frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 20 January 2023
DOI 10.3389/feart.2022.1051184

OPEN ACCESS

EDITED BY
Shengnan Chen,
University of Calgary, Canada

REVIEWED BY
Shaojie Zhang,
China University of Mining and Technology, China
Qiang Chen,
Chongqing University, China
Wenjun Xu,
Yangtze University, China
Li Zhiqiang,
Chongqing University of Science and Technology, China

*CORRESPONDENCE
Zhenhua Wang,
wzh_swpu@126.com

SPECIALTY SECTION
This article was submitted to Solid Earth Geophysics, a section of the journal Frontiers in Earth Science

RECEIVED 22 September 2022
ACCEPTED 21 November 2022
PUBLISHED 20 January 2023

CITATION
Zhang L, Wang Z, Xu R, Cheng H, Ren L and Lin R (2023), Modeling and analysis of hydraulic fracture complexity index in sandy conglomerate reservoirs based on genetic expression programming—A case study in Xinjiang Oilfield. *Front. Earth Sci.* 10:1051184. doi: 10.3389/feart.2022.1051184

Modeling and analysis of hydraulic fracture complexity index in sandy conglomerate reservoirs based on genetic expression programming—A case study in Xinjiang Oilfield

Long Zhang[1], Zhenhua Wang[2]*, Rui Xu[1], Hao Cheng[1], Lan Ren[2] and Ran Lin[2]

[1]Development Company, Xinjiang Oilfield Company, China National Petroleum Corporation (CNPC), Karamay, Xinjiang, China, [2]State Key Laboratory of Oil & Gas Reservoir Geology and Exploitation, Southwest Petroleum University, Chengdu, Sichuan, China

The stimulation effect of oil wells is seriously affected by the complexity of hydraulic fractures, and the analysis of the factors that control the fracture complexity index has become the key to fracturing design in sandy conglomerate reservoirs. Based on the intrinsic relationship between geological engineering parameters and the fractures complexity index, a Genetic Expression Programming (GEP) method, which has broad advantages in solving multi-factor nonlinear fitting and black-box prediction problems, is proposed to analyze the hydraulic fracture complexity index. Combined with the geoengineering factors that affect the hydraulic fractures propagation, a comprehensive index calculation method is established to analyze the relative importance of these features and 18 reconstructed features were obtained by collecting the geoengineering parameter data of 118 fracturing sections in 8 fracturing wells in Jinlong oilfield. The principal component analysis was performed to eliminate the interaction between the features, and then a GEP-based fractures complexity index calculation model was developed. The partial dependence plot is used to analyze the influence of the main control feature (variable) on the hydraulic fracture complexity index. It showed that GEP model can achieve satisfactory performance (Training set: R = 0.861; Test set: R = 0.817) by statistical parameters. The results showed that the model can calculate the hydraulic fracture complexity index quickly and precisely. The influence of geological engineering control factors can be obtained. It proved that the GEP method can effectively analyze and evaluate the complexity in sandy conglomerate reservoirs.

KEYWORDS

sandy conglomerate reservoir, hydraulic fracturing, fractures complexity index, controlling factors, genetic expression programming (GEP)

1 Introduction

In recent years, hydraulic fracturing technology has been the key reservoir stimulation technology (Zhao et al., 2018; Zhao et al., 2022). Fracture network fracturing has proven to be an effective technology for the development of tight shale reservoirs (Ren et al., 2017; Ren et al., 2018; Ren et al., 2022). Predicting the production performance of multistage fractured horizontal wells is essential for developing unconventional resources such as shale gas and oil. It is essential to accurately characterize the fracture morphology on the reservoir scale (Wang et al., 2020; Wang et al., 2021). The sandy conglomerate reservoirs that are widely distributed in China have poor permeability and strong heterogeneity (Xv et al., 2019). In the Junggar Basin, there have been made important breakthroughs in exploration and development of sandy conglomerate reservoirs, which have proved that the region is rich in oil and gas resources. In order to develop the glutenite reservoirs in Jinlong Oilfield efficiently, it is necessary to conduct in-depth research on glutenite reservoir fracturing and analyze the geological engineering factors that control the formation of complex fractures after fracturing. It is of great value to optimize the fracturing well section of the glutenite reservoir and design the mining scheme for the glutenite reservoir.

The distribution of gravel particles in a glutenite reservoir is often heterogeneous, which the fracturing effect is seriously affected. So the fracture propagation law of conglomerate was studied. It found that the fracture propagation speed is relatively low under low pumping rate. Most of the fractures propagate along the cementation surface around the gravel (Xv et al., 2019). Hydraulic fracture propagation shows different propagation modes when encountering conglomerate, including: directly crossing the conglomerate, turning along the conglomerate, and expanding around the conglomerate (Xv et al., 2020), and the degree of fracture extension is closely related to the number of gravels. Besides, the breakdown pressure of the conglomerate formations was improved due to the existence of conglomerate particles (Shentu et al., 2015). The pressure would lose a lot with the increase in proppant concentration, pump rate, and the volumetric fraction of conglomerates (Wen et al., 2015). As opposed to shale, the unique rock fabric and strong heterogeneities of tight conglomerate formation are favorable factors for forming complex fractures. A small space well pattern can proactively control and make use of interwell interference to increase the complexity of the fracture network, and the "optimum-size and distribution" hydraulic fracturing can be achieved through synergetic optimization (Li et al., 2020). The slickwater and guanidine gum gelout were used to study the mechanisms of formation damage. It found that the gaps in the edges were closed because of slick water. In contrast, the particles exited in the pores due to the flocculent deposits during the damage of guanidine gum gelout (Cheng et al., 2022). In the conglomerate reservoir, it is difficult for proppants to transport in complex fractures networks due to hydraulic fractures tending to bypass the gravels. So the width is changed frequently. Due to the more conglomerate and low hardness, the proppant is seriously embedded in the fracture surface, so quartz sand is more broken up (Zou et al., 2021). At present, in order to develop glutenite reservoirs, fracturing operation parameters are optimized by many scholars through numerical simulation, which can increase the control range of reservoir transformation and form complex hydraulic fracture networks, greatly improving the seepage capacity of reservoirs (Guo, 2019).

