The initial value in each row quantifies model performance attenuation. The values behind ± denotes the standard deviation across multiple tests. Negative values signify improved prediction accuracy when features are perturbed than the real data.

5.4. Explainable Hybrid ML Model

5.4.1. Feature Selection

Generally, features that are poorly correlated with the target variable will increase the prediction model error. Therefore, we first use the Pearson algorithm to conduct correlation analysis on numerous HPMI parameters, mineral contents, and fractal dimensions, combined with expert knowledge, and 18 features are screened out. Then we utilize the PI algorithm with four different validation models, RF, DT, LightGBM and XGBoost algorithms to evaluate the permutation feature importance, thereby achieving feature selection and dimensionality reduction (Figure 5). The RF and LightGBM validation models are the optimal validation algorithms for feature selection due to the consistency of the results and consistent with the geological cognition. Thus, we yield two optimal feature sets for prediction: *D6*, *P1*, *φ*, *Dr*, *Carbonate*, *Clay*, *Feldspar*, *Pcd*, *Sp*, *R25* (10 features) and *D6,*

Table 3
Errors of the Permeability Prediction Model

Model	MSE (mD^2)	RMSE (mD)	Error reduction percent
Winland (*R35*)	0.312	0.559	–
Linear (6 Features)	0.262	0.451	↓19.32%
SVR (6 Features)	0.256	0.500	↓10.55%
XGBoost (6 Features)	0.294	0.474	↓15.21%
LightGBM (6 Features)	0.319	0.488	↓12.70%
Radom Forest (6 Features)	0.245	0.448	↓19.86%
FM (6 Features)	0.288	0.536	↓4.11%
DNN (6 Features)	About 100	About 10	–
DeepFM (18 Features)	0.397	0.542	↓3.04%
PI-DeepFM (10 Features)	0.334	0.475	↓15.03%
PI-DeepFM (6 Features)	**0.240**	**0.415**	**↓25.76%**

Note. Error reduction percent: The percentage reduction of error (RMSE) between the three DeepFM models and the best classical Winland (*R*35) model. Unsatisfactory DNN. The bold values represent the results of the optimal model.

Skp, *Clay*, φ, *Carbonate Sp* (6 features). Notably, consistently ranked as the most influential parameter across the four validation models, with DT and XGBoost attributing importance results even greater than 1-a statistical anomaly underscoring its dominance to a certain extent. This aligns with the above analysis.

5.4.2. Prediction Model

Considering that there may be low- and high-order feature interaction and to assess the rationality and reliability of feature selection again, we apply the DeepFM algorithm using features after and before PI feature selection. A total of 50 rounds of training were conducted, with 100 iterations per round, and strictly apply the following criteria to select the best model: the losses of both the training and test sets must decrease simultaneously, and the difference between the two should be the smallest. The error value of the best model for 50 rounds does not fluctuate notably. The arithmetic mean of error values are shown in Table 3. The DeepFM model with 18 features achieves an RMSE of 0.542, representing a 3.04% improvement over the Winland (*R35*) model. After PI feature selection, the PI-DeepFM models demonstrate a significantly enhanced performance: the RF validation version (10 Features) reduces the RMSE to 0.475 (an improvement of 15.03%), whereas the LightGBM validation version (6 Features) achieves RMSE of 0.415 (an improvement of 25.76%). This progressive enhancement with feature reduction suggests improved model stability and convergence. In addition, the PI-DeepFM model outperforms classical empirical model, Liner model, SVR model, FM model, Tree-based models, and DNN model by effectively automated mining of both low- and high- order feature interactions without manual engineering. In conclusion, these results validate the distinctive advantages of DeepFM in modeling permeability under geologically complex and interaction-intensive conditions.

This approach offers a novel ML method and idea for permeability prediction of rock porous media with complex pore structure. Moreover, most geological features often interact and influence each other; thus, it is difficult to discover potential feature interaction information and model them in detail manually. Therefore, the introduction of advanced explainable feature selection and feature interactive information mining ML algorithms is necessary for geological exploration and development, and we believe that the PI-DeepFM algorithm has significant potential in the geological field.

6. Conclusions

1. *D6* has the best correlation with petrophysical properties and can be directly calculated to obtain a single fractal dimension representing the entire pore size range. Therefore, the use Model VI to characterize the pore structure of tight sandstone is recommended. A higher *D6* indicates a more complicated pore structure and a reduction in porosity, whereas permeability increases due to the development of microfractures and dissolution pores with complex geometric shapes.

