energy&fuels

pubs.acs.org/EF

Article

# Predicting the Molecular Models, Types, and Maturity of Kerogen in Shale Using Machine Learning and Multi-NMR Spectra

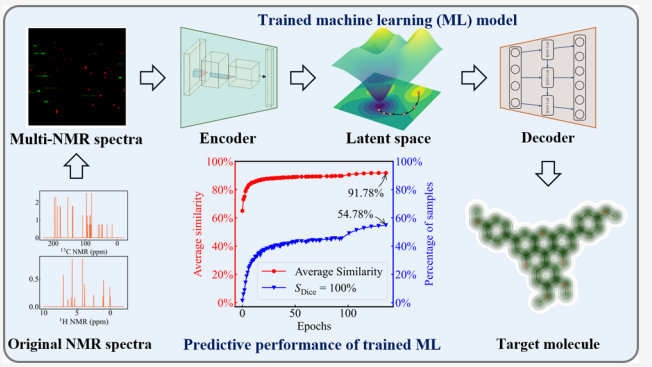
Dongliang Kang and Ya-Pu Zhao*

Cite This: *Energy Fuels* 2022, 36, 5749−5761

Read Online

ACCESS | Metrics & More | Article Recommendations

**ABSTRACT:** Kerogen is the primary hydrocarbon source of shale oil/gas. The kerogen types and maturity are the two most crucial indicators that can reflect the hydrocarbon generation potential of shale oil/gas reservoirs. These indicators and the other mechanochemical properties can be effectively studied in a bottom-up strategy using kerogen molecular models. Thus, the rapid construction of kerogen molecular models is the cornerstone of shale oil/gas exploitation research. Because of the combinatorial explosion problem, there are two inherent disadvantages of traditional methods: being time- and material-consuming and labor-intensive. We propose a new method that combines machine learning with multiple nuclear magnetic resonance spectra to intelligently and with a high throughput predict the kerogen structures, types, and maturity. Neither the manual analysis of experimental spectra nor the enormous trial-and-error process is required in our method. The 650,000 groups of samples are annotated as the sample datasets. Various spectral types can be analyzed comprehensively using the multi-spectral form, and the predictive capability beyond that of the single input form is obtained. The results demonstrate that the average similarity of prediction molecules and the targets is 91.78%. The prediction accuracy of kerogen components, types, and maturity indexes is better than 92.4%, and the coefficients of determination $R^2$ are all over 0.934. The results exhibit the excellent comprehensive performance and effectiveness of our method. Thus, we anticipate that this work will shorten the research cycle and tremendously reduce costs in constructing kerogen models and predicting kerogen properties.




## 1. INTRODUCTION

Kerogen is the most abundant source of organic compounds on earth and the primary hydrocarbon source of shale oil/gas.[1−4] The two most fundamental indicators of kerogen: type and maturity, can be directly characterized by using the information on the kerogen molecular structure.[5−9] The kerogen types are used to indicate the origin and the hydrocarbon generation type of kerogen. Generally, according to the distribution of the kerogen component H/C and O/C (H for hydrogen, O for oxygen, and C for carbon) atomic ratio in the van Krevelen diagram, kerogen can be divided into three types: Type I (oil-prone), Type II (oil- and gas-prone), and Type III (gas-prone).[10] The maturity is to represent the predominant hydrocarbon potential of kerogen. The maturity indicator, such as vitrinite reflectance %Ro, the molecule maturity index (MMI), and the orbital hybridization maturity index (OrbHMI), also can be evaluated from the kerogen molecular models.[11−14] Furthermore, kerogen models are the cornerstone of exploring the mechanism of adsorption/desorption, maturity evolution, pyrolysis behavior, and generation of oil/gas through molecular simulation. Therefore, the high-efficiency and high-quality construction of the kerogen structural models is the bridge of shale oil/gas exploitation.[15−25]

Traditionally, the construction of the kerogen models is based on experiments, such as nuclear magnetic resonance (NMR), X-ray diffraction (XRD), X-ray photoelectron spectroscopy (XPS), and pyrolysis gas chromatography and mass spectrometry to approximate the natural structures. Initially, only the main functional groups and the skeleton can be determined roughly. Subsequently, the complete kerogen molecular models begin to be constructed through a combined analysis of various experiments.[26−29] The kerogen models at different evolution stages and in different mining areas are established.[30] Although the models are still displayed in the two-dimensional (2D) form, the influence of the three-dimensional (3D) configuration is considered.[31] Finally, with

Received: March 14, 2022
Revised: April 20, 2022
Published: May 16, 2022






https://doi.org/10.1021/acs.energyfuels.2c00738
*Energy Fuels* 2022, 36, 5749−5761

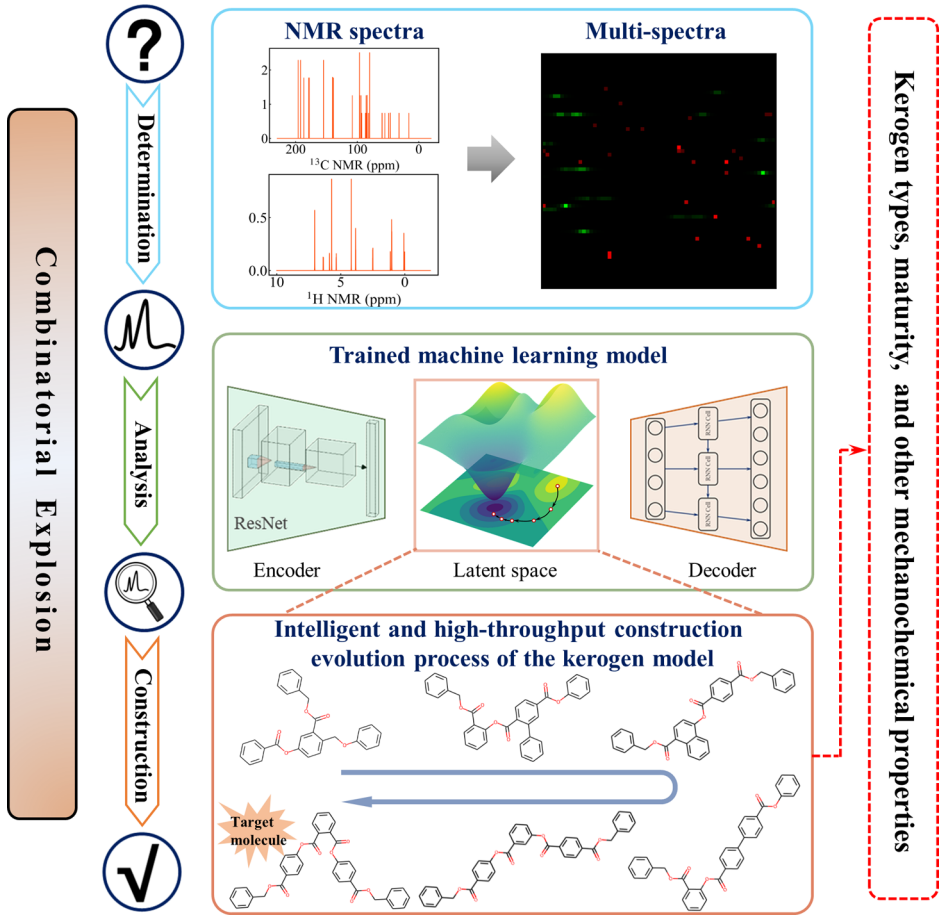



**Figure 1.** Schematic diagram of using ML to construct the kerogen molecular models intelligently. The multi-spectral input form and the matching ML model are designed to solve the combinatorial explosion problem. Also, 650,000 samples are labeled for the training process. The latent space, which connects the multi-spectral and molecular models, is established in the trained ML model. Hence, the target structures can be constructed. The intelligent high-throughput analysis and construction of the target molecules will be carried out in the latent space. Neither manual analysis nor the trial-and-error process is required. Then, the prediction of kerogen types, maturity, and other mechanochemical properties can be obtained through further molecular simulations.

the development of high-performance simulation methods such as ab initio, density functional theory, and molecular dynamics, 3D molecular construction methods are developed.[32−36] Unlike the previous traditional techniques, the molecular dynamics hybrid reverse Monte Carlo (MD-HRMC) method is developed to construct the kerogen model conveniently. The 2D model is not necessary to be built as an intermediate medium in producing a 3D molecular aggregate model using the MD-HRMC method. Only the comprehensive analysis of experimental data is required. However, the MD-HRMC method may cause isolated atoms in the constructed structures, which will affect the mechanical and chemical properties.[37]

There is no doubt that the above-mentioned molecular construction methods are excellent. Nevertheless, due to the combinatorial explosion problem, the theoretically existing structures will explode to an astronomical number steeply with the expansion of the molecular scale.[38] However, there is only one that fits the experimental spectra. As a result, an enormous trial-and-error process based on experiments is required in the traditional method, and the efficiency is extremely low. Two inherent disadvantages exist in these methods. First, the comprehensive analysis of experimental data requires powerful professional ability and experience. Thus, these methods are not conducive to engineering promotion. Second, tremendous time and material resources are consumed during the repeated construction and comparison processes. In the exploitation of shale oil/gas, it is essential to construct the kerogen molecular models for the mining areas before characterizing the mechanochemical properties. However, because of the two disadvantages, only a few kerogen molecular models of mining samples (Green River, Duvernay, etc.) have been able to be used to study the kerogen properties until now. Consequently, it is imperative to develop an intelligent reverse construction method for the kerogen models. Then, the researchers can focus on the study of the molecular mechanochemical properties.[39−41]

Artificial intelligence based on machine learning (ML) neural networks has recently achieved remarkable and fruitful results in many fields.[42] The ML methods have powerful intelligent information extraction and feature learning capabilities and can convincingly solve high-complexity problems.[43] Researchers try to introduce state-of-art ML methods into the analysis of molecular properties.[44−46] Although the data collection and training of ML models are challenging, once the successful ML models are trained, the kerogen molecular models can be constructed from the experimental data intelligently without any intervention by researchers. Thus,


https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

the ML method is the most suitable method to solve the two inherent shortcomings in the construction of the kerogen model currently. In 2020, Kang et al. predicted the molecular skeleton components and types of kerogen by combining ML with the $^{13}C$ NMR spectral dataset.[47] This work proves the feasibility of the ML method to predict the kerogen molecular information intelligently based on the experimental spectra.

