Front. Earth Sci. 2025, 19(2): 322–339
https://doi.org/10.1007/s11707-025-1149-y

RESEARCH ARTICLE

A new method for total organic carbon prediction of marine-continental transitional shale based on multivariate nonlinear regression

Xinyu ZHANG[1], Yanjun MENG (✉)[1,2,3], Taotao YAN[1,3], Jinzhi ZHONG[4], Zhen QIU[5], Weibo ZHAO[6], Liangliang YIN[6], Haojie MA[1], Qin ZHANG[5]

1 College of Geological and Surveying Engineering, Taiyuan University of Technology, Taiyuan 030024, China
2 Provincial Center of Technology Innovation for Coal Measure Gas Co-production, Taiyuan 030082, China
3 Shanxi Key Laboratory of Fine Exploration of Coal-based Critical Mineral Resources, Taiyuan 030024, China
4 School of Resources and Environmental Engineering, Hefei University of Technology, Heifei 230009, China
5 PetroChina Research Institute of Petroleum Exploration and Development, Beijing 100083, China
6 Research Institute of Exploration and Development of PetroChina Changqing Oilfield Company, Xi'an 710020, China

Abstract Total organic carbon (TOC) content is a crucial evaluation parameter in the process of shale gas exploration and development. Marine-continental transitional shale is characterized by strong heterogeneity and thin single-layer thickness. The discrete TOC data measured by experimental methods are unable to accurately reflect the reservoir characteristics of marine-continental transitional shale. In this paper, a multivariate nonlinear regression prediction model (R-MNR) was established, and the model was applied to predict the TOC content of shale for the first time. The ΔlgR model, multiple linear regression model (MLR), BP neural network model (BP model), and R-MNR model were built to predict the TOC of shale in Benxi Formation. The coefficient of determination (R^2), mean-absolute-percentage-error (MAPE), root-mean-square-error (RMSE), and the number of input layer parameters (NILP) were employed to assess the efficacy of the model through the analytic hierarchy process (AHP) method. The total weight of R-MNR is 0.361, and that of BP model is 0.336. The weights of the two traditional models are 0.104 and 0.199, respectively. The results indicate that the R-MNR is comparable to the BP model in terms of prediction accuracy, and both models are significantly more accurate than the traditional prediction model. The R-MNR is capable of obtaining a clear TOC prediction formula, which is convenient for verification and promotion. During the training process of the R-MNR, the influence of each parameter and coupling relationship on the prediction results is elucidated, which enables researchers to gain a deeper understanding of the geophysical significance and geological process of the model. The result of this study suggests that the R-MNR can be employed to predict the TOC content of marine-continental transitional shale effectively in the future.

Keywords TOC prediction, shale reservoir, unconventional oil and gas resources, R Language, multiple nonlinear regression

1 Introduction

Natural gas is the cleanest form of fossil energy, and it is the "bridge" from fossil energy to new energy (Zou et al., 2016, 2018; Safari et al., 2019; Bugaje et al., 2022). The world is increasingly interested in the exploitation and utilization of shale oil and gas resources. The total organic carbon (TOC) content of shale describes the abundance of organic matter in shale, which is an important parameter in the evaluation of shale gas reservoirs, and this parameter can effectively reflect the hydrocarbon potential of shale (Zhao et al., 2018; Nie et al., 2021; He et al., 2023; Wang et al., 2024b). At the same time, organic matter, as part of the mineral composition, also directly affects the quality, pore development, and microstructure of shale, thus controlling important parameters such as porosity, permeability, and rock mechanical properties of shale reservoirs (Kleber et al., 2021; Wang and Yang, 2024).

According to the different depositional environments,

Received September 13, 2024; accepted December 30, 2024

E-mail: mengyanjun@tyut.edu.cn

shale can be divided into marine shale, continental shale, and marine-continental transitional shale. The world's known shale oil and gas-producing reservoirs are mainly marine shales. The shales in the basins of the United States are dominated by marine shales, while the marine-continental transitional shales in the basins of China are widely distributed. Compared with marine shale, marine-continental transitional shale has the disadvantages of strong heterogeneity, a low degree of thermal evolution, etc (Wang et al., 2022a; Cao et al., 2023).

Currently, shale TOC is mainly measured by the geochemical method, which is the most reliable way to obtain the TOC data. However, in the process of practical application, there are problems such as the high cost of core collection, discontinuous core samples, and the inability to finely characterize the distribution of TOC in the reservoir. Meanwhile, due to the poor continuity of the distribution of marine-continental transitional shale planarly, the use of discrete TOC data will cause large errors in the shale reservoirs evaluation in a large area. Therefore, the development of an accurate, efficient, and continuous TOC prediction method is of paramount importance for the exploration and production of shale reservoirs, particularly for the exploration and production of marine-continental transitional shale reservoirs (Asante-Okyere et al., 2021; Zhang et al., 2022).

The response characteristics of logging are found to be related to the total organic carbon (TOC) content of shale (Wang et al., 2022b). The prediction of shale TOC can be achieved by analyzing the mapping relationship between the logging curves and the measured TOC. Schmoker identified the correlation between lithological density logging and TOC as early as 1979, and effectively predicted the TOC of hydrocarbon source rock based on the density logging curves (Schmoker, 1981; Wang et al., 2024a). Passey et al. (2010) proposed a TOC estimation model based on acoustic time difference logging, neutron logging, density logging, and resistivity logging. The CARBOLOG® (Carbon Organic Log) company employs a combination of density logging, acoustic time difference logging, porosity logging, and natural gamma logging to predict the organic matter characteristics of a given sample (Bessereau et al., 1991). These methodologies have been implemented in actual production settings and have yielded some outcomes (Hu et al., 2021; Liu et al., 2021; Liu et al., 2023). However, these traditional prediction methods have inherent limitations, including a lack of consideration of the coupling relationship between logging parameters and the error introduced by manually selecting the baseline.

The advancement of artificial intelligence technology has led to the increasing utilization of machine learning algorithms in the prediction of shale TOC. Compared with traditional prediction methods, various algorithms are very effective in solving the multicollinearity problem among logging data. A substantial amount of evidence demonstrates that neural networks are capable of attaining superior prediction outcomes (Yu et al., 2017; Zhu et al., 2018; Mahmoud et al., 2019; Mandal et al., 2021; Chan et al., 2022). Nevertheless, these techniques present significant challenges in optimizing the network structure and adjusting the prediction model. Additionally, the neural network is unable to derive an explicit formula between the logging curves and the TOC.

In this paper, a multivariate nonlinear regression TOC prediction model (R-MNR) based on the R language is proposed. It was applied to predict the TOC content of shale in Benxi Formation in the Ordos Basin for the first time. The multivariate nonlinear regression model is capable of considering the complex coupling relationship between the independent variable parameters and the dependent variables, rather than just considering the linear relationship of the parameters.The R-MNR model can provide a clear causal relationship and parameter weights and can explain the specific impact of each independent variable on the target variable. This enables researchers to gain a profound understanding of the physical or geological processes in the system, not just the results of predicted parameters. To evaluate the effectiveness of the prediction model, this paper adopts the analytic hierarchy process (AHP) method to compare the different aspects of four models, the ΔlgR model, multiple linear regression model (MLR), BP neural network model (BP Model), and R-MNR model, so as to validate the effectiveness of the model.

2 Geological setting and data

2.1 Geological setting

The Ordos Basin is located on the western edge of the North China Craton, starting from the Yinshan Mountains in the north and reaching the Qinling Mountains in the south. The basin is bounded by the Lvliang Mountains in the west and the Tengger Desert in the east. It is the second largest sedimentary basin within the land of China, with an area of about 37×10^4 km^2 (Fig. 1). The evolution of the Ordos Prototype Basin has gone through multiple stages, and the basin has evolved mainly through the Indo-Chinese, Yanshan, and Himalayan movements and other phases of the tectonic movement cycle. During the Late Paleozoic, the Ordos Basin was deposited with the descent of the North China Plate, and the Permo-Carboniferous sediments were mainly deposited in the basin (Bao et al., 2014; Ju et al., 2017; Shi et al., 2019; Liu et al., 2022; Yan et al., 2023a; Liu et al., 2024). The main coal-bearing seams are the Carboniferous Benxi Formation, Permian Taiyuan Formation, and Shanxi Formation (Yan et al., 2023b; Zhao et al., 2024). Among them, the Benxi Formation at the bottom of the Carboniferous is a typical marine-continental transitional

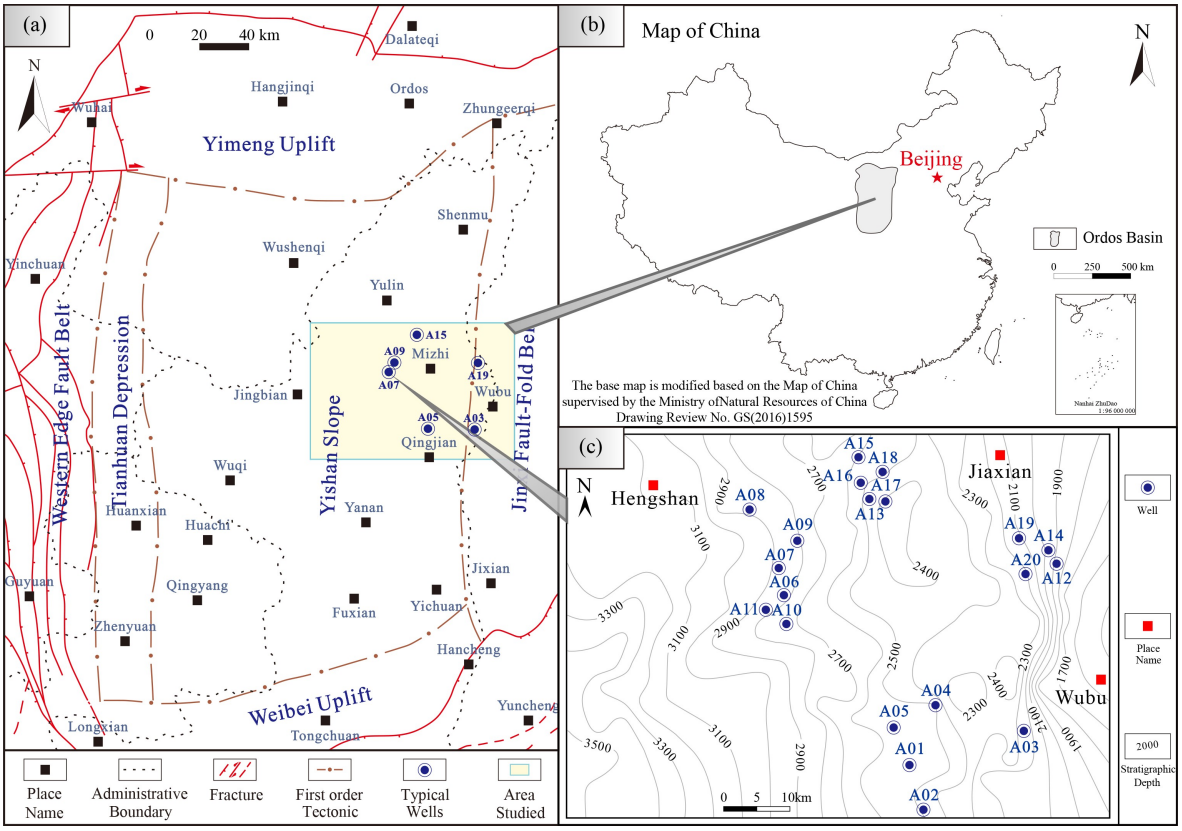

Fig. 1 Map of the study area. (a) The location of the study area in Ordos Basin. (b) The location of Ordos Basin in the national map. (c) The buried depth and typical well location map of Benxi Formation in the study area.

sedimentary environment (Fig. 2). The Iron-aluminum mudstone is present at the bottom of the Benxi Formation, and the No.8 coal seam is developed at the top. The mudstone, sandy mudstone, and limestone occur in the middle section. The Benxi Formation has good gas source addition and sealing conditions, which is conducive to the occurrence of shale gas.

