Petroleum Science 22 (2025) 3599–3612

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

Coaly source rock evaluation using well logs: The Jurassic Kezilenuer Formation in Kuqa Depression, Tarim Basin, China

Fei Zhao [a,b], Jin Lai [a,b,*], Zong-Li Xia [a,b], Zhong-Rui Wang [a,b], Ling Li [c], Bin Wang [c], Lu Xiao [a,b], Yang Su [a,b], Gui-Wen Wang [a,b]

[a] State Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum (Beijing), Beijing, 102249, China
[b] College of Geosciences, China University of Petroleum (Beijing), Beijing, 102249, China
[c] Research Institute of Petroleum Exploration and Development, Tarim Oilfield Company, CNPC, Korla, 841000, Xinjiang, China

ARTICLE INFO

Article history:
Received 5 January 2025
Received in revised form
14 April 2025
Accepted 29 May 2025
Available online 3 June 2025

Edited by Meng-Jiao Zhou

Keywords:
Source rock
Well logs
Kuqa Depression
Kezilenuer formation
Machine learning

ABSTRACT

Coaly source rocks have attracted considerable attention for their significant hydrocarbon generation potential in recent years. However, limited study is performed on utilizing geochemical data and well log data to evaluate coaly hydrocarbon source rocks. In this study, geochemical data and well log data are selected from two key wells to conduct an evaluation of coaly hydrocarbon source rocks of Jurassic Kezilenuer Formation in Kuqa Depression of Tarim Basin. Initially, analysis was focused on geochemical parameters to assess organic matter type, source rock quality, and hydrocarbon generation potential. Lithology types of source rocks include mudstone, carbonaceous mudstone and coal. The predominant organic matter type identified was Type III and Type II_2, indicating a favorable hydrocarbon generation potential. Well log data are integrated to predict total organic carbon (TOC) content, and the results indicate that multiple regression method is effective in predicting TOC of carbonaceous mudstone and coal. However, the ΔlgR method exhibited limited predictive capability for mudstone source rock. Additionally, machine learning methods including multilayer perceptron neural network (MLP), random forest (RF), and extreme gradient boosting (XGBoost) techniques are employed to predict TOC of mudstone source rock. The XGBoost performs best in TOC prediction with correlation coefficient (R^2) of 0.9517, indicating a close agreement between measured and predicted TOC values. This study provides a reliable prediction method of coaly hydrocarbon source rocks through machine learning methods, and will provide guidance for resource assessment.

1. Introduction

Source rocks are the essential petroleum systems element (Hunt, 1996; Sahoo et al., 2021). As global energy demand continues to grow, the exploration and development of coal-bearing formation have become increasingly important (Lee et al., 2022; Mkono et al., 2023). Therefore, evaluation of quality and spatial distribution of coaly source rocks is essential for resource assessment and petroleum system analysis (Bolandi et al., 2015; Goliatt et al., 2023). The content of total organic carbon (TOC) is a crucial parameter in source rock evaluation (Aziz et al., 2020; Goliatt et al., 2023; Lu et al., 2025). TOC not only reflects the hydrocarbon potential of source rocks, but also provides important inputs for petroleum system analysis (Hood et al., 1975; Mulashani et al., 2021; Gao et al., 2022; Gordon et al., 2022). Conventional methods for measuring TOC primarily rely on core analysis and geochemical experiments such as pyrolysis. Though these experiment approaches are commonly accurate, experiment analysis are limited to cored intervals, limiting the comprehensive assessment of vertical continuity and changes (Zhu et al., 2018; Zhao et al., 2019; Tenaglia et al., 2020; Sahoo et al., 2021). Additionally, pyrolysis is time consuming and expensive (Sahoo et al., 2021; Gao et al., 2022). As a result, the efficient and precise estimation of TOC has become a key area of interest in hydrocarbon source rock analysis (Bolandi et al., 2015; Wang et al., 2020).

* Corresponding author.
E-mail address: laijin@cup.edu.cn (J. Lai).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2025.05.029

In recent years, advancements in logging technology have sparked increasing interest in predicting TOC using well log data (Zhao et al., 2016; Zhu et al., 2018; Sêco et al., 2019; Tenaglia et al., 2020; Mkono et al., 2023). Well log is characterized by vertical continuity and high resolution, providing valuable information about underground lithology, fluids, TOC and other parameters (Mahmoud et al., 2017; Handhal et al., 2020; Mulashani et al., 2021; Zeng et al., 2021; Lee et al., 2022). However, relationship between well log data and TOC is complex, making it challenging for traditional linear regression and empirical formulas to capture its non-linear features (Zhu et al., 2018; Sêco et al., 2019; Nyakilla et al., 2022; Ochoa et al., 2022). For instance, Schmoker (1979) proposed a model for predicting organic carbon content based on density logging, while Passey et al. (1990) introduced the ΔlgR method for predicting organic carbon content (Passey et al., 1990). These widely adopted methods have shown limited applicability in coaly hydrocarbon source rocks compared to shallow buried sedimentary basins (Hu et al., 2015; Shi et al., 2016; Bolandi et al., 2017; Lai et al., 2022).

In recent studies, with the increasing complexity and demands of source rock evaluation, there has been a growing interest in combining machine learning or deep learning techniques with well logs to assess hydrocarbon source rock quality (Rui et al., 2020; Aziz et al., 2020; Maroufi and Zahmatkesh, 2023; Mkono et al., 2023). Numerous studies have highlighted the effectiveness and accuracy of machine learning techniques in predicting TOC content (Sfidari et al., 2012; Lee et al., 2022; Goliatt et al., 2023). Machine learning-based TOC prediction methods have the advantages of accuracy and efficiency than traditional approaches (El Sharawy and Gaafar, 2012; Mkono et al., 2023; Maroufi and Zahmatkesh, 2023; Goliatt et al., 2023). Algorithms such as back propagation (BP), group method of data handling (GMDH) neural network, linear regression (LR), random forest (RF), and deep learning (DL) have exhibited strong predictive performance for TOC prediction across various geological conditions (Zhu et al., 2018; Wang et al., 2019; Mulashani et al., 2021; Zhang et al., 2023; Zhang et al., 2023).

This study focuses on the Jurassic source rocks located in the Kuqa Depression of the Tarim Basin, China. The aim of this study is to optimize suitable methods for evaluating coaly hydrocarbon source rocks. The organic matter type and hydrocarbon generation potential are assessed from three different lithology types (mudstone, coal and carbonaceous mudstone) (Zhao et al., 2005; Huang et al., 2019). Various methods, including the multivariate regression, ΔlgR method, MLP, RF, and XGBoost model, are applied to predict TOC for different lithologies. This study will provide insights into the coaly hydrocarbon source rocks evaluation using well logs, and has implication for resource assessment and hydrocarbon exploration.

2. Geological setting

The Kuqa Depression, situated in the northern region of the Tarim Basin and adjacent to the Tianshan Mountains, and is a Mesozoic to Cenozoic foreland depression (Fig. 1) (Guo et al., 2018; Gao et al., 2022). The Kuqa Depression is categorized by four distinct structural belts: namely the Northern Monocline Belt, the Kelasu Structural Belt, the Qiulitage Structural Belt, and the Southern Gentle Slope (Fig. 1) (Zhao et al., 2005; Lai et al., 2017, 2023a). Additionally, it includes three sags: Wushi Sag, Baicheng

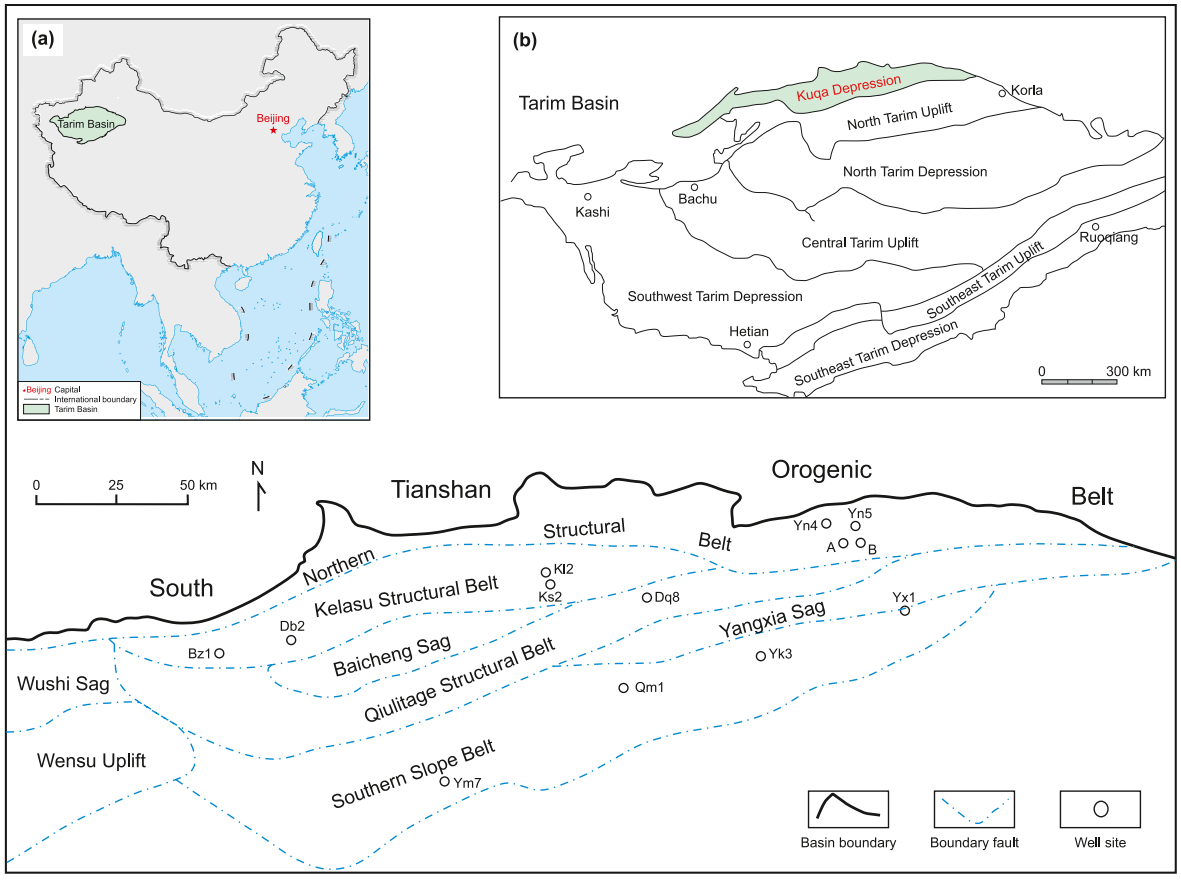

Fig. 1. Maps show the structural characteristics of the Kuqa Depression in the Tarim Basin of western China, where the study area is located.