This work achieves the intelligent and high-throughput reverse construction of the kerogen molecular models via the comprehensive analysis of multi-spectral experimental data (Figure 1). A sample dataset containing 650,000 groups of $^{13}C$ NMR and $^{1}H$ NMR spectra and the corresponding molecular labels is established to address this challenge. The spectral combination method that the ML models can recognize is proposed, and the predictive capability beyond the input form of a single spectral type is achieved. The prediction results of the trained ML model show that the average similarity between the constructed molecules and the targets is 91.78%. The prediction accuracy and goodness-of-fit about kerogen components, types, and maturity indexes are superb and significantly improved compared to those of the previous ML models applied to predict the single structural information. The above results prove that the trained ML model can address the intelligent high-throughput reverse construction for kerogen models without the manual analysis of experimental data. Therefore, this work exhibits the effectiveness and excellent performance of the ML method to solve the two disadvantages of traditional molecular construction methods. We estimate that our research is an essential exploration of intelligent reverse construction of the kerogen molecular models from the experimental data and will shorten the research cycle and tremendously reduce costs in constructing kerogen models and predicting kerogen properties.

## 2. METHODOLOGY AND MODELS

**2.1. Combinatorial Explosion of Molecular Structures.** There are several kinds of isomerism because of the different bonding modes and binding sites for the same number of atoms, such as chain isomerism, positional isomerism, functional isomerism, optical isomerism, and so forth. With the increase of atoms, the number of theoretically existing structures increases exponentially. This phenomenon is called combinatorial explosion and is the root of the inefficiency of traditional techniques. There is no universal way to calculate the number of isomers yet. We conservatively estimate the magnitude by using the most superficial functional groups. Only the position isomerism and the carbon skeleton are considered. It is assumed that there are two functional groups whose skeleton is composed of four carbon atoms. Two binding sites are in one, and three binding sites are in another. Considering that the molecule is only composed of two functional groups, each with half, the number of structures that meet the theoretical hypothesis is

$$C_a^{a/2} \times 3^{a/2} \times 2^{a/2} = \frac{a!}{[(a/2)!]^2} \times \sqrt{6}^{a} \approx 5^{a} = 5^{n/4} \tag{1}$$

where $a$ is the number of functional groups and $n$ is the number of atoms in the molecule. In fact, the combination of molecules is much more complicated and has more potential configurations. More radically, considering that the four binding sites of the carbon atom are all unique, the number of possible structures is $4^n$. Even so, it has not evaluated all the situations, such as ring formation. Therefore, constructing molecular structures in reverse is an extraordinarily complex and challenging problem.

**2.2. Kerogen Maturity Indexes.** Kerogen maturity is an essential parameter for evaluating the hydrocarbon generation potential. In this work, three maturity indicators are selected to test the accuracy of the trained ML model. The Easy%Ro assumes that vitrinite reflectance is related to the kerogen's H/C and O/C atomic ratio. Also, the calculation formula of Easy %Ro maturity is given based on a large number of experiments[11]

$$\%\mathrm{Ro} = 12\exp(-3.2r_{\mathrm{H/C}}) - 1.2r_{\mathrm{O/C}} \tag{2}$$

where $r_{H/C}$ and $r_{O/C}$ are the ratios of the kerogen component atoms H/C and O/C, respectively. The MMI is also based on the kerogen components and is given in a more concise form

$$\mathrm{MMI} = \frac{1}{1 + r_{\mathrm{H/C}} + r_{\mathrm{O/C}}} \tag{3}$$

The thermal evolution experiments indicate that the MMI positively correlates with vitrinite reflectance.[13] Unlike the above two indexes, the OrbHMI is based on atomic hybridization. Thus, the OrbHMI is closer to the physical basis of maturity. Also, the OrbHMI is expressed as

$$\mathrm{OrbHMI} = \frac{1}{2.85 - 1.1r_{\mathrm{c}} + 0.1r_{\mathrm{o}}} \tag{4}$$

where $r_C = n_{sp^2}^{C}/(n_{sp^2}^{C} + n_{sp^3}^{C})$ and $r_O = n_{sp^2}^{O}/(n_{sp^2}^{O} + n_{sp^3}^{O})$. The $n$ is the number of hybrid orbitals of atoms, superscript C or O denotes the carbon or oxygen atom, respectively, and subscript $sp^2$ or $sp^3$ represents the type of atomic orbital hybridization.[14] Thus, the Easy%Ro and MMI are directly related to the chemical formula, but the OrbHMI is associated with the structural information. All of the information is contained in the molecular structural model.

**2.3. NMR Spectra.** NMR is one of the most effective methods for analyzing the components and structures of unknown substances. The formation of the NMR spectra is due to the vibration frequency shift of the atomic nuclei under the influence of adjacent functional groups. Resonance occurs under strong magnetic fields of different frequencies, and then, the NMR signals are generated. The shift positions of the NMR spectral peaks reflect the types of functional groups, and the peak area reflects the number. Especially in the $^{1}H$ NMR spectra, the number of functional groups is directly proportional to the integrated value of the NMR peak. The process of constructing the molecular models is to comprehensively combine the functional groups according to the various experimental spectra.[48,49] The relation is reflected by the connection between shift peaks in NMR spectra.[50] NMR spectra of different measured nuclei such as $^{1}H$, $^{11}B$ (boron), $^{13}C$, $^{17}O$ and so forth are commonly used. The research object of this work is the kerogen organic molecules, and the main components are C, H, a small amount of O, nitrogen (N), and sulfur (S). The properties (type, maturity, etc.) of kerogen are only closely related to the skeleton structure, consisting of C, H, and O. Thus, $^{13}C$ NMR and $^{1}H$ NMR are used for the analysis and construction.

**2.4. Simplified Molecular Input Line Entry System.** The simplified molecular input line entry system (SMILES) was initiated by Weininger in 1987.[51] In this way, the molecule


https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

can be written as a single-line string. Generally, the single bonds and hydrogen atoms are not displayed in SMILES, and the double and triple bonds are denoted by = and #, respectively. For example, ethylene can be written as C=C. The benzene molecule can be composed as c1ccccc1, where the lowercase letters indicate that the atom is aromatic and the numbers represent the closed-loop position. The linear representation of molecules also has the International Chemical Identifier (InChI) form, but InChI is far less legible than SMILES. Therefore, SMILES is chosen as the tool of molecular representation. There are many SMILES standards, and the SMILES characters of the same molecule under different standards may differ. Thus, the canonical SMILES is selected to eliminate the impact of nonuniform standards. Canonical SMILES can establish a one-to-one correspondence between molecules and their International Union of Pure and Applied Chemistry (IUPAC) names.

**2.5. Molecular Fingerprints and Similarity.** The molecular fingerprints are designed to characterize the substructures of molecules in the chemical database for searching. With the expansion of fingerprint types and the development, the fingerprints are beginning to be used to analyze the similarity between molecules, predict the molecular activity and virtual screening, and so forth.[52−54] The most famous fingerprints are the Morgan fingerprints, Molecular ACCess System (MACCS) key fingerprints, and topological fingerprints. The Morgan fingerprints are also known as the circle fingerprints, which are obtained by searching the molecular substructures with a defined radius $r$. The MACCS key fingerprints compare molecular substructures with the predefined substructural dictionary. If the defined substructure is contained in the molecule, the value of the corresponding key will be set as 1; otherwise, it is 0. These fingerprints are the binary array containing only 0 and 1.

The similarity calculation methods based on molecular fingerprints include Dice, Tanimoto, Cosine, and so forth.[55] The Dice similarity coefficient ($S_{\text{Dice}}$) represents the ratio between the double of the intersection and the union of the two fingerprints. The calculation method is as follows

$$S_{\text{Dice}} = \frac{2 \times n_{AB}}{n_A + n_B} \tag{5}$$

where $n_{AB}$ means the number of elements of the intersection between fingerprints $A$ and $B$. The $n_A$ and $n_B$ are the total numbers of elements in fingerprints $A$ and $B$, respectively. While the intersection is equal to the union, $S_{\text{Dice}} = 100\%$, the two molecules are same; on the contrary, if there is no intersection between the fingerprints of the two molecules, $S_{\text{Dice}} = 0$, the two molecules are completely different. The Dice coefficient is proportional to the size of the intersection and is associated with a clear physical meaning. Therefore, the Dice coefficient of the Morgan fingerprint (ECFP2) is used to characterize the similarity between the constructed molecules and the targets.