The current study of glutenite fracturing confirmed that the fracture has a high complexity index, but the complexity of fracture morphology in engineering research is mainly achieved by microseismic monitoring technology, which can timely guide the fracturing parameters adjustment and optimization design (Fu et al., 2021). However, this technology cannot describe the correlation between fracture complexity and geological-engineering parameters, and it is difficult to clarify the influence law of geological-engineering parameters, which brings great challenges to fracturing optimization. The GEP method (Ferreira, 2001) is a genetic algorithm that can easily implement symbolic regression. Compared with the black box model of neural networks and other methods, the GEP method can clearly give the model equation, so it can be easily applied to engineering (Weatheritt et al., 2017; Akolekar et al., 2019). Through field practice, it is found that the fracture complexity index is affected by many factors, and there is no formula to directly calculate the fracture complexity index after fracturing. Therefore, based on the actual acquisition of microseismic data in the field and combined with GEP technology, a mathematical model for calculating the fracture complexity index is established in the paper, the control factors affecting complexity index of glutenite is studied, and quantitatively analyzes how the control factors affect the complexity index is analyzed according to permutation importance and partial dependence plot. The research results are of great significance for understanding the complexity of glutenite fracturing.

2 Mathematical model

The Jinlong 2 block of Jinlong Oilfield is about 41 km southeast of Karamay City. The northern part of the Jinlong 2 block is adjacent to the proven Permian Upper Wuerhe Formation in the Ke 79 well area. It is located in the triangular fault block formed by the Ke-Wu fault zone in the western uplift of the Junggar Basin and the northern fault of Baijiantan. In the Jinlong 2 block, a 100 m small well spacing three-dimensional development test area was opened up, and 8 horizontal wells were deployed. The three-dimensional well spacing of the two oil layers was 50 m, the horizontal section was 1,300 ~ 1,600 m, and the production was 76,200 tons. In order to study the fracture rupture and propagation in the process of hydraulic fracturing in the small well spacing three-

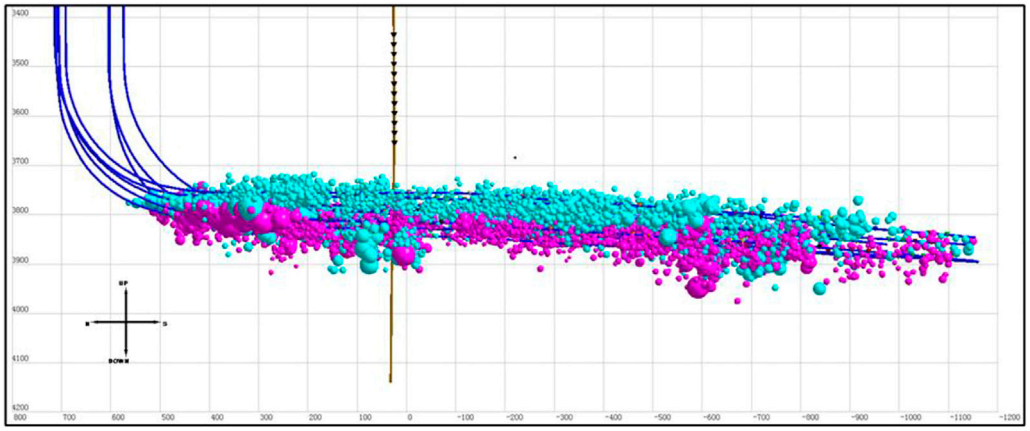

FIGURE 1
Microseismic monitoring map of well JL1.

dimensional development demonstration area, and provide guidance for the analysis of fracturing effect and scheme optimization of this platform, 8 horizontal wells were monitored by microseismic in the area. The spatial distribution characteristics of fracture network orientation, height, and length formed by fracturing were determined, and the fracture propagation morphology was monitored in real time, shown in Figure 1.

Based on the fracture length and width obtained by on-site microseismic monitoring, the fracture complexity index can be preliminarily calculated (Cipolla et al., 2008), and the calculation result is used as the target value in the machine learning. The accuracy of the GEP calculation model is determined by the sample size and the selection of the geological engineering parameters that affect the fracture complexity index. By collecting the data of a total of 118 fracturing sections of 8 fracturing wells on site, a total of 14 possible influencing factors are considered in each section as the data samples for modeling and calculation analysis in this paper, including fracturing engineering parameters such as displacement and operation pressure and reservoir geological parameters.

2.1 Data collection

A total of 118 fracture stages from 8 wells were obtained as a data set to calculate the complex index of the conglomerate. Each stage is used as a data sample. The input includes 14 variables (engineering parameters and geological parameters), and the output variable is the fracture complexity index value, so there are a total of 14 input variables and one output variable. The distribution of the dataset is described in Table 1. It can be seen that there are obvious differences in the data distribution (such as data range) of each variable.

With different unit dimensions and data distribution, efficient machine learning models are sensitive to the distribution of features, so data preprocessing is important for machine learning. The data is moved by the minimum value unit, and will be converged to between [0,1], and it is called data normalization. The normalization formula (Qi et al., 2019) is:

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where $\min(x)$ is the minimum value of data x; $\max(x)$ is the maximum value of data x; x^* is the normalized value of data x.

2.2 Calculation and evaluation of basic data

2.2.1 Judgment of divergence

For the designed features, the coefficient of variation (C_V) is used to judge the degree of dispersion, which is defined as the ratio between the standard deviation of the feature and the mean value, and the expression is as follows:

$$C_V = \frac{\sigma_X}{\mu_X} \tag{2}$$

After calculation, the results and ranking of the 14 feature dispersion coefficients designed are shown in Table 2. Generally speaking, the larger the dispersion coefficient C_V, the more dispersed the data are, and the more effective for prediction. It can be seen from Table 2 that the dispersion coefficient C_V of the five characteristics of mud content, amount of crosslinking agent, total amount of net liquid, average sand ratio, and total proppant are relatively large, and the values of the data are relatively scattered and strongly divergent.

TABLE 1 Descriptive statistical table of characteristic parameters.

Variable	Feature code	Mean	Standard deviation	Partial degrees	Coefficient of variation	Minimum value	Maximum value
Staged length	X_1	66.021	10.003	2.783	0.152	47.000	122.000
Average pumping rate	X_2	7.341	1.375	−0.246	0.187	4.200	9.700
Average operation pressure	X_3	58.924	6.112	0.230	0.104	45.500	79.500
Amount of crosslinking agent	X_4	7.251	2.948	1.500	0.407	3.000	20.000
Total proppant	X_5	61.915	11.987	−1.873	0.194	6.400	90.000
30–50 mesh ceramsite proportion	X_6	0.943	0.047	−6.881	0.050	0.531	0.983
Average sand ratio	X_7	13.911	3.374	−0.255	0.243	3.147	22.222
Total amount of net liquid	X_8	946.086	312.383	2.098	0.330	571.300	2,412.000
Guar gum/jelly proportion	X_9	0.697	0.119	−0.386	0.171	0.386	0.918
Breakdown pressure	X_{10}	73.822	8.001	−0.337	0.108	52.000	89.000
Pump-stopped pressure	X_{11}	29.941	4.246	2.046	0.142	21.800	55.000
Mud content	X_{12}	3.031	2.701	1.780	0.891	0.230	17.430
Porosity	X_{13}	10.123	1.736	−0.005	0.171	6.420	14.810
Oil saturation	X_{14}	49.966	7.652	−0.206	0.153	30.560	72.150
Fracture network complexity index	Y	0.454	0.113	1.029	0.248	0.179	0.958

TABLE 2 Characteristic parameter calculation results.

