Petroleum Science 21 (2024) 4446–4453

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

# Impact of molecular composition on viscosity of heavy oil: Machine learning based on semi-quantitative analysis results from high-resolution mass spectrometry

Qian-Hui Zhao, Jian-Xun Wu *, Tian-Hang Zhou, Suo-Qi Zhao, Quan Shi

*State Key Laboratory of Heavy Oil Processing, Petroleum Molecular Engineering Center (PMEC), China University of Petroleum, Beijing, 102249, China*

ARTICLE INFO

*Article history:*
Received 11 October 2023
Received in revised form
27 March 2024
Accepted 29 March 2024
Available online 30 March 2024

Edited by Min Li

*Keywords:*
Heavy oil
High resolution mass spectrometry
Machine learning
Viscosity

ABSTRACT

The primary impediment to the recovery of heavy oil lies in its high viscosity, which necessitates a deeper understanding of the molecular mechanisms governing its dynamic behavior for enhanced oil recovery. However, there remains a dearth of understanding regarding the complex molecular composition inherent to heavy oil. In this study, we employed high-resolution mass spectrometry in conjunction with various chemical derivatization and ionization methods to obtain semi-quantitative results of molecular group compositions of 35 heavy oils. The gradient boosting (GB) model has been further used to acquire the feature importance rank (FIR). A feature is an independently observable property of the observed object. Feature importance can measure the contribution of each input feature to the model prediction result, indicate the degree of correlation between the feature and the target, unveil which features are indicative of certain predictions. We have developed a framework for utilizing physical insights into the impact of molecular group compositions on viscosity. The results of machine learning (ML) conducted by GB show that the viscosity of heavy oils is primarily influenced by light components, specifically small molecular hydrocarbons with low condensation degrees, as well as petroleum acids composed of acidic oxygen groups and neutral nitrogen groups. Additionally, large molecular aromatic hydrocarbons and sulfoxides also play significant roles in determine the viscosity.


## 1. Introduction

Heavy oil accounts for over 70% of global oil resources and is one of the most important fossil energy sources in the future. However, the exploitation of heavy oil poses a worldwide challenge to its high viscosity (Guo et al., 2016; Santos et al., 2014). The development of various recovery techniques has been undertaken to mitigate the excessively high viscosity of heavy oils. It mainly includes hot injection technologies such as steam flooding (Zhao et al., 2015) and in situ combustion (Mahinpey et al., 2007), gas injection technologies (Sun et al., 2017) using natural gas, $CO_2$ and $N_2$ as injection gases, and chemical agent injection technologies, such as catalysts (Tang et al., 2019; Zhou et al., 2017), surfactant (Wang and Lai, 2019), alkalis (Zhang et al., 2016), nanoparticles (Anto et al., 2020), and polymers (Zhang et al., 2021). However, these technologies have inherent limitations, and heavy oil still possess significant untapped potential development (Guo et al., 2016; Sun et al., 2017; Zhao et al., 2013). Therefore, it is imperative to thoroughly investigate the factors contributing to high viscosity in order to enhance the recovery of heavy oil.

The high viscosity of petroleum is generally attributed to the significant role played by asphaltenes, which possesses the highest polarity and molecular weight. Asphaltenes exhibit a remarkable structural complexity originating from highly condensed aromatic rings with alkyl moieties of different sizes and functional groups, typically containing oxygen, nitrogen, sulfur, and metal elements. The aggregation phenomenon arises due to the interaction between heteroatoms and aromatic rings (McKenna et al., 2019). Luo and Gu (2007) prepared 11 reconstituted heavy oil samples by adding precipitated asphaltenes into deasphalted heavy oils with varying asphaltene contents. The study revealed that the viscosity is significantly influenced by the state of asphaltene particles, which undergo changes based on variations in asphaltene content and temperature. Hasan and Shaw (2010) proposed that differences

* Corresponding author.
*E-mail address:* wjx@cup.edu.cn (J.-X. Wu).

https://doi.org/10.1016/j.petsci.2024.03.026

in the solubility of asphaltenes in different oils would change the aggregation form of asphaltenes, which means low solubility of asphaltenes would result in high viscosity of crude oil. However, some studies (Ilyin and Strelets, 2018; Li et al., 2018a, 2018b) have shown that crude oils with low asphaltene content can exhibit extremely high viscosity as well. This suggests that the underlying mechanism governing heavy oil viscosity cannot be comprehended solely through composition property. Undoubtedly, the incorporation of asphaltene into crude oil will result in an increase in viscosity. However, the difference in viscosity among different crude oils does not solely depend on the asphaltene content, but essentially relies on the composition and molecular interactions within the crude oil, including π-π interaction of polycyclic aromatics, hydrogen bonding between heteroatoms, acid-base interaction, coordination interaction involving metals, as well as the intertwine of long-chain aliphatic and naphthenic hydrocarbons (Alomair and Almusallam, 2013; Ghanavati et al., 2013; Larter et al., 2008; Luo and Gu, 2007; Muraza, 2015; Zhu et al., 2004). Therefore, a comprehensive analysis of the molecular composition of heavy oil becomes imperative for elucidating the underlying mechanism driving changes in viscosity.

