energy&fuels

pubs.acs.org/EF

Review

Perspectives of Machine Learning Development on Kerogen Molecular Model Reconstruction and Shale Oil/Gas Exploitation

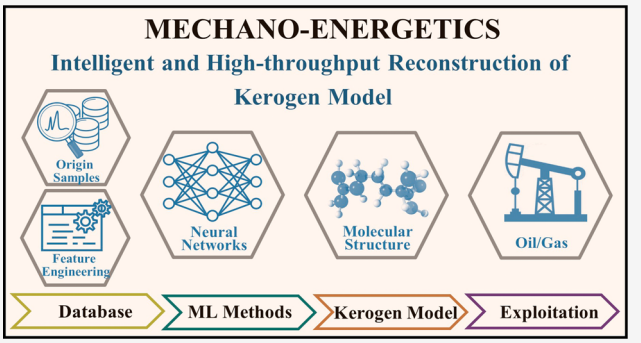
Dongliang Kang, Jun Ma, and Ya-Pu Zhao*

Cite This: *Energy Fuels* 2023, 37, 98−117

Read Online

ACCESS | Metrics & More | Article Recommendations

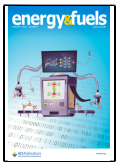
ABSTRACT: The shale revolution has provided abundant shale oil/gas resources for the world, but the efficient, sustainable, and environmentally friendly exploitation of shale oil/gas is still challenging. Kerogen is the primary hydrocarbon source of shale oil/gas. The research on the kerogen chemo-mechanical properties significantly influences the development of shale oil/gas extraction technology. Rapid reconstruction of the kerogen molecular models is the most effective way to study the generation mechanism of shale oil/gas from the bottom-up molecular level. However, due to the combinatorial explosion problem, the reconstruction complexity of kerogen increases sharply because of the kerogen's characteristics of complex origin, large molecular weight, and diverse functional groups. The traditional kerogen molecular reconstruction methods require professionals to comprehensively analyze various experimental information to approximate the actual kerogen molecular models through trial-and-error. So, the traditional methods are time and material-consuming and extremely inefficient. These shortcomings make researchers spend too much strength on the reconstruction of kerogen molecular models and cannot focus on the study of kerogen chemo-mechanical properties. For the past few years, state-of-the-art machine learning (ML) methods have been applied to intelligently reconstruct the kerogen molecular models through high-throughput and predict shale oil/gas production mechanisms. Although the current work is still in the infancy stage, ML methods are believed to be the most promising way to solve the drawbacks of traditional methods and reconstruct kerogen in reliable and large molecular weight. Hence, mechano-energetics is proposed to study the efficient development and utilization of energy based on mechanics and ML. This paper briefly reviews the development history of kerogen molecular model reconstruction methods and the research of ML in the fields of kerogen reconstruction and shale oil/gas exploitation. Some recommendations for further ML-based work are also suggested. We are convinced that the ML methods will accelerate the research of kerogen and promote the significant development of unconventional oil/gas exploitation technologies.



1. INTRODUCTION

With the development of horizontal drilling and hydraulic fracturing technologies, the shale revolution occurred in the United States. The shale oil/gas can be extracted from previously inaccessible reservoirs and play an essential role in the field of energy.[1−4] However, the sustainability and environmental risks of shale oil/gas have prompted researchers to develop new extraction technologies continuously.[5−7] Kerogen is the insoluble macromolecular organic matter in sedimentary rocks, formed by the degradation of ancient algae, plankton, higher plants, etc., through geological sedimentary diagenesis. As the parent material of shale oil/gas, more than half of the hydrocarbons in shale are adsorbed in kerogen. Therefore, research on the chemo-mechanical properties of kerogen, such as oil/gas migration, maturation, in situ ripening, pyrolysis, etc., are the basis for increasing the production and extraction efficiency of shale oil/gas.[8−10] And the information on the environment, climate, and biota in ancient geological times can also be provided by kerogen.[11−13] Bottom-up simulation analysis of kerogen from the molecular level is the most direct and effective way.[14,15] Thus, the rapid reconstruction of qualified kerogen molecular models is the cornerstone of the research on the chemo-mechanical properties of kerogen.[16] Since the 1940s, research on the kerogen formation, molecular model reconstruction, and pyrolysis evolution mechanism is out to predict the sweet spots distribution and production of the reservoirs, then achieve the in situ ripening, low-cost extraction, and environ-

Received: September 30, 2022
Revised: November 19, 2022
Published: December 9, 2022



mental protection of shale oil/gas. But there is an intractable problem because of the combinatorial explosion in the molecular structure reconstruction.[17] The obstacle of molecular reconstruction will be sharply increased with molecular weight. Kerogen has the complex origin, significant molecular weight, and various types of functional groups. The problem of the combinatorial explosion is particularly prominent. So, the combinatorial explosion problem is the essential reason for the tedious of reconstructing kerogen molecular models.

Researchers have almost exhausted all the molecular structural experimental measurement methods to detect the molecular structure of kerogen. By using element analysis (EA), nuclear magnetic resonance (NMR), Fourier transform infrared (FTIR) spectroscopy, X-ray photoelectron spectroscopy (XPS), etc., the information on chemical elements, functional groups is analyzed and then used to reconstruct the entire kerogen structural models. With the advancement of experimental technology, the determination of the molecular structure information on kerogen is more and more accurate. The three-dimensional (3D) kerogen molecular models are reconstructed with molecular dynamics (MD) simulations. However, the traditional reconstruction methods need to adjust the molecular structures repeatedly based on the experiment/simulation to approach the actual molecular structures. There are two inherent drawbacks to the traditional trial-and-error reconstruction methods. First, the experimental data require a comprehensive analysis by experienced professionals, which is not conducive to engineering promotion and application. Second, tremendous time and material resources are consumed during the repeated trial-and-error process, and the reconstruction efficiency will be extremely low. The two inherent drawbacks have greatly limited the progress of studying kerogen chemo-mechanical properties from the molecular level. Consequently, it is imperative to develop an intelligent kerogen reconstruction method and liberate researchers from the predicament of kerogen molecular models.[18]

Recent years have witnessed the rapid blossom of machine learning (ML). And ML-based methods have achieved remarkable success in geological exploration, medical health, natural language processing, and so forth.[19−24] Artificial intelligence ML methods have powerful analysis capability of big data and can adapt to various high-complexity problems. Therefore, ML technology is a promising way to address the problem of intelligent and high-throughput reconstruction of kerogen molecular models. Researchers have developed some ML-based methods to predict the kerogen molecular models and properties. But the relevant work is still in the infancy stage. There are still many tasks that should be further solved. Zhao proposes the concept of "mechano-energetics" to address the further challenges for shale oil/gas exploration (Figure 1). The mechano-energetics is coined to study the efficient development and utilization of energy based on mechanics, coupling with a force field, radiation field, temperature field, and electric field by theory, experiment, simulation, and artificial intelligence methods.[25] This review briefly introduces the fundamental importance of kerogen molecular models for shale oil/gas research, the development history of kerogen molecular model reconstruction methods, and the application of ML methods in kerogen molecular model reconstruction and shale oil/gas production. Finally, some challenges and directions of developing ML to reconstruct molecular models of the kerogen macromolecule are also summarized.

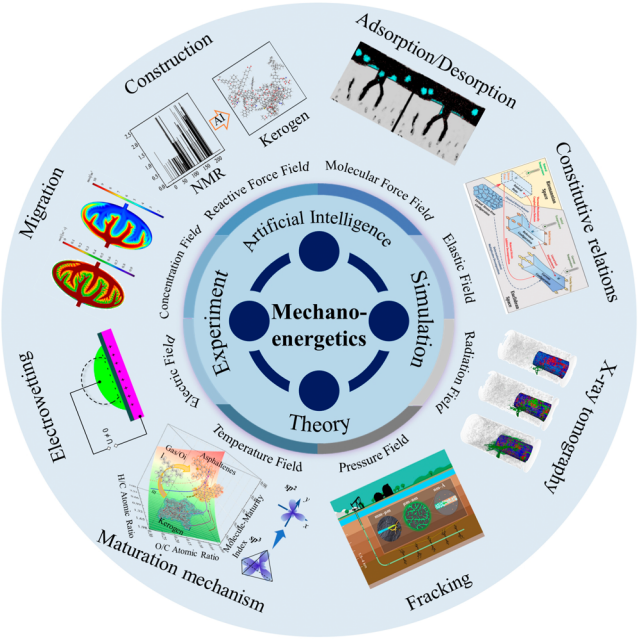


Figure 1. Schematic of mechano-energetics. It is coined to study the efficient development and utilization of energy based on mechanics, coupling with a force field, radiation field, temperature field, and electric field by theory, experiment, simulation, and artificial intelligence methods.[25] Reproduced with permission from ref 25. Copyright 2021 Springer Nature.

2. NECESSITY OF KEROGEN MOLECULAR MODEL

Kerogen was named by Alexander Crum Brown in 1906 to describe the substance that can produce waxy oil from Scottish oil shale. In Greek, "keros" means "waxy oil", while "-gen" means "birth".[26] According to the formation sources, kerogen can be divided into three types: lacustrine shale, marine shale, and terrestrial shale, is the most abundant form of organic matter on the earth.[27] The carbon in the form of kerogen on the earth is about 10^{16} tons, compared to only 10^{12} tons in living organisms.[28] With the change of pressure and temperature during deposition, the structures of kerogen will undergo degradation, isomerization, aromatization, etc., accompanied by the generation of oil/gas. The mechanical and chemical properties of kerogen also change, such as chemical structure, molecular density, maturity, and organic pore distribution. Therefore, kerogen structure models play an essential role in studying the mechanism of shale oil/gas generation and efficient exploitation, which is attractive to various fields such as geology and petroleum. Further research on kerogen's mechanical and chemical properties can effectively promote the development of in situ ripening of oil shale, optimization of pore network, and oil/gas migration technologies.[29]

2.1. Formation of Kerogen. It is widely recognized that kerogen is formed due to the geological evolution of ancient algae, plant, and animal remains in the anaerobic sedimentary environment. The chemical structures of kerogen vary significantly with the precursor and sedimentary environment. Generally, the biological organic matter preserved in sedimentary rocks accounts for only about 0.1−1%, and 90% of the source rocks are contained in strata with the warm climate and high water table. There are four explanations for kerogen formation: selective preservation, degradation-recondensation, natural sulphurization, and sorptive protection.[30]

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

Table 1. Comparison of Various Kerogen Maturity Models

model	expression[a]	advantages	limitations
Vitrimat[51,52]	%Ro = 12 exp [−3.2(H/C)] − 1.2(O/C)	easy to be calculated with initial composition	not a rigorous mechanistic model
MMI[53]	MMI = 1/(1 + H/C + O/C N/C + S/C)	based on the change in molecular structure	cannot explain functional groups' evolution
	OrbHMI = $1/(2.85 - 1.1r_C + 0.1r_O)$		
OrbHMI[54]	$r_C = C_{sp^2}/(C_{sp^2} + C_{sp^3})$	closer to the physical bottom mechanism	challenging to obtain hybridization information
	$r_O = O_{sp^2}/(O_{sp^2} + O_{sp^3})$		
Easy %Ro[55]	%Ro = exp (−1.6 + 3.7F)	can compute vitrinite maturation with time and temperature	unsuitable for experiments with short heating times
Basin %Ro[56]	%Ro = %Roo exp (3.7$F(t)$)	uncertainty of thermal reconstruction is considered	insensitive to inconsistent calibration data

[a]F is the fraction of reactant converted; H/C is the hydrogen/carbon atomic ratio; O/C oxygen/carbon atomic ratio; N/C is the nitrogen/carbon atomic ratio; S/C is the sulfur/carbon atomic ratio; % Roo is the vitrinite reflectance of immature vitrinite; $F(t)$ is a function of time; C_{sp^2}, C_{sp^3} are sp^2, sp^3 hybridized carbons, respectively; and O_{sp^2}, O_{sp^3} are sp^2, sp^3 hybridized oxygens, respectively.

Selective preservation assumes that kerogen is the compound with anti-degradation properties in ancient organisms, such as microbial cell walls. And the anti-degradation organic matter is selectively preserved and enriched during geological evolution.[31] Morphological observations of the kerogen microstructure provide direct evidence for selective preservation. Tegelaar and co-workers observed the biological structures (debris of spores, pollen, plant, etc.) contained in kerogen through transmission electron microscopy. Compared with the biomacromolecular structures, they concluded that kerogen can be formed from a small amount of specific, insoluble, nonhydrolyzable, and resistant degradation macromolecular structures during sedimentary diagenesis. The corresponding relationships between the initial and expected kerogen molecular structures are given. Although the initial content of these macromolecular structures is low, it can be increased by two to three orders of magnitude during enrichment.[32] The homogeneity of the carbon isotope analysis of kerogen pyrolysis products also supports this opinion.[33] However, selective preservation is lacking in identifying amorphous components of kerogen.

The degradation-recondensation assumes that the organic matter is decomposed into small molecular structures (monosaccharides, amino acids) during evolution. The tiny units are recondensed as the precursors of humic substances and kerogen: melanoidins.[34] It is generally believed that degradation-recondensation is responsible for the amorphous aliphatic structure of kerogen. But why degradable organic molecules are preserved has puzzled researchers until the sorptive protection of minerals is discovered.[35] In the pyrolysis study of North American kerogen, the products of kerogen exhibit high aliphatic and phenolic characteristics. These substances are not found in plant structures and come from the polymerization and transformation of animal tissues.[36] Poirier and co-workers analyzed the refractory organic matter of an ancient soil near Pointe Noire in Congo and found that the degradation of organic matter in deep soil under natural conditions is significantly different from that in the laboratory.[37] Unstable organic molecules are adsorbed on minerals and retained during degradation, then undergo subsequent condensation reactions to form kerogen.[38−40]

Natural sulphurization refers to the reaction between the inorganic sulfur element and the biomolecules in the early diagenesis stage. This opinion explains the source of the sulfur element in sulfur-rich kerogen.[41,42] The structure and content of sulfur in macromolecular organic matter such as kerogen can be determined by flash pyrolysis. It is speculated that the sulfur moieties in kerogen come from the abiogenic sulfur.[43] The process of incorporating inorganic sulfur into organic matter is affected by the reactive iron element in the sedimentary environment. It is generally believed that the reactivity between iron and sulfur is higher than the organic matter.[44] Therefore, in the environment where the reducing sulfur content is higher than the active iron, organic molecules can extensively react with inorganic sulfur to generate such sulfur-rich macromolecular organic matter. Most kerogen with high sulfur content occurs in marine sedimentary environments that are rich inorganic sulfur but less in lake environments.[45] In addition, the polysulfides are generated during the reduction of iron hydroxides and react with organic matter. So, some organic matter of lacustrine mines may also be rich in sulfur.[46]

2.2. Structural Transformation during Kerogen Maturation. Since the oil/gas production properties of kerogen attracted researchers in the 1940s, the maturation and pyrolysis mechanisms of kerogen have been the focus of petroleum.[47−50] It is necessary to explore kerogen structure to promote oil/gas production. The kerogen structure is also the basis for studying kerogen origin, type, and maturity. The maturity and type are important indicators of kerogen. Combining the maturity and type of kerogen, the development stage and the oil/gas extraction potential of the reservoir can be analyzed. The commonly used kerogen maturity models are summarized in Table 1. The Vitrimat was established based on the analysis of a large number of experimental data.[51,52] The maturity can be calculated by the atomic ratio of O/C and H/C in the kerogen molecular structure. The Vitrimat maturity index is extremely terse, which is beneficial for application in engineering. Wang et al. proposed the kerogen molecular maturity index (MMI) through thermal evolution experiments, then established a dynamic model of kerogen thermal maturity evolution through MMI.[53] The MMI is positively correlated with the vitrinite reflectance %Ro and can accurately reflect the loss rate of kerogen weight during thermal evolution. Ma et al. suggested the orbital hybridization maturity index (OrbHMI) based on the hybrid orbital of atoms in the kerogen molecular model.[54] Compared with other maturity models, OrbHMI can be more helpful in understanding the underlying mechanism of kerogen maturity evolution. Thus, maturity is closely related to the kerogen molecular structure and can be applied to guide kerogen ripening studies.