The similarity map is drawn by coloring the same and the different functional groups between molecules based on fingerprints. The same and the different fragments between molecules can be displayed in detail and intuitively.[56] In the analysis process of this work, the open-source toolkit RDKit is used to calculate molecular fingerprints and similarity.[57]

**2.6. ML Model.** The ML neural network model is inspired by how neurons are connected in the human brain. The intelligent models are established with a self-learning ability by combining computer science, probability statistics, optimization theory, and so forth.[58] Generally, the model includes four parts: feedforward neural network, loss function, optimization method,[59] and backpropagation algorithm.[60]

The recurrent neural network (RNN) models are designed to solve the sequence analysis problems.[61] The RNN models are mainly used in point-to-sequence,[62] sequence-to-point,[63] and sequence-to-sequence forms.[64,65] In this study, the structural construction through NMR multi-spectra is a sequence-to-sequence problem essentially. The reconstructed NMR spectra in this study and the canonical SMILES can be seen as the sequence formed according to specific rules. Therefore, referring to the natural language process model, which is created to solve the sequence-to-sequence problem, our ML model is also designed in two parts: an encoder and a decoder. The encoder is built based on the residual neural network (ResNet).[66−68] The multi-NMR spectral data are abstracted and encoded into a high-dimensional learning molecular fingerprint,[69] which contains 512 elements with values between −1 and 1. The ResNet is developed based on the convolutional neural network (CNN).[70,71] It improves the efficiency of information dissemination by connecting different convolutional layers directly. Following this, the performance of deep convolutional layers is improved. In the ResNet, the ResNet unit can be expressed as

$$\boldsymbol{A}^{l+1} = \sigma[g(\boldsymbol{A}^l) + f(\boldsymbol{A}^l, \boldsymbol{W}^l)] \tag{6}$$

where $\boldsymbol{A}^l$ represents the matrix value of the natural network nodes in layer $l$; the $g(\boldsymbol{A}^l)$ is the direct connect item, where the input information passes through several layers of the network to the output directly; the $f(\boldsymbol{A}^l,\boldsymbol{W}^l)$ is the trainable mapping function, in which the target features are extracted; and $\sigma(\cdot)$ is the nonlinear activation function. From eq 6, the difference between the ResNet and the ordinary neural network is that the ResNet adds a direct connection path that can pass the input information into the output position directly. In this work, the rectified linear unit activation function is used, $h(\boldsymbol{A}^l) \equiv \boldsymbol{A}^l$. Then, eq 6 can be simplified to $\boldsymbol{A}^{l+1} = \boldsymbol{A}^l + f(\boldsymbol{A}^l,\boldsymbol{W}^l)$ in the efficient nodes. Finally, the trainable mapping function is

$$f(\boldsymbol{A}^l, \boldsymbol{W}^l) = \boldsymbol{A}^{l+1} - \boldsymbol{A}^l \tag{7}$$

and eq 7 shows that the prediction object is the residual between the predicted value and the input value. This is also the origin of the ResNet name. For the residual unit connected across neural network layers ($n$) directly, the output is

$$\boldsymbol{A}^{l+n} = \boldsymbol{A}^l + \sum_{i=l}^{l+n-1} f(\boldsymbol{A}^i, \boldsymbol{W}^i) \tag{8}$$

according to eq 8, in the ML backpropagation algorithm, the gradient of the loss function $\boldsymbol{L}$ is

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{A}^l} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{A}^{l+n}} \frac{\partial \boldsymbol{A}^{l+n}}{\partial \boldsymbol{A}^l} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{A}^{l+n}} \left\{ 1 + \frac{\partial}{\partial \boldsymbol{A}^l} \left[ \sum_{i=l}^{l+n-1} f(\boldsymbol{A}^i, \boldsymbol{W}^i) \right] \right\} \tag{9}$$

The mean absolute error (MAE) loss function is selected to evaluate the distance between the predicted value $\boldsymbol{y}^{\text{pred}}$ and the target $\boldsymbol{y}^{\text{true}}$

$$L_{\text{MAE}} = \frac{1}{n_{\text{batch}}} |\boldsymbol{y}^{\text{true}} - \boldsymbol{y}^{\text{pred}}| \tag{10}$$



https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

where $n_{batch}$ represents the batch size, which is the number of samples put into the ML model at each time. During the training, the value $\partial[\sum_{i=l}^{l+n-1} f(\boldsymbol{A}^i, \boldsymbol{W}^i)]/\partial \boldsymbol{A}^l$ will not always be −1, which ensures that the gradient of the deep neural network always exists and avoids the problem of gradient disappearance. At the beginning of constructing the ML model, the encoders with varying scales of the fully connected neural network, CNN, and ResNet are established. The test results prove that the ResNet with eight residual layers combined with one fully connected layer obtains the best performance.

The decoder implements transfer learning based on the pretrained model of Winter et al.[72] The encoded learning molecular fingerprint is translated into the corresponding SMILES formula in the RNN decoder. The RNN is designed to process time-series data with specific arrangement rules. The calculation process of the RNN can be expressed as

$$\begin{aligned} \boldsymbol{h}_t &= \sigma(\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}_t) \\ \hat{\boldsymbol{y}}_t &= \mathrm{softmax}(\boldsymbol{V}\boldsymbol{h}_t + \boldsymbol{c}_t) \end{aligned} \tag{11}$$

where $\boldsymbol{h}_t$ is the hidden nodes of the RNN; $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{W}$ represent the weight matrix; $\boldsymbol{b}$ and $\boldsymbol{c}$ represent the bias parameters; $\hat{\boldsymbol{y}}_t$ is the predicted value of step $t$; and the softmax function ($\mathrm{softmax} = \exp(a_i)/\sum\exp(a_k)$) reconstructs the output of the neural network as a probability distribution between 0 and 1 with a sum of 1. The negative log likelihood is used as the loss function $\boldsymbol{L}^*$ in the decoder model to measure the distance between the target sequence $\boldsymbol{y}$ and the predicted sequence $\hat{\boldsymbol{y}}$

$$\boldsymbol{L}^*(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{t=1}^{\tau} [P(y_t) \ln P(\hat{y}_t | \hat{y}_1, ..., \hat{y}_{t-1}, x_1, ..., x_T)] \tag{12}$$

During the RNN training process, it is required to take the gradient of the loss function relative to the trainable parameters ($\boldsymbol{W}$, $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{b}$, and $\boldsymbol{c}$) and adjust the parameters according to the optimization strategy. Combining eqs 11 and 12, the gradient of the loss function to matrix $\boldsymbol{W}$ is

$$\frac{\partial \boldsymbol{L}^*}{\partial \boldsymbol{W}} = \sum_{t=1}^{\tau} \sum_{k=1}^{t} \left( \frac{\partial \boldsymbol{L}_t}{\partial \boldsymbol{Z}_k} \frac{\partial \boldsymbol{Z}_k}{\partial \boldsymbol{W}} \right) \tag{13}$$

where $\boldsymbol{Z}_k = \boldsymbol{W}\boldsymbol{h}_{k-1} + \boldsymbol{U}\boldsymbol{x}_k + \boldsymbol{b}_k$ represents the linear output of the neural network nodes. Then, the error term of the loss function is obtained as

$$\boldsymbol{\delta}_{t,k} = \frac{\partial \boldsymbol{L}_t}{\partial \boldsymbol{Z}_k} = \frac{\partial \boldsymbol{L}_t}{\partial \boldsymbol{Z}_{k+1}} \frac{\partial \boldsymbol{Z}_{k+1}}{\partial \boldsymbol{h}_k} \frac{\partial \boldsymbol{h}_k}{\partial \boldsymbol{Z}_k} = \mathrm{diag}[\sigma'(\boldsymbol{Z}_k)] \boldsymbol{W}^{\mathrm{T}} \boldsymbol{\delta}_{t,k+1} \tag{14}$$

Then, the error term of previous layer nodes can be calculated using the next layer, and the information is propagated backward from the back to the front. This is the principle of the RNN backpropagation algorithm. In addition, $\partial \boldsymbol{Z}_k/\partial \boldsymbol{w}_{ij} = \boldsymbol{h}_{k-1}$. Finally, the $\partial \boldsymbol{L}^*/\partial \boldsymbol{W}$ is obtained as

$$\frac{\partial \boldsymbol{L}^*}{\partial \boldsymbol{W}} = \sum_{t=1}^{\tau} \sum_{k=1}^{t} (\boldsymbol{\delta}_{t,k} \boldsymbol{h}_{k-1}^{\mathrm{T}}) \tag{15}$$

and in the same way, the gradients of the other parameters are

$$\begin{aligned} \frac{\partial \boldsymbol{L}^*}{\partial \boldsymbol{U}} &= \sum_{t=1}^{\tau} \sum_{k=1}^{t} (\boldsymbol{\delta}_{t,k} \boldsymbol{x}_k^{\mathrm{T}}), \\ \frac{\partial \boldsymbol{L}^*}{\partial \boldsymbol{b}} &= \sum_{t=1}^{\tau} \sum_{k=1}^{t} \boldsymbol{\delta}_{t,k}, \\ \frac{\partial \boldsymbol{L}^*}{\partial \boldsymbol{V}} &= \sum_{t=1}^{\tau} [(\hat{\boldsymbol{y}}_t - \boldsymbol{y}_t) \boldsymbol{h}_t^{\mathrm{T}}], \\ \frac{\partial \boldsymbol{L}^*}{\partial \boldsymbol{c}} &= \sum_{t=1}^{\tau} \left( \frac{\partial \boldsymbol{L}_t}{\partial \boldsymbol{O}_k} \frac{\partial \boldsymbol{O}_k}{\partial \boldsymbol{c}} \right) = \sum_{t=1}^{\tau} (\hat{\boldsymbol{y}}_t - \boldsymbol{y}_t) \end{aligned} \tag{16}$$

then, the loss gradient information in the RNN can be transmitted backward to update the training iteration parameters.