It is difficult to quantitatively characterize such a complex system contains billions of molecules (Beens and Brinkman, 2000), hence most of the results are qualitative. Li et al. (2020) characterized the molecular composition of a fluid catalytic cracking decant oil using a high-resolution Orbitrap MS coupled with ESI and APPI ionization sources. The authors provided a semi-quantitative result of 7001 molecules of 20 class species. Li et al. (2023) analyzed the molecular composition of heavy petroleum fractions through high-resolution mass spectrometry. More than 5000 molecules were quantitatively characterized from four heavy petroleum fractions. The accuracy of the analysis was acceptable according to the H/C ratio comparation between that derived from element analysis and the semi-quantitative molecular composition. High resolution mass spectrometry (HRMS) enables the identification of thousands to tens of thousands of petroleum components at a molecular scale (Hughey et al., 2002; Qian et al., 2001) providing unparalleled analytical capabilities for characterization complex mixtures. Although HRMS currently allows for only semi-quantitative analysis of these molecules, it remains an important method for investigating the molecular composition of heavy petroleum. This study referred to Li's (Li et al., 2020) work to use high resolution mass spectrometry to analyze compositions of heavy oils. Different from the traditional classification method of SARA group composition, this study quantitatively calculates the molecular mass according to the ionization properties of the compounds in the ionization source, and the obtained molecular group composition has specific properties, so the classification is more refined and more conducive to the study of viscosity mechanism.

Machine learning (ML) techniques are utilized to extract actionable insights from big data generated from simulations, not only from limited experimental data so that can greatly save time and cost, make a transformative impact on chemical sciences. Achieving this goal requires a fusion of computer science and chemical science knowledge (Keith et al., 2021). The extent that ML terms appear in scientific papers aligned by American Chemical Society (ACS) technical divisions show that analytical division has No. 2 occurrences in the last two decades. The combination of ML and spectrometry characterization is the hotspot of spectrometry data visualization (Keith et al., 2021). It is currently most widely used in biochemistry, metabolomics, medicines and other related fields (Liebal et al., 2020; Mortier et al., 2021; Mowbray et al., 2021; van Oosten and Klein, 2020; Zien et al., 2009), but has relatively few applications in petroleum chemistry field. Raljević et al. (2021) used statistical multi-way analysis and extensive ML multivariate linear regression methods to model crude oil stability in relation to NMR spectra and other measured properties, such as aromaticity, API gravity, percentage of aliphatic chains, asphaltene content and relative diffusivities. Kirch et al. (2020) combined ML techniques with classical molecular dynamics simulations (MD) to predict oil/brine interfacial tensions (IFT). They built a consistent IFT data set through MD simulations for gradient boosted (GB) algorithms ML training. The obtained model had error 2% and 9% against MD and experimental data from the literature, respectively.

In this study, the molecular composition of 35 heavy crude oils from various oilfields in China was characterized semi-quantitatively using high-resolution Orbitrap mass spectrometry. Subsequently, the ML gradient boosting algorithm was employed to investigate the feature-importance ranking of elements, SARA compositions, and molecular group compositions with respect to the viscosity of heavy oils.

## 2. Experiment

### 2.1. Materials

Analytical grade *n*-hexane (*n*-$C_6$), toluene, methanol carbon tetrachloride ($CCl_4$), acetonitrile ($CH_3CN$), chloroform ($CHCl_3$), and tetrahydrofuran (THF) were obtained from Beijing Chemical Reagents Company, subjected to distillation for purification, and stored in glass containers before use. Ruthenium trichloride ($RuCl_3$) was obtained from J&K Chemical Ltd. Sodium periodate ($NaIO_4$) and potassium hydroxide (KOH) were purchased from Beijing Chemical Reagents Company.

Silver tetrafluoroborate ($AgBF_4$) and methyl iodide ($CH_3I$) were purchased from J&K Chemical, Ltd. A total of 35 heavy crude oils were from Xinjiang, Shengli, Liaohe, and Henan oilfields in China. The element composition and viscosity of each oil are shown in Fig. 3.

### 2.2. Chemical and property analysis

The elemental analysis of carbon, hydrogen, sulfur, oxygen, and nitrogen was performed according to ASTM D5291, ASTM D5453, ASTM D5622, and ASTM D57621 methods, respectively.

The saturated and aromatic fractions were obtained by eluting the adsorbed oil sample on the $Al_2O_3$ column with solvents *n*-$C_6$ and toluene. The mass ratio of saturated fraction to aromatic fraction was required for the semi-quantitative calculation.

The viscosity was determined by rotary viscometer according to a Chinese standard method SY/T 0520−2008.

### 2.3. Semi-quantitative analysis of oils

The semi-quantitative analysis refers to the published work by Li et al. (2020). A brief summary is as follows:

The polar compounds, including acidic oxygen-containing compounds, neutral nitrogen-containing compounds, and basic nitrogen-containing compounds were analyzed by ±ESI HRMS directly. The aromatic hydrocarbons were analyzed by +APPI HRMS.