During the burial, immature kerogen would be recombined into more stable structures with temperature and pressure changes. In the early stage of diagenesis, the original molecule removes N, S, and O heteroatom functional group structures.

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

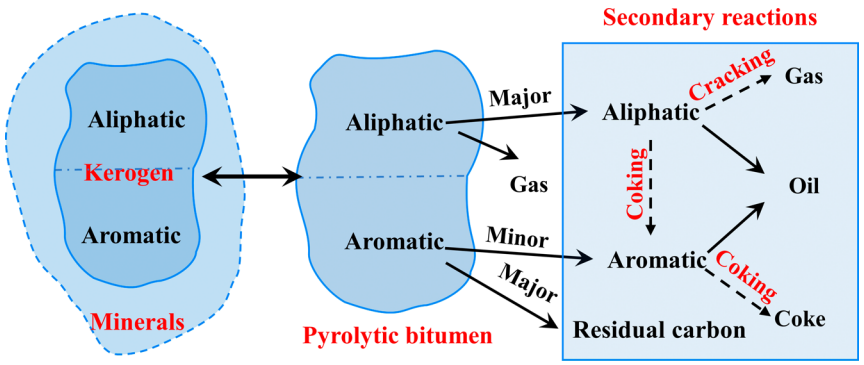


Figure 2. Carbon transformation during the process of kerogen pyrolysis.[62] Reproduced with permission from ref 62. Copyright 2017 Elsevier.

After that, aliphatic carbon structures are stripped or transformed during the catagenesis stage.[57] This is also the main stage of oil/gas generation, and the maturity of kerogen gradually increases. Finally, the aromatic ring structures evolve into larger, closer-packed groups during metagenesis.[11] Hou and co-workers[58] showed that lacustrine Type-II kerogen removed O-functional groups and short aliphatic chains when %Ro was below 0.6. The peak of hydrocarbon production is reached in the maturity range of 0.6 < %Ro < 1.09, after which the highly aromatic kerogen residue structure is formed.

In 1950, Hubbard et al.[59] heated the Colorado oil shale from 623 to 798 K under anaerobic conditions. They proposed the empirical mechanism of kerogen pyrolysis products. Kerogen produces oil/gas in the form of eq 1 during pyrolysis.

$$\text{kerogen} \rightarrow \text{bitumen} \rightarrow \text{oil} + \text{gas} + \text{char} \quad (1)$$

In subsequent studies, the oil/gas types and yield were established by multistage pyrolysis of oil shale kerogen.[60] On the basis of hydrolytic reactivity in hot water an explanation has been proposed for the formation and decomposition mechanisms of natural kerogen.[61] And a comprehensive pyrolysis mechanism is mentioned based on analyzing the structural transformation of products during pyrolysis. In the opinion of Lai et al., the oil/gas products in the pyrolysis process are mainly derived from aliphatic structures, while the aromatic structures are converted to carbon residue. As is shown in Figure 2, the partial aliphatic carbon chains in kerogen are cracked into small fragments, then aromatized and condensed with the original aromatic carbon skeleton. After the first-order reaction, the residual aliphatic and partial aromatic carbon experienced coking and cracking. The mechanism systematically described the chemical structure transformation process of oil shale pyrolysis at the molecular level.[62]

The above kerogen pyrolysis mechanism is only the empirical summary based on experiments. Since the thermal transformation process of kerogen involves several reactions, it is troublesome to observe the details through experiments. Therefore, the pyrolysis mechanism of kerogen needs to be elucidated by simulation based on the molecular model. The pyrolysis order of the kerogen structure is related to the dissociation energy of local bonds.[63] The density functional theory (DFT) was used to analyze the order of breaking chemical bonds in the molecular structure of kerogen. The result is exhibited in eq 2:

$$\text{C—S} < \text{C—N} < \text{C—O} < \text{C—C} < \text{C—H} < \text{C=C} \quad (2)$$

The pyrolysis behavior was simulated at 300−3000K based on the Siskin Green River Type-I kerogen molecular model.[64] With the temperature increasing, the pyrolysis process is divided into three stages: cleavage of weak bonds, generation of long-chain hydrocarbons, and formation of hydrogen-containing gases.[65] Wang et al. studied the pyrolysis mechanism of the Erdos Type-III and Songliao Type-I kerogens with the hybrid molecular dynamics/force-biased Monte Carlo method. And the bond dissociation energy in the molecular structure was found to be the main reason affecting the pyrolysis products. As for the Type-III kerogen, the products are less at low temperatures, and the production of methane (CH_4) increases at high temperatures.[66] Furthermore, the kinetic equation expressing the relationship between activation energy and maturity is established based on molecular pyrolysis simulation. Then the maturation mechanism of kerogen at different temperatures can be clarified.[53] Thus, the pyrolysis simulation with the kerogen model plays a crucial role in studying the intrinsical mechanism of oil/gas production at the molecular level, further illustrating the necessity of the kerogen molecular model in shale oil/gas research. Recently, the ML methods are also proposed to reduce the computational complexity of pyrolysis simulations and estimate the product yields.[67]

2.3. Kerogen Structure and Mechanical Properties.

The structural changes during maturation directly affect the mechanical properties of kerogen, such as microscopic pores and fracture mechanisms, and play a vital role in establishing oil/gas transport channels in shale oil/gas reservoirs.[68−73] However, kerogen exists at the nanoscale, and the mechanical properties through experiments are challenging to be measured. Therefore, molecular simulation is widely used to the mechanical properties of kerogen. It is significant in guiding oil/gas extraction and is one of the key directions of kerogen research.[74−76]

There are a large number of organic pores in the kerogen, and the pores form the channel for oil/gas migration during evolution.[77] However, due to the adsorption, many oil/gas molecules are confined in the pores, obstructing the oil/gas migration fracture network.[78−82] Zhu et al. explained the influence of size and curvature effects of nanopores on CH_4 adsorption, then established the state equation of CH_4 adsorption phase based on the simulation.[83] Lin et al. proposed the CH_4 displacement kinetic equation by investigating the angle of nitrogen (N_2), water (H_2O), and carbon dioxide (CO_2) in displacing adsorbed CH_4.[84] The displacement mechanism of CO_2, N_2, and H_2O was determined by studying entropy and enthalpy. The results

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

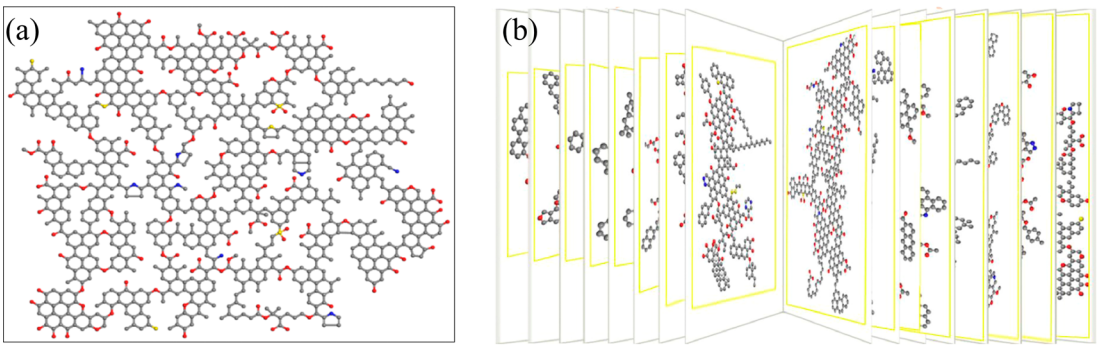


Figure 3. Sample of (a) kerogen monomer[66] and (b) kerogen matrix.[95] Panel b was reproduced with permission from ref 95. Copyright 2021 Elsevier.

show that CO_2 is driven by both entropy and enthalpy, and N_2 is driven by entropy. Moreover, H_2O is driven by reducing the partial pressure of CH_4. The mechanism explains the distinctions in displacement efficiency in different supercritical fluids.[85] An in situ adsorption simulation of CH_4 is carried out to study the adsorption state of shale oil/gas under natural geological conditions based on the scanning electron microscope (SEM) images of shale pore structure. The mechanism of excess adsorption isotherm crossing and desorption hysteresis is clarified.[86] Research on shale oil/gas adsorption properties has been carried out with the development of kerogen molecular models. The adsorption of shale oil in kerogen slits was simulated by MD. It was found that the adsorption capacity of heavy components (e.g., asphaltene) was greater than that of light components. And the absorption density of hydrocarbons reached a maximum value of 2 nm aperture.[87] With increasing water content, the pores of kerogen matrix volume and CH_4 desorption amount decreased based on the Ungerer kerogen model. It is not conducive to the transportation of oil/gas in kerogen.[88]

The pore network formation and characteristics of kerogen under confined volume reservoir conditions are studied through reactive force field-molecular dynamics (ReaxFF-MD) simulation, and the features of pore networks are affected by kerogen maturity and pyrolysis temperature.[89] The establishment and maintenance of shale oil/gas migration channels in microscopic pores are influenced by the kerogen mechanical behavior. So, the research of kerogen mechanical properties is widely concerned. Zeszotarski and co-workers measured the hardness and reduced modulus of Woodford Type-II kerogen by atomic force microscope and nanoindenter. They concluded that the isotropic mechanical behavior is exhibited in kerogen. The hardness is about 550 MPa higher than that of common polymers (about 200 MPa). Since there is no Poisson's ratio for kerogen, Zeszotarski only obtained an indentation modulus of about 10−11 GPa. And they proved that kerogen has both viscoelasticity and plasticity.[90] Jakob et al. demonstrated that the aromaticity of kerogen is positively correlated with the local mechanical modulus of the surrounding inorganic matrix by peak force infrared microscopy. Their work improves the understanding of the effect of kerogen heterogeneity on the mechanical properties of source rock.[91]

Currently, the research on the mechanical properties of kerogen with experiments can only be roughly estimated. The detailed characteristics of kerogen under stress can only be calculated on the basis of the kerogen molecular models. The structure and mechanical properties of 3D kerogen molecules during pyrolysis were studied by Spiro. They thought aliphatic molecular function groups with weak covalent bonds are first decomposed. Then the aliphatic fragments and adsorbed oil/gas molecules are lubricated in the planar aromatic molecular groups. Finally, the molecular cross-linking and intermolecular van der Waals interactions are inhibited, resulting in the thermoplastic behavior of kerogen.[92] Bousige et al. applied MD to calculate kerogen's bulk, shear, and elastic modulus with different maturity. They observed an exponential increase of these mechanical parameters with kerogen density, and the fracture behavior is determined by the sp^2/sp^3 ratio.[93] The elastic modulus increases with the increase of pressure through molecular simulation of the kerogen model. The Mohr−Coulomb failure criterion and tensile strength criterion are used to describe the fracture behavior of the kerogen matrix.[94] Wang et al. obtained the stress−strain dynamic response curves under different strain rates via tensile simulation. And the hyper-viscoelastic constitutive model is established to describe the mechanical behavior of kerogen.[95] The extensive application of molecular simulation in the study of the kerogen chemo-mechanical mechanism further illustrates the necessity of the kerogen molecular model in shale oil/gas research.

3. TRADITIONAL RECONSTRUCTION METHODS OF THE KEROGEN MODEL

The determination experiment of kerogen is the basis for analyzing the macroscopic and microscopic structure, exploring the components, and constructing the molecular models. Due to the high compositional complexity, there is no fixed molecular structure of kerogen and no repeating simple molecular units like polymers. Almost all the experimental methods have been used to characterize macroscopic and microscopic structural characteristics. The methods include transmitted and fluorescent light microscopy, SEM, EA, pyrolysis-gas chromatography−mass spectrometry (Py-GC/MS), FTIR, and NMR experiments.[96−98] Furthermore, kerogen structural models based on experiments and statistics are constructed to study kerogen's mechanical properties and transport network. Therefore, the kerogen model can effectively promote the production research of shale oil/gas.

3.1. Kerogen Monomer and Kerogen Matrix. Generally, the reconstruction methods of kerogen structural models have undergone a development process from rough to detailed. And the reconstructed kerogen models are from skeleton to complete structure, from two-dimensional (2D) to 3D. Dozens of various mining areas' kerogen molecular models have been reconstructed in different ways up to now. In accordance with the reconstruction ideas, these models can be roughly divided into two categories: kerogen monomer[99] and kerogen matrix (a group of kerogen molecules).[87] The kerogen monomer is based on the assumption that kerogen is formed from the self-polymerization of molecular structures. So, a kerogen monomer is

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

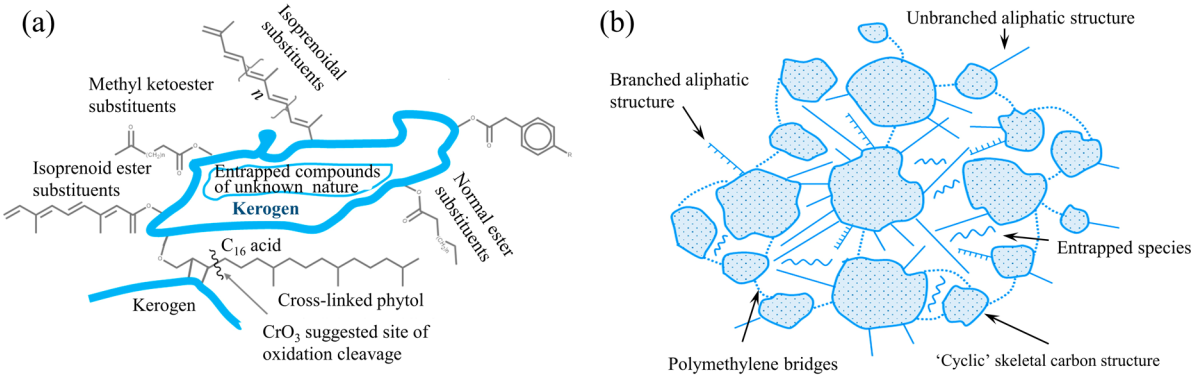


Figure 4. Initial two kerogen molecular structure conjecture models of the Green River formation. (a) Schematic diagram of the kerogen structure of the Green River shale inferred from the oxidative cracking of CrO_3 by Burlingame et al.[103] (b) Schematic diagram of the 3D kerogen structure conjecture model of the Green River formation is given by Young et al.[106] Panel a was reproduced with permission from ref 103. Copyright 1969 Elsevier. Panel b was reproduced with permission from ref 106. Copyright 1977 Elsevier.