The Adam optimizer is used during the training process and sets the initial learning rate to $5 \times 10^{-5}$. The ReduceLROnPlateau strategy of PyTorch[73] is adopted for the learning rate, which is dropped to 0.7 times of the current learning rate for each adjusting step. The early stopping and batch regularization strategies are adopted to prevent overfitting during the training process. The ML model is built on two open-source artificial intelligence frameworks: PyTorch and TensorFlow.[74] In the evaluation of kerogen components, types, and maturity indexes, the prediction accuracy $P_{acc}$ and coefficient of determination $R^2$ are chosen to express the degree of fit between the predicted value and the actual value

$$P_{acc} = \frac{|y_i - y_i'|}{y_i} \times 100\% \tag{17}$$

$$R^2 = 1 - \frac{\sum_i (y_i - y_i')^2}{\sum_i (y_i - \bar{y}_i)^2} \tag{18}$$

where $y_i$ and $y_i'$ denote the true value and the prediction value, respectively, and $\bar{y}$ is the mean value of the actual value. The range of $R^2$ is 0 to 1, and the closer the $R^2$ value is to 1, the better the performance of the model is.

## 3. RESULTS AND DISCUSSION

**3.1. ML Datasets.** Establishing qualified datasets is prerequisite for ML methods. All the target features need to be contained in the training dataset and distributed reasonably. In this way, the generalization ability of target features can be learned by the ML model during the iterative training. Tens of thousands of qualified samples are often required for the ML models. It is challenging to collect massive kerogen molecular samples as the training dataset. However, according to the characteristics of the ML model, the chemical bonding rules, which are common to all molecules, are learned by the ML model during the training process. The ML model can obtain the predictive ability of the kerogen molecular model through the training of the other organic molecules. Thus, the sample molecules are derived from the open-source molecular database ChemPub,[75] historical paper,[4,32,50] and experiments.[34,47] The NMR spectra are calculated using software MestRenova 14.2.[76] The 650,000 groups of $^{13}$C NMR and $^{1}$H NMR spectra and their corresponding molecular SMILES structural labels are marked in this study to improve the generalization ability as much as possible. Also, the NMR spectra automatic labeling algorithm, in which the NMR


https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

**Table 1. All-Atom-Scale Information of Molecules in the Sample Datasets**

| name | number | average | variance | minimum | maximum | median |
|---|---|---|---|---|---|---|
| total dataset | 650,000 | 41.35 | 14.05 | 5 | 119 | 39 |
| training dataset | 550,000 | 41.35 | 14.06 | 5 | 119 | 39 |
| validation dataset | 50,000 | 41.33 | 14.00 | 5 | 107 | 39 |
| test dataset | 50,000 | 41.34 | 14.00 | 6 | 108 | 39 |

**Table 2. Number Information of Carbon and Oxygen Atoms in the Sample Datasets**

| name | number | average | variance | minimum | maximum | median |
|---|---|---|---|---|---|---|
| total dataset | 650,000 | 20.14 | 7.06 | 2 | 46 | 19 |
| training dataset | 550,000 | 20.14 | 7.06 | 2 | 46 | 19 |
| validation dataset | 50,000 | 20.17 | 7.03 | 4 | 46 | 19 |
| test dataset | 50,000 | 20.13 | 7.05 | 4 | 44 | 19 |

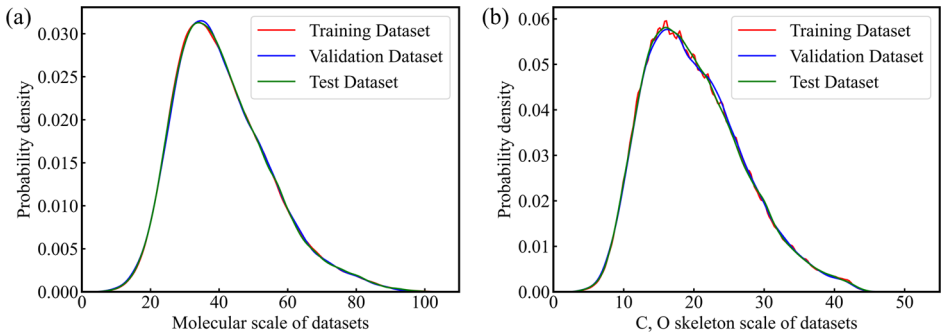



**Figure 2.** Distribution of the total molecular scale and C, O skeleton scale in the training, validation, and test datasets. (a) Molecular scale of datasets. The molecular scale is the number of all atoms. (b) C, O scale of datasets. The C, O skeleton scale, which is the measure of SMILES, representes only the number of C and O in the molecule.

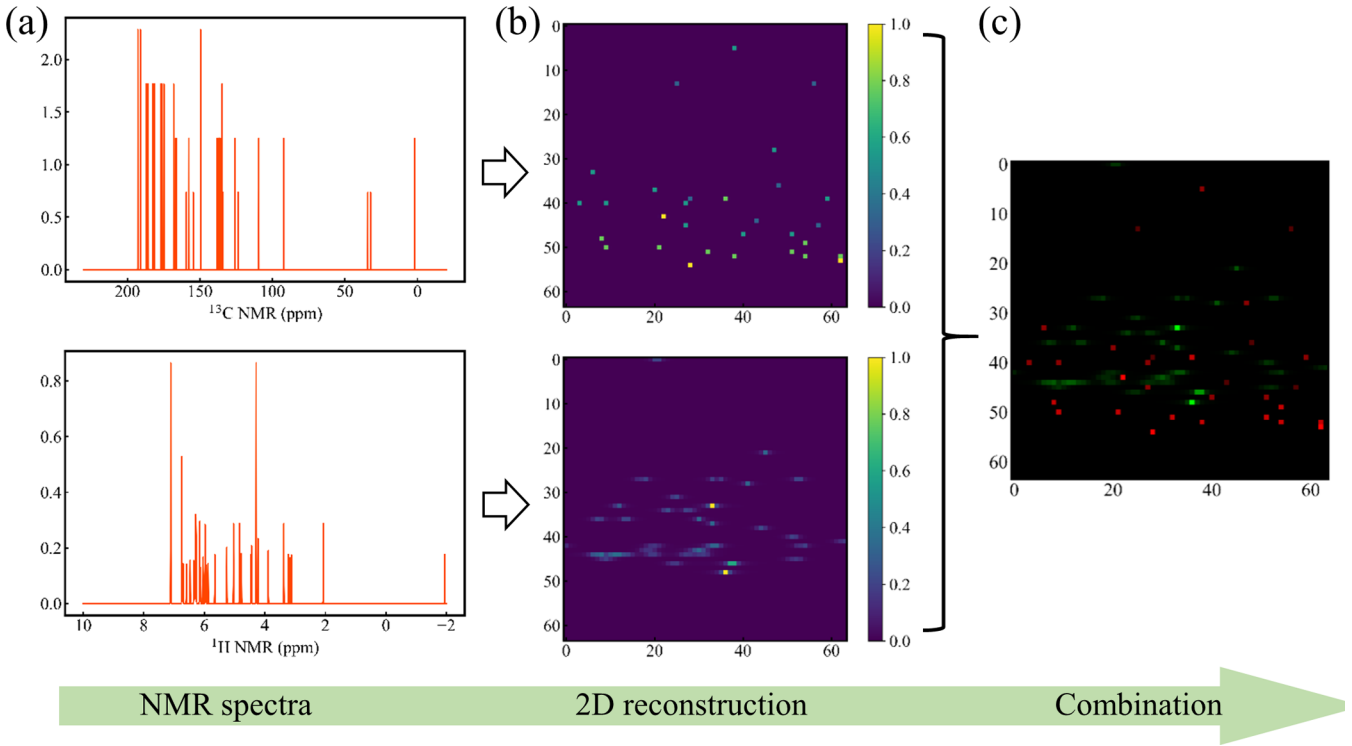



**Figure 3.** (a) $^{13}$C NMR and $^{1}$H NMR spectra. (b) 2D reconstruction form of the NMR spectra. (c) Visual form of the combined multi-spectra.

spectra of the molecules can be labeled automatically, is compiled. Thus, it is possible to label massive samples for the training of the ML model. Since there is no practical way to convert the 2D spectra into the form that can be input into the ML model, we creatively standardize the $^{13}$C NMR spectrum and $^{1}$H NMR spectrum into a one-dimensional (1D) array on the labeling algorithm through fitting and equally spaced sampling.[47] The horizontal axis, which is the shift position of



https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

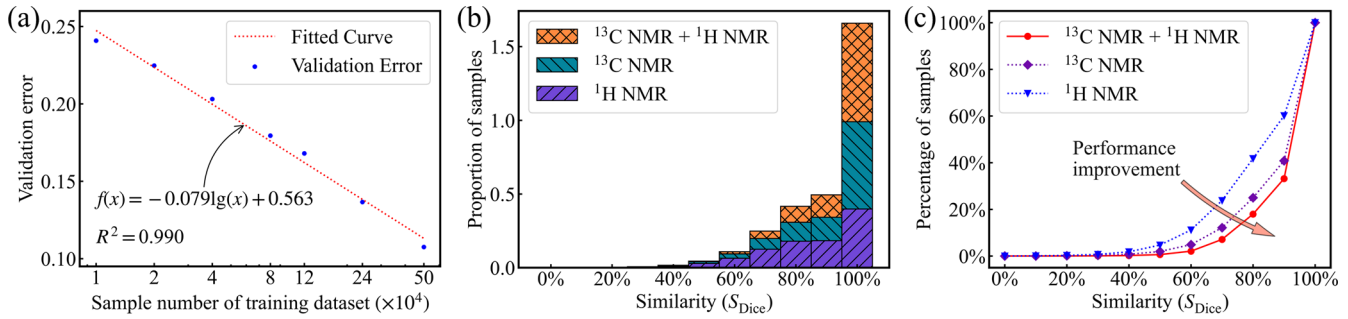



**Figure 4.** Influence of the model generalization error due to the number of samples in the training dataset. (a) Change of the validation error with the training sample number. (b,c) Constructed sample number proportion histogram and the cumulative curve of different similarities in the models that are trained using $^{1}$H NMR, $^{13}$C NMR, and the combined NMR spectra.

the NMR peaks, is implied in the 1D array index. A total of 4096 sampling points are set during conversion in this work, and the horizontal axes of $^{13}$C NMR and $^{1}$H NMR spectra range from −20 ppm to 230 ppm and −2 ppm to 10 ppm, respectively.