The saturated hydrocarbons were oxidized through ruthenium ion-catalyzed oxidation (RICO) derivatization, converting saturated hydrocarbons to monohydric alcohols, to ensure the response in −ESI HRMS (Zhou et al., 2012). Similarly, non-polar sulfur-containing compounds were methylated with methyl iodide ($CH_3I$) in the presence of silver tetrafluoroborate ($AgBF_4$), and the product methyl-sulfonium was characterized through + ESI HRMS (Muller and Andersson, 2005; Shi et al., 2010).

The mass percentage of molecules in each oil sample was

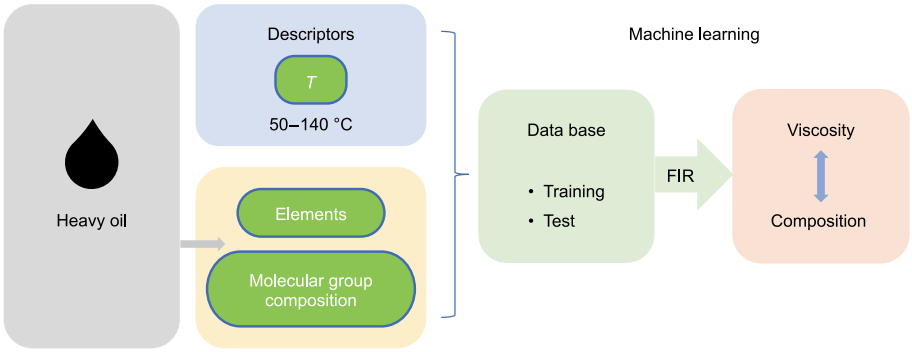


**Fig. 1.** Contribution of composition to viscosity is obtained through ML and HRMS.

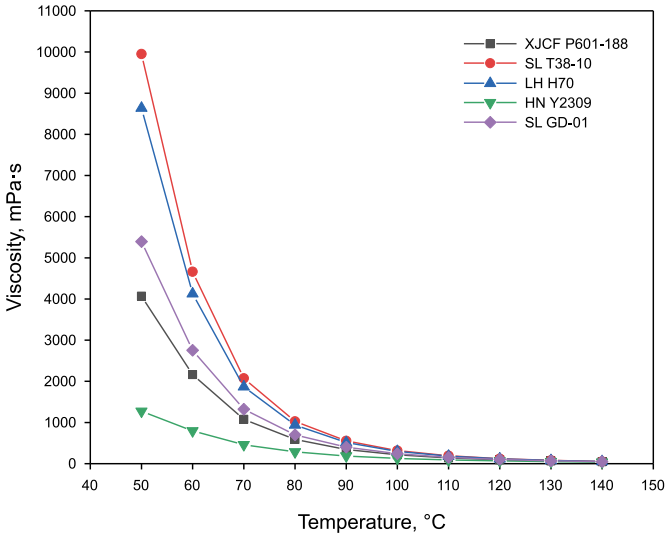


**Fig. 2.** Viscosity-temperature curves of representative heavy oils.

calculated and normalized based on the mass peak intensity, the mass ratio of the saturated and aromatic fractions, and the elemental composition. Additionally, we mandatorily specify that one-third of the total nitrogen is basic nitrogen.

### 2.4. Orbitrap MS analysis

The heavy crude oils, RICO derivatizations, and methylation products were firstly dissolved in toluene (10 mg $mL^{-1}$). Then they were diluted with toluene/methanol (1:3, v/v) for −ESI HRMS, toluene/methanol (1:1, v/v) for +ESI HRMS, and pure toluene for +APPI HRMS (0.02 mg $mL^{-1}$), respectively. The MS characterization was carried out using an Orbitrap mass spectrometer (Orbitrap Fusion, Thermo Scientific, USA). The test samples were injected directly into the ESI through an injection pump. The ion transfer tube temperature was 300 °C, and the vaporizer temperature was 100 °C. The resolution was up to 500,000 at m/z 200 Da. The ions in the range from m/z 150 to m/z 1000 were recorded in a 0.5 min detection period. The sheath, auxiliary, and sweep gas flow rates were 5.0, 2.0, 0.1 arbitrary units for ± ESI and 8.0, 3.0, 0.1 arbitrary units for +APPI, respectively. The MS data analysis was carried out using Thermo. Xcalibur Qual Browser software.

### 2.5. Modeling of viscosity-composition through machine learning

ML can be divided into four categories: supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning (Schmidt et al., 2019). Supervised learning is to train an optimal model based on the existing data set and the relationship between the input (feature) and output (label). Supervised learning tasks mainly include classification and regression. In this process, the samples in the data set are called training samples, and each sample has an input feature and corresponding label (classification task) or target value (regression task). Gradient boosting algorithm, proposed by Friedman (2001), is a typical classification algorithm. The basic principle is to train newly added weak learners base on the negative gradient of the current model loss function, to improve the accuracy and robustness of the final model, overcome the shortcomings of existing weak learners through iteration and reducing the bias (Konstantinov and Utkin, 2021). The loss function is defined as the residual between the real results and ML predictions.