Sag, and Yangxia Sag (Fig. 1) (Guo et al., 2018; Huang et al., 2019; Lai et al., 2023a).

During the Late Triassic to Middle Jurassic, the relatively warm and humid climate favored the formation of lacustrine and lacustrine–swamp transitional sediments, which were accumulated in the Kuqa Depression (Zhao et al., 2005; Lai et al., 2017; Huang et al., 2019). The source rocks in the Kuqa Depression mainly consist of six sets, including the Triassic Kelamayi ($T_{2\text{-}3}k$), Huangshanjie (T_3h), Taliqike Formation (T_3t), as well as the Jurassic Yangxia (J_1y), Kezilenuer (J_2kz), and Qiakemake Formation (J_2q) (Fig. 2) (Gao et al., 2022; Wan et al., 2022; Li et al., 2025). It is worth noting that the Karamay, Huangshanjie, and Qiakemake Formation are interpreted as lacustrine source rocks, while the Taliqike, Yangxia, and Kezilenuer Formation are coaly source rocks (Zhao et al., 2005; Guo et al., 2018; Gao et al., 2022). Laterally, the petroleum source rocks are characterized by their extensive distribution, substantial thicknesses (up to 320 m), and high organic matter content (average TOC up to 2.15%) in the Kuqa Depression (Zhao et al., 2005; Wang et al., 2022). The petroleum source rocks of the Kezilenuer Formation have attained an overall mature stage (average R_o > 0.7%), indicating that they are entering the oil generation phase (Guo et al., 2018).

This study focuses on source rocks of the Jurassic Kezilenuer Formation, which was deposited in a braided river delta-swamp depositional system, and contain coaly strata (Zhao et al., 2005; Gao et al., 2022). The lithology of Jurassic Kezilenuer source rocks comprises mudstone, carbonaceous mudstone, and coal (Fig. 2) (Zhao et al., 2005; Huang et al., 2019).

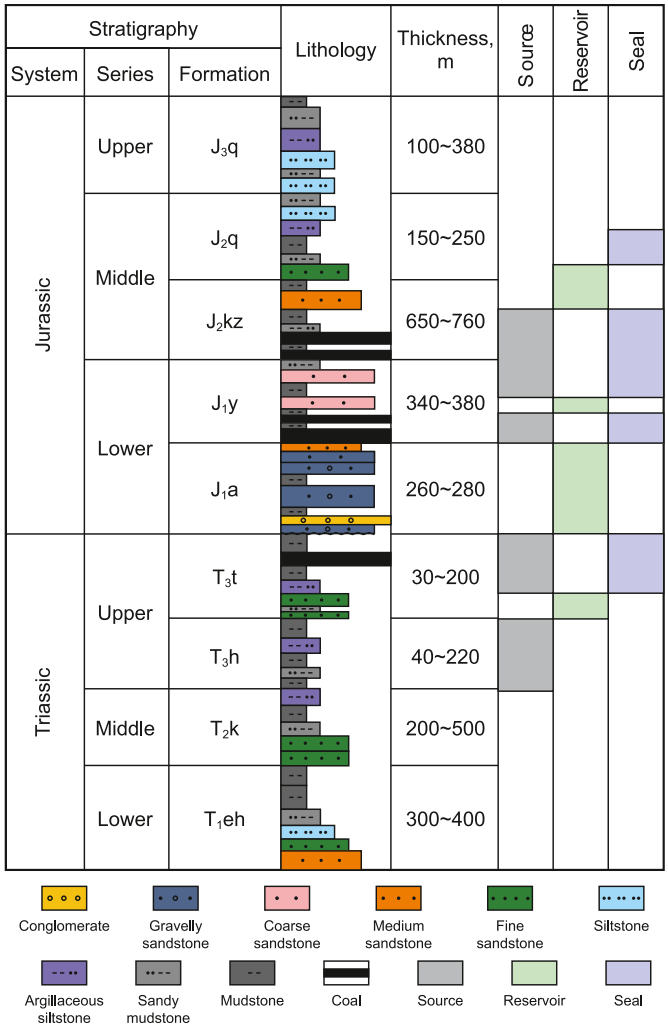

Fig. 2. Generalized Mesozoic stratigraphy of the Kuqa Depression, Tarim Basin, showing major oil and gas combinations.

3. Data and methods

This study evaluates the source rocks of the Jurassic Kezilenuer Formation using two primary datasets: geochemical analysis and well log data. Geochemical analysis provides key parameters of studied source rocks, while well log data is utilized for TOC prediction for intervals without core control. Then, a correlation analysis was conducted between TOC and logging parameters for various lithology types, followed by the utilization of multiple regression methods to predict the TOC content of source rocks. Finally, this study attempted to employ ΔlgR, MLP, RF, and XGBoost methods to predict total organic carbon content in mudstone while assessing the predictive effects of these diverse methodologies. We used the coefficient of determination (R^2) and root mean square error (RMSE) to evaluate the performance of the different models.

3.1. Data

A total of 103 samples from Wells A and B in the Kuqa Depression were collected to study the source rocks of the Kezilenuer Formation. The lithology of the samples includes mudstone, carbonaceous mudstone, and coal. Pyrolysis analysis was conducted using a Rock Eval OGE-V instrument under standard conditions, yielding parameters such as TOC, S1 (free hydrocarbon), S2 (pyrolysed hydrocarbon), T_{max} (temperature at the highest yield of S2) and hydrogen index (HI) (Alizadeh et al., 2018; Shalaby et al., 2019; Wang et al., 2022).

Geophysical well logs data include open-hole caliper (CAL), acoustic transit time logs (DT), deep and medium resistivity (RT, RM), natural gamma-ray (GR), bulk density (DEN), compensated neutron porosity (CNC), and spontaneous potential (SP).

3.2. ΔlgR method

The ΔlgR method predicts TOC by overlaying deep resistivity curve and porosity curve (Passey et al., 1990). The model is used to predict TOC in shallow clastic and carbonate rocks (Hu et al., 2015; Bolandi et al., 2017; Alizadeh et al., 2018; Lai et al., 2024). It calculates ΔlgR, which correlates linearly with TOC and varies with thermal maturity. The equations are as follows:

$$\Delta\mathrm{lgR} = \log(\mathrm{RT}/\mathrm{RT}_{\mathrm{Baseline}}) + k(\Delta t - \Delta t_{\mathrm{Baseline}}) \quad (1)$$

$$\mathrm{TOC} = \Delta\mathrm{lgR}\cdot 10^{(2.297-0.1688\mathrm{LOM})} \quad (2)$$

where RT is the resistivity log, Ω·m, and Δt is the sonic transit time log, μs/m; $\mathrm{RT}_{\mathrm{Baseline}}$ is the baseline of RT, Ω·m; and $\Delta t_{\mathrm{Baseline}}$ is the baseline of DT, μs/m; LOM, linked to thermal maturity, can be determined by the vitrinite reflectance; k is the coefficient (Passey et al., 1990; Aziz et al., 2020; Lai et al., 2022).

3.3. Multilayer perceptron neural network model (MLP)

The MLP is a type of feedforward neural network model that comprises an input layer, one or more hidden layers, and an output layer, with each layer containing interconnected neurons (Aziz et al., 2020; Goliatt et al., 2023). MLP model adjusts weights and thresholds through backpropagation to minimize error. Due to its adaptability and ability to model nonlinear relationships, MLP can

be employed in classification and prediction tasks (Bolandi et al., 2017; Liu et al., 2021; Zhang et al., 2023). Additionally, the input data includes six logging curves (GR, DT, CNC, RT, DEN, and SP), and the output data is the measured TOC (Fig. 3). In this study, a two-hidden-layer MLP was employed for TOC prediction (Fig. 3(a)).

3.4. Random forest (RF)

Random forest is an ensemble model composed of decision trees (Breiman, 2001). It uses Bootstrap resampling to create multiple subsets of the dataset, which are then used to construct decision trees through random feature selection (Gordon et al., 2022) (Fig. 3(b)). Predictions are aggregated through majority voting or averaging, enhancing robustness and reducing overfitting (Safaei-Farouji and Kadkhodaie, 2022). This approach improves prediction accuracy and is particularly effective for handling large datasets with complex relationships (Cappuccio et al., 2021).

3.5. Extreme gradient boosting (XGBoost)

XGBoost, proposed by Chen and Guestrin (2016), is a gradient boosting algorithm known for its high efficiency and flexibility. It constructs multiple weak learners iteratively, combining them into a strong predictive model. XGBoost reduces overfitting through regularization and handles large datasets with high computational efficiency (Liu et al., 2021). Its ability to model complex nonlinear relationships has been demonstrated in various regression and classification tasks (Liu et al., 2021). The model architecture used to predict TOC is shown in Fig. 3(c).

3.6. Workflow of TOC prediction

The workflow for TOC prediction is illustrated in Fig. 3. The process begins with extracting well log curves (GR, DT, CNC, RT, DEN, and SP) and TOC parameters derived from Pyrolysis. Subsequently, the dataset designated for TOC prediction is randomly divided into a training and a test subset, adhering to a 7:3 ratio. After normalization, the data are input into MLP, RF, and XGBoost models. A 5-fold cross-validation approach is utilized during training to enhance model generalization and minimize overfitting. Meanwhile, Bayesian optimization are used to optimize the hyperparameters for each model. Finally, the models are evaluated using the test set, and predictive results are generated.

