Oil Shale, 2025, Vol. 42, No. 4, pp. 414–433
doi: https://doi.org/10.3176/oil.2025.4.04
ISSN 0208-189X

Logging-based prediction of organic geochemical parameters in oil shale during thermal evolution using the XGBoost algorithm

Lianxin Tao[a,b], Xin Liu[c,d], Zhisheng Luan[e], Ling Jiang[f], Hongliang Dang[a,b,g,h], Pingchang Sun[a,b]*

[a] College of Earth Sciences, Jilin University, Changchun, Jilin 130061, China
[b] Key-Lab for Oil Shale and Paragenetic Minerals of Jilin Province, Changchun, Jilin 130061, China
[c] State Key Laboratory of Continental Shale Oil, Daqing, Heilongjiang 163712, China
[d] Exploration and Development Research Institute, Daqing Oilfield Company Limited, Daqing, Heilongjiang 163712, China
[e] Liaohe Oilfield Company, PetroChina, Panjin, Liaoning 124010, China
[f] Exploration and Development Research Institute, Northeast Oil & Gas Branch, SINOPEC, Changchun, Jilin 130062, China
[g] Technology Innovation Center for Exploration and Exploitation of Strategic Mineral Resources in Plateau Desert Region, Ministry of Natural Resources, Xining, Qinghai 810000, China
[h] Qinghai Geological Survey, Xining, Qinghai 810000, China

Received 17 January 2025, accepted 15 October 2025, available online 20 October 2025

Abstract. *Oil shale in large basins undergoes multiple evolutionary stages, limiting the applicability of a single logging-based prediction model. This study focuses on the oil shale of the Qingshankou Formation in the Songliao Basin, using gamma ray (GR), deep resistivity (LLD), acoustic travel time (DT), neutron porosity (CNL), density (DEN), and depth data as input features. The XGBoost algorithm is employed to develop predictive models for total organic carbon (TOC) content, free hydrocarbon (S_1), pyrolyzable hydrocarbon (S_2), and maximum pyrolysis peak temperature (T_{max}). TOC predictions are further stratified for low-maturity, mature, and high-maturity oil shale intervals. The results show that S_2 achieves the highest prediction accuracy ($R^2 = 0.91$), due to its strong correlation with hydrogen index (HI) driven by thermal evolution. TOC prediction accuracy ($R^2 = 0.75$) is influenced by combined changes in porosity and organic matter evolution. T_{max} prediction ($R^2 = 0.74$) depends mainly on depth and CNL. S_1 correlates weakly with all well logs, yielding the lowest accuracy ($R^2 = 0.29$). Shale maturity plays a critical role in determining the reliability of TOC prediction models. Low-*

* Corresponding author, sunpingchang711@126.com

maturity oil shale exhibits the best TOC accuracy (R^2 = 0.83), as wellpreserved organic matter and high porosity correlate closely with LLD, DT, CNL, and DEN. In mature oil shale, retained hydrocarbon and reduced porosity weaken logging signals, lowering accuracy to R^2 = 0.63. In high-maturity intervals, hydrocarbon expulsion and porosity rebound improve accuracy (R^2 = 0.69). Our approach provides a cost-effective, continuous method for evaluating lacustrine oil shale resources. It is particularly applicable to the evaluation of uncored wells.

Keywords: *oil shale, prediction models, logging responses, machine learning, Songliao Basin.*

1. Introduction

Unconventional oil and gas research increasingly leverages big data and artificial intelligence. Integrating high-resolution well log data with machine learning and optimized logging technologies significantly improves evaluation accuracy [1, 2]. This approach not only aligns with current developmental trends but also represents the future direction of the field. Logging data, with its continuity, accuracy, and high resolution, contain rich information that can effectively reveal the geochemical characteristics of oil and gas [3, 4]. Consequently, it supports the secondary development of uncored wells and helps reduce exploration and production costs.

Conventional logging techniques, including gamma ray (GR), resistivity (RT), acoustic travel time (DT), neutron porosity (CNL), and density (DEN), are widely used in oil and gas exploration and development [5–7]. Oil shale differs from surrounding formations in terms of mineral composition, structural characteristics, reservoir properties, and organic matter content, resulting in distinct logging responses [8, 9]. Typical logging responses for oil shale include high GR, high RT, high DT, high CNL, and low DEN values [10–15]. Consequently, various methods have been developed to identify oil shale, such as calculating total organic carbon (TOC) content [10–12], Rock-Eval parameters (S_1, S_2, T_{max}) [13–15], oil saturation [16], kerogen type [17], thermal maturity [5, 18], and reservoir properties [3, 6, 7, 18, 19].

Currently, common methods for predicting organic geochemical parameters include the ΔLogR method [11] and its various improvements [20–22], multiple regression methods [23–25], and machine learning approaches [25–30]. The ΔLogR method calculates TOC content by determining the amplitude difference between the DT and RT curves, effectively eliminating the influence of porosity on organic carbon content [11]. However, this method involves fewer variables, and the selection of the baseline is susceptible to human manipulation, requiring careful well section selection and segmentation [11, 20–22]. Multiple regression methods establish relationships between several logging curves and organic geochemical parameters, which can

significantly improve predictive correlation. However, these formulas need to be manually derived and are not universally applicable across different regions. Machine learning approaches, on the other hand, improve algorithm performance through data-driven training, essentially replacing manual derivation in multiple regression with intelligent algorithms [23–25].

There are various types of machine learning algorithms, each with its strengths, weaknesses, and suitable conditions [12, 13, 25–30]. For example, support vector machine (SVM) is suitable for small datasets and nonlinear problems, excelling in binary classification, but it is inefficient when dealing with large datasets [12, 13, 27]. Random forest (RF) performs well on highdimensional classification tasks but is less effective for regression tasks and on small datasets [27]. Artificial neural networks (ANN) can handle complex nonlinear relationships but require large datasets [15, 28, 29]. Extreme gradient boosting (XGBoost) is a powerful and efficient algorithm, widely used for both classification and regression tasks due to its high accuracy and flexibility [30]. Nevertheless, its application in the geological field is still relatively limited.

It is important to note that previous studies on predicting organic geochemical parameters have predominantly been conducted under similar maturity conditions. In basins with significant depth variations, the evolution of organic matter must be considered, as it can limit the applicability of logging-based prediction models. Moreover, research on logging-based predictions of S_1, S_2, and T_{max} is limited. This study focuses on the oil shale of the Qingshankou Formation (Fm) in the Songliao Basin, applying the XGBoost algorithm to predict the TOC content of oil shale at different maturity stages. A TOC estimation system was established, and efforts were made to predict S_1, S_2, and T_{max}, along with an analysis of the factors influencing these predictions. This research provides valuable insights for the secondary development of legacy wells and has significant implications for oil and gas basins transitioning to unconventional oil and gas exploitation.

2. Regional setting

The Songliao Basin, located in northeastern China, is a major terrestrial oil and gas basin, and a significant area for unconventional resource exploration. Spanning approximately 820 km north–south and 350 km east–west, the basin covers an area of 2.6×10^5 km² [31, 32]. It is divided into six tectonic units: the western slope, northern plunge, central depression, and three uplift zones (Fig. 1a) [31–40]. The basin has a rift–sag composite structure, shaped by both extensional and compressional forces. Its tectonic evolution includes stages of mantle uplift, continental rifting, thermal subsidence, and compressional inversion [33].