Data split is a necessary step in ML models to evaluate prediction accuracy by splitting data into training and test datasets. The sample size for training and testing can be determined using the learning curve approach (Pedregosa et al., 2011). The loss function of ML models regarding the training data is minimized to improve the prediction accuracy on training data.

The value of determination coefficient ($R^2$, from 0 to 1) can be used to assess how well a regression equation fits the observed values, mathematically as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{p} (y_i - f_i)^2}{\sum_{i=1}^{p} (y_i - \overline{y})^2}$$

where $y_i$ and $f_i$ represent the "real" results in the data set and the predicted values from ML models, respectively; $p$ is the number of data points in the search space.

Hyperparameter optimization can improve the fitting ability of the ML model. By changing the size of the training set, the corresponding $R^2$ score is obtained, and the optimal training set size, that is, the largest $R^2$ value, is selected to obtain the ML model with high prediction accuracy.

The ML gradient boosting algorithms used in this study is implemented with the Scikit-Learn package using Python language. The flow chart in this work is shown in Fig. 1. The temperature, elements composition and molecular group composition are input features.

## 3. Results and discussion

### 3.1. Viscosities and compositions of oils

Fig. 2 shows the viscosity variation with temperature of five representative heavy oils from Xinjiang, Shengli, Liaohe, and Henan

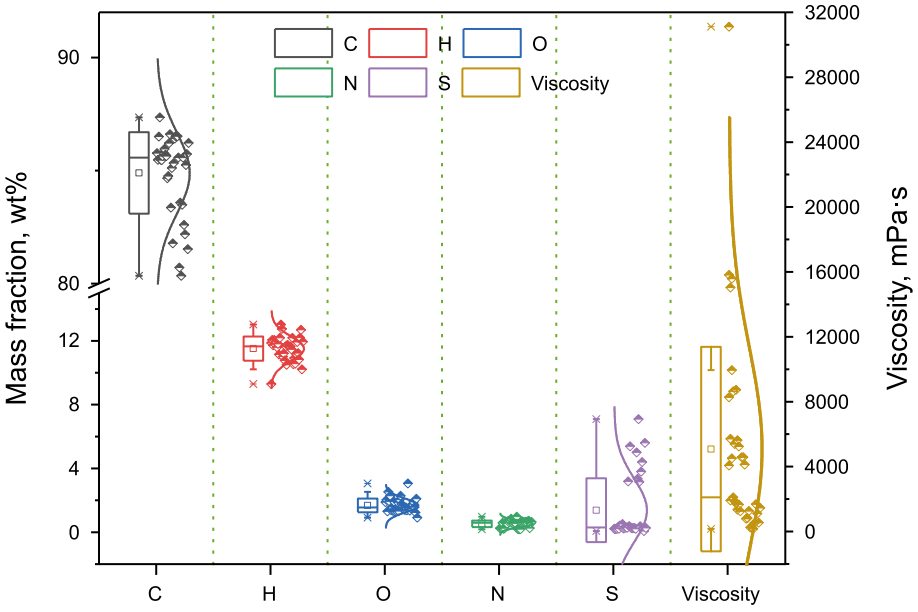


**Fig. 3.** Box diagram of element composition and viscosity of 35 heavy oils.

**Table 1**
Elements compositions and hydrogen-to-carbon ratios of representative heavy oils.

| Heavy oil | C, wt% | H, wt% | O, wt% | N, wt% | S, wt% | H/C |
|---|---|---|---|---|---|---|
| XJCF P601-188 | 85.78 | 11.87 | 1.89 | 0.25 | 0.22 | 1.66 |
| SL T38-10 | 85.71 | 11.14 | 1.92 | 0.71 | 0.52 | 1.56 |
| LH H70 | 86.21 | 11.02 | 1.88 | 0.58 | 0.30 | 1.53 |
| HN Y2309 | 85.33 | 11.75 | 1.65 | 0.98 | 0.29 | 1.65 |
| SL GD-01 | 83.58 | 10.92 | 1.53 | 0.62 | 3.35 | 1.57 |

oilfields. The viscosities of 35 heavy oils range from 167 to 9951 mPa s at 50 °C. The viscosity decreases exponentially with temperature, in accordance with the non-Newtonian fluidity law.

The organic elemental compositions and H/C of representative heavy oils are shown in Table 1. Fig. 3 visually illustrates the differences among 35 heavy oil samples. The large dispersion observed in both element composition and viscosity data of heavy oils indicates that the ML results obtained in this study hold universal applicability.

