Fuel 290 (2021) 120006

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel

Full Length Article

Predicting the components and types of kerogen in shale by combining machine learning with NMR spectra

Dongliang Kang [a,b], Xiaohe Wang [a,b], Xiaojiao Zheng [a], Ya-Pu Zhao [a,b,*]

[a] State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Machine learning
Kerogen and shale
Molecular structure
High-throughput prediction
NMR spectra datasets

ABSTRACT

This study aims to develop a new method that combines machine learning with nuclear magnetic resonance (NMR) spectra to predict the kerogen components and types. Kerogen is the primary hydrocarbon source of shale oil/gas, and nearly half of the hydrocarbons in shale are adsorbed in kerogen. The adsorption and hydrocarbon generation capacity of kerogen is directly related to its types, molecular components, and structures. Fruitful researches studying kerogen at the molecular level have been conducted. Unfortunately, these methods are complicated, time-consuming, and labor-intensive. Our method has the advantages of high-throughput prediction, high accuracy, and time savings compared with the existing methods. Additionally, this method simplifies the operations from repetitive trial and error. This study proposes a solution to convert non-uniform two-dimensional (2D) graph into a uniform one-dimensional (1D) matrix, which makes 2D graph data available for machine learning models. An automatic labeling platform is constructed that annotated over 22,000 groups of organic matter molecules and their NMR spectra. The results show that the carbon, hydrogen, and oxygen element prediction accuracy reach 96.1%, 94.8%, and 81.7%, respectively. In addition, the accuracy of the three kerogen types is approximately 90% in total. These results reflect the excellent performance of the machine learning method. Therefore, our work provides an automated and intelligent prediction and analysis method, which is a powerful and superior tool in kerogen studies at the molecular level.

1. Introduction

In the past two decades, with the development of hydraulic fracturing and horizontal well technology [1–3], the contribution of shale oil/gas to the energy supply has continued to rise [4]. As the hydrocarbon source organic matter of shale oil/gas, kerogen has become the focus of research. Kerogen is a general term for the organic matter in sedimentary rocks that is insoluble in conventional solvents and is the most abundant source of organic compounds on earth. Oil and natural gas are mainly formed from kerogen [5]. In order to improve the extraction efficiency and yield, it is imperative to have a deep understanding of the generation and evolutionary mechanism of shale oil/gas [6].

The type of kerogen is one of the crucial indicators. The type of kerogen reflects the geological age, formation modes, and even structural characteristics [7,8], which are essential bases for analyzing the geological conditions and reserves of shale oil/gas reservoirs. According to element analysis (EA), the elements contained in kerogen are carbon (C), hydrogen (H), oxygen (O), and a small number of heteroatoms such as nitrogen (N) and sulfur (S). As shown in Table 1, kerogen is generally roughly divided into three types in terms of the H/C and O/C atomic ratios according to the van Krevelen diagram [9]. Type I kerogen is also known as lacustrine organic matter, Its O/C is lower than other types, and the main product is oil. Type II kerogen is an intermediate type, mainly formed by the deposition of algae and other organisms in a hypoxic marine environment, and the product is oil/gas. Type III kerogen contains the most oxygen but a low hydrogen content. It is mainly produced by the deposition of higher plants and typically generates gas [10]. By considering the definition of kerogen classification, bottom-up analysis of kerogen types using molecular information is a simple and direct way. Kelemen et al. analyzed kerogen types and maturity by component characteristics, which were obtained by nuclear magnetic resonance (NMR) and X-ray photoelectron spectroscopy (XPS) [11].

Maturity, which is the degree of development of kerogen, is the other crucial index of organic matter, and it can be used to predict the

[*] Corresponding author at: State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: yzhao@imech.ac.cn (Y.-P. Zhao).

https://doi.org/10.1016/j.fuel.2020.120006
Received 29 September 2020; Received in revised form 19 November 2020; Accepted 13 December 2020
Available online 26 December 2020

Table 1
The ranges of the H/C and O/C in the three types of Kerogen [12].

	H/C	O/C
Type I	>1.5	<0.15
Type II	<1.4	0.03 – 0.18
Type III	<1	0.03 – 0.3

hydrocarbon production rate in the reservoir. There are abundant adsorption sites in the kerogen structure, and nearly half of the hydrocarbons are adsorbed by organic matter [13–15]. The adsorption state of kerogen changes with the environment, further affecting the permeability and reserves of bulk shale oil/gas [16,17]. Maturity can be roughly divided into three stages from low to higher: the diagenesis stage, the catagenesis stage, and the metagenesis stage [12]. The adsorption capacity of kerogen is stronger when the maturity is higher [18,19]. The maturity of kerogen is closely related to the surface functional groups, the proportion of organic matter components, etc. [20], but there is no concise and effective index to reflect this relationship. To solve this problem, the molecule-maturity index (MMI), in which the maturity is directly measured by H/C and O/C atomic ratios, is proposed as a new evaluation index [21,22].

Overall, using microscopic molecular structures is a very effective way to study the characteristics of kerogen types and maturity. In addition, using the molecular structures or components, fruitful studies on the mechanical, adsorption/desorption, and pyrolysis properties of kerogen have been conducted worldwide [23]. Combining NMR with experimental analysis, Lille [24] and Orendt [25] established two-dimensional (2D) and three-dimensional (3D) kerogen monomer molecular models, respectively. Then, quantum chemistry, molecular dynamics (MD), Monte Carlo methods with ^{13}C NMR, and XPS experimental were used by Ungerer et al. [26], to construct the kerogen molecular groups and verify the consistency of molecular model density and thermal maturity change with experimental data. On the basis of the above works, Coasene et al. reconstructed kerogen molecular models of different geological conditions and maturity without restricting molecular functional groups and other information. Mechanical properties such as Young's modulus and elastic modulus were tested, and the relationship between the carbon atom hybridization modes with porosity and maturity was discussed [27]. Wang et al. combined ^{13}C NMR, XPS, and Fourier transform infrared spectroscopy (FT-IR) to construct the 3D monomeric kerogen molecular structure, which achieved results consistent with the experiments. Then, the pyrolysis process of this kerogen molecular model was simulated by hybrid molecular dynamics/force-biased Monte Carlo (MD/fbMC). In addition, the pyrolysis mechanism was also explored [28]. Using kerogen molecular structure models, Yu et al. studied the migration mechanism of methane gas in the pores of kerogen and proved that the continuous model is quantitatively satisfied in the pores of different kerogen types [29].

Above all, the research on kerogen's components and structures provides excellent support for the exploration of shale oil/gas reserves and industrial exploitation yields. Currently, there are two main ways to obtain kerogen components and molecular structures. One way is to collect kerogen components and functional group information based on experiments and then reconstruct kerogen molecules empirically. The other way is to simulate the kerogen molecular group by MD, based on the information obtained from experiments. Both methods are excellent and practical. The only problem, which is a severe problem, is that a sufficient knowledge reserve of theory, simulation, and even molecular reconstruction experience is necessary to analyze the components and molecular structures. Furthermore, tremendous efforts will be spent on repetitive calculations between the trial and error. Therefore, developing a more convenient and effective method to analyze kerogen's components and structures is significant work that will save enormous resources from repetitive trials and errors.

In recent years, the rapid development of machine learning has made it possible to realize the aforementioned idea [30]. Machine learning neural networks can automatically analyze, extract, and record the target features from massive training samples and then use the recorded features to predict the target features in the application [31]. Unfortunately, searching for samples and creating the sample sets requires considerable effort and materials, and it is related to the success or failure of the model. Nevertheless, once the neural network models are successfully trained, they can be used directly as an analysis tool. Neither theoretical reserves nor simulation skills are required for the operators, the prediction process can be complete within a few minutes. Additionally, thousands of sample predictions can be performed simultaneously to achieve high-throughput prediction. This significantly shortens the simulation calculation and analysis time compared to before.

