Petroleum Science 21 (2024) 752–764

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

Intelligent geochemical interpretation of mass chromatograms: Based on convolution neural network

Kai-Ming Su [a, b, *], Jun-Gang Lu [c], Jian Yu [d], Zi-Xing Lu [d], Shi-Jia Chen [c]

[a] Hubei Key Laboratory of Petroleum Geochemistry and Environment, Wuhan, 430100, Hubei, China
[b] College of Resources and Environment, Yangtze University, Wuhan, 430100, Hubei, China
[c] School of Geoscience and Technology, Southwest Petroleum University, Chengdu, 610500, Sichuan, China
[d] Changqing Oilfield Company, PetroChina, Xi'an, 710021, Shaanxi, China

ARTICLE INFO

Article history:
Received 26 February 2022
Received in revised form
23 May 2023
Accepted 8 November 2023
Available online 11 November 2023

Edited by Jie Hao and Teng Zhu

Keywords:
Organic geochemistry
Biomarker
Mass chromatographic analysis
Automated interpretation
Convolution neural network
Machine learning

ABSTRACT

Gas chromatography-mass spectrometry (GC-MS) is an extremely important analytical technique that is widely used in organic geochemistry. It is the only approach to capture biomarker features of organic matter and provides the key evidence for oil-source correlation and thermal maturity determination. However, the conventional way of processing and interpreting the mass chromatogram is both time-consuming and labor-intensive, which increases the research cost and restrains extensive applications of this method. To overcome this limitation, a correlation model is developed based on the convolution neural network (CNN) to link the mass chromatogram and biomarker features of samples from the Triassic Yanchang Formation, Ordos Basin, China. In this way, the mass chromatogram can be automatically interpreted. This research first performs dimensionality reduction for 15 biomarker parameters via the factor analysis and then quantifies the biomarker features using two indexes (i.e. *MI* and *PMI*) that represent the organic matter thermal maturity and parent material type, respectively. Subsequently, training, interpretation, and validation are performed multiple times using different CNN models to optimize the model structure and hyper-parameter setting, with the mass chromatogram used as the input and the obtained *MI* and *PMI* values for supervision (label). The optimized model presents high accuracy in automatically interpreting the mass chromatogram, with R^2 values typically above 0.85 and 0.80 for the thermal maturity and parent material interpretation results, respectively. The significance of this research is twofold: (i) developing an efficient technique for geochemical research; (ii) more importantly, demonstrating the potential of artificial intelligence in organic geochemistry and providing vital references for future related studies.

1. Introduction

Biomarker, also known as “molecular fossil”, refers to the compound that originates from living organisms and presents some typical features (Peters et al., 2007). By investigating biomarkers, petroleum geochemists can extract information such as age, parent material type, sedimentary environment, and thermal maturity of organic matter from hydrocarbon fluids and deposits, which provides important references for hydrocarbon generation potential assessment, thermal maturity evaluation, and oil-source correlation. Biomarker analysis is regarded as one of the most representative achievements in modern petroleum exploration (Kaufman et al., 1990; Isaksen and Bohacs, 1995).

Gas chromatography-mass spectrometry (GC-MS) is the main approach for evaluating biomarkers (Seifert and Moldowan, 1978; Lin and Abbas, 1990; Peters et al., 2007). It can identify and quantify typical biomarkers by separating the organic mixture with the chromatograph and determining molecular structures of compounds with the mass spectrometer. Mass chromatogram, as the primary data produced by GC-MS, displays itself as a fluctuating curve, with each hump (peak) representing an individual compound and the peak area (or height) representing the corresponding compound content. Therefore, various biomarker features are associated with different shapes of mass chromatograms. Mass chromatograms of sterane ($m/z = 217$) and terpane ($m/

[*] Corresponding author.
E-mail address: sukaiming@yangtzeu.edu.cn (K.-M. Su).

https://doi.org/10.1016/j.petsci.2023.11.010

$z = 191$) frequently used in the geochemical analysis are presented in Fig. 1a and b.

However, it is hard to directly apply such mass chromatograms to evaluation and investigation related to petroleum exploration, because most petroleum explorationists are not proficient in handling such data that need to be processed and interpreted by technicians of organic geochemistry or analytical chemistry. The typical analysis process includes labeling peaks, identifying compounds, calculating peak areas, and summarizing the analysis results (Fig. 2), which will eventually produce the parameters that can characterize the biomarkers, such as $C_{29}\alpha\alpha\alpha S/(S+R)$, $Ts/(Ts+Tm)$, and $Ga/C_{30}\alpha\beta$. With the current approach and technique, the wide application of GC-MS is constrained due to its high dependence on geochemical/chemical professionals. In this context, it is crucial to develop a method that can directly and automatically investigate and extract information of interest from mass chromatograms such as thermal maturity and parent material type of organic matter.

Linking the mass chromatogram and the corresponding biomarker feature is, in essence, an issue of supervised learning, which can be addressed by applying the convolution neural network (CNN) (Koeshidayatullah et al., 2020). The deep neural network is incredibly competent to capture the variation pattern of massive data (Reichstein et al., 2019) and performs better than conventional data analysis methods (Bergen et al., 2019), which makes it one of the cutting-edge techniques for probing geological problems, including thin section analysis (Koeshidayatullah et al., 2020), solid mineral deposit prediction (Li et al., 2020a; Zhang et al., 2021a), elemental regularity analysis of underground water (Yu et al., 2020), geophysical prospecting (Ho, 2009), and engineering geology (Wei et al., 2021). Nonetheless, its application in organic geochemistry is still rare.

The CNN method builds the correlation model between the known data (independent variables, and in this case, mass chromatograms) and labels (dependent variables, and in this case, biomarker features), which can achieve classification, regression, and prediction of unknown data. Our initial idea is to apply machine learning to the dataset of the conventional biomarker parameters (e.g. $Ts/(Ts+Tm)$ and $Ga/C_{30}\alpha\beta$), and yet this practice seems to be similar to existing chemometrics methods in terms of the analysis performance (Zumberge, 1987; Pan et al., 2017). Therefore, it is more straightforward to apply artificial intelligence techniques in interpreting the original mass chromatogram, which can automatically interpret biomarker features, eliminate numerous time-consuming and labor-intensive steps in conventional methods, and more importantly, improve the profundity and accuracy of the research. For instance, the proposed method in this paper introduces extra information, like the baseline shape of the mass chromatogram, into the analysis, thereby contributing to good analysis performances (Fig. 3).

2. Geological background

The Ordos Basin is one of the most important onshore petroliferous basins in China (Fig. 4a). Tectonically, it is located in the western part of the North China craton, with stable internal structures and rarely developed faults (Yang et al., 2006; Liu et al., 2019). It consists of six major structural units, namely the northern Yimeng Uplift, the southern Weibei Uplift, the eastern Jinxi Flexure Belt, the western Tianhuan Depression and thrust belt, and the central wide Yishan Slope (Li et al., 2020b).

The Triassic Yanchang Formation is divided into ten members according to sedimentary cycles and lithological features (named as Chang 10–1 from bottom to top) (Fig. 4b). This formation is the sedimentary product of a complete lacustrine transgression and regression cycle (Yang, 2004; Qu et al., 2020; Zhang et al., 2021b). The Chang 7 member is a set of organic-matter-rich mudstone and shale with a thickness up to 20–60 m, and it is interpreted to be formed during the peak development of the lake (Li et al., 2020b). This set of source rocks is featured by high organic matter abundance, in which the shale is mostly of Type I and II_1 organic matter, with a TOC range of 8.0%–16.0%, where the mean value is 13.8%, while the mudstone is mainly of Type II_1 and II_2 organic matter, with a TOC range of 2.0%–6.0%, where the mean value is 3.7%. The vitrinite reflectance (R_o) is in the range of 0.8%–1.0% for both mudstone and shale (Yang et al., 2016), suggesting high thermal maturity. In addition, controlled by the secondary sedimentary cycles, smaller-scale sources are also developed to different degrees in other members (Fig. 4b), particularly in the Chang 9 member (Li et al., 2012; Yang et al., 2017; Zou et al., 2017). Numerous tight and low-permeability oil reservoirs are developed in the good source-reservoir rock assemblage formed by high-quality Chang 7 and Chang 9 source rocks as well as widely distributed Yanchang Formation deltaic deposits (Zou, 2014).

3. Dataset preparation and model training

3.1. Sample types, pre-treatment, and GC-MS

A total of 108 core samples are collected, including 74 sandstone samples and 34 mudstone samples. These samples are first cleaned using distilled water, then fully dried, and grounded into 80-mesh

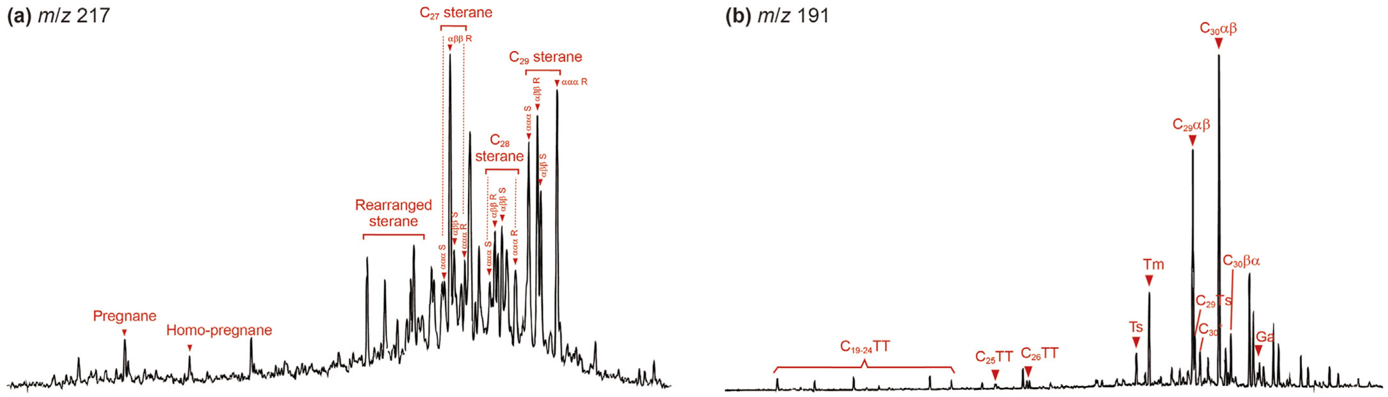
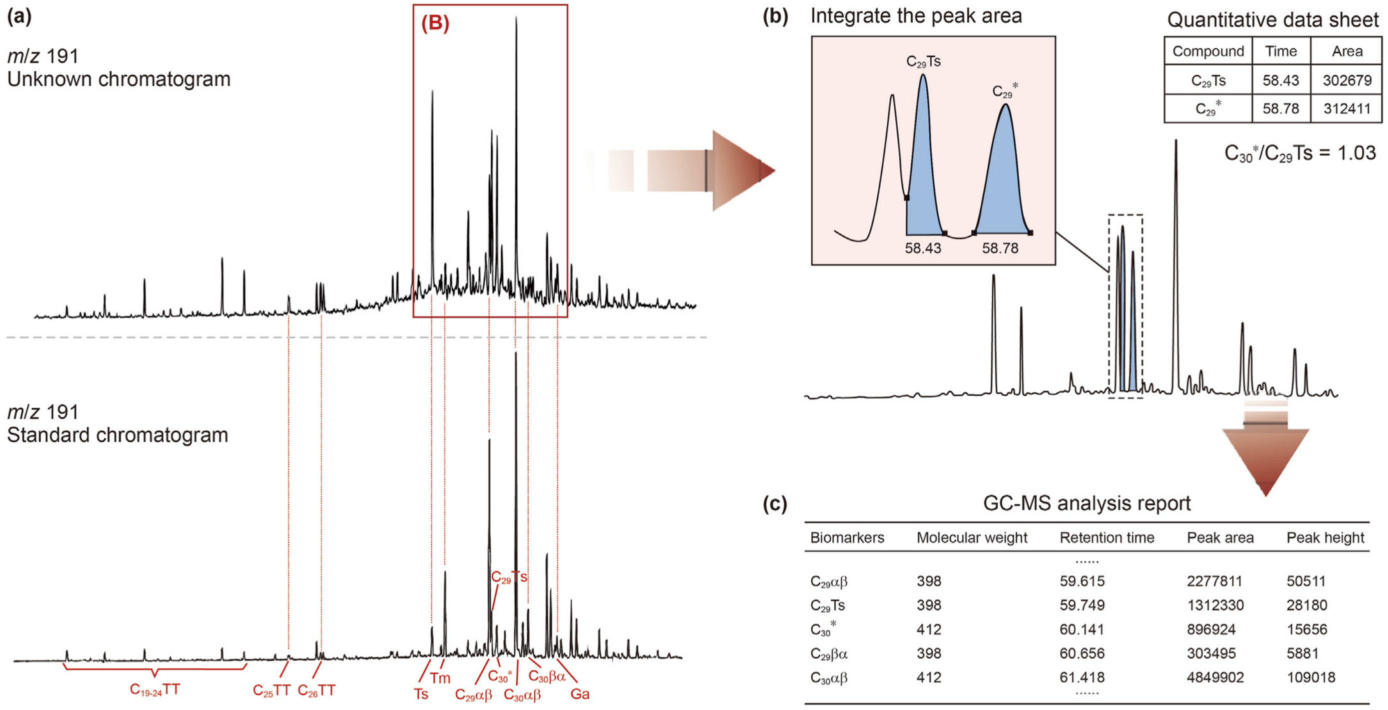