Feature code	Parameter	Dispersion coefficient	Correlation coefficient	Importance	Comprehensive scores	Rank
X_1	Staged length	0.149	0.047	10.173	4.649	3
X_2	Average pumping rate	0.194	−0.122	3.743	1.784	10
X_3	Average operation pressure	0.107	0.194	4.810	2.232	9
X_4	Amount of crosslinking agent	0.401	0.083	12.780	5.940	2
X_5	Total proppant	0.205	−0.246	3.651	1.760	11
X_6	30–50 mesh ceramsite proportion	0.092	−0.608	2.900	1.407	13
X_7	Average sand ratio	0.253	−0.363	5.845	2.781	8
X_8	Total amount of net liquid	0.326	0.079	6.385	3.028	6
X_9	Guar gum/jelly proportion	0.170	0.185	3.440	1.643	12
X_{10}	Breakdown pressure	0.107	−0.026	2.030	0.964	14
X_{11}	Pump-stopped pressure	0.160	0.227	7.681	3.551	5
X_{12}	Mud content	0.882	−0.069	9.200	4.544	4
X_{13}	Porosity	0.166	−0.010	21.104	9.572	1
X_{14}	Oil saturation	0.156	0.039	6.252	2.887	7
Y	Fracture network complexity index	—	—	—	—	—

2.2.2 Judgment of correlation

Pearson's correlation coefficient (PCC) was used to judge the correlation between 14 features and fracture complexity index. PCC is a classical statistic used to reflect the degree of linear correlation between two variables. The correlation coefficient is represented by $\rho_{X,Y}$, which is the covariance cov(X,Y) of two variables X,Y divided by their standard deviations σ_X and σ_Y. The result is between -1 and 1, and the larger the absolute value of ρ, the stronger the correlation.

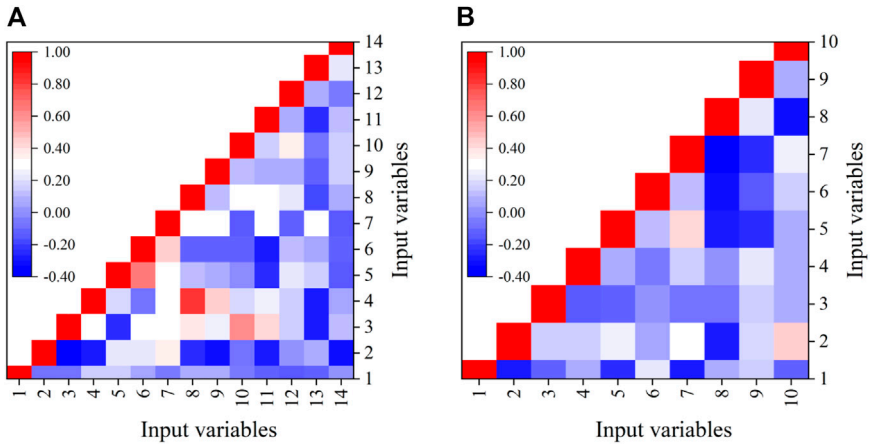

FIGURE 2
Independence among 14 variables **(A)** and 10 reconstructed variables **(B)**.

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (3)$$

where X is feature; Y is predictor variable; σ_x is the standard deviation of X; σ_Y is the standard deviation of Y.

The pearson correlation coefficient between the characteristics and the calculated value of fracture complexity index is shown in Table 2.

2.2.3 Feature importance judgment

Random forest is a classic ensemble model based on bagging to improve the performance of basic decision tree models. CatBoost is a parallel computing model based on boosting to gradually iterate decision tree models to improve the fitting effect (BuKhamseen et al., 2017). The CatBoost model is selected for importance judgment in this paper. After the training of the CatBoost model, the relative importance score of each feature is directly output by the model. The importance of the fracture complex index is shown in Table 2.

2.2.4 Feature synthesis calculation

Regarding the correlation between the feature and the complexity index, a weighted evaluation of the three evaluation indexes (the sum of the weights is 1) is defined. A comprehensive fracture complexity index correlation is given to the feature score by integrating the results of different evaluation methods. The C_V score reflects the divergence of the feature, that is, the amount of information. A weight of 0.1 is assigned to the C_V score (w_3), the remaining weights are equally distributed to PCC and CatBoost. The comprehensive scores are shown in Table 2.

$$Score = w_1 \cdot S_{PCC} + w_2 \cdot S_{CatBoost} + w_2 \cdot S_{CV} \quad (4)$$

2.2.5 Independent judgment

There is not only a correlation between the feature and the fracture complex index, but also a correlation within the feature, that is, between the input variables. The correlation between the 14 input variables is judged by the PCC. As shown in Figure 2A, some of the correlation coefficients are greater than 0.5, indicating that there is a strong correlation between the input variables (Koo et al., 2016). Strong correlation will make GEP biased in variable selection during training and prediction, resulting in poor model correlation. Therefore, it is necessary to combine the importance and independence of features to re-divide the features, and the poor independence features are classified into one category. The selection can not only reflect the important factors, but also ensure the independence between features and improve the accuracy of the model. Based on the comprehensive score of features, the important factors are first selected and then classified according to their independence. The classification results are shown in Table 3 and the reconstructed features are F_1~F_{10}. The importance distribution histogram are shown in Figure 2B.

2.2.6 Principal component analysis

Some strongly correlated variables can be excluded and features that carry important information can be selected by judging the importance and independence of the features above. All the information that affects the complex index is carried in the remaining reconstructed features (F_1~F_{10}), but it is inevitable that there will be poor independence between the reconstructed features, which will cause some

TABLE 3 Reconstruction feature parameter correspondence table.