Fig. 4(a) shows the semi-quantitative molecular compositions of five representative heavy oils of the 35 heavy oils. We firstly calculate the content of each molecule through semi-quantitative approach and then added together to obtain the molecular group compositions. The types of molecular group compositions are classified according to different HRMS characterization methods. Each type of compound consists of hundreds of molecules. For example, the saturated hydrocarbons in XJCF P601-188 account for 45% in Fig. 4(a), which is the sum of the content each saturate molecule. Fig. 4(b) shows the composition of saturate molecules in XJCF P601-188 oil. The horizontal coordinate is the number of carbons, and the vertical coordinate is the number of hydrogens. The colors, as shown in the illustration, represent the percentage of the content of the molecule. The dot marked by the dotted line in the figure corresponds to the carbon number of 29, the hydrogen number of 46, and the content corresponding to the color is 0.69%, that is the saturated hydrocarbon of molecular formula $C_{29}H_{46}$ in

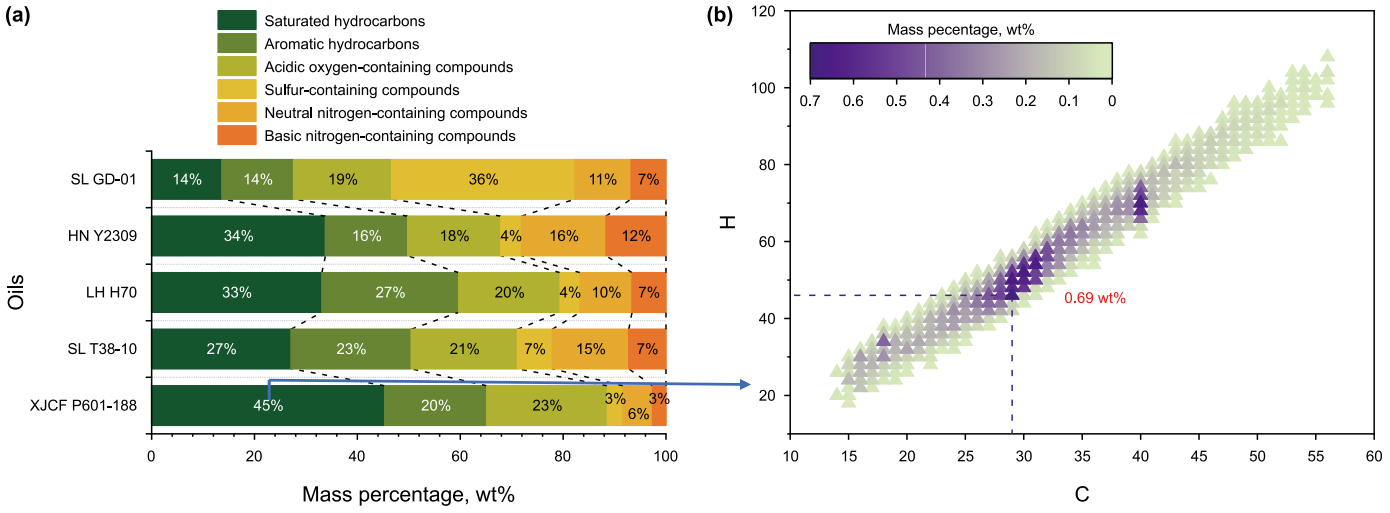


**Fig. 4.** Molecular group compositions of representative heavy oils **(a)** and the semi-quantitative molecular composition of saturated hydrocarbons of XJCF P601-188 oil **(b)**.

XJCF P601-188 oil accounts for 0.69 wt%.

### 3.2. Correlations between molecular composition and viscosity

Fig. 5 shows the feature importance of elements C, H, O, N and S simulated by ML. The $R^2$ is 0.83 illustrates that the model fits well. The FIR represents the contribution of features to viscosity, including both positive and negative contributions. The H element exhibits the highest FIR among all the elements. Generally, the H/C ratio serves as an indicator for assessing condensation degree and determining whether the oil is light or heavy. In addition, our findings indicate that heteroatoms exhibit a lesser contribution compared to hydrogen, which has not been previously documented in the existing literature Notably, nitrogen demonstrates the highest contribution among heteroatoms, followed by oxygen and sulfur. Heteroatoms predominantly reside within the heavier fractions of crude oil, and their abundance can serve as an indicator of crude oil quality. The viscosity of heavy oil is affected by the dilution effect of light components, the π-π interaction of polycyclic aromatic hydrocarbons, and the interaction between heteroatomic compounds (acid-base interaction, hydrogen bond interaction, etc.); however, their respective contributions remain unclear. This conclusion indicates that the impact of dilution and the degree of molecular condensation outweighs the influence of interaction among heteroatomic compounds.

Fig. 6 shows the correlations between group composition and viscosity. Viscosity of heavy oil essentially depends on intermolecular interaction. In this study, the molecules are divided into different groups according to their responses in different ionization sources. It ensures that each group of molecules has similar chemical properties, similar intermolecular interaction mechanisms, so as to similar contributions to viscosity. SH, AH, AO, S, NN and BN represents to saturated hydrocarbons, aromatic hydrocarbons, acidic oxygen-containing compounds, sulfur-containing compounds, neutral nitrogen-containing compounds and basic nitrogen-containing compounds, respectively. Compared with the correlations of element composition, the only difference in group composition's contribution to viscosity is sulfur-containing compounds. Element S contributes less than O, while sulfur-containing compounds contribute more than acidic oxygen-containing compounds. Part of the contribution of sulfur compounds comes from sulfide, and the other parts come from multi-heteroatomic compounds containing both sulfur and oxygen. Methylation reagents also react with these compounds, and they are classified into sulfur-containing compounds. A part of the contribution of oxygen is transferred to the sulfur-containing compounds, making the sulfur-containing compounds contribute more. Saturated and aromatic hydrocarbons, entirely composed of H and C, contribute the first and second, respectively. Among heteroatomic compounds, nitrogen compounds contribute the most, the same as N element. The contribution of neutral nitrogen-containing compounds (such as pyrrole compounds) is greater than that of basic nitrogen-containing compounds (such as pyridines).