As shown in Table 1, the number of atoms in the sample molecules labeled in this work is between 5 and 119. The median atomic scale of the molecules is 39, and the average is 41.35. In the SMILES, the H atoms are implied to be coordinating the molecules and do not appear directly. The complexity of the molecular SMILES is only affected by the number of the C and O atoms and their combination forms. Thus, the number of C and O atoms in the sample molecules is counted. As shown in Table 2, the number of C and O atoms is between 2 and 46, the average number of atoms is 20.14, and the median number is 19.

The total 650,000 samples are randomly divided into training, validation, and test datasets according to 550,000:50,000:50,000 (Tables 1 and 2). The training dataset is made for training the ML models. The validation dataset is responsible for adjusting the hyperparameter. The test dataset is built to verify the generalization ability of the trained model. It should be pointed out that no intersection exists among the three datasets. In order to ensure the uniformity of the feature distribution, both in terms of the molecular atom scale and the C and O atom scale, the average, variance, and median parameters of the three datasets are almost the same. Figure 2a,b show the molecular proportions of different scales in the training, validation, and test datasets in detail to illustrate the distribution between the datasets. The same distribution of the molecular scale and C, O skeleton atoms proves the effectiveness of the datasets. Also, the scales of the molecular C, O skeleton in the datasets are mainly concentrated between 10 and 30 (the full-atom scale of molecules is 20−60), and only a few of the samples are between 2−10 and 30−50.

**3.2. Preprocessing of NMR Spectra.** The origin NMR spectra (Figure 3a) are processed as a 1D matrix during the sample labeling, and the horizontal axis information is implicit in the index. The 1D data are generally processed using the 1D convolution technique in the ML neural network. However, limited by the perceptual domain of the 1D convolutional layer, the contact information between the peaks of the NMR spectra cannot be obtained by the shallow network layers. Then, part of the implicit information will be lost, which is harmful for the construction of the structures. Because of this, a 2D method that aims to fold the NMR spectral sequence is designed. As displayed in Figure 3b, the NMR spectral features of a certain distance away will be extracted simultaneously during the training process. The connection between the peaks will also be reflected in the shallow layers.

In order to comprehensively analyze the multi-spectra, we also design a combination of spectral input strategies (Figure 3c). Each reconstructed 2D spectrum will occupy one single channel of the combined multi-spectrum, and the combined NMR spectrum will be entered into the neural network model. Then, the molecule can be predicted by the combined spectrum of its $^{13}$C and $^{1}$H NMR spectra. The test results show that the combined input can significantly improve the performance of the ML model (the details are discussed in Section 3.3). Not only is this method suitable for $^{13}$C and $^{1}$H NMR spectra, but also the other spectra that are used to characterize the molecular structures, such as XRD, XPS, and infrared spectra, may be combined to train the ML models. Also, the interfaces are reserved in the model, which can be used for other spectra directly.

**3.3. Impact of the Training Sample Number and Multi-Spectra.** The ML models are trained by using the training datasets containing 10,000, 20,000, 40,000, 80,000, 120,000, 240,000, and 500,000 samples, respectively. Also, the influence of samples number during the molecular construction is explained in detail. Figure 4a exhibits the validation results of the trained models on the same validation dataset. The relationship between the validation error and the training sample number is

$$f(x) = -0.079\lg(x) + 0.563 \tag{19}$$

where $x$ is the number of samples included in the training dataset and $f(x)$ is the validation error. The value of $R^2$ is 0.990, which proves the extremely strong correlation between the fitted formula and the original data. The validation error is proportional to the logarithm of the sample number in the training dataset. Therefore, the training samples will increase exponentially, while the same improvement is obtained using the ML model. The relationship between the training sample number and the validation error is described in eq 19. When the number of training samples reaches about $1.33 \times 10^7$ groups, the validation error will be close to zero. In fact, the generalization ability will eventually stabilize with the samples improving, and it is impossible to develop according to the linear relationship expressed by eq 19 indefinitely. Obtaining such an astronomical number of qualified molecules and labeling them are considerable workload. Therefore, it is more feasible to improve the performance by optimizing the ML model itself than by adding more samples while the generalization ability reaches a certain accuracy, especially in the fields where collecting samples is complex.



https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

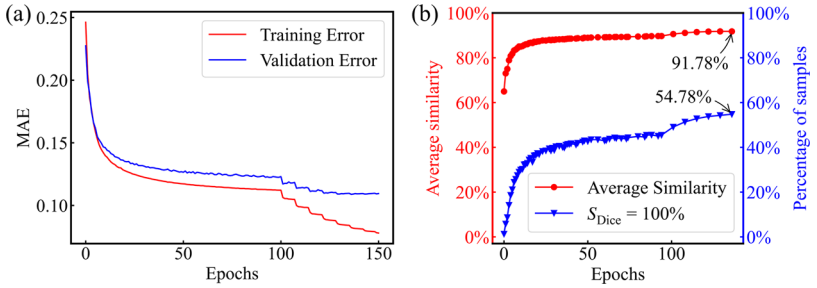



**Figure 5.** (a) Change of the training error and validation error during the training process. (b) Average similarity and the entirely matched sample ($S_{Dice}$ = 100%) percentage with epochs.

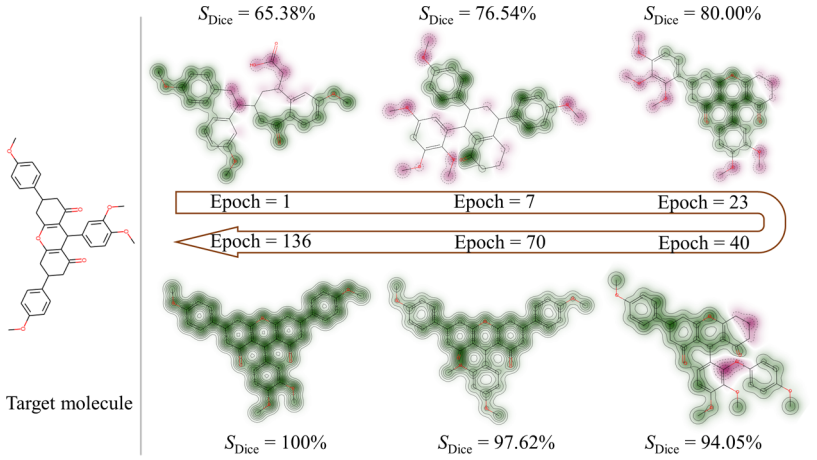



**Figure 6.** Evolution of the target molecular sample [SMILES: COc1ccc(C2CC(=O)C3=C(C2)OC2=C(C(=O)CC(c4ccc(OC)cc4)C2)-C3c2ccc(OC)c(OC)c2)cc1] at epoch = 1, 7, 23, 40, 70, and 136.

The advantages of the combined input form of spectra over the single input are introduced in Figure 4b,c. The ML models are trained using a single $^{1}$H NMR spectral training dataset, a single $^{13}$C NMR spectral training dataset, and a combined form of $^{13}$C NMR and $^{1}$H NMR spectral training dataset, respectively. The same 500,000 sample molecules are contained in the three training datasets during the training process. All the three trained models are validated in the same 50,000 group samples. There is no intersection between the training dataset and the validation dataset. Thus, all influencing factors except the input form of spectra are excluded.

The validation results are shown in Figure 4b. The construction accuracy of the three input forms shows the same distribution trend. The proportion of the sample number gradually increases with the similarity degree, and the maximum value is reached at the similarity $S_{Dice}$ = 100%. The result means that the single $^{1}$H NMR or $^{13}$C NMR spectral dataset can enable the ML model to obtain the molecular construction ability. Even so, the single input form accuracy is not as good as that of the multi-spectral input form, especially the single $^{1}$H NMR spectrum. The construction accuracy is much lower than that of the multi-spectral input form. It can be clearly seen from Figure 4c that the molecules predicted by the multi-spectral trained model of $S_{Dice}$ < 80% are significantly less than that by the single input model. However, the case for $S_{Dice}$ > 80% is the opposite. Hence, the effectiveness of the multi-spectral input form and the trained model is proved. The trained ML model can comprehensively analyze the combined spectra and obtain the predictive ability that is difficult to be achieved with a single spectral type. In addition to NMR spectra, the XPS spectra, XRD spectra, and so forth can also be used to characterize the structures. The different types of spectra usually indicate unique molecular characteristics. Consequently, increasing the number of spectral types is an effective way to improve the performance of the ML model where the number of samples is insufficient. The spectral reconstruction method is entirely adaptable to these 2D spectra. Compared with the previous single-input model, our model has more outstanding expansion capabilities and potential.

**3.4. Training Process of the Optimal ML Model.** The training process of the optimal ML model with $^{13}$C NMR and $^{1}$H NMR multi-spectra is shown in Figure 5. The early stopping strategy is set during the training process, and a total of 151 iterative epochs are carried out (Figure 5a). The optimal model is achieved at the 136th epoch. The corresponding training error is 0.082, and the validation error is 0.108. The element value of learning a molecular fingerprint is between −1 and 1. According to the definition of the MAE loss function (eq 10), the validation error of the model for each element is about 5.4%. The training error decreases with epochs in subsequent training iterations, but the validation error oscillates at 0.109. The ML model has not achieved a better performance during the following training.