Feature code	Parameter	Feature code	Strongly correlated features	Reconstructed features
X_{13}	Porosity	—	—	F_1
X_4	Amount of crosslinking agent	X_7	Average sand ratio	F_2
		X_8	Total amount of net liquid	
X_1	Staged length	—	—	F_3
X_{12}	Mud content	—	—	F_4
X_{11}	Pump-stopped pressure	—	—	F_5
X_{14}	Oil saturation	—	—	F_6
X_3	Average operation pressure	X_{10}	Breakdown pressure	F_7
X_2	Average pumping rate	—	—	F_8
X_5	Total proppant	X_6	30–50 mesh ceramsite proportion	F_9
X_9	Guar gum/jelly proportion	—	—	F_{10}

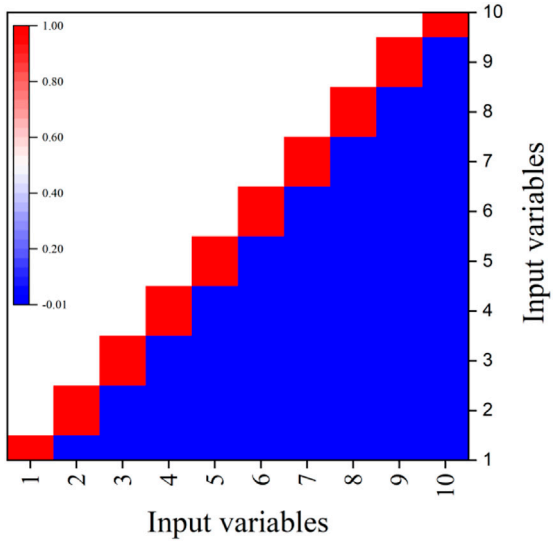

FIGURE 3
Independence among 10 variables after PCA processing.

information to be ignored or covered. And then the accuracy and generalization ability of the GEP calculation model will seriously be affected. Therefore, the principal component analysis method (Sircar et al., 2021) is used to eliminate the weak independence between the reconstructed variables. The input reconstruction features are recombined into a new set of mutually unrelated variables, while retaining the information carried by the original variables. After the dealing with PCA, the correlation between the variables is shown in Figure 3. It can be seen that the variables are independent of each other, and the GEP model can be trained after the PCA processing.

2.3 Gene expression programming

2.3.1 GEP principle

The evolutionary algorithm is a method to search for the maximum or minimum value of a function by simulating Darwin's evolutionary theory of the survival of the fittest in natural organisms. It is suitable for solving complex problems and is used in various fields. The model has strong robustness and is especially suitable for establishing complex functional relationships between variables. Compared with other machine learning algorithms (such as random forests, neural networks, etc.), evolutionary algorithms are interpretable. The advantages of GA and GP are inheritted by GEP, expressing structures of different sizes and shapes using simple, linear, fixed-length individuals. The main genetic operators used in GEP include mutation, inversion, transposition, crossover/recombination, and gene crossover (Ferreira, 2001).

2.3.2 Assessment of the results

To evaluate the performance of the trained GEP model, a statistical index is introduced, which is the mean squared error(*RMSE*) (Emamgolizadeh et al., 2015). It is used to describe the difference between the model calculated value and the field actual value, which is shown in Eq. 5. The smaller the *RMSE*, the better the performance of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - Y_i)^2} \tag{5}$$

where Y_i is fracture complex index value obtained in field, dimensionless; y_i is predicted fracture complexity index value, dimensionless; Y is fracture complex index mean value obtained in field, dimensionless; y is predicted fracture complexity index mean value, dimensionless; n is the number of data samples, dimensionless.

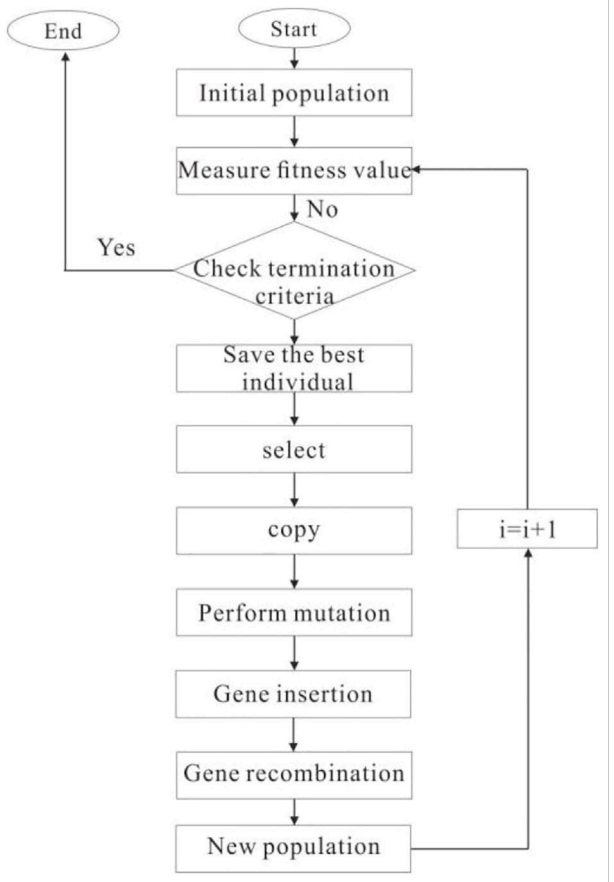

FIGURE 4
Flow chart of GEP evolution.

2.3.3 Fitness function

If the chosen error is the absolute error, the fitness f_i of a single program i is calculated by Eq. 6. If the selected error is a relative error, then the Eq. 7 is used to calculate this value.

$$f_i = \sum_{C_i}^{j=1}\left(M - \left|C_{(i,j)} - T_{(j)}\right|\right) \tag{6}$$

$$f_i = \sum_{C_i}^{j=1}\left(M - \left|\frac{C_{(i,j)} - T_{(j)}}{T_{(i)}} \cdot 100\right|\right) \tag{7}$$

where M is the selection range, $C_{(i,j)}$ is the value returned by individual chromosome i for fitness case j (outside of C_t fitness case), and T_j is the target value for fitness case j. Note that for full adaptation, $C_{(i,j)} = T_j$, $f_i = f_{max} = C_t{\cdot}M$。

2.3.4 GEP algorithm flow

Combined with the principle of the GEP algorithm, the GEP evolution process is shown in Figure 4.

TABLE 4 Equation coefficient table.

a_1	3.651	f_1	5.007
a_2	−10.007	f_2	−8.074
a_3	−6.091	f_3	−0.634
a_4	−5.078	g_1	3.572
b_1	−7.118	g_2	−0.893
b_2	2.729	g_3	1.497
c_1	5.652	h_1	3.042
c_2	8.654	h_2	1.619
d_1	6.924	h_3	0.216
d_2	8.807	k_1	1.996
e_1	4.666	k_2	−3.348
e_2	2.346	k_3	3.054

3 Applications and analysis

At present, there are many studies on the selection of training set and test set size in machine learning. Generally, 70% is selected as the train set size (Qi et al., 2018). Therefore, the data of 83 fracturing stages is selected as training set samples, and the remaining 35 stages are used as test set samples for GEP fitting.