Our previous study showed that the separation or addition of petroleum acid could significantly reduce or increase the viscosity of heavy oil (Zhao et al., 2023). In this study, AO and NN ionized in −ESI ionization source both showed acidity. The sum of their feature importance to viscosity is much greater than that of other compounds, indicating that petroleum acids are the most important factors in viscosity of heavy oil.

To further study the correlations between specific substances and viscosity in each group, the molecules of each group were classified according to their carbon number or heteroatomic composition. The correlations were showed in Fig. 7. The feature

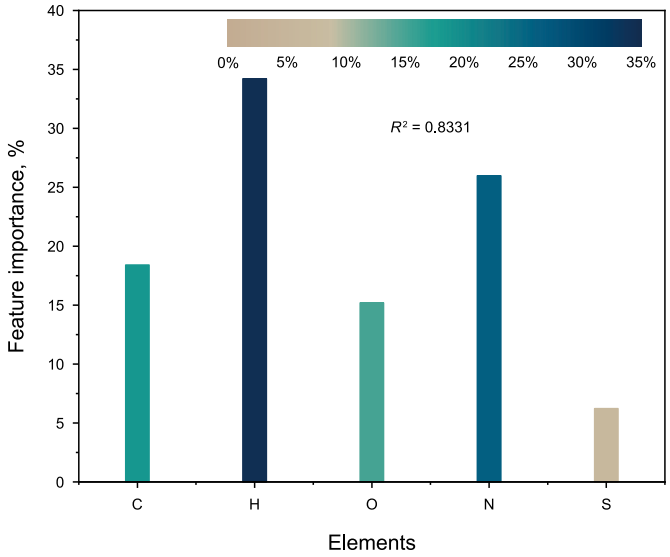


**Fig. 5.** Correlations between element composition and viscosity.

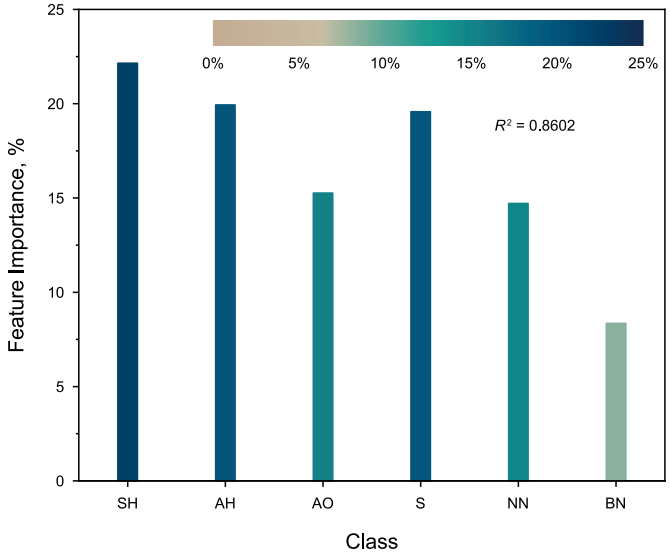


**Fig. 6.** Correlations between group composition and viscosity.

importance of each group was calculated separately, and the results were independent of each other. In Fig. 7(a), the saturated hydrocarbons with carbon numbers between 20 and 30 contribute more to the viscosity. The impact of saturated hydrocarbon with a low carbon number on viscosity is higher than that with a high carbon number. The conclusion is different from that of aromatic hydrocarbons. Aromatic hydrocarbons with carbon numbers less than 20 and more than 50 contribute more, as shown in Fig. 7(b). The saturated hydrocarbons with carbon number less than 40 and the aromatic hydrocarbons with carbon number less than 20 are natural light components in heavy oil. According to Figs. 6 and 7(a) and (b), light components are especially critical for viscosity. It is true that the content of asphaltene has a great influence on the viscosity. In our forthcoming work, however, we find that in sufficient amounts of light components, the aggregation of asphaltenes is inhibited and the effect on viscosity is negligible. This is due to the dilution and dispersion of the light component, so that the viscosity