4. Results

4.1. Organic matter type, source rock quantity, and hydrocarbon generation potential

The lithology types of the Kezilenuer Formation samples consist of mudstone, carbonaceous mudstone, and coal (Zhao et al., 2005; Huang et al., 2019). Pyrolysis analysis results are shown in Supplementary Table. Hydrogen index (HI) was plotted against the temperature at which S2 hydrocarbons yield is highest (T_{max}) to classify organic matter type categories. As shown in Fig. 4(a), the source

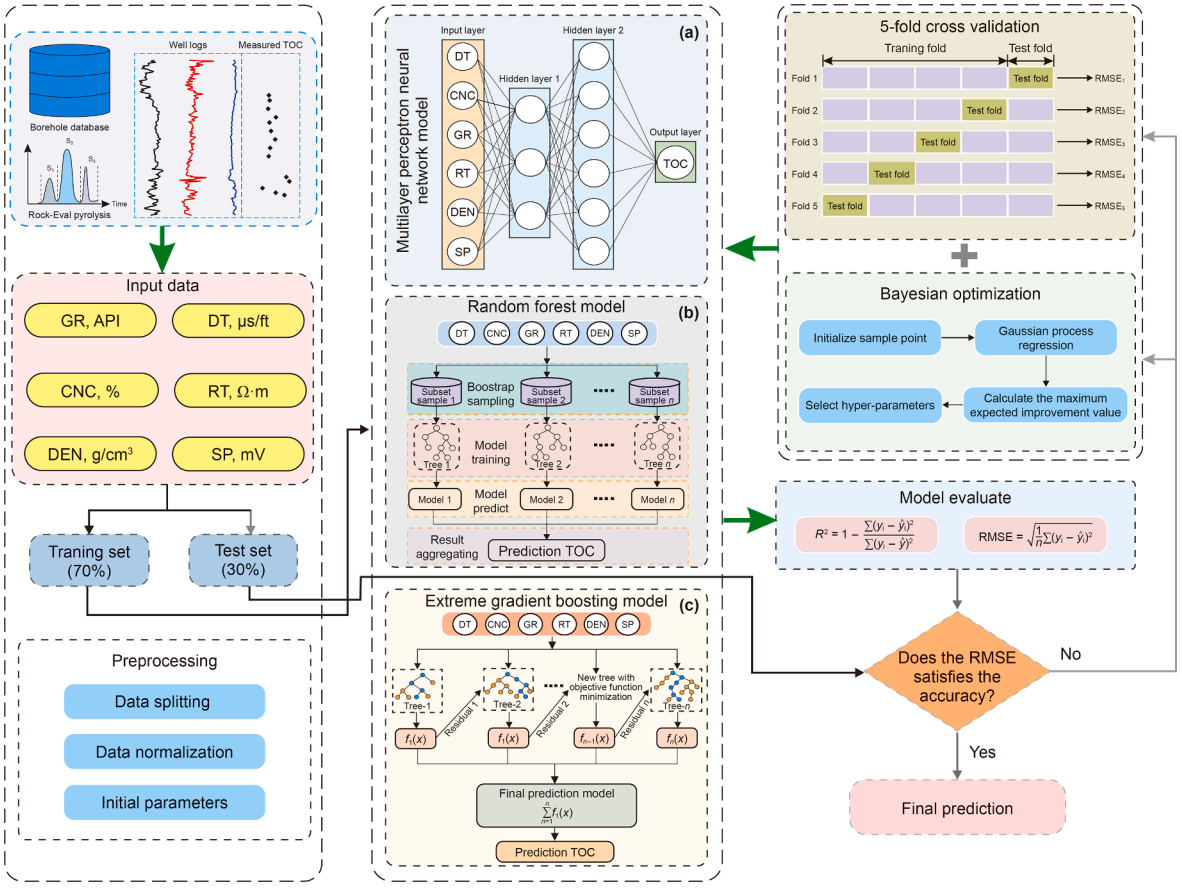

Fig. 3. Schematic diagram for predicting mudstone total organic carbon content using machine learning methods. **(a)** Schematic diagram of MLP structure; **(b)** schematic diagram of RF model; **(c)** schematic diagram of XGBoost model.

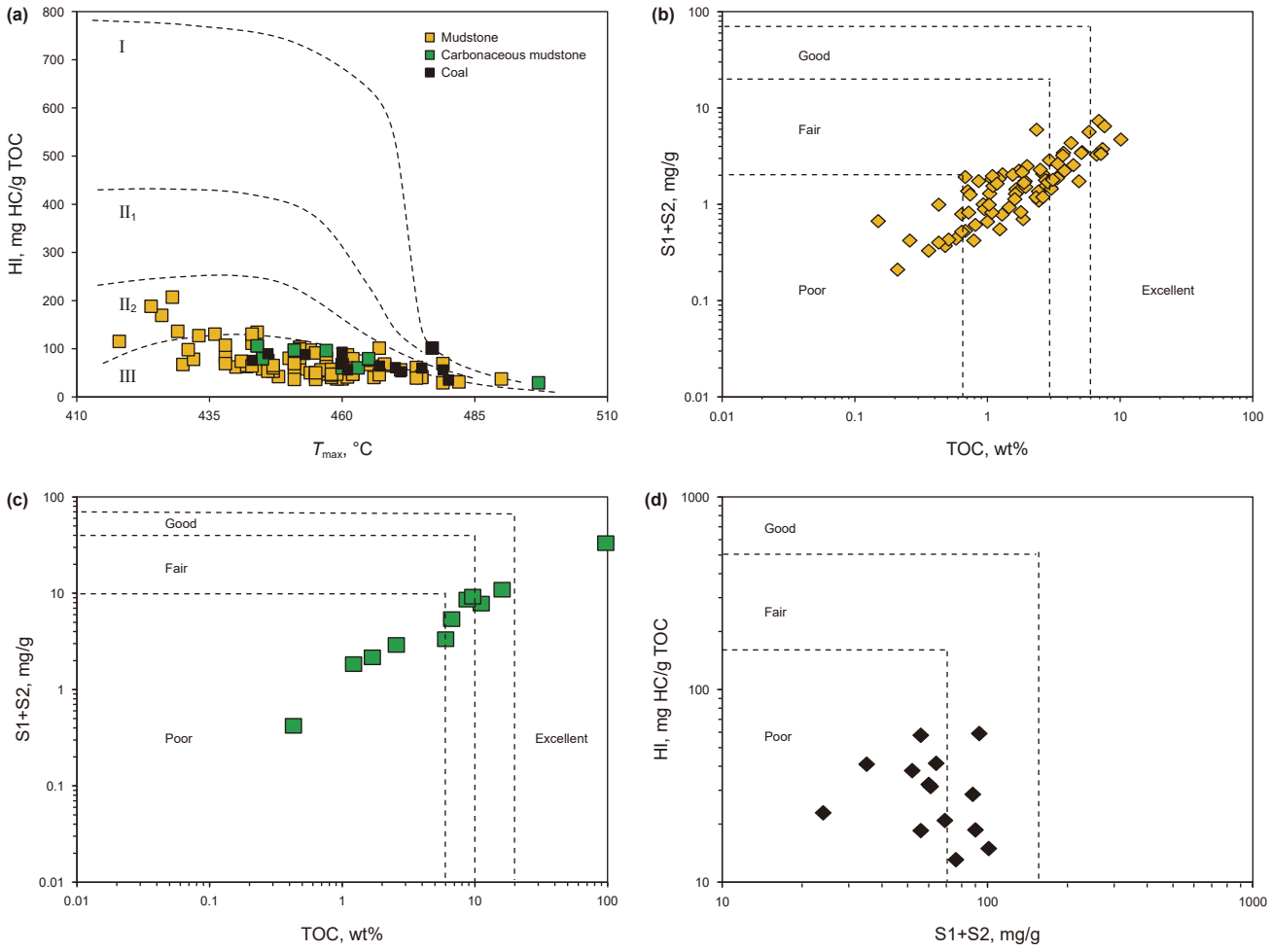

Fig. 4. Geochemical characteristics of Kezilenuer source rock. **(a)** Cross-plots of HI versus T_{max} showing organic matter type; **(b)** mudstone hydrocarbon potential distributions; **(c)** carbonaceous mudstone hydrocarbon potential distribution; **(d)** coal hydrocarbon potential distribution.

rocks predominantly contain Type III kerogen and minor occurrences of Type II_2. Type III kerogen indicates gas-prone organic matter (Huang et al., 2019; Mkono et al., 2023; Hu et al., 2024).

Fig. 4(b), (c), and (d) illustrate the relationships between S1+S2 and TOC, and HI and S1+S2, to evaluate the hydrocarbon generation potential of mudstone, carbonaceous mudstone, and coal. These cross-plots reveal variations in hydrocarbon production potential across different lithologies. Mudstone samples range from poor to excellent source rocks (Fig. 4(b)), while carbonaceous mudstone samples are mostly classified as poor to good, with one sample rated as excellent (Fig. 4(c)). Coal samples, however, are categorized as poor to fair source rocks in the Kezilenuer Formation (Fig. 4(d)) (Aziz et al., 2020; Gao et al., 2022; Nyakilla et al., 2022).

Overall, mudstone and carbonaceous mudstone exhibit fair to excellent hydrocarbon generation potential, which is better than coal. This phenomenon is mainly attributed to the intrinsically lower HI values characteristic of Type III kerogen constituents in the coal matrix, as well as post-sampling hydrocarbon volatilization (S1 escape), exhibiting reduced hydrocarbon generation potential of coal samples (Huang et al., 2019; Gao et al., 2022; Mkono et al., 2023).

4.2. Well log responses of source rocks

Mature source rocks are characterized by typical well log responses, such as high gamma-ray, resistivity, acoustic transit time, and compensated neutron porosity, along with reduced bulk density (Tan et al., 2015; Sahoo et al., 2021; Lai et al., 2022; Goliatt et al., 2023). These features result from the physical and chemical attributes of organic matter (Bolandi et al., 2015; Aziz et al., 2020; Lee et al., 2022). The Kezilenuer Formation, contains three lithologies—mudstone, carbonaceous mudstone, and coal, resulting in varying well log response characteristics.

Carbonaceous mudstone and coal, characterized by high organic content, show strong correlations with well log curves (Fig. 5). They exhibit negative correlations with gamma-ray, bulk density, and spontaneous potential curves, while positive correlations with acoustic transit time, neutron porosity, and resistivity curves (Fig. 5). Among these, acoustic transit time and bulk density are particularly sensitive to TOC (Fig. 5(b) and (c)), while showing weaker correlations with resistivity and spontaneous potential (Fig. 5(e) and (f)).

In contrast, mudstone TOC shows weaker correlations with most well log curves due to its lower organic content and intense compaction (Fig. 6). However, resistivity and spontaneous potential curves display stronger correlations with TOC compared to other curves (Fig. 6(e) and (f)).