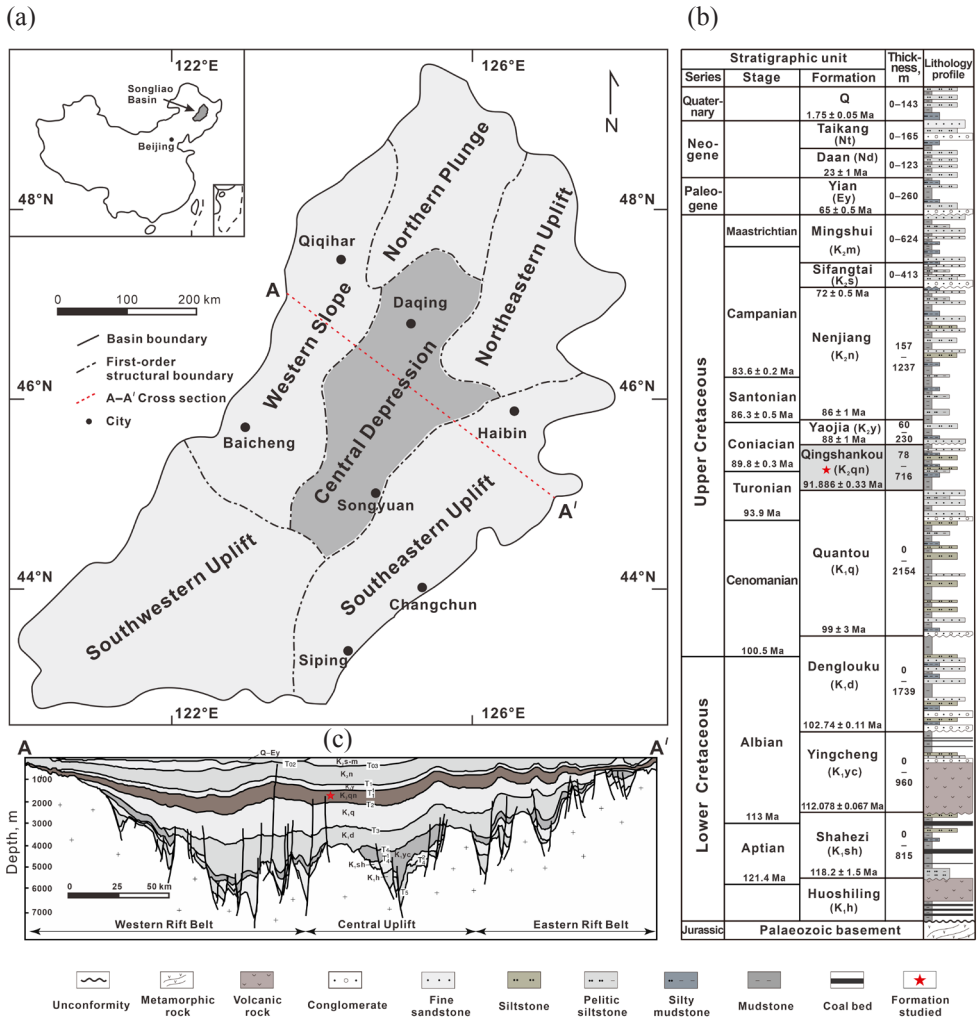

Fig. 1. Regional geological setting of the Songliao Basin [31, 33].

The basin's basement consists of Precambrian and Paleozoic strata, which are overlain by thick Mesozoic deposits [31–33]. The basin exhibits a thin margin and a thick central section, indicative of long-term lacustrine sedimentation [31]. The Jurassic, Cretaceous, and Paleogene deposits can exceed 10 km in thickness, with the Cretaceous deposits reaching up to 7 km. The Cretaceous stratigraphy is subdivided into lower (Huoshiling, Shahezi, Yingcheng, Denglouku, Quantou) and upper (Qingshankou, Yaojia, Nenjiang, Sifangtai, Mingshui) formations (Fig. 1b) [31–33].

The Qingshankou Fm, a primary source rock, ranges from 260 to 500 m in thickness (Fig. 1c) [33]. Deposition occurred in semi-deep to deep lake

settings in the central basin, characterized by dark shale, siltstone, and thin carbonate layers, with shallow lake and deltaic deposits at the margins [31–40]. The Qingshankou Fm exhibits high organic richness, overpressure, and significant hydrocarbon potential, making it a key focus for commercial shale oil and gas development [31].

3. Materials and methods

3.1. Materials

We collected organic geochemical data and logging data from the Qingshankou Fm in the central depression of the Songliao Basin. The dataset includes 1,240 TOC values and 520 Rock-Eval values (S_1, S_2, T_{max}), along with corresponding logging data, such as GR, deep resistivity (LLD), DT, CNL, and DEN values (Fig. 2). These data were obtained from wells located within the same region but at varying depths.

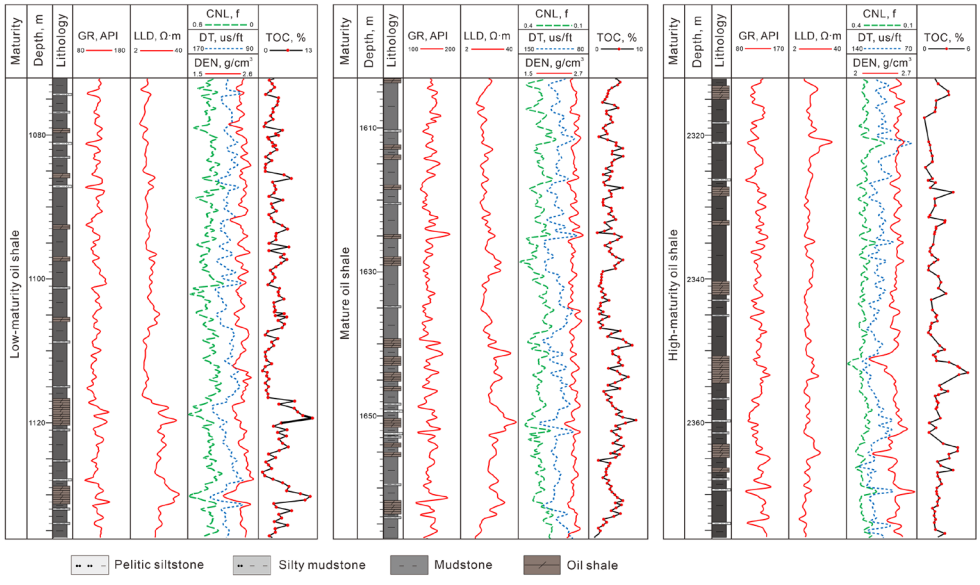

Fig. 2. Logging curves of oil shale at different maturity stages [34–40].