In this study, the machine learning artificial intelligence method is applied to NMR spectra to predict the skeleton components of kerogen. On this basis, the types of kerogen are predicted and analyzed. We propose a method to convert non-uniform 2D graph into the uniform 1D matrix, which makes 2D graph data available for machine learning models. Then, in order to solve the problem of massive training samples, a sample automatic labeling platform is built, and the sample sets containing 22,000 NMR spectra and their molecular labels are constructed for neural network models. Additionally, eight machine learning models are built and tested to find the optimal model. The prediction results show that the trained machine learning model performs well in the prediction process, and the accuracy of the various indicators can reach more than 90% in total. This demonstrates the superior performance that the machine learning method achieves in kerogen prediction.

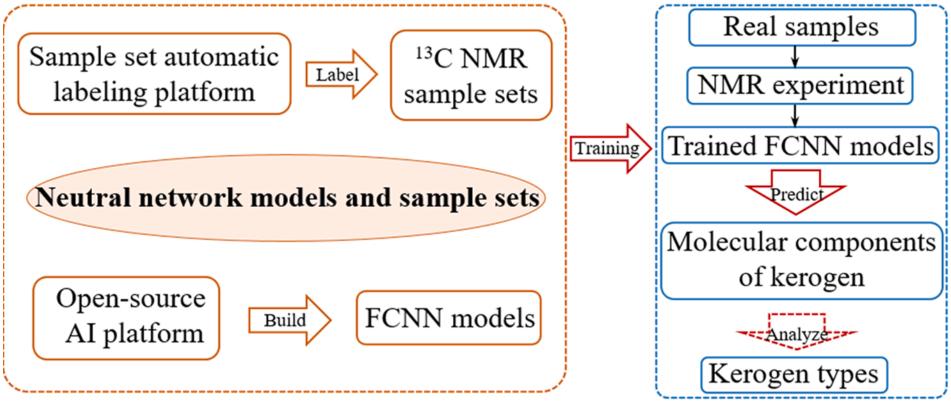

Fig. 1. The framework of machine learning models to predict the components and types of kerogen.

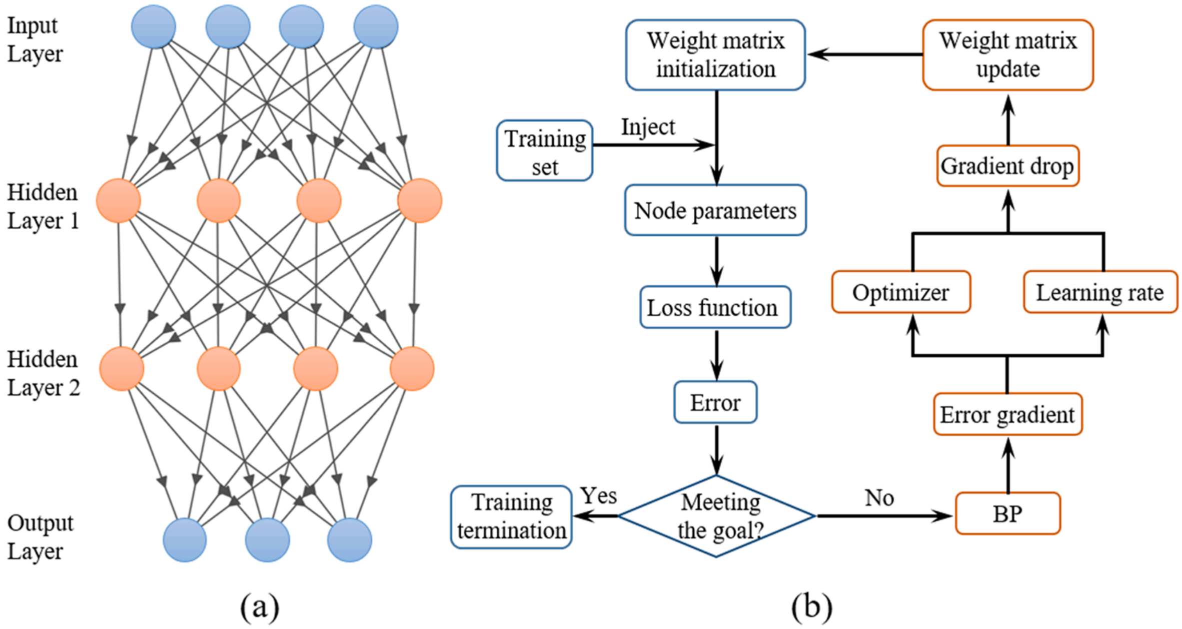

Fig. 2. (a) The schematic diagram of a two-layer FCNN. (b) The training process for machine learning.

2. Methodology and models

2.1. Neural networks

The establishment of artificial neural networks (ANNs) is inspired by the structure of human brain neurons [32]. In 1949, Hebb proposed the Hebbian theory, which explains the working principle of brain neurons, and shows that there is stimulation and strengthening the relationship between neighboring neurons [33]. On this basis, Ivakhnenko and Lapa released the first general working learning algorithm for a feedforward multilayer perceptron for supervised learning in 1967 [34]. After the backpropagation (BP) algorithm was invented and applied to neural networks in 1986 [35,36], neural networks attracted broad attention and became a popular research topic. With the development of big data technologies such as cloud computing and the internet of things, especially after Google's AlphaGo deep learning algorithm defeated the world's top Go player Lee Se-dol in 2016, a new wave of research on artificial neural networks was triggered [37,38]. Artificial intelligence (AI) neural networks have been significantly utilized in geological prospecting [39], medicine and health [40], natural language processing (NLP) [41], and smart agriculture [42].

Neural network models and sample sets are the main factors that affect the prediction accuracy of machine learning neural networks. First, if the model algorithm's design is unreasonable, the neural networks' learning ability would be insufficient. Then, the algorithm would not be able to effectively extract and learn the target features in the training sets. In addition, if the sample sets do not contain all the research target features, the final trained neural networks would not have a sufficient ability to predict the features in the application, resulting in a low generalization ability. According to the degree of human intervention from weak to strong, artificial neural network methods can be divided into three types: unsupervised learning, semi-supervised learning, and supervised learning [43]. Fig. 1 briefly exhibits the framework of this work. The research is based on fully connected neural networks (FCNNs) and supervised learning. Therefore, only supervised learning is introduced here.

2.1.1. Feedforward neural networks

FCNNs adopt a net structure system, in which each neuron node is connected with all the upper and lower layer neurons (Fig. 2(a)). Therefore, the neuron nodes' value is affected by each upper layer neuron, and its value also affects each lower layer neuron, correspondingly.

Structurally, the FCNN model is divided into two parts: the feedforward neural network and the backpropagation algorithm. The feedforward algorithm process of the fully connected neural network algorithm can be expressed as

$$\begin{matrix} \boldsymbol{Z}^l = \boldsymbol{W}^l \boldsymbol{A}^{l-1} + \boldsymbol{b}^l \\ \boldsymbol{A}^l = \sigma(\boldsymbol{Z}^l) \end{matrix}, \tag{1}$$

where $\boldsymbol{A}$ is the value matrix of neural network nodes for each layer; $\boldsymbol{W}$ is the weight matrix of the connections between nodes, which indicates the strength of the influence; $\boldsymbol{b}$ represents the model bias parameter; $\boldsymbol{Z}$ is the linear combination value of weights and biases of each node; and superscript l denotes that the parameter is in the lth layer of the neural networks. Furthermore, σ denotes the activation function of each layer in the neural networks. In this work, the rectified linear unit (ReLU) activation function is selected:

$$\sigma = \mathrm{ReLU}(x) = max(0, x). \tag{2}$$

According to the Eq. (2), the ReLU activates the node value only if it is greater than zero, which is an approach that corresponds to the state of unilateral inhibition and wide excitation boundary in the propagation of biological nerve signals [44]. After the prediction results obtained by the feedforward neural networks, loss function $L\left(y_{\mathrm{pred}}, y_{\mathrm{true}}\right)$ needs to be used to evaluate the accuracy of the model. In this case, the loss function is used to measure the difference between the model prediction $\left(y_{\mathrm{pred}}\right)$ of the number of kerogen skeleton atoms and the actual value y_{true}. The mean squared error (MSE) is defined as

$$L\left(y_{\mathrm{pred}}, y_{\mathrm{true}}\right) = \mathrm{MSE} = \frac{1}{n}\sum_{n}\left(y_{\mathrm{pred}} - y_{\mathrm{true}}\right)^2. \tag{3}$$

As shown in Eq. (3), the MSE represents the mean value of the square of the difference between the actual value and the prediction. The MSE is selected as the loss function for our models.