Actually, GR will show high correlation relationships with TOC content, and high GR readings will imply a high TOC content (Tan et al., 2015; Zhao et al., 2016). However, the source rocks (mudstone) in the Kezilenuer Formation are deposited in shore-shallow lake, and no enough radioactive elements (U) will be absorbed in the source rocks, and the GR readings will not reflect

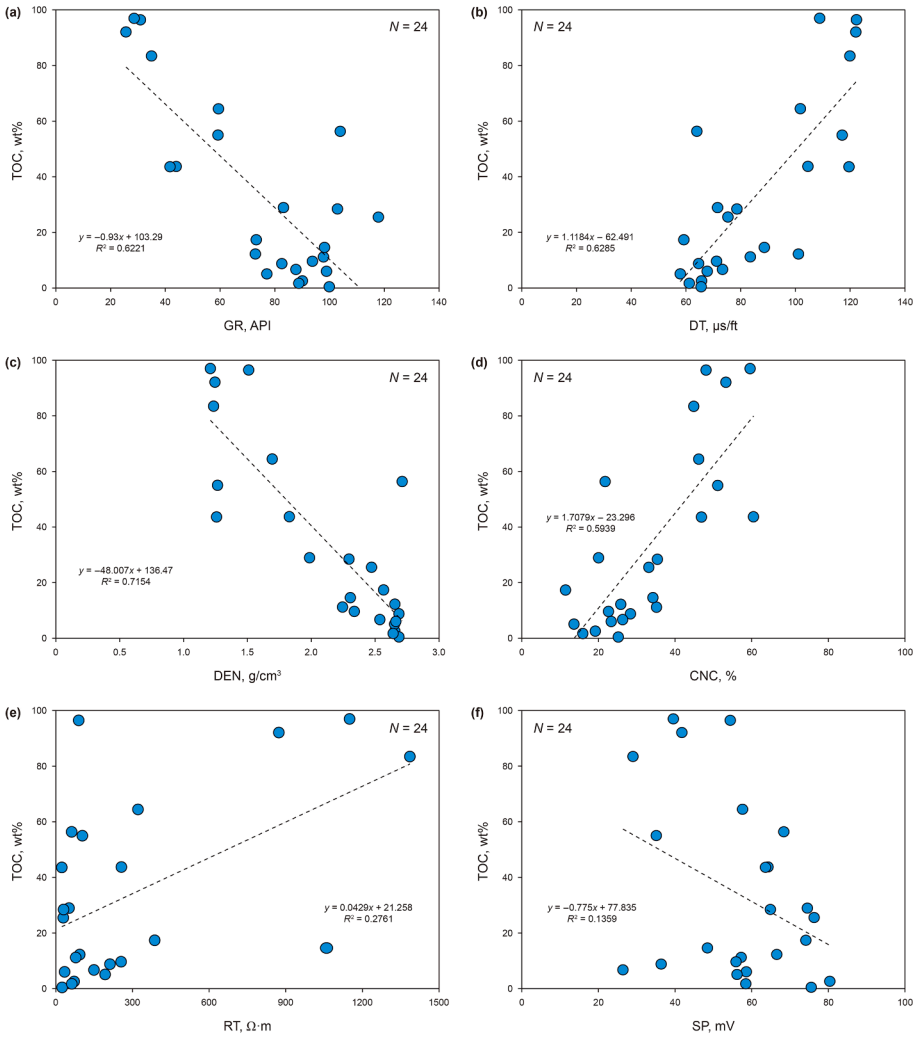

Fig. 5. Cross-plots of logging curves versus TOC content of carbonaceous mudstone and coal.

the TOC content, but the shale content (Huang et al., 2019; Lai et al., 2024; Lu et al., 2025). Actually, U enrichment is not common in lacustrine source rocks (Tenaglia et al., 2020; Zheng et al., 2021). Previous studies also indicate GR is not a good indicator of TOC for lacustrine source rocks (Lai et al., 2024). In addition, the mudstone will display low SP amplitudes, and the organic matter-rich mudstone will further reduce the SP amplitudes (Shi et al., 2016). Therefore, high TOC content will result in low SP readings in the dark mudstone of the Kezilenuer Formation. As can be seen From Fig. 6, SP shows a negative trend with TOC values.

As burial depth increases, the physical and chemical properties of mudstones undergo significant transformations (Zhao et al., 2005; Lai et al., 2023b). Intense compaction causes grain to become more closely contact, thereby reducing sonic travel time while increasing bulk density and resistivity (Hu et al., 2015; Lai et al., 2022; Lai et al., 2023b). As illustrated in Fig. 7, the same relatively stable mudstone interval in the study area occurs at shallower depths in Fig. 7(a) but extends deeper in Fig. 7(b). A direct comparison of these intervals reveals that Fig. 7(b) exhibits higher resistivity and bulk density values, along with reduced sonic travel time. Source rocks will display markedly different log responses compared to their different buried (Zhao et al., 2005; Hu et al., 2015; Lai et al., 2023b).

5. Discussion

5.1. Multivariate regression prediction

Multivariate regression, which integrates information from multiple well log curves, is employed for TOC prediction as it overcomes the limitations of relying on single curve (Shi et al., 2016; Lai et al., 2024). In this study, multivariate regression was applied to predict the TOC of carbonaceous mudstone and coal, utilizing their strong correlations with density, gamma-ray, and acoustic transit time curves. The TOC prediction results of all methods are shown in Supplementary Table. The regression formula is as follows:

$$TOC = -0.2953 \cdot GR - 31.5827 \cdot DEN + 0.1245 \cdot DT + 112.7018 \quad (3)$$

The predicted TOC values show a moderate to strong

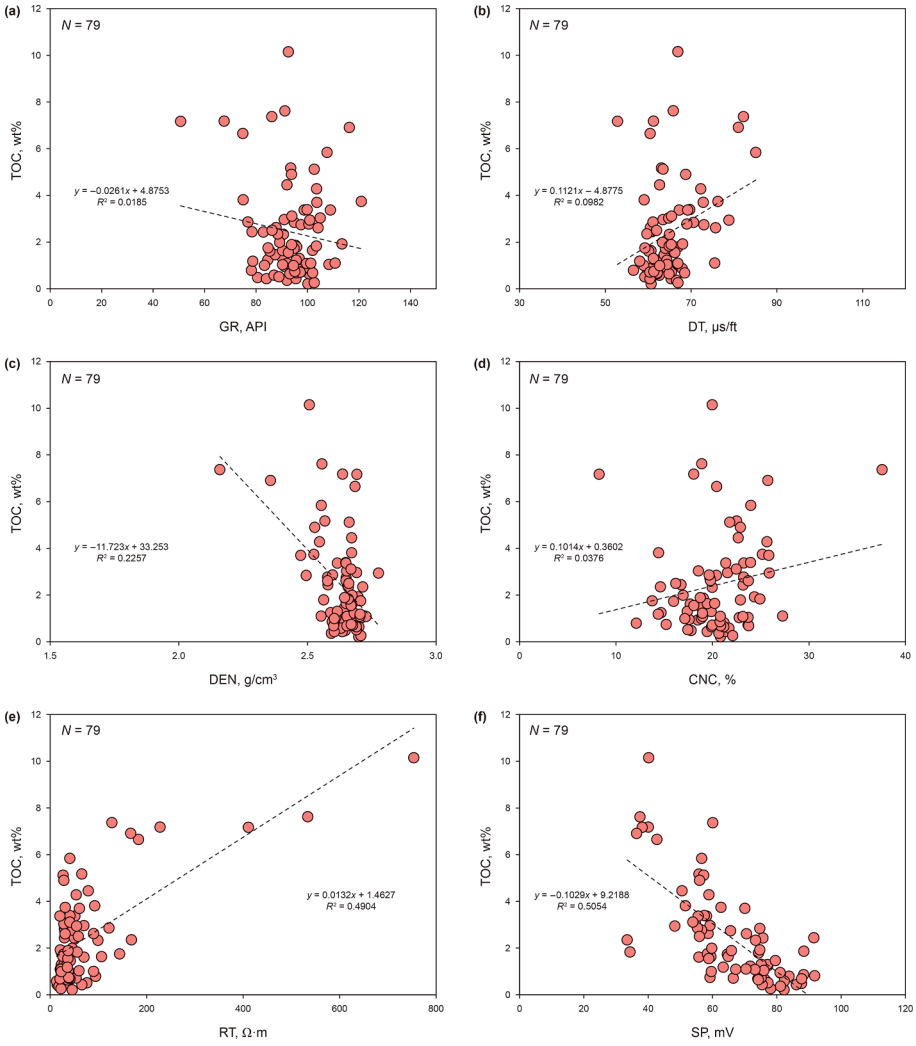

Fig. 6. Cross-plots of logging curves versus TOC content of mudstone.

correlation with measured TOC (R^2 = 0.7381), with most data points aligning closely along the $y = x$ line (Fig. 8(a)). This indicates the reliability of the multivariate regression model for TOC prediction for carbonaceous mudstone and coal.

5.2. ΔlgR method

In this study, acoustic and resistivity well log curves were used with the ΔlgR method (Eqs. (1) and (2)) to predict mudstone TOC. As shown in Fig. 8(b), the ΔlgR method yielded a determination coefficient of R^2 = 0.5466, indicating significant data dispersion and poor predictive accuracy (Fig. 8(b)). The ΔlgR method tends to underestimate low TOC values (<2 wt%) and fails to capture high TOC values (>6 wt%) (Fig. 8(b)).

Though widely used for shallow clastic and carbonate rocks, the ΔlgR method faces limitations in coaly source rocks of Kezilenuer Formation due to compaction and fluid effects, which weaken the responses of resistivity and acoustic logs, reducing predictive accuracy (Hu et al., 2015; Liu et al., 2021; Lee et al., 2022). The poor performance of the ΔlgR method is attributed to the weak correlation between acoustic transit time and TOC in this study (Fig. 6(b)). The coaly source rocks of Kezilenuer Formation have undergone significant compaction and stress, which reduces the interaction between porosity curves, resistivity, and TOC (Fig. 6). Consequently, the ΔlgR method, which relies on the overlay of acoustic and resistivity curves, is less effective for coaly source rocks of Kezilenuer Formation (Hu et al., 2015; Rui et al., 2020; Lai et al., 2024). To address these challenges, machine learning techniques have been increasingly employed to predict TOC by leveraging nonlinear relationships in well log data (Rui et al., 2020; Ochoa et al., 2022; Sahoo et al., 2021; Goliatt et al., 2023).