The study area is located in the eastern Ordos Basin, and the research object is the marine-continental transition shale in the Benxi Formation. The thickness of the Benxi Formation in the study area ranges from 17 to 80 m (Fig. 3(a)), and the thickness of mudstone in the Benxi Formation ranges from 5 to 62 m (Fig. 3(b)). The thickness of the Benxi Formation in the study area gradually increases from west to east. The main shale layer in the Benxi Formation is the Jinci member, and the development of No. 8 and No. 9 coal seams in the Jinci member provides a stable gas source for the shale reservoirs. The coal seams play an important role in the physical closure or hydrocarbon enrichment closure of the shale reservoirs. The thickness of the Jinci member is distributed between 9 and 55 m (Fig. 3(c)), and the thickness of mudstone in the Jinci member varies from 2 to 40 m (Fig. 3(d)). The Jinci member is similar to the Benxi Formation and shows the same stratigraphic thickness variation characteristics.

2.2 Data source

In this study, 20 wells with 105 core samples were collected from the Hengshan-Wubu area in the Ordos Basin. The sampled wells are distributed throughout the study area, and the locations of the sampling wells are shown in Fig. 1(c).

The collected shale samples were geochemically tested at the Key Laboratory of Unconventional Oil and Gas, Research Institue of Petroleum Exploration and Development (RIPED), PetroChina. The TOC of the shale sample was obtained by the Chinese national standard “Determination of total organic carbon in sedimentary rocks” (GB/T 19145-2022). Leco Carbon and Sulfur Tester CS230 was used in the experiment, and the testing temperature was 20°C. The experimental steps were as follows. First, each shale sample was ground to 200 mesh, weighed at 1.0 g, and put into a quartz crucible. 5% dilute hydrochloric acid was added and the mixture was then heated at a temperature of 80°C to remove inorganic carbon. The samples were washed with purified water and dried at 60°C for treatment. Finally, the treated samples were put into a CS230 carbon and sulfur meter to determine the TOC.

The logging curve data were provided by the Research Institute of Exploration and Development of PetroChina Changqing Oilfield Company, and the logging data of

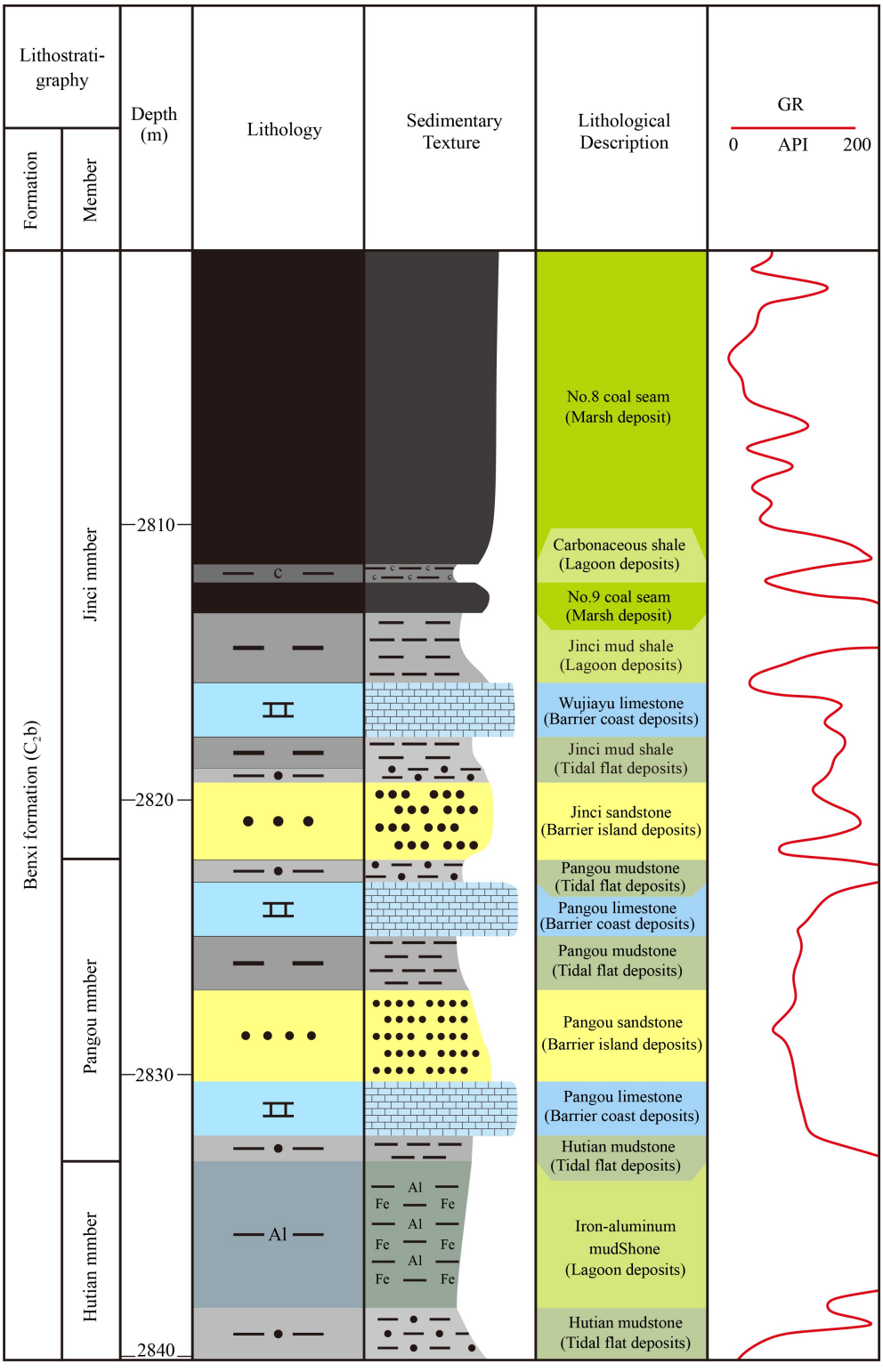

Fig. 2 Lithostratigrphic histogram of the study area.

sampling wells A1−A20 were extracted by Explorer software. The extracted logging curves mainly include acoustic time difference logging (AC), natural gamma logging (GR), natural potential logging (SP), density logging (DEN), shallow lateral resistivity logging (RLLS), deep lateral resistivity logging (RLLD), potassium logging (K), uranium logging (U), thorium logging (TH), and neutron logging (CNL).

The Origin, Matlab, SPSS, and R Studio software were used for data analysis and processing. In the model-building process, the measured TOC of shale was randomly divided into a training group and a test group,

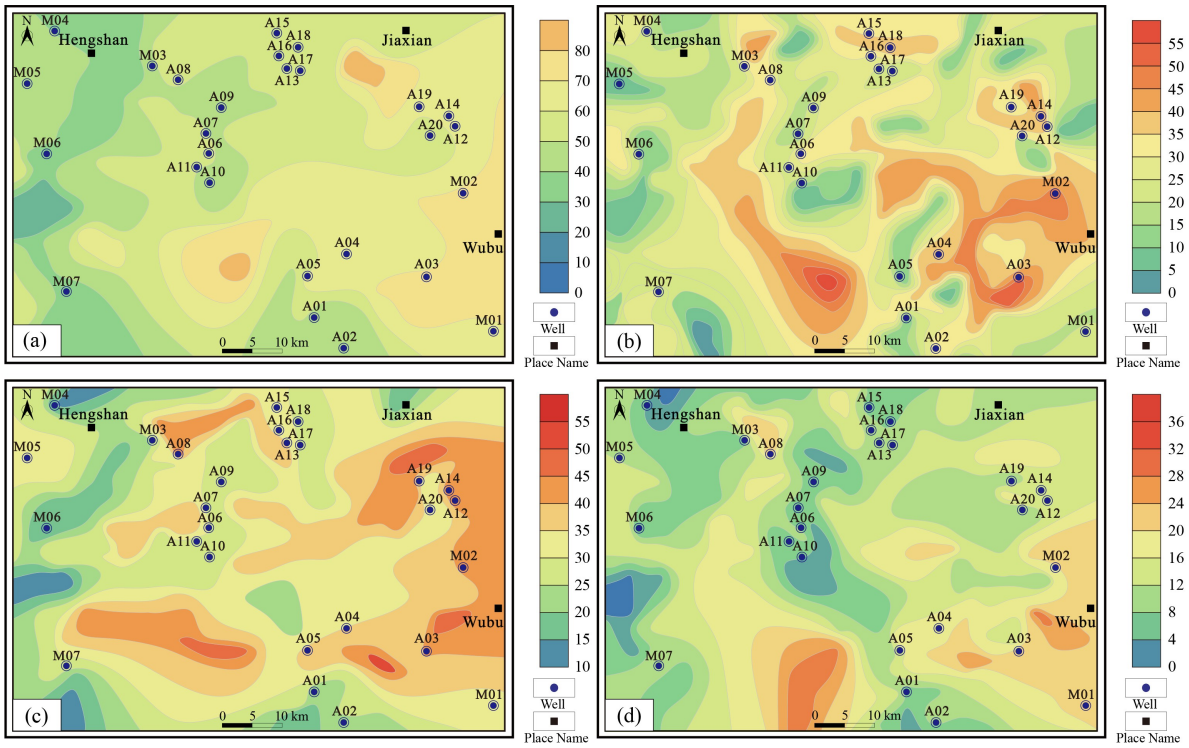

Fig. 3 Strata thickness distribution of the study area. (a) Benxi Formation. (b) Shale of the Benxi Formation. (c) Jinci Member. (d) Shale of the Jinci Member.

where the training group contained 70 data sets and the test group contained 20 data sets. The test group was not involved in the model-building process.

2.3 Organic matter characteristics and well logging data

Sample kerogen maceral identification and type in the Jinci member of the Benxi Formation in the study area shows that the macerals of shale kerogen are mainly inertinite and sapropelic, containing a small amount of vitrinite (Fig. 4). The distribution of TI-type indices ranges from −87 to 69.3, which shows that the organic matter type is dominated by Type II_1 and Type III. The results of vitrinite reflectance measurements show that the shale in the study area has a high degree of maturity, with R_o ranging from 1.60% to 2.41% and an average value of 1.95%. According to the evaluation criteria of the maturity of organic matter evolution in hydrocarbon source rocks, the shale samples have a high degree of thermal evolution and have entered the stage of forming gas, while individual samples have reached the stage of over-maturity (Faiz et al., 2022).