3.2. Methods

3.2.1. Correlation analysis

The organic carbon content in oil shale demonstrates a linear relationship with various logging curves. Specifically, as TOC increases, logging responses show systematic increases or decreases depending on curve type. The Pearson

correlation coefficient is employed to quantify the linear correlation between variables. For two variables (X and Y), the Pearson correlation coefficient ($\rho(X,Y) = cov\ (X,Y)/(\sigma X\ \sigma Y) = E[(X - \mu X)(Y - \mu Y)]/(\sigma X\ \sigma Y)$) is defined as the covariance divided by the product of their standard deviations; it ranges from –1 to +1. A coefficient less than 0 indicates a negative correlation, while a value greater than 0 suggests a positive correlation. The closer the absolute value is to 1, the stronger the correlation. A coefficient of 0 indicates no correlation. We use the Pearson correlation coefficient to assess linear relationships between organic geochemical parameters and logging curves.

3.2.2. Model environment setup

In this study, Python is utilized as the development environment for model construction, leveraging its powerful third-party libraries to support all necessary model-building functionalities. Pandas and NumPy are core libraries for data analysis, providing efficient tools for data manipulation, cleaning, transformation, and computation, thereby enabling rapid data processing. Pandas is employed to read and preprocess the data, facilitating the calculation of correlations between various organic geochemical parameters and well log curves. Matplotlib, a widely used plotting library, is used to generate visualizations such as learning curves and scatter plots of model predictions, offering an intuitive means to assess model performance. Scikit-learn (Sklearn), the most widely adopted machine learning library, provides a comprehensive suite of algorithms, along with modules for feature extraction, data preprocessing, and model evaluation. In this study, we employ the XGBoost implementation from Scikit-learn, leveraging its efficient gradient boosting decision tree algorithm for predictive modeling.

3.2.3. Model tuning

XGBoost is an ensemble learning algorithm based on gradient-boosted decision trees. It enhances model performance by iteratively optimizing the second-order derivative of the loss function and incorporating regularization terms to reduce model complexity and mitigate overfitting. Moreover, its parallelized design significantly improves training efficiency [30]. The hyperparameters of the XGBoost algorithm are categorized into general parameters, booster parameters, and task parameters [30]. General parameters determine the weak learners used in the ensemble, with decision trees selected as the booster in this study. The number of trees is controlled by the number of iterations (*n_estimators*). An excessive number of trees may lead to overfitting, while too few trees may result in insufficient learning capacity. Therefore, selecting an appropriate number of iterations is crucial for model training.

During the hyperparameter tuning process, the number of iterations is optimized first. Booster parameters correspond to the decision trees and include hyperparameters that significantly affect model performance. In this study, we tune the following hyperparameters: *max_depth*, *sub_sample*,

learning_rate, and *gamma*, due to their significant impact on performance. *max_depth* controls the maximum depth of the tree. Larger depths may lead to overfitting. Given the six features, *max_depth* is set within the range of [1, 6]. *sub_sample* controls the fraction of training instances used to build each tree. While it helps prevent overfitting, a value that is too small may cause underfitting. Its range is [0, 1]. *learning_rate* controls the step size of each iteration. Smaller learning rates improve the model's generalization ability, with a range of [0, 1]. *gamma* serves as a penalty term for tree complexity, controlling the minimum information gain required for a tree split. Larger *gamma* values reduce the risk of overfitting, with its range set to [0, +∞].

Task parameters specify the learning task and evaluation metrics. We use the default *squarederror* objective for regression.

4. Results and discussion

4.1. Data preparation

Based on R_o, the Qingshankou Fm oil shale is classified into low-maturity (R_o < 0.7%), mature (R_o = 0.7–1.2%), and high-maturity (R_o > 1.2%). TOC predictions were conducted for each maturity stage, and the TOC, S_1, S_2, and T_{max} data are summarized in Table 1.

The TOC values of the Qingshankou Fm oil shale are generally high but decrease with increasing maturity (Fig. 3). The Rock-Eval parameters indicate that the S_1 value initially increases and then decreases as maturity increases, while the S_2 value decreases consistently, reflecting the processes of hydrocarbon generation and expulsion during burial (Fig. 3). The T_{max} values clearly distinguish oil shales at different maturity stages. Due to the influence of low S_2 values on T_{max}, samples with S_2 values less than 0.5 mg/g were excluded from the T_{max} analysis [41].

In this study, the data were neither normalized nor subjected to outlier elimination using the three-sigma method [25, 28]. XGBoost, a tree-based algorithm capable of handling both classification and regression tasks, is relatively insensitive to parameter scaling [30], eliminating the need for normalization. The three-sigma method, commonly used to remove outliers

Table 1. Statistics of organic geochemical parameters [34–40]

	TOC, %	Low-maturity TOC, %	Mature TOC, %	High-maturity TOC, %	S_1, mg/g	S_2, mg/g	T_{max}, °C
Min	0.09	0.26	0.59	0.09	0.19	6.8	429.5
Max	15.25	15.25	7.67	5.23	5.12	137.39	554
Average	2.45	3.65	2.49	1.74	1.42	48.34	452