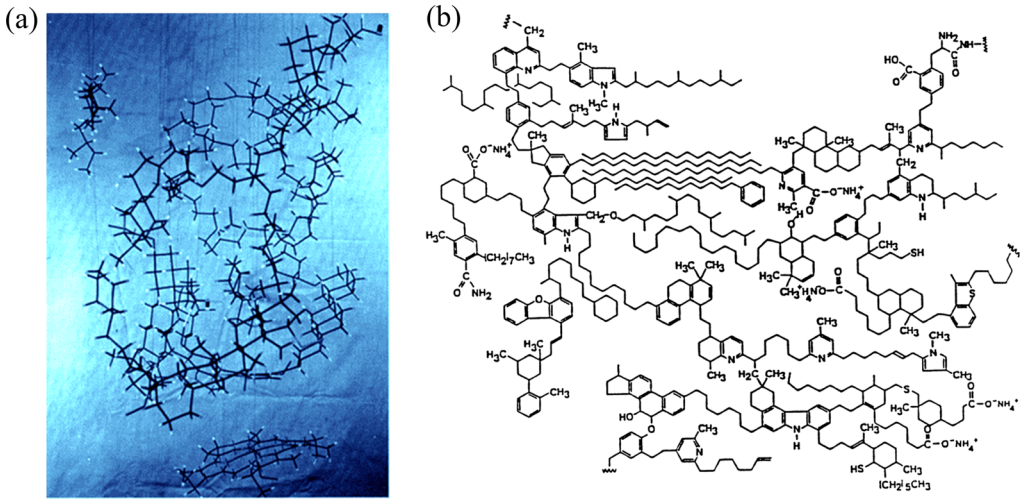


Figure 5. Molecular models of the Green River Formation kerogen with detailed structure. (a) Molecular model constructed by Yen.[107] (b) Molecular model constructed by Siskin.[100] Panel a was reproduced with permission from ref 107. Copyright 1976 Elsevier. Panel b was reproduced with permission from ref 100. Copyright 1995 Springer Nature.

an average structure of a batch of kerogen molecular components measured simultaneously in the experiment and can represent the mechanical and chemical properties of the whole kerogen (Figure 3a). The kerogen matrix points out that kerogen is aggregated from many molecules of different scales (Figure 3b). This opinion is relatively consistent with the actual situation in the natural environment. Both methods try to contain more functional groups while reconstructing the kerogen structural model to make the kerogen models have more stable statistical characteristics. In this way, the molecular models can better reflect kerogen's properties through MD simulation.

Both methods have their own pros and cons, and they have been used in reconstructing the kerogen molecular models of mining areas until now. The benefit of the kerogen monomer is that it can better characterize the polymer properties of kerogen due to its considerable molecular weight. Nevertheless, the combinatorial explosion problem leads to an exponential increase in construction intricacy with increasing molecular weight. This is also the fundamental reason for the challenge of constructing a kerogen structure model with a significant molecular weight. So, the methods of the kerogen monomer always require lots of labor and material resources. The giant Erods Type-III kerogen monomer molecular model ($C_{861}H_{750}O_{124}N_8S_8$) constructed by Wang et al. based on experiments is a typical representative of this method.[66] Unlike kerogen monomer models, the kerogen matrix aims to represent true kerogen through a group of interacting multimolecular mixtures rather than sizable molecular weight. Thus, it is only necessary to construct a series of small kerogen molecular structures based on the experimental data. These kerogen submolecules are relatively easily reconstructed but cannot effectively reflect the polymer properties of kerogen. The most famous kerogen matrix is the Green River kerogen model created by Siskin et al.[100] The chemical formula of the Sisken kerogen model is $C_{645}H_{1017}O_{17}N_{19}S_4$. The smallest molecule in the group is $C_{18}H_{30}$, while the largest is $C_{367}H_{547}O_{10}S_2$. In fact, it is unrealistic and unnecessary to construct a molecular model that is entirely consistent with the real kerogen. Depending on the research goals, several properties such as elemental/molecular composition and pyrolysis products are traded off while reconstructing a kerogen molecular model. Consider the physical/chemical information comprehensively to build a molecular model that meets the research needs.[27]

3.2. Qualitative Reconstruction Methods of the Kerogen Skeleton Model. The study of the kerogen molecular model draws on the reconstruction method of coal in the early stage.[101] In fact, many of the kerogen chemo-mechanical properties testing and characterization methods are based on previous studies on coal minerals, such as the classification of kerogen. Initially, the kerogen molecular model is constructed by analyzing residues, products, and small molecule extracts (e.g., bitumen) in source rocks during degradation. The residue molecules are used as a kerogen skeleton, and the low molecular weight hydrocarbon products are integrated into the skeleton to reconstruct cross-linked macromolecular structures.[102] With the advancement of experimental technology, the detection accuracy of unknown molecular structures has become

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

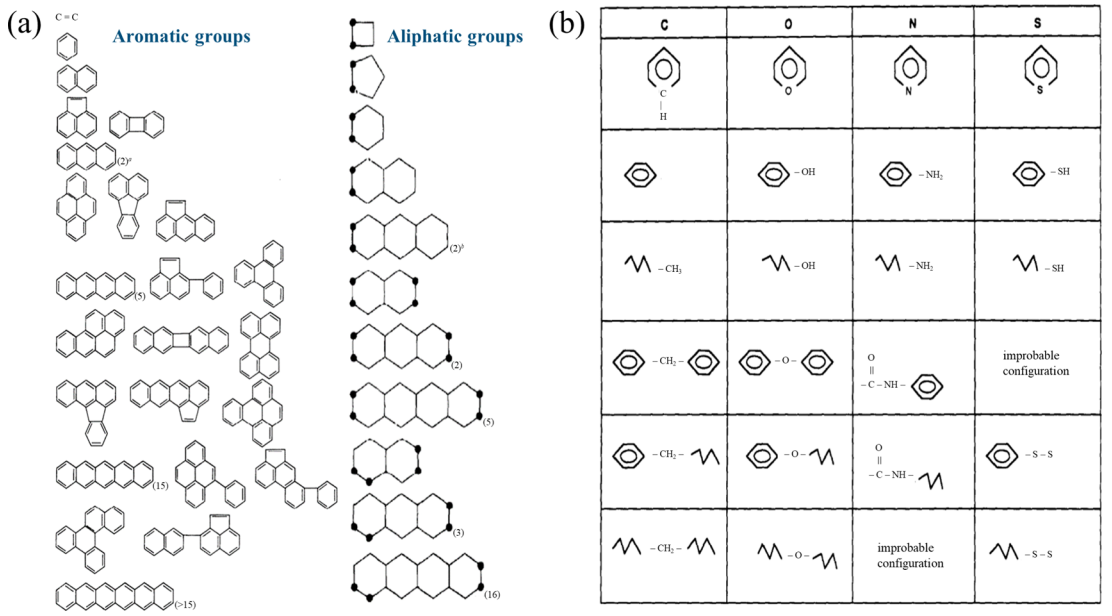


Figure 6. Part of the basic structural units in (a) CAMSC[108] and (b) Xmol.[109] Panel a was reproduced with permission from ref 108. Copyright 1977 Elsevier. Panel b was reproduced with permission from ref 109. Copyright 1990 Elsevier.

higher, but the way of constructing kerogen macromolecules is still used today.

In 1969, Burlingame et al. combined high-resolution mass spectrometry with organic debris in shale, and proposed the possible cross-linking model of Green River kerogen skeleton.[103] The content and form of organic fatty acids in Green River kerogen were determined. They deduced that these organic matters are connected to the kerogen skeleton as easily degradable functional groups.[104] Subsequently, the degradation products of chromic acid oxidation via high-resolution mass spectrometry are determined. The result exhibits that the Green River kerogen is a highly cross-linked polymer composed of random C_4−C_{25} aromatic nucleus, heteroatoms, and long aliphatic side-chains.[105] The kerogen molecular models of the Green River formation are shown in Figure 4a. Although Burlingame's kerogen model is rough without specific molecular structures, it indicates kerogen composition and specific cross-linking form. Further research on kerogen molecular models is mainly carried out according to the reconstruction idea of this model. Young and Yen prevented the generation of intermediates during the degradation of kerogen by mild stepwise oxidation. They applied gas−liquid chromatography, gas chromatography−mass spectrometry, and proton-NMR experiments to determine the straight-chain aliphatic structures in Green River kerogen. On the basis of this, the conjectured model of the Green River kerogen is given as Figure 4b.[106] Compared with the aliphatic structure polymerization model, this model focuses on cyclic organic molecular groups, such as fused-ring aliphatic and aromatic structures. A large number of branched structures during the oxidation are attached to the "core" of the kerogen model. In addition, Young's model also considers the 3D structural information of kerogen.

Yen constructed a kerogen skeleton model of the Green River formation based on the molecular structure of coal tar pitches. It is a 3D kerogen molecular model with the detailed structure containing 20 small molecules. The largest molecular structure is $C_{32}H_{58}$ and the smallest is CO_2 (Figure 5a). Yen's model explains the structure of chemically degraded compounds and matches the results of chemical analysis such as X-ray diffraction, FTIR, etc.[107] Meanwhile, this work indicated that some free asphaltic molecules may be contained in the kerogen. In addition to covalent bonds, hydrogen bonds also exist between molecules, which cause folding cross-links in the kerogen model. The solid-state NMR is used to analyze the molecular structure of kerogen without degradation. The elements and functional groups are determined via ^{13}C and ^{29}Si NMR spectra. Siskin et al. then combined mass spectrometry with the pyrolysis products to construct the final kerogen molecular model (Figure 5b). The kerogen molecular models of Green River oil shale in the United States and Rundle Ramsay Crossing oil shale in Australia are successfully constructed by this method.[100] Both models fit well with the experimental results on elemental components, aromaticity, chain length distribution, hydrocarbon/heteroatom functional groups, etc. The advantage of the Siskin method is innovatively applying the solid-state NMR for characterizing the molecular structure to reconstruct the kerogen model without destruction. And the reconstructed structural models are successfully matched with the original information on the products and residues during degradation. Therefore, the kerogen models reconstructed by this method are closer to the actual state of kerogen.

All in all, the reconstruction methods of the kerogen molecular model in the initial half-century were mainly started from the following four aspects: (1) Degradation of kerogen molecules by physical and chemical methods. (2) Determination of kerogen degradation products. (3) Quantitative analysis of kerogen degradation residues and products. (4) Kerogen combinatorial reconstruction with residue molecules (core) and degradation products.

The constructed accuracy of the above methods depends upon the precision of experimental technology directly. The understanding of kerogen structure is evolved with the development of experimental methods. However, the reliance on experiments makes the reconstruction of the kerogen molecular structure inherently flawed in the method. Since original morphological structures will be destroyed during the determination experiments, there is no way to verify the constructed kerogen model effectively. Hence, the kerogen models reconstructed with the above-mentioned methods are the molecular structures that are random combinations and cross-linking between kerogen pyrolysis products and residues. This process relies heavily on the builder's experience of the determination experiments of the organic matter molecular structures. Therefore, these methods can be described as empirical methods, and the reconstructed structure's accuracy and rationality are challenging to verify. To solve this problem, simulation calculation methods based on DFT, MD, etc., are developed with the improvement of computers. The computer-based techniques are widely applied in reconstructing and verifying kerogen molecular structural models.

3.3. Quantitative Digital Reconstruction Methods of the Kerogen Model. In 1977, the computer-assisted molecular structure construction (CAMSC) program was designed to rapidly construct

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

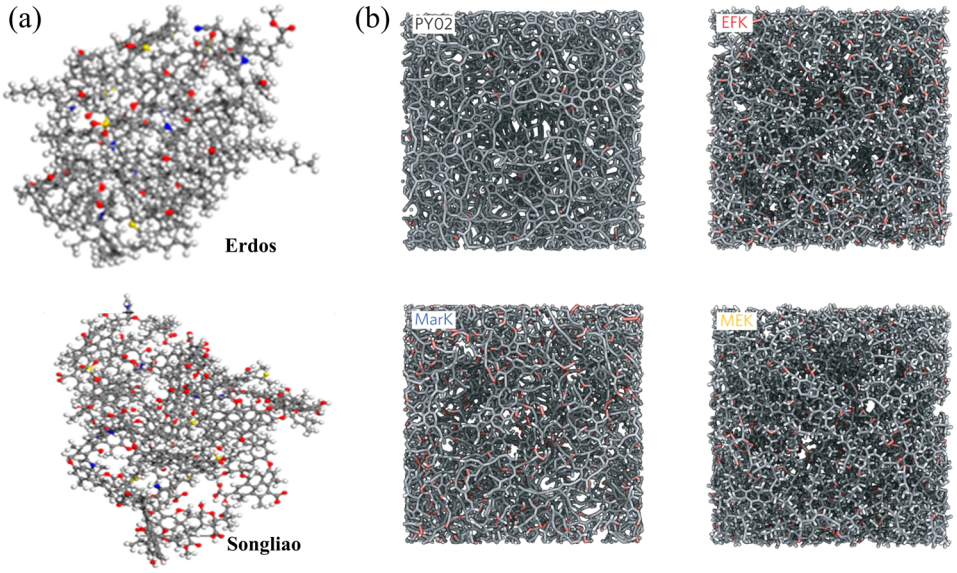


Figure 7. 3D kerogen molecular models reconstructed by (a) Wang et al.[66] and (b) Bousige et al.[93] Panel a was reproduced with permission from ref 66. Copyright 2019 Wiley. Panel b was reproduced with permission from ref 93. Copyright 2016 Springer Nature.

the coal molecular structural model.[108] The process of this program to build a molecular structure can be divided into three steps. First, the elemental components and the ratio of each functional group are obtained via elemental analysis and NMR spectroscopy experiments. Second, a set of building blocks consistent with the experimental information was selected from the assigned ten aromatic and 24 aliphatic basic structures. Finally, these structural units are manually combined to form the molecular structure of coal. This computer-assisted method is extremely simple and can only screen the basic structural units from a minimal structural database (Figure 6a). It is rough to deal with the conditions that complex functional groups need to be combined into molecular models. However, this work is a practical attempt to apply computer technology to reconstruct the molecular model of coal and kerogen. The design idea of this method can be expressed as eq 3:

$$\text{molecule model} = \text{basic structural units} + \text{bonds} \tag{3}$$

Faulon and co-workers developed the kerogen molecular construction software Xmol based on this idea. And Xmol has evolved significantly. On the one hand, the basic structural units used by Xmol are simpler and can adapt to the needs of more functional groups during the reconstruction process (Figure 6b). On the other hand, the molecular structural units and the types of bonds are parametrized. The reconstruction process can be transformed into an equation based on the parametrized basic structural units and bonds as eq 4:

$$s(M) = \sum_{i=1}^{G} x_i s(g_i) + \sum_{j=1}^{B} y_j s(b_j) \tag{4}$$

where $s(M)$ represents the final constructed kerogen molecule, $s(g_i)$ represents the basic structural unit groups, $s(b_j)$ represents the basic bonding type of the basic groups, x_i is the number of each basic unit, y_j is the number of each bond type, G and B are the total number of basic structural units and bonds, respectively. The atomic information is obtained from experiments, and the basic structural units are preset. Thus, only y_j needs to be calculated. Then, the basic structural units are randomly combined by the calculated bonds via Xmol. Finally, the coordinates of each atom are determined through molecular mechanics (MM) and computational geometry.[109] Therefore, the method can directly reconstruct the kerogen molecular models based on the experimental information, and the 3D molecular structure also can be given. However, the combination of substructural units in the reconstruction process is still random. The reconstructed molecular models are matched with the chemical bonding mechanism and experimental information. But it is unlikely to reflect the pyrolysis properties of kerogen, and the information, such as porosity and density, that is obtained from 3D models is also inaccurate.