2. *D6* can characterize the pore structure well. In addition, the quantitative evaluation results of the PI model indicate that the most important factor affecting permeability prediction is *D6*. Overall, the PI model results verify the geological understanding well and the model is explainable and quantitative.
3. We evaluate classical permeability prediction models, and the best model is the Winland (*R35*) model with an RMSE of 0.559. The RMSE of the PI-DeepFM (6 Features) model based on the PI-DeepFM algorithm is 25.76% lower than that of the Winland (*R35*) model, and obviously better than the FM and DNN models.
4. The novel hybrid explainable ML permeability prediction model we proposed provides a new way of thinking for geological feature prediction problems with feature interaction concealment, difficult prior identification, and challenging manual modeling. We believe that it is both necessary and promising to introduce more advanced explainable feature selection and feature interactive information mining AI algorithms into geological exploration and development.

Data Availability Statement

The data and code associated with this work are available at Huo (2024a, 2024b).

Acknowledgments
This work was supported by the National Natural Science Foundation of China (NSFC) (41872128), National Natural Science Foundation of China (NSFC) (42302142) and China Postdoctoral Science Foundation (2022M723487). We appreciate the Taxinan Exploration and Development Company, PetroChina Tarim Oilfield Company for providing rock samples. We also extend our gratitude to the reviewers for their valuable comments and suggestions.

References

Abbaszadeh, M., Fujii, H., & Fujimoto, F. (1996). Permeability prediction by hydraulic flow units - theory and applications. *SPE Formation Evaluation*, *11*(4), 263–271. https://doi.org/10.2118/30158-PA

Al-Kharra'a, H. S., Wolf, K.-H. A. A., AlQuraishi, A. A., Mahmoud, M. A., Deshenenkov, I., AlDuhailan, M. A., et al. (2023). Impact of clay mineralogy on the petrophysical properties of tight sandstones. *Geoenergy Science and Engineering*, *227*, 211883. https://doi.org/10.1016/j.geoen.2023.211883

Altmann, A., Tolosi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Amaefule, J. O., Altunbay, M., Tiab, D., Kersey, D. G., & Keelan, D. K. (1993). Enhanced reservoir description: Using core and log data to identify hydraulic (flow) units and predict permeability in uncored intervals/wells. *Paper presented at the SPE Annual Technical Conference and Exhibition*.

Behrang, A., Mohammadmoradi, P., Taheri, S., & Kantzas, A. (2016). A theoretical study on the permeability of tight media; effects of slippage and condensation. *Fuel*, *181*, 610–617. https://doi.org/10.1016/j.fuel.2016.05.048

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. *Biometrics*, *40*, 874. https://doi.org/10.1038/nmeth.4370

Cai, J., Zhang, Z., Wei, W., Guo, D., Li, S., & Zhao, P. (2019). The critical factors for permeability-formation factor relation in reservoir rocks: Pore-throat ratio, tortuosity and connectivity. *Energy*, *188*, 116051. https://doi.org/10.1016/j.energy.2019.116051

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, X., & Yao, G. (2017). An improved model for permeability estimation in low permeable porous media based on fractal geometry and modified Hagen-Poiseuille flow. *Fuel*, *210*, 748–757. https://doi.org/10.1016/j.fuel.2017.08.101

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., et al. (2016). Wide & deep learning for recommender systems.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cui, H., Zhu, S., Wang, J., Gao, Y., Wan, C., & Tong, H. (2022). Physical properties, pore-throat structure, fractal characteristics and their effects on the gas-bearing capacity of tight sandstone: A case study from the northern Tianhuan depression, Ordos Basin, China. *Natural Resources Research*, *31*(3), 1559–1584. https://doi.org/10.1007/s11053-022-10059-2

Dai, Q., Wang, G., Zhao, X., Han, Z., Lu, K., Lai, J., et al. (2021). Fractal model for permeability estimation in low-permeable porous media with variable pore sizes and unevenly adsorbed water lay. *Marine and Petroleum Geology*, *130*, 105135. https://doi.org/10.1016/j.marpetgeo.2021.105135