The logging data corresponding to the samples tested for TOC were extracted, and the specific parameters of the logging data are shown in Table 1. Excluding the anomalous measurements, the TOC of 105 core samples in the study area ranges from 0.12% to 19.1%, with an average value of 3.5% (Fig. 5). The medium and high organic matter content dominates in all samples. The three test results of organic matter type, organic matter maturity, and TOC indicate that the shale in the study area has high hydrocarbon potential.

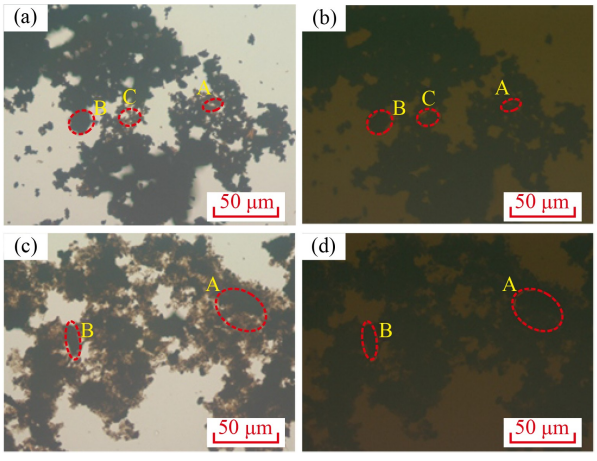

Fig. 4 Organic matter characteristics of shale in the area studied. (a) Well A19, 2099.48 m, mudstone. The kerogen type is type III. (b) Well A19, 2099.48 m, carbonaceous mudstone. The kerogen type is type III. (c) Well A19, 2102.20 m. The kerogen type is II_1. (d) Well A19, 2102.20 m, Silty mudstone. The kerogen type is II_1.

3 Model and methodology

3.1 ΔlgR model

The ΔlgR model was initially proposed by Exxon and

Table 1 Statistical details of ten logging parameters for all samples

Logging parameter	Mean	Min	Max
CNL (%)	36.91	10.30	73.94
SP (MV)	59.75	48.13	70.62
K (10^{-6})	1.23	0.17	3.80
GR (API)	130.31	39.70	413.33
AC (μs/m)	240.70	172.69	337.57
DEN (g/cm^3)	2.46	1.70	2.82
RLLS (Ω·m)	126.39	7.00	3377.08
RLLD (Ω·m)	194.05	7.67	8276.24
TH (10^{-6})	16.56	2.94	45.64
U (10^{-6})	7.61	1.77	39.43

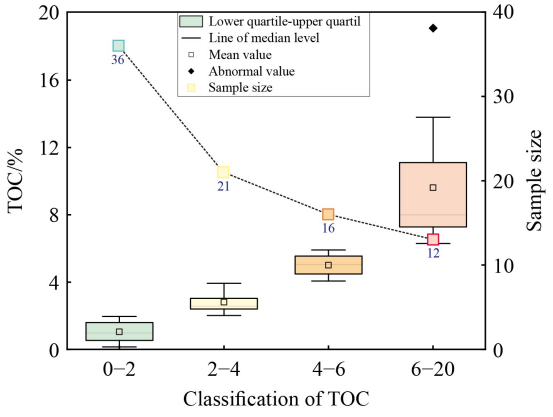

Fig. 5 Distribution of measured TOC in the area studied.

Esso, and subsequently refined and applied by Passey et al. (2010). In comparison to non-hydrocarbon source rocks, organic matter in hydrocarbon source rocks is distinguished by a low acoustic time difference logging and a high resistivity logging response. This results in the two logging curves exhibiting disparate morphological characteristics within the layers of varying organic matter enrichment (Fig. 6). The model is founded upon the linear correlation between TOC and Δlg$R$. The specific formula is as follows:

$$\Delta \lg R = \lg(R/R_{\text{baseline}}) + K(\Delta t - \Delta t_{\text{baseline}}), \quad (1)$$

$$\text{TOC} = 10^{(2.297-0.1688\times L_{\text{OM}})\times \Delta \lg R}. \quad (2)$$

In Eq. (1), R is the shallow lateral resistivity, R_{baseline} is the resistivity of the non-hydrocarbon-sourced rock section, Δt is the acoustic time difference, and $\Delta t_{\text{baseline}}$ is the acoustic time difference of the non-hydrocarbon-sourced rock section. In Eq. (2), L_{OM} is the maturity of the organic matter in the formation.

The advantage of this model is that the two selected logging curves are highly responsive to organic matter. However, a potential limitation is the lack of consideration of the influences from other logging

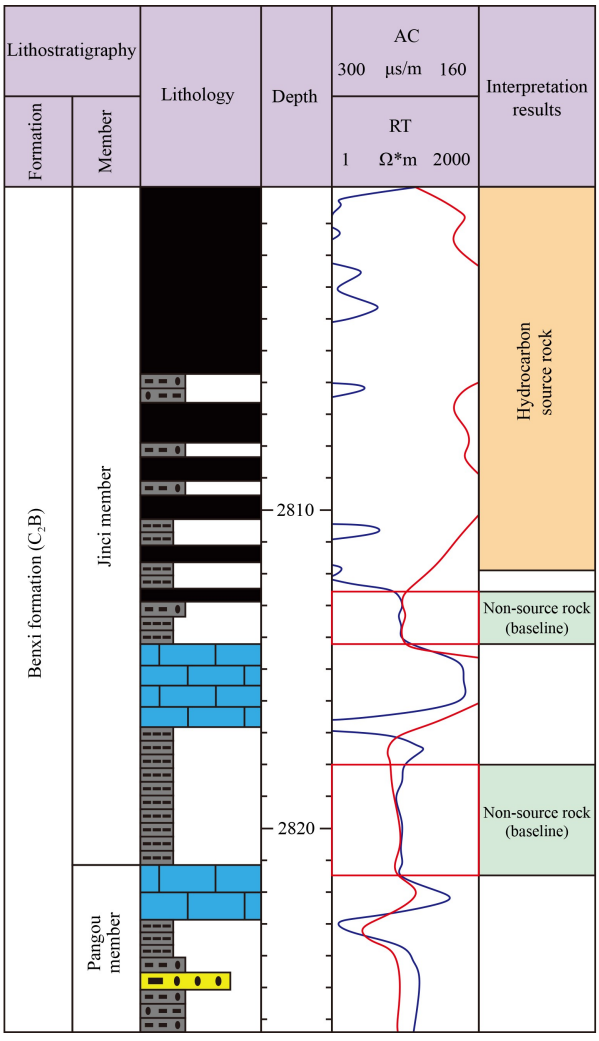

Fig. 6 Schematic diagram of ΔlgR model.

parameters. Furthermore, the method necessitates the manual identification of non-hydrocarbon source rock segments. Meanwhile, the frequent interbedding of lithologies in marine-continental transitional shale results in non-unique baseline values at the same well location. Hu et al. improved on the original Δlg$R$ model and proposed an improved Δlg$R$ model (Hu et al., 2021; Liu et al., 2023). This is based on the observation that density logging data of hydrocarbon source rocks are significantly lower than those of non-hydrocarbon source rocks, and Eq. (3) can be simplified as follows:

$$\text{TOC} = (\Delta \lg R + B\Delta t + C)/\text{DEN}. \quad (3)$$

The improved Δlg$R$ method avoids the inaccuracy caused by artificial baseline selection, and the addition of density logging improves the accuracy of prediction results. By leveraging the TOC data acquired from the samples and the well logging data, the Δlg$R$ model was constructed according to the specifications outlined in Eq. (3). This was achieved using Origin software. This led to the formulation of the expression encapsulating the shale

TOC prediction model based on the improved ΔlgR method, as detailed in Eq. (4):

$$\mathrm{TOC} = (2.57044\lg R + 0.12618\Delta t - 24.67401)/\mathrm{DEN}, \quad (4)$$

where 2.57044, 0.12618, and −24.6701 are constants obtained from least squares analysis of the samples from the study area using Origin software.

3.2 Multiple Linear Regression Model (MLR)

Shale TOC is controlled by a variety of factors, and the prediction using a multivariate model is significantly better than that using a single-variable model (Saporetti et al., 2023). Multiple linear regression, as a classical multivariate statistical analysis method, can be used to analyze the linear relationship between different logging curves and TOC. The mathematical model of multiple linear regression is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon, \quad (5)$$

where y is the dependent variable, x_1, x_2,...,x_n are the independent variables, β_0, β_1,...,β_n are the biased regression coefficients, ε is the random error, the residual variance, which independently obeys the normal distribution $N(0, \sigma^2)$.

Based on the geophysical significance of various logging curves, a total of 10 logging parameters including GR, AC, CNL, DEN, SP, RLLD, RLLS, U, K, and TH were selected. The correlation and P-value test between TOC content and different logging data were analyzed by SPSS software. The results demonstrate a significant relationship between the TOC content of shale and AC, DEN, CNL, U, and TH (Table 2). The correlation with GR, SP, K, RLLD, and RLLS was found to be weak (Fig. 7). Table 3 presents the statistical parameters for the aforementioned five logging curves. Therefore, this study selected five logging parameters (AC, DEN, CNL, U, and TH) with strong correlations as the main parameters for predicting shale organic carbon content.

With the above five logging parameters as independent variables and TOC data as the dependent variable, the TOC prediction model was obtained by multiple regression analysis using Origin software:

$$\mathrm{TOC} = 17.331 + 0.047 \times \mathrm{AC} - 9.738 \times \mathrm{DEN} - 0.034 \times \mathrm{CNL} - 0.045 \times \mathrm{U} + 0.098 \times \mathrm{TH}, \quad (6)$$

where AC is acoustic time difference logging. DEN is density logging. CNL is compensate neutron logging. U is uranium elemental logging. TH is thorium elemental logging.

3.3 BP Neural Network Model (BP)

The BP neural network is capable of learning and storing a substantial number of input-output mapping relationships. Consequently, it can be employed to address nonlinear issues with copious input data that are challenging to express in explicit formulas. The learning rule for BP neural networks is to utilize gradient descent, which entails the continuous adjustment of the network

Table 2 The P-value test between logging parameters and TOC

	AC/(μs·m^{-1})	DEN/(g·cm^{-3})	CNL/%	U/10^{-6}	TH/10^{-6}	GR/API	SP/MV	K/%	RLLD/OMM	RLLS/OMM
P	0.000***	0.000***	0.001***	0.006***	0.024**	0.849	0.521	0.123	0.213	0.110

Fig. 7 Correlations of measured TOC and logging parameters. Note: The values displayed in the figure represent the correlation coefficients. The correlation between the parameters is primarily represented by the color depth.