The computer-assisted methods in the early stage for the kerogen molecular model are essentially digital simulation of the artificial reconstruction process. Structural analysis methods such as MM are added, the reconstruction efficiency is improved, and the 3D molecular models can be constructed. However, due to the lack of verification methods, the results are still doubtful to be regarded as reliable in various chemo-mechanical property analyses. In 2003, a 2D kerogen molecular model of the Estonian kukersite oil shale was reconstructed by combining oxidation pyrolysis and ^{13}C magic-angle spinning nuclear magnetic resonance (MAS NMR). During the reconstruction process, the software ACD/CNMR was used repeatedly to calculate the molecular ^{13}C MAS NMR spectra. A molecular model is obtained, compared, and adjusted with the experimental spectra, and the molecular formula is $C_{421}H_{638}O_{44}S_4NCl$.[110] The structural model with a verification process during reconstruction is more reasonable than that obtained by random combination of small molecular units based on the experimental spectra. The reliability is improved with the development of computer simulation and its successful application in reconstructing kerogen molecular models. Meanwhile, the improved reliability of molecular model building also brings a challenging problem. Comparing and adjusting the molecular structure information with the simulation results in the reconstruction process is necessary. And this trial-and-error process is particularly cumbersome while reconstructing a large molecular structure because of the combinatorial explosion problem, which makes the reconstruction methods being time- and materials-consuming and labor-intensive.

3.4. Quantitative Reconstruction Methods to Approximate the Actual Kerogen Model. The trial-and-error methods based on experiments and simulations are still the primary choice for reconstructing the 2D kerogen molecular models. The addition of the verification process solves the reliability problem of the 2D kerogen molecule. However, the actual molecule is cross-linked and folded in 3D space, and there are interactions between molecules and functional groups. The chemo-mechanical properties of the kerogen are also affected by the 3D folded form. The folded 3D kerogen molecular structure was obtained based on the 2D Siskin Green River model using an annealing algorithm with ab initio and MM methods. The solid-state NMR spectroscopy and pair distribution function of

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

Table 2. Comparison of Different Kerogen Reconstruction Methods

methods	refs	advantages	limitations
qualitative skeleton reconstruction	103−106	indicates kerogen skeleton and cross-linking form	unsuitable for molecular simulation
random cross-linking	100,107−109	reconstruct the complete molecular model in the early stage of kerogen research	depends upon precision of experimental technology and lacks verification; the results are difficult to be reliable
trial-and-error	53,110−112	more reasonable and higher reliability	high cost but low reconstruction efficiency
MD-HRMC	93	repeated trial-and-error process is avoided	chemical structural defects may occur
ML-based	113	high-throughput, intelligent and efficient, expected to reconstruct macromolecular models that conform to functional groups directly	in the initial stage, and the reconstructed molecules are still smaller than common big kerogen monomer

the 3D molecular model are calculated. And the 3D kerogen models are in good agreement with the experiments.[111] The work demonstrates the feasibility of obtaining a 3D molecular structure of kerogen by minimizing molecular energy through the annealing algorithm. This method can be used not only to verify the accuracy of the 3D kerogen molecular models during the reconstruction process but also as a general method for developing existing 2D models into 3D models. With the development of MD simulation technology, Wang et al. developed a physicomechanical inversion method (PMIM) for kerogen molecular reconstruction based on ReaxFF-MD.[53] PMIM is a modification and extension of the traditional reconstruction methods introduced above. On the basis of experiments, the kerogen molecular skeleton is built according to the NMR spectroscopy, XPS, and pyrolysis information. The experimental spectra are compared with the ReaxFF-MD pyrolysis simulation results. And the possible cleavage sites are adjusted by trial-and-error according to the bond dissociation energies. Finally, the kerogen molecular model that conforms to each experiment is determined (Figure 7a). Therefore, compared with the previous methods, the kerogen molecular model constructed by PMIM has a more reasonable distribution of chemical bonds and higher reliability. Obviously, increasing the reliability of the kerogen molecular model by continuously adding verification items is very effective. But the amount of trial-and-error will dramatically increase in the kerogen reconstruction process. Naturally, the costs of labor and material will also be greatly increased, and the reconstruction efficiency will become lower. Thus, the traditional kerogen reconstruction methods are unsustainable for improving the reliability of molecular models. Repeated trial-and-error is required by traditional methods. And this process directly leads to the inherent disadvantages of traditional methods: low reconstruction efficiency and high cost.

Many novel methods have been explored to eliminate the demerits of traditional methods. Bousige et al. developed a molecular dynamics-hybrid reverse Monte Carlo (MD-HRMC) reconstruction method based on MD simulation.[93] Unlike the conventional methods, only three elements and the density are taken as the input to reconstruct the kerogen structure. The atoms are automatically combined to form macromolecular structures by the MD-HRMC method with ReaxFF. The annealing algorithm is used to screen out the molecular model that matches the density information (Figure 7b). The merit of the MD-HRMC method is that there is no need to construct a 2D molecular structure as the intermediate model and then obtain the 3D target molecule. Thus, the repeated trial-and-error process is avoided. And they believe this strategy can be used to model the molecular structure of any heterogeneous and disordered material. The basic idea of the MD-HRMC method can be considered to hand over the trial-and-error process to the high-performance simulation algorithm, and only the simple constraint conditions are relied on to construct a set of candidate molecules. Finally, the target molecule is searched reversely to avoid the cumbersome trial-and-error process according to the additional experimental information. The MD-HRMC method has high requirements for computing power and simulation comparison methods. Due to the limitation of the ReaxFF and the quality of the phase space sampling, there are chemical structural defects (about 0−10%) in the molecular model constructed by MD-HRMC, and the function groups in reconstructed molecular models cannot be accurately indicated.[29]

All in all, during the development of the reconstruction method from conjectured kerogen skeleton models to refined 3D molecular models, researchers have made remarkable efforts to develop more efficient and reliable reconstruction methods (Table 2). At present, the reconstruction method of the kerogen molecular model is overall developing in two directions: more reliable and more significant in molecular weight, which are complementary but incompatible with each other. Reconstructing a more reliable molecular model requires repeated trial-and-error, so the molecular weight is challenging to be achieved on a large scale. The method based on MD simulation can quickly reconstruct a sizable molecular weight kerogen model, but the functional group structure is doubtful to match with the actual sample. ML-based methods have powerful analysis capabilities for complex problems and can be efficient and accurate. The kerogen molecular models can be reconstructed based on experimental data with high-throughput by ML. Therefore, we believe that the ML-based method is most likely to integrate the benefits of the two major directions of kerogen reconstruction and makeup shortcomings of each other.

4. ML RECONSTRUCTION METHODS OF KEROGEN MODEL

In accordance with the above statement, it can be seen that the kerogen molecular model occupies the most effective pathway for the study of kerogen's chemo-mechanical properties. The researchers have been working hard to explore reconstruction methods that are more reliable, simple, and effective. However, because of the complexity of the kerogen structure, current molecular model reconstruction methods are still too complicated. The dilemma is mainly manifested in the following two aspects: on the one hand, determining the kerogen structure generally requires many experiments such as NMR, XPS, FTIR, and pyrolysis. The comprehensive analysis of the experimental information is required to reconstruct organic molecules by experienced professionals. On the other hand, the traditional reconstruction methods are time- and material-consuming and labor-intensive because of the trial-and-error processing. For these reasons, the kerogen in each mining area has different structural characteristics due to the various origins and geological evolution conditions. But only the kerogen models of a few mining areas, such as Green River, are reconstructed. Almost all the studies that have been carried out are based on these few kerogen models. It is exceptionally unfavorable for exploring shale oil/gas reserves and developing oil/gas in situ ripening technology. Therefore, more intelligent and effective methods are urgent to be developed for reconstructing the kerogen molecular models. In the past decades, ML methods have grown rapidly and obtained outstanding achievements in many fields.[114] This novel method makes it possible to realize the intelligent and high-throughput reconstruction of kerogen. ML methods can automatically extract the target features from massive training samples and establish the implicit connection between input and target in the application. Of course, a lot of time and material resources are required during the process of labeling enormous samples. But once the ML neural

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

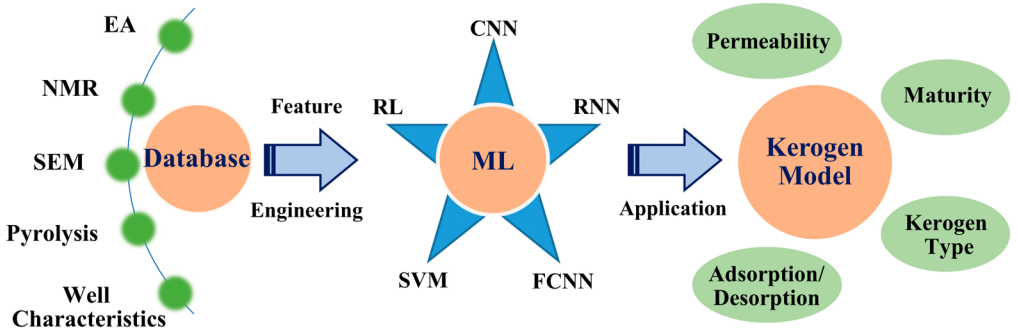


Figure 8. Application of ML methods in reconstructing kerogen model and extracting shale oil/gas.

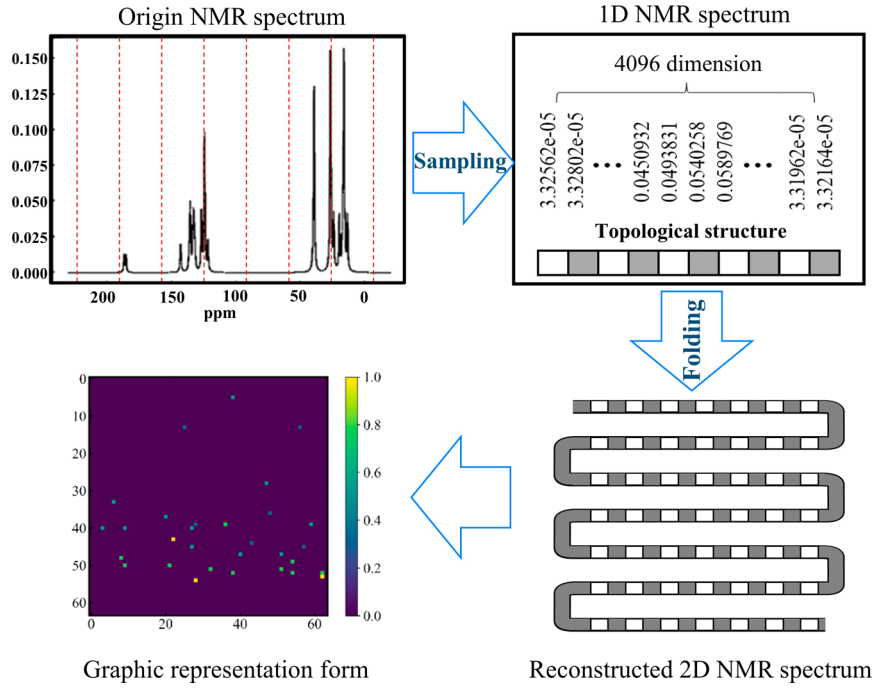


Figure 9. Schematic diagram of two-dimensional equidistant folding and reconstruction of the NMR spectrum.

network is trained, it can be used directly as an analysis tool without the operator's experience or theoretical and simulation skills.

4.1. Synopsis of ML Methods. The essence of ML is to capture implicit rules from massive labeled samples and apply the predictive capabilities for similar unknown problems. Therefore, the labeled sample dataset is the essential prerequisite for the training of the ML model. The designed ML neural network is the computational framework, and the dataset can be regarded as the soul of the ML method. The neural network model can be trained successfully only when the two parts work together. During the training process, it is necessary to input more than tens of thousands of qualified samples. However, establishing a qualified database for the training of ML is an extremely challenging project. Generally speaking, more than 80% of the effort is spent on building the database while solving practical problems by ML.

Generally, solving practical problems through ML methods is mainly divided into three parts, data collection, feature engineering, and suitable ML model design. Currently, many methods with different characteristics are developed, such as support vector machine (SVM), reinforcement learning (RL), fully connected neural network (FCNN), convolutional neural network (CNN), recurrent neural network (RNN), etc. The appropriate ML models are designed ML model according to the characteristics of the research target.[115,116] As is exhibited in Figure 8, except for the reconstruction of kerogen molecular models, the works of well completion designs and mechanism of oil/gas production are also carried out based on ML methods directly or indirectly. Therefore, ML methods have broader application prospects in shale oil/gas research.

4.2. Feature Engineering and Database of ML. Feature engineering in ML refers to the process of extracting the features from the original sample information and reconstructing the selected features to the form that can be used by the ML model.[117,118] Reconstructing the original data through feature engineering is the most critical part of the ML method. If the key sample features are missing during the feature engineering, the performance of the trained ML model will decrease or even fail. On the contrary, the computational complexity will increase significantly if the redundant features are contained in the processed samples. Therefore, the better feature engineering scheme can effectively reduce the amount of computation and improve the performance of ML algorithms.[119] Since ML methods require large batches of labeled samples for training, the following three characteristics should be considered: (1) can be collected and labeled quickly with low cost, (2) can be processed digitally, and (3) the rules between sample features and the training targets are universal.