Dou, W., Liu, L., Jia, L., Xu, Z., Wang, M., & Du, C. (2021). Pore structure, fractal characteristics and permeability prediction of tight sandstones: A case study from Yanchang formation, Ordos Basin, China. *Marine and Petroleum Geology*, *123*, 104737. https://doi.org/10.1016/j.marpetgeo.2020.104737

Elmorsy, M., El-Dakhakhni, W., & Zhao, B. (2022). Generalizable permeability prediction of digital porous media via a novel multi-scale 3D convolutional neural network. *Water Resources Research*, *58*(3). https://doi.org/10.1029/2021wr031454

Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*. https://doi.org/10.48550/arXiv.1801.01489

Friesen, W. I., & Mikula, R. J. (1987). Fractal dimensions of coal particles. *Journal of Colloid and Interface Science*, *120*(1), 263–271. https://doi.org/10.1016/0021-9797(87)90348-1

Golab, A. N., Knackstedt, M. A., Averdunk, H., Senden, T., Butcher, A. R., & Jaime, P. (2010). 3D porosity and mineralogy characterization in tight gas sandstones. *The Leading Edge*, *29*(12), 1476–1483. https://doi.org/10.1190/1.3525363

Guo, H., Tang, R., Ye, Y., Li, Z., & He, X. (2017). DeepFM: A factorization-machine based neural network for CTR prediction. *Paper presented at the Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1725–1731. https://doi.org/10.24963/ijcai.2017/239

Guo, R., Xie, Q., Qu, X., Chu, M., Li, S., Ma, D., & Ma, X. (2020). Fractal characteristics of pore-throat structure and permeability estimation of tight sandstone reservoirs: A case study of Chang 7 of the Upper Triassic Yanchang Formation in Longdong area, Ordos Basin, China. *Journal of Petroleum Science and Engineering*, *184*, 106555. https://doi.org/10.1016/j.petrol.2019.106555

Guo, X., Liu, K., Jia, C., Song, Y., Zhao, M., Zhuo, Q., & Lu, X. (2016). Fluid evolution in the Dabei gas field of the Kuqa Depression, Tarim Basin, NW China: Implications for fault-related fluid flow. *Marine and Petroleum Geology*, *78*, 1–16. https://doi.org/10.1016/j.marpetgeo.2016.08.024

Guo, Y., Sima, L., Wang, L., & Guo, Y. (2022). Prediction of ultra-tight sandstone reservoir permeability by capillary pressure curve based on partial least squares regression method. *Petroleum Geology and Recovery Efficiency*, *29*(6), 67–76. https://doi.org/10.13673/j.cnki.cn37-1359/te.202108009

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422. https://doi.org/10.1023/A:1012487302797

Hastie, T. J., Tibshirani, R., & Wainwright, M. J. (2015). *Statistical learning with sparsity: The Lasso and generalizations*. CRC Press.

He, C., & Hua, M. (1998). Fractal geometry description of reservoir pore structure. *Oil & Gas Geology*, *19*(1), 17–25.

Huo, L. (2024a). Code for permeability prediction model [Software]. *figshare*. https://doi.org/10.6084/m9.figshare.25567236.v1

Huo, L. (2024b). Data availability [Dataset]. *figshare*. https://doi.org/10.6084/m9.figshare.25565973.v1

Jia, A., Wei, Y., Guo, Z., Wang, G., Meng, D., & Huang, S. (2022). Development status and prospect of tight sandstone gas in China. *Natural Gas Industry*, *42*(1), 83–92. https://doi.org/10.1016/S1876-3804(12)60026-3

Jiang, F., Guo, Y., Tsuji, T., Kato, Y., Shimokawara, M., Esteban, L., et al. (2023). Upscaling permeability using multiscale X-ray-CT images with digital rock modeling and deep learning techniques. *Water Resources Research*, *59*(3). https://doi.org/10.1029/2022wr033267

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Katz, A. J., & Thompson, A. H. (1986). Quantitative prediction of permeability in porous rock. *Physical Review B*, *34*(11), 8179–8181. https://doi.org/10.1103/PhysRevB.34.8179

Katz, A. J., & Thompson, A. H. (1987). Prediction of rock electrical conductivity from mercury injection measurements. *Journal of Geophysical Research*, *92*(B1), 599–607. https://doi.org/10.1029/JB092iB01p00599

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Paper presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Khaki, S., Khalilzadeh, Z., & Wang, L. (2020). Predicting yield performance of parents in plant breeding: A neural collaborative filtering approach. *PLoS One*, *15*(5), e0233382. https://doi.org/10.1371/journal.pone.0233382

Kolodzie, S., Jr. (1980). Analysis of pore throat size and use of the waxman-smits equation to determine ooip in spindle field, Colorado. *Paper presented at the SPE Annual Technical Conference and Exhibition*.