5.3. Machine learning models

The TOC of mudstone shows a certain degree of correlation with well log curves; however, these correlations are generally weak (Fig. 6). Initially, the traditional ΔlgR method was applied to predict TOC, but it demonstrated poor predictive performance in the study area (Fig. 8(b)). To address this limitation, machine learning methods, including MLP, RF, and XGBoost, were

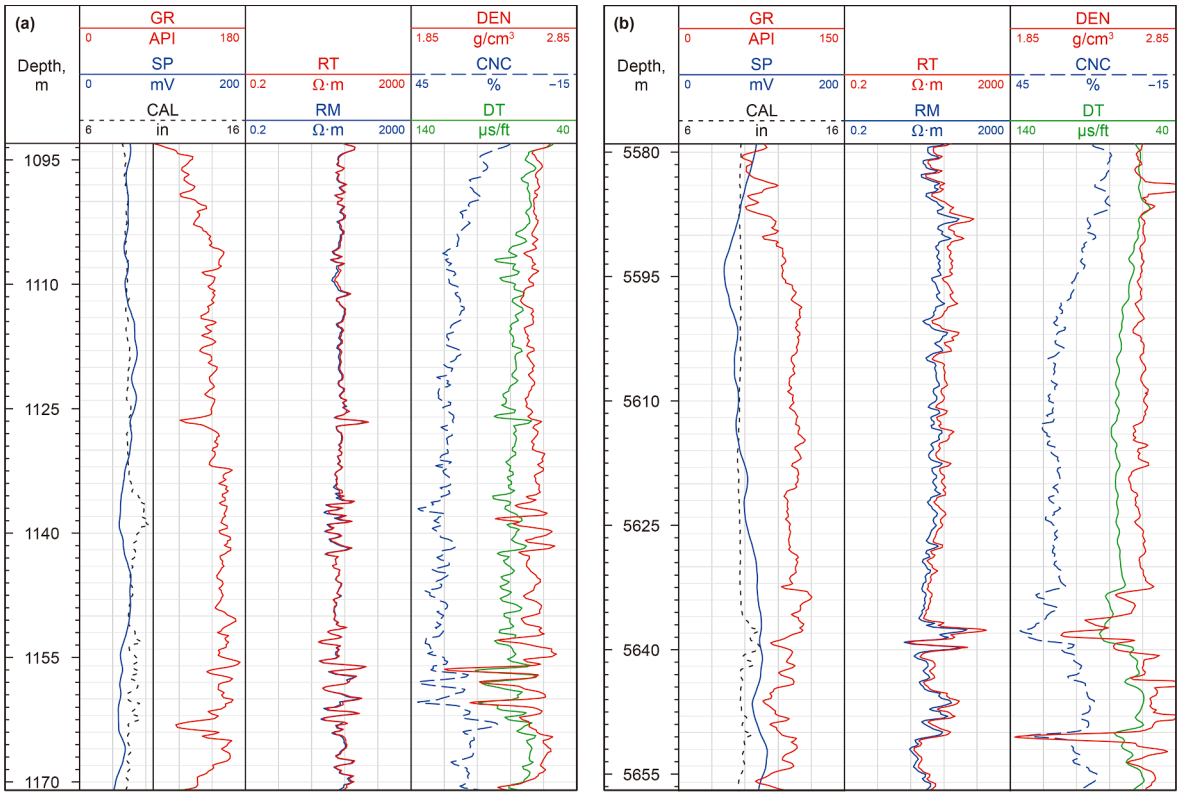

Fig. 7. Comparison of the well logging response characteristics of source rocks at different buried depth.

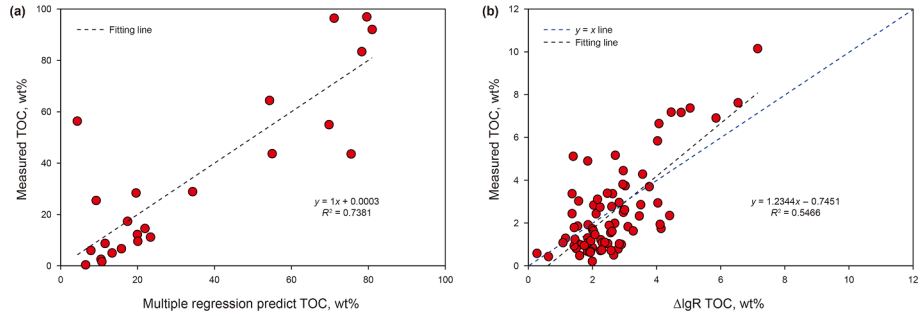

Fig. 8. **(a)** Cross-plots of measured TOC content of carbonaceous mudstone and coal versus TOC predicted from multiple regression, **(b)** cross-plots of measured TOC content of mudstone versus TOC predicted from ΔlgR method.

Table 1
Comparative analysis of mudstone TOC prediction models.

Methods	Well log parameters	Measured TOC	All predicted TOC	Train RMSE	Test RMSE	All samples RMSE
ΔlgR	RT, AC	0.21 ~ 10.15 2.43(79)	0.27 ~ 7.15 2.58(79)			1.3996
MLP	DT, CNC, GR, RT, DEN, SP	0.21 ~ 10.15 2.43(79)	0.44 ~ 9.32 2.35(79)	0.6614	0.8441	0.7211
RF	DT, CNC, GR, RT, DEN, SP	0.21 ~ 10.15 2.43(79)	0.37 ~ 8.93 2.36(79)	0.4833	0.7988	0.5957
XGBoost	DT, CNC, GR, RT, DEN, SP	0.21 ~ 10.15 2.43(79)	0.39 ~ 9.57 2.38(79)	0.3848	0.6003	0.4602

Note: The fraction represents the minimum ~ maximum/average value (number of samples) of TOC.

employed. These methods are able to capture complex nonlinear relationships among variables, which are often beyond the scope of the ΔlgR model (Zhu et al., 2018; Mulashani et al., 2021; Goliatt et al., 2023; Lai et al., 2024). In addition, a 5-fold cross-validation approach and Bayesian optimization algorithm are used to prevent overfitting and optimize the better parameters for each model (Fig. 3). The RMSE values for the training set, testing set, and all samples are summarized in Table 1.

Table 2
MLP neural network parameter settings.

Parameter	Value
Number of neurons in input layer	6
Hidden layer 1	3
Hidden layer 2	6
Learning rate	0.1419
Number of iterations	2845
Number of neurons in out layer	1

Table 3
RF parameter settings.

Parameter or function	Value
Number of iterations	22
Regression tree maximum depth	8
Minimum sample of leaf nodes	1
Minimum sample size required for leaf node division	3

5.3.1. MLP model

The MLP model was optimized using Bayesian optimization to adjust parameters. The optimal parameters for MLP are listed in Table 2, and the MLP structure is illustrated in Fig. 3(a).

The MLP model demonstrated good predictive performance, with R^2 values of 0.912, 0.8249, and 0.8846 for the training set, test set, and all samples, respectively. Corresponding RMSE values were 0.6614, 0.8414, and 0.7211 (Fig. 9(a)–(d), (g); Table 1). While the model showed slight underperformance in predicting high TOC values during training, the predicted and measured values were generally well-aligned along the $y = x$ line, indicating good

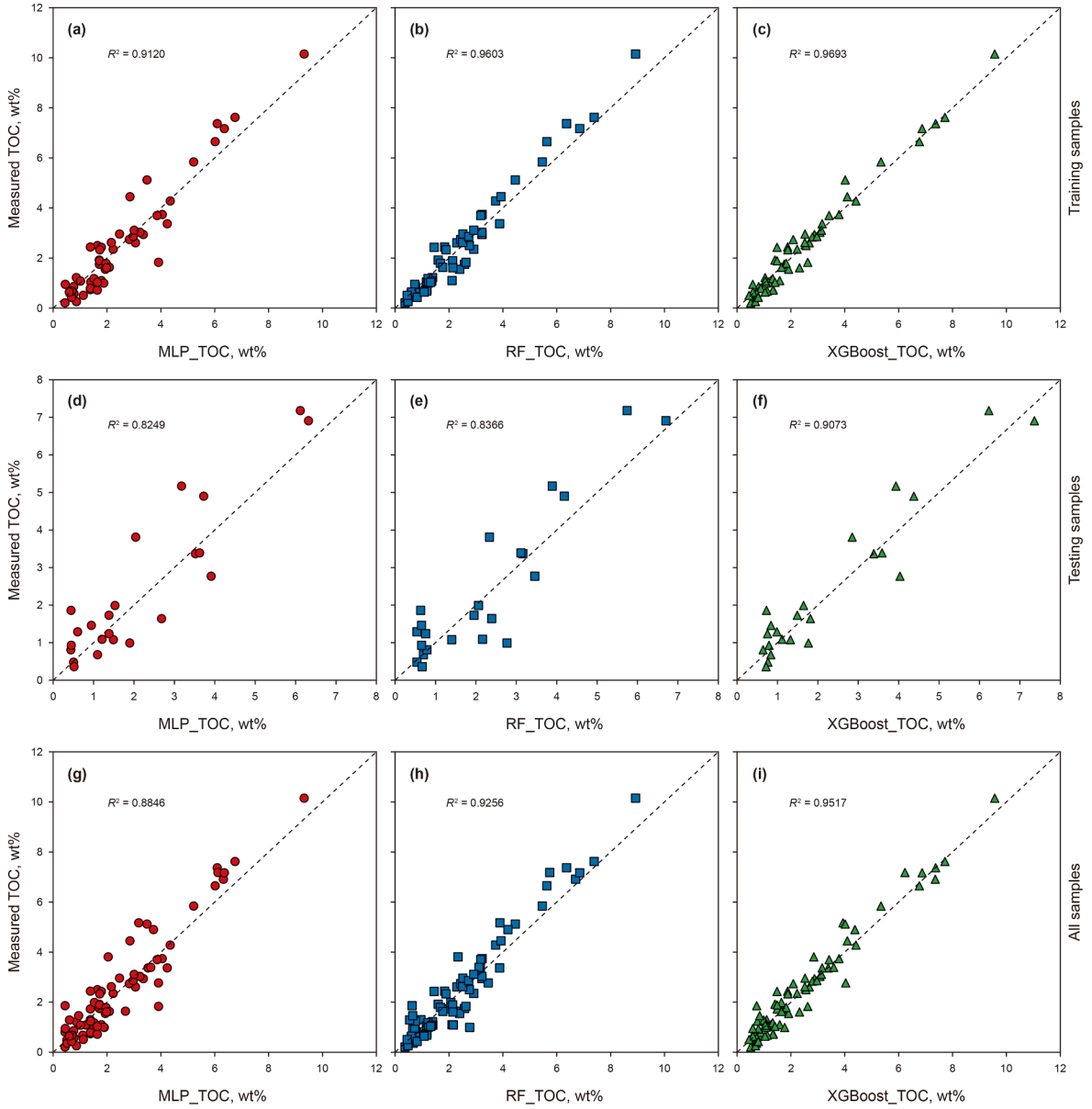

Fig. 9. A comparison of the predicted TOC and measured TOC for the different methods. **(a), (d)**, and **(g)** are the cross-plots of the measured and predicted TOC values by the MLP model for the training, testing, and all samples, respectively. **(b)**, **(e)**, and **(h)** are the cross-plots of the measured and predicted TOC values by the RF model for the training, testing, and all samples, respectively. **(c)**, **(f)**, and **(i)** are the cross-plots of the measured and predicted TOC values by the XGBoost model for the training, testing, and all samples, respectively.