2.1.2. Backpropagation algorithm

The learning process of neural networks is realized by updating the weight parameters between the network nodes constantly. Currently, training neural networks generally use the gradient descent method. In this process, the gradient of the loss function of the weight value in the current step needs to be calculated and then updated iteratively according to the optimizer optimization plan. The gradient vectors are obtained by the BP algorithm, which means, starting from the output results, the BP algorithm calculates the gradient of each parameter layer by layer in reverse. Hence, BP is the core algorithm for training neural networks.

The calculation process of the backpropagation algorithm is as follows:

$$\begin{aligned}\frac{\partial \boldsymbol{L}}{\partial w_{ij}^{l}} &= \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{Z}^{l}}\frac{\partial \boldsymbol{Z}^{l}}{\partial w_{ij}^{l}}, \\ \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{b}^{l}} &= \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{Z}^{l}}\frac{\partial \boldsymbol{Z}^{l}}{\partial \boldsymbol{b}^{l}}\end{aligned} \tag{4}$$

where w_{ij}^{l} is the components of the matrix $\boldsymbol{W}$. Combined with the calculation process of the feedforward neural networks in Eq. (1), the relationship between the two levels of error terms $\boldsymbol{\delta}$ can be obtained as

$$\boldsymbol{\delta}^{l} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{Z}^{l}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{Z}^{l+1}}\frac{\partial \boldsymbol{Z}^{l+1}}{\partial \boldsymbol{A}^{l}}\frac{\partial \boldsymbol{A}^{l}}{\partial \boldsymbol{Z}^{l}} = \left(\boldsymbol{W}^{l+1}\right)^{T}\boldsymbol{\delta}^{l+1} \odot \sigma^{'}\left(\boldsymbol{Z}^{l}\right), \tag{5}$$

where $\odot$ is the Hadamard product of matrices, which represents the product of the elements corresponding to the matrix position. According to Eq. (5), the lth layer's error term can be calculated directly from the $(l+1)$th error term. This shows that the BP algorithm allows information to flow from back to front through the networks and reduces the computational complexity from $O(n^2)$ to $O(n)$. This is the origin of the name of the BP algorithm. Combining Eqs. (1) and (4), the parameter update of each layer is obtained as follows:

$$\begin{aligned}\boldsymbol{W}_{t+1}^{l} &= \boldsymbol{W}_{t}^{l} - \alpha\boldsymbol{\delta}^{l}\left(\boldsymbol{A}^{(l-1)}\right)^{\mathrm{T}}, \\ \boldsymbol{b}_{t+1}^{l} &= \boldsymbol{b}_{t}^{l} - \alpha\boldsymbol{\delta}^{l}\boldsymbol{I}^{l}\end{aligned} \tag{6}$$

where the subscript t and α denote the position of iterative time step and learning rate, respectively. The learning rate is responsible for adjusting the amplitude of the gradient drop of each iteration parameter during training.

2.1.3. Optimizer

During machine learning training, to obtain the optimal model, an optimization strategy algorithm is created to adjust the amplitude of the gradient descent of the parameters in each step and update the corresponding parameters. This optimization strategy algorithm is called the optimizer. Eq. (6) explains the stochastic gradient descent optimizer (SGD), one of the commonly used optimizers. The learning rate of the SGD always remains constant, causing problems such as slow convergence and oscillation at the saddle point of the gradient [45].

In this work, the adaptive moment estimation optimizer (Adam) is used to control gradient changes [46]. Compared with the SGD, the learning rate of the Adam optimizer will be adaptively adjusted, relying on the current time step and historical parameters, to ensure that the model can quickly and effectively converge to the optimum point.

The neural network models are constructed by two AI open-source frameworks: Google TensorFlow [47] and Baidu PaddlePaddle [48].

2.2. Overfitting and underfitting

During the machine learning training process, in addition to the prediction accuracy, whether overfitting and underfitting appear is another essential indicator of whether the model is suitable for the target. Overfitting refers to the phenomenon in which the model's training error decreases but the test error increases again during the training. This phenomenon occurs mainly because the neural network model's complexity is too high compared to the target problem. The invalid features in the training set are noticed by the model, which obstructs the prediction process of the test set. Therefore, the prediction accuracy is reduced again. Underfitting occurs mainly because the complexity of neural networks or training set is too low compared to the target problem. Thus, the model cannot accurately and effectively identify the target features during the training process, which results in the stagnation of the training accuracy. Hence, to solve the overfitting and underfitting problems, the model's complexity and the training set must be considered comprehensively. Regularization and early stopping methods are used in our work to defend against overfitting.

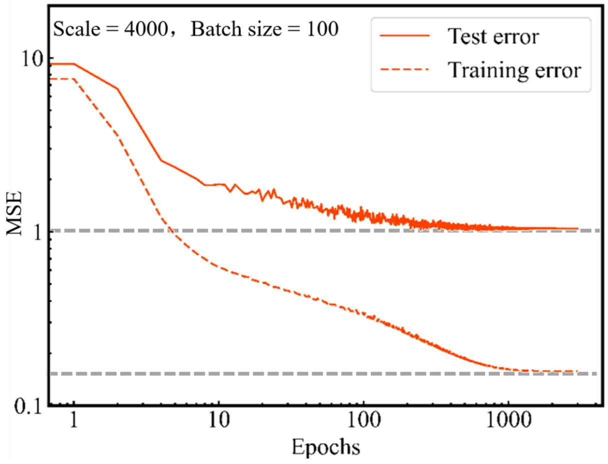

Fig. 3. Training process of neural network model.

The change conditions of the training error and test error for the optimal model obtained during the 3000 epochs training process are shown in Fig. 3. The optimal neural network shows stable convergence after 1000 training epochs, proving that there is no overfitting problem. The minimum training error and test error are 1.01 and 0.07, respectively. Following the definition of the MSE in Eq. (3), it can be approximated that the error of the model in the test set is about one atom per kerogen molecule on average. The discussion of solving the underfitting problem will be further explained in Section 3.2, which analyzes the findings regarding the optimal neural network model.

2.3. NMR spectra

In 1938, Rabi discovered the phenomenon of nuclear magnetic resonance [49]. Since the 1980s, NMR technology has been widely used in the structural analysis of complex molecules such as proteins and nucleic acids [50,51]. The nuclear magnetic resonance spectrum is a 2D graph that records the nuclear magnetic resonance frequency offset of the atoms in a molecule under the influence of the surrounding atoms or functional groups [52]. Based on the offset information, the compound's functional groups and other characteristics can be identified, and the complex molecular structure can be further inferred. In addition, NMR technology has also played a pivotal role in the analysis and construction of kerogen molecules, and NMR spectroscopy is an excellent tool to predict the structural information of kerogen molecules. Usually, the widely used NMR spectrum types include ^{13}C, ^{1}H, ^{15}N, etc. Since the kerogen molecule's carbon skeleton information is mainly concerned in the study, the ^{13}C NMR spectrum is the best choice compared with the other types. During the NMR spectrum calculation, the NMR distinguishability is set to 150 MHz. Because the NMR spectrum obtained at higher NMR frequencies can reflect molecular information more accurately, a higher frequency can be assigned if the conditions permit.

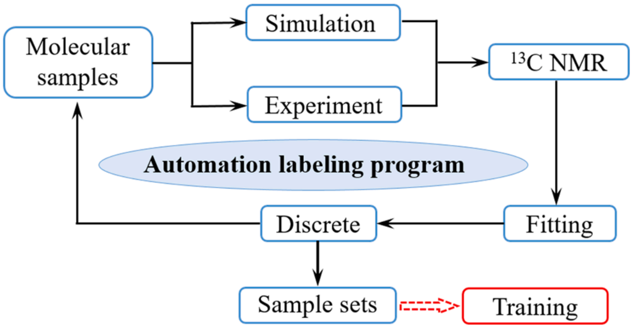

Fig. 4. Schematic diagram of the dataset automatic labeling platform.