First of all, easy to be collected and labeled is the premise of establishing ML datasets. The samples that can only be obtained through experiments are often strenuous in large quantities. With the development of simulation technology, more and more samples can be obtained by simulation. The basis of ML model designing and

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

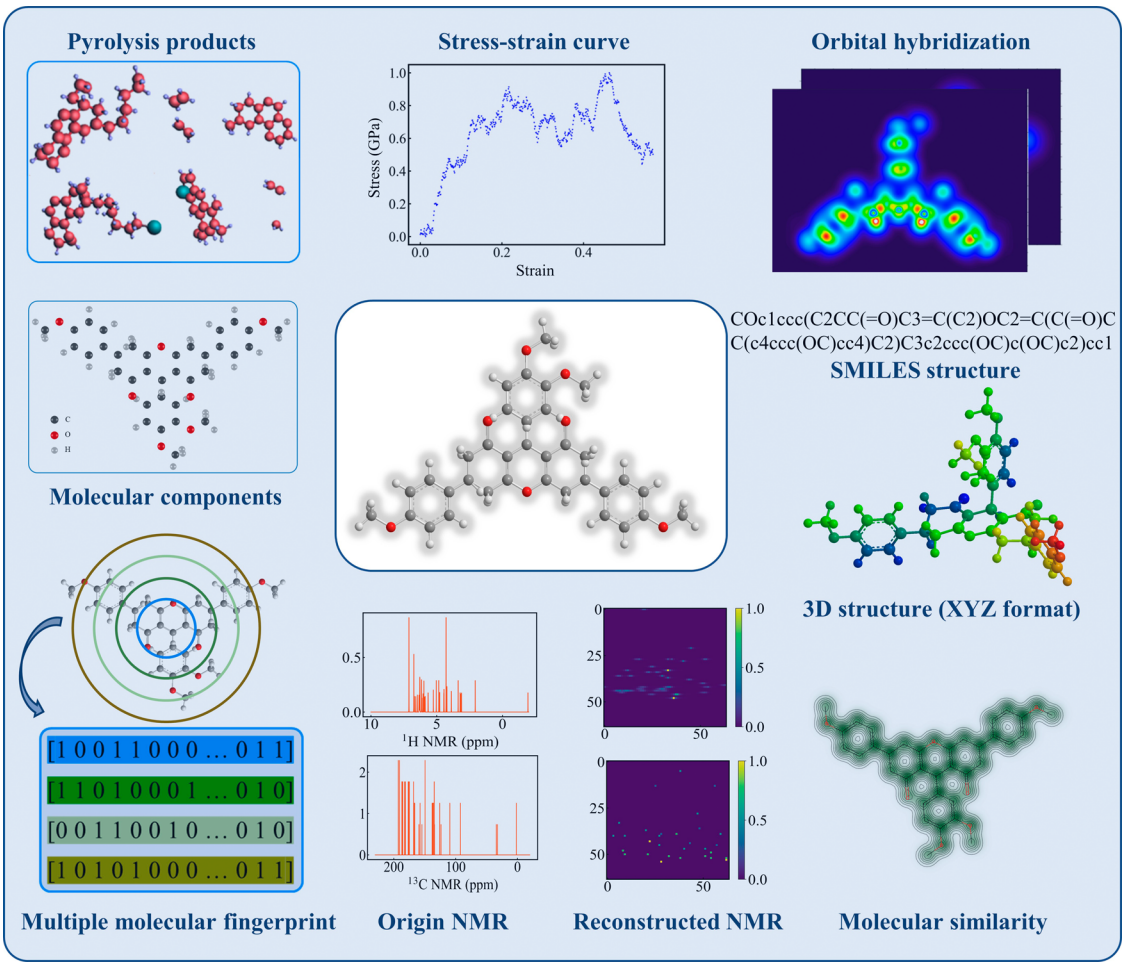


Figure 10. Composition of sample molecular information in ML database.

training is that sample features can be processed digitally. Finally, the ML methods refer to the deduction based on the training dataset. The universal laws implied in the training dataset are learned and taken to predict the unknown samples. But ML methods are powerless if there are special cases in the problem.

NMR spectroscopy is the primary sample information for reconstructing kerogen molecular models via ML. It is one of the most commonly used tools for chemical structural analysis, as it contains information on nuclei's chemical shifts and peaks. The chemical shifts of atoms are only affected by the adjacent functional groups. The peak position represents the functional group type in the molecule. The peak value denotes the content of a certain structural unit.[120−122] In general, only the sequential input feature information on the same dimension can be entered into the ML models. And the location of input features is practical for ML. Hence, Kang et al. developed the one-dimensional (1D) and 2D reconstruction methods of NMR spectral features according to the properties of NMR spectra.[123] As shown in Figure 9, the original NMR spectrum is normalized and discretized into a 1D array. The information on the abscissa is implicit in the position index of the sequenced NMR spectra. Combined with the physical meaning of the NMR spectra, it can be known that in the processed array, the index and the corresponding value represent the structure and content of different molecular functional groups, respectively. The merits of this NMR spectral reconstruction method are the clear physical meaning, no redundant features, and ease to be stored. The processing of unifying NMR spectral features through discretization and sampling also results in the loss of original sample information. However, this is an inevitable process to support the NMR spectral features from different sources successfully learned by neural networks.

The 1D reconstituted NMR spectra is suitable for the extraction of simple molecular structural information but cannot meet the needs of building the whole kerogen molecular model. The dimension of the 1D reconstructed NMR spectra is too large for the CNN. It is a barrier for CNN to obtain correlation information between the distant peaks in the NMR spectra. The convolution kernels are set in the CNN to identify local parts of input features by sliding according to different step size parameters. The computational cost of neural network model training is significantly reduced. And the strong local feature analysis capability of CNN is provided by the convolution kernels. However, the connection of the extracted local information can only be expressed in the deeper neural network layer because of the limited receptive field, which weakens the ability to extract the remote contact between features. So, if the target problem is affected significantly by the relationship between the farther apart features, the CNN is arduous to achieve the desired effect. And inferring molecular structure from NMR spectra is one of these problems. In the analysis of NMR spectra, information is processed at multiple levels as it passes from the shallow CNN layer to the deep. Part of the NMR spectral information will be lost, and it is thin for the deep network to obtain the predictive ability of the entire molecular model from the incoming information. Therefore, a method to fold the 1D sequence of the NMR spectra into 2D is developed, as shown in Figure 9.[113] The local and distant features of NMR spectra will be simultaneously extracted by the 2D convolution kernels during the training process. Thus, the connection between features will be reflected in the shallow CNN layers, and information loss will not occur prematurely. The better prediction accuracy of the trained ML model is performed when reconstructing the molecular structural models.

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

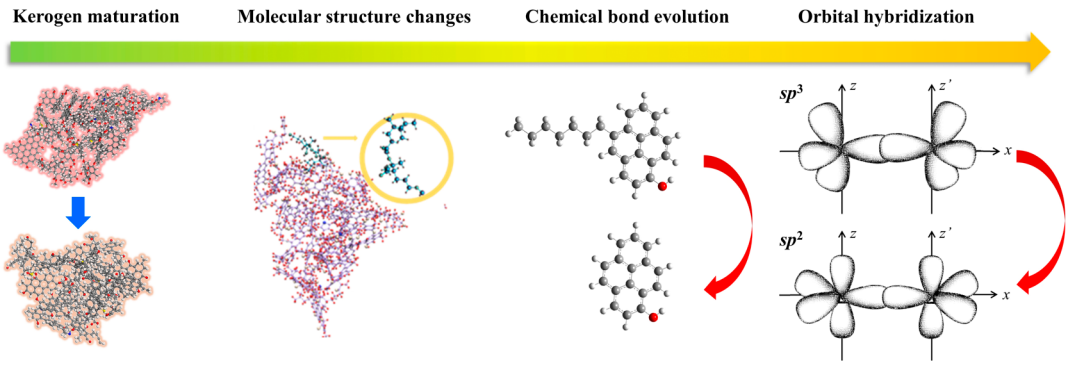


Figure 11. Mechanism of kerogen thermal maturation based on orbital hybridization.

In section 3, various molecular information characterization experiments are analyzed comprehensively in the traditional models to improve the accuracy of reconstructed kerogen structural models. Similarly, the multi-NMR spectra method can also be designed through feature engineering to analyze different spectra comprehensively. Kang et al. designed an input method and ML model for combining ^{13}C and ^{1}H NMR spectra based on 2D NMR reconstructed spectra. The prediction performance beyond any single input spectral type is obtained. Theoretically, this spectral NMR reconstruction method is suitable for other 2D spectra such as XPS, FTIR, etc. Therefore, compared with the single-input form, the multi-spectra way has more outstanding expansion capabilities and potential.

Researchers have explored the exploitation, exploration, and maturation of unconventional oil/gas reserves through ML methods.[124−127] However, there is no database of kerogen molecules that can be directly applied to ML until now. The knots of building a database are mainly manifested in three aspects. First, it is complicated to obtain samples from mining areas. Shale oil/gas reservoirs are mainly buried hundreds to thousands of meters underground. Collecting sampling is laborious and expensive. However, constructing a ML database requires many samples from different mining areas, which is unrealistic. Second, the purification and experimental determination of kerogen samples are expensive. Current methods for determining unknown molecular structures include FTIR, NMR, XPS, and Py-GC/MS. The manpower and material resources to repeat tens of thousands of experiments are too outrageous to estimate. As an amorphous substance, kerogen has a complex molecular structure and various functional groups, which often take much time to reconstruct a single kerogen molecular model. Therefore, it is impossible to reconstruct tens of thousands of qualified kerogen models to build the database by traditional methods.

The chemical structural rules between NMR spectral features and the corresponding molecular functional groups are learned by the ML model during training. And all the molecules follow the same chemical rules. Therefore, in the condition without sufficient kerogen molecular samples, the existing other molecular structures can be selected as samples to train the ML model. Then the predictive ability for various spectral features is obtained by ML models. Hundreds of millions of molecular samples are recorded in the open-source database, such as PubChem, SuperNature II, etc.[128−130] The molecular structures with kerogen molecular characteristics and the constructed kerogen models in published literature can be collected in the database by screening. But the collected molecules cannot be used directly by the ML model before labeling. In the research of intelligent and high-throughput reconstruction of kerogen models by ML, more than one million samples for the training of ML models are labeled by the Zhao group. The NMR spectra, structural formulas, multimolecular fingerprints, molecular hybridization information, etc. are included in the database, as shown in Figure 10. It should be pointed out that the samples need to be labeled one by one. So, establishing a database should take a lot of time and patience to accumulate qualified samples. Therefore, it is meaningful work to collect various information on kerogen through feature engineering. The database can be used for intelligent high-throughput reconstruction of the kerogen molecular model and is also helpful for the study of various kerogen chemo-mechanical properties.

4.3. Kerogen Molecular Model Reconstruction by ML Methods. With the development of artificial intelligence and computing power, ML methods have obtained outstanding achievements in many fields.[131] Researchers have tried to apply ML methods to the reverse reconstruction of molecules.[132,133] Due to the inherent complexity of kerogen molecular models, the research is still in the initial stage. There are few published works on the reverse reconstruction of molecular models based on ML. However, it is believed that the ML methods have broad application prospects and application value in the intelligent reconstruction of molecular models.[114,132] Duvenaud et al. extracted the ML fingerprints end-to-end by the CNN model. They demonstrated that these new fingerprints are more interpretable and have better predictive performance on various tasks.[134] The data-driven molecular characterization model based on RNN was established to realize the encoding and decoding of simplified molecular input line entry specification (SMILES)[135] structural formula and validated on molecular models with less than nine heavy atoms.[136] Winter et al. extended the model to structures with molecular weights from 12 to 600 Da by training on 7.2×10^7 groups of samples.[137] And the conversion between different SMILES standards is achieved.[138] Subsequently, the molecular models were reconstructed reversely based on extended-connectivity fingerprints (ECFPs).[139,140] Although the work belongs to the reverse reconstruction of molecules, the molecular fingerprint information should be obtained through existing molecular models, which seems to be a paradox.

The above-mentioned molecular characterization models based on ML laid the foundation for the intelligent reconstruction of kerogen via NMR spectra. In 2021, Kang et al. analyzed the structural components of the kerogen molecular skeleton by combining ML with the ^{13}C NMR spectra and predicted kerogen types.[123] The prediction accuracy of each component is C: 96.1%, H: 94.8%, and O: 81.7%. The prediction accuracy of the three kerogen types can reach about 90%. Ma et al. extracted the orbital hybridization and chemical bond information from ^{13}C NMR spectra to predict thermal maturity via ML. The results showed that the average prediction error for kerogen maturity was less than 5%. This work proves that the proportion of sp^2 carbons increases while sp^3 carbons decrease during kerogen maturation. Thus, the molecular structure of kerogen gradually changes from the aliphatic to the aromatic structure (Figure 11). A new maturity index called OrbHMI is proposed based on the relationship between orbital hybridization and maturity:[54]

$$\mathrm{OrbHMI} = \frac{1}{2.85 - 1.1C_{sp^2}/(C_{sp^2} + C_{sp^3}) + 0.1O_{sp^2}/(O_{sp^2} + O_{sp^3})} \quad (5)$$

Here, C_{sp^2} and C_{sp^3} represent sp^2 and sp^3 hybridized carbons, respectively. O_{sp^2} and O_{sp^3} represent sp^2 and sp^3 hybridized oxygens, respectively. As a result, the feasibility of ML methods to analyze

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

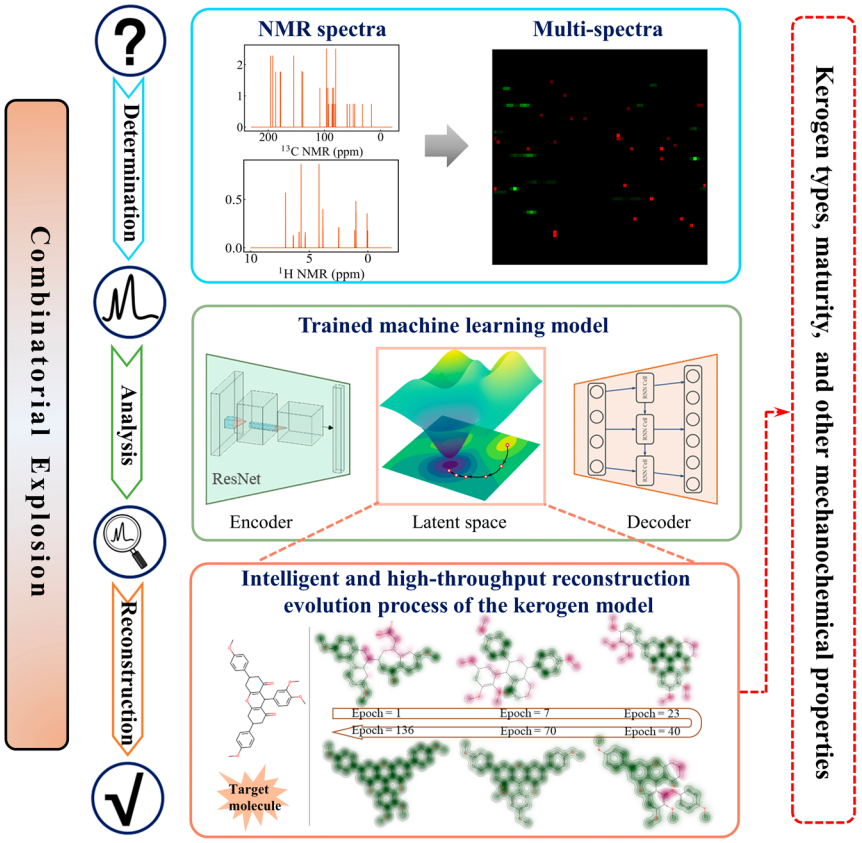


Figure 12. Schematic diagram of using machine learning to construct the kerogen molecular models intelligently.[113] Reproduced from ref 113. Copyright 2022 American Chemical Society.