Fig. 1. The mass chromatograms for **(a)** $m/z = 217$ and **(b)** $m/z = 191$, with involved typical biomarkers in the mudstone sample collected at the depth of 2041 m in Well G21, Ordos Basin.



Quantitative data sheet

Compound	Time	Area
C_{29}Ts	58.43	302679
C_{29}^*	58.78	312411

GC-MS analysis report

Biomarkers	Molecular weight	Retention time	Peak area	Peak height
		......		
$C_{29}\alpha\beta$	398	59.615	2277811	50511
C_{29}Ts	398	59.749	1312330	28180
C_{30}^*	412	60.141	896924	15656
$C_{29}\beta\alpha$	398	60.656	303495	5881
$C_{30}\alpha\beta$	412	61.418	4849902	109018
		......		

Fig. 2. The conventional processing workflow of the mass chromatogram is complex, time-consuming, and labor-intensive, typically, including **(a)** correlating the peaks of a measured spectrum to the corresponding compound, according to the standard spectrum; **(b)** calculating the area of each peak via integration; **(c)** preparing the analysis report. Only an extremely small part of the typical tasks is presented above, and in most cases nearly 100 peaks need to be identified and calculated via integration for a single sample.

powders. Subsequently, the powder samples are put through the Soxhlet extraction for 72 h using the trichloromethane as the solvent, and alkanes are separated via chromatography. The alkane analysis is performed via the GC-MS/MS method using the Agilent 7890A gas chromatography system connected in series with the 5975C mass spectrometer. The carrier gas is helium, and the chromatographic column is the HP-5MS elastic quartz capillary column (30 m × 250 μm × 0.25 μm). The ion source temperature is 230 °C, the inlet temperature is 250 °C, the quadrupole temperature is 150 °C, and the filament current is 35 μA. Heating is programed as holding at the initial temperature of 100 °C for 2 min, then heating to 300 °C at a rate of 3 °C/min, and holding at 300 °C for 20 min. The mass chromatograms of terpane ($m/z = 191$) and sterane ($m/z = 217$) are exported from the Agilent MassHunter Workstation. All preparation and analysis are completed at the State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation of the Southwest Petroleum University.

3.2. Labeling samples (factor analysis)

The supervised learning approach, including the CNN method, requires a set of known labels for the algorithm to probe the correlation between the data (independent variables) and labels (dependent variables). In this research, the labels are the biomarker features of the samples, such as thermal maturity and parent material types. However, it is challenging to find individual indicators that can sufficiently represent the biomarker features. For instance, R_o is a reliable indicator for thermal maturity, but it is only applicable to mudstone samples, which are not abundantly sampled in most commercial oil fields. Biomarker parameters that can characterize biomarkers are composite and thus cannot be directly used for machine learning.

Given the above mentioned, this research performs dimensionality reduction and combination on 15 typical biomarker parameters via the factor analysis (Table 1) (Rubinstein et al., 1975; Seifert and Moldowan, 1978, 1980, 1986; Huang and Meinschein, 1979; Sieskind et al., 1979; Moldowan et al., 1985; Connan et al., 1986; Kruge et al., 1990; Grande et al., 1993; Grice et al., 2001; Peters et al., 2007). The resultant factor scores are used as the values of the labels representing various biomarker features. The nomenclature of the compounds of interest is shown in Table 2.

The factor analysis, derived from the principal component analysis (PCA), is an important multivariate statistical analysis method, and it has been extensively applied in various disciplines including chemometrics (Park and Tauler, 2020). It groups parameters according to their variation patterns and develops a comprehensive representation (namely the factor score) of multiple parameters in the form of their linear combination. In this paper, the factor analysis is performed using the IBM SPSS Statistics according to the correlation matrix; the common factor extraction criterion is set as the eigenvalue above one; the rotation is implemented using the Varimax (max variance) with Kaiser Normalization; the factor score is obtained via regression.

3.3. Construction and processing of the dataset

3.3.1. Mass chromatogram digitalization

The mass chromatograms of terpane ($m/z = 191$) and sterane ($m/z = 217$) are the most important biomarker mass chromatograms, and most of the current classic and important biomarker parameters are built based on them. Due to the limited sample quantity, an appropriate width of the mass chromatogram fed to CNN is required to ensure the robustness and significance of statistics. The main parts of sterane ($m/z = 217$) and terpane ($m/z = 191$) mass chromatograms are separated from the whole mass chromatograms (Fig. 5), including the interval from the left side of

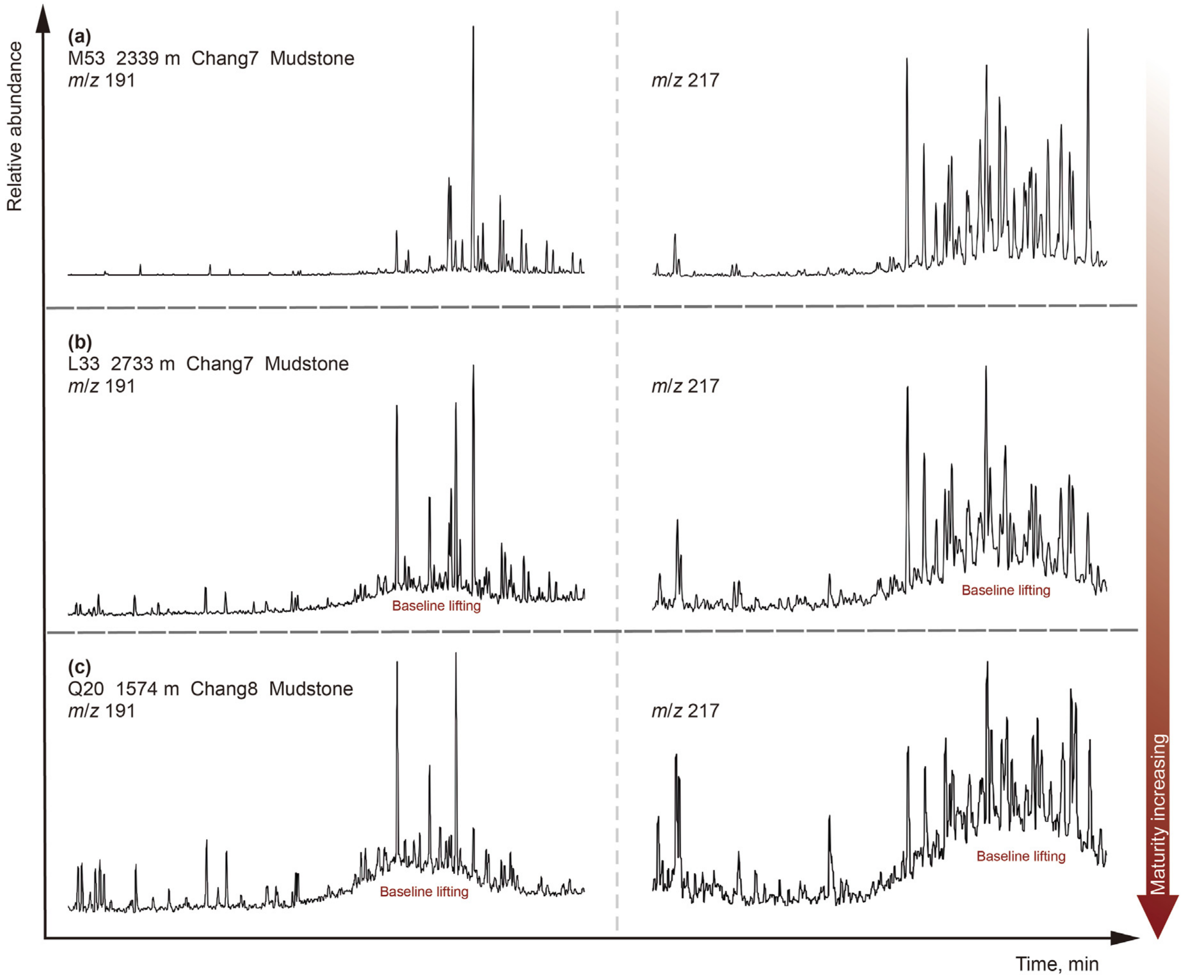


Fig. 3. The mass chromatogram contains abundant information, and yet the feature extraction based on the conventional method is limited. For instance, as the thermal maturity grows (from **(a)** to **(c)**), the baseline of the spectrum is gradually elevated and becomes increasingly irregular, besides the variations of the biomarker peaks. This is because the relevant biomarker is gradually degraded underground due to high temperature, which reduces the biomarker abundance and relatively expands the baseline morphology (namely the effects of noise signals). The conventional parameter system cannot characterize such variations, which thus cannot be compared among samples. However, with the original mass chromatogram used as the input, the CNN method can fully incorporate such features, and this helps to improve the robustness of research.

the C_{27} rearranged sterane (20S) peak to the right side of the C_{29} sterane (ααα20R) peak and the interval from the left side of the Ts peak to the right side of the C_{30}βα peak, respectively. These separated main parts are believed to preserve most biomarker information of the two types of mass chromatograms.