3.1 Fitting of equations

The accuracy and validity of the GEP model depend on many factors. Whether the input variables are independent of each other plays an important role in the selection of important variables during equation fitting. Therefore, the reconstruction variables (F_1, F_2, F_3...F_{10}) are used as the input variables of the GEP model to calculate the fracture complex index of the sandy conglomerate reservoir based on the previous analysis. In this study, basic operators such as +、 -、 *、 /、 exp、 Inv、 Min、 Max are used to implement the GEP model. The fitted equation is shown in Eq. 8, in which y_1~y_{10} are shown in Eq. 9, and the equation coefficients are shown in Table 4.

$$y = y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10} \tag{8}$$

$$\begin{cases}
y_1 = (x_{10}{*}(\max(a_4, gep(\max(a_3, a_2))) - e^{(x_1{*}a_1+\max(x_4,x_1))/2}) \\
\quad +(gep((a_4 - x_3){*}(x_8 - x_{10}){*}x_8{*}x_6){*}x_2))/2 \\
y_2 = x_9^2/(((x_8{*}b_1{*}gep(b_1))^2 + b_1{*}x_1)/2) \\
\quad -(((b_2^2 + b_2 + x_3)/2) - x_2){*}x_{10}^2) \\
y_3 = (((1/((((((1 - (((c_1 - x_6){*}x_2{*}x_6 + (x_7 + c_2)/2)/2)) + (x_{10} + x_1/2))) \\
\quad +(x_6 + x_7)/2))))){*}x_7) + (1 - x_3))/2 \\
y_4 = x_9 - (x_9 - \min \\
\quad (x_{10}, (((1 - ((((((x_4 + x_{10}) + (d_2 + x_4)/2)/2) + (d_1{*}x_1)/2) - \min((x_6 + d_1), (x_3 + x_5)/2)))^2 \\
\quad +x_9)/2)))) \\
y_5 = 2{*}x_9{*}e^{(((x_4-x_6){*}e_2+(x_1+e_1)/2)/2+(1-x_9)/2)^2} * \min(x_4, (1 - gep(1 - (x_{10} - x_1)))) \\
y_6 = x_1{*}gep(((((f_1{*}x_{10} - f_3){*}(f_2 + x_7)/2) \\
\quad +x_{10}/x_8)/2{*}\max(x_3, (x_1{*}x_2)^2)) - x9){*}x_3), \\
y_7 = \min((x_3 - 2{*}x_6^2{*}(1/((((g_1 + x_6) + x_6) \\
\quad +g_2 + x_5) + (x_3^2 + x_7/x_6)))), g_3) \\
y_8 = x_8 \\
y_9 = (((1 - x_3){*}x_{10}{*}(gep(h_1{*}x_3){*}e^{h_2}/((1 - h_3) - 1/x_{10}) - x_{10}){*}e^{x_6}) + x_{10})/2 \\
y_{10} = \min(((((1 - x_3) + 2{*}x_9) + 1/(1 - x_7)) \\
\quad -(2{*}x_{10} + x_2^2)) + (k_2 + x_2)/2{*}x_5{*}k_3/gep((k_1 + x_9)/2), x_9)
\end{cases} \tag{9}$$

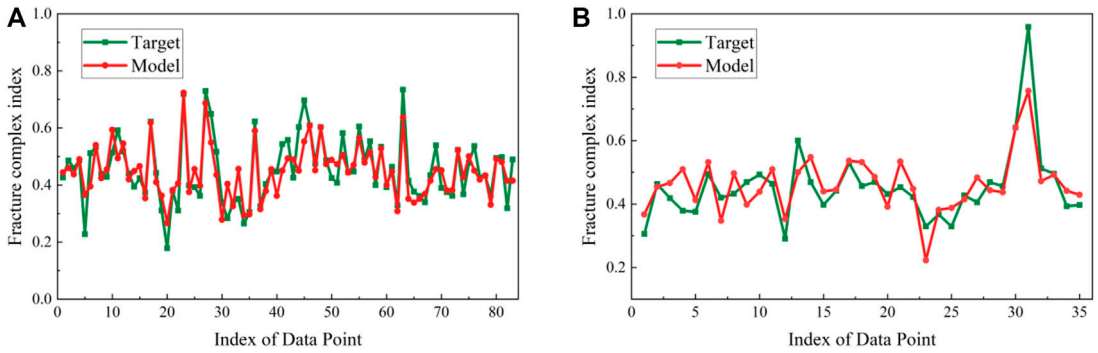

FIGURE 5
Simultaneous plot of outcomes of the GEP model and target data against index of data: **(A)** training set and **(B)** test set.

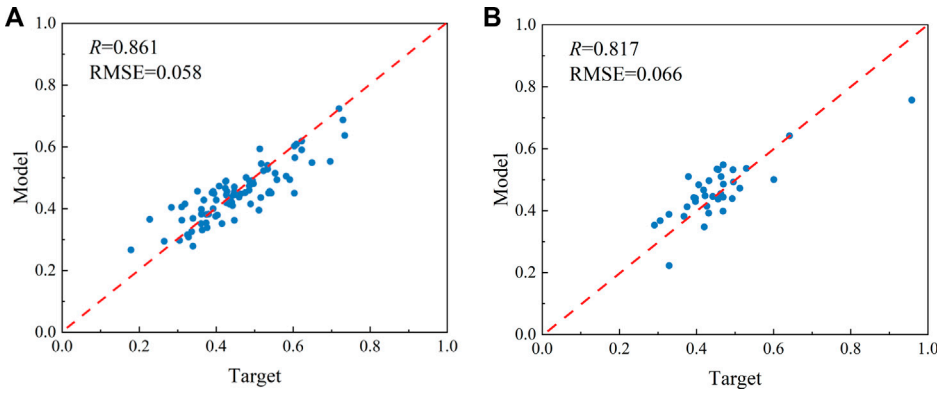

FIGURE 6
Comparison of target value of fracture complexity index and calculated value of GEP model: **(A)** training set and **(B)** test set.

$$gep(x) = \begin{cases} -(-x)^{\frac{1}{3}} & , x < 0 \\ x^{\frac{1}{3}} & , x > 0 \end{cases} \quad (10)$$

3.2 Result analysis

According to the obtained GEP model, the training and test set of fracture complex index are analyzed. In order to get a clearer understanding of the accuracy and reliability of the model, a comparison chart of the fracture complexity index and data point sequence of the actual field results and the model calculation results is also drawn, as shown in Figure 5. It demonstrates the accuracy of the developed model and the GEP model correctly calculate the trend of the field data by comparing the target value and the calculated value in the training and test set.