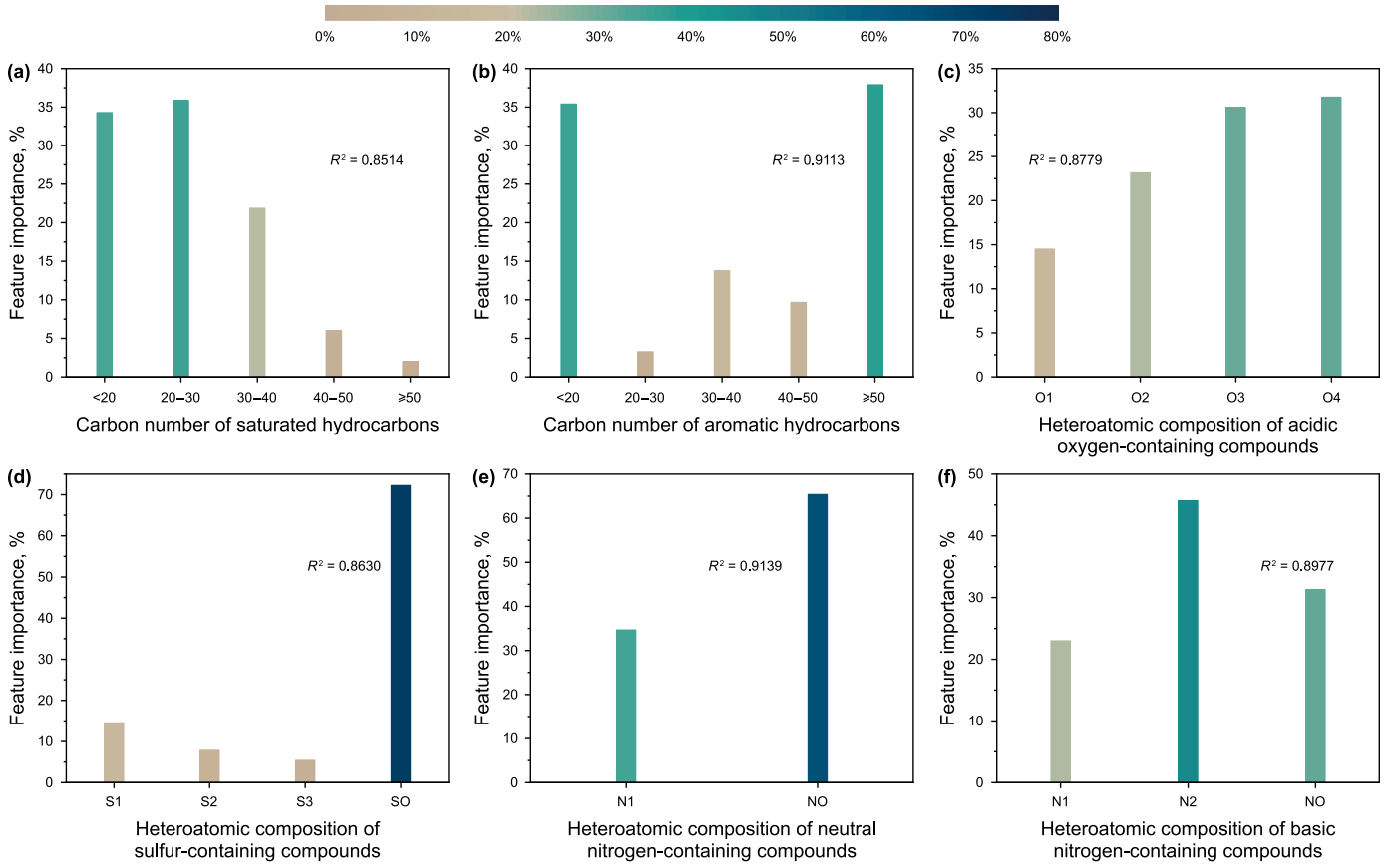


**Fig. 7.** Correlations between viscosity and carbon number of saturated hydrocarbons **(a)**, carbon number of aromatic hydrocarbons **(b)**, heteroatomic composition of acidic oxygen-containing compounds **(c)**, heteroatomic composition of sulfur-containing compounds **(d)**, heteroatomic composition of neutral nitrogen-containing compounds **(e)**, heteroatomic composition of basic nitrogen-containing compounds **(f)**.

is close to the light component, and the viscosity of the asphaltene is no longer a decisive factor. This echoes the conclusion of this study that light components are one of the most important factors affecting the viscosity of heavy oil.

For heteroatomic compounds, oxygen is an important factor. In Fig. 7(c), as the oxygen number increases, the feature importance increases. Oxygen also plays an important role in other groups, especially in sulfur-containing compounds (Fig. 7(d)) and neutral nitrogen-containing compounds (Fig. 7(e)). Poly-sulfur containing compounds contribute little to viscosity compared with mono-sulfur containing compounds (Fig. 7(d)). The nitrogen atoms in basic nitrogen-containing compounds are just the opposite, with compounds containing two nitrogen atoms contributing more than one nitrogen (Fig. 7(f)).

The SO class compounds in Fig. 7(d) ionized in the +ESI ionization source in oils are generally considered to be mainly sulfoxides. The thionyl functional groups (>S═O) contained are highly polar and can generate large intermolecular forces, so their contribution to viscosity is much greater than that of sulfide containing only S element.