Because the exported data formats are different for varied models and brands of GC-MS systems, it is ideal to use universal JPG files as the input data. Although the CNN method can efficiently analyze image data, most area of the input image is empty (with no effective information), because only the curve in the mass chromatogram is the information of interest. Given this, converting the curve of the mass chromatogram into one-dimensional data can effectively reduce the proportions of ineffective and redundant data and may result in higher implementation efficiency of the algorithm.

The curve displayed in the image can be well converted into one-dimensional data by scanning pixels in a column-wise manner and recording the Y-coordinates of the curve according to a pre-specified gray threshold (Fig. 5a). Considering the widths of the used mass chromatograms, the lateral pixel count for sterane is set as 900, and that for terpane is set as 400. Correspondingly, the input data fed to CNN is a first-order tensor composed of 1300 elements (Fig. 5b).

The dataset with data and labels is constructed by correlating the first-order tensor set to the factor scores obtained in Section 3.2. Subsequently, the ranking of the dataset is sufficiently disorganized by randomly re-arranging. The first 70% of the data are used for training the CNN model, while the rest 30% are used for testing the interpretation capacity of the CNN model.

3.3.2. Normalization

Normalization of data is a necessary step before machine learning so that the differences of data in dimensions and magnitudes are eliminated to speed up the model training and improve accuracy (Rojas, 1996; Anysz et al., 2016). The *Min-Max* normalization (Eq. (1)) is performed in this paper, which converts all data into the range between zero and one.

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

where x and x' are the original and normalized data, respectively; $\min(x)$ and $\max(x)$ are the minimum and maximum of the same type of elements for the whole dataset, respectively.

The label values (i.e. the obtained factor scores) are also normalized.

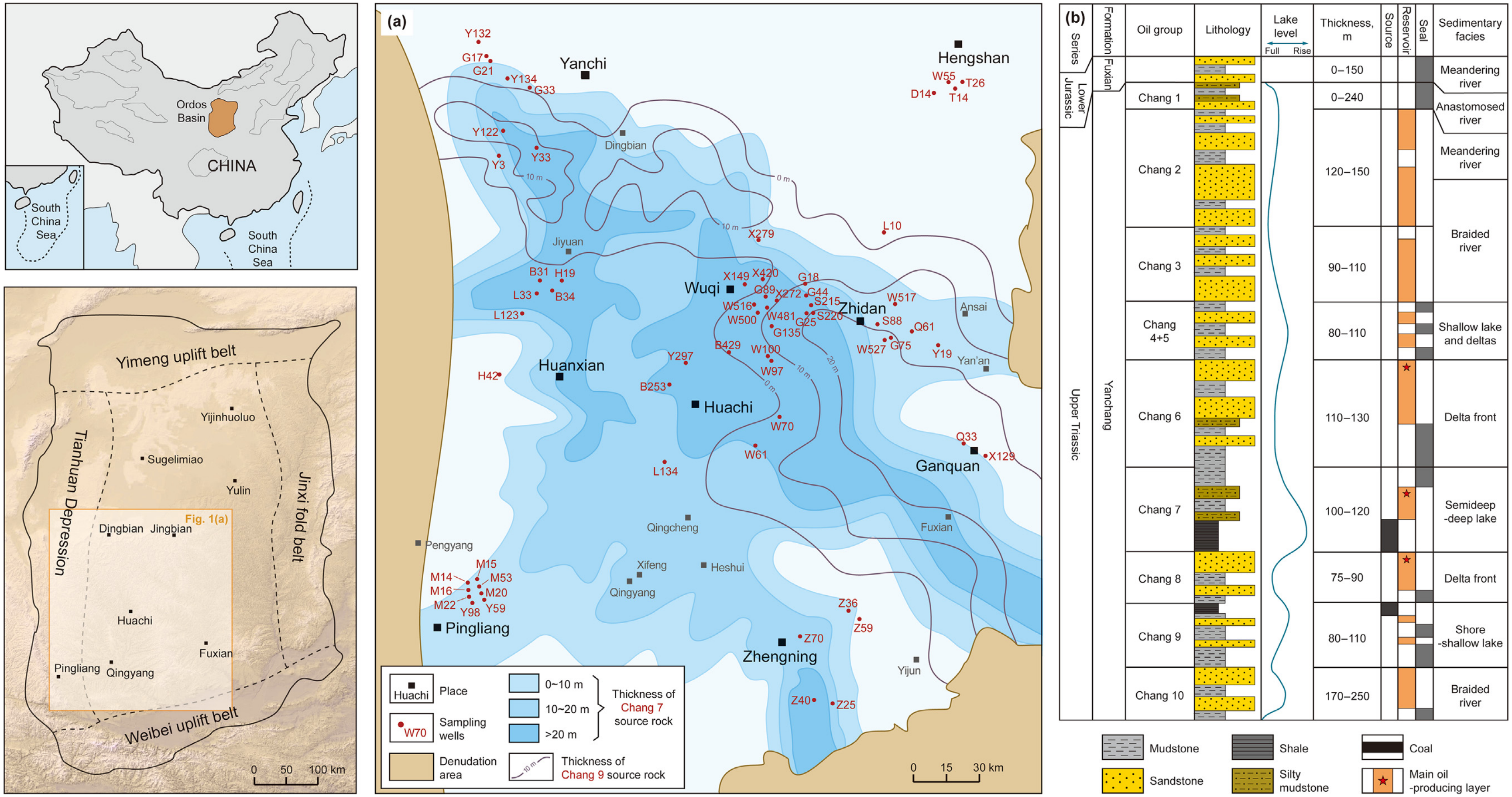


Fig. 4. (**a**) Distributions of source rocks and locations of sampling wells in the study area (modified from Li et al. (2012) and Yang et al. (2016)). (**b**) The stratigraphic column, source-reservoir-cap rock assemblage (modified from Qu et al. (2020) and Zhang et al. (2021b)).

Table 1
Biomarker parameters used in this research and their geochemical implications.

Biomarker parameters	*m*/*z*	Major geochemical implications	References
$C_{29}\alpha\beta\beta/(\alpha\alpha\alpha+\alpha\beta\beta)$	217	Maturity	Seifert and Moldowan (1986)
$C_{29}\alpha\alpha\alpha S/(S+R)$	217	Maturity	
$C_{30}\beta\alpha/C_{30}\alpha\beta$	191	Maturity	Seifert and Moldowan (1980)
Ts/(Ts+Tm)	191	Maturity	Seifert and Moldowan (1978)
C_{30}*/C_{29}Ts	191	More complex, but related to maturity	Peters et al. (2007)
C_{30}*/$C_{30}\alpha\beta$	191	More complex, but related to maturity	
C_{27}/C_{27-29} sterane	217	Parent material types	Huang and Meinschein (1979) and Moldowan et al. (1985)
C_{28}/C_{27-29} sterane	217	Parent material types	
C_{29}/C_{27-29} sterane	217	Parent material types	
Rearranged sterane/sterane	217	Clay-rich environment?	Rubinstein et al. (1975) and Sieskind et al. (1979)
C_{24}TET/$C_{30}\alpha\beta$	191	Salinity?	Connan et al. (1986) and Grice et al. (2001)
Ga/$C_{30}\alpha\beta$	191	Salinity	
ΣC_{19-26}TT/C30$\alpha\beta$	191	More complex, mainly maturity or salinity?	Kruge et al. (1990) and Grande et al. (1993)
C_{23}TT/$C_{30}\alpha\beta$	191	More complex, mainly maturity or salinity?	
$C_{29}\alpha\beta/C_{30}\alpha\beta$	191	Anoxic carbonate or marl environment	Peters et al. (2007)

Note: "?" indicates that the relevant knowledge has not been fully confirmed.

Table 2
Nomenclature of related biomarkers.

Abbreviations	Full names	Abbreviations	Full names
$C_{27}\alpha\alpha\alpha$ S	5α(H),14α(H),17α(H)—C_{27} sterane (20S)	TT	Tricyclic terpane
$C_{27}\alpha\beta\beta$ R	5α(H),14β(H),17β(H)—C_{27} sterane (20R)	TET	Tetracyclic terpane
$C_{27}\alpha\beta\beta$ S	5α(H),14β(H),17β(H)—C_{27} sterane (20S)	Ts	18α(H)—C_{27} trisnorhopane
$C_{27}\alpha\alpha\alpha$ R	5α(H),14α(H),17α(H)—C_{27} sterane (20R)	Tm	17α(H)—C_{27} trisnorhopane
$C_{28}\alpha\alpha\alpha$ S	24-methyl-5α(H),14α(H),17α(H)—C_{28} sterane (20S)	$C_{29}\alpha\beta$	17α(H), 21β(H)—C_{29} norhopane
$C_{28}\alpha\beta\beta$ R	24-methyl-5α(H),14β(H),17β(H)—C_{28} sterane (20R)	C_{29}Ts	18α(H), 21β(H)—C_{29} norneohopane
$C_{28}\alpha\beta\beta$ S	24-methyl-5α(H),14β(H),17β(H)—C_{28} sterane (20S)	C_{30}*	17α(H)—C_{30} rearranged hopane
$C_{28}\alpha\alpha\alpha$ R	24-methyl-5α(H),14α(H),17α(H)—C_{28} sterane (20R)	$C_{29}\beta\alpha$	17β(H), 21α(H)—C_{29} norhopane
$C_{29}\alpha\alpha\alpha$ S	24-ethyl-5α(H),14α(H),17α(H)—C_{29} sterane (20S)	$C_{30}\alpha\beta$	17α(H), 21β(H)—C_{30} hopane
$C_{29}\alpha\beta\beta$ R	24-ethyl-5α(H),14β(H),17β(H)—C_{29} sterane (20R)	$C_{30}\beta\alpha$	17β (H), 21α(H)—C_{30} hopane
$C_{29}\alpha\beta\beta$ S	24-ethyl-5α(H),14β(H),17β(H)—C_{29} sterane (20S)	Ga	Gammacerane
$C_{29}\alpha\alpha\alpha$ R	24-ethyl-5α(H),14α(H),17α(H)—C_{29} sterane (20R)		