2.4. Sample sets for neural network training and testing

The training process of deep learning models often requires tens of thousands of sample data. However, it is unrealistic to either obtain such large quantities of kerogen samples from mining areas or build NMR spectra through experiments. Moreover, it often takes months to make one monomer molecular model of kerogen that meets experimental data. Thus, creating tens of thousands of molecular structure models is also an unrealistic job. Therefore, smaller organic matter molecules are used to train the neural networks as a prediction example. Since the peak offset of an atom is only related to its surrounding atoms or functional groups in NMR spectroscopy, kerogen molecules can be regarded as a group of small molecular functional groups to a certain extent, although the molecules are enormous in scale. Moreover, regardless of the size of the molecules, the basic chemical organization mechanisms are same. The machine learning model will learn the mechanisms during training and then apply them to the prediction. In the traditional method of constructing the 2D and 3D microscopic molecular structure of kerogen, the small functional groups are determined from the 13C NMR spectrum to assemble the macromolecular structure. Each small functional group reflects part of the kerogen properties, and overall functional groups repeat the nature of kerogen [24–26,53]. The NMR spectra of macromolecules generally have different degrees of stacking [54], which may affect the prediction accuracy. However, the machine learning accuracy can still be improved by increasing the distinguishability and setting a reasonable sample scale distribution. Therefore, constructing the sample sets using similar small organic matter molecules is a feasible way.

During the construction of the dataset, the organic matter molecular structures are obtained from PubChem [55], and part of the samples obtained from the laboratory are added. The MestReNova 14 software is used to calculate the NMR spectra of the molecules [56,57]. Because the narrow range and low distinguishability will affect the model's generalization ability and prediction accuracy, a more comprehensive NMR spectrum range can be set to obtain the molecular spectrum in experiments and simulations. Since the range and frequency of the NMR spectra obtained by different molecules in the sample sets may be diverse, 2D graphs cannot be used directly for machine learning. A new method for constructing the tabular data needs to be proposed, and the NMR spectra data must be reconstructed before training. Thus, interpolation fitting is performed on the NMR spectrum, and then the spectrum is discretized on a fixed abscissa. In this way, all spectral data are normalized to a unified dimension, and their abscissas are entirely consistent. The sequenced NMR spectrum can be reorganized into a one-dimensional matrix. The index corresponds to the abscissa of the spectrogram, and the array values correspond to the NMR spectrum values. A total of 8192 sampling points are set during normalization. Hence, the

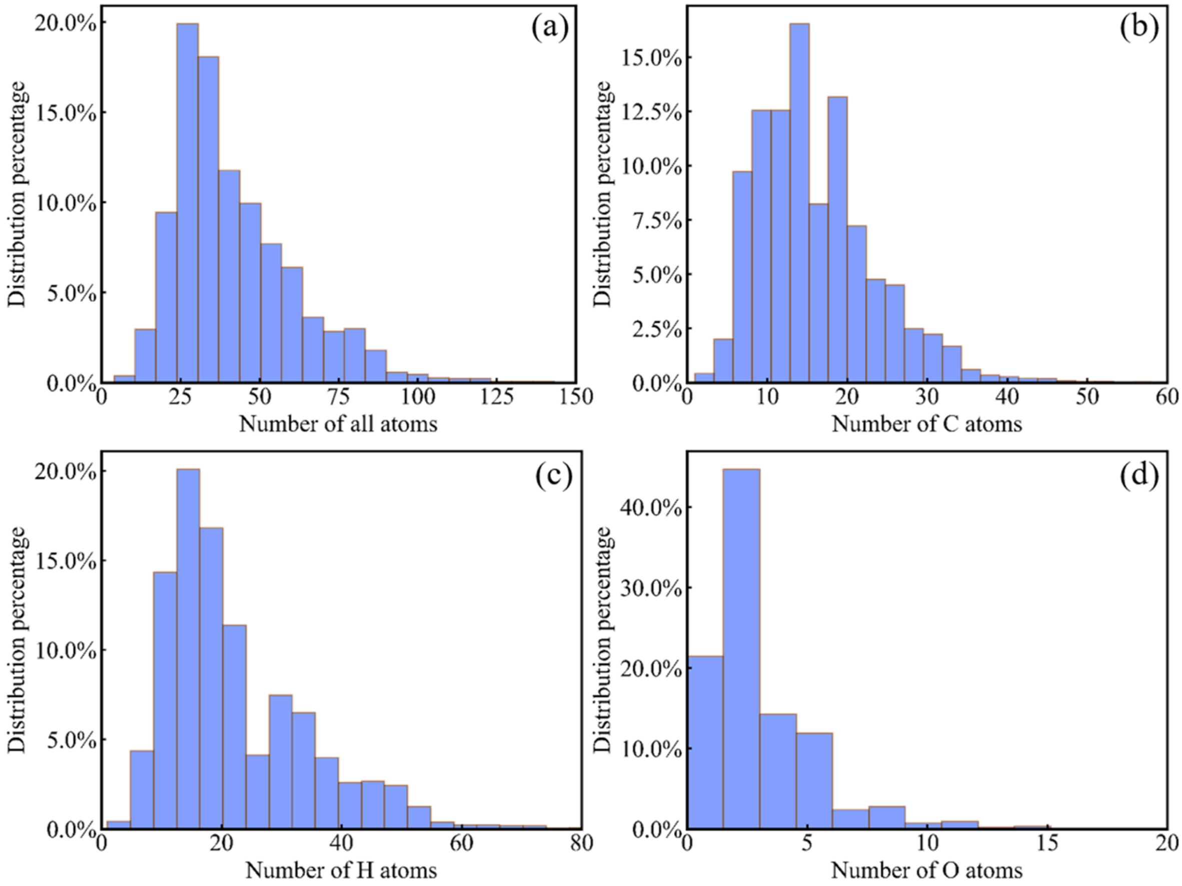

Fig. 5. Molecular-scale distribution of the sample sets. (a) The distribution of all atoms. (b) The distribution of C atoms. (c) The distribution of H atoms. (d) The distribution of O atoms.

Table 2
Statistical information of each component in the sample sets.

	minimum	maximum	average	median
C	1	108	16.2	15
H	1	174	22.8	20
O	0	53	16.2	3

NMR spectrum will be normalized to an array with 8192 columns and injected into the neural network models during training.

Even with the help of NMR spectra calculation software, extensive computation is required to obtain tens of thousands of matched molecules. In addition, calculating the NMR spectra one by one is still too overloaded. Therefore, an automatic molecular annotation program, which combines the post-processing process with automatic annotation for sample sets, is written in the Python language. The workflow is exhibited in Fig. 4. During the labeling, the main python program takes over the MestReNova 14 software thread and calls it as a subroutine. Then it is spliced together with NMR spectra reading, fitting, and discrete subroutines. In other words, the main python program exists as the "glue" in the platform, and the functional subroutines are glued together to automatic cycle operation. Approximately 800 groups of molecules can be labeled in one day on the platform.

3. Results and discussion

3.1. Molecular scale distribution of sample sets

The development of big data and computing power is a prerequisite for machine learning success. This means that prediction accuracy of the neural network model depends on massive samples and their labels. The closer the samples are to the target features, the better the performance of the trained model. The analysis of kerogen molecular components shows that the main elements in kerogen are C, H, and O with small amounts of N and S. The kerogen type depends on the H/C and O/C atomic ratios, which represent the relative contents of hydrogen and oxygen, respectively. Therefore, the dataset is organized by the skeleton atoms (C, H, and O) of kerogen.