Table 3 Statistical details of five evaluation indexes for the prediction model

	AC/ ($\mu s \cdot m^{-1}$)	DEN/ ($g \cdot cm^{-3}$)	CNL/ %	U/ 10^{-6}	TH/ 10^{-6}
Mean	246.13	2.49	37.95	7.22	17.57
Min	172.69	1.70	10.30	1.77	3.86
Max	337.57	2.52	73.94	39.43	45.64

weights and thresholds by back propagation with the objective of minimizing the sum of the squared errors of the network. The network structure of the BP neural network comprises input, hidden, and output layers. Neurons within each layer are independent of one another, while neurons between layers are connected. A diagram of the network structure is provided in Fig. 8. The input layer of the neural network represents the initial point of information input, the hidden layer constitutes the processing stage, and the output layer represents the final stage of information output. The learning process is divided into two distinct phases: forward computation and backpropagation. In the forward propagation phase, the weights of the hidden layer are applied to the input data in a layer-by-layer manner, resulting in a propagation of the output to the output layer. In the event that the discrepancy between the predicted and actual outputs is not within the desired range, the error signal will be transmitted in a layer-by-layer manner from the output layer back to the input layer, where it will then be transmitted once more to the output layer after the hidden layer has been re-weighted. These steps should be repeated until the error is within reasonable limits (Wang et al., 2019; Li et al., 2022).

The forward computation process is divided into two parts: from the input layer to the hidden layer, as shown in Eq. (6), and from the hidden layer to the output layer as shown in Eq. (7). The network computation process uses the values of each layer multiplied by the corresponding weights and bias variables. The backpropagation process makes the error smaller by continuously calculating the error between the output layer and the desired value and adjusting the network parameters calculated by Eq. (5). Finally, the weights are recalculated and updated back to the input layers Eqs. (9) and (10), cycling the above steps until the training reaches the desired value.

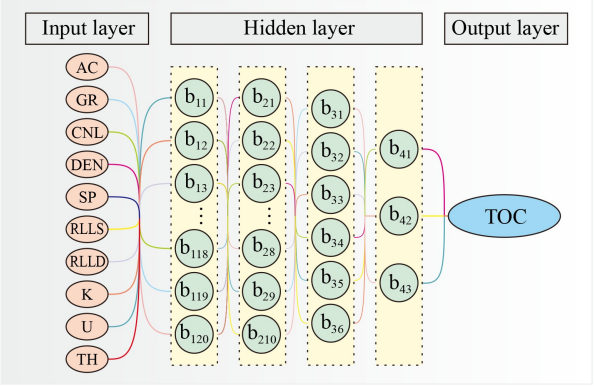

Fig. 8 Structure diagram of BP Model.

Input layer to hidden layer:

$$a_h = \sum_{i=1}^{d} v_{ih} x_i + \theta_h. \tag{7}$$

Hidden layer to output layer:

$$\beta_h = \sum_{h=1}^{q} \omega_{hj} b_h + \theta_j, \tag{8}$$

where x is the input layer, b is the bias term, i is the number of input layers, h is the number of hidden layers, j is the number of neurons in the hidden layer, v is the weight from the input layer to the hidden layer, w is the weight from the hidden layer to the output layer, and θ is the activation function (the activation function tends to be a nonlinear function used to achieve a nonlinear mapping of the network):

$$E = \frac{1}{2} \sum_{k=1}^{2} (y_k - T_k)^2, \tag{9}$$

where x is the output value and T is the desired value;

$$\Delta\omega_{ij} = (l) E_y k, \tag{10}$$

$$\omega_{ij} = \Delta\omega_{ij} + \omega_{ij}, \tag{11}$$

where ω_{ij} is the weight, l is the learning rate (the control objective function converges to a minimum value in a suitable time), $\Delta\omega_{ij}$ and is the value of weight change.

The BP model was constructed using Matlab software. As the neural network is capable of adjusting the weight of each input layer parameter autonomously, this paper elects to utilize the logging parameters that are most commonly employed for TOC prediction as the input variables of the model. The 70 data points of the training group were selected as the data set. By continuously adjusting the network, the number of hidden layers is set to 4, with the number of neurons in each hidden layer being 20, 10, 6, and 3, respectively, and the number of output layers is 1. The L-M (Levenberg-Marquardt) algorithm is employed to learn the data and construct the neural network model. The number of nodes in the hidden layer was calculated using Eq. (12):

$$\sum_{i}^{n} C_M^i > k. \tag{12}$$

In Eq. (12), n is the number of neurons in the input layer, M is the number of neurons in the hidden layer, i is a positive integer taking the value $0-n$, and k is the

number of samples. The specific parameters of the neural network model are shown in Table 4.

3.4 Multiple Nonlinear Regression Model Based on R Language (R-MNR)

In regression analysis, a regression model comprising two or more independent variables is designated as multiple regression. Eq. (6) represents a multivariate linear prediction model of shale TOC, established through linear regression. However, the intricate geological implications between TOC and logging response parameters indicate that their relationship is not a simple linear one. MLR model merely constructs a prediction formula by analyzing the linear relationship and significance characteristics between several independent variable parameters and dependent variables, without considering the complex coupling relationship between parameters (Xu et al., 2022). Although the MLR model is still the most commonly used prediction method in engineering, it is deficient in its inability to consider the geological significance between TOC and logging response parameters. Therefore, this paper built a multivariate nonlinear TOC prediction model with logging parameters as independent variables. Statistical analysis was also used to study the effect of logging parameters on the model's significance and the interaction between the parameters. The mathematical model of multivariate nonlinear regression is shown in Eq. (13):

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n+\beta_{n+1}x_1x_2+\cdots+\beta_{\frac{n(n+1)}{2}}x_{n-1}x_n+\cdots+\beta_{\frac{n^3+2n}{3}}x_1\cdots x_n+\varepsilon. \quad (13)$$

In Eq. (13), y is the dependent variable, x_1, x_2,..., x_n are the independent variables, β_0, β_1,..., β_n are the biased regression coefficients, ε is the random error, i.e., the residuals.

The relationship between TOC and ten logging parameters, including GR, AC, CNL, DEN, SP, RLLS, RLLD, U, K, and TH, was analyzed by R Studio software. To guarantee the precision of the R-MNR model and to prevent the formula from becoming unduly complex, this study examined a range of logging parameters and their interrelationships. The multivariate nonlinear regression model constructed by five parameters such as GR, AC, DEN, CNL, and SP (Case 1) and six parameters such as GR, AC, DEN, CNL, SP, and U (Case 2) has similar prediction accuracy. However, due to too many parameters selected, the number of characters in the prediction formula of Case 2 is far more than that of Case 1, and there is no significant improvement in the prediction accuracy. The prediction accuracy of the model built by selecting five parameters such as AC, DEN, SP, U, and TH (Case 3) and four parameters such as GR, AC, DEN, and CNL (Case 4) is much lower than that of Case 1 and Case 2 (Fig. 9). Therefore, this study selected the R-NMR model built in Case 1 to predict the TOC. The statistical analysis results of Case 1 are shown in Table 5. Some logging curves have little contribution to the prediction of TOC in the process of linear analysis, but considering the coupling relationship between the curve and other logging curves, the logging parameters have a significant impact on the prediction of total organic carbon content.

The statistical significance of the model is usually determined using the F-statistic, which is shown in Table 5, and the model is significant. The P-value is used to determine the significance of the parameters in the model. The smaller the P-value is, the greater the contribution to the predictive effect of the model. After statistical analysis, the logging parameters used to build the model and their coupling relationships have a P-value of less than 0.01, which is considered to be extremely significant. The correlation coefficient R^2 of the training group reaches 0.8549, which indicates that the fitting relationship between shale TOC and the input logging parameters is significant.

The specific formula of the R-MNR built with the above five logging parameters combined with the selected parameter combination as the independent variable and the organic carbon content as the dependent variable is shown in Eq. (14):

Table 4 BP neural network parameter settings

Parameter	Value of function
Number of neurons input layer	10
Number of neurons in hidden layer	4
Number of neurons in output layer	20, 10, 6, 3
Training function	Levenberg-Marquardt
Iterative accuracy	10^{-8}
Number of iterations	10^4
Learning rate	5×10^{-5}

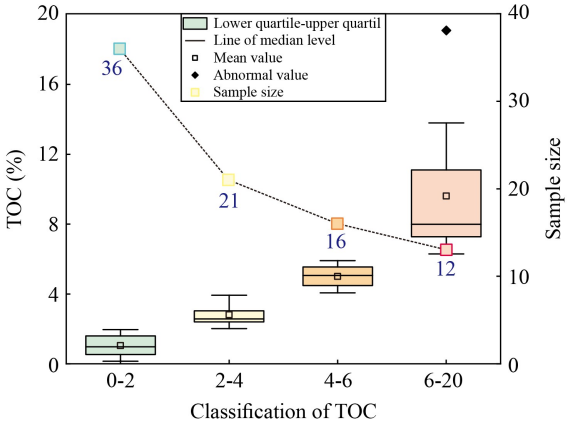

Fig. 9 Comparison of the application effects of R-MNR model.

Table 5 Statistical analysis for TOC

	Estimate	Std. error	*t* value	*P* value
(Intercept)	−1.414e+04	2.706e+03	−5.225	3.73e−06
SP	2.385e+02	4.521e+01	5.275	3.14e−06
AC	4.681e+01	9.765e+00	4.793	1.62e−05
DEN	5.385e+03	1.085e+03	4.964	9.13e−06
CNL	4.450e+02	7.767e+01	5.729	6.48e−07
SP×AC	−7.993e−01	1.632e−01	−4.898	1.14e−05
SP×DEN	−9.079e+01	1.807e+01	−5.026	7.39e−06
AC×DEN	−1.785e+01	3.984e+00	−4.479	4.62e−05
SP×CNL	−7.711e+00	1.315e+00	−5.863	4.05e−07
AC×CNL	−1.326e+00	2.696e−01	−4.916	1.07e−05
DEN×CNL	−1.580e+02	3.035e+01	−5.206	3.98e−06
SP×AC×DEN	3.045e−01	6.628e−02	4.594	3.17e−05
SP×AC×CNL	2.319e−02	4.539e−03	5.109	5.55e−06
SP×DEN×CNL	2.745e+00	5.111e−01	5.370	2.26e−06
AC×DEN×CNL	4.575e−01	1.079e−01	4.239	0.000102
GR×DEN×CNL	−4.688e−02	6.906e−03	−6.789	1.55e−08
GR×SP×CNL	2.061e−03	2.945e−04	6.999	7.39e−09
GR×AC×CNL	−8.914e−04	1.367e−04	−6.523	3.97e−08
SP×AC×DEN×CNL	−8.031e−03	1.805e−03	−4.450	5.09e−05
GR×SP×AC×CNL	8.539e−06	1.598e−06	5.343	2.48e−06
GR×AC×DEN×CNL	5.363e−04	7.861e−05	6.822	1.38e−08
GR×SP×AC×DEN×CNL	−6.568e−06	1.004e−06	−6.538	3.76e−08

Notes: *F*-statistic: 13.46 on 21 and 48; *R*-sq: 0.8549.