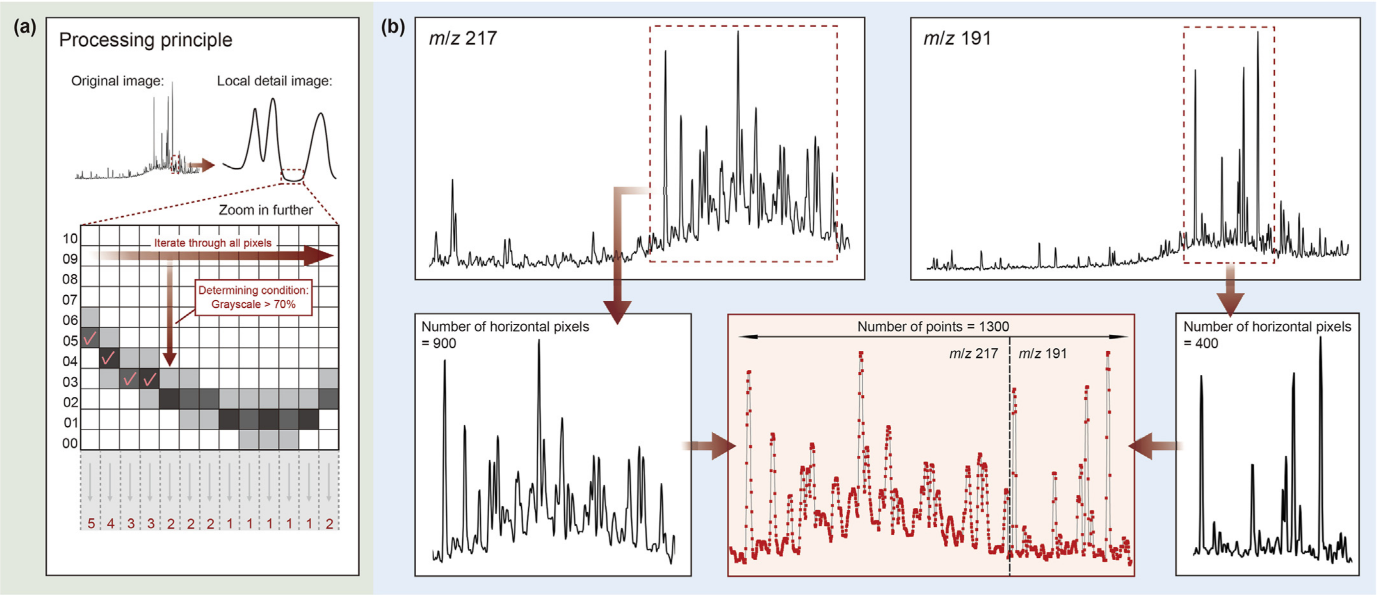


Fig. 5. **(a)** Workflow to convert the mass chromatogram into a first-order tensor. **(b)** The spectrum ranges for $m/z = 217$ and $m/z = 191$ used in this research and the final input for the CNN model. Each red dot in the image is converted into a corresponding number.

3.4. CNN architecture introduction

CNN is considered an important step forward for neural network technology (Lecun et al., 1998) and is one of the most important deep learning models. It has great capacities of feature extraction and generalization and presents the analysis accuracy even higher than those of manual works for some datasets (Niu and Suen, 2012; Lin et al., 2020). Compared with the conventional neural network,

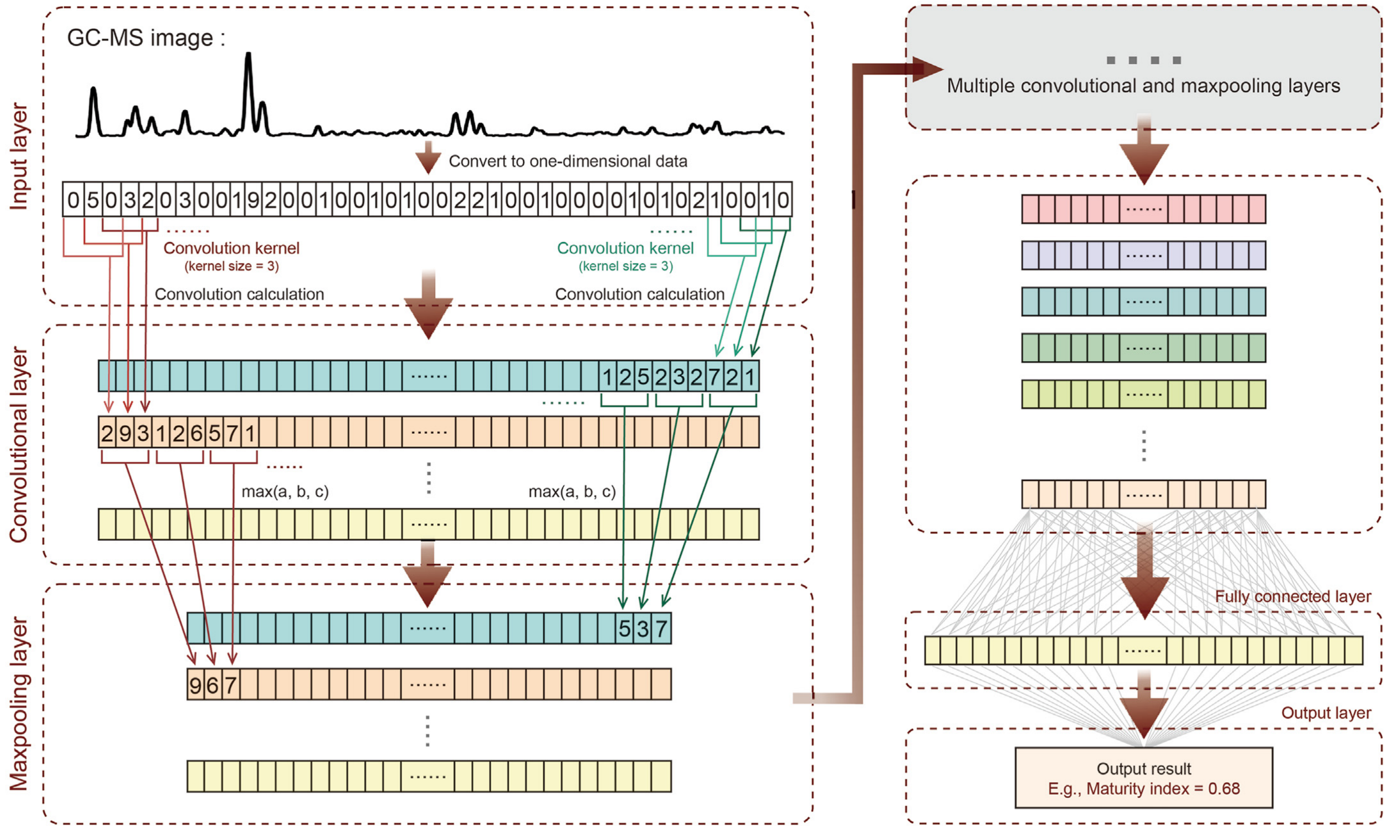


Fig. 6. The overall architecture of 1D-CNN.

the local connectivity and weight-sharing characteristics of CNN effectively reduce the model parameter quantity (Lecun et al., 1998) and improve the model trainability. Since the input data of this research is a first-order tensor, the 1D-CNN structure is adopted, which includes the input layer, convolution layer, pooling layer, fully-connected layer, and output layer (Fig. 6).

3.5. CNN training

The model of this paper is constructed based on Python (v3.7.6) and TensorFlow (v2.0.0) and the calculation is performed on a PC with Intel Xeon E5-2678 v3 CPU and NVIDIA GeForce RTX 2080 Ti GPU.

Table 3
Factor score coefficient matrix.

Serial number	Biomarker parameters	*MI*	*PMI*
1	C_{30}*/C_{29}Ts	0.195	0.041
2	C_{30}*/$C_{30}\alpha\beta$	0.139	−0.004
3	$C_{29}\alpha\beta\beta/(\alpha\alpha\alpha+\alpha\beta\beta)$	0.231	0.188
4	$C_{29}\alpha\alpha\alpha S/(S+R)$	0.273	0.036
5	$C_{30}\beta\alpha/C_{30}\alpha\beta$	−0.083	0.116
6	Ts/(Ts+Tm)	0.24	−0.029
7	C_{27}/C_{27-29} sterane	−0.038	−0.349
8	C_{28}/C_{27-29} sterane	−0.03	0.065
9	C_{29}/C_{27-29} sterane	0.074	0.376
10	Rearranged sterane/sterane	0.118	−0.214
11	$\Sigma C_{19-26}TT/C_{30}\alpha\beta$	−0.083	−0.024
12	$C_{24}TET/C_{30}\alpha\beta$	0.024	0.04
13	$Ga/C_{30}\alpha\beta$	−0.015	−0.028
14	$C_{23}TT/C_{30}\alpha\beta$	−0.075	−0.029
15	$C_{29}\alpha\beta/C_{30}\alpha\beta$	−0.16	0.136

The hyper-parameters of the neural network greatly affect the training and interpretation performances of the model (Du et al., 2021; Li et al., 2021), which typically includes the architecture complexity (the layer quantity of the model), epoch, and optimizer type. Generally, a model with a more complex network structure can fit a more complicated variation pattern, and yet is also prone to over-fitting and thus accuracy compromising. Therefore, the hyper-parameters shall be set scientifically and properly. Nonetheless, the practice is mostly empirical or based on a trial-and-error manner, since no thorough insights have been gained at present for setting hyper-parameters (Pan et al., 2010; Zhu et al., 2012). Given this, it is meaningful to test the effects of typical hyper-parameters, including layer quantity, epoch quantity, batch size, and optimizer type.

The interpretation performance is assessed using the average absolute error (*MAE*, Eq. (2)) and the coefficient of determination (R^2, Eq. (3)):

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - p_i| \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - p_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{3}$$

where m is the quantity of the output results; y_i is the actual result of the i-th sample; p_i is the corresponding interpretation; and $\bar{y}$ is the average of the actual results.

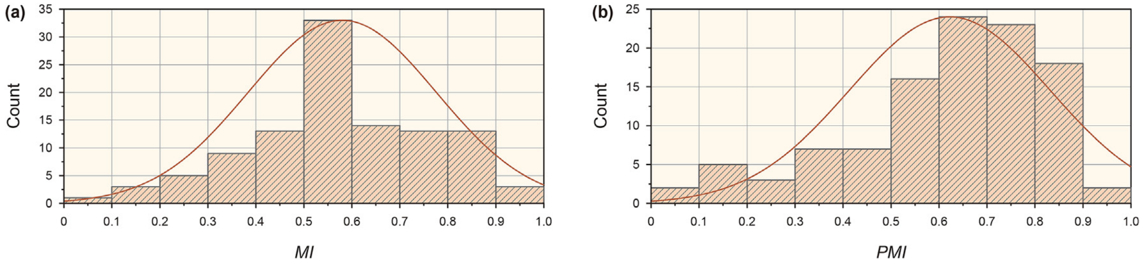


Fig. 7. Histograms of *MI* (**a**) and *PMI* (**b**) of the collected samples (after normalization).