Figure 5b shows the molecular percentage of the average similarity and the complete accuracy in the test dataset during the training process. In the first 15 epochs, the performance of the training model is rapidly improved. The validation error is reduced from 0.227 to 0.139, and the average similarity is increased from 64.96% to 86.56%. The proportion of



https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

molecules, whose predictions are exact, is increased sharply from 1.24% to 34.40%. Then, the training enters the plateau period, and the performance is improved slowly. In the last 16−151 epochs of training, the validation error is only reduced by 0.031 to 0.108. Correspondingly, the average similarity increases to 91.78%, and the proportion of entirely accurate molecules increases to 54.78%.

The molecular model evolution process is exhibited in Figure 6 to explain the model performance changes during training. The matched functional group distribution between the constructed molecules at epoch = 1, 7, 23, 40, 70, and 136 and the target are exhibited detailly. The similarities ($S_{Dice}$) are 65.38%, 76.54%, 80.00%, 94.05%, 97.62%, and 100%, respectively. With the accuracy of the training models increasing, the number of identical functional groups (marked green) between the constructed molecule and the target structure increases. Also, the different functional groups (purple area) reduce correspondingly. The predicted structures evolve in the direction of the target, and finally, the molecules that are completely consistent with the target are constructed. Figure 6 shows that only a few functional groups are different between the molecular structures with the similarity of over 80.00% and the target. Also, only individual characters are shifted or missing in the SMILES formula. The reason is that the completely accurate canonical SMILES expression is unique, and the syntax is rigorous. As a result, the performance at the level of 80% average similarity can be quickly achieved, but it is challenging to improve further. By recalling the training process of the 11th to 136th epochs, the results exhibit that the similarity of constructed molecules is maintained to be in the range from 80% to 100% until they match the target completely. These intermediate structures only have individual functional groups that do not match the target molecule. Hence, the difficulty of constructing molecules based on experimental data is illustrated further.

**3.5. Construction Accuracy of the Trained Model.** Although the generalization ability can be indirectly reflected by the validation error, it cannot explain the exact situation of the constructed molecular structures in the test dataset. Therefore, the percentage information of constructed molecules with different similarities is counted in Figure 7. The

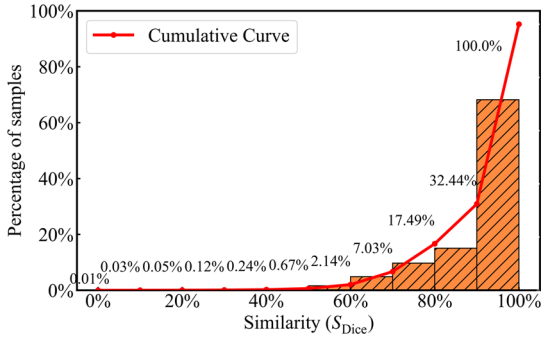



**Figure 7.** Prediction similarity distribution of the test dataset.

proportion of the constructed molecules increases with the similarity, and the maximum is reached at $S_{Dice}$ = 100%. The molecules with the similarity of $S_{Dice}$ < 60% account for only 2.14%. Compared with this, the proportion of molecules with the similarity of $S_{Dice}$ > 80% reaches 82.51%, and the complete accurate molecules account for 54.78%. Consequently, these parameters exhibit the superb generalization ability of the trained model.

Due to the combinatorial explosion, the difficulty of constructing molecules for different scales varies greatly. Figure 8a,b show the construction accuracy distribution of the trained model for the test samples in different C, O scales. It can be seen from Figure 8a that the molecular similarity with the C, O skeleton scale smaller than 30 is distributed in the range of 90% < $S_{Dice}$ ≤ 100%, and the similarity of 30−50 is distributed in the range of 60% < $S_{Dice}$ ≤ 100%. Figure 8b shows the proportion of the construction accuracy with the different scales in the test dataset. At the accuracy level of the similarity $S_{Dice}$ > 80%, although the proportion of constructed molecules decreases as the molecular skeleton scale increases, the overall degree of decline is not significant. The ratio of up-to-standard molecules with the skeleton scale less than 30 is between 78.4% and 86.3%, and the ranges 30−40 and 40−50 can also reach 73.8% and 62.0%, respectively. Unlike the case of $S_{Dice}$ > 80%, the proportion of wholly matched molecules ($S_{Dice}$ = 100%) gradually decreases with the increase of the C, O skeleton scale. On the scale of 0−10, the ratio of completely accurate molecules is achieved at 70.0%. However, while the skeleton scale expands to 40−50, only 9.0% is left.

Combining with the molecular scale distribution of the training dataset and validation dataset in Figure 2, the samples with the C, O skeleton scale distributed between 10 and 40 are the highest, accounting for about 96.1%. The skeleton scale in the range of 0−10 is 2.9% and that in the range of 40−50 is only 1.0%. Although a few samples are in ranges 0−10 and 40−50, the construction ability, which is analogous to that of 10−40 at the accuracy level of $S_{Dice}$ > 80%, is still obtained. This shows that the fine analytical capabilities for the functional groups in the NMR spectral peaks are learned from the training dataset. The learned ability can be accurately applied to other unknown molecules. At the accuracy level of entirely accurate, the predictive ability gradually decreases with the skeleton scale, indicating that the predictive ability for the large-scale molecules is greatly affected by the explosive increase in the construction complexity caused by the expansion of the molecular scale. The correlation with the proportion of the corresponding scale samples in the training dataset is weaker. After all, the skeleton scales from 10 to 30 account for the most samples. The proportion of molecules whose constructed structures are completely accurate still decreases with the scale increasing. The conclusion points out that the construction ability of the larger molecules can be learned from the training dataset where a sufficient amount of the smaller molecules is included. Increasing the number of small-molecule samples can make up for the difficulty of obtaining the large-molecule samples to a certain extent.

**3.6. Prediction Accuracy of Kerogen Parameters.** The kerogen types directly depend on the structure. Three thousand groups of kerogen fragment samples are reconstructed to further verify the effectiveness of the trained ML model. Part of these kerogen samples are obtained from the pyrolysis products of Erdos and Songliao kerogen models,[34,47] and others are the fragments of the published models.[32,33,50] The prediction accuracy of C, H, and O skeleton components is shown in Figure 9a−c. The accuracies of C, H, and O can reach 99.1%, 98.1%, and 97.1%, respectively, and the $R^2$ values are 0.990, 0.970, and 0.957, respectively, which proves the excellent goodness-of-fit between the predicted value and the true value. Compared with ref 47 (C: 96.1%, H: 94.8%, and O:


https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

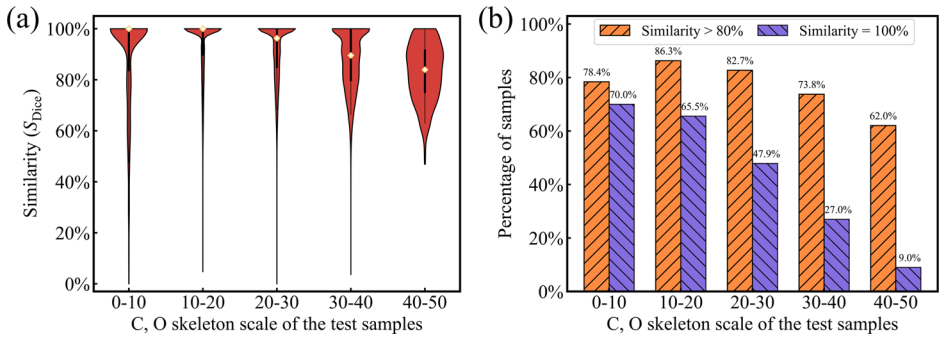



**Figure 8.** (a) Violin figure of similarity distribution with the C, O skeleton scale. The red area represents the probability density distribution. The diamond-shaped point represents the median value, the thick straight line is the 25%−75% distribution interval of samples, and the thin straight line is the 5%−95% distribution interval of samples. (b) Percentage of samples with the C, O skeleton scale.