Kuhn, M., & Johnson, K. (2018). *Applied predictive modeling*. Springer.

Lai, J., Wang, G., Fan, Z., Chen, J., Wang, S., Zhou, Z., & Fan, X. (2016). Insight into the pore structure of tight sandstones using NMR and HPMI measurements. *Energy & Fuels*, *30*(12), 10200–10214. https://doi.org/10.1021/acs.energyfuels.6b01982

Lai, J., Wang, G., Fan, Z., Zhou, Z., Chen, J., & Wang, S. (2018). Fractal analysis of tight shaly sandstones using nuclear magnetic resonance measurements. *AAPG Bulletin*, *102*(2), 175–193. https://doi.org/10.1306/0425171609817007

Li, K. (2010). Analytical derivation of Brooks–Corey type capillary pressure models using fractal geometry and evaluation of rock heterogeneity. *Journal of Petroleum Science and Engineering*, *73*(1–2), 20–26. https://doi.org/10.1016/j.petrol.2010.05.002

Li, P., Zheng, M., Bi, H., Wu, S., & Wang, X. (2017). Pore throat structure and fractal characteristics of tight oil sandstone: A case study in the ordos basin, China. *Journal of Petroleum Science and Engineering*, *149*, 665–674. https://doi.org/10.1016/j.petrol.2016.11.015

Li, Z., Liu, D., Cai, Y., Wang, Y., & Si, G. (2020). Evaluation of coal petrophysics incorporating fractal characteristics by mercury intrusion porosimetry and low-field NMR. *Fuel*, *263*, 116802. https://doi.org/10.1016/j.fuel.2019.116802

Liu, M., Xie, R., Wu, S., Zhu, R., Mao, Z., & Wang, C. (2018). Permeability prediction from mercury injection capillary pressure curves by partial least squares regression method in tight sandstone reservoirs. *Journal of Petroleum Science and Engineering*, *169*, 135–145. https://doi.org/10.1016/j.petrol.2018.05.020

Ma, K., Pang, X., Pang, H., Lv, C., Gao, T., Chen, J., et al. (2022). A novel method for favorable zone prediction of conventional hydrocarbon accumulations based on RUSBoosted tree machine learning algorithm. *Applied Energy*, *326*, 119983. https://doi.org/10.1016/j.apenergy.2022.119983

Mahdaviara, M., Rostami, A., Keivanimehr, F., & Shahbazi, K. (2021). Accurate determination of permeability in carbonate reservoirs using Gaussian Process Regression. *Journal of Petroleum Science and Engineering*, *196*, 107807. https://doi.org/10.1016/j.petrol.2020.107807

Mandelbrot, B. B., & Wheeler, J. A. (1982). The fractal geometry of nature. *American Journal of Physics*, *51*(3), 286–287. https://doi.org/10.1119/1.13295

Neimark, A. V. (1990). Calculating surface fractal dimensions of adsorbents. *Adsorption Science and Technology*, *7*(4), 210–219. https://doi.org/10.1177/026361749000700402

Oluwadebi, A. G., Taylor, K. G., & Ma, L. (2019). A case study on 3D characterisation of pore structure in a tight sandstone gas reservoir: The Collyhurst Sandstone, East Irish Sea Basin, northern England. *Journal of Natural Gas Science and Engineering*, *68*, 102917. https://doi.org/10.1016/j.jngse.2019.102917

Paz Ferreiro, J., & Vidal Vázquez, E. (2010). Multifractal analysis of Hg pore size distributions in soils with contrasting structural stability. *Geoderma*, *160*(1), 64–73. https://doi.org/10.1016/j.geoderma.2009.11.019

Pearson, K. (1901). LIII. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pittman, E. D. (1992). Relationship of porosity and permeability to various parameters derived from mercury injection-capillary pressure curves for Sandstone1. *AAPG Bulletin*, *76*(2), 191–198. https://doi.org/10.1306/bdff87a4-1718-11d7-8645000102c1865d