$$\begin{aligned}\mathrm{TOC} = {} & 2.39\times e^{2}\times \mathrm{SP} + 4.68\times e\times \mathrm{AC} + 5.39\times e^{3}\times \\ & \mathrm{DEN} + 4.45\times e^{2}\times \mathrm{CNL} - 7.99\times e^{-1}\times \mathrm{SP}\times \mathrm{AC} - 9.08\times \\ & e\times \mathrm{SP}\times \mathrm{DEN} - 1.78\times e\times \mathrm{AC}\times \mathrm{DEN} - 7.71\times \mathrm{SP}\times \\ & \mathrm{CNL} - 1.33\times \mathrm{AC}\times \mathrm{CNL} - 1.58\times e^{2}\times \mathrm{DEN}\times \mathrm{CNL} + 3.04\times \\ & e^{-1}\times \mathrm{SP}\times \mathrm{AC}\times \mathrm{DEN} + 2.32\times e^{-2}\times \mathrm{SP}\times \mathrm{AC}\times \mathrm{CNL} + \\ & 2.74\times \mathrm{SP}\times \mathrm{DEN}\times \mathrm{CNL} + 4.58\times e^{-1}\times \mathrm{AC}\times \mathrm{DEN}\times \\ & \mathrm{CNL} - 4.69\times e^{-2}\times \mathrm{DEN}\times \mathrm{CNL}\times \mathrm{GR} + 2.06\times e^{-3}\times \\ & \mathrm{SP}\times \mathrm{CNL}\times \mathrm{GR} - 8.91\times e^{-4}\times \mathrm{AC}\times \mathrm{CNL}\times \mathrm{GR} - 8.03\times \\ & e^{-3}\times \mathrm{SP}\times \mathrm{AC}\times \mathrm{DEN}\times \mathrm{CNL} + 8.54\times e^{-6}\times \mathrm{SP}\times \mathrm{AC}\times \\ & \mathrm{CNL}\times \mathrm{GR} + 5.36\times e^{-4}\times \mathrm{AC}\times \mathrm{DEN}\times \mathrm{CNL}\times \mathrm{GR} \\ & - 6.57\times e^{-6}\times SP\times AC\times DEN\times CNL\times GR - 1.413736\times e^{4}. \end{aligned} \quad (14)$$

3.5 Analytic Hierarchy Process (AHP)

To evaluate the efficacy of each prediction model, this study employed the Analytic Hierarchy Process (AHP) (Vaidya and Kumar, 2006; Meng et al., 2014). The AHP method initially identifies the pertinent factors associated with the problem to be solved. The problem to be solved and the related factors are categorized into a multilevel structural model comprising target, criterion, and solution layers. In the second step, weights are assigned to each factor within the hierarchy through the application of expert prior knowledge or empirical values. The elements of the criterion layer are then compared with each other to establish a judgment matrix. The weight of each index is then calculated for according to the judgment matrix. The hierarchical single ranking and its consistency test are completed. The total weight value is obtained through the total hierarchical ordering, thus providing the comprehensive multi-objective evaluation results. The optimal solution to the problem is then obtained (Fig. 10).

In this paper, we use the TOC prediction model evaluation as the target layer, the number of input logging parameters (NILP), the coefficient of determination (R^2), mean-absolute-percentage-error (MAPE), and root-mean-square-error (RMSE) as the criterion layer, and four TOC prediction models as the solution layer to establish the AHP evaluation model.

The three most frequently employed metrics for the assessment of correlation are RMSE, MAPE and R^2. The mean-squared-error (MSE) represents the mean of the squares of the differences between the measured and predicted values. It is used to measure the variance of the residual. The root-mean-square-error (RMSE) is defined

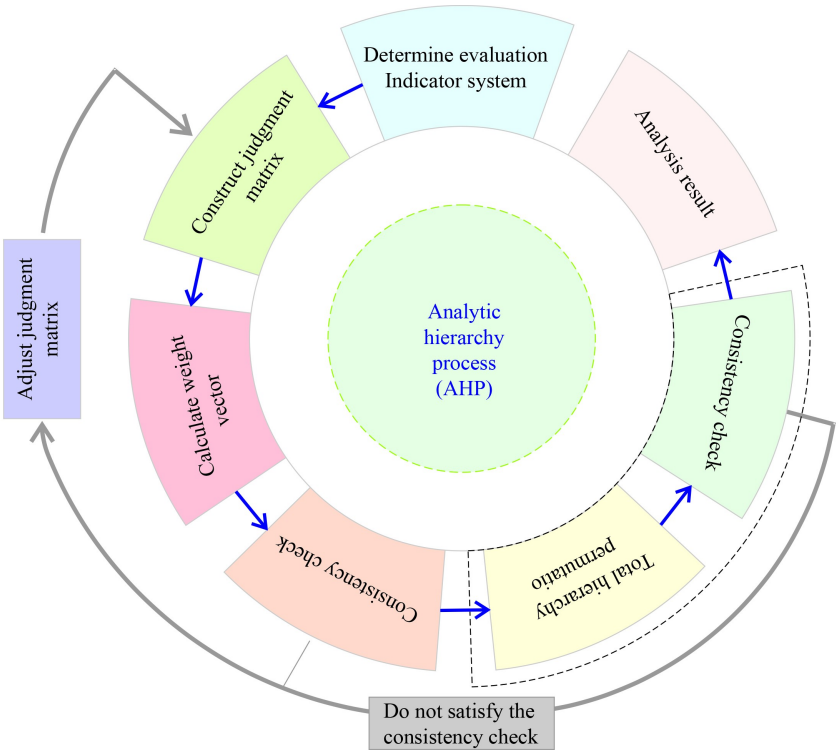

Fig. 10 Structure diagram of AHP.

as the square root of the mean-square-error (MSE). The root-mean-square-error (RMSE) is the most frequently employed metric for the assessment of regression models. The root-mean-square-error (RMSE) is a more commonly used metric than the mean-squared-error (MSE). A lower RMSE value indicates a superior model fit. The mean-absolute-error (MAE) represents the mean of the difference between the measured and predicted values, which is employed as a direct measure of the mean of the residuals. The mean-absolute-percentage-error (MAPE) is a variant of the mean-absolute-error (MAE) that is expressed as a percentage and is not affected by outliers. In addition to calculating the discrepancy between the fitted value and the true value, MAPE also considers the ratio between the two. A lower value of MAPE indicates a superior model. Both are employed for the assessment of the model's performance. A lower value indicates a reduced level of error. The coefficient of determination (R^2) indicates the proportion of the dependent variable that can be explained by the model. It is the most common evaluation indicator in multiple regression models (Eq. (17)). The closer the value of R^2 is to 1, the stronger the relationship between the predictor variable (s) and the response variable.

$$\mathrm{RMSE} = \sqrt{\frac{i}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{15}$$

$$\mathrm{MAPE} = \frac{i}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}, \tag{16}$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2}, \tag{17}$$

where y_i denotes true value, $\hat{y}_i$ denotes predicted value, $\bar{y}_i$ denotes denotes average true value.

4 Results and model comparisons

In this study, the collected shale samples were first tested for TOC content, and the logging data of the corresponding samples were extracted. Subsequently, the shale samples were randomly divided into a training group and a testing group. Using the measured TOC data from the training group and the well logging data, the TOC prediction model was established by applying the ΔlgR, MLR, BP Model, and R-MNR, respectively. The predicted TOC data of the four methods were compared with the measured TOC of the training group and the test group. The model exhibiting the optimal predictive performance was selected through a comparative analysis of the RMSE, MAPE, and R^2.

4.1 Traditional Predictive Model (Δlg$R$, MLR Model)

The improved Δlg$R$ model is established using three parameters: RLLS, AC, and DEN. Because kerogen has the characteristics such as low density, low sound velocity, and high resistivity, these three logging curves

can show different morphological characteristics in sections with organic-rich layer and exhibit strong logging response. The prediction model using the Δlg$R$ model could achieve an R^2 value of 0.5788 (Fig. 11(a)), a MAPE value of 0.4095, and an RMSE value of 0.5822 in the training group. The model applied in the test group was comparable to that of the training group with the R^2 value of 0.5927 (Fig. 11(b)), MAPE value of 0.3273, and RMSE value of 0.4600.

The improved Δlg$R$ model has a general impact on the prediction of shale TOC content. The model demonstrates a relatively robust prediction efficacy in the middle ($2 \leqslant$ TOC < 4) and high organic matter layers (TOC $\geqslant 4$) (Fig. 12(a)). In the low organic matter shale section (TOC < 2), the predicted value is significantly lower than the measured value, and the error value accounts for up to 30% of the measured value. The reason for this phenomenon may be that the organic-low shale has a poor response to the density logging curve, which is difficult to be captured by the density logging, resulting in a high predicted value (Wang et al., 2022b). Conversely, due to the good response of density logging to organic matter in shale, the predicted value is low. Furthermore, the majority of the sample data lie within the range of 2% to 6%, which results in the model having limited constraints on samples with either excessively high or low TOC content. This leads to significant model distortion.

Through significance analysis, the multiple linear regression model (MLR) selects the parameters that exhibit a robust correlation with TOC prediction, leading to a slight improvement over the Δlg$R$ model in terms of TOC prediction. Figure 13(a) shows the R^2 of the MLR model in the training group is 0.6101, MAPE is 0.4033, and RMSE is 0.5051. The model applied in the test group is comparable to the training group, and the R^2 between the measured value and the predicted value in the test group is 0.6896 (Fig. 13(b)). The MAPE is 0.3329 and the RMSE is 0.4145.

The MLR model has a good prediction effect on the organic carbon content of shale, and the advantages are obvious in the middle organic matter layer and the high organic matter layer (Fig. 12(b)). Similarly to the diminished efficacy of the improved Δlg$R$ model in organic-low and organic-rich shale, the applicability of the model is diminished in these contexts. The underlying cause of this phenomenon can be attributed to the notable discrepancy between the logging parameters and the varying organic matter content. The linear regression model considers only the linear relationship between the

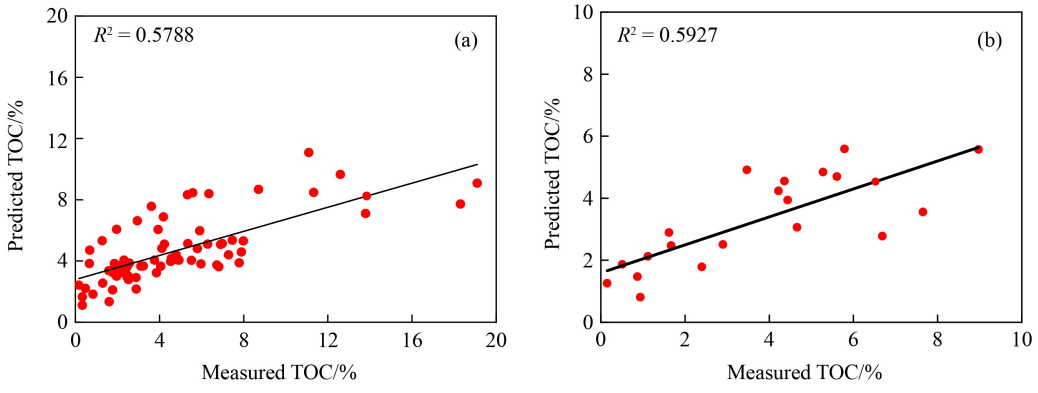

Fig. 11 Prediction performance of ΔlgR Model. (a) Training group. (b) Testing group.