The atoms of the sample sets and their proportions are shown in Fig. 5. The total number of atoms in the molecules of the sample sets ranges from 5 to 150. The number of carbon atoms ranges from 1 to 60, and the average scale is 16.2. The number of hydrogen atoms ranges from 1 to 80, and the average scale is 22.8. The number of oxygen atoms ranges from 0 to 20, and the average scale is 3.2. Table 2 gives the maximum and the minimum numbers of each element in the sample sets in detail. Because kerogen is a hydrocarbon source material, its main components are carbon and hydrogen. Compared with the carbon and hydrogen contents, the oxygen content is minimal. Therefore, when the samples are selected, the average number of oxygen atoms in a molecule is lower than those of carbon and hydrogen atoms.

The sample data for training the neural networks are the ^{13}C NMR spectrum and its corresponding molecular structure label. Before training, the dataset is randomly shuffled and then divided into a training set and test set at ratio of 80%:20%. The training set contains 17,600 samples, which are responsible for training the neural network models; and the test set contains 4400 samples, which are responsible for testing the prediction performance of the trained neural networks. There is no crossover between the two data sets to avoid leakage of the test set information and ensure the validity of the test set.

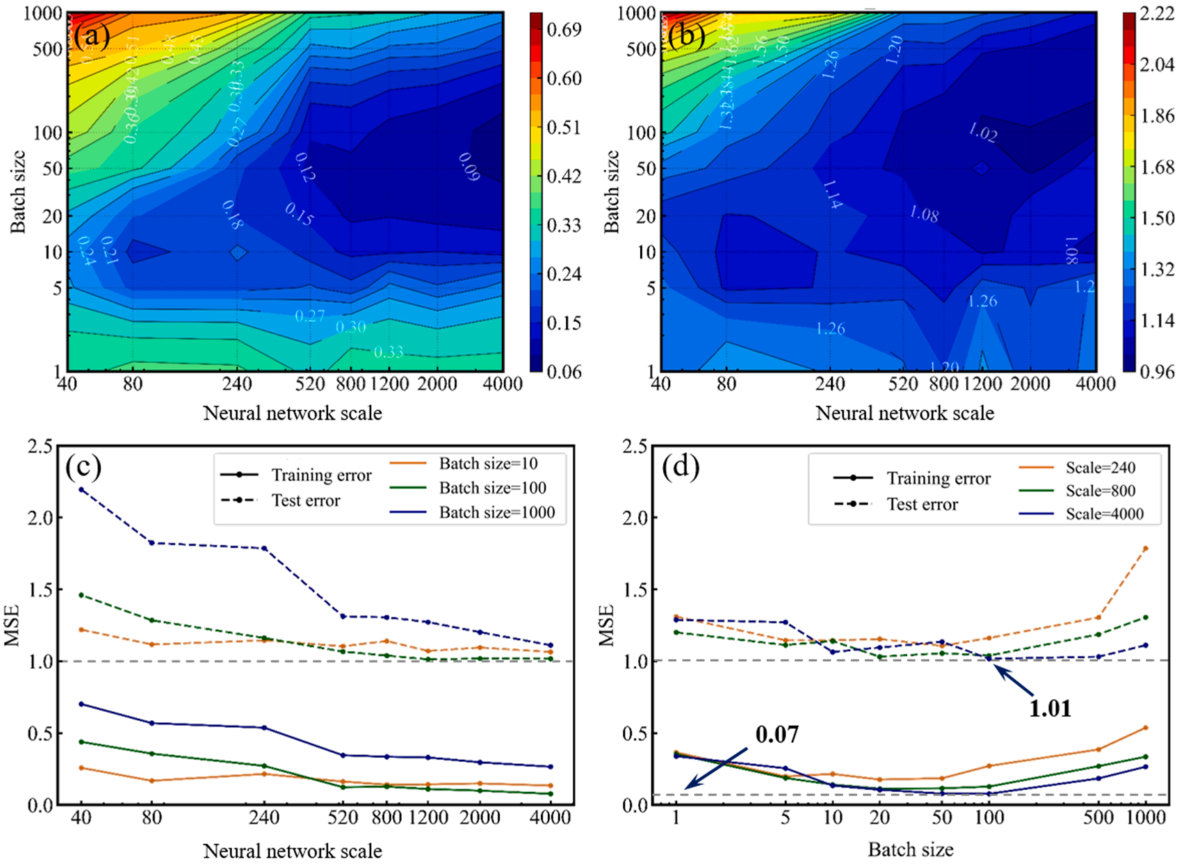

Fig. 6. (a) Training errors of different models and parameters. (b) Test errors of different models and parameters. (c) Errors of different models. (d) Errors of different batch sizes.

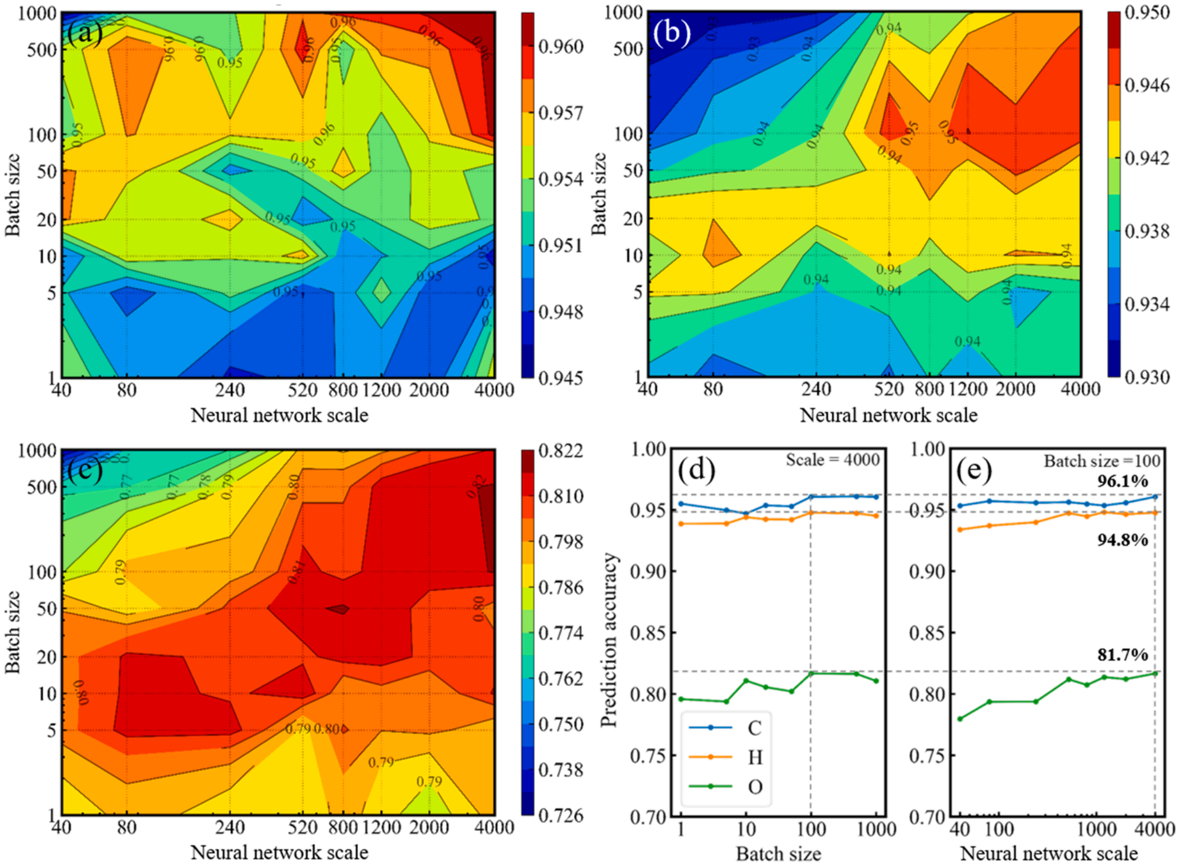

Fig. 7. (a) The accuracy of C. (b) The accuracy of H. (c) The accuracy of O. (d) The accuracy of C/H/O at scale = 4000. (e) The accuracy of C/H/O at batch size = 100.

3.2. *Training and test accuracy of the neural networks*

The training error reflects the performance of the neural network models in the training set. The test error represents the generalization performance on unknown samples. Generally, the focus of the research is on the generalization performance of the model on unknown samples.