Qiao, J., Zeng, J., Chen, D., Cai, J., Jiang, S., Xiao, E., et al. (2022). Permeability estimation of tight sandstone from pore structure characterization. *Marine and Petroleum Geology*, *135*, 105382. https://doi.org/10.1016/j.marpetgeo.2021.105382

Qiao, J., Zeng, J., Ma, Y., Jiang, S., Feng, S., & Hu, H. (2020). Effects of mineralogy on pore structure and fluid flow capacity of deeply buried sandstone reservoirs with a case study in the Junggar Basin. *Journal of Petroleum Science and Engineering*, *189*, 106986. https://doi.org/10.1016/j.petrol.2020.106986

Rasmussen, C. E., & Williams, C. K. I. (2005). Gaussian processes for machine learning. https://doi.org/10.7551/mitpress/3206.001.0001

Rendle, S. (2010). Factorization machines. *Paper presented at the 2010 IEEE International Conference on Data Mining*.

Rezaee, R., Saeedi, A., & Clennell, B. (2012). Tight gas sands permeability estimation from mercury injection capillary pressure and nuclear magnetic resonance data. *Journal of Petroleum Science and Engineering*, *88–89*, 92–99. https://doi.org/10.1016/j.petrol.2011.12.014

Rootare, H. M., & Prenzlow, C. F. (1967). Surface areas from mercury porosimeter measurements. *Journal of Physical Chemistry*, *71*(8), 2733–2736. https://doi.org/10.1021/J100867A057

Shen, P., & Li, K. (1994). A new method for determining the fractal dimensions of pore structures and its application. *Paper present at the 10th Offshore South East Asia Conference*.

Shen, P., Li, K., & Jia, F. (1995). Quantitative description for the heterogeneity of pore structure by using mercury capillary pressure curves. *Paper presented at the SPE International Meeting Held in Beijing, China*.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Su, P., Xia, Z., Qu, L., Yu, W., Wang, P., Li, D., & Kong, X. (2018). Fractal characteristics of low-permeability gas sandstones based on a new model for mercury intrusion porosimetry. *Journal of Natural Gas Science and Engineering*, *60*, 246–255. https://doi.org/10.1016/j.jngse.2018.10.008

Sun, L., Zou, C., Zhu, R., Zhang, Y., Zhang, S., Zhang, B., et al. (2013). Formation, distribution and potential of deep hydrocarbon resources in China. *Petroleum Exploration and Development*, *40*(6), 587–695. https://doi.org/10.1016/S1876-3804(13)60093-2

Tian, J., Yang, H., Wu, C., Mo, T., Zhu, W., & Shi, L. (2020). Discovery of Well Bozi 9 and ultra-deep natural gas exploration potential in the Kelasu tectonic zone of the Tarim Basin. *Natural Gas Industry*, *40*(1), 11–19.

Wang, F., & Wang, L. (2022). Pore structure analysis and permeability prediction of shale oil reservoirs with HPMI and NMR: A case study of the Permian Lucaogou formation in the Jimsar Sag, Junggar Basin, NW China. *Journal of Petroleum Science and Engineering*, *214*, 110503. https://doi.org/10.1016/j.petrol.2022.110503

Wang, F., Yang, K., You, J., & Lei, X. (2019). Analysis of pore size distribution and fractal dimension in tight sandstone with mercury intrusion porosimetry. *Results in Physics*, *13*, 102283. https://doi.org/10.1016/j.rinp.2019.102283

Wang, L., & Fu, X. (2018). Data-driven analyses of low salinity water flooding in sandstones. *Fuel*, *234*, 674–686. https://doi.org/10.1016/j.fuel.2018.07.063

Wang, L., Tian, Y., Yao, B., & Yu, X. (2020). Machine learning analyses of low salinity effect in sandstone porous media. *Journal of Porous Media*, *23*(7), 731–740. https://doi.org/10.1615/jpormedia.2020033000

Wang, L., Zeng, X., Yang, H., Lv, X., Guo, F., Shi, Y., & Hanif, A. (2021). Investigation and application of fractal theory in cement-based materials: A review. *Fractal and Fractional*, *5*(4), 247. https://doi.org/10.3390/fractalfract5040247