The fracture complexity index values from actual field data and the calculated results is compared in the Figure 6. It can be seen that the GEP model successfully learns the relationship between the nonlinear fracture complexity index and its influencing variables. The data points are mainly distributed around the diagonal, implying that there is a proper coordination between the calculated data and the target data. Computational performance is evaluated using R and RMSE, and in the case of the training set, the statistical parameters obtained by GEP are: R=0.861, RMSE=0.058. According to statistical suggestions, when R>0.8, it means that the calculation result is better (Roy et al., 2008). Therefore, the computational performance of GEP is satisfactory in the training set. Likewise, the statistical parameters of the test set are: R=0.817 and RMSE=0.066, indicating that the model trained by GEP can be used to calculate the fracture complexity index.

In order to analyze the influencing factors of the fracture complex index, the Permutation Importance (PI) method was used to determine the importance of the factors. It provides a model-independent method for calculating feature importance by selecting a feature and using the test set to calculate a score

TABLE 5 The ranking of main control factors.

Feature name	Feature code	Standard deviation	Ranking
Porosity	F_1	−0.135	6
Amount of crosslinking agent	F_2	−0.369	2
Staged length	F_3	−0.103	8
Mud content	F_4	−0.278	4
Pump stop pressure	F_5	−0.284	3
Oil saturation	F_6	−0.108	7
Average operation pressure	F_7	−0.036	10
Average pumping rate	F_8	−0.227	5
Total proppant	F_9	−0.101	9
Guar gum/jelly proportion	F_{10}	−0.916	1

(standard deviation). Randomly shuffle the values of the feature column of the test set, and calculate the score (standard deviation) of the feature. The effect of the feature on the calculation can be obtained by taking the difference of the score, and then the effect of each feature on the calculation can be obtained. If there is little difference between the old and new results, it means that the feature is of low importance. If the difference between them is significant, then the effect on the model is also significant. Finally, the scores of all features are ranked to get the importance.

The relative importance of the feature variables to the fracture complexity index calculated using the feature importance evaluation method PI is shown in Table 5. Among them, the more important engineering control factors are the ratio of guar gum/jelly, the amount of crosslinking agent and the pump stop pressure, etc. And the more important geological control factor is the mud content.

4 Discussion

Two important issues are discussed in this section. The first is to analyze how the controlling factors, such as geological parameters and engineering parameters, affect the fracture complexity index by using the model. The second is whether the model trained using GEP can be used to calculate the fracture complex index.

The Partial Dependency Plot (PDP) shows the marginal effect of a feature on the results of a previously fitted model calculation, reflecting how this feature affects the calculation. To obtain a partial correlogram (Friedman, 2001), several values of the input variable are first selected, and then, for all cases of the other input variables, each of these values is used to calculate the output. Finally, the average output is calculated and then compared with the corresponding input value. And the partial dependence relationship between some important engineering and geological parameters in the reconstruction variables and the fracture complex index is drawn as shown in the figures.

It can be seen from Figure 7 that the engineering parameters have different degrees of influence on the fracture complexity index. According to the PI calculation results, the proportion of guar gum/jelly is an important controlling factor. With the increase of the proportion, the fracture complex index shows an upward trend, but when the proportion of guar gum/jelly is about 70%, the fracture complexity index begins to decrease. The increasing of the proportion can effectively promote the net pressure in the fracture, which is beneficial to form complex fractures. In addition, with the increase of the crosslinking agent, the fracture complexity index first increases and then decreases, and there is an optimal value.

It can be seen from Figure 8 that the fracture complexity index increases with the increase of the pump-stop pressure, and the higher the pump-stop pressure, the higher the net pressure in the fracture, which can significantly promote the fracture complexity. At the same time, it can be seen that the fracture complexity index increases first and then decreases with the variety of the fracturing stage length, indicating that when the fracturing stage is short, although the length of the stage is fully stimulated, the stimulated area has a low degree of fracture complexity. When the fracturing stage is long, it may not be completely stimulated, the fracture width of microseismic monitoring is smaller than the fracturing stage length, and the fracture complexity is still small. The optimal fracture complexity can be obtained while the fracture length is extended, the lateral stimulated degree is maximized, and the fracture width monitored is approximately equal to the stage length after fracturing by optimizing a reasonable stage length.

Whether the hydraulic fracture can be effectively supported depends to a large extent on the proppant. The selection of proppant specifications is mainly based on the comprehensive consideration of formation sand feeding capacity, proppant conductivity, and proppant breakage rate under the closing stress condition of the target layer. The 30/50 mesh ceramsite and 40/70 mesh quartz sand is used, and it can be seen from the Figure 9 that the amount of

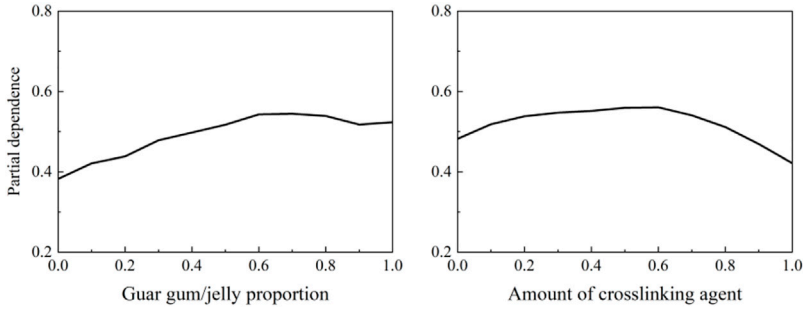

FIGURE 7
Partial dependency plot of engineering parameters(A) of GEP model.

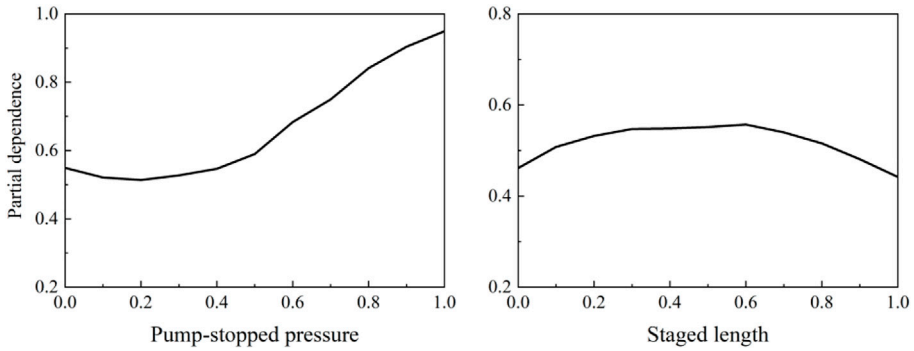

FIGURE 8
Partial dependency plot of engineering parameters(B) of GEP model.