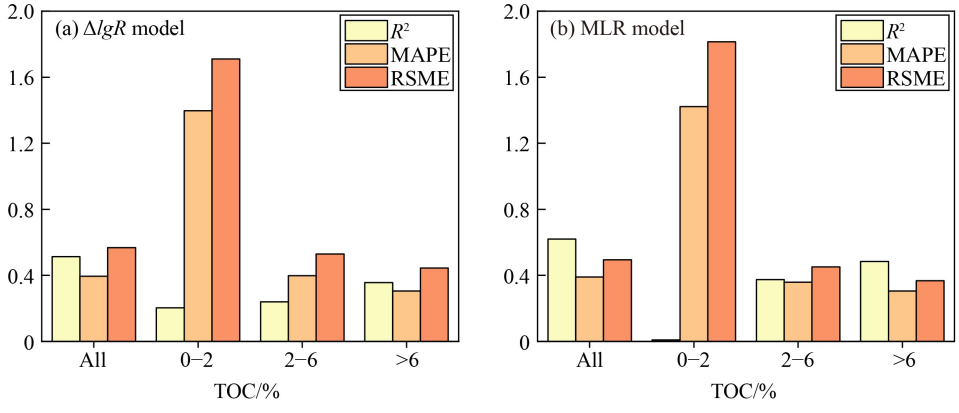

Fig. 12 Performance diagrams of traditional models. (a) Performance diagrams of ΔlgR Model. (b) Performance diagrams of MLR Model (Where, 'All' represents all samples, 0–2 represents samples with TOC between 0 and 2%, and so on. The closer R^2 is to 1, the higher the correlation of the model is. The closer the MAPE and RSME are to 0, the smaller the error of model is.).

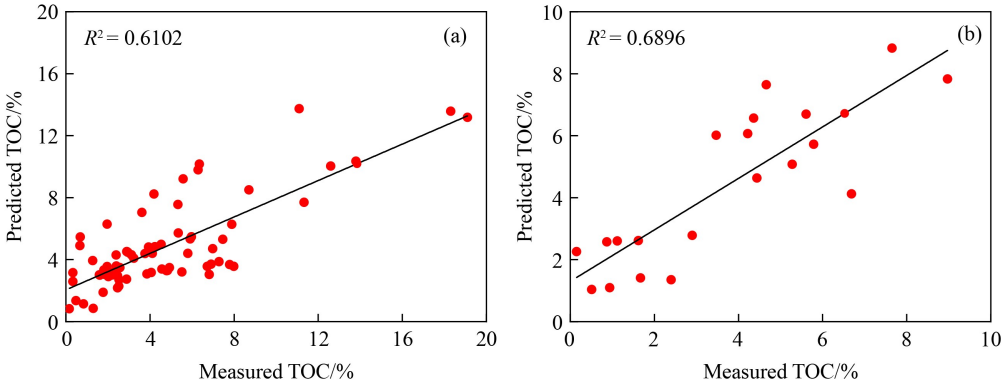

Fig. 13 Prediction performance of MLR Model. (a) Training group. (b) Testing group.

parameters, with the partial regression coefficients being less constrained to the logging parameters with large differences. This results in a distorted model in the low and high organic matter layers.

It is noteworthy that the two traditional models in the test group demonstrate superior prediction efficacy compared to models in the training group. This phenomenon can be attributed to the limited sample size of the test group. The predicted values for some data points exhibit a close alignment with the actual values, resulting in an enhanced evaluation effect for the test group. This phenomenon is commonly observed in predictions based on small samples.

4.2 BP Neural Network (BP Model)

A total of nine logging parameters are used in the BP neural network model, and unlike the traditional methods, the BP model can establish a nonlinear mapping relationship between multiple parameters. Therefore, the BP model has better performance compared with the improved ΔlgR and MLR model. The R^2 of the BP neural network model in the training group can reach 0.8764 (Fig. 14(a)), the MAPE is 0.2335, and the RMSE is 0.2878. The application effect of the model in the test group decreases significantly, and some predicted values shows obvious deviations. The correlation coefficient between the measured TOC content and the predicted value of the test group is 0.6576 (Fig. 14(b)). The MAPE is 0.339, and the RMSE is 0.427.

The deviation between the predicted value of organic carbon content of the BP neural network and the measured value is small. The model demonstrates an optimal predictive capability across a range of organic matter abundance layers. The predictive efficacy observed in organic-medium shale samples is comparable to that observed in all samples (Fig. 15(a)). The prediction effect of the organic-high layer is superior to the average value. Nevertheless, the BP neural network is susceptible to overfitting due to the limited number of samples. The specific performance is that the prediction effect of the test group is significantly lower than that of the training group. To address this issue, it is necessary to increase the number of samples, which is challenging given the limitations of the model itself. To overcome the over-fitting phenomenon of neural networks, it is necessary to increase the number of samples. However, this is not a straightforward process and it is difficult to improve the model itself.

4.3 R-MNR

The R-MNR uses five logging parameters combined with the coupling relationship between them to establish a

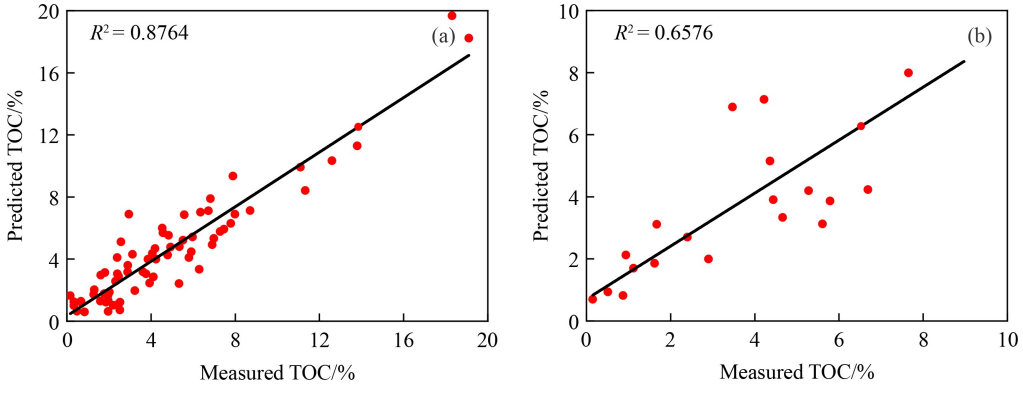

Fig. 14 Prediction performance of BP Model. (a) Training group. (b) Testing group.

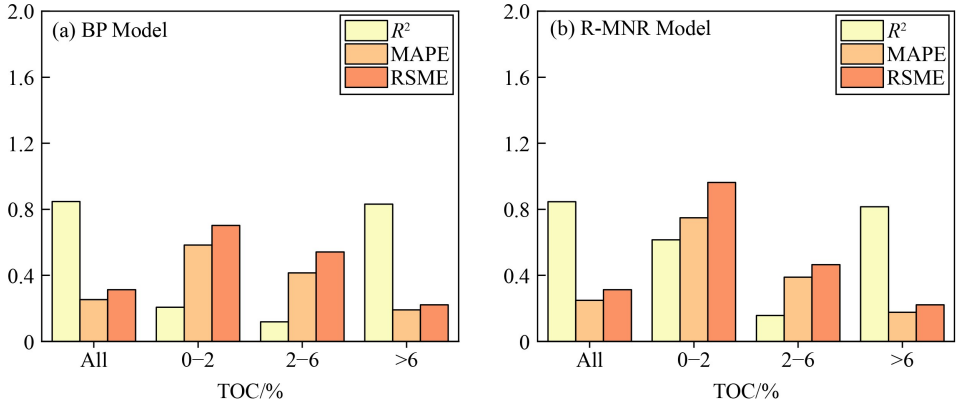

Fig. 15 Performance diagrams of BP Model, R-MNR. (a) Performance diagrams of BP Model. (b) Performance diagrams of R-MNR (Where, 'All' represents all samples, 0−2 represents samples with TOC between 0 and 2%, and so on. The closer R^2 is to 1, the higher the correlation of the model is. The closer the MAPE and RSME are to 0, the smaller the error of model is.).

prediction model for TOC. Compared with the traditional method, the nonlinear regression model analyzes the interaction between parameters by considering the synergistic effect of multi-parameters. Compared with the BP model, the R-MNR can obtain an explicit formula, which is easy to validate and apply. The R^2 of the model in the training group reaches 0.8549 (Fig. 16(a)), the MAPE is 0.2456, and the RMSE is 0.3075. The effect of the model applied in the test group is comparable to that of the training group, and the R^2 between the measured and predicted value is 0.7908 in the test group (Fig. 16(b)). The MAPE is 0.2662 and the RMSE is 0.3369.

The prediction effect of the R-MNR is demonstrably superior to that of the traditional prediction model, which is nearly indistinguishable from the prediction effect of the BP neural network model in the training group (Fig. 15(b)). The performance of the R-MNR in the test group was demonstrably superior to that of the BP neural network. In contrast to the BP model, the R-MNR is capable of analyzing the coupling relationship between logging parameters in a limited number of samples and deriving a discernible prediction formula. This enables the R-MNR to be readily applied to the prediction of organic carbon content, obviating the necessity for the use of large workstations or professional software to complete the application of the model.

4.4 Model comparison

In this paper, four prediction models were evaluated using AHP. The TOC prediction model evaluation is taken as the target layer (A_1). The number of input logging parameters, R^2, MAPE, and RMSE are the criterion layers (A_{11}, A_{12}, A_{13}, and A_{14}), and four prediction models are taken as the solution layer (A_{111}, A_{112},..., A_{144}). The AHP comparison system is shown in Fig. 17, and the specific parameters are shown in Table 6.

Among the four parameters of the criterion layer, the NILP is directly responsible for determining the cost of application and the complexity of model extension. The R^2 is one of the most commonly employed correlation evaluation indices, and it can, to a certain extent, reflect the accuracy of the model. The RMSE is a statistical measure that quantifies the accuracy of a model's prediction of continuous data. It provides a means of assessing the average discrepancy between the predicted and actual values, so as to clarify the prediction accuracy of the model. The MAPE was employed to assess the discrepancy between the actual and fitted values, as well as to ascertain the ratio between the error and the true

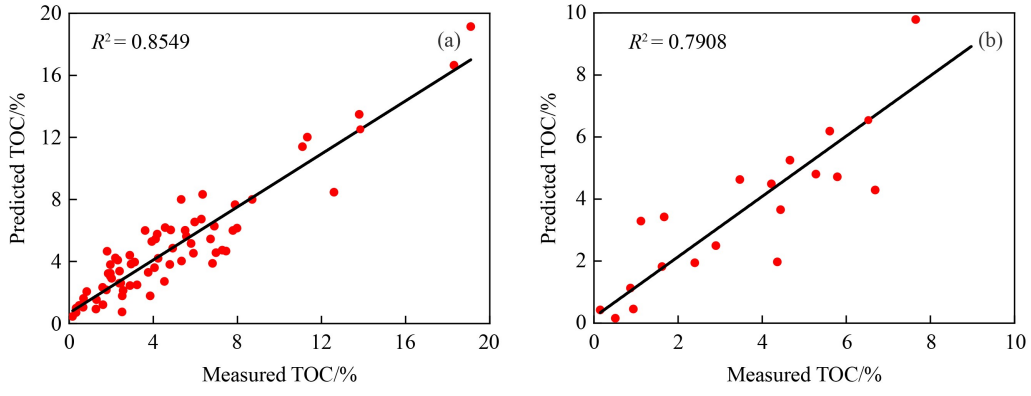

Fig. 16 Prediction performance of R-MNR Model. (a) Training group. (b) Testing group.