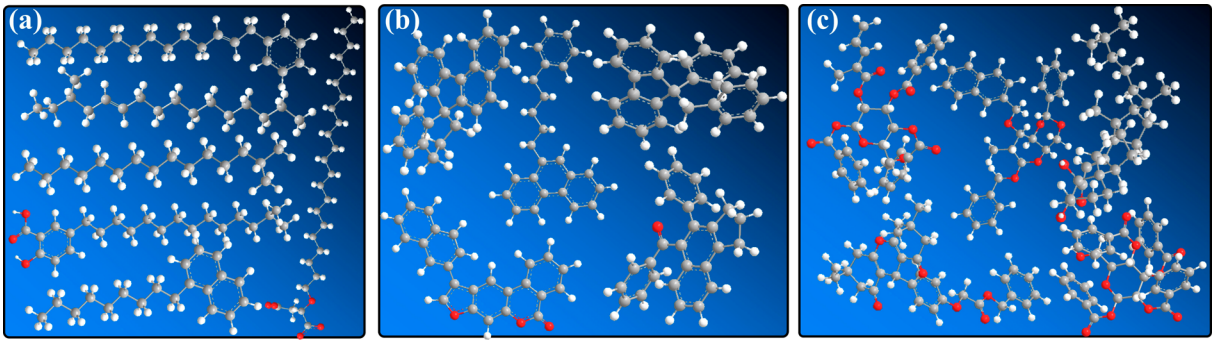


Figure 13. Different characters of the predicted kerogen molecular structures. (a) Chains of aliphatic carbon. (b) Condensed aromatic ring structures. (c) Aromatic and aliphatic ring structures. (The gray, white, and red represent C, H, O atoms, respectively.).

molecular structure information based on experimental spectra is demonstrated. The ML neural network model was established with ^{1}H NMR and ^{13}C NMR 2D multi-spectra. The comprehensive analysis of different spectra is realized to solve the combinatorial explosion in kerogen reconstruction (Figure 12). The correlation between multi-spectral and molecular models is established in ML latent space during the training process. Thus, the ability to predict the kerogen molecular models is obtained by ML neural network. During the testing and predicting process, the multi-spectral of unknown molecules are fed into a trained ML model. The input spectral features are analyzed in latent space through a high-throughput way. Then the reconstructed molecular SMILES structures are outputted at the end of the ML model. Neither manual intervention nor trial-and-error is demanded in the molecular reconstruction. Finally, the maturity, types, and other chemo-mechanical properties of kerogen can be acquired from further analysis and simulations.

The predictive that the performance exceeded any single spectral input form is learned. The molecular similarity[141,142] of 82.51% reconstructed unknown kerogen molecules is greater than 80%, and 54.78% of the total molecules are matched with the target. The parameters such as kerogen skeleton components, types, and maturity are analyzed based on the reconstructed molecular models. The prediction accuracy is between 92.1% to 99.5%, and the determination coefficient R^2 exceeds 0.932.[113] Compared with the previous models for single structural information, the prediction accuracy of the new model has been significantly improved. As is shown in Figure 13, different types of kerogen molecular models can also be reconstructed by the ML method, which illustrates the effectiveness of the ML method in reconstructing the kerogen structure more clearly. The results exhibit wonderful performance for the intelligent high-throughput reconstruction of kerogen molecules by the ML method. The molecular weight of the reconstructed model is about 1/4−1/8 of the commonly used maximum molecule in the

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

kerogen matrix until now. Still, it is suitable for the asphalt model[143−145] and part of the kerogen molecular matrix. Also, the predicted molecules can be regarded as the functional groups of the super-large kerogen monomer molecular model. Then the kerogen monomer model can be obtained by combining the predicted functional groups.

The molecular structural model contains all the structural information of kerogen. Hence, compared with the ML method of predicting the single kerogen structural parameters, the higher prediction accuracy, and wider applicability are obtained via the reconstruction of kerogen structural models based on ML. Even more, the chemical bonding rules of components are learned by ML models during the training of reconstructing kerogen models. Under the guidance of unified laws, the number of atoms, bonds, and the bonding position of each group will be corrected with each other. The predictive accuracy of obtaining structural information is improved. It is an essential reason that the predictive performance of constructing the whole kerogen models is better than the ML model of predicting single-parameter.

4.4. Application of ML Methods in Shale Oil/Gas Exploitation. Researchers have developed various prediction models based on ML methods to evaluate the impact of oil well properties, reservoir characteristics, and well production behaviors. With the help of ML simulation and modeling, the exploitation characteristics of shale oil/gas reservoirs are quickly analyzed, saving the exploration cost.[146−148] Due to the shale oil/gas production characteristics of rapid decay and gradual recovery, various parameters such as oil/gas well location, geological conditions, petrophysics, etc., must be considered comprehensively.[149] Hence, oil/gas production is challenging to be predicted even with ML-based methods.[150] In addition, the training of ML models requires massive data from shale oil/gas reservoirs. Collecting and labeling a large number of qualified training samples is also a very tough task. The well geological characteristics, well completion design, location of well, shale wettability, and reservoir quality are often adopted as dataset to optimize the well design and oil/gas production.[151−153] Shahkarami et al. collected data from more than 800 wells under different drilling and hydraulic fracturing parameters and normalized the 25 input characteristics to estimate the production behavior of oil wells.[154] Also, the completion and stimulation parameters of nearly 2700 wells are used to predict the Marcellus shale's initial production and optimize the oil wells.[155] The SVM algorithm is proposed to evaluate the oil/gas saturation of the Ordos shale reservoir.[156]

Various components are contained in the shale, and the heterogeneous distribution of each component in the shale will affect its mechanical parameters, permeability, etc. Therefore, it is challenging to establish a simple mathematical model to describe the heterogeneous shale. The study of the shale mechanical properties is helpful in improving the understanding of the formation mechanism of fracture networks in shale oil/gas reservoirs. ML provides a new way for oil/gas shale research because of the robust analysis and modeling capabilities for complex issues.[157] Currently, the research on shale exploration using ML is mainly based on SEM images, and image analysis is one of the strong fields in ML.[158] Thus, selecting SEM images to study the shale characters is a very convenient and practical cut-in point. The CNN and conditional generative adversarial network method are proposed to enhance damaged shale SEM images. More importantly, the images similar to SEM can be predicted through nondestructive transmission X-ray microscopy images. The technique can preserve samples for further tests without damaging shale samples.[159] Pores and mineral composition in shale SEM images were divided by the ML method.[160] The pore distribution in shale and the transportation capacity of oil/gas transport channels were further analyzed.[161] ML and finite element simulation were combined to establish a model for predicting the elastic modulus of shale. It is believed that the ML method could be extended to predict the elastic modulus of other heterogeneous materials.[162] In addition, image recognition technology was used to develop a strategy for predicting permeability based on low-resolution scanning microscope images of porous media.[163] And the porous media can be built via generative adversarial networks, the efficiency is higher than the traditional numerical methods.[164]

Besides the shale image-based studies, researchers also apply ML to establish methane adsorption models in shale.[165] The methane adsorption curve was successfully predicted by ML, which can be easily applied to optimize the shale gas production curve. It is noted that the new implicit prediction model based on ML is a new model that is different from the traditional adsorption models.[166] The 352 groups of samples from the literature are analyzed and an explicit adsorption model is established through gene expression programming, and the correlation coefficient of predicted results is 0.9837. The parameters of pressure, temperature, water content, and total organic carbon content are concluded in this model.[167] Although Amar exhibits the mathematical expression of the adsorption model, the mathematical form of this model is extremely complex. It is still the forward propagation process of parameters in the ML neural network. The relationship between the terms is still chaotic to explain, so it is essentially a ML implicit model.

Compared with traditional experimental and finite element simulation methods, ML is more efficient with lower labor and material resources costs. In addition, it can quickly and implicitly establish a model for complex problems, which is convenient for promotion in engineering. The benefits have led to the rapid development of ML methods as a new and powerful tool beyond experimental and simulation. However, the implicit model established by ML usually does not have a concise mathematical explanation. Hence, even if high prediction accuracy is obtained by the trained ML models, it is challenging to explain the actual physical meaning and role of each neuron in the ML models. Thus, the ML model may not be conducive to establishing a model of abstract physical concepts.[168]

5. CHALLENGES AND PERSPECTIVES

The ML methods provide the novel way for kerogen reconstruction and oil/gas exploitation. The pros of the trial-and-error and MD methods can be combined to reconstruct the kerogen macromolecular model with accurate functional groups through ML-based methods. It is expected to completely solve the cons of traditional reconstruction with the high cost and low efficiency. And there is no need for professional intervention during the reconstruction process. So, ML-based methods are more intelligent and conducive to industrial promotion. Also, the excellent application potential in complex shale reservoir exploitation and drilling design is exhibited. Preliminary results of kerogen reconstruction and shale oil/gas exploration are achieved by ML. However, the development of ML-based methods in shale oil/gas research is still in the initial stage. The complex kerogen and other macromolecules are beyond the ability of the current ML models. The analysis of in situ maturation of oil/gas reservoirs and the mechanism of oil/gas generation and migration also need to be further developed. The key challenges of ML-based methods are mainly concerned in three directions: database building, ML model, and feature engineering design.

In terms of the sample database, new simulation tools need to be developed with high-performance computers to expand the training samples and types of spectra. With the molecular scale expansion, the predictive ability of the current models to reconstruct the unknown molecules completely matched with the target will decrease gradually. The ability can still be improved by collecting and labeling more training samples or adding the spectral types of current samples. However, an exponentially increased sample number is required and is time- and cost-consuming under the present experimental and simulated conditions. Thus, it is necessary to collect as many training samples as possible through continuous accumulation.

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

It should be emphasized that a robust database is a prerequisite for the realization of ML methods.

In terms of ML model design, as the scale of the molecule increases, the number of neural network layers in the decoder will increase accordingly. The gradient disappearance/explosion may appear in the network's deep layers, failing training. There is a lot of research on gradient disappearance/explosion, which can be alleviated by weight initialization, gradient clipping, etc. Unfortunately, there is still no effective method to solve it completely.[20] For the result of this, the SMILES-style character length of the target molecule is limited to approximately 100 or less. Therefore, the better ML neural network models must be designed to improve the decoding length. AlphaFold successfully predicted the 3D protein structure of the determined amino acid sequence and caused a considerable sensation in the biological study. Protein has only 22 amino acids, and each amino acid can only be connected one-to-one, so it is very different from the reconstruction of the kerogen molecular structure. However, the idea of the AlphaFold combining expert system, residual neural network, and the self-attention mechanism to build the ML model provides a new idea to reconstruct larger-scale kerogen molecular models.[169,170] Among the developed neural network models, the authors believe that Transformer[171] is the most promising way for reconstructing larger molecular weight kerogen molecular models. In Transformer, the multi-head attention is used to replace the RNN and the sequences are calculated in parallel, which improves the computing efficiency. The transformer effectively solves the problem of remote forgetting of RNN and reduces the risk of gradient disappearance/explosion. However, the prior information (such as sequence order) is not taken in Transformer, and the computational complexity is increased with the sequence length n by square times ($O(n^2)$). Transformer has significant sequence prediction ability, but more samples are required during training. It still needs to be selected according to the number of collected samples in the application.

The new experimental spectral reconstruction plan in the design of feature engineering will also be helpful for the ML methods. One is to compromise on the intelligent index from the perspective of ease of implementation. The methods should be developed to divide the NMR spectrum into several sub-spectra reasonably, according to the characteristics of the NMR. After the sub-spectra is constructed by the trained model, the complete structures can be combined with the original spectrum by traditional methods. The constructed structures are finally fine-tuned by traditional methods for the larger molecules according to the relationship between similarity and structures. The other is to extend the ML-based technical route completely. However, this way is more challenging, and the new reconstruction plans may need to be designed in combination with new ML methods.

As for the complex problems, such as sweet spots prediction, well completion design, and so forth, that are affected by multiple factors in the research of shale oil/gas exploitation, the data-driven supervised ML methods are indeed the best choice at present. However, there are two different ways in the mechanism study of the in situ ripening and oil/gas adsorption, desorption, and migration combined with MD and ML. One is to combine unsupervised ML methods to analyze the oil/gas generation behavior, even establish the theoretical model via ML. This way may be more convenient and effective than the data-driven supervised ML methods. On the other hand, ML methods can be used to develop the new MD potential fields or directly combine with simulation methods to improve the computational efficiency and accuracy of simulation methods and then indirectly enhance research on oil/gas extraction.

6. CONCLUSIONS

In this work, the fundamental importance of kerogen molecular models, the development of kerogen reconstruction methods, and the application of ML are introduced briefly. Some recommendations for further research are suggested. Generally, it is necessary to determine the mechanism of adsorption/desorption, maturity evolution, and in situ ripening bottom-up through kerogen molecular models while exploring shale oil/gas. Thus, the high-efficiency and high-accuracy reconstruction of kerogen molecular models are the cornerstone of kerogen chemo-mechanical research. For decades, with the blossom of simulation and experiment methods, researchers have continuously extended more accurate reconstruction methods of kerogen and achieved excellent achievements. However, the traditional methods require experienced professionals to adjust the molecular structure through repeated trial-and-error on the basis of various experimental spectra and approach the reconstructed kerogen molecular model closer to the actual sample. Therefore, the reconstruction methods of trial-and-error not only consume a lot of time and material resources but also have extremely low reconstruction efficiency. Recent years have witnessed the rapid development of ML on high-complexity problems because of the vital analysis capabilities via big data. Intelligent and high-throughput prediction without human intervention is the most prominent advantage of ML methods. The ML-based methods are designed to predict the kerogen molecular models and the mechanism of oil/gas generation and exploration. Although the ML research on unconventional oil/gas is still in the exploratory stage, superior achievements have been obtained. It is believed that state-of-the-art ML is the most promising method to realize intelligent high-throughput reconstruction of kerogen molecular models and can be widely used in oil/gas production predicting, in situ ripening, etc. Thus, developing ML-based methods has remarkable significance for the exploration of shale oil/gas.

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

Authors

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*

Jun Ma − *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.2c03307

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

Notes

The authors declare no competing financial interest.

Biographies

Dongliang Kang now is a Ph.D. student under the supervision of Professor Ya-Pu Zhao at the Institute of Mechanics, Chinese Academy of Sciences. His research concerns machine learning in intelligently constructing the kerogen model and chemo-mechanical properties of kerogen.

Jun Ma now is a M.S. student under the supervision of Professor Ya-Pu Zhao at the Institute of Mechanics, Chinese Academy of Sciences. His research focuses on machine learning and the kerogen maturation mechanism.

Ya-Pu Zhao received his Ph.D. degree in solid mechanics in 1994 from Peking University (China) and now is a professor at the Institute of Mechanics, Chinese Academy of Sciences. He served as Director of the State Key Laboratory of Nonlinear Mechanics from 2000 to 2006 and obtained The National Science Fund for Distinguished Young Scholars in 2002. His research interests include mechano-energetics, physical mechanics, impact dynamics, and surface forces in MEMS and NEMS.