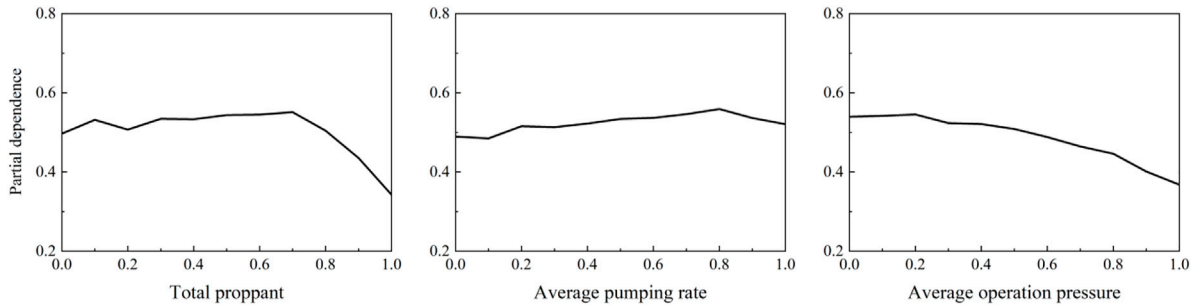

FIGURE 9
Partial dependency plot of engineering parameters(C) of GEP model.

proppant within a certain range can promote the formation of complex fractures, and the proppant with small particle size can play the role of temporary blocking and turning.

The practice of hydraulic fracturing shows that the larger the pumping rate, the greater the net pressure, which is Figure 9 conducive to the propagation of hydraulic fractures. However, the sandy conglomerate reservoir is different from the shale and other natural fracture-developed reservoirs, so the effect of increasing the pumping rate on the fracture complexity index is relatively slow.

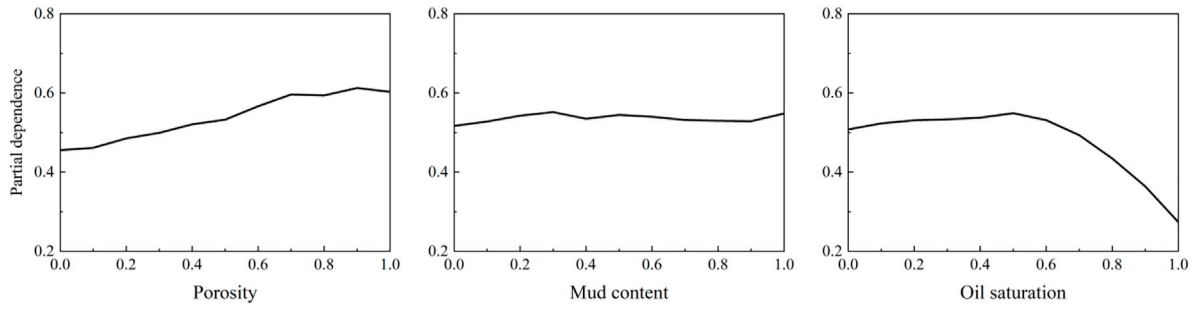

FIGURE 10
Partial dependence plot of geological parameters of GEP model.

The operation pressure is predicted under different pumping rate, considering the need for full stimulation of a single cluster. There are 2 clusters in a single stage, each cluster is 1 m/16 holes, a total of 32 holes, when the pumping rate is 8~10 m^3/min, the predicted operation pressure is 56~59 MPa. The average pumping rate is 4.2~9.7 m^3/min in this paper. Therefore, the pumping rate within this range will not lead to excessive operation pressure. Different from the pump stop pressure, the operation pressure shows the opposite trend. Before fracturing, the operation pressure limit at different sand concentration can be calculated to ensure the safe operation. When the operation pressure varies greatly, it is difficult for the hydraulic fracture to propagate, resulting in a low fracture complexity index.

As shown in Figure 10, for low-porosity sandy conglomerate reservoirs, the fracture complexity index shows an upward trend, which is related to the high elastic characteristics of low-porosity rocks. For high-porosity sandy conglomerate reservoirs, the effect of porosity on the fracture complexity index can be ignored. In the range of mud content in the target block, the effect on the complexity is low. Under low oil saturation, the fracture complex index can be improved due to the small content of pore-liquid phase and strong rock elastic characteristics. When the oil saturation is too high, the complex index is suppressed.

In general, the above discussion shows that the trained calculation model can be used for the calculation and characterization of fracture complex index in this area to achieve better results.

5 Conclusion

(1) Based on the internal relationship between reservoir geoengineering parameters and fracture complexity index, the data of 118 fracturing stages in Jinlong Oilfield were collected. The Genetic Expression Programming (GEP) method, which has extensive advantages in solving multi-factor nonlinear fitting was introduced. And then a hydraulic fracture complex index calculation model was developed. It showed that the model can be extended to calculate the fracture complex index in sandy conglomerate reservoirs.

(2) The controlling factors affecting the fracture complex index were obtained, and the influence on the fracture complex index was analyzed by the partial dependence plot (PDP). It is found that engineering parameters have a greater impact on the fluctuation of the fracture complex index, followed by geological parameters. The influence law of the factors on fracture complex index was obtained in the sandy conglomerate.

(3) The intelligent method is a useful tool for solving complex mechanism problems, especially in the process of hydraulic fracturing. The fracture complex index calculation model established in this paper can be used to analyze the influencing factors of the sandy conglomerate reservoir in Jinlong Oilfield.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data comes from the actual data on site, which is confidential. Requests to access these datasets should be directed to WZ, wzh_swpu@126.com.

Author contributions

ZL: Conceptualization, methodology, design. WZ: Software, methodology, analysis. XR: Data collecting, reviewing and editing. CH: Writing—original draft preparation, revision. RL: Writing, validation, design. LR: Visualization, investigation.

Conflict of interest

Authors ZL, XR, and CH were employed by the company China National Petroleum Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Akolekar, H. D., Weatheritt, J., Hutchins, N., Sandberg, R. D., Laskowski, G., and Michelassi, V. (2019). Development and use of machine-learnt algebraic Reynolds stress models for enhanced prediction of wake mixing in low-pressure turbines. *J. Turbomach.* 141 (4), 041010. doi:10.1115/1.4041753

BuKhamseen, N. Y., and Ertekin, T. (2017). Validating hydraulic fracturing properties in reservoir simulation using artificial neural networks. *SPE.* 9 188093.