Table 4
XGBoost parameter settings.

Parameter or function	Value
Number of iterations	75
Learning rate	0.1
Regression tree maximum depth	2
Minimum loss reduction	0.1922
Minimum sample of leaf nodes	1

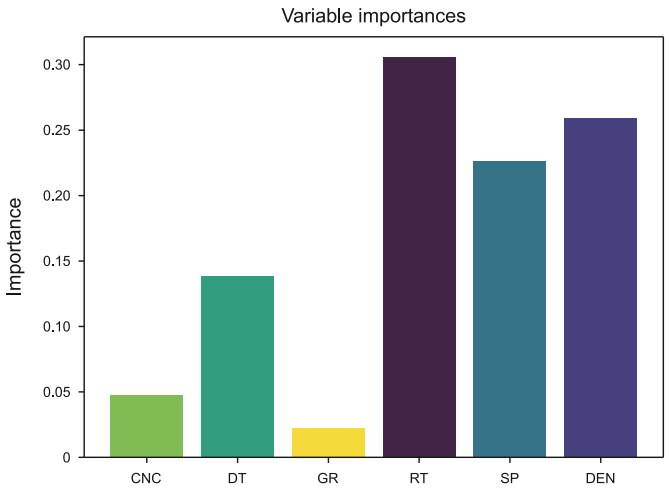

Fig. 10. Permutation feature importance of XGBoost model for each input parameter.

generalization ability. In contrast to the ΔlgR method, the MLP model successfully captured high TOC values, as illustrated by the resulting TOC distribution (Fig. 9(g)).

5.3.2. RF model

The RF model, optimized using Bayesian methods, achieved strong predictive performance with R^2 values of 0.9603, 0.8366, and 0.9256 for the training set, test set, and all samples, respectively. The optimal parameters for RF are listed in Table 3 corresponding RMSE values were 0.4833, 0.7988, and 0.5967 (Fig. 9(b)–(e), (h); Table 1). These results highlight the RF model's robust predictive capabilities.

However, the model struggled to accurately predict high TOC values, as evidenced by deviations from the $y = x$ line for data points corresponding to higher TOC values (Fig. 9(h)). This limitation may be due to the scarcity of high TOC samples in the dataset, which restricted the model's ability to learn from these instances.

5.3.3. XGBoost model

The XGBoost model, with optimized parameters listed in Table 4, outperformed all other methods in this study. It achieved R^2 values of 0.9693, 0.9073, and 0.9517 for the training set, testing set, and all samples, respectively, with RMSE values of 0.3848, 0.6003, and 0.4602 (Fig. 9(c)–(f), (i); Table 1).

The XGBoost model demonstrated the highest predictive accuracy and lowest error among all models. Feature importance analysis revealed that resistivity (RT) had the greatest influence on TOC prediction, followed by SP, DEN, DT, CNC, and GR (Fig. 10). Notably, SP made a significant contribution, while GR had the least impact, aligning with the relationship between TOC and well log parameters illustrated in Fig. 6. This highlights the importance of incorporating SP in TOC prediction for coaly hydrocarbon source rocks.

Overall, the XGBoost model exhibited superior performance in handling complex nonlinear relationships and multivariable interactions, significantly outperforming the ΔlgR, MLP, and RF models shown in Figs. 11 and 12. In addition, lithologies of mudstone, carbonaceous mudstone and coal are frequently interbedded; however, these figures present a comparison between measured TOC and predicted TOC in the mudstone interval (Figs. 11 and 12).

5.4. Implication for source rock prediction

Accurately predicting TOC in hydrocarbon source rocks is critical for evaluating hydrocarbon generation potential and guiding oil and gas exploration (Zhao et al., 2005; Sahoo et al., 2021; Gao et al., 2022; Mkono et al., 2023). However, complex geological conditions and the limitations of traditional methods, such as the ΔlgR model, hinder prediction accuracy (Shi et al., 2016; Lai et al., 2024). The ΔlgR method, relying solely on acoustic and resistivity curves, fails to capture the multifaceted relationship between TOC and well log parameters, resulting in poor performance in coaly hydrocarbon source rocks (Hu et al., 2015; Liu et al., 2021; Lee et al., 2022; Lai et al., 2024). This study shows that incorporating more comprehensive logging data, such as SP, DEN, and CNC, into machine learning models significantly enhances TOC prediction accuracy (Figs. 8(b) and 9).

Among the models employed, XGBoost demonstrated the best predictive performance, accurately capturing both low and high TOC values due to its capability of handling complex nonlinear interactions and multivariable relationships (Figs. 11 and 12). While MLP and RF models also outperformed the ΔlgR method, they showed deficiencies in predicting high TOC values, likely due to limited representation of high-value samples during training (Fig. 9). These findings highlight the potential of advanced machine learning models, particularly XGBoost, for improving TOC prediction in coaly hydrocarbon source rocks, providing more reliable guidance for coal-bearing formation exploration.

In addition to predicting TOC, other source rock parameters—such as R_o, T_{max}, and hydrocarbon generation potential—are also important indicators for source rock evaluation. Integration of machine learning with numerous geological data enables accurate prediction of key reservoir indicators. These data-driven models significantly enhance decision-making precision in coal-bearing formation hydrocarbon exploration.

6. Conclusion

The evaluation of source rocks is critical for both conventional and unconventional oil and gas exploration, as well as for coaly hydrocarbon source rocks. This study focused on the Kezilenuer Formation in the Kuqa Depression and yielded the following key findings.

(1) Mudstone, carbonaceous mudstone, and coal are the primary lithology types of source rocks. The organic matter is predominantly Type III, with minor Type II_2. Mudstone exhibits fair to good hydrocarbon generation potential, while carbonaceous mudstone and coal range from fair to poor source rocks.

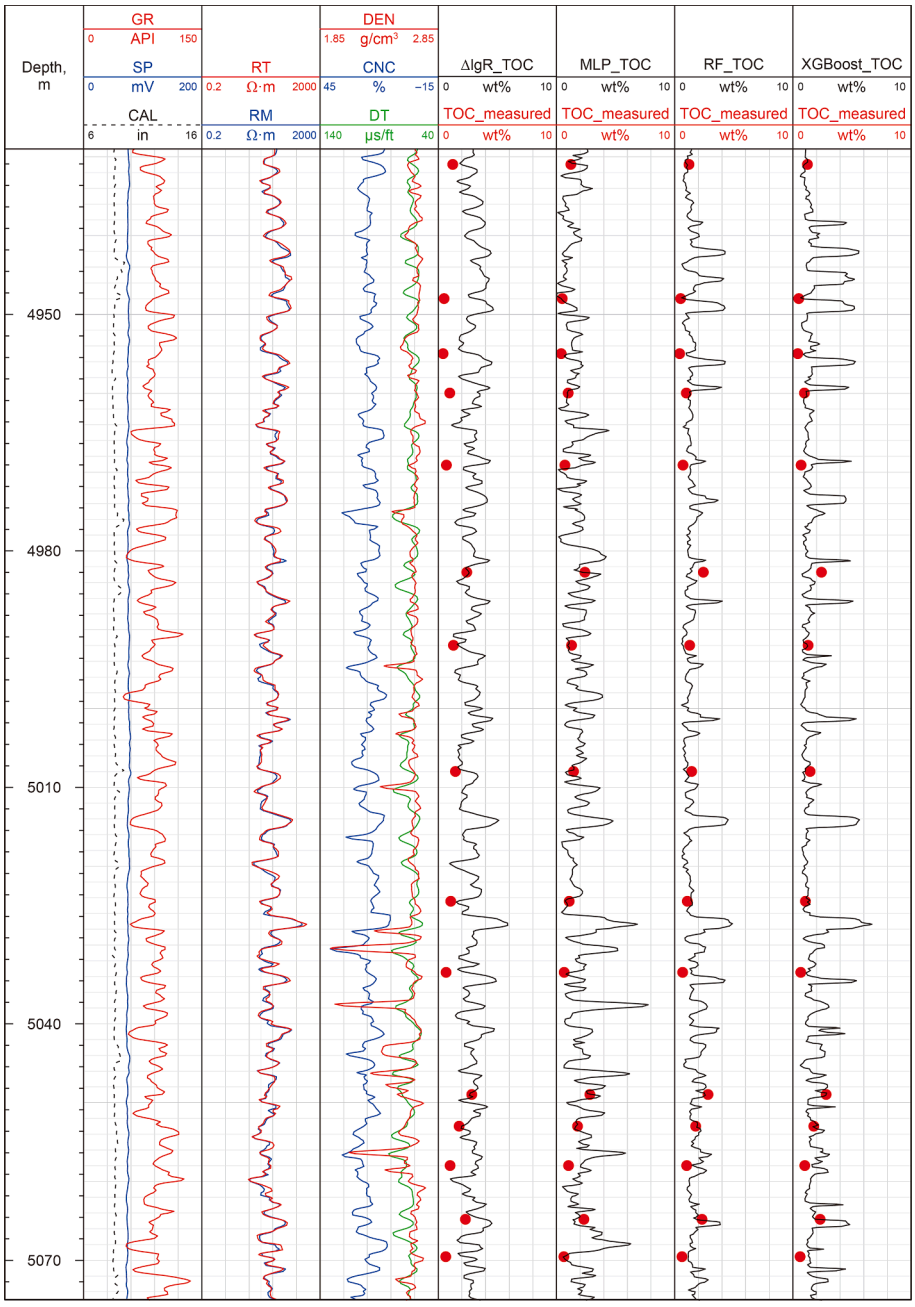

Fig. 11. TOC content predicted using ΔlgR method, MLP, RF, and XGBoost model of mudstone in Well A.