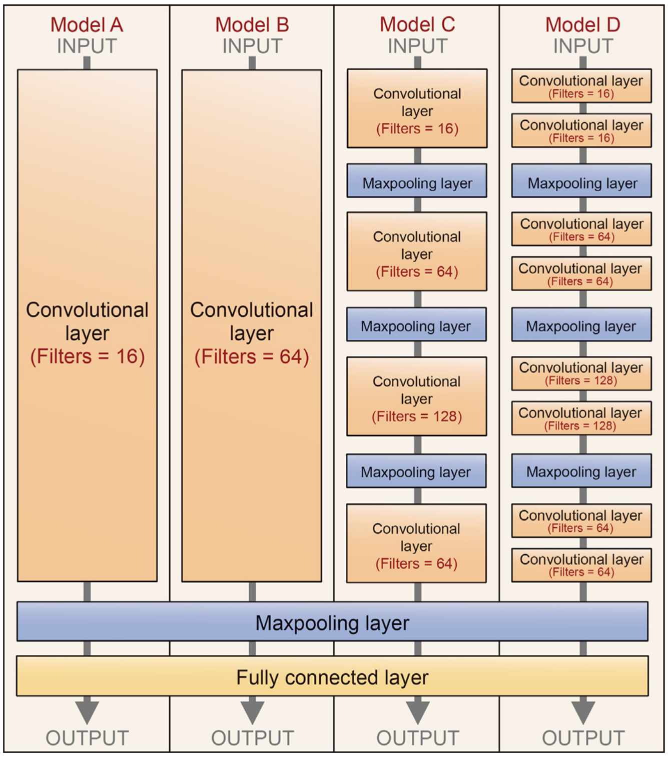

Model A
INPUT
Convolutional layer (Filters = 16)
Model B
INPUT
Convolutional layer (Filters = 64)
Model C
INPUT
Convolutional layer (Filters = 16)
Maxpooling layer
Convolutional layer (Filters = 64)
Maxpooling layer
Convolutional layer (Filters = 128)
Maxpooling layer
Convolutional layer (Filters = 64)
Model D
INPUT
Convolutional layer (Filters = 16)
Convolutional layer (Filters = 16)
Maxpooling layer
Convolutional layer (Filters = 64)
Convolutional layer (Filters = 64)
Maxpooling layer
Convolutional layer (Filters = 128)
Convolutional layer (Filters = 128)
Maxpooling layer
Convolutional layer (Filters = 64)
Convolutional layer (Filters = 64)
Maxpooling layer
Fully connected layer
OUTPUT
OUTPUT
OUTPUT
OUTPUT

Fig. 8. Four CNN models with varied architectures used for testing, and Models A, B, C, and D are found with ascending architecture complexity.

4. Results and discussion

4.1. Geochemical characteristics of the samples (based on factor analysis)

The factor analysis extracts four common factors from the original 15 biomarker parameters. In other words, the original parameters are combined to form four comprehensive parameter groups. The cumulative variance percentage stands for the ratio of the information reflected by the four common factors to that contained by the original data, and it reaches 79.751% (Table S1), which indicates satisfactory dimensionality reduction. The factor loading matrix (Table S2) shows:

(1) The original parameters related to Factor 1 are $C_{23}TT/C_{30}\alpha\beta$, $\Sigma C_{19-26}TT/C_{30}\alpha\beta$, and $C_{24}TET/C_{30}\alpha\beta$.
(2) The original parameters related to Factor 2 are $C_{29}\alpha\alpha\alpha S/(S+R)$, $Ts/(Ts+Tm)$, and $C_{30}^{*}/C_{29}Ts$.
(3) The original parameters related to Factor 3 are C_{27}/C_{27-29} sterane and C_{29}/C_{27-29} sterane.
(4) The original parameters related to Factor 4 are C_{28}/C_{27-29} sterane and $C_{30}\beta\alpha/C_{30}\alpha\beta$.

According to the current understandings of organic geochemistry (Peters et al., 2007), the original parameters related to Factor 2 are the typical parameters for thermal maturity; those to Factor 3, for parent material types (i.e. aquatic organisms or terrestrial higher plants). In contrast, it is tricky to determine the meanings of Factors 1 and 4. Factor 1 presents a positive correlation with the abundance of the tricyclic terpane, which is influenced by both

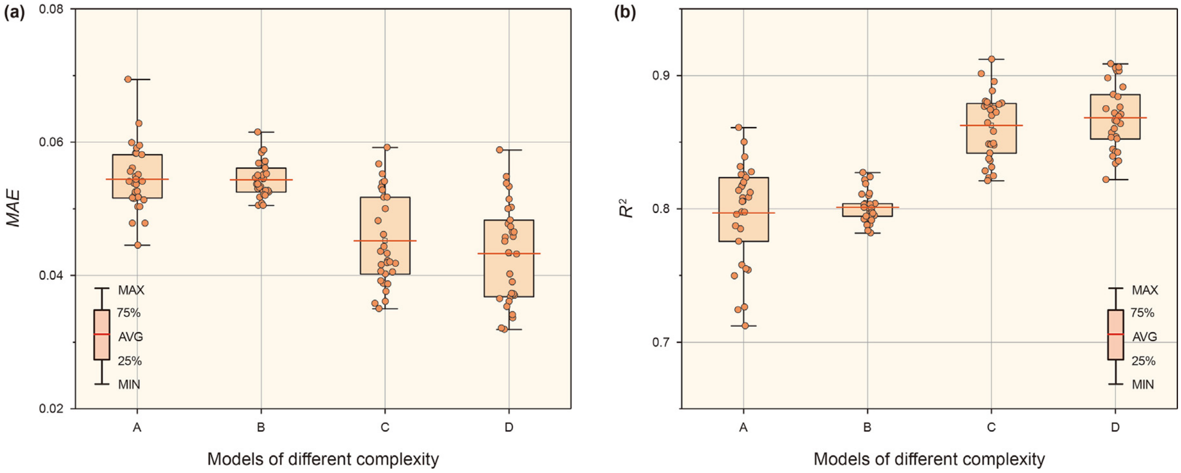


Fig. 9. Scatter-box plots showing the differences of the interpretation performances among the CNN models with varied architectures: **(a)** *MAE* and **(b)** R^2. The letters listed below the X-axis are consistent with those in Fig. 8.

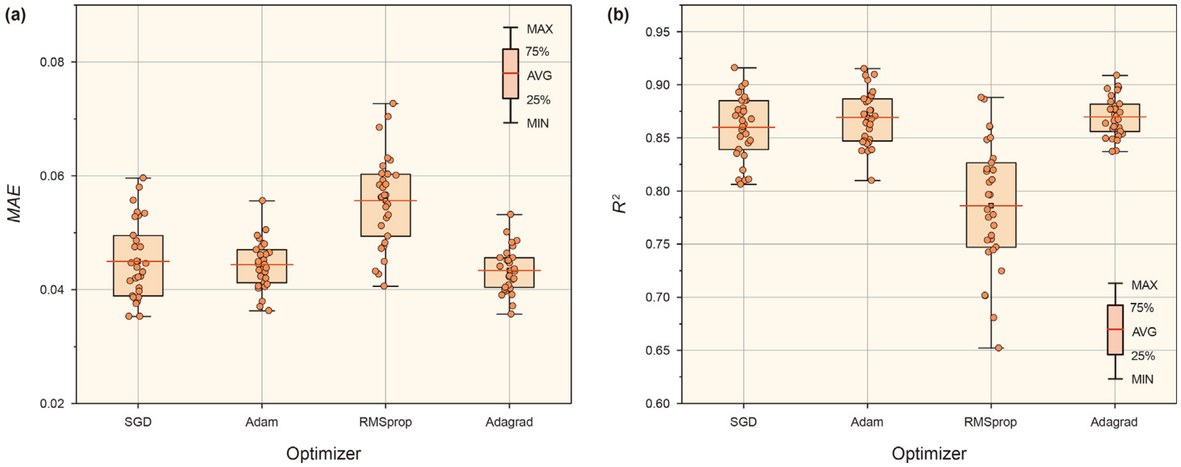


Fig. 10. Scatter-box plots showing the differences of the interpretation performances among the CNN models with varied optimizers: **(a)** *MAE* and **(b)** R^2.

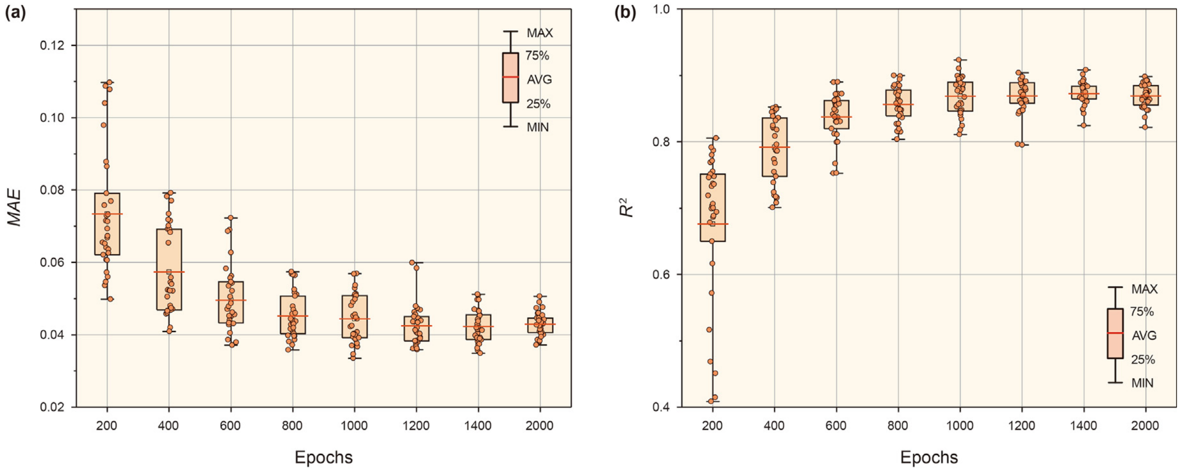


Fig. 11. Scatter-box plots showing the differences of the interpretation performances among the CNN models with varied epoch quantities: **(a)** *MAE* and **(b)** R^2.

water salinity and thermal maturity (Kruge et al., 1990; Grande et al., 1993). Thus, it is hard to assign a deterministic geochemical significance to Factor 1. As for Factor 4, unfortunately, we have no clues on its deterministic geochemical significance. Given this, we have to discard Factors 1 and 4. Nonetheless, we believe that the proposed CNN method is effective and valid if it can well capture the information on the thermal maturity and parent material type of the samples since these are important geochemical parameters for source rock evaluation.

For the convenience of discussion, scores of Factors 2 and 3 are named maturity index (*MI*, Eq. (4)) and parent material type index (*PMI*, Eq. (5)), respectively. These two indexes are used as the labels for training the CNN model, and their values are normalized into the range from zero to one. The signs of the factor loadings (Table S2) show that for *MI*, zero indicates low maturity and one, high maturity; for *PMI*, zero demonstrates the dominance of aquatic organisms, and one, the dominance of terrestrial higher plants. It should be noted that these indexes indicate relative values within the specific sample set.

$$MI = \sum_{j=1}^{n} \alpha_j x_j \tag{4}$$

$$PMI = \sum_{j=1}^{n} \beta_j x_j \tag{5}$$

where *j* is the serial number of the biomarker parameter (Table 3); *n* is the number of the biomarker parameters, which is 15 in this research; α and β are the factor score coefficients for *MI* and *PMI*, respectively (Table 3); *x* is the value of the corresponding biomarker parameter (Table S3).

The *MI* and *PMI* distributions of our sample set both follow the Gaussian distribution, and they are yet both somewhat left-skewed (Fig. 7).