To obtain the optimal parameter settings of the neural networks and avoid underfitting, eight models from small to large are built: (20, 20), (40, 40), (200, 40), (400, 120), (400, 400), (600, 600), (1000, 1000) and (2000, 2000), In these models, the first value in parentheses represents the number of nodes in the first layer. The second value represents the number of nodes in the second layer for the FCNN models. For convenience, the model scales are represented by the sum of nodes. During the training process, eight different batch sizes of 1, 5, 10, 20, 50, 100, 500, and 1000 are selected from small to large for each model. The batch size is the number of samples injected into the neural networks per time step. During the prediction process, high-throughput prediction can be achieved by inputting the sample data into the trained model in the same way, and the batch size can be adjusted as needed. The contour figures (Fig. 6(a, b)) exhibit the training and test error distribution of each model and batch size, respectively.

Overall, the training and test error decrease as the model scale increases, and they first decrease and then increase as the batch size increases. The optimal model occurs at the scale = 4000 and batch size = 100, where the training error is 1.01 and the test error is 0.07, which indicates that the training performance and generalization performance of the models under these parameters are in the optimal state. To explain the trend clearly, the error lines at scales = 240, 800, and 4000, batch sizes = 10, 100, and 1000 are plotted in Fig. 6(c, d). According to Fig. 6 (c), the overall trend of the training error and test error of the neural network models decreases. When the scale exceeds 520, the improvement of the model performance tends to be flat, and the optimal value is obtained at the maximum scale. Besides, as the model scale increases, more computing power and training time are necessary. Using this method to improve the performance is not unlimited, and a comprehensive consideration of the computing power and performance is required. As shown in Fig. 6(d), the training and test error changes are slight for batch sizes of 5 – 100 and increase sharply for batch sizes larger than 100. It should be noted that when the batch size increases, the memory consumption of the training machine's GPU will also increase. If the memory exceeds the limit, the training process will be terminated.

3.3. *Prediction accuracy of the kerogen skeleton atoms*

Compared with the average of each component's prediction error in the kerogen molecule using the training error and the test error, the kerogen skeleton atoms' prediction accuracy can better reflect the predictive ability of the model on unknown samples. The prediction accuracy of carbon, hydrogen, and oxygen atoms in different neural network scales and batch sizes are shown in Fig. 7(a, b, c). Overall, the trend of accuracy matches the trends of the training and test errors. The larger the model scale is, the better the model prediction performance. In addition, the prediction accuracy first increases and then decreases again as the batch size increase. The optimal value is obtained when the batch size = 100 – 500. The accuracy curve lines of C, H, and O at the model scale = 4000 and batch size = 100 are drawn in Fig. 7(d). As the scale of the model increases, the accuracy increases accordingly and gradually stabilizes. The highest prediction accuracy is also obtained at batch size = 100. The optimal model's prediction accuracy for carbon atoms and hydrogen atoms reaches 96.1% and 94.8%, respectively, and the accuracy of oxygen atoms is slightly worse than that of carbon and hydrogen, at 81.7%. Because the oxygen content in kerogen molecules is

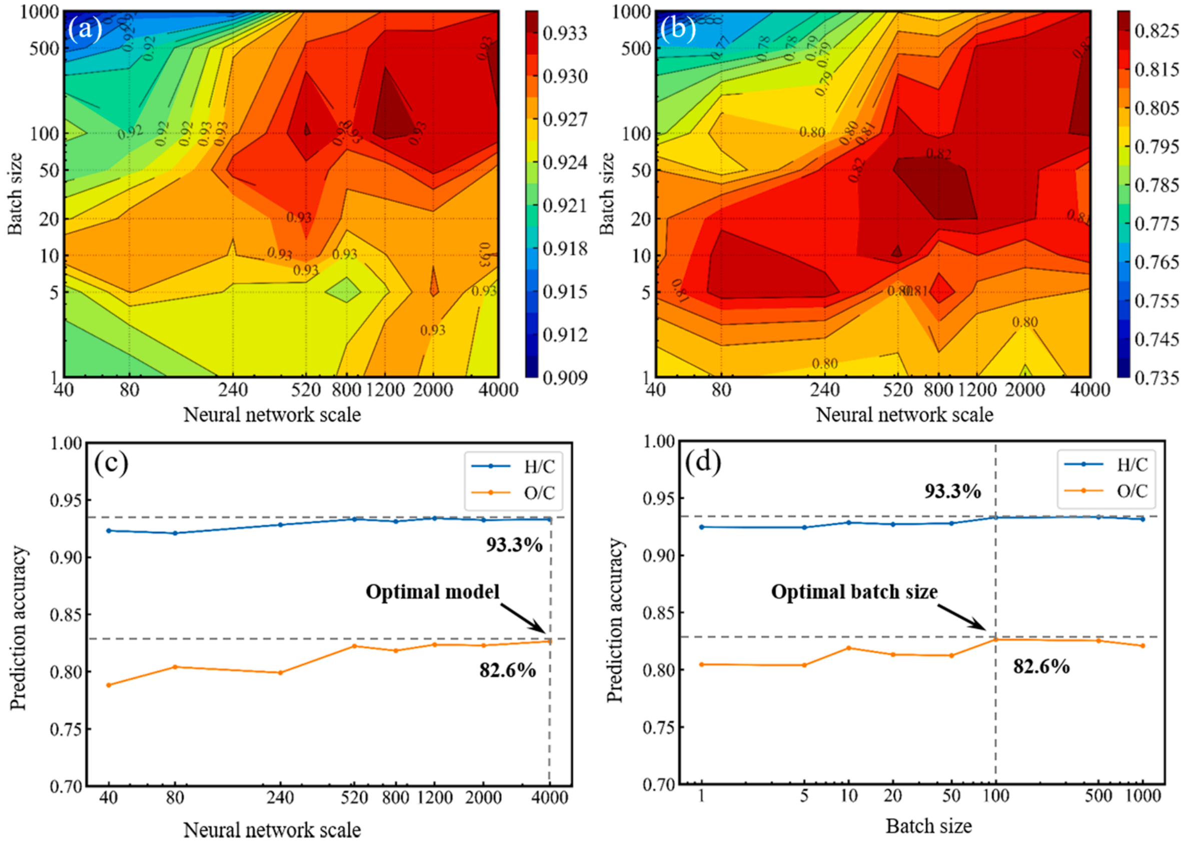

Fig. 8. (a) The prediction accuracy of H/C. (b) The prediction accuracy of O/C. (c) The prediction accuracy of H/C and O/C at batch size = 100. (d) The prediction accuracy of H/C and O/C at scale = 4000.

extremely low. Generally, the O/C atomic ratio ranges from 0.03 to 0.3, which leads to differences in the number of elements in the sample sets. Thus it is difficult for neural networks to obtain the characteristics of oxygen atoms with the same accuracy as carbon and hydrogen from the training set.

3.4. Prediction accuracy of the kerogen types

According to the context in Chapter 1, O/C and H/C are the most suitable prediction performance indicators to evaluate neural network models. The prediction accuracy of O/C and H/C is shown in Fig. 8. It can be seen that the H/C atomic ratio prediction accuracy of the test set achieves the optimal value at batch size = 100 and model scale = 520, 1200, and 4000. The optimal accuracy of the H/C atomic ratio approximately occurs when the batch size is in the range of 20 – 500, and the model scale is in the range of 520 – 4000. In Section 3.2, the optimal model parameters are obtained at batch size = 100 and scale = 4000, so these parameters are also selected as the optimal model. Fig. 8 (c, d) prove that the parameters are suitable for the study. Under this model, the prediction accuracy of the O/C atomic ratio is 82.6%, and that of the H/C is 93.3%.