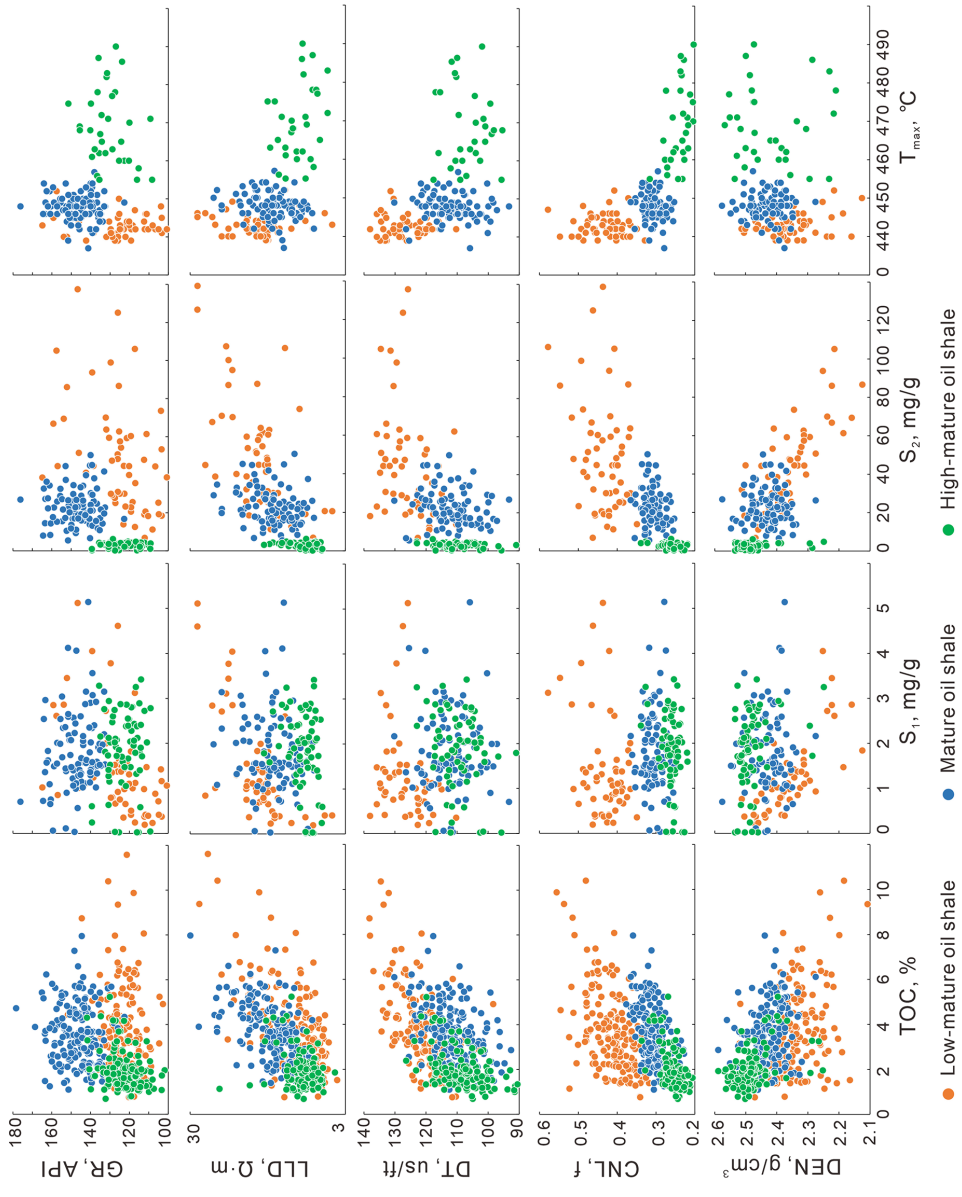

Fig. 3. Crossplot of well log data and organic geochemical parameters [34–40].

from normally distributed data, was not applied because geochemical parameters such as TOC in oil shale do not follow a normal distribution. Instead, the dataset was expanded, additional predictive variables were introduced, and hyperparameters were optimally tuned, allowing the machine learning model to fully exploit its strengths.

4.2. Correlation analysis

Given the substantial depth variation in the study area, well log values are impacted by burial depth. Therefore, depth was included as a feature variable in the correlation analysis (Table 2). Specifically, TOC is negatively correlated with depth, positively correlated with LLD, CNL, and DT, and weakly correlated with GR and DEN. The S_1 value shows weak correlations with all well logs, with only a slight correlation with GR and DT. The S_2 value is negatively correlated with depth, positively correlated with LLD, DT, and CNL, and weakly correlated with GR and DEN. The $S_1 + S_2$ correlation pattern closely matches that of S_2. T_{max} is strongly correlated with depth, negatively correlated with CNL, and shows negligible correlations with other logs. These findings indicate that TOC, S_2, $S_1 + S_2$, and T_{max} can be reliably predicted using well logs, while predicting S_1 proves to be more challenging.

The analysis further reveals that the correlation between well log values and TOC varies significantly with maturity. For low-maturity oil shale, TOC is strongly correlated with LLD, DT, CNL, and DEN, suggesting that TOC in low-maturity oil shale can be reliably predicted using well logs. In mature oil shale, the correlation between TOC and most well logs is generally weak, with only LLD and depth showing stronger correlations, which may result in

Table 2. Pearson correlation coefficients between organic geochemical parameters and logging curve data (bold font indicates Pearson correlation coefficients >0.3)

	Depth, m	GR, API	LLD, Ω·m	DT, us/ft	CNL, f	DEN, g/cm^3
TOC	**–0.501**	0.126	**0.396**	**0.436**	**0.520**	–0.283
Low-maturity TOC	–0.251	0.162	**0.642**	**0.451**	**0.530**	**–0.486**
Mature TOC	**–0.311**	0.249	**0.354**	0.251	0.171	–0.022
High-maturity TOC	**–0.427**	–0.185	–0.059	**0.434**	**0.531**	–0.249
S_1	0.018	–0.108	–0.084	0.114	0.034	0.038
S_2	**–0.821**	0.262	**0.414**	**0.320**	**0.645**	–0.157
$S_1 + S_2$	**–0.804**	0.246	**0.398**	**0.327**	**0.637**	–0.150
T_{max}	**0.735**	–0.073	–0.104	–0.179	**–0.399**	0.042

lower prediction accuracy. For high-maturity oil shale, TOC exhibits stronger correlations with DT, CNL, and depth, but shows a weak correlation with GR and DEN, and no correlation with LLD. The correlation patterns between TOC and well logs vary significantly across different maturities, which may lead to substantial differences in the prediction results.

4.3. Model training and prediction results

XGBoost models are highly sensitive to hyperparameters, necessitating careful tuning to achieve optimal performance. Parameter tuning is a critical aspect of model training, as the choice of hyperparameters significantly influences model performance. Since XGBoost employs an ensemble of decision trees, the number of trees (i.e., the number of boosting iterations) directly affects prediction outcomes. Consequently, we first optimize the number of boosting iterations (0 to 300) via cross-validation to maximize the average R^2 (Fig. 4).

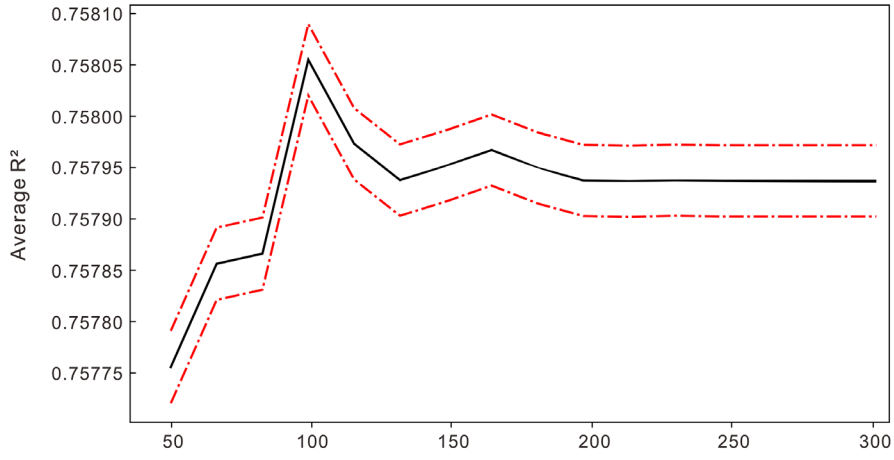

Fig. 4. Optimization of the number of boosting iterations (0–300) using cross-validation.

Grid search is employed to tune the remaining hyperparameters. This method exhaustively explores all possible combinations of predefined parameter ranges and step sizes, selecting the best set of parameters based on cross-validation results. The parameter settings, including value ranges and step sizes, are detailed in Table 3.

The optimized XGBoost model is then applied to prediction. The dataset is split into training and test sets at a 7:3 ratio. The input feature variables include six well log parameters: GR, LLD, DT, CNL, DEN, and depth. For each target parameter (TOC, S_1, S_2, T_{max}), 70% of the data are used for training and 30% for testing. We train the model on the training set and evaluate on the test set using R^2 (coefficient of determination) as the metric.

Table 3. Ranges and steps of hyperparameters

Hyperparameter	Value range	Step size
max_depth	[1,6]	1
sub_sample	[0,1]	0.01
learning_rate	[0,1]	0.01
gamma	[0,1]	0.01

The results demonstrate strong predictive performance for TOC, S_2, and T_{max}, with R^2 values of 0.75, 0.91, and 0.74, respectively. However, prediction of S_1 is less accurate, yielding an R^2 of 0.29 (Figs 5–7).