4.2. Determination of the model structure and hyper-parameters

Due to the randomness of the neural network algorithm, the model may produce slightly different results for different runs (Wei et al., 2021). Taking *MI* as an example, training and interpretation of each model are all repeated 30 times, and the interpretation performance of each time is recorded. By doing so, we manage to obtain the distribution characteristics of *MAE* and R^2 of different models and optimize the CNN model structure and hyper-parameter setting.

CNN models with varied architectures (Fig. 8) present different

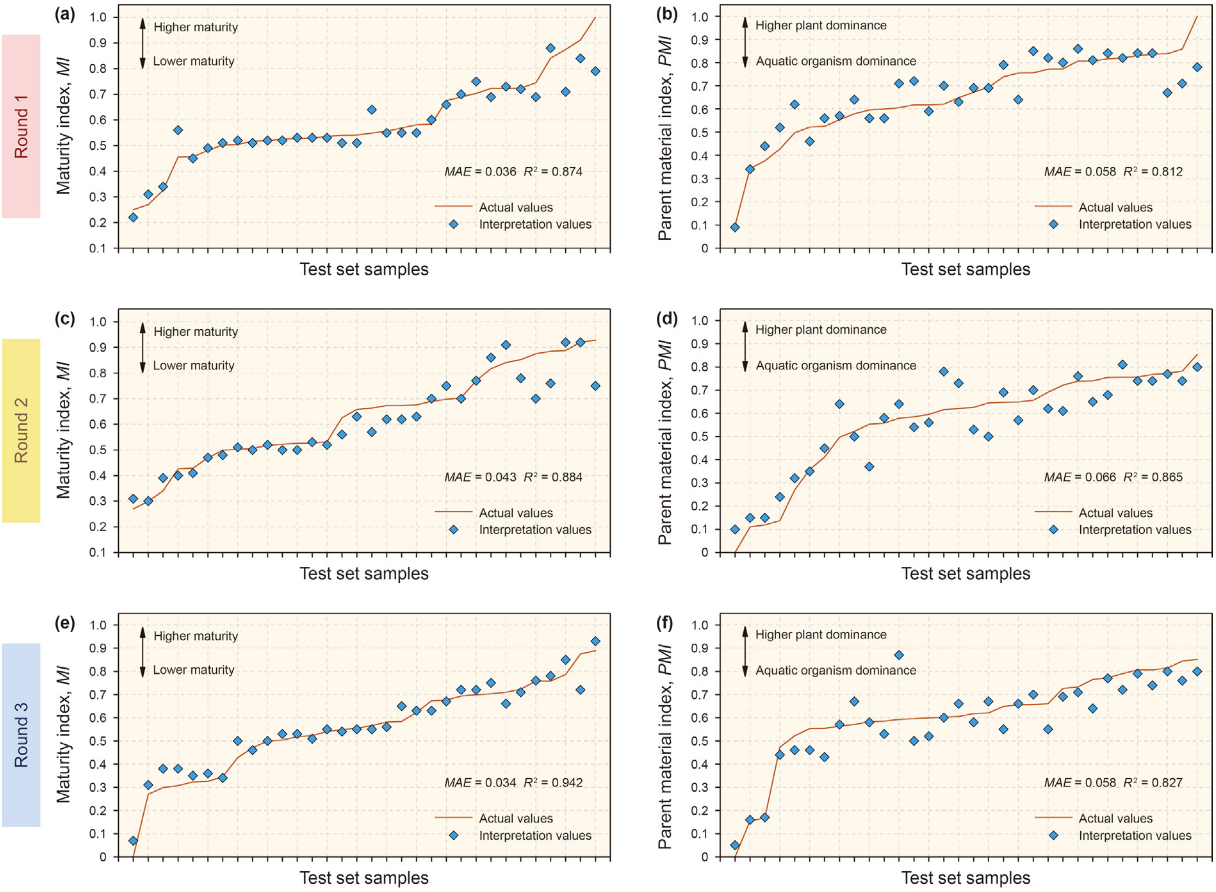


Fig. 12. Comparison between the interpreted (scattered dots in blue) and actual (broken lines in red) values of the samples in the testing set. For observation convenience, the samples are rearranged along the X-axis, with the ascending Y values. The results of three rounds of training and interpretation are shown in subfigures (**a**)–(**f**), respectively. The data are sufficiently shuffled for each round to avoid discrepancies in the interpretation performance caused by the potential differences between the training and testing sets.

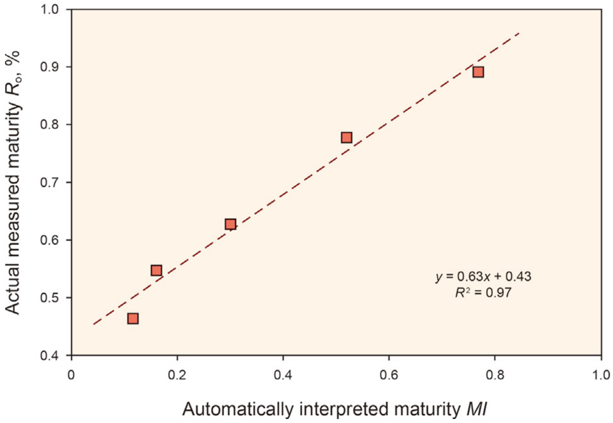


Fig. 13. Since the automatically interpreted maturity (*MI*) and the actual analysis maturity (R_o, %) appear to be correlated, it is possible to determine the R_o of unidentified samples from the *MI*.

interpretation performances (Fig. 9). Generally, the multi-layer models C and D present the interpretation accuracy considerably higher than those of the single-layer models A and B. Moreover, properly expanding the size of the model (i.e. the kernel quantity) can improve the interpretation robustness, which is manifested as the considerably narrowed distribution ranges of *MAE* and R^2 as indicated by the comparison between Models A and B. However, this has no notable contributions to improving interpretation accuracy. Given the above analysis, a multi-layer CNN model is preferred for this research.

The various optimizers for CNN models tested in this research can all deliver relatively satisfactory interpretation performances, except the RMSprop optimizer (Fig. 10). However when the computation time is nearly equal to that of other optimizers, the Adagrad optimizer (Duchi et al., 2011) can produce the best interpretation results (in model D, SGD is 13 s, Adam is 12 s, RMSprop is 13 s, Adagrad is 13 s).

The interpretation performance of the CNN model varies with the epoch quantity (Fig. 11), and the relatively good interpretation performance occurs with 1000 epochs (R^2 averages 0.87). With more epochs, the interpretation is not further improved, and instead, R^2 gradually declines, which in most cases suggests overfitting (Chuang et al., 2000).

To sum up, the preferred structure and hyper-parameters of the CNN model are determined: the multi-layer CNN structure (i.e. Model D in Fig. 8), the Adagrad optimizer, 1000 epochs and batch size of 32.

4.3. Automated analysis performance and error analysis

With the above-determined model structure and hyper-parameters, the CNN model is trained using the mass chromatograms of sterane ($m/z = 217$) and terpane ($m/z = 191$) of the 76 samples in the training set. Then, the data of 32 samples are used for performance presentation and model validation.