(2) The traditional ΔlgR method demonstrated poor accuracy for predicting TOC in coaly source rocks due to its limited input variables and inability to capture complex relationships. The ΔlgR method is unsuitable for TOC prediction in coaly source rocks of Kezilenuer Formation.

(3) Machine learning models (MLP, RF, and XGBoost) significantly improved TOC prediction accuracy. Among them, XGBoost demonstrated the highest predictive accuracy due to its ability to handle complex and nonlinear relationships and multivariable interactions, making it the most suitable method for TOC prediction in coaly source rocks of Kezilenuer Formation. For coaly source rocks, machine learning methods have the advantages for TOC prediction, and will provide technical guidance for coal-bearing formation hydrocarbon exploration.

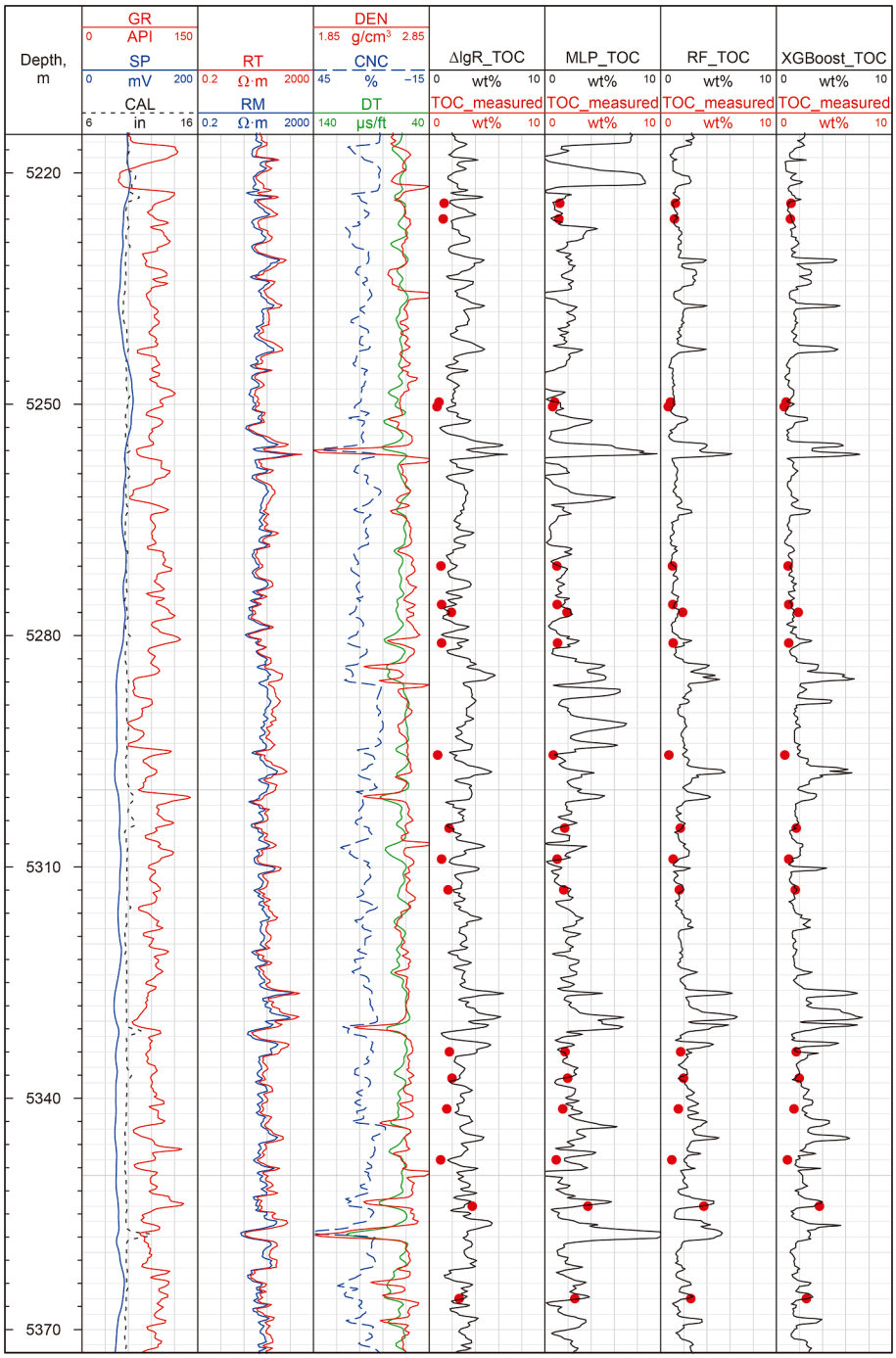

Fig. 12. TOC content predicted using ΔlgR method, MLP, RF, and XGBoost model of mudstone in Well B.

CRediT authorship contribution statement

Fei Zhao: Writing – original draft, Visualization, Validation, Software, Methodology. **Jin Lai:** Writing – review & editing, Resources, Methodology, Funding acquisition. **Zong-Li Xia:** Visualization, Validation, Supervision, Formal analysis. **Zhong-Rui Wang:** Visualization, Supervision, Formal analysis. **Ling Li:** Resources, Conceptualization. **Bin Wang:** Resources, Conceptualization. **Lu Xiao:** Validation, Data curation. **Yang Su:** Visualization, Investigation. **Gui-Wen Wang:** Writing – review & editing, Resources, Project administration.

Declaration of interests

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.

Acknowledgement

This work is financially supported by Science Foundation of China University of Petroleum (Beijing) (No.2462023QNXZ010). The authors extend their sincere gratitude to the editors for their

enthusiasm, patience, and tireless efforts, as well as to the four reviewers for their constructive suggestions, which greatly enhanced the quality of this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petsci.2025.05.029.

References

Alizadeh, B., Maroufi, K., Heidarifard, M.H., 2018. Estimating source rock parameters using wireline data: an example from Dezful Embayment, South West of Iran. J. Petrol. Sci. Eng. 167, 857–868. https://doi.org/10.1016/j.petrol.2017.12.021.

Aziz, H., Ehsan, M., Ali, A., et al., 2020. Hydrocarbon source rock evaluation and quantification of organic richness from correlation of well logs and geochemical data: a case study from the sembar formation, Southern Indus Basin, Pakistan. J. Nat. Gas Sci. Eng. 81, 103433. https://doi.org/10.1016/j.jngse.2020.103433.

Bolandi, V., Kadkhodaie-Ilkhchi, A., Alizadeh, B., et al., 2015. Source rock characterization of the Albian Kazhdumi formation by integrating well logs and geochemical data in the Azadegan oilfield, Abadan plain, SW Iran. J. Petrol. Sci. Eng. 133, 167–176. https://doi.org/10.1016/j.petrol.2015.05.022.

Bolandi, V., Kadkhodaie, A., Farzi, R., 2017. Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation, the Persian Gulf basin, offshore Iran. J. Petrol. Sci. Eng. 151, 224–234. https://doi.org/10.1016/j.petrol.2017.01.003.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cappuccio, F., Porreca, M., Omosanya, K.O., et al., 2021. Total organic carbon (TOC) enrichment and source rock evaluation of the Upper Jurassic-Lower Cretaceous rocks (Barents Sea) by means of geochemical and log data. Int. J. Earth Sci. 110, 115–126. https://doi.org/10.1007/s00531-020-01941-6.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

El Sharawy, M.S., Gaafar, G.R., 2012. Application of well log analysis for source rock evaluation in the Duwi Formation, Southern Gulf of Suez, Egypt. J. Appl. Geophys. 80, 129–143. https://doi.org/10.1016/j.jappgeo.2011.12.005.

Gao, T.Z., Ding, X.J., Yang, X.Z., et al., 2022. Geochemical characteristics and depositional environment of coal-measure hydrocarbon source rocks in the northern tectonic belt, Kuqa depression. Appl. Sci. 12 (19), 9464. https://doi.org/10.3390/app12199464.

Goliatt, L., Saporetti, C.M., Pereira, E., 2023. Super learner approach to predict total organic carbon using stacking machine learning models based on well logs. Fuel 353, 128682. https://doi.org/10.1016/j.fuel.2023.128682.

Gordon, J.B., Sanei, H., Pedersen, P.K., 2022. Predicting hydrogen and oxygen indices (HI, OI) from conventional well logs using a Random Forest machine learning algorithm. Int. J. Coal Geol. 249, 103903. https://doi.org/10.1016/j.coal.2021.103903.

Guo, S., Lyu, X.X., Zhang, Y., 2018. Relationship between tight sandstone reservoir formation and hydrocarbon charging: a case study of a Jurassic reservoir in the eastern Kuqa Depression, Tarim Basin, NW China. J. Nat. Gas Sci. Eng. 52, 304–316. https://doi.org/10.1016/j.jngse.2018.01.031.

Handhal, A.M., Al-Abadi, A.M., Chafeet, H.E., et al., 2020. Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. Mar. Petrol. Geol. 116, 104347. https://doi.org/10.1016/j.marpetgeo.2020.104347.

Hood, A., Gutjahr, C.C.M., Heacock, R.L., 1975. Organic metamorphism and the generation of petroleum. AAPG Bull. 59 (6), 986–996. https://doi.org/10.1306/83d91f06-16c7-11d7-8645000102c1865d.

Hu, H.T., Liu, C., Lu, et al., 2015. The method and application of using generalized-ΔLgR technology to predict the organic carbon content of continental deep source rocks. Acta Geol. Sin. 89. https://doi.org/10.1111/1755-6724.12306_14.

Hu, Y., Jia, C.Z., Chen, J.Q., et al., 2024. Restoration of hydrocarbon generation potential of the highly mature Lower Cambrian Yuertusi Formation source rocks in the Tarim Basin. Pet. Sci. 22 (2), 588–606. https://doi.org/10.1016/j.petsci.2024.12.001.

Huang, W.K., Zeng, L.F., Pan, C.C., et al., 2019. Petroleum generation potentials and kinetics of coaly source rocks in the Kuqa Depression of Tarim Basin, northwest China. Org. Geochem. 133, 32–52. https://doi.org/10.1016/j.orggeochem.2019.04.007.

Hunt, M.J., 1996. Petroleum Geochemistry and Geology, second ed. WH Freeman and company, New York. 743–743.

Lai, J., Bai, T.Y., Zhao, Y.D., et al., 2023b. Unusually petrophysical behavior and geological significance of mudrocks. Geoenergy Sci. Eng. 230, 212171. https://doi.org/10.1016/j.geoen.2023.212171.