Three hundred groups of molecular samples, which are the pyrolysis products of kerogen, with O/C atomic ratios ranging from 0 to 0.35 and H/C atomic ratios between 0.25 and 2 are selected for the validation set. The kerogen samples are taken from the Erdos and Songliao basins in China, and the purified kerogen samples are decomposed at 650 °C via

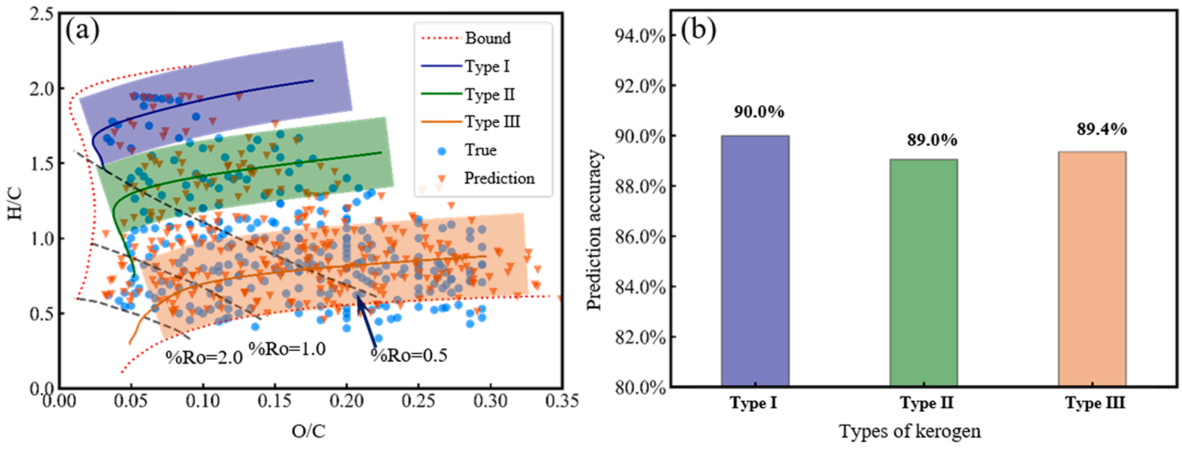

Fig. 9. (a) van Krevelen diagram of the validation set. (b) The prediction accuracy of kerogen types.

the platinum wire cracking experiment [28]. And the platinum wire pyrolysis experiment has the advantages of a fast heating rate, less secondary reactions, and stable pyrolysis products. The validation set is constructed further to test the generalization ability of the optimal neural network and determine the prediction ability of the neural network model for the three kerogen types. It should be noted that the structures of the molecules in the verification set are entirely different from those in the training set and the test set. In other words, there is no intersection between the three sets.

The distribution of 300 molecules in the validation set and the three kerogen types ranges are given in Fig. 9(a). The blue dots represent the true values of the molecules, and the brown triangles represent the predicted values. Type I, type II, and type III kerogen are marked in purple, green, and brown, respectively. The red dotted lines represent the upper and lower bounds of the kerogen type distributions. The black dotted lines denote the maturity (Ro) of kerogen. As shown in Fig. 9(b), the accuracy of the optimal neural network model in the verification set is 90.0%, 89.0%, and 89.4% for type I, type II, and type III kerogen, respectively, demonstrating the excellent predictive ability of the neural network model for the types of kerogen.

4. Conclusions

In summary, we propose a machine learning neural network method to predict the components and types of kerogen based on NMR spectra. Compared with conventional techniques, the machine learning method has the advantages of high-throughput and accurate prediction and does not require any knowledge accumulation of the operators.

The sample sets and neural network models are the two most important parts of machine learning. In terms of sample sets, to make it so that the 2D graphs can be read by machine learning models, we propose a method to reconstruct the non-uniform 2D graph into the uniform 1D matrix. Additionally, a sample set automatic labeling platform is built. Over 22,000 group organic matter molecular samples are labeled on this platform. In terms of the models, eight different scales of neural network models are constructed, and eight input parameters are set during training. The results show that the optimal model is achieved at batch size = 100 and scale = 4000. The model's training and test error stably converge to 0.07 and 1.01, respectively, which means that the prediction error is approximately one atom per molecule in unknown samples. The prediction accuracy of kerogen skeleton elements is 96.1%, 94.8%, and 81.7% for carbon, hydrogen, and oxygen, respectively. Based on this, the accuracy of the H/C and O/C atom ratios is analyzed. The H/C is 93.3% and O/C is 82.6% in the optimal model.

Finally, the validation set is built to test the generalization performance of the optimal model. Three hundred molecules, which are cracked from the mined kerogen samples, are contained in the validation set. The results show that the accuracy of the optimal model for the three kerogen types is 90.1%, 89.0%, and 89.4%, respectively. Although the sample set molecules are smaller than real kerogen molecules, as an example to verify the feasibility of the machine learning method, organic matter molecules are a suitable choice. In addition, the optimal model plays an excellent performance in the prediction of kerogen skeleton components and types. It is believed that this new method will be a meaningful attempt not only to simplify the repetitive work in the analysis between trial and error but also to provide a solid foundation to build kerogen molecular groups automatically.

CRediT authorship contribution statement

Dongliang Kang: Investigation, Methodology, Software, Data curation, Visualization, Writing - original draft. **Xiaohe Wang:** Investigation, Methodology. **Xiaojiao Zheng:** Software. **Ya-Pu Zhao:** Conceptualization, Resources, Funding acquisition, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was jointly supported by the National Natural Science Foundation of China (NSFC, Grant No. 12032019, 11872363, 51861145314), the Chinese Academy of Sciences (CAS) Key Research Program of Frontier Sciences (Grant No. QYZDJ-SSW-JSC019), and the CAS Strategic Priority Research Program (Grant No. XDB22040401).