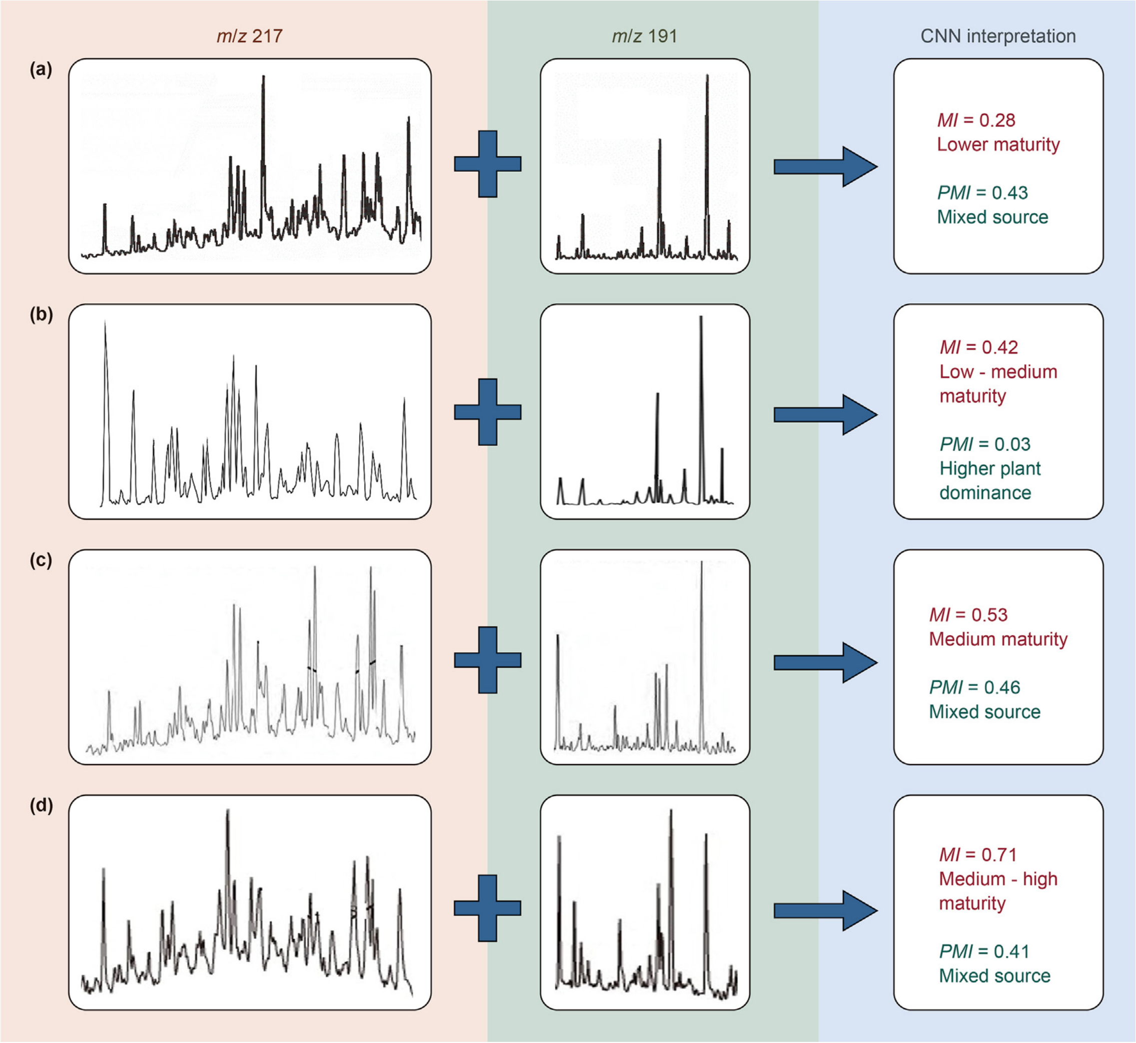


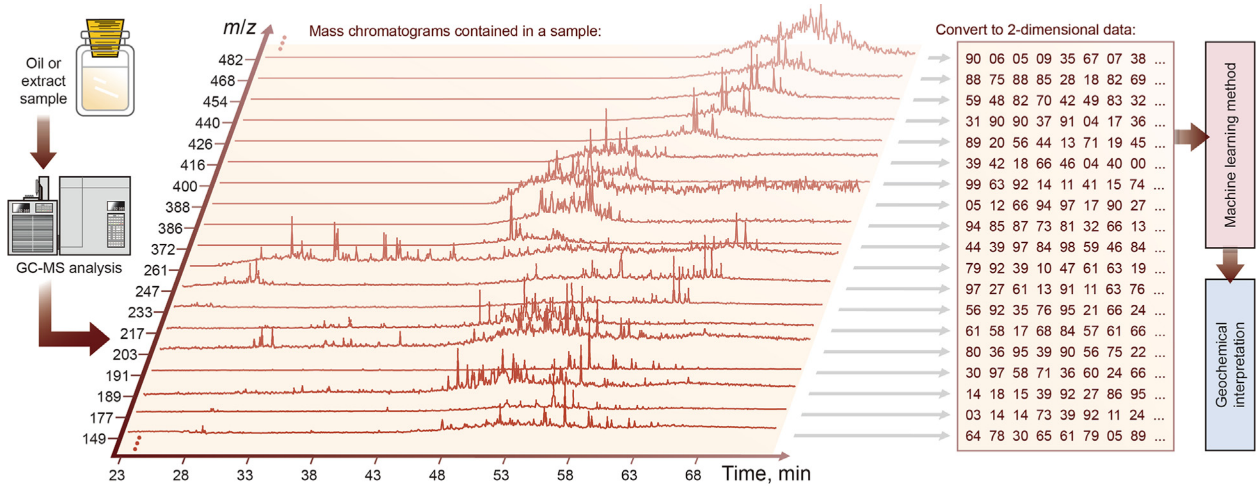
Fig. 14. Automated interpretation results using the proposed CNN model for the publicly-reported mass chromatograms. The corresponding sample attributes are **(a)** Hilba C-1 Source rock 2760–2795 m (Xiao et al., 2019); **(b)** Fahdene Fm. (Hallek and Montacer, 2021); **(c)** Chang 6–Chang 10 members of the Yanchang Fm. (Bai et al., 2013a); **(d)** Well Feng189 (Bai et al., 2013b). No modification is made to the images, except for eliminating the markers added by their authors. It is seen that low image quality does not affect the analysis of the CNN model. It should be noted that the results produced by the CNN method are relative values, applicable to the used specific dataset. Therefore, to produce the results of globally universal value needs to build a dataset with a gargantuan volume on a global basis, which obviously requires joint efforts of researchers from all over the world.

The differences between the interpreted and actual scores are shown in Fig. 12. The results for three rounds of training and interpretation are presented to avoid the interpretation performance discrepancy attributed to inappropriately dividing the samples into the training and testing sets. The data are shuffled in each round. No considerable discrepancies are found among the results of the three rounds (the distributions of *MAE* and R^2 values are consistent with each other) and therefore the interpretation performance of the model is robust.

In terms of the two indexes, the interpretation performance of the model for *MI* is apparently better than that of *PMI*. Although their R^2 values are both above 0.8, the distributions of the interpreted *PMI* values (Fig. 12b, d, f) are visually more irregular than those of the *MI* interpretation results (Fig. 12a, c, e). We believe that this is mainly attributed to the labels of the dataset. According to the factor analysis result (Table S1), *MI* is dependent on six biomarker parameters (e.g. sterane $C_{29}\alpha\alpha\alpha S/(S+R)$), with a variance contribution of 23.4%; *PMI* is related to three parameters, with a variance contribution of 18.5%. These suggest that *PMI* is intrinsically prone to more errors than *MI*, and thus it is natural that the CNN model delivers an interpretation performance of *PMI* that is inferior to that of *MI*. Consequently, the future study shall focus on increasing the number of samples and introducing extra dimensions into data analysis.

When the findings of automatic interpretation are compared to the actual measured maturity (using R_o as a reference), it appears that there is a correlation between the two, which indicates the results of automatic interpretation have excellent accuracy (Fig. 13). In conclusion, the CNN model shows significant promise for geochemical interpretation of GC-MS data, and it is anticipated that increasing the amount and quality of data sets would further improve its performance. We also apply the proposed model to interpreting the publicly-reported mass chromatograms (Bai et al., 2013a, 2013b; Xiao et al., 2019; Hallek and Montacer, 2021), and the results are shown in Fig. 14. The CNN-based automated biomarker interpretation technique and software are expected to be soon comprehensively applied in the research of petroleum geology (a demo software is designed for this research, see the Supplementary Video).

In the long run, the significance of this method is not only



```
90 06 05 09 35 67 07 38 ...
88 75 88 85 28 18 82 69 ...
59 48 82 70 42 49 83 32 ...
31 90 90 37 91 04 17 36 ...
89 20 56 44 13 71 19 45 ...
39 42 18 66 46 04 40 00 ...
99 63 92 14 11 41 15 74 ...
05 12 66 94 97 17 90 27 ...
94 85 87 73 81 32 66 13 ...
44 39 97 84 98 59 46 84 ...
79 92 39 10 47 61 63 19 ...
97 27 61 13 91 11 63 76 ...
56 92 35 76 95 21 66 24 ...
61 58 17 68 84 57 61 66 ...
80 36 95 39 90 56 75 22 ...
30 97 58 71 36 60 24 66 ...
14 18 15 39 92 27 86 95 ...
03 14 14 73 39 92 11 24 ...
64 78 30 65 61 79 05 89 ...
```

Fig. 15. The technical principles of the "full-compounds oil-source correlation", which is the development orientation of our proposed method. The mass chromatogram curves corresponding to multiple *m*/*z* values for a single oil sample are converted into a two-dimensional dataset and fed into the CNN model for deep learning. Theoretically, with sufficient mass chromatogram curves (importing the mass chromatograms at an identical interval of the *m*/*z* value), the converted two-dimensional dataset can fully represent the features of all compounds in the oil sample. Therefore, it is referred to as the "full-compounds correlation". However, this goal requires extremely high-quality data.

providing convenience for relevant studies but also enabling more in-depth analysis of data from optimized points of view (depths and dimensions). Due to limited data availability, this research only uses the mass chromatograms of sterane ($m/z = 217$) and terpane ($m/z = 191$), which should be viewed as a pilot study that can be followed by much more relevant and detailed research. The future study obviously can utilize the mass chromatograms for more m/z values. The mass chromatogram for each m/z will be represented by a row of data, and thus the curves for various m/z values will be integrated into the form of data in different rows. By doing so, mass chromatograms of each sample can be converted into two-dimensional data that can be learned by the CNN model (Fig. 15). Theoretically, if the number of mass chromatographic curves are sufficiently large (e.g. importing the mass chromatograms at a constant m/z interval), the converted two-dimensional data will be able to fully represent the characteristics of all compounds in the oil sample, and the full-compounds oil-source correlation will be realized, which will have profound effects upon the organic geochemistry.

5. Conclusions

Conventional GC-MS methods are time-consuming, labor-intensive, and more importantly, of generally poor performance. They fail to efficiently and comprehensively extract the information from the mass chromatogram and thus many meaningful features cannot be interpreted, which to some extent restrains the development of the theory and technique of organic geochemistry.

Given this, this research makes tentative efforts to apply the convolution neural network (CNN) to link the original mass chromatogram to the biomarker feature. The mass chromatograms of the samples collected from the Triassic Yanchang Formation in the Ordos Basin are used to build the dataset, accompanied by the automated interpretation of the thermal maturity and parent material type features of organic matter.

The significance of this work is not only enabling automatic interpretation of the mass chromatogram and providing convenience for relevant studies but also presenting a preliminary attempt for applications of artificial intelligence to organic geochemistry. The developed method is of tremendous commercial value and is expected to have major effects upon the oil-source correlation and favorable target prediction based on geochemical data.

Future research will be significantly facilitated by increasing the quantity and quality of the data.

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.

Acknowledgments

This research was financially supported by China Postdoctoral Science Foundation (Grant No. 2023M730365), Natural Science Foundation of Hubei Province of China (Grant No. 2023AFB232).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petsci.2023.11.010.

References

Anysz, H., Zbiciak, A., Ibadov, N., 2016. The influence of input data standardization method on prediction accuracy of artificial neural networks. Pro. Eng. 153, 66–70. https://doi.org/10.1016/j.proeng.2016.08.081.

Bai, Y.B., Luo, J.L., Hao, X.R., Wang, L., Chen, X.P., Zhang, A.C., 2013a. Geochemical characteristics of source rocks of Yanchang Formation in panlong oilfield,Ordos Basin. Geol. Sci. Technol. Inf. 32, 19–24 (in Chinese).

Bai, Y.B., Luo, J.L., Liu, X.J., Jin, W.Q., Wang, X.J., 2013b. Geochemical characteristics of crude oil and oil-source correlation in Yanchang Formation (upper triassic) in Wubao area, Ordos Basin. Acta Sedimentol. Sin. 31, 374–383 (in Chinese).

Bergen, K.J., Johnson, P.A., Hoop, M.V.D., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363. https://doi.org/10.1126/science.aau0323 eaau0323.

Chuang, C.C., Su, S.F., Hsiao, C.C., 2000. The annealing robust backpropagation (ARBP) learning algorithm. IEEE Trans. Neural Network. 11, 1067–1077. https://doi.org/10.1109/72.870040.

Connan, J., Bouroullec, J., Dessort, D., Albrecht, P., 1986. The microbial input in carbonate-anhydrite facies of a sabkha palaeoenvironment from Guatemala: a molecular approach. Org. Geochem. 10, 29–50. https://doi.org/10.1016/0146-6380(86)90007-0.

Du, X., Xu, H., Zhu, F., 2021. Understanding the effect of hyperparameter optimization on machine learning models for structure design problems. Comput. Aided Des. 135, 103013. https://doi.org/10.1016/j.cad.2021.103013.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12, 2121–2159.

Grande, S.M.B.D., Neto, F.R.A., Mello, M.R., 1993. Extended tricyclic terpanes in sediments and petroleums. Org. Geochem. 20, 1039–1047. https://doi.org/10.1016/0146-6380(93)90112-O.

Grice, K., Audino, M., Boreham, C.J., Alexander, R., Kagi, R.I., 2001. Distributions and stable carbon isotopic compositions of biomarkers in torbanites from different palaeogeographical locations. Org. Geochem. 32, 1195–1210. https://doi.org/10.1016/S0146-6380(01)00087-0.

Hallek, T., Montacer, M., 2021. Occurrences and origin of oil seeps and new marks of petroleum impregnations in Northwestern Tunisia: implications from aliphatic biomarkers and statistical modelling. J. Afr. Earth Sci. 182, 104278. https://doi.org/10.1016/j.jafrearsci.2021.104278.

Ho, T.L., 2009. 3-D inversion of borehole-to-surface electrical data using a back-propagation neural network. J. Appl. Geophys. 68, 489–499. https://doi.org/10.1016/j.jappgeo.2008.06.002.