Lai, J., Wang, G.W., Fan, Q.X., et al., 2022. Geophysical well-log evaluation in the era of unconventional hydrocarbon resources: a review on current status and prospects. Surv. Geophys. 43 (3), 913–957. https://doi.org/10.1007/s10712-022-09705-4.

Lai, J., Li, D., Bai, T.Y., Z, et al., 2023a. Reservoir quality evaluation and prediction in ultra-deep tight sandstones in the Kuqa depression, China. J. Struct. Geol. 170, 104850. https://doi.org/10.1016/j.jsg.2023.104850.

Lai, J., Wang, G.W., Chai, Y., et al., 2017. Deep burial diagenesis and reservoir quality evolution of high-temperature, high-pressure sandstones: examples from Lower Cretaceous Bashijiqike Formation in Keshen area, Kuqa Depression, Tarim Basin of China. AAPG Bull. 101 (6), 829–862. https://doi.org/10.1306/08231614008.

Lai, J., Zhao, F., Xia, Z.L., et al., 2024. Well log prediction of total organic carbon: a comprehensive review. Earth Sci. Rev., 104913 https://doi.org/10.1016/j.earscirev.2024.104913.

Lee, J., Lumley, D.E., Lim, U.Y., 2022. Improving total organic carbon estimation for unconventional shale reservoirs using Shapley value regression and deep machine learning methods. AAPG Bull. 106 (11), 2297–2314. https://doi.org/10.1306/02072221021.

Li, D., Wang, G.W., Bie, K., et al., 2025. Formation mechanism and reservoir quality evaluation in tight sandstones under a compressional tectonic setting: the Jurassic Ahe Formation in Kuqa Depression, Tarim Basin, China. Pet. Sci. 22 (3), 998–1020. https://doi.org/10.1016/j.petsci.2024.12.026.

Liu, X.Z., Tian, Z., Chen, C., 2021. Total organic carbon content prediction in lacustrine shale using extreme gradient boosting machine learning based on bayesian optimization. Geofluids 2021, 1–18. https://doi.org/10.1155/2021/6155663.

Lu, M., Duan, G.Q., Zhang, T.X., et al., 2025. Influences of paleoclimatic changes on organic matter enrichment mechanisms in freshwater and saline lacustrine oil shales in China: a machine learning approach. Earth Sci. Rev., 105061 https://doi.org/10.1016/j.earscirev.2025.105061.

Mahmoud, A.A.A., Elkatatny, S., Mahmoud, M., et al., 2017. Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network. Int. J. Coal Geol. 179, 72–80. https://doi.org/10.1016/j.coal.2017.05.012.

Maroufi, K., Zahmatkesh, I., 2023. Effect of lithological variations on the performance of artificial intelligence techniques for estimating total organic carbon through well logs. J. Petrol. Sci. Eng. 220, 111213. https://doi.org/10.1016/j.petrol.2022.111213.

Mkono, C.N., Shen, C.B., Mulashani, A.K., et al., 2023. Deep learning integrated approach for hydrocarbon source rock evaluation and geochemical indicators prediction in the Jurassic - paleogene of the Mandawa basin, SE Tanzania. Energy 284, 129232. https://doi.org/10.1016/j.energy.2023.129232.

Mulashani, A.K., Shen, C.B., Asante-Okyere, S., et al., 2021. Group method of data handling (GMDH) neural network for estimating total organic carbon (TOC) and hydrocarbon potential distribution (S1, S2) using well logs. Nat. Resour. Res. 30 (5), 3605–3622. https://doi.org/10.1007/s11053-021-09908-3.

Nyakilla, E.E., Silingi, S.N., Shen, C.B., et al., 2022. Evaluation of source rock potentiality and prediction of total organic carbon using well log data and integrated methods of multivariate analysis, machine learning, and geochemical analysis. Nat. Resour. Res. 31 (1), 619–641. https://doi.org/10.1007/s11053-021-09988-1.

Ochoa, R.I., Birgenheier, L.P., Schwarz, E., et al., 2022. Calibrated petrophysical model for elevated organic matter intervals and mineralogical variability in the Agrio Formation, Neuquen Basin, Argentina. Mar. Petrol. Geol. 146, 105913. https://doi.org/10.1016/j.marpetgeo.2022.105913.

Passey, Q.R., Creaney, S., Kulla, J.B., et al., 1990. A practical model for organic richness from porosity and resistivity logs. AAPG Bull. 12 (74), 1777–1794. https://doi.org/10.1016/0148-9062(91)90313-b.

Rui, J.W., Zhang, H.B., Ren, Q., et al., 2020. TOC content prediction based on a combined Gaussian process regression model. Mar. Petrol. Geol. 118, 104429. https://doi.org/10.1016/j.marpetgeo.2020.104429.

Safaei-Farouji, M., Kadkhodaie, A., 2022. Application of ensemble machine learning methods for kerogen type estimation from petrophysical well logs. J. Petrol. Sci. Eng. 208, 109455. https://doi.org/10.1016/j.petrol.2021.109455.

Sahoo, T.R., Funnell, R.H., Brennan, S.W., et al., 2021. Delineation of coaly source rock distribution and prediction of organic richness from integrated analysis of seismic and well data. Mar. Petrol. Geol. 125, 104873. https://doi.org/10.1016/j.marpetgeo.2020.104873.

Schmoker, J.W., 1979. Determination of organic content of Appalachian Devonian shales from formation-density logs: geologic notes. AAPG Bull. 63, 1504–1509. https://doi.org/10.1306/2f9185d1-16ce-11d7-8645000102c1865d.

Sêco, S.L.R., Silva, R.L., Watson, N., et al., 2019. Application of petrophysical methods to estimate total organic carbon in Lower Jurassic source rocks from the offshore Lusitanian Basin (Portugal). J. Petrol. Sci. Eng. 180, 1058–1068. https://doi.org/10.1016/j.petrol.2019.05.065.

Sfidari, E., Kadkhodaie-Ilkhchi, A., Najjari, S., 2012. Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems. J. Petrol. Sci. Eng. 86–87, 190–205. https://doi.org/10.1016/j.petrol.2012.03.024.

Shalaby, M.R., Jumat, N., Lai, D., et al., 2019. Integrated TOC prediction and source rock characterization using machine learning, well logs and geochemical analysis: case study from the Jurassic source rocks in Shams Field, NW Desert, Egypt. J. Petrol. Sci. Eng. 176, 369–380. https://doi.org/10.1016/j.petrol.2019.01.055.

Shi, X., Wang, J., Liu, G., et al., 2016. Application of extreme learning machine and neural networks in total organic carbon content prediction in organic shale with wire line logs. J. Nat. Gas Sci. Eng. 33, 687–702. https://doi.org/10.1016/j.jngse.2016.05.060.

Tan, M.J., Song, X.D., Yang, X., et al., 2015. Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: a comparative study. J. Nat. Gas Sci. Eng. 26 (1), 792–802. https://doi.org/10.1016/j.jngse.2015.07.008.

Tenaglia, M., Eberli, G.P., Weger, R.J., et al., 2020. Total organic carbon quantification from wireline logging techniques: a case study in the Vaca Muerta Formation, Argentina. J. Petrol. Sci. Eng. 194, 107489. https://doi.org/10.1016/j.petrol.2020.107489.

Wan, J.L., Gong, Y.J., Huang, W.H., et al., 2022. Characteristics of hydrocarbon migration and accumulation in the lower jurassic reservoirs in the tugerming area of the eastern Kuqa Depression, Tarim Basin. J. Petrol. Sci. Eng. 208, 109748. https://doi.org/10.1016/j.petrol.2021.109748.

Wang, B., Qiu, N.S., Amberg, S.T., et al., 2022. Modelling of pore pressure evolution in a compressional tectonic setting: the Kuqa Depression, Tarim Basin, northwestern. China. Mar. Petrol. Geol. 146, 105936. https://doi.org/10.1016/j.marpetgeo.2022.105936.

Wang, H.J., Wu, W., Chen, T., et al., 2019. An improved neural network for TOC, S1 and S2 estimation based on conventional well logs. J. Petrol. Sci. Eng. 176, 664–678. https://doi.org/10.1016/j.petrol.2019.01.096.

Wang, J.B., Gao, Z.Q., Kang, Z.H., et al., 2020. Geochemical characteristics, hydrocarbon potential and depositional environment of the Yangye Formation source rocks in Kashi Sag, southwestern Tarim Basin, NW China. Mar. Petrol. Geol. 112, 104084. https://doi.org/10.1016/j.marpetgeo.2019.104084.

Zeng, B., Li, M.J., Zhu, J.Q., et al., 2021. Selective methods of TOC content estimation for organic-rich interbedded mudstone source rocks. J. Nat. Gas Sci. Eng. 93, 104064. https://doi.org/10.1016/j.jngse.2021.104064.

Zhang, H.Y., Wu, W.S., Wu, H., 2023. TOC prediction using a gradient boosting decision tree method: a case study of shale reservoirs in Qinshui Basin. J. Petrol. Sci. Eng. 221, 111271. https://doi.org/10.1016/j.petrol.2022.111271.

Zhao, P.Q., Mao, Z.Q., Huang, Z.H., et al., 2016. A new method for estimating total organic carbon content from well logs. AAPG Bull. 100 (8), 1311–1327. https://doi.org/10.1306/02221615104.

Zhao, P.Q., Ostadhassan, M., Shen, B., et al., 2019. Estimating thermal maturity of organic-rich shale from well logs: case studies of two shale plays. Fuel 235, 1195–1206. https://doi.org/10.1016/j.fuel.2018.08.037.

Zhao, W.Z., Zhang, S.C., Wang, F.Y., et al., 2005. Gas systems in the Kuche Depression of the Tarim Basin: source rock distributions, generation kinetics and gas accumulation history. Org. Geochem. 36 (12), 1583–1601. https://doi.org/10.1016/j.orggeochem.2005.08.016.

Zheng, D.Y., Wu, S.X., Hou, M.C., 2021. Fully connected deep network: an improved method to predict TOC of shale reservoirs from well logs. Mar. Petrol. Geol. 132, 105205. https://doi.org/10.1016/j.marpetgeo.2021.105205.

Zhu, L.Q., Zhang, C., Zhang, C.M., et al., 2018. Prediction of total organic carbon content in shale reservoir based on a new integrated hybrid neural network and conventional well logging curves. J. Geophys. Eng. 15 (3), 1050–1061. https://doi.org/10.1088/1742-2140/aaa7af.