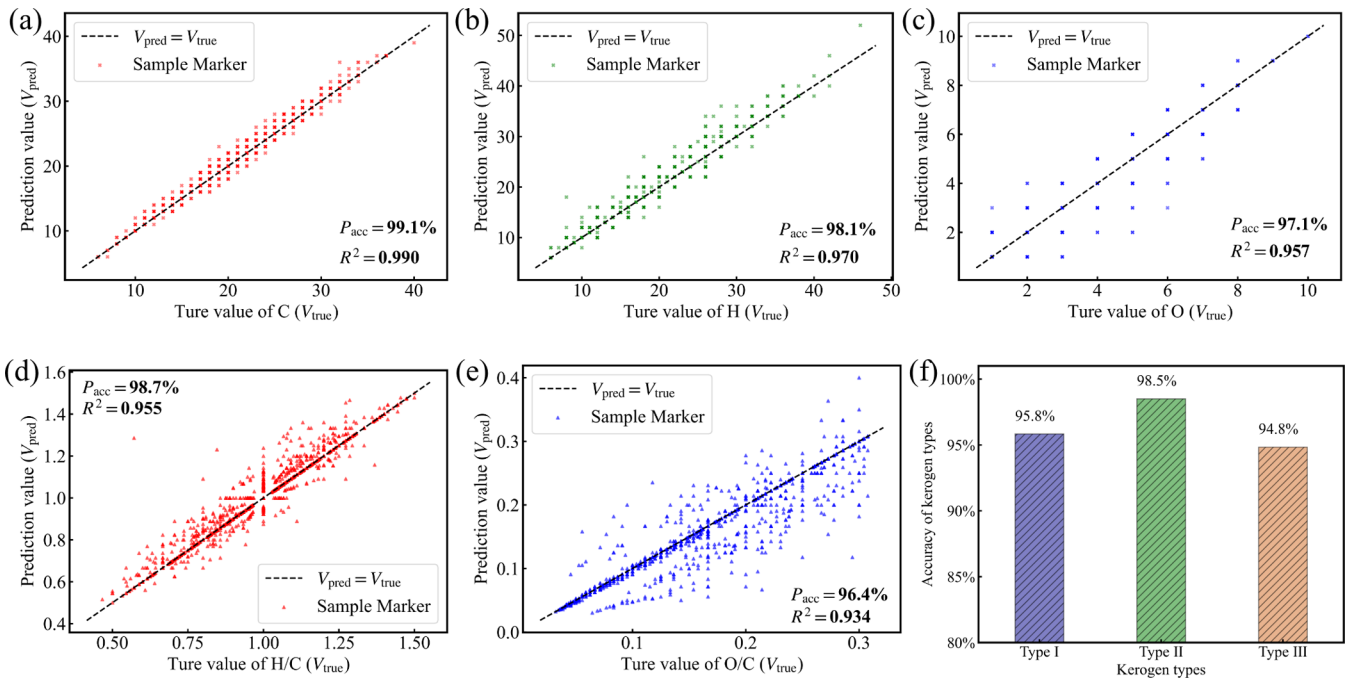



**Figure 9.** Prediction accuracy of kerogen skeleton components and types.

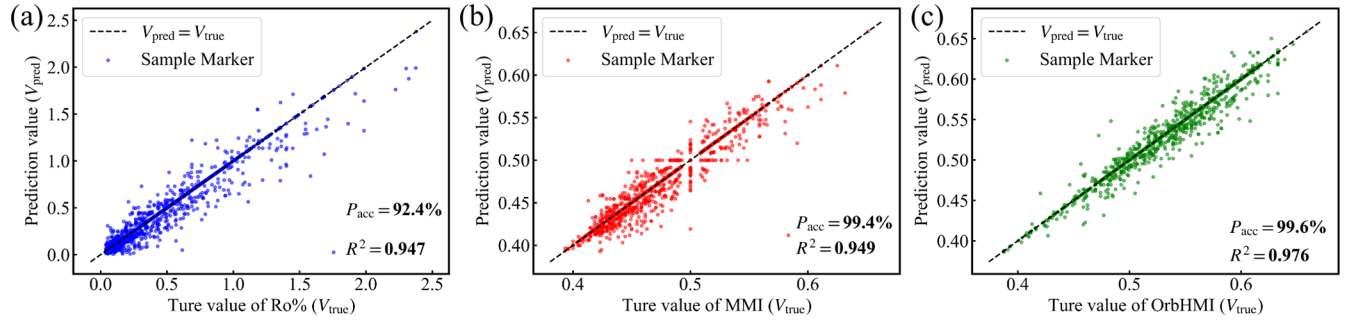



**Figure 10.** Prediction accuracy of kerogen maturity indexes.

81.7%), the prediction accuracy of each component is significantly improved. The high-precision prediction of the components makes the kerogen types more accurate. The H/C and O/C atomic ratios are calculated in Figure 9d,e based on the skeleton components, and then, the prediction accuracy of kerogen types is analyzed (Figure 9f). The prediction accuracy of Type I kerogen is 95.8% (increase by 5.8%), of Type II is 98.5% (increase by 9.5%), and of Type III is 94.8% (increase by 5.4%).

Maturity is another important inherent parameter of kerogen. The %Ro and MMI parameters can also be calculated from the component information. These two indexes are analyzed in Figure 10a,b. The prediction accuracy of %Ro maturity indexes is 92.4%, with an $R^2$ of 0.947, and the MMI's accuracy is 99.4%, with an $R^2$ of 0.949. Unlike %Ro and the



https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

MMI, the OrbHMI is based on the hybrid orbital of the kerogen structure. It is directly related to the molecular bond and is closer to the physical nature of kerogen cracking to oil/gas. The hybrid orbital information can also be obtained from the molecular structure. Figure 10c shows the accuracy of the constructed molecules to predict the OrbHMI maturity. The accuracy is 99.6%, and $R^2$ is 0.976. This result is significantly higher than that in ref 14: $P_{acc}$ = 95.1% and $R^2$ = 0.6837. The results of kerogen components, types, and maturity prove that the molecular structures constructed using the ML method can exhibit excellent performance for predicting the kerogen properties. Also, the prediction accuracy is also better than that of the previous ML models, which are applied to predict the single structural information.

Limited by the training dataset, our method can only exhibit excellent performance in the molecular model where the skeleton atom number is less than 50 (about 600 Da). The molecular scale is about 1/4−1/8 of the commonly used kerogen monomer molecules. However, it is believed that the ML method is the most promising method to completely solve the complex problems in kerogen model construction by far, and our work has laid a solid foundation for the construction of larger-scale models in the future.

## 4. CONCLUSIONS

In summary, we propose an intelligent high-throughput reverse construction method of kerogen molecular models. The kerogen models can be constructed using the trained ML model with combined experimental data directly. Neither the manual analysis of experimental spectra nor the enormous trial-and-error process is required. Thus, this study will save much time and materials in the fields involving the reverse construction of molecules and accelerate the kerogen ripening mechanism research.

The 2D spectral combination input method is designed in this study. The different types of spectra can be comprehensively analyzed, and the method is verified using the $^1$H NMR and $^{13}$C NMR spectra. The performance beyond that of a single spectral input is achieved by the ML model trained with multi-spectra. Therefore, the more robust expansion capabilities and higher development potential are contained in our ML model. Combined with the multi-spectral input method, we annotate a sample dataset containing 650,000 groups of $^1$H NMR, $^{13}$C NMR, and their corresponding labels independently. The ML model is trained by 550,000 group samples of the training dataset and 50,000 group samples of the validation dataset. The test results, in which 50,000 molecular models are constructed reversely, show that the average similarity is 91.78%, and the accuracy of the trained model is 82.51% with a similarity of $S_{Dice}$ > 80% and 54.78% with completely matched similarity ($S_{Dice}$ = 100%). Finally, 3000 kerogen molecules are used to verify the prediction accuracy and the effectiveness of the trained model. The results exhibit that the prediction accuracy of kerogen components, types, and maturity indexes can be achieved at 92.4%−99.6%, and the $R^2$ coefficients are all over 0.934. The prediction accuracy and goodness-of-fit about kerogen components, types, and maturity indexes are superb and significantly improved than the previous ML models applied to predict the single structural information. Thus, the results prove the effectiveness and superb comprehensive performance of the ML method.

Overall, this work realizes the comprehensive analysis of different types of spectra, and excellent predictive ability is achieved. Our work proves the feasibility of constructing the kerogen structural models based on the experimental spectra using ML methods reversely. Although the constructed molecular scale is less than the commonly used kerogen monomer molecules, we believe that this research is an essential exploration of reverse construction of the kerogen molecular models from the experimental data and will shorten the research cycle and tremendously reduce costs in constructing kerogen models and predicting kerogen properties.

## AUTHOR INFORMATION


### Corresponding Author

**Ya-Pu Zhao** − *State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China; School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China;* orcid.org/0000-0001-9269-7404; Email: yzhao@imech.ac.cn

### Author

**Dongliang Kang** − *State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China; School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.energyfuels.2c00738


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was jointly supported by the National Natural Science Foundation of China (NSFC Grant Nos. 12032019, 11872363, and 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) Durand, B. *Kerogen: Insoluble Organic Matter from Sedimentary Rocks*; Editions Technip: Paris, 1980.

(2) Wang, H.; Chen, L.; Qu, Z.; Yin, Y.; Kang, Q.; Yu, B.; Tao, W.-Q. Modeling of multi-scale transport phenomena in shale gas production — A critical review. *Appl. Energy* **2020**, *262*, 114575.

(3) Qing, W.; Xinmin, W.; Shuo, P. Study on the structure, pyrolysis kinetics, gas release, reaction mechanism, and pathways of Fushun oil shale and kerogen in China. *Fuel Process. Technol.* **2022**, *225*, 107058.

(4) Vandenbroucke, M.; Largeau, C. Kerogen origin, evolution and structure. *Org. Geochem.* **2007**, *38*, 719−833.

(5) Radke, M.; Welte, D. H.; Willsch, H. Maturity parameters based on aromatic hydrocarbons: Influence of the organic matter type. *Org. Geochem.* **1986**, *10*, 51−63.

(6) Zhang, T.; Ellis, G. S.; Ruppel, S. C.; Milliken, K.; Yang, R. Effect of organic-matter type and thermal maturity on methane adsorption in shale-gas systems. *Org. Geochem.* **2012**, *47*, 120−131.

(7) Wang, X.; Huang, X.; Gao, M.; Zhao, Y.-P. Mechanical response of kerogen at high strain rates. *Int. J. Impact Eng.* **2021**, *155*, 103905.

(8) Wu, T.; Firoozabadi, A. Effect of microstructural flexibility on methane flow in kerogen matrix by molecular dynamics simulations. *J. Phys. Chem. C* **2019**, *123*, 10874−10880.

(9) Xu, H.; Yu, H.; Fan, J.; Xia, J.; Wang, F.; Wu, H. Enhanced gas recovery in kerogen pyrolytic pore network: Molecular simulations and theoretical analysis. *Energy Fuels* **2021**, *35*, 2253−2267.