In predicting TOC in oil shale at various maturities, the best results are obtained for low-maturity oil shale, with an R^2 of 0.83. The worst prediction accuracy is observed for mature oil shale, with an R^2 of 0.63, while the R^2 value for high-maturity oil shale is 0.69 (Figs 5 and 7).

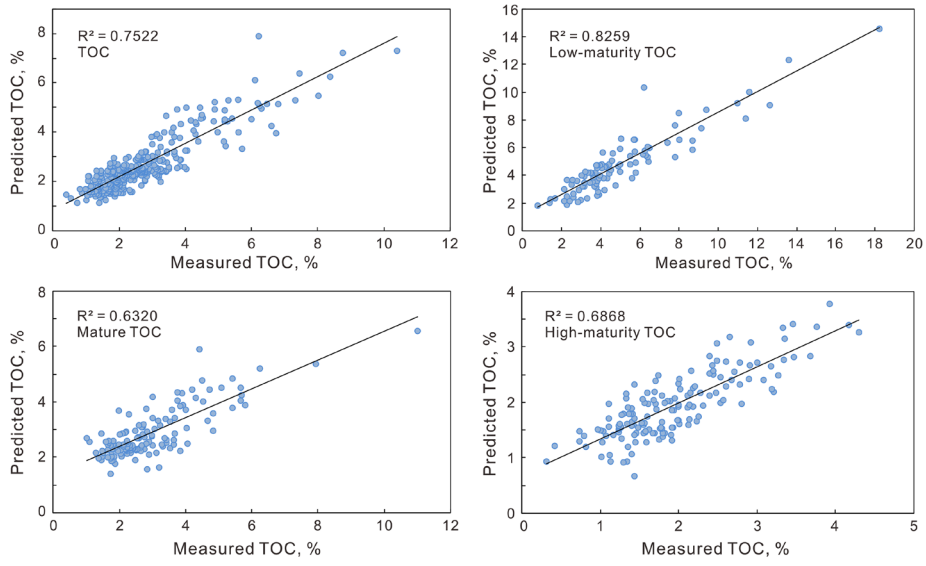

Fig. 5. TOC prediction results for oil shale at different maturity stages.

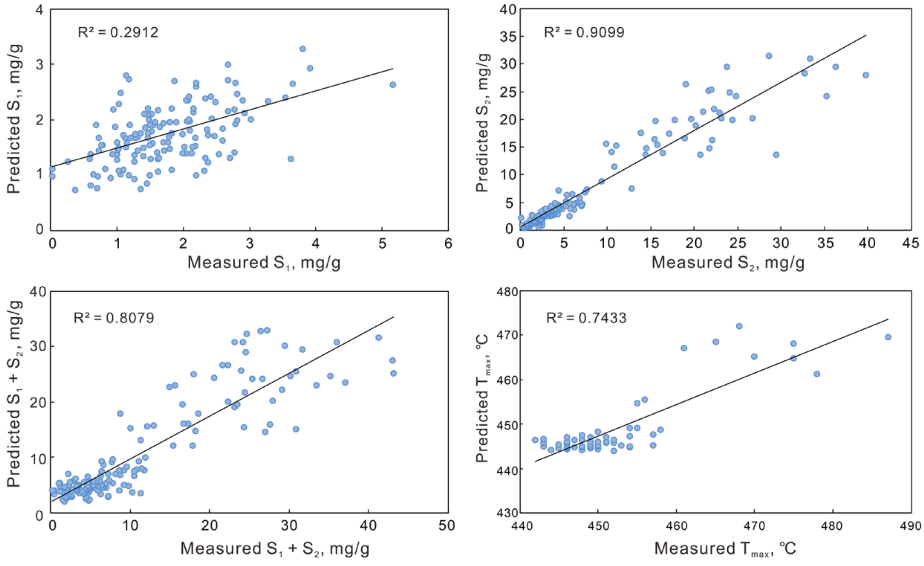

Fig. 6. Prediction results for S_1, S_2, $S_1 + S_2$, and T_{max}.

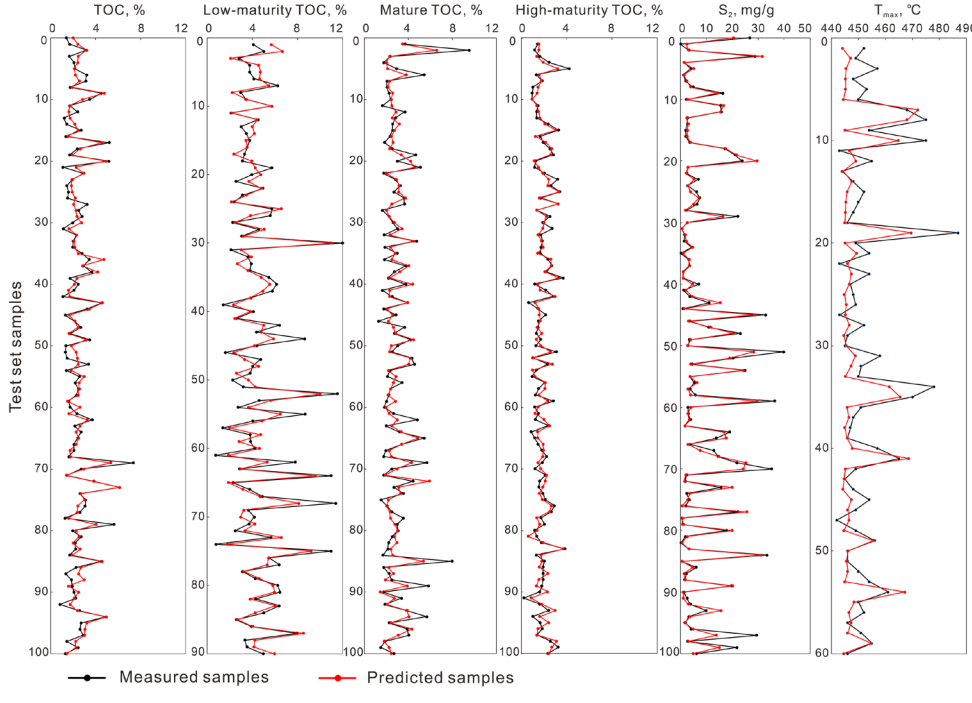

Fig. 7. Comparison of measured and predicted results for the test set.

The predictive performance of TOC using thermal maturity-stratified models was lower than expected, particularly for mature oil shales. One contributing factor is that machine learning models typically require large datasets for effective training. In this study, dividing the dataset into three subsets based on maturity stage substantially reduced the number of samples available for each model, resulting in insufficient training for each subset. Another factor is the sensitivity of R^2 values to outliers. When high TOC values from low-maturity samples and low TOC values from high-maturity samples are combined in a single training set, R^2 can be inflated even if the overall prediction accuracy remains low. Moreover, these results are consistent with the findings of the correlation analysis. Therefore, prediction accuracy primarily depends on the strength of the correlation between geochemical parameters and logging curves. During diagenesis, organic matter and porosity in oil shale undergo significant changes, which weakens their correlation with logging responses.