■ 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) Pang, W.; Wang, Y.; Jin, Z. Comprehensive review about methane adsorption in shale nanoporous media. *Energy Fuels* **2021**, *35*, 8456−8493.
(2) Wu, J.; Prausnitz, J. M.; Firoozabadi, A. Molecular-thermodynamic framework for asphaltene-oil equilibria. *AIChE J.* **1998**, *44*, 1188−1199.
(3) Buenrostro-Gonzalez, E.; Lira-Galeana, C.; Gil-Villegas, A.; Wu, J. Asphaltene precipitation in crude oils: Theory and experiments. *AIChE J.* **2004**, *50*, 2552−2570.
(4) Zhou, J.; Jin, Z.; Luo, K. H. Insights into recovery of multi-component shale gas by CO2 injection: A molecular perspective. *Fuel* **2020**, *267*, 117247.
(5) Wang, Q.; Chen, X.; Jha, A. N.; Rogers, H. Natural gas from shale formation − The evolution, evidences and challenges of shale gas revolution in United States. *Renewable Sustainable Energy Rev.* **2014**, *30*, 1−28.
(6) Stevens, P.*The shale gas revolution: Developments and changes*; Chatham House: London, 2012.
(7) Hughes, J. D. A reality check on the shale revolution. *Nature* **2013**, *494*, 307−308.
(8) Li, W.; Zhang, M.; Nan, Y.; Pang, W.; Jin, Z. Molecular dynamics study on CO2 storage in water-filled kerogen nanopores in shale reservoirs: Effects of kerogen maturity and pore sizex. *Langmuir* **2021**, *37*, 542−552.
(9) Kang, Z.; Zhao, Y.; Yang, D. Review of oil shale in-situ conversion technology. *Appl. Energy* **2020**, *269*, 115121.
(10) Afagwu, C.; Al-Afnan, S.; Patil, S.; Aljaberi, J.; Mahmoud, M. A.; Li, J. The impact of pore structure and adsorption behavior on kerogen tortuosity. *Fuel* **2021**, *303*, 121261.
(11) Tissot, B. P.; Welte, D. H.*Petroleum formation and occurrence*; Springer: Heidelberg, 1984.
(12) Cornelissen, G.; Gustafsson, Ö.; Bucheli, T. D.; Jonker, M. T.; Koelmans, A. A.; van Noort, P. C. Extensive sorption of organic compounds to black carbon, coal, and kerogen in sediments and soils: mechanisms and consequences for distribution, bioaccumulation, and biodegradation. *Environ. Sci. Technol.* **2005**, *39*, 6881−6895.
(13) Kontorovich, A. E.; Bogorodskaya, L. I.; Borisova, L. S.; Burshtein, L. M.; Ismagilov, Z. P.; Efimova, O. S.; Kostyreva, E. A.; Lemina, N. M.; Ryzhkova, S. V.; Sozinov, S. A.; Fomin, A. N. Geochemistry and catagenetic transformations of kerogen from the Bazhenov horizon. *Geochem. Int.* **2019**, *57*, 621−634.
(14) Huang, X.; Zhao, Y.-P. Characterization of pore structure, gas adsorption, and spontaneous imbibition in shale gas reservoirs. *J. Pet. Sci. Eng.* **2017**, *159*, 197−204.
(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) 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.
(17) Schuster, P. Taming combinatorial explosion. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 7678−7680.
(18) Yan, Y.; Borhani, T. N.; Subraveti, S. G.; Pai, K. N.; Prasad, V.; Rajendran, A.; Nkulikiyinka, P.; Asibor, J. O.; Zhang, Z.; Shao, D.; Wang, L.; Zhang, W.; Yan, Y.; Ampomah, W.; You, J.; Wang, M.; Anthony, E. J.; Manovic, V.; Clough, P. T. Harnessing the power of machine learning for carbon capture, utilisation, and storage (CCUS) − a state-of-the-art review. *Energy. Environ. Sci.* **2021**, *14*, 6122−6157.
(19) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85−117.
(20) Goodfellow, I.; Bengio, Y.; Courville, A.*Deep learning*; MIT Press: Cambridge, 2016.
(21) Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55−75.
(22) Kamilaris, A.; Prenafeta-Boldú, F. X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70−90.
(23) Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **2018**, *9*, 1−8.
(24) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—A review of the state of the art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828−849.
(25) Zhao, Y.-P. Physical mechanics investigation into carbon utilization and storage with enhancing shale oil and gas recovery. *Sci. China Technol. Sci.* **2022**, *65*, 490−492.
(26) Hutton, A.; Bharati, S.; Robl, T. Chemical and petrographic classification of kerogen/macerals. *Energy Fuels* **1994**, *8*, 1478−1488.
(27) Vandenbroucke, M.; Largeau, C. Kerogen origin, evolution and structure. *Org. Geochem.* **2007**, *38*, 719−833.
(28) Durand, B.*Kerogen: Insoluble organic matter from sedimentary rocks*; Editions Technip: Paris, 1980.
(29) Zhang, H.; Ahmed, M.; Zhan, J.-H. Recent advances in molecular simulation of oil shale kerogen. *Fuel* **2022**, *316*, 123392.
(30) Salmon, V.; Derenne, S.; Lallier-Vergès, E.; Largeau, C.; Beaudoin, B. Protection of organic matter by mineral matrix in a Cenomanian black shale. *Org. Geochem.* **2000**, *31*, 463−474.
(31) Hatcher, P. G.; Spiker, E. C.; Szeverenyi, N. M.; Maciel, G. E. Selective preservation and origin of petroleum-forming aquatic kerogen. *Nature* **1983**, *305*, 498−501.
(32) Tegelaar, E.; De Leeuw, J.; Derenne, S.; Largeau, C. A reappraisal of kerogen formation. *Geochim. Cosmochim. Acta* **1989**, *53*, 3103−3106.
(33) Eglinton, T. I. Carbon isotopic evidence for the origin of macromolecular aliphatic structures in kerogen. *Org. Geochem.* **1994**, *21*, 721−735.
(34) Philp, R.; Calvin, M. Possible origin for insoluble organic (kerogen) debris in sediments from insoluble cell-wall materials of algae and bacteria. *Nature* **1976**, *262*, 134−136.
(35) Salmon, V.; Derenne, S.; Largeau, C.; Beaudoin, B.; Bardoux, G.; Mariotti, A. Kerogen chemical structure and source organisms in a Cenomanian organic-rich black shale (Central Italy) — Indications

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

for an important role of the "sorptive protection" pathway. *Org. Geochem.* **1997**, *27*, 423−438.
(36) Stankiewicz, B.; Scott, A.; Collinson, M. E.; Finch, P.; Mösle, B.; Briggs, D.; Evershed, R. Molecular taphonomy of arthropod and plant cuticles from the Carboniferous of North America: implications for the origin of kerogen. *J. Geol. Soc.* **1998**, *155*, 453−462.
(37) Poirier, N.; Derenne, S.; Balesdent, J.; Rouzaud, J.-N.; Mariotti, A.; Largeau, C. Abundance and composition of the refractory organic fraction of an ancient, tropical soil (Pointe Noire, Congo). *Org. Geochem.* **2002**, *33*, 383−391.
(38) Ransom, B.; Kim, D.; Kastner, M.; Wainwright, S. Organic matter preservation on continental slopes: importance of mineralogy and surface area. *Geochim. Cosmochim. Acta* **1998**, *62*, 1329−1345.
(39) Zimmerman, A. R.; Goyne, K. W.; Chorover, J.; Komarneni, S.; Brantley, S. L. Mineral mesopore effects on nitrogenous organic matter adsorption. *Org. Geochem.* **2004**, *35*, 355−375.
(40) Mayer, L. M.; Schick, L. L.; Hardy, K. R.; Wagai, R.; McCarthy, J. Organic matter in small mesopores in sediments and soils. *Geochim. Cosmochim. Acta* **2004**, *68*, 3863−3872.
(41) Schouten, S.; van Driel, G. B.; Sinninghe Damsté, J. S.; de Leeuw, J. W. Natural sulphurization of ketones and aldehydes: A key reaction in the formation of organic sulphur compounds. *Geochim. Cosmochim. Acta* **1993**, *57*, 5111−5116.
(42) Kohnen, M. E. L.; Sinninghe Damste, J. S.; De Leeuw, J. W. Biases from natural sulphurization in palaeoenvironmental reconstruction based on hydrocarbon biomarker distributions. *Nature* **1991**, *349*, 775−778.
(43) Sinninghe Damsté, J. S.; Eglinton, T. I.; De Leeuw, J. W.; Schenck, P. A. Organic sulphur in macromolecular sedimentary organic matter: I. Structure and origin of sulphur-containing moieties in kerogen, asphaltenes and coal as revealed by flash pyrolysis. *Geochim. Cosmochim. Acta* **1989**, *53*, 873−889.
(44) Berner, R. A. Sedimentary pyrite formation: An update. *Geochim. Cosmochim. Acta* **1984**, *48*, 605−615.
(45) Emmings, J. F.; Hennissen, J. A. I.; Stephenson, M. H.; Poulton, S. W.; Vane, C. H.; Davies, S. J.; Leng, M. J.; Lamb, A.; Moss-Hayes, V. Controls on amorphous organic matter type and sulphurization in a Mississippian black shale. *Rev. Palaeobot. Palynol.* **2019**, *268*, 1−18.
(46) Filley, T. R.; Freeman, K. H.; Wilkin, R. T.; Hatcher, P. G. Biogeochemical controls on reaction of sedimentary organic matter and aqueous sulfides in holocene sediments of Mud Lake, Florida. *Geochim. Cosmochim. Acta* **2002**, *66*, 937−954.
(47) Al-Ayed, O. S.; Matouq, M.; Anbar, Z.; Khaleel, A. M.; Abu-Nameh, E. Oil shale pyrolysis kinetics and variable activation energy principle. *Appl. Energy* **2010**, *87*, 1269−1272.
(48) Pan, C.; Geng, A.; Zhong, N.; Liu, J.; Yu, L. Kerogen pyrolysis in the presence and absence of water and minerals. 1. Gas components. *Energy Fuels* **2008**, *22*, 416−427.
(49) Mao, J.; Fang, X.; Lan, Y.; Schimmelmann, A.; Mastalerz, M.; Xu, L.; Schmidt-Rohr, K. Chemical and nanometer-scale structure of kerogen and its change during thermal maturation investigated by advanced solid-state 13C NMR spectroscopy. *Geochim. Cosmochim. Acta* **2010**, *74*, 2110−2127.
(50) Espitalié, J.; Ungerer, P.; Irwin, I.; Marquis, F. Primary cracking of kerogens. Experimenting and modelling C1, C2−C5, C6−C15 and C15+ classes of hydrocarbons formed. *Org. Geochem.* **1988**, *13*, 893−899.
(51) Burnham, A. K. Kinetic models of vitrinite, kerogen, and bitumen reflectance. *Org. Geochem.* **2019**, *131*, 50−59.
(52) Burnham, A. K.; Sweeney, J. J. A chemical kinetic model of vitrinite maturation and reflectance. *Geochim. Cosmochim. Acta* **1989**, *53*, 2649−2657.
(53) 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.
(54) Ma, J.; Kang, D.; Wang, X.; Zhao, Y.-P. Defining kerogen maturity from orbital hybridization by machine learning. *Fuel* **2022**, *310*, 122250.
(55) Sweeney, J. J.; Burnham, A. K. Evaluation of a simple model of vitrinite reflectance based on chemical kinetics. *AAPG Bulletin* **1990**, *74*, 1559−1570.
(56) Nielsen, S.; Clausen, O.; McGregor, E. basin % Ro: A vitrinite reflectance model derived from basin and laboratory data. *Basin Res.* **2017**, *29*, 515−536.
(57) Mango, F. D. The light hydrocarbons in petroleum: a critical review. *Org. Geochem.* **1997**, *26*, 417−440.
(58) Hou, L.; Ma, W.; Luo, X.; Tao, S.; Guan, P.; Liu, J. Chemical structure changes of lacustrine Type-II kerogen under semi-open pyrolysis as investigated by solid-state 13C NMR and FT-IR spectroscopy. *Mar. Pet. Geol.* **2020**, *116*, 104348.
(59) Hubbard, A. B.; Robinson, W. E.*A thermal decomposition study of Colorado oil shale*; US Department of the Interior, Bureau of Mines, 1950.
(60) Campbell, J. H.; Gallegos, G.; Gregg, M. Gas evolution during oil shale pyrolysis. 2. Kinetic and stoichiometric analysis. *Fuel* **1980**, *59*, 727−732.
(61) Siskin, M.; Katritzky, A. R. Reactivity of organic compounds in hot water: geochemical and technological implications. *Science* **1991**, *254*, 231−237.
(62) Lai, D.; Zhan, J.-H.; Tian, Y.; Gao, S.; Xu, G. Mechanism of kerogen pyrolysis in terms of chemical structure transformation. *Fuel* **2017**, *199*, 504−511.
(63) Guan, X.-H.; Wang, D.; Wang, Q.; Chi, M.-S.; Liu, C.-G. Estimation of various chemical bond dissociation enthalpies of large-sized kerogen molecules using DFT methods. *Mol. Phys.* **2016**, *114*, 1705−1755.
(64) Van Duin, A. C.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: a reactive force field for hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396−9409.
(65) 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. Hydrog. Energy* **2016**, *41*, 12093−12100.
(66) Wang, X.; Huang, X.; Lin, K.; Zhao, Y.-P. The constructions and pyrolysis of 3D kerogen macromolecular models: Experiments and simulations. *Glob. Chall.* **2019**, *3*, 1900006.
(67) Kostetskyy, P.; Broadbelt, L. J. Progress in modeling of biomass fast pyrolysis: A review. *Energy Fuels* **2020**, *34*, 15195−15216.
(68) Shen, W.; Zhao, Y.-P. Quasi-static crack growth under symmetrical loads in hydraulic fracturing. *J. Appl. Mech.* **2017**, *84*, 081009.
(69) Shen, W.; Zhao, Y.-P. Combined effect of pressure and shear stress on penny-shaped fluid-driven cracks. *J. Appl. Mech.* **2018**, *85*, 031003.
(70) Du, S. Anisotropic rock poroelasticity evolution in ultra-low permeability sandstones under pore pressure, confining pressure, and temperature: experiments with Biot's coefficient. *Acta Geol. Sin.* **2021**, *95*, 937−945.
(71) Du, S. Potential laws on the changes of shale in acid erosion process based on the fast matching method of dimensional analysis. *Int. J. Hydrog. Energy* **2021**, *46*, 7836−7847.
(72) Yu, H.; Xu, H.; Fan, J.; Zhu, Y.-B.; Wang, F.; Wu, H. Transport of shale gas in microporous/nanoporous media: Molecular to pore-scale simulations. *Energy Fuels* **2021**, *35*, 911−943.
(73) Zhou, W.; Zhu, J.; Wang, H.; Kong, D. Transport diffusion behaviors and mechanisms of CO2/CH4 in shale nanopores: Insights from molecular dynamics simulations. *Energy Fuels* **2022**, *36*, 11903−11912.
(74) 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.
(75) Huang, J.; Jin, T.; Barrufet, M.; Killough, J. Evaluation of CO2 injection into shale gas reservoirs considering dispersed distribution of kerogen. *Appl. Energy* **2020**, *260*, 114285.