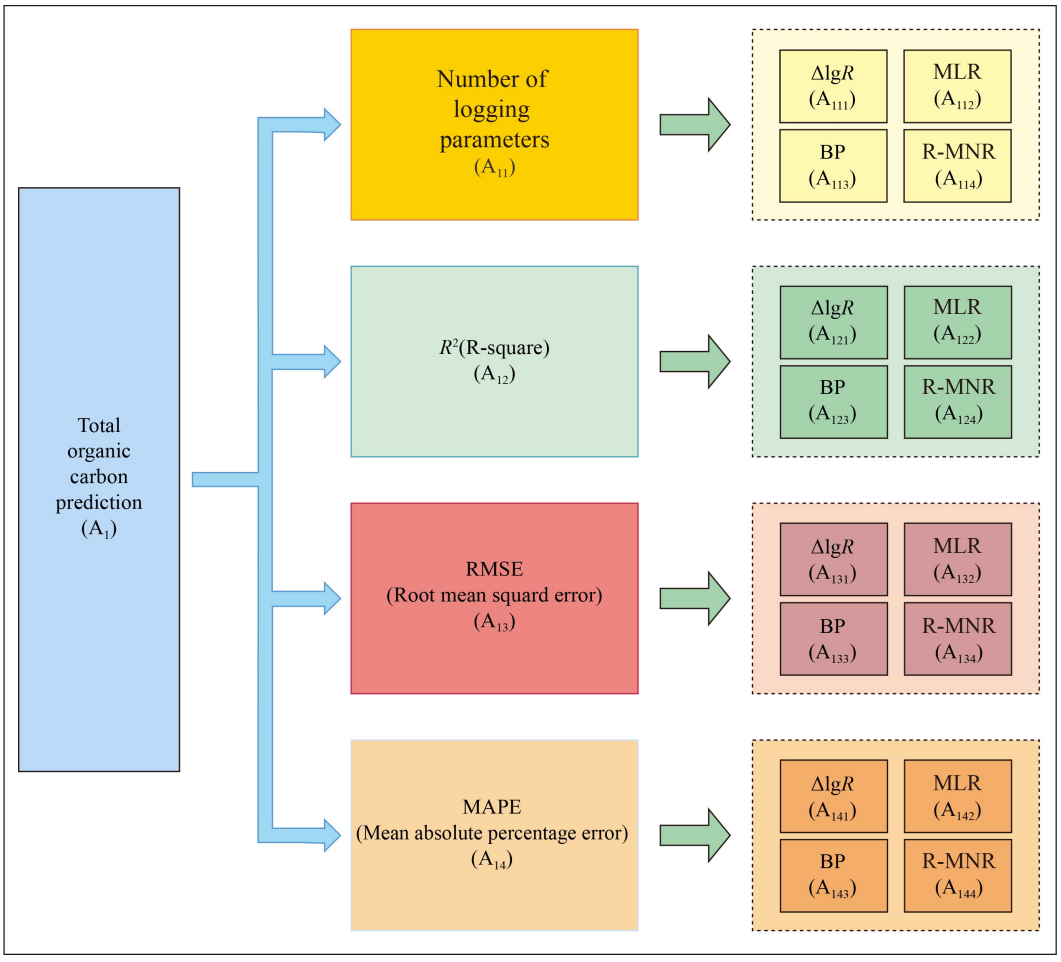

Fig. 17 Structure of prediction model evaluated by AHP.

Table 6 Evaluation parameters of different models

Model	R^2	MAPE	RMSE	Number of logging parameters
ΔlgR	0.513	0.394	0.568	3
MLR	0.619	0.390	0.495	5
BP	0.849	0.253	0.314	8
R-MNR	0.846	0.246	0.313	5

value. The aforementioned four parameters are of equal importance with regard to the evaluation of the prediction model. Consequently, the weights of the four parameters in the criterion layer were set at 0.25 each.

The weights of the solution layer parameters are determined by the consistency matrix method proposed by Saaty. The specific steps are as follows: first, the importance of the parameters is compared by using the 1−9 scale method to establish the judgment matrix, then, the maximum eigenvalue and the corresponding eigenvector of this matrix are obtained, and the judgment matrix consistency ratio *CR* is computed to determine whether the matrix has satisfactory consistency or not (Franěk and Kresta, 2014). Take layer A_{12} as an example:

The relative goodness of the factors within layer A_{12} was compared using the 1−9 scale method, and the judgment matrix U_{A12} was constructed to calculate the weights using Eq. (18). The maximum eigenvalue of this matrix is calculated to be 4.1545, and the corresponding eigenvectors are (0.401, 0.801, 1.577, 1.221). To ensure the credibility and relative accuracy of the calculation results, according to Eq. (19) and Eq. (20), the judgment matrix consistency ratio *CR* is calculated to be 0.0572, which is less than 0.10, indicating that the matrix has a satisfying consistency (Xu and Xu, 2020; Pant et al., 2022).

$$U_{A_{11}} = \begin{pmatrix} A_{11} & A_{111} & A_{112} & A_{113} & A_{114} \\ A_{111} & 1 & 0.33 & 0.33 & 0.67 \\ A_{112} & 3 & 1 & 1 & 0.67 \\ A_{113} & 3 & 1 & 1 & 0.67 \\ A_{114} & 1.5 & 1.5 & 1.5 & 1 \end{pmatrix}, \quad (18)$$

$$CR = CI/RI, \quad (19)$$

$$CI = (\lambda_{\max} - n)/(n-1), \quad (20)$$

where $\lambda_{\max}$ is the maximum eigenvalue of the judgment matrix, n is the order of the matrix, *CI* is the consistency

index, and *RI* is the consistency index obtained by Saaty, as shown in Table 7.

The weights of the solution layer parameters are determined by the consistency matrix method proposed by Saaty (Xu and Xu, 2020). The specific steps are as follows: first, the importance of the parameters is compared by using the 1−9 scale method to establish the judgment matrix, then, the maximum eigenvalue and the corresponding eigenvector of this matrix are obtained, and the judgment matrix consistency ratio *CR* is computed to determine whether the matrix has satisfactory consistency or not. Take layer A_{12} as an example:

The importance coefficients of the indicators in the solution layer are weighted and synthesized with the corresponding importance coefficients in the guideline layer to obtain the weights of the solution layer related to the target layer. The formula for calculating the synthesized weights is as follows:

$$P_{ij} = P_i \omega_{ij}, \tag{21}$$

where P_{ij} is the weight of the *j*-th element in the solution layer on the *i*-th parameter in the target layer, P_i is the weight of the *i*-th parameter in the criterion layer on the target layer, and $\boldsymbol{\omega}_{ij}$ is the vector of weights of the *j*-th element in the solution layer on the *i*-th element in the criterion layer.

According to the above steps, the weights of the elements within layer A_{12} for the target layer are obtained as shown in Table 8.

Repeat the above steps to calculate the weights of each factor in the criterion layer separately, and the results are shown in Table 9. Among the four prediction models established in this paper, the total weight of the R-MNR is the highest, followed by the BP model, and the weights of the Δlg*R* and MLR models are lower. The R-MNR is more suitable to be applied and popularized in the prediction of the TOC content of the marine-continental transitional shale of the Benxi Formation, in the Ordos Basin.

4.5 Discussion

During the process of model building and prediction effect comparison, the traditional models (Δlg*R* model and MLR model) are relatively straightforward to construct, however, it is necessary to enhance the precision of the prediction. As a widely used model in the field of TOC prediction, BP model has a significantly higher prediction effect than traditional models. In this study, the BP model has the same problems as other small sample predictions. Improving the prediction accuracy of the training group in the network model can easily lead to over-fitting. As a result, the prediction accuracy on the training group will be significantly reduced. The inability of the BP neural network model to derive a definitive prediction formula renders it challenging to ascertain the extent to which each input parameter influences the prediction outcomes.

The R-MNR provides a clear causal relationship and parameter weights. It can explain the specific influence of each input parameter and the coupling relationship between parameters on the prediction results. This allows researchers to deeply understand the physical or geological factors in the system, rather than just obtain the predicted results. At the same time, due to the simple structure of the multivariate nonlinear regression model, the application effect in small sample prediction is better, and the over-fitting phenomenon is not easy to occur.

Combined with the results of AHP analysis, the R-MNR model is suitable for the prediction of TOC content in marine-continental transitional shale with small sample data sets.

Table 7 Random consistency index

Order *n*	1	2	3	4	5	6	7	8	9
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

Table 8 Plan parameter weight of criterion A_{12}.

Parameter	A_{121}	A_{122}	A_{123}	A_{124}
Weight	0.1	0.2	0.394	0.306

Table 9 Total weight of prediction model

Model	ΔlgR	MLR	BP	R-MNR
Total weight	0.104	0.199	0.336	0.361

5 Conclusions

Four prediction models of TOC were established based on the logging response characteristics of shale by combining different regression analysis methods. The four models were employed to predict the TOC content of marine-continental transitional shale in the Ordos Basin. The prediction effect of the models were evaluated using AHP method. The comprehensive evaluation results demonstrate that the R-MNR model exhibits superior predictive performance, with a comprehensive evaluation index of 0.361. The R^2, RMSE, and MAPE parameters of the model are better than those of other models. The BP model had a comprehensive evaluation index of 0.306. The traditional prediction models performed relatively poor, with comprehensive evaluation indeces of 0.10 and 0.20.

R-MNR combines the advantages of traditional models and neural network model. It can not only achieve the prediction accuracy close to the neural network model, but also obtain an explicit prediction formula. Other researchers can verify the prediction effect of the model

according to the formula, which is helpful to the promotion and application of the model. The R-MNR model can also perform better in the training of small sample data sets, which can avoid the occurrence of over-fitting of neural network models.

The prediction formula of R-MNR model and the process of parameter significance analysis can assist decision makers and researchers to carry out scenario analysis and hypothesis testing based on the model. Obtaining logging parameters that have a significant impact on the prediction of TOC can provide theoretical support for development work and policy formulation. Especially for the marine-continental transitional shale with strong heterogeneity, R-MNR can intuitively reflect the response characteristics of shale organic carbon and logging parameters. Furthermore, the process of regression analysis allows for a deeper understanding of the physical or geological processes at play within the system, rather than merely the prediction results.

Competing Interests The authors declare that they have no competing interests.