(10) Van Krevelen, D. W. *Coal: Typology, Physics, Chemistry, Constitution*; Elsevier: Amsterdam, 1993.
(11) Burnham, A. K. Kinetic models of vitrinite, kerogen, and bitumen reflectance. *Org. Geochem.* **2019**, *131*, 50−59.
(12) Burnham, A. K.; Sweeney, J. J. A chemical kinetic model of vitrinite maturation and reflectance. *Geochim. Cosmochim. Acta* **1989**, *53*, 2649−2657.
(13) 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*, 118264.
(14) Ma, J.; Kang, D.; Wang, X.; Zhao, Y.-P. Defining kerogen maturity from orbital hybridization by machine learning. *Fuel* **2022**, *310*, 122250.
(15) Huang, L.; Ning, Z.; Wang, Q.; Qi, R.; Zeng, Y.; Qin, H.; Ye, H.; Zhang, W. Molecular simulation of adsorption behaviors of methane, carbon dioxide and their mixtures on kerogen: effect of kerogen maturity and moisture content. *Fuel* **2018**, *211*, 159−172.
(16) 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.
(17) 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*, 9552−9562.
(18) Zhao, Y.-P. *Physical Mechanics of Surfaces and Interfaces*; Science Press: Beijing, 2012.
(19) Zhao, Y.-P. *Nano and Mesoscopic Mechanics*; Science Press: Beijing, 2014.
(20) Huang, L.; Ning, Z.; Wang, Q.; Zhang, W.; Cheng, Z.; Wu, X.; Qin, H. Effect of organic type and moisture on CO2/CH4 competitive adsorption in kerogen with implications for CO2 sequestration and enhanced CH4 recovery. *Appl. Energy* **2018**, *210*, 28−43.
(21) Collell, J.; Galliero, G.; Gouth, F.; Montel, F.; Pujol, M.; Ungerer, P.; Yiannourakou, M. Molecular simulation and modelisation of methane/ethane mixtures adsorption onto a microporous molecular model of kerogen under typical reservoir conditions. *Microporous Mesoporous Mater.* **2014**, *197*, 194−203.
(22) Qian, Y.; Zhan, J.-H.; Lai, D.; Li, M.; Liu, X.; Xu, G. Primary understanding of non-isothermal pyrolysis behavior for oil shale kerogen using reactive molecular dynamics simulation. *Int. J. Hydrogen Energy* **2016**, *41*, 12093−12100.
(23) Shi, G.; Kou, G.; Du, S.; Wei, Y.; Zhou, W.; Zhou, B.; Li, Q.; Wang, B.; Guo, H.; Lou, Q.; Li, T. What role would the pores related to brittle minerals play in the process of oil migration and oil-water two-phase imbibition? *Energy Rep.* **2020**, *6*, 1213−1223.
(24) Zeng, Y.; Du, S.; Zhang, X.; Zhang, B.; Liu, H. The crucial geometric distinctions of microfractures as the indispensable transportation channels in hydrocarbon-rich shale reservoir. *Energy Rep.* **2020**, *6*, 2056−2065.
(25) Du, S. H.; Shi, G. X.; Yue, X. J.; Kou, G.; Zhou, B.; Shi, Y. M. Imaging-based characterization of perthite in the upper triassic yanchang formation tight sandstone of the Ordos basin. *China. Acta Geol. Sin.* **2019**, *93*, 373−385.
(26) Burlingame, A. L.; Simoneit, B. R. High resolution mass spectrometry of Green River formation kerogen oxidations. *Nature* **1969**, *222*, 741−747.
(27) Dow, W. G. Kerogen studies and geological interpretations. *J. Geochem. Explor.* **1977**, *7*, 79−99.
(28) Bandurski, E. Structural similarities between oil-generating kerogens and petroleum asphaltenes. *Energy Sources* **1982**, *6*, 47−66.
(29) Young, D. K.; Yen, T. F. The nature of straight-chain aliphatic structures in green river kerogen. *Geochim. Cosmochim. Acta* **1977**, *41*, 1411−1417.
(30) Behar, F.; Vandenbroucke, M. Chemical modeling of kerogen. *Org. Geochem.* **1987**, *11*, 15−24.
(31) Siskin, M.; Scouten, C.; Rose, K.; Aczel, T.; Colgrove, S.; Pabst, R. Detailed structural characterization of the organic material in Rundle Ramsay Crossing and Green River oil shales. *Composition, Geochemistry and Conversion of Oil Shales*; Springer: Dordrecht, 1995; pp 143−158.
(32) Orendt, A. M.; Pimienta, I. S. O.; Badu, S. R.; Solum, M. S.; Pugmire, R. J.; Facelli, J. C.; Locke, D. R.; Chapman, K. W.; Chupas, P. J.; Winans, R. E. Three-dimensional structure of the Siskin Green River oil shale kerogen model: A comparison between calculated and observed properties. *Energy Fuels* **2013**, *27*, 702−710.
(33) 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.
(34) Wang, X.; Huang, X.; Lin, K.; Zhao, Y. P. The constructions and pyrolysis of 3D kerogen macromolecular models: Experiments and simulations. *Global Chall.* **2019**, *3*, 1900006.
(35) Tong, J.; Jiang, X.; Han, X.; Wang, X. Evaluation of the macromolecular structure of Huadian oil shale kerogen using molecular modeling. *Fuel* **2016**, *181*, 330−339.
(36) Liu, Y.; Liu, S.; Zhang, R.; Zhang, Y. The molecular model of Marcellus shale kerogen: Experimental characterization and structure reconstruction. *Int. J. Coal Geol.* **2021**, *246*, 103833.
(37) Bousige, C.; Ghimbeu, C. M.; Vix-Guterl, C.; Pomerantz, A. E.; Suleimenova, A.; Vaughan, G.; Garbarino, G.; Feygenson, M.; Wildgruber, C.; Ulm, F.-J.; Pellenq, R. J.-M.; Coasne, B. Realistic molecular model of kerogen's nanostructure. *Nat. Mater.* **2016**, *15*, 576−582.
(38) Schuster, P. Taming combinatorial explosion. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7678−7680.
(39) Zhou, J.; Zhang, J.; Yang, J.; Jin, Z.; Luo, K. H. Mechanisms for kerogen wettability transition from water-wet to CO2-wet: Implications for CO2 sequestration. *Chem. Eng. J.* **2022**, *428*, 132020.
(40) Faisal, H. M. N.; Katti, K. S.; Katti, D. R. An insight into quartz mineral interactions with kerogen in Green River oil shale. *Int. J. Coal Geol.* **2021**, *238*, 103729.
(41) Nan, Y.; Li, W.; Zhang, M.; Jin, Z. Ethanol blending to improve reverse micelle dispersity in supercritical CO2: A molecular dynamics study. *J. Phys. Chem. B* **2021**, *125*, 9621−9628.
(42) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.
(43) Zhao, Y.-P. *A Course in Rational Mechanics*; Science Press: Beijing, 2020.
(44) Brown, N. Chemoinformatics—An introduction for computer scientists. *ACM Comput. Surv.* **2009**, *41*, 1−38.
(45) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(46) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(47) Kang, D.; Wang, X.; Zheng, X.; Zhao, Y.-P. Predicting the components and types of kerogen in shale by combining machine learning with NMR spectra. *Fuel* **2021**, *290*, 120006.
(48) Kelemen, S. R.; Afeworki, M.; Gorbaty, M. L.; Sansone, M.; Kwiatek, P. J.; Walters, C. C.; Freund, H.; Siskin, M.; Bence, A. E.; Curry, D. J.; Solum, M.; Pugmire, R. J.; Vandenbroucke, M.; Leblond, M.; Behar, F. Direct characterization of kerogen by X-ray and solid-state 13C nuclear magnetic resonance methods. *Energy Fuels* **2007**, *21*, 1548−1561.
(49) Zweckstetter, M. NMR: Prediction of molecular alignment from structure using the PALES software. *Nat. Protoc.* **2008**, *3*, 679−690.
(50) Lille, Ü.; Heinmaa, I.; Pehk, T. Molecular model of Estonian kukersite kerogen evaluated by 13C MAS NMR spectra☆. *Fuel* **2003**, *82*, 799−804.
(51) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(52) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.


https://doi.org/10.1021/acs.energyfuels.2c00738
Energy Fuels 2022, 36, 5749−5761

(53) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(54) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons: New York, 2008.
(55) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* **2014**, *57*, 3186−3204.
(56) Riniker, S.; Landrum, G. A. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminf.* **2013**, *5*, 43.
(57) Rdkit: Open-source cheminformatics 2013. http://www.rdkit.org. (accessed 26 July, 2021).
(58) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Network.* **2015**, *61*, 85−117.
(59) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. **2014**, arXiv preprint:1412.6980.
(60) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.
(61) Elman, J. L. Finding structure in time. *Cognit. Sci.* **1990**, *14*, 179−211.
(62) Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D.; Wierstra, D. Draw: A recurrent neural network for image generation. *International Conference on Machine Learning*, 2015; pp 1462−1471.
(63) Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015; pp 1422−1432.
(64) Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to sequence learning with neural networks. **2014**, arXiv preprint:1409.3215.
(65) Xiong, W.; Wu, L.; Alleva, F.; Droppo, J.; Huang, X.; Stolcke, A. The Microsoft 2017 conversational speech recognition system. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*; ICASSP, 2018; pp 5934−5938.
(66) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; pp 770−778.
(67) He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. *European Conference on Computer Vision*, 2016; pp 630−645.
(68) Kingma, D. P.; Welling, M. Auto-encoding variational bayes. **2013**, arXiv preprint:1312.6114.
(69) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. **2015**, arXiv preprint:1509.09292.
(70) Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. **2014**, arXiv preprint:1409.1556.
(71) Zhang, H.; Yu, H.; Yuan, X.; Xu, H.; Micheal, M.; Zhang, J.; Shu, H.; Wang, G.; Wu, H. Permeability prediction of low-resolution porous media images using autoencoder-based convolutional neural network. *J. Pet. Sci. Eng.* **2022**, *208*, 109589.
(72) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10*, 1692−1701.
(73) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026−8037.
(74) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. Tensorflow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI'16)*, 2016; pp 265−283.
(75) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem substance and compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.
(76) Willcott, M. R. MestRe Nova. *J. Am. Chem. Soc.* **2009**, *131*, 13180.




https://doi.org/10.1021/acs.energyfuels.2c00738
*Energy Fuels* 2022, 36, 5749−5761