## 4. Conclusion

35 heavy oils from different sources were semi-quantitatively characterized by Orbitrap high-resolution mass spectrometry. ML performed the correlation analysis between semi-quantitative molecular group compositions and viscosity data. The determination coefficients of the model are greater than 0.83, indicating that the ML model in this study has a high accuracy.

We find that the order of feature importance of elements to heavy oil viscosity is: H > N > C > O > S, and the order of feature importance of different molecular group compounds is: saturated hydrocarbons > aromatics hydrocarbons > sulfur-containing compounds > acidic oxygen-containing compounds > neutral nitrogen-containing compounds > basic nitrogen-containing compounds. The viscosity contribution of basic nitrogen-containing compounds, sulfur-containing compounds and neutral nitrogen-containing compounds is partly due to the oxygen-containing poly-heteroatomic compounds in them. The light components (include the saturated hydrocarbons with carbon number<30 and the aromatic hydrocarbons with carbon number<20) and the petroleum acids are first two most important compositional factors to viscosity of heavy oil. The aromatic hydrocarbons with high carbon number and the sulfoxides are also important.

Viscosity is an inherent property of heavy oils, resulting from interactions between heteroatomic compounds. However, we found that the contribution of light components to viscosity is greater than that of heteroatomic compounds, and light components are important factors determining the viscosity of heavy oil. This discovery can provide guidance for viscosity reduction of heavy oil, and increase the content of light components by a series of physical and chemical methods, such as: direct injection of light components into heavy oil, adding modifier or injecting energy to produce light components, etc., to improve heavy oil recovery. At the same time, in view of the great contribution of petroleum acid

and sulfoxides to viscosity, the development of targeted viscosity reducing agents can also effectively reduce the viscosity of heavy oil.

The process of semi-quantitative analysis of molecular composition is very complicated. A single heavy oil sample needs to be pretreated 5 times and analyzed by high-resolution mass spectrometry, and then integrated with quantitative algorithm to obtain the results of semi-quantitative analysis of molecular composition. The amount of work here is enormous, so in this study, only 35 heavy oil samples are treated this way. More samples will be analyzed in our future work.

## CRediT authorship contribution statement

**Qian-Hui Zhao:** Writing – original draft, Visualization, Methodology, Investigation, Data curation. **Jian-Xun Wu:** Writing – review & editing, Methodology, Investigation, Data curation. **Tian-Hang Zhou:** Software, Methodology, Investigation, Data curation. **Suo-Qi Zhao:** Investigation, Conceptualization. **Quan Shi:** Writing – review & editing, Conceptualization.

## Declaration of competing interest

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

## Acknowledgement

This work was supported by the National Key R&D Program of China (2018YFA0702400).

## References

Alomair, O.A., Almusallam, A.S., 2013. Heavy crude oil viscosity reduction and the impact of asphaltene precipitation. Energy Fuels 27 (12), 7267–7276. https://doi.org/10.1021/ef4015636.

Anto, R., Deshmukh, S., Sanyal, S., et al., 2020. Nanoparticles as flow improver of petroleum crudes: study on temperature-dependent steady-state and dynamic rheological behavior of crude oils. Fuel 2020 (275), 117873. https://doi.org/10.1016/j.fuel.2020.117873.

Beens, J., Brinkman, U.A.T., 2000. The role of gas chromatography in compositional analyses in the petroleum industry. Trends Anal. Chem. 19 (4), 260–275. https://doi.org/10.1016/S0165-9936(99)00205-8.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Ghanavati, M., Shojaei, M.-J., S. A, A.R., 2013. Effects of asphaltene content and temperature on viscosity of Iranian heavy crude oil: experimental and modeling study. Energy Fuels 27 (12), 7217–7232. https://doi.org/10.1021/ef400776h.

Guo, K., Li, H., Yu, Z., 2016. In-situ heavy and extra-heavy oil recovery: a review. Fuel 2016 (185), 886–902. https://doi.org/10.1016/j.fuel.2016.08.047.

Hasan, M.D.A., Shaw, J.M., 2010. Rheology of reconstituted crude oils: artifacts and asphaltenes. Energy Fuels 24 (12), 6417–6427. https://doi.org/10.1021/ef101185x.

Hughey, C.A., Rodgers, R.P., Marshall, A.G., 2002. Resolution of 11000 compositionally distinct components in a single electrospray ionization Fourier Transform Ion Cyclotron Resonance mass spectrum of crude oil. Anal. Chem. 74, 4145–4149. https://doi.org/10.1021/ac020146b.

Ilyin, S.O., Strelets, L.A., 2018. Basic fundamentals of petroleum rheology and their application for the investigation of crude oils of different natures. Energy Fuels 32 (1), 268–278. https://doi.org/10.1021/acs.energyfuels.7b03058.

Keith, J.A., Vassilev-Galindo, V., Cheng, B., et al., 2021. Combining machine learning and computational chemistry for predictive insights into chemical systems. Chem. Rev. 121 (16), 9816–9872. https://doi.org/10.1021/acs.chemrev.1c00107.

Kirch, A., Celaschi, Y.M., de