Acknowledgments This research was funded by the National Natural Science Foundation of China (Grant No. 42372194), the Natural Science Foundation of Shanxi Province, China (No. 20210302123165), the Chinese Postdoctoral Science Foundation (No. 2024T170634) and the Open Fund Project of Provincial Center of Technology Innovation for Coal Measure Gas Co-production (ZZGSSASMCYJ2024-0306). We would like to express our gratitude to the PetroChina Research Institute of Petroleum Exploration and Development (RIPED), and the Research Institute of Exploration and Development of PetroChina Changqing Oilfield Company for providing shale samples and well-logging data for this study.

CRediT author statement Zhang Xinyu built the prediction model and wrote the first draft of the article. Meng Yanjun provided the computer equipment support and participate in the revision of the paper. Zhong Jinzhi provided the technical support for the construction of the model. Zhang Qin and Qiu Zhen collected experimental samples and experimental data. Zhao Weibo and Yin Liangliang provided the logging data. Ma Haojie participated in the revision of the article.

Data statement The data that support the finding of this study are within this paper. Logging information is confidential. Other data are available from the corresponding author upon reasonable request.

References

Asante-Okyere S, Ziggah Y Y, Marfo S A (2021). Improved total organic carbon convolutional neural network model based on mineralogy and geophysical well log data. Unconventional Resources, 1: 1–8

Bao C, Chen Y, Li D, Wang S (2014). Provenances of the Mesozoic sediments in the Ordos Basin and implications for collision between the North China Craton (NCC) and the South China Craton (SCC). J Asian Earth Sci, 96: 296–307

Bessereau G, Carpentier B, Huc A Y (1991). Wireline logging and source rocks - Estimation of organic carbon content by the Carbolbg@ method. Log Anal, 32: 279–297

Bugaje A B, Dioha M O, Abraham‐Dukuma M C, Wakil M A (2022). Rethinking the position of natural gas in a low-carbon energy transition. Energy Res Soc Sci, 90: 102604

Cao T, Deng M, Xiao J, Liu H, Pan A, Cao Q (2023). Reservoir characteristics of marine–continental transitional shale and gas-bearing mechanism: understanding based on comparison with marine shale reservoir. J Nat Gas Geosci, 8(3): 169–185

Chan S A, Hassan A, Usman M, Humphrey J D, Alzayer Y, Duque F (2022). Total organic carbon (TOC) quantification using artificial neural networks: Improved prediction by leveraging XRF data. J Petrol Sci Eng, 208: 109302

Faiz M, Altmann C, Baruch E, Côté A, Gong S, Schinteie R, Ranasinghe P (2022). Organic matter composition and thermal maturity evaluation of Mesoproterozoic source rocks in the Beetaloo Sub-Basin, Australia. Org Geochem, 174: 104513

Franěk J, Kresta A (2014). Judgment scales and consistency measure in AHP. Procedia Econ Finance, 12: 164–173

He Y, He Z, Tang Y, Xu Y, Long J, Sepehrnoori K (2023). Shale gas production evaluation framework based on data-driven models. Petrol Sci, 20(3): 1659–1675

Hu H, Lu S, Liu C, Wang W, Wang M, Li J, Shang J (2021). Models for calculating organic carbon content from logging information: comparison and analysis. Acta Sediment Sin, 29(6): 1199–1205 (in Chinese)

Ju W, Shen J, Qin Y, Meng S, Wu C, Shen Y, Yang Z, Li G, Li C (2017). In-situ stress state in the Linxing region, eastern Ordos Basin, China: implications for unconventional gas exploration and production. Mar Pet Geol, 86: 66–78

Kleber M, Bourg I C, Coward E K, Hansel C M, Myneni S C, Nunan N (2021). Dynamic interactions at the mineral–organic matter interface. Nat Rev Earth Environ, 2(6): 402–421

Li Y, Yang S, Lu Y, Ma Z, Song F, Zheng K, Li X, Wang Y, Tittel F K, Zheng C (2022). Multi-parameter methane measurement using near-infrared tunable diode laser absorption spectroscopy based on back propagation neural network. Infrared Phys Technol, 125: 104275

Liu C, Zhao W C, Sun L, Zhang Y, Chen X, Li J (2021). An improved ΔlogR model for evaluating organic matter abundance. J Petrol Sci Eng, 206: 109016

Liu D, Yao Y, Chang Y (2022). Measurement of adsorption phase densities with respect to different pressure: potential application for determination of free and adsorbed methane in coalbed methane reservoir. Chem Eng J, 446: 137103

Liu D, Zhao Z, Cai Y, Sun F (2024). Characterizing coal gas reservoirs: a multiparametric evaluation based on geological and geophysical methods. Gondwana Res, 133: 91–107

Liu Z, Tang S, Zhang P, Zhang Q, Zhang K, Yang X, Mei X (2023). Organic matter characteristics and total organic carbon content prediction of coal measure shale: a case study of the south Ningwu block. Sci Techn Eng, 23(27): 11593–11604 (in Chinese)

Mahmoud A A, Elkatatny S, Ali A, Abouelresh M, Abdulraheem A (2019). Evaluation of the total organic carbon (TOC) using different artificial intelligence techniques. Sustainability (Basel), 11(20): 5643

Mandal P, Rezaee R, Emelyanova I V (2021). Ensemble learning for predicting TOC from well-logs of the unconventional goldwyer shale. Energies, 15(1): 216

Meng Y, Tang D, Xu H, Li C, Li L, Meng S (2014). Geological controls and coalbed methane production potential evaluation: a case study in Liulin area, eastern Ordos Basin, China. J Nat Gas Sci Eng, 21: 95–111

Nie X, Wan Y K, Gao D, Zhang C, Zhang Z (2021). Evaluation of the in-place adsorbed gas content of organic-rich shales using wireline logging data: a new method and its application. Front Earth Sci, 15(2): 301–309

Pant S, Kumar A, Ram M, Klochkov Y, Sharma H K (2022). Consistency indices in analytic hierarchy process: a review. Mathematics, 10(8): 1206

Passey Q R, Bohacs K M, Esch W L, Klimentidis R E, Sinha S (2010). From Oil-Prone Source Rock to Gas-Producing Shale Reservoir - Geologic and Petrophysical Characterization of Unconventional Shale-Gas Reservoirs. In： International Oil and Gas Conference and Exhibition in China, Beijing, China, June 2010

Safari A, Das N, Langhelle O, Roy J, Assadi M (2019). Natural gas: a transition fuel for sustainable energy system transformation. Energy Sci Eng, 7(4): 1075–1094

Saporetti C M, Fonseca D L, Oliveira L C, Pereira E, Goliatt L (2023). Machine learning with model selection to predict TOC from mineralogical constituents: case study in the Sichuan Basin. Int J Environ Sci Technol, 20(2): 1585–1596

Schmoker J W (1981). Determination of organic-matter content of Appalachian Devonian shales from gamma-ray logs. AAPG Bull, 65(7): 1285–1298

Shi X, Yang Z, Dong Y, Zhou B (2019). Tectonic uplift of the northern Qinling Mountains (Central China) during the late Cenozoic: evidence from DEM-based geomorphological analysis. J Petrol Sci Eng, 184: 104005

Vaidya O S, Kumar S (2006). Analytic hierarchy process: an overview of applications. Eur J Oper Res, 169(1): 1–29

Wang E, Guo T, Li M, Xiong L, Dong X, Zhang N, Wang T (2022a). Depositional environment variation and organic matter accumulation mechanism of marine–continental transitional shale in the Upper Permian Longtan Formation, Sichuan Basin, SW China. ACS Earth Space Chem, 6(9): 2199–2214

Wang H, Wu W, Chen T, Dong X, Wang G (2019). An improved neural network for TOC, S1 and S2 estimation based on conventional well logs. J Petrol Sci Eng, 176: 664–678

Wang J, Xu Y, Sun P, Liu Z, Zhang J, Meng Q, Zhang P, Tang B (2022b). Prediction of organic carbon content in oil shale based on logging: a case study in the Songliao Basin, Northeast China. Geomechan Geophys Geo-Energy Geo-Resour, 8(2): 44

Wang X, Liu G, Wang X, Ma J, Wang Z, Wang F, Song Z, Fan C (2024a). Geophysical prediction of organic matter abundance in source rocks based on geochemical analysis: a case study of southwestern Bozhong Sag, Bohai Sea, China. Petrol Sci, 21(1): 31–53

Wang Y, Wang Z, Zhang Z, Yao S, Zhang H, Zheng G, Luo F, Feng L, Liu K, Jiang L (2024b). Recent techniques on analyses and characterizations of shale gas and oil reservoir. Energy Reviews, 3(2): 100067

Wang Y, Yang J (2024). Origin of organic matter pore heterogeneity in oil mature Triassic Chang-7 mudstones, Ordos Basin, China. Int J Coal Geol, 283: 104458

Xu J, Li M, Zhong J, Hou Y, Xia S, Yu P (2022). Process parameter modeling and multi-response optimization of wire electrical discharge machining NiTi shape memory alloy. Mater Today Commun, 33: 104252

Xu K, Xu J (2020). A direct consistency test and improvement method for the analytic hierarchy process. Fuzzy Optim Decis Making, 19(3): 359–388

Yan T, He S, Zheng S, Bai Y, Chen W, Meng Y, Jin S, Yao H, Jia X (2023a). Critical tectonic events and their geological controls on deep buried coalbed methane accumulation in Daning-Jixian Block, eastern Ordos Basin. Front Earth Sci, 17(1): 197–217

Yan T, Yang C, Zheng S, Bai Y, Chen W, Liu Y, Tian W, Sun S, Jin S, Wang J, Liu Z, Yao H (2023b). Geochemical characteristics of produced fluids from CBM wells and their indicative significance for gas accumulation in Daning-Jixian block, Ordos Basin. Front Earth Sci, 17(3): 661–678

Yu H, Rezaee R, Wang Z, Han T, Zhang Y, Arif M, Johnson L M (2017). A new method for TOC estimation in tight shale gas reservoirs. Int J Coal Geol, 179: 269–277

Zhang H, Wu W, Wu H (2022). TOC prediction using a gradient boosting decision tree method: a case study of shale reservoirs in Qinshui Basin. J Petrol Sci Eng, 221: 111271

Zhao L, Qin X, Zhang J, Liu X, Han D, Geng J, Xiong Y (2018). An effective reservoir parameter for seismic characterization of organic shale reservoir. Surv Geophys, 39(3): 509–541

Zhao Z, Xu W, Zhao Z, Yi S, Yang W, Zhang Y, Sun Y, Zhao W, Shi Y, Zhang C, Gao J (2024). Geological characteristics and exploration breakthroughs of coal rock gas in Carboniferous Benxi Formation, Ordos Basin, NW China. Pet Explor Dev, 51(2): 262–278

Zhu L, Zhang C, Zhang C, Wei Y, Zhou X, Cheng Y, Huang Y, Zhang L (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

Zou C, Zhao Q, Chen J, Li J, Yang Z, Sun Q, Lu J, Zhang G (2018). Natural gas in China: development trend and strategic forecast. Natural Gas Industry B, 5(4): 380–390

Zou C, Zhao Q, Zhang G, Xiong B (2016). Energy revolution: from a fossil energy era to a new energy era. Nat Gas Indust B, 3(1): 1–11