Huang, W.Y., Meinschein, W.G., 1979. Sterols as ecological indicators. Geochem. Cosmochim. Acta 43, 739–745. https://doi.org/10.1016/0016-7037(79)90257-6.

Isaksen, G.H., Bohacs, K.M., 1995. Geological controls on source rock geochemistry through relative sea level, Triassic, Barents Sea. In: Katz, B.J. (Ed.), Petroleum Source Rocks. Springer - Verlag, New York, pp. 25–50.

Kaufman, R.L., Ahmed, A.S., Elsinger, R.J., 1990. Gas chromatography as a development and production tool for fingerprinting oils from individual reservoirs: applications in the Gulf of Mexico. In: Schumacher, D., Perkins, B.F. (Eds.), Proceedings of the 9th Annual Research Conference of the Society of Economic Paleontologists and Mineralogists. Society of Paleontologists and Mineralogists, Tulsa, pp. 263–282.

Koeshidayatullah, A., Morsilli, M., Lehrmann, D.J., Al-Ramadan, K., Payne, J.L., 2020. Fully automated carbonate petrography using deep convolutional neural networks. Mar. Petrol. Geol. 122, 104687. https://doi.org/10.1016/j.marpetgeo.2020.104687.

Kruge, M.A., Hubert, J.F., Akes, R.J., Meriney, P.E., 1990. Biological markers in Lower Jurassic synrift lacustrine black shales, Hartford basin, Connecticut. U.S.A. Org. Geochem. 15, 281–289. https://doi.org/10.1016/0146-6380(90)90006-L.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86. https://doi.org/10.1109/5.726791.

Li, H., Li, X.H., Yuan, F., Jowitt, S.M., Zhang, M.M., Zhou, J., Zhou, T.F., Li, X.L., Ge, C., Wu, B.C., 2020a. Convolutional neural network and transfer learning based mineral prospectivity modeling for geochemical exploration of Au mineralization within the Guandian–Zhangbaling area, Anhui Province, China. Appl. Geochem. 122, 104747. https://doi.org/10.1016/j.apgeochem.2020.104747.

Li, J.J., Wu, H., Lu, S.F., Xue, H.T., Huang, Z.K., Wang, K., Shi, L., Wang, X.F., 2012. Development and hydrocarbon expulsion efficiency of source rock in 9th member of Yanchang Formation,Ordos Basin. J. Jilin Unv. (Earth Science Edition) 42, 26–32 (in Chinese).

Li, Q., Wu, S.H., Xia, D.L., You, X.L., Zhang, H.M., Lu, H., 2020b. Major and trace element geochemistry of the lacustrine organic-rich shales from the Upper Triassic Chang 7 Member in the southwestern Ordos Basin, China: implications for paleoenvironment and organic matter accumulation. Mar. Petrol. Geol. 111, 852–867. https://doi.org/10.1016/j.marpetgeo.2019.09.003.

Li, W., Ng, W.W.Y., Wang, T., Pelillo, M., Kwong, S., 2021. HELP: an LSTM-based approach to hyperparameter exploration in neural network learning. Neurocomputing 442, 161–172. https://doi.org/10.1016/j.neucom.2020.12.133.

Lin, D.P., Abbas, N.M., 1990. Use of a GC/MS/MS technique in determination of biomarkers in regional petroleum. Talanta 37, 731–734. https://doi.org/10.1016/0039-9140(90)80102-L.

Lin, J.D., Wu, X.Y., Chai, Y., Yin, H.P., 2020. Structure optimization of convolutional neural networks: a survey. Acta Autom. Sin. 46, 24–37. https://doi.org/10.16383/j.aas.c180275 (in Chinese).

Liu, X.P., Zhao, H.T., Yan, X.X., Jia, Y.N., 2019. The geological characteristics of tight sandstone gas and exploration target evaluation in the craton basin: case study of the Upper Palaeozoic of Ordos Basin. Nat. Gas Geosci. 30, 331–343. https://doi.org/10.11764/j.issn.1672-1926.2018.12.015 (in Chinese).

Moldowan, J.M., Seifert, W.K., Gallegos, E.J., 1985. Relationship between petroleum composition and depositional environment of petroleum source rocks. AAPG Bull. 69, 1255–1268.

Niu, X.X., Suen, C.Y., 2012. A novel hybrid CNN–SVM classifier for recognizing handwritten digits. Pattern Recogn. 45, 1318–1325. https://doi.org/10.1016/j.patcog.2011.09.021.

Pan, F., Zhu, P., Zhang, Y., 2010. Metamodel-based lightweight design of B-pillar with TWB structure via support vector regression. Comput. Struct. 88, 36–44. https://doi.org/10.1016/j.compstruc.2009.07.008.

Pan, S., Horsfield, B., Zou, C., Yang, Z., Gao, D., 2017. Statistical analysis as a tool for assisting geochemical interpretation of the upper triassic Yanchang Formation, Ordos Basin, Central China. Int. J. Coal Geol. 173, 51–64. https://doi.org/10.1016/j.coal.2017.02.009.

Park, E.S., Tauler, R., 2020. 2.17 - bayesian methods for factor analysis in chemometrics. In: Brown, S., Tauler, R., Walczak, B. (Eds.), Comprehensive Chemometrics, second ed. Elsevier, pp. 355–369.

Peters, K.E., Walters, C.C., Moldowan, J.M., 2007. The Biomarker Guide: Volume 2, Biomarkers and Isotopes in Petroleum Systems and Earth History, 2 ed. Cambridge University Press, New York.

Qu, H.J., Yang, B., Gao, S.L., Zhao, J.F., Han, X., Chen, S., Hayat, K., 2020. Controls on hydrocarbon accumulation by facies and fluid potential in large-scale lacustrine petroliferous basins in compressional settings: a case study of the Mesozoic Ordos Basin, China. Mar. Petrol. Geol. 122, 104668. https://doi.org/10.1016/j.marpetgeo.2020.104668.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Rojas, R., 1996. Neural Networks: A Systematic Introduction. Springer, Berlin.

Rubinstein, I., Sieskind, O., Albrecht, P., 1975. Rearranged sterenes in a shale: occurrence and simulated formation. J. Chem. Soc., Perkin Trans. 1 1, 1833–1836. https://doi.org/10.1039/P19750001833.

Seifert, W.K., Moldowan, J.M., 1978. Applications of steranes, terpanes and monoaromatics to the maturation, migration and source of crude oils. Geochem. Cosmochim. Acta 42, 77–95. https://doi.org/10.1016/0016-7037(78)90219-3.

Seifert, W.K., Moldowan, J.M., 1980. The effect of thermal stress on source-rock quality as measured by hopane stereochemistry. Phys. Chem. Earth 12, 229–237. https://doi.org/10.1016/0079-1946(79)90107-1.

Seifert, W.K., Moldowan, J.M., 1986. Use of biological markers in petroleum exploration. In: Johns, R.B. (Ed.), Methods in Geochemistry and Geophysics. Elsevier, Amsterdam, pp. 261–290.

Sieskind, O., Joly, G., Albrecht, P., 1979. Simulation of the geochemical transformations of sterols: superacid effect of clay minerals. Geochem. Cosmochim. Acta 43, 1675–1679. https://doi.org/10.1016/0016-7037(79)90186-8.

Wei, X., Zhang, L.L., Yang, H.Q., Zhang, L.M., Yao, Y.P., 2021. Machine learning for pore-water pressure time-series prediction: application of recurrent neural networks. Geosci. Front. 12, 453–467. https://doi.org/10.1016/j.gsf.2020.04.011.

Xiao, H., Li, M.J., Liu, J.G., Mao, F.J., Cheng, D.S., Yang, Z., 2019. Oil-oil and oil-source rock correlations in the Muglad Basin, Sudan and South Sudan: new insights from molecular markers analyses. Mar. Petrol. Geol. 103, 351–365. https://doi.org/10.1016/j.marpetgeo.2019.03.004.

Yang, H., 2004. Deposition System and Oil Accumulation Research of Yanchang Formation in Triassic, Ordos Basin. Chengdu University of Technology, Chengdu.

Yang, H., Niu, X.B., Xu, L.M., Feng, S.B., You, Y., Liang, X.W., Wang, F., Zhang, D.D., 2016. Exploration potential of shale oil in Chang7 member, upper triassic Yanchang Formation, Ordos Basin, NW China. Petrol. Explor. Dev. 43, 560–569. https://doi.org/10.1016/S1876-3804(16)30066-0.

Yang, H., Xi, S.L., Wei, X.S., Li, Z.H., 2006. Evolution and natural gas enrichment of multicycle superimposed basin in Ordos. China petrol. explor. 11, 17–24. https://doi.org/10.3969/j.issn.1672-7703.2006.01.004 (in Chinese).

Yang, Y.N., Zhou, S.X., Li, J., Li, C.C., Li, Y.J., Ma, Y., Chen, K.F., 2017. Geochemical characteristics of source rocks and oil-sourcecorrelation of Yanchang Formation in southern Ordos Basin, China. Nat. Gas Geosci. 28, 550–565. https://doi.org/10.11764/j.issn.1672-1926.2017.02.012 (in Chinese).

Yu, H., Wang, Z., Rezaee, R., Zhang, Y., Nwidee, L.N., Liu, X., Verrall, M., Iglauer, S., 2020. Formation water geochemistry for carbonate reservoirs in Ordos basin, China: implications for hydrocarbon preservation by machine learning. J. Petrol. Sci. Eng. 185, 106673. https://doi.org/10.1016/j.petrol.2019.106673.

Zhang, C.J., Zuo, R.G., Xiong, Y.H., 2021a. Detection of the multivariate geochemical anomalies associated with mineralization using a deep convolutional neural network and a pixel-pair feature method. Appl. Geochem. 130, 104994. https://doi.org/10.1016/j.apgeochem.2021.104994.

Zhang, K., Liu, R., Liu, Z.J., 2021b. Sedimentary sequence evolution and organic matter accumulation characteristics of the Chang 8–chang 7 members in the upper triassic Yanchang Formation, southwest Ordos Basin, central China. J. Petrol. Sci. Eng. 196, 107751. https://doi.org/10.1016/j.petrol.2020.107751.

Zhu, P., Pan, F., Chen, W., Zhang, S., 2012. Use of support vector regression in structural optimization: application to vehicle crashworthiness design. Math. Comput. Simulat. 86, 21–31. https://doi.org/10.1016/j.matcom.2011.11.008.

Zou, C.N., 2014. Unconventional Petroleum Geology, 1 ed. Geological Publishing House, Beijing.

Zou, X.L., Chen, S.J., Lu, J.G., Zhang, H., Wang, L., Zhou, S.Y., 2017. Composition and distribution of 17α(H)-diahopane in the Yanchang Formation source rocks, Ordos Basin. Geochimica 46, 252–261. https://doi.org/10.3969/j.issn.0379-1726.2017.03.005 (in Chinese).

Zumberge, J.E., 1987. Prediction of source rock characteristics based on terpane biomarkers in crude oils: a multivariate statistical approach. Geochem. Cosmochim. Acta 51, 1625–1637. https://doi.org/10.1016/0016-7037(87)90343-7.