4.4. Factors influencing prediction results

The relationship between organic matter content in oil shale and well log data is significant, with various organic geochemical parameters showing distinct correlations with well log curves (Fig. 3 and Table 2). High TOC values are generally associated with high GR, LLD, DT, and CNL values, and with low DEN values. While S_2 values exhibit a similar pattern, their correlations with depth and CNL are notably stronger, due to the significant influence of thermal evolution on organic matter. Neutron logging reflects the hydrogen index ($HI = S_2/TOC$) of rock layers [6, 9]. As hydrocarbon generation and expulsion occur during burial, the hydrocarbon generation potential of organic matter declines, leading to a reduction in HI [42]. This results in a strong correlation between CNL and S_2 values, yielding the highest prediction accuracy for S_2. T_{max} follows a similar trend but correlates only with depth and CNL. In contrast, S_1 values show weak correlations with well log curves.

Oil shale undergoes various transformations during deep burial, including porosity evolution, organic matter maturation, and hydrocarbon generation and expulsion [43, 44]. These changes result in distinct logging responses and prediction outcomes for oil shales at different maturity stages.

In low-maturity oil shale, only a small amount of hydrocarbons is expelled [45]. TOC and S_2 values are typically high, while S_1 values remain low, reflecting well-preserved organic matter. Consequently, TOC values exhibit a strong correlation with LLD. Furthermore, low-maturity oil shale is less affected by compaction, resulting in higher porosity [43, 45], which enhances the correlation between TOC and DT as well as CNL. Due to minimal impacts from compaction, dissolution, cementation, and hydrocarbon generation, the correlation between TOC and DEN is also significant [46]. As a result, TOC prediction for low-maturity oil shale demonstrates high accuracy.

In mature oil shale, S_1 values increase significantly, while S_2 and TOC values decrease, reflecting extensive hydrocarbon generation and expulsion [45, 47]. This transformation weakens the correlation between TOC and LLD. Organic matter exerts an adsorptive effect on radioactive elements, but the complex interplay between clay minerals, organic matter, and radioactive elements remains unclear. Consequently, the correlation between TOC and GR increases slightly in mature oil shale, but remains weak in both low- and high-maturity oil shale, suggesting that elevated S_1 values influence the GR response.

During hydrocarbon generation, organic acids form dissolution pores, while the expulsion and retention of hydrocarbons, coupled with the strong sealing properties of shale, lead to abnormal pressure and increased porosity [48, 49]. However, these pores are simultaneously reduced by compaction and cementation, resulting in lower porosity in mature oil shale [49, 50]. As a result, correlations between TOC and CNL, DT, and DEN are weak. The complex interplay among hydrocarbon generation, pore system evolution, and elevated S_1 collectively weakens correlations between TOC and well log curves, thus explaining the lower prediction accuracy for mature oil shale.

In high-maturity oil shale, organic matter has largely lost its hydrocarbon generation potential [42, 51], resulting in low S_1 content. Most of the organic matter has been expelled, leading to a negligible correlation between TOC and LLD. The correlation with DEN also remains weak, potentially influenced by residual light oil. At this stage, the shrinkage of organic matter following hydrocarbon generation, combined with increased organic acids, produces numerous organic matter and dissolution pores [46, 48]. This increase in porosity strengthens the correlations between TOC and both CNL and DT. Notably, TOC exhibits a weak negative correlation with GR, suggesting that expelled hydrocarbons are the primary factor influencing GR responses. Overall, TOC prediction for highmaturity oil shale is primarily controlled by porosity, yielding better prediction accuracy compared to mature oil shale.

5. Conclusion

Conventional logging curves (GR, RT, DT, CNL, DEN) combined with depth data enable effective prediction of TOC, S_2, and T_{max} in the Qingshankou Fm. S_2 achieves the highest prediction accuracy ($R^2 = 0.91$), primarily due to its strong correlation with thermal evolution-driven variations in the HI. TOC prediction accuracy ($R^2 = 0.75$) is influenced by the combined effects of porosity and organic matter evolution, while T_{max} prediction ($R^2 = 0.74$) primarily depends on depth and CNL. S_1 prediction remains challenging, with lower accuracy ($R^2 = 0.29$) due to its weak correlations with logging curves.

Shale maturity significantly impacts TOC prediction accuracy. Low-maturity oil shale exhibits the highest accuracy ($R^2 = 0.83$), owing to well-

preserved organic matter and high porosity that correlate strongly with logging responses. In mature oil shale, hydrocarbon retention and reduced porosity obscure logging signals, decreasing model accuracy (R^2 = 0.63). High-maturity shale shows improved accuracy (R^2 = 0.69), following hydrocarbon expulsion and porosity rebound.

This model offers a low-cost, continuous approach for predicting parameters in lacustrine oil shale resource evaluation, which is particularly beneficial for assessing uncored wells. However, its generalizability is currently limited to the Qingshankou Fm in the Songliao Basin, necessitating further validation for application in other basins or formations.

Data availability statement

All data used in this study were obtained from published literature and have been duly cited at the corresponding locations in the text. No original data were generated in this study.

Acknowledgments

We acknowledge the funding from the National Natural Science Foundation of China (grants Nos 42372125 and 41772092). The publication costs of this article were partially covered by the Estonian Academy of Sciences.

References

1. Wylie, A. S., Jr., Huntoon, J. E. Log-curve amplitude slicing: visualization of log data and depositional trends in the Middle Devonian Traverse Group, Michigan basin, United States. *AAPG Bulletin*, 2003, **87**(4), 581–608.
2. Jin, X. C., Shah, S. N., Roegiers, J.-C., Zhang, B. An integrated petrophysics and geomechanics approach for fracability evaluation in shale reservoirs. *SPE Journal*, 2015, **20**(3), 518–526. https://doi.org/10.2118/168589-pa
3. Qi, H., Su, J., Hu, X., Ma, A., Dong, Y., Li, A. Study on well logging technology for the comprehensive evaluation of the "seven properties" of shale oil reservoirs – an example of shale oil in the Lucaogou Formation in the Jimsar Sag, Junggar Basin. *Frontiers in Earth Science*, 2022, **9**, 827380. https://doi.org/10.3389/feart.2021.827380
4. Zhao, J., Ge, X., Fan, Y., Liu, J., Chen, Y., Xing, L. A genetic algorithm-driven support vector machine to discriminate the kerogen type using conventional geophysical logging data. *AAPG Bulletin*, 2023, **107**(11), 1837–1849. https://doi.org/10.1306/08022320102
5. Zhao, H., Givens, N. B., Curtis, B. Thermal maturity of the Barnett Shale deter-

mined from well-log analysis. *AAPG Bulletin*, 2007, **91**(4), 535–549. https://doi.org/10.1306/10270606060