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117

(76) 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.
(77) Vasileiadis, M.; Peristeras, L. D.; Papavasileiou, K. D.; Economou, I. G. Modeling of bulk kerogen porosity: methods for control and characterization. *Energy Fuels* **2017**, *31*, 6004−6018.
(78) Ho, T. A.; Wang, Y.; Criscenti, L. J. Chemo-mechanical coupling in kerogen gas adsorption/desorption. *Phys. Chem. Chem. Phys.* **2018**, *20*, 12390−12395.
(79) Zhao, Y.-P.*Physical mechanics of surfaces and interfaces*; Science Press: Beijing, 2012.
(80) Zhao, Y.-P.*Nano and mesoscopic mechanics*; Science Press: Beijing, 2014.
(81) 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.
(82) 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.
(83) Zhu, X.; Zhao, Y.-P. Atomic mechanisms and equation of state of methane adsorption in carbon nanopores. *J. Phys. Chem. C* **2014**, *118*, 17737−17744.
(84) Lin, K.; Yuan, Q.; Zhao, Y.-P.; Cheng, C. Which is the most efficient candidate for the recovery of confined methane: Water, carbon dioxide or nitrogen? *Extreme Mech. Lett.* **2016**, *9*, 127−138.
(85) Lin, K.; Zhao, Y.-P. Entropy and enthalpy changes during adsorption and displacement of shale gas. *Energy* **2021**, *221*, 119854.
(86) Lin, K.; Huang, X.; Zhao, Y.-P. Combining image recognition and simulation to reproduce the adsorption/desorption behaviors of shale gas. *Energy Fuels* **2020**, *34*, 258−269.
(87) Yang, Y.; Liu, J.; Yao, J.; Kou, J.; Li, Z.; Wu, T.; Zhang, K.; Zhang, L.; Sun, H. Adsorption behaviors of shale oil in kerogen slit by molecular simulation. *Chem. Eng. J.* **2020**, *387*, 124054.
(88) Wang, T.; Tian, S.; Li, G.; Sheng, M.; Ren, W.; Liu, Q.; Zhang, S. Molecular simulation of CO2/CH4 competitive adsorption on shale kerogen for CO2 sequestration and enhanced gas recovery. *J. Phys. Chem. C* **2018**, *122*, 17009−17018.
(89) Xu, H.; Yu, H.; Fan, J.; Xia, J.; Liu, H.; Wu, H. Formation mechanism and structural characteristic of pore-networks in shale kerogen during in-situ conversion process. *Energy* **2022**, *242*, 122992.
(90) Zeszotarski, J. C.; Chromik, R. R.; Vinci, R. P.; Messmer, M. C.; Michels, R.; Larsen, J. W. Imaging and mechanical property measurements of kerogen via nanoindentation. *Geochim. Cosmochim. Acta* **2004**, *68*, 4113−4119.
(91) Jakob, D. S.; Wang, L.; Wang, H.; Xu, X. G. Spectro-mechanical characterizations of kerogen heterogeneity and mechanical properties of source rocks at 6 nm spatial resolution. *Anal. Chem.* **2019**, *91*, 8883−8890.
(92) Spiro, C. L. Space-filling models for coal: a molecular description of coal plasticity. *Fuel* **1981**, *60*, 1121−1126.
(93) 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.; Coasne, B. Realistic molecular model of kerogen's nanostructure. *Nat. Mater.* **2016**, *15*, 576−582.
(94) Wu, T.; Firoozabadi, A. Mechanical properties and failure envelope of kerogen matrix by molecular dynamics simulations. *J. Phys. Chem. C* **2020**, *124*, 2289−2294.
(95) Wang, X.; Huang, X.; Gao, M.; Zhao, Y.-P. Mechanical response of kerogen at high strain rates. *Int. J. Impact Eng.* **2021**, *155*, 103905.
(96) 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.
(97) Zhao, Y.-P.*Modern continuum mechanics*; Science Press: Beijing, 2016.
(98) Du, S.; Shi, Y. Rapid determination of complete distribution of pore and throat in tight oil sandstone of Triassic Yanchang Formation in Ordos Basin, China. *Acta Geol. Sin.* **2020**, *94*, 822−830.
(99) Perez, F.; Devegowda, D. Spatial distribution of reservoir fluids in mature kerogen using molecular simulations. *Fuel* **2019**, *235*, 448−459.
(100) 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. In *Composition, Geochemistry and Conversion of Oil Shales*; Springer, Dordrecht, 1995; pp 143−158, DOI: 10.1007/978-94-011-0317-6_9.
(101) Vandenbroucke, M. Kerogen: from types to models of chemical structure. *Oil Gas Sci. Technol.* **2003**, *58*, 243−269.
(102) Vandenbroucke, M.Structure of kerogens as seen by investigations on soluble extracts. In *Kerogen*; Editions Technip: Paris, 1980; pp 415−444.
(103) Burlingame, A.; Haug, P. A.; Schnoes, H. K.; Simoneit, B. R. Fatty acids derived from the Green River Formation oil shale by extractions and oxidations—A review. *Adv. Org. Geochem.* **1969**, 85−129.
(104) Burlingame, A.; Simoneit, B. Analysis of the mineral entrapped fatty acids isolated from the Green River Formation. *Nature* **1968**, *218*, 252−256.
(105) Burlingame, A.; Simoneit, B. High resolution mass spectrometry of Green River formation kerogen oxidations. *Nature* **1969**, *222*, 741−747.
(106) Young, D.; Yen, T. The nature of straight-chain aliphatic structures in green river kerogen. *Geochim. Cosmochim. Acta* **1977**, *41*, 1411−1417.
(107) Yen, T.Structural aspects of organic components in oil shales. In *Developments in Petroleum Science*; Elsevier: New York, 1976; pp 129−148.
(108) Oka, M.; Hsueh-Chia, C.; Gavalas, G. R. Computer-assisted molecular structure construction for coal-derived compounds. *Fuel* **1977**, *56*, 3−8.
(109) Faulon, J.; Vandenbroucke, M.; Drappier, J.; Behar, F.; Romero, M. 3D chemical model for geological macromolecules. *Org. Geochem.* **1990**, *16*, 981−993.
(110) Lille, Ü.; Heinmaa, I.; Pehk, T. Molecular model of Estonian kukersite kerogen evaluated by 13C MAS NMR spectra☆. *Fuel* **2003**, *82*, 799−804.
(111) 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.
(112) 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.
(113) Kang, D.; Zhao, Y.-P. Predicting the molecular models, types, and maturity of kerogen in shale using machine learning and multi-NMR spectra. *Energy Fuels* **2022**, *36*, 5749−5761.
(114) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.
(115) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255−260.
(116) Zhou, Z. H.; Liu, S.*Machine Learning*; Springer Nature: Singapore, 2021.
(117) Heaton, J.An empirical analysis of feature engineering for predictive modeling. *SoutheastCon 2016*; IEEE, 2016; pp 1−6, .
(118) Zheng, A.; Casari, A.*Feature engineering for machine learning: principles and techniques for data scientists*; O'Reilly Media: Boston, 2018.
(119) Kuhn, M.; Johnson, K.*Feature engineering and selection: A practical approach for predictive models*; CRC Press: New York, 2019.
(120) Bovey, F. A.; Mirau, P. A.; Gutowsky, H.*Nuclear magnetic resonance spectroscopy*; Elsevier: San Diego, 1988.

(121) Lambert, J. B.; Mazzola, E. P.; Ridge, C. D.*Nuclear magnetic resonance spectroscopy: An introduction to principles, applications, and experimental methods*; John Wiley & Sons: Hoboken, 2019.
(122) Cao, X.; Yang, J.; Mao, J. Characterization of kerogen using solid-state nuclear magnetic resonance spectroscopy: A review. *Int. J. Coal Geol.* **2013**, *108*, 83−90.
(123) 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.
(124) Tahmasebi, P.; Javadpour, F.; Sahimi, M. Data mining and machine learning for identifying sweet spots in shale reservoirs. *Expert Syst. Appl.* **2017**, *88*, 435−447.
(125) Luo, G.; Tian, Y.; Bychina, M.; Ehlig-Economides, C.Production optimization using machine learning in Bakken shale. In *Unconventional Resources Technology Conference*, Houston, Texas, July 23−25, 2018; pp 2174−2197, DOI: 10.15530/urtec-2018-2902505.
(126) Asala, H.; Chebeir, J.; Zhu, W.; Gupta, I.; Taleghani, A. D.; Romagnoli, J.A machine learning approach to optimize shale gas supply chain networks. In *SPE Annual Technical Conference and Exhibition*; 2017; pp 1−28, DOI: 10.2118/187361-MS.
(127) Kasyap, S.; Senetakis, K. Characterization of two types of shale rocks from Guizhou China through micro-indentation, statistical and machine-learning tools. *J. Pet. Sci. Eng.* **2022**, *208*, 109304.
(128) 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.
(129) Banerjee, P.; Erehman, J.; Gohlke, B.-O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II—A database of natural products. *Nucleic Acids Res.* **2015**, *43*, D935−D939.
(130) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(131) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.
(132) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(133) Brown, N. Chemoinformatics—An introduction for computer scientists. *ACM Comput. Surv.* **2009**, *41*, 1−38.
(134) 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. *Proceedings of Advances in Neural Information Processing Systyems 28*, Montreal, Canada, Dec 7−12, 2015; pp 2215−2223.
(135) Bajusz, D.; Rácz, A.; Héberger, K. Chemical data formats, fingerprints, and other molecular descriptions for database analysis and searching. *Comprehensive Medicinal Chemistry III* **2017**, 329−378.
(136) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(137) Winter, R.; Montanari, F.; Noe, F.; Clevert, D. A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10*, 1692−1701.
(138) Honda, S.; Shi, S.; Ueda, H. R.SMILES Transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv***2019**, arXiv preprint:1911.04738.
(139) Le, T.; Winter, R.; Noé, F.; Clevert, D.-A. Neuraldecipher − Reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chem. Sci.* **2020**, *11*, 10378−10389.
(140) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(141) Cereto-Massague, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallve, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58−63.
(142) Riniker, S.; Landrum, G. A. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminf.* **2013**, *5*, 1−7.
(143) Yao, H.; Dai, Q.; You, Z. Molecular dynamics simulation of physicochemical properties of the asphalt model. *Fuel* **2016**, *164*, 83−93.
(144) Yao, H.; Liu, J.; Xu, M.; Bick, A.; Xu, Q.; Zhang, J. Generation and properties of the new asphalt binder model using molecular dynamics (MD). *Sci. Rep.* **2021**, *11*, 9890.
(145) Li, D. D.; Greenfield, M. L. Chemical compositions of improved model asphalt systems for molecular simulations. *Fuel* **2014**, *115*, 347−356.
(146) Luo, G.; Tian, Y.; Bychina, M.; Ehlig-Economides, C. Production-strategy insights using machine learning: Application for bakken shale. *SPE Reservoir Eval. Eng.* **2019**, *22*, 800−816.
(147) Calderón, A. J.; Pekney, N. J. Optimization of enhanced oil recovery operations in unconventional reservoirs. *Appl. Energy* **2020**, *258*, 114072.
(148) Lee, K. J. Characterization of kerogen content and activation energy of decomposition using machine learning technologies in combination with numerical simulations of formation heating. *J. Pet. Sci. Eng.* **2020**, *188*, 106860.
(149) Sun, F.; Du, S.; Zhao, Y.-P. Fluctuation of fracturing curves indicates in-situ brittleness and reservoir fracturing characteristics in unconventional energy exploitation. *Energy* **2022**, *252*, 124043.
(150) Syed, F. I.; Alnaqbi, S.; Muther, T.; Dahaghi, A. K.; Negahban, S. Smart shale gas production performance analysis using machine learning applications. *Pet. Res.* **2022**, *7*, 21−31.
(151) Bhattacharya, S.; Ghahfarokhi, P. K.; Carr, T. R.; Pantaleone, S. Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: A case study from the Marcellus Shale, North America. *J. Pet. Sci. Eng.* **2019**, *176*, 702−715.
(152) Arif, M.; Zhang, Y.; Iglauer, S. Shale wettability: Data sets, challenges, and outlook. *Energy Fuels* **2021**, *35*, 2965−2980.
(153) Sharifigaliuk, H.; Mahmood, S. M.; Al-Bazzaz, W.; Khosravi, V. Complexities driving wettability evaluation of shales toward unconventional approaches: A comprehensive review. *Energy Fuels* **2021**, *35*, 1011−1023.
(154) Shahkarami, A.; Ayers, K.; Wang, G.; Ayers, A.Application of machine learning algorithms for optimizing future production in marcellus shale, case study of southwestern pennsylvania.*SPE/AAPG Eastern Regional Meeting*, Oct 7−11, 2018; pp 1−14, .
(155) Al-Alwani, M.; Britt, L.; Dunn-Norman, S.; Alkinani, H. H.; Al-Hameedi, A. T.; Al-Attar, A.Production performance estimation from stimulation and completion parameters using machine learning approach in the Marcellus Shale. *53rd US Rock Mechanics/Geomechanics Symposium*, June 23−26, 2019; pp 1−14.
(156) Chen, S.; Yu, H.; Lu, M.; Lebedev, M.; Li, X.; Yang, Z.; Cheng, W.; Yuan, Y.; Ding, S.; Johnson, L. A new approach to calculate gas saturation in shale reservoirs. *Energy Fuels* **2022**, *36*, 1904−1915.
(157) Chandra, D.; Vishal, V. A critical review on pore to continuum scale imaging techniques for enhanced shale gas recovery. *Earth Sci. Rev.* **2021**, *217*, 103638.
(158) Misra, S.; Wu, Y.Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. In: *Machine Learning for Subsurface Characterization*; Gulf Professional Publishing, 2019; pp 289−298.
(159) Anderson, T. I.; Vega, B.; Kovscek, A. R. Multimodal imaging and machine learning to enhance microscope images of shale. *Comput. Geosci.* **2020**, *145*, 104593.
(160) Wu, Y.; Misra, S.; Sondergeld, C.; Curtis, M.; Jernigen, J. Machine learning for locating organic matter and pores in scanning electron microscopy images of organic-rich shales. *Fuel* **2019**, *253*, 662−676.

(161) Goral, J.; Walton, I.; Andrew, M.; Deo, M. Pore system characterization of organic-rich shales using nanoscale-resolution 3D imaging. *Fuel* **2019**, *258*, 116049.
(162) Li, X.; Liu, Z.; Cui, S.; Luo, C.; Li, C.; Zhuang, Z. Predicting the effective mechanical property of heterogeneous materials by image based modeling and deep learning. *Comput. Methods Appl. Mech. Eng.* **2019**, *347*, 735−753.
(163) 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.
(164) Zhang, H.; Yu, H.; Meng, S.; Huang, M.; Micheal, M.; Su, J.; Liu, H.; Wu, H. Fast and accurate reconstruction of large-scale 3D porous media using deep learning. *J. Pet. Sci. Eng.* **2022**, *217*, 110937.
(165) Wang, L.; Liu, M.; Altazhanov, A.; Syzdykov, B.; Yan, J.; Meng, X.; Jin, K.Data driven machine learning models for shale gas adsorption estimation. *SPE Europec*, Dec 1−3, 2020; pp 1−13, .
(166) Meng, M.; Zhong, R.; Wei, Z. Prediction of methane adsorption in shale: Classical models and machine learning based models. *Fuel* **2020**, *278*, 118358.
(167) Nait Amar, M.; Larestani, A.; Lv, Q.; Zhou, T.; Hemmati-Sarapardeh, A. Modeling of methane adsorption capacity in shale gas formations using white-box supervised machine learning techniques. *J. Pet. Sci. Eng.* **2022**, *208*, 109226.
(168) Georgescu, I. How machines could teach physicists new scientific concepts. *Nat. Rev. Phys.* **2022**, DOI: 10.1038/s42254-022-00497-5.
(169) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(170) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706−710.
(171) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I.Attention is all you need. *Advances in Neural Information Processing Systems 30*, Long Beach, CA, Dec 4−9, 2017; Vol. *30*, pp 1−11.

https://doi.org/10.1021/acs.energyfuels.2c03307
Energy Fuels 2023, 37, 98−117