References

[1] Yew CH, Weng X. Mechanics of hydraulic fracturing. 2nd ed. Houston: Gulf Professional Publishing; 2015. DOI:10.1016/C2013-0-12927-3.
[2] Shen W, Zhao Y-P. Quasi-static crack growth under symmetrical loads in hydraulic fracturing. J Appl Mech 2017;84:81009. https://doi.org/10.1115/1.4036988.
[3] Shen W, Zhao Y-P. Combined effect of pressure and shear stress on penny-shaped fluid-driven cracks. J Appl Mech 2018;85:31003. https://doi.org/10.1115/1.4038719.
[4] Montgomery CT, Smith MB. Hydraulic fracturing: History of an enduring technology. J Pet Technol 2010;62:26–40. https://doi.org/10.2118/1210-0026-JPT.
[5] Durand B. Kerogen: Insoluble organic matter from sedimentary rocks. 1st ed. Paris: Editions Technip; 1980.
[6] Lin K, Yuan Q, Zhao Y-P. Using graphene to simplify the adsorption of methane on shale in MD simulations. Comput Mater Sci 2017;133:99–107. https://doi.org/10.1016/j.commatsci.2017.03.010.
[7] Radke M, Welte D, Willsch H. Maturity parameters based on aromatic hydrocarbons: influence of the organic matter type. Org Geochem 1986;10:51–63. https://doi.org/10.1016/0146-6380(86)90008-2.
[8] Zhang T, Ellis GS, Ruppel SC, Milliken K, Yang R. Effect of organic-matter type and thermal maturity on methane adsorption in shale-gas systems. Org Geochem 2012;47:120–31. https://doi.org/10.1016/j.orggeochem.2012.03.012.
[9] Van Krevelen DW. Coal: typology, physics, chemistry, constitution. 3rd ed. Amsterdam: Elsevier Science; 1993.
[10] Dow WG. Kerogen studies and geological interpretations. J Geochem Explor 1977;7:79–99. https://doi.org/10.1016/0375-6742(77)90078-4.
[11] Kelemen S, Afeworki M, Gorbaty M, Sansone M, Kwiatek P, Walters C, et al. Direct characterization of kerogen by X-ray and solid-state 13C nuclear magnetic resonance methods. Energy Fuels 2007;21:1548–61. https://doi.org/10.1021/ef060321h.
[12] Tissot BP, Welte DH. Petroleum formation and occurrence. 2nd ed. New York: Springer Science & Business Media; 1984. DOI:10.1007/978-3-642-87813-8.
[13] Gasparik M, Bertier P, Gensterblum Y, Ghanizadeh A, Krooss BM, Littke R. Geological controls on the methane storage capacity in organic-rich shales. Int J Coal Geol 2014;123:34–51. https://doi.org/10.1016/j.coal.2013.06.010.
[14] Huang X, Zhao Y-P. Characterization of pore structure, gas adsorption, and spontaneous imbibition in shale gas reservoirs. J Pet Sci Eng 2017;159:197–204. https://doi.org/10.1016/j.petrol.2017.09.010.
[15] Zhao Y-P. Physical mechanics of surfaces and interfaces. 1st ed. Beijing: Science Press; 2012.
[16] Lin K, Huang X, Zhao Y-P. Combining image recognition and simulation to reproduce the adsorption/desorption behaviors of shale gas. Energy Fuels 2019;34:258–69. https://doi.org/10.1021/acs.energyfuels.9b03669.
[17] Zhao Y-P. Lectures on mechanics. 1st ed. Beijing: Science Press; 2018. (in Chinese).
[18] Okiongbo KS, Aplin AC, Larter SR. Changes in type II kerogen density as a function of maturity: evidence from the kimmeridge clay formation. Energy Fuels 2005;19:2495–9. https://doi.org/10.1021/ef050194+.
[19] Busch A, Gensterblum Y. CBM and CO2-ECBM related sorption processes in coal: a review. Int J Coal Geol 2011;87:49–71. https://doi.org/10.1016/j.coal.2011.04.011.
[20] Huang L, Ning Z, Wang Q, Qi R, Zeng Y, Qin H, et al. Molecular simulation of adsorption behaviors of methane, carbon dioxide and their mixtures on kerogen: effect of kerogen maturity and moisture content. Fuel 2018;211:159–72. https://doi.org/10.1016/j.fuel.2017.09.060.
[21] Burnham AK. Kinetic models of vitrinite, kerogen, and bitumen reflectance. Org Geochem 2019;131:50–9. https://doi.org/10.1016/j.orggeochem.2019.03.007.
[22] Wang X, Zhao Y-P. The time-temperature-maturity relationship: A chemical kinetic model of kerogen evolution based on a developed molecule-maturity index. Fuel 2020;278. DOI:10.1016/j.fuel.2020.118264.
[23] Vandenbroucke M, Largeau C. Kerogen origin, evolution and structure. Org Geochem 2007;38:719–833. https://doi.org/10.1016/j.orggeochem.2007.01.001.
[24] Lille Ü, Heinmaa I, Pehk T. Molecular model of Estonian kukersite kerogen evaluated by 13C MAS NMR spectra☆. Fuel 2003;82:799–804. https://doi.org/10.1016/S0016-2361(02)00358-7.
[25] Orendt AM, Pimienta ISO, Badu SR, Solum MS, Pugmire RJ, Facelli JC, et al. Three-dimensional structure of the siskin green river oil shale kerogen model: a

comparison between calculated and observed properties. Energy Fuels 2013;27:702–10. https://doi.org/10.1021/ef3017046.

[26] Ungerer P, Collell J, Yiannourakou M. Molecular modeling of the volumetric and thermodynamic properties of kerogen: influence of organic type and maturity. Energy Fuels 2014;29:91–105. https://doi.org/10.1021/ef502154k.

[27] Bousige C, Ghimbeu CM, Vix-Guterl C, Pomerantz AE, Suleimenova A, Vaughan G, et al. Realistic molecular model of kerogen's nanostructure. Nat Mater 2016;15:576–82. https://doi.org/10.1038/nmat4541.

[28] Wang X, Huang X, Lin K, Zhao YP. The constructions and pyrolysis of 3D kerogen macromolecular models: experiments and simulations. Glob Chall 2019;3:1900006. https://doi.org/10.1002/gch2.201900006.

[29] Yu H, Xu H, Xia J, Fan J, Wang F, Wu H. Nanoconfined transport characteristic of methane in organic shale nanopores: the applicability of the continuous model. Energy Fuels 2020;34:9552–62. https://doi.org/10.1021/acs.energyfuels.0c01789.

[30] Alpaydin E. Introduction to machine learning. 4th ed. Cambridge: MIT press; 2020.

[31] Yegnanarayana B. Artificial neural networks. 1st ed. New Delhi: Prentice-Hall of India Pvt Ltd; 2006.

[32] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

[33] Hebb DO. The organization of behavior: a neuropsychological theory. 1st ed. New York: Wiley; 1949.

[34] Ivakhnenko AGe, Lapa VGe. Cybernetics and forecasting techniques. 1st ed. New York: Elsevier; 1967.

[35] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–6. https://doi.org/10.1038/323533a0.

[36] Hecht-Nielsen R. Theory of the backpropagation neural network. In: Wechsler H. Neural networks for perception, San Diego: Academic Press; 1992, p. 65-93. DOI:10.1016/C2013-0-11676-5.

[37] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9. https://doi.org/10.1038/nature16961.

[38] Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 2018;362:1140–4. https://doi.org/10.1126/science.aar6404.

[39] Liu Y, Wu L. Geological disaster recognition on optical remote sensing images using deep learning. Procedia Comput Sci 2016;91:566–75. https://doi.org/10.1016/j.procs.2016.07.144.

[40] Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inf 2017;21:4–21. https://doi.org/10.1109/JBHI.2016.2636665.

[41] Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag 2018;13:55–75. https://doi.org/10.1109/MCI.2018.2840738.

[42] Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: a survey. Comput Electron Agric 2018;147:70–90. https://doi.org/10.1016/j.compag.2018.02.016.

[43] Zhao Y-P. A course in rational mechanics. 1st ed. Beijing: Science Press; 2020. (in Chinese).

[44] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: ICML'10: Proceedings of the 27th international conference on machine learning; 2010. p. 807–14.

[45] Ruder S. An overview of gradient descent optimization algorithms. arXiv 2016; arXiv preprint:1609.04747.

[46] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv 2014;arXiv preprint:1412.6980.

[47] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI'16); 2016. p. 265–83.

[48] Ma Y, Yu D, Wu T, Wang H. PaddlePaddle: an open-source deep learning platform from industrial practice. Front Data Comput 2019;1:105–15. https://doi.org/10.11871/jfdc.issn.2096.742X.2019.01.011.

[49] Rabi II, Zacharias JR, Millman S, Kusch P. A new method of measuring nuclear magnetic moment. Phys Rev 1938;53:318. https://doi.org/10.1103/PhysRev.53.318.

[50] Zweckstetter M. NMR: prediction of molecular alignment from structure using the PALES software. Nat Protoc 2008;3:679–90. https://doi.org/10.1038/nprot.2008.36.

[51] Robien W. A critical evaluation of the quality of published 13C NMR data in natural product chemistry. In: Prog Chem Org Nat Prod 105. Springer; 2017. p. 137–215. https://doi.org/10.1007/978-3-319-49712-9_3.

[52] Jackman LM, Sternhell S. Application of Nuclear Magnetic Resonance Spectroscopy in Organic Chemistry. 2nd ed. Oxford: Pergamon Press; 1969. DOI:10.1016/C2013-0-03028-9.

[53] Siskin M, Scouten CG, Rose KD, Aczel T, Colgrove SG, Pabst RE. Detailed Structural Characterization of the Organic Material in Rundle Ramsay Crossing and Green River Oil Shales. In: Snape C. Composition, Geochemistry and Conversion of Oil Shales, Dordrecht: Springer; 1995, p. 143-158. DOI:10.1007/978-94-011-0317-6_9.

[54] Bovey FA, Mirau PA, Gutowsky H. Nuclear magnetic resonance spectroscopy. 2nd ed. San Diego: Academic Press; 1988.

[55] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res 2016;44:D1202–13. https://doi.org/10.1093/nar/gkv951.

[56] Bremser W. HOSE—a novel substructure code. Anal Chim Acta 1978;103:355–65. https://doi.org/10.1016/S0003-2670(01)83100-7.

[57] Willcott MR. MestRe Nova. J Am Chem Soc 2009;131:13180-13180. DOI:10.1021/ja906709t.