6. Fu, Q., Horvath, S. C., Potter, E. C., Roberts, F. Log-derived thickness and porosity of the Barnett Shale, Fort Worth basin, Texas: implications for assessment of gas shale resources. *AAPG Bulletin*, 2015, **99**(1), 119–141. https://doi.org/10.1306/07171413018
7. Chen, X., Chen, L., Jiang, S., Liu, A., Luo, S., Li, H. et al. Evaluation of shale reservoir quality by geophysical logging for Shuijingtuo Formation of lower Cambrian in Yichang Area, Central Yangtze. *Journal of Earth Science*, 2021, **32**(4), 766–777. http://dx.doi.org/10.1007/s12583-020-1051-1
8. Mathur, N., Raju, S. V., Kulkarni, T. G. Improved identification of pay zones through integration of geochemical and log data: a case study from Upper Assam Basin, India. *AAPG Bulletin*, 2001, **85**(2), 309–323. https://doi.org/10.1306/8626C7CB-173B-11D7-8645000102C1865D
9. Hu, S., Liu, W., Liu, Y., Liu, K. Acoustic logging response law in shales based on petrophysical model. *Petroleum Science*, 2022, **19**(5), 2120–2130. https://doi.org/10.1016/j.petsci.2022.03.015
10. Schmoker, J. W., Hester, T. C. Organic carbon in Bakken Formation, United States portion of Williston Basin. *AAPG Bulletin*, 1983, **67**(12), 2165–2174. https://doi.org/10.1306/AD460931-16F7-11D7-8645000102C1865D
11. Passey, Q. R., Moretti, F. J., Kulla, J. B., Moretti, F. J., Stroud, J. D. A practical model for organic richness from porosity and resistivity logs. *AAPG Bulletin*, 1990, **74**(12), 1777–1794. https://doi.org/10.1306/0C9B25C9-1710-11D7-8645000102C1865D
12. Bolandi, V., Kadkhodaie, A., Farzi, R. 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. *Journal of Petroleum Science and Engineering*, 2017, **151**, 224–234. https://doi.org/10.1016/j.petrol.2017.01.003
13. Alizadeh, B., Maroufi, K., Heidarifard, M. H. Estimating source rock parameters using wireline data: an example from Dezful Embayment, south west of Iran. *Journal of Petroleum Science and Engineering*, 2018, **167**, 857–868. https://doi.org/10.1016/j.petrol.2017.12.021
14. Wang, H., Wu, W., Chen, T., Dong, X., Wang, G. An improved neural network for TOC, S_1 and S_2 estimation based on conventional well logs. *Journal of Petroleum Science and Engineering*, 2019, **176**, 664–678. https://doi.org/10.1016/j.petrol.2019.01.096
15. Barham, A., Ismail, M. S., Hermana, M., Padmanabhan, E., Baashar, Y., Sabir, O. Predicting the maturity and organic richness using artificial neural networks (ANNs): a case study of Montney Formation, NE British Columbia, Canada. *Alexandria Engineering Journal*, 2021, **60**(3), 3253–3264. https://doi.org/10.1016/j.aej.2021.01.036
16. Deaf, A. S., Omran, A. A., El-Arab, E. S. Z., Maky, A. B. F. Integrated organic geochemical/petrographic and well logging analyses to evaluate the hydrocarbon

source rock potential of the Middle Jurassic upper Khatatba Formation in Matruh Basin, northwestern Egypt. *Marine and Petroleum Geology*, 2022, **140**, 105622. https://doi.org/10.1016/j.marpetgeo.2022.105622

17. Zhao, J., Ge, X., Fan, Y., Liu, J., Chen, Y., Xing, L. A genetic algorithm-driven support vector machine to discriminate the kerogen type using conventional geophysical logging data. *AAPG Bulletin*, 2023, **107**(11), 1837–1849. https://doi.org/10.1306/08022320102
18. Kadkhodaie, A., Rezaee, M. R. Estimation of vitrinite reflectance from well log data. *Journal of Petroleum Science and Engineering*, 2017, **148**, 94–102. https://doi.org/10.1016/j.petrol.2016.10.015
19. Ye, Y., Tang, S., Xi, Z. et al. A new method to predict brittleness index for shale gas reservoirs: Insights from well logging data. *Journal of Petroleum Science and Engineering*, 2022, **208**, 109431. https://doi.org/10.1016/j.petrol.2021.109431
20. Kamali, M. R., Mirshady, A. A. Total organic carbon content determined from well logs using *ΔLogR* and neuro fuzzy techniques. *Journal of Petroleum Science and Engineering*, 2004, **45**(3–4), 141–148. https://doi.org/10.1016/j.petrol.2004.08.005
21. Wang, P., Chen, Z., Pang, X., Hu, K., Sun, M., Chen, X. Revised models for determining TOC in shale play: example from Devonian Duvernay Shale, Western Canada Sedimentary Basin. *Marine and Petroleum Geology*, 2016, **70**, 304–319. https://doi.org/10.1016/j.marpetgeo.2015.11.023
22. Rahmani, O., Khoshnoodkia, M., Kadkhodaie, A., Pour, A. B., Tsegab, H. Geochemical analysis for determining total organic carbon content based on Δ*LogR* technique in the South Pars Field. *Minerals*, 2019, **9**(12), 735. https://doi.org/10.3390/min9120735
23. Meyer, B. L., Nederlof, M. H. Identification of source rocks on wireline logs by density/resistivity and sonic transit time/resistivity crossplots. *AAPG Bulletin*, 1984, **68**(2), 121–129. https://doi.org/10.1306/AD4609E0-16F7-11D7-8645000102C1865D
24. Zhao, P., Mao, Z., Huang, Z., Zhang, C. A new method for estimating total organic carbon content from well logs. *AAPG Bulletin*, 2016, **100**(8), 1311–1327. https://doi.org/10.1306/02221615104
25. Wang, J., Xu, Y., Sun, P., Liu, Z., Zhang, J., Meng, Q. et al. 2022. Prediction of organic carbon content in oil shale based on logging: a case study in the Songliao Basin, Northeast China. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, 2022, **8**, 44. http://dx.doi.org/10.1007/s40948-022-00355-9
26. Tang, B., Meng, Q., Liu, Z., Hu, F., Zhang, P., Dang, W. et al. Logging identification of lithology in fine-grained sedimentary rocks based on the FSSA-HKELM model: a case study of the Qingshankou Formation in the Songliao Basin (NE China). *Oil Shale*, 2024, **41**(3), 163–188. https://doi.org/10.3176/oil.2024.3.02
27. Cracknell, M. J., Reading, A. M. The upside of uncertainty: identification of lithology contact zones from airborne geophysics and satellite data using random forest and support vector machines. *Geophysics*, 2013, **78**(3), 113–126.