Washburn, E. W. (1921). The dynamics of capillary flow. *Physical Review*, *17*(3), 273–283. https://doi.org/10.1103/PhysRev.17.273

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Worden, R. H., Bukar, M., & Shell, P. (2018). The effect of oil emplacement on quartz cementation in a deeply buried sandstone reservoir. *AAPG Bulletin*, *102*(1), 49–75. https://doi.org/10.1306/02071716001

Wu, J., Fan, T., Gomez-Rivas, E., Gao, Z., Yao, S., Li, W., et al. (2019). Impact of pore structure and fractal characteristics on the sealing capacity of Ordovician carbonate cap rock in the Tarim Basin, China. *Marine and Petroleum Geology*, *102*, 557–579. https://doi.org/10.1016/j.marpetgeo.2019.01.014

Yi, X., Duan, Z., Li, R., Zhang, J., Li, T., & Zheng, Y. (2022). Predicting fine-grained air quality based on deep neural networks. *IEEE Transactions on Big Data*, *8*(5), 1326–1339. https://doi.org/10.1109/TBDATA.2020.3047078

Zhang, B., & Li, S. (1995). Determination of the surface fractal dimension for porous media by mercury porosimetry. *Industrial & Engineering Chemistry Research*, *34*(4), 1383–1386. https://doi.org/10.1021/ie00043a044

Zhang, Q., Liu, Y., Wang, B., Ruan, J., Yan, N., Chen, H., et al. (2022). Effects of pore-throat structures on the fluid mobility in chang 7 tight sandstone reservoirs of longdong area, Ordos Basin. *Marine and Petroleum Geology*, *135*, 135. https://doi.org/10.1016/j.marpetgeo.2021.105407

Zhang, W., Qin, J., Guo, W., Tang, R., & He, X. (2021). Deep learning for click-through rate estimation. In *Proceedings of the 30th international joint conference on artificial intelligence, IJCAI 2021, virtual event/Montreal, Canada, 19–27 August 2021* (pp. 4695–4703). ijcai.org.

Zhang, Z., & Weller, A. (2014). Fractal dimension of pore-space geometry of an Eocene sandstone formation. *Geophysics*, *79*(6), D377–D387. https://doi.org/10.1190/geo2014-0143.1

Zhao, P., Wang, Z., Sun, Z., Cai, J., & Wang, L. (2017). Investigation on the pore structure and multifractal characteristics of tight oil reservoirs using NMR measurements: Permian Lucaogou Formation in Jimusaer Sag, Junggar Basin. *Marine and Petroleum Geology*, *86*, 1067–1081. https://doi.org/10.1016/j.marpetgeo.2017.07.011

Zhao, S., Fu, Q., Fu, J., Liu, X., Li, S., Zhang, G., & Teng, J. (2022). Effect of authigenic clay minerals and carbonate cements on quality of tight sandstone reservoirs: Insight from Triassic tight sandstones in the Huaqing area, Ordos Basin, Northern China. *Journal of Asian Earth Sciences*, *229*, 105099. https://doi.org/10.1016/j.jseaes.2022.105099

Zhou, N., Lu, S., Wang, M., Liu, W., Guan, Y., Tan, H., & Wang, Z. (2020). Applicability of fractal capillary pressure models to sandstones. *Journal of Petroleum Science and Engineering*, *185*, 106626. https://doi.org/10.1016/j.petrol.2019.106626

Zhou, X., Wei, J., Zhao, J., Zhang, X., Fu, X., Shamil, S., et al. (2024). Study on pore structure and permeability sensitivity of tight oil reservoirs. *Energy*, *288*, 129632. https://doi.org/10.1016/j.energy.2023.129632

Zhu, F., Hu, W., Cao, J., Sun, F., Liu, Y., & Sun, Z. (2018). Micro/nanoscale pore structure and fractal characteristics of tight gas sandstone: A case study from the Yuanba area, northeast Sichuan Basin, China. *Marine and Petroleum Geology*, *98*, 116–132. https://doi.org/10.1016/j.marpetgeo.2018.08.013

Zhu, H., Zhang, T., Zhong, D., Li, Y., Zhang, J., & Chen, X. (2019). Binary pore structure characteristics of tight sandstone reservoirs. *Petroleum Exploration and Development*, *46*(6), 1297–1306. https://doi.org/10.1016/S1876-3804(19)60283-1