Cheng, B., Li, J., Li, J., Su, H., Tang, L., Yu, F., et al. (2022). Pore-scale formation damage caused by fracturing fluids in low-permeability sandy conglomerate reservoirs. *J. Petroleum Sci. Eng.* 208, 109301. doi:10.1016/j.petrol.2021.109301

Cipolla, C. L., Warpinski, N. R., and Mayerhofer, M. J., The relationship between fracture complexity, reservoir properties, and fracture treatment design, SPE annual technical conference and exhibition, 2008. Denver, Colorado, USA, 21–24. doi:10.2118/115769-MS

Emamgolizadeh, S., Bateni, S. M., Shahsavani, D., Ashrafi, T., and Ghorbani, H. (2015). Estimation of soil cation exchange capacity using genetic expression programming (GEP) and multivariate adaptive regression splines (MARS). *J. Hydrology* 529, 1590–1600. doi:10.1016/j.jhydrol.2015.08.025

Ferreira, C. (2001). Gene expression programming: A new adaptive algorithm for solving problems. *Complex Syst.* 13 (2), 87–129.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. statistics*, 43 1189–1232.

Fu, H., Cai, B., Xiu, N., Wang, X., Liang, T., Liu, Y., et al. (2021). The study of hydraulic fracture vertical propagation in unconventional reservoir with beddings and field monitoring. *Nat. Gas. Geosci.* 32 (11), 1610–1621 (Chinese).

Guo, J. (2019). *Optimization design of volume fracturing parameters of horizontal wells in mahu glutenite reservoir*. Beijing, china: China University of Petroleum.

Guoxin, L. I., Jianhua, Q. I. N., Chenggang, X., Fan, X., Zhang, J., and Ding, Y. (2020). Theoretical understandings, key technologies and practices of tight conglomerate oilfield efficient development: A case study of the mahu oilfield, Junggar Basin, NW China. *Petroleum Explor. Dev.* 47 (6), 1275–1290. doi:10.1016/s1876-3804(20)60135-0

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi:10.1016/j.jcm.2016.02.012

Qi, C., Fourie, A., Ma, G., and Tang, X. (2018). A hybrid method for improved stability prediction in construction projects: A case study of stope hangingwall stability. *Appl. Soft Comput.* 71, 649–658. doi:10.1016/j.asoc.2018.07.035

Qi, C., Tang, X., Dong, X., Chen, Q., Fourie, A., and Liu, E. (2019). Towards intelligent mining for backfill: A genetic programming-based method for strength forecasting of cemented paste backfill. *Miner. Eng.* 133, 69–79. doi:10.1016/j.mineng.2019.01.004

Ren, L., Lin, R., Zhao, J., Rasouli, V., and Yang, H. (2018). Stimulated reservoir volume estimation for shale gas fracturing: Mechanism and modeling approach. *J. Petroleum Sci. Eng.* 166, 290–304. doi:10.1016/j.petrol.2018.03.041

Ren, L., Lin, R., Zhao, J., and Wu, L. (2017). An optimal design of cluster spacing intervals for staged fracturing in horizontal shale gas wells based on the optimal SRVs. *Nat. Gas. Ind. B* 4 (5), 364–373. doi:10.1016/j.ngib.2017.10.001

Ren, L., Wang, Z., Zhao, J., Wu, J., Lin, R., Wu, J., et al. (2022). Shale gas load recovery modeling and analysis after hydraulic fracturing based on genetic expression programming: A case study of southern sichuan basin shale. *J. Nat. Gas Sci. Eng.* 107, 104778. doi:10.1016/j.jngse.2022.104778

Roy, P. P., and Roy, K. (2008). On some aspects of variable selection for partial least squares regression models. *QSAR Comb. Sci.* 27 (3), 302–313. doi:10.1002/qsar.200710043

Shentu, J., Lin, B., Dong, J., Yu, H., Shi, S., and Ma, J., (2019).Investigation of hydraulic fracture propagation in conglomerate reservoirs using discrete element method, ARMA-CUPB Geothermal International Conference. Washington, DC, USA. ARMA-CUPB-19-7769.

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

Wang, S., Qin, C., Feng, Q., Javadpour, F., and Rui, Z. (2021). A framework for predicting the production performance of unconventional resources using deep learning. *Appl. Energy* 295, 117016. doi:10.1016/j.apenergy.2021.117016

Wang, S., Wang, X., Bao, L., Feng, Q., and Xu, S. (2020). Characterization of hydraulic fracture propagation in tight formations: A fractal perspective. *J. Petroleum Sci. Eng.* 195, 107871. doi:10.1016/j.petrol.2020.107871

Weatheritt, J., and Sandberg, R. D. (2017). The development of algebraic stress models using a novel evolutionary algorithm. *Int. J. Heat Fluid Flow* 68, 298–318. doi:10.1016/j.ijheatfluidflow.2017.09.017

Wen, Q., Zhang, J., and Li, M. (2015). A new correlation to predict fracture pressure loss and to assist fracture modeling in sandy conglomerate reservoirs. *J. Nat. Gas Sci. Eng.* 26, 1673–1682. doi:10.1016/j.jngse.2015.04.007

Xv, C. Z., Zhang, G. Q., Lyu, Y. J., Xv, Q. S., et al. (2020). *The influence of gravels on hydraulic fracture propagation of conglomerate, Rock Mechanics/Geomechanics Symposium*. Washington, DC, USA. ARMA-2020-1644.

Xv, C., Zhang, G., Yanjun, L., Wang, P., et al. (2019). *Experimental study on hydraulic fracture propagation in conglomerate reservoirs, Rock Mechanics/Geomechanics Symposium*. Washington, DC, USA. ARMA-2019-1844.

Yushi, Z. O. U., Shanzhi, S. H. I., Zhang, S., Yu, T., Tian, G., Ma, X., et al. (2021). Experimental modeling of sanding fracturing and conductivity of propped fractures in conglomerate: A case study of tight conglomerate of mahu sag in Junggar Basin, NW China. *Petroleum Explor. Dev.* 48 (6), 1383–1392. doi:10.1016/s1876-3804(21)60294-x

Zhao, J., Ren, L., Jiang, T., Hu, D., Wu, L., Wu, J., et al. (2022). Ten years of gas shale fracturing in China: Review and prospect. *Nat. Gas. Ind. B*, 9, 158, 175, doi:10.1016/j.ngib.2022.03.002

Zhao, J., Ren, L., Shen, C., and Li, Y. (2018). Latest research progresses in network fracturing theories and technologies for shale gas reservoirs. *Nat. Gas. Ind. B* 5 (5), 533–546. doi:10.1016/j.ngib.2018.03.007