28. Mahmoud, A. A. A., Elkatatny, S., Mahmoud, M., Abouelresh, M., Abdulraheem, A., Ali, A. Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network. *International Journal of Coal Geology*, 2017, **179**, 72–80. https://doi.org/10.1016/j.coal.2017.05.012
29. Johnson, L. M., Rezaee, R., Kadkhodaie, A., Smith, G., Yu, H. Geochemical property modelling of a potential shale reservoir in the Canning Basin (Western Australia), using artificial neural networks and geostatistical tools. *Computers & Geosciences*, 2018, **120**, 73–81. https://doi.org/10.1016/j.cageo.2018.08.004
30. Chen, T., Guestrin, C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 785–794. https://doi.org/10.1145/2939672.2939785
31. Liu, B., Wang, H., Fu, X., Bai, Y., Bai, L., Jia, M. et al. Lithofacies and depositional setting of a highly prospective lacustrine shale oil succession from the Upper Cretaceous Qingshankou Formation in the Gulong sag, northern Songliao Basin, northeast China. *AAPG Bulletin*, 2019, **103**(2), 405–432. https://doi.org/10.1306/08031817416
32. Liu, B., Liu, L., Fu, J., Lin, T., He, J., Liu, X. et al. The Songliao Super Basin in northeastern China. *AAPG Bulletin*, 2023, **107**(8), 1257–1297. https://doi.org/10.1306/02242321181
33. Feng, Z. Q., Jia, C. Z., Xie, X. N., Zhang, S., Feng, Z. H., Cross, T. A. Tectonostratigraphic units and stratigraphic sequences of the nonmarine Songliao basin. *Basin Research*, 2010, **22**(1), 79–95. https://doi.org/10.1111/j.1365-2117.2009.00445.x
34. Liu, B., Zhao, X., Fu, X., Yuan, B., Bai, L., Zhang, Y. et al. Petrophysical characteristics and log identification of lacustrine shale lithofacies: a case study of the first member of Qingshankou Formation in the Songliao Basin, Northeast China. *Interpretation*, 2020, **8**(3), 45–57. https://doi.org/10.1190/INT-2019-0254.1
35. Zhang, X., Zou, C., Zhao, J., Li, N. Organic-rich source rock characterization and evaluation of the Cretaceous Qingshankou Formation: results from geophysical logs of the second scientific drilling borehole in the Songliao Basin, NE China. *Geosciences Journal*, 2018, **23**, 119–135. http://dx.doi.org/10.1007/s12303-018-0013-4
36. Li, C., Yan, W., Wu, H., Tian, H., Zheng, J., Yu, J. et al. Calculation of oil saturation in clay-rich shale reservoirs: a case study of Qing 1 Member of Cretaceous Qingshankou Formation in Gulong Sag, Songliao Basin, NE China. *Petroleum Exploration and Development*, 2022, **49**(6), 1351–1363. https://doi.org/10.1016/S1876-3804(23)60354-4
37. Liu, B., Sun, J., Zhang, Y., He, J., Fu, X., Yang, L. et al. Reservoir space and enrichment model of shale oil in the first member of Cretaceous Qingshankou Formation in the Changling Sag, southern Songliao Basin, NE China. *Petroleum Exploration and Development*, 2021, **48**(3), 608–624. https://doi.org/10.1016/S1876-3804(21)60049-6
38. Xu, J., Bechtel, A., Sachsenhofer, R. F., Liu, Z., Gratzer, R., Meng, Q. et al.

High resolution geochemical analysis of organic matter accumulation in the Qingshankou Formation, Upper Cretaceous, Songliao Basin (NE China). *International Journal of Coal Geology*, 2015, **141–142**, 23–32. https://doi.org/10.1016/j.coal.2015.03.003

39. Wu, H., Xu, H., Zhou, H., Jiang, P., Shang, F., Wang, L. Astronomical control on organic matter enrichment of lacustrine mudstones in the first member of the Late Cretaceous Qingshankou Formation, the Songliao Basin, NE China. *Journal of Asian Earth Sciences*, 2024, **259**, 105906. https://doi.org/10.1016/j.jseaes.2023.105906
40. Fu, X., Meng, Q., Bai, Y., Su, Y., Jin, M., Huo, Z. et al. Quantitative analysis of paleoenvironment of Qingshankou Formation in northern Songliao Basin, Northeastern China. *Interpretation*, 2022, **10**(3), 75–87. https://doi.org/10.1190/INT-2021-0153.1
41. Sun, Y., Wang, Y., Liao, L., Shi, S., Liu, J. How grain size influences hydrocarbon generation and expulsion of shale based on Rock-Eval pyrolysis and kinetics? *Marine and Petroleum Geology*, 2023, **155**, 106369. https://doi.org/10.1016/j.marpetgeo.2023.106369
42. Li, C., Pang, X., Huo, Z., Wang, E., Zue, N. A revised method for reconstructing the hydrocarbon generation and expulsion history and evaluating the hydrocarbon resource potential: example from the first member of the Qingshankou Formation in the Northern Songliao Basin, Northeast China. *Marine and Petroleum Geology*, 2020, **121**, 104577. https://doi.org/10.1016/j.marpetgeo.2020.104577
43. Milliken, K. L., Zhang, T., Chen, J., Ni, Y. Mineral diagenetic control of expulsion efficiency in organic-rich mudrocks, Bakken Formation (Devonian-Mississippian), Williston Basin, North Dakota, USA. *Marine and Petroleum Geology*, 2021, **127**, 104869. https://doi.org/10.1016/j.marpetgeo.2020.104869
44. Li, X. S., Zhong, H. J., Zhang, K. X., Li, Z., Yu, Y. X., Feng, X. Q. et al. Pore characteristics and pore structure deformation evolution of ductile deformed shales in the Wufeng-Longmaxi Formation, southern China. *Marine and Petroleum Geology*, 2021, **127**, 104992. https://doi.org/10.1016/j.marpetgeo.2021.104992
45. Mastalerz, M., Drobniak, A., Stankiewicz, A. B. Origin, properties, and implications of solid bitumen in source-rock reservoirs: a review. *International Journal of Coal Geology*, 2018, **195**, 14–36. https://doi.org/10.1016/j.coal.2018.05.013
46. Wu, S. T., Zhu, R. K., Cui, J. G., Cui, J. W., Bai, B., Zhang, X. X. et al. Characteristics of lacustrine shale porosity evolution, Triassic Chang 7 Member, Ordos Basin, NW China. *Petroleum Exploration and Development*, 2015, **42**(2), 185−195. https://doi.org/10.1016/S1876-3804(15)30005-7
47. Huang, W. B., Hersi, O. S., Lu, S. F., Deng, S. W. Quantitative modelling of hydrocarbon expulsion and quality grading of tight oil lacustrine source rocks: case study of Qingshankou 1 member, central depression, Southern Songliao Basin, China. *Marine and Petroleum Geology*, 2017, **84**, 34–48. https://doi.org/10.1016/j.marpetgeo.2017.03.021
48. He, W., Wang, M., Wang, X., Meng, Q., Wu, Y., Lin, T. et al. Pore structure characteristics and affecting factors of shale in the First Member of the

Qingshankou Formation in the Gulong Sag, Songliao Basin. *ACS Omega*, 2022, **7**(40), 35755–35773. http://dx.doi.org/10.1021/acsomega.2c03804

49. Han, Y. J., Horsfield, B., Wirth, R., Mahlstedt, N., Bernard, S. Oil retention and porosity evolution in organic-rich shales. *AAPG Bulletin*, 2017, **101**(6), 807−827. https://doi.org/10.1306/09221616069
50. Curtis, M. E., Cardott, B. J., Sondergeld, C. H., Rai, C. S. Development of organic porosity in the Woodford Shale with increasing thermal maturity. *International Journal of Coal Geology*, 2012, **103**, 26–31. https://doi.org/10.1016/j.coal.2012.08.004
51. Zhang, P., Misch, D., Meng, Q., Bechtel, A., Sachsenhofer, R., Liu, Z. et al. Comprehensive thermal maturity assessment in shales: a case study on the upper cretaceous Qingshankou formation (Songliao Basin, NE China). *International Journal of Earth Sciences*, 2021, **110**, 943–962. http://dx.doi.org/10.1007/s00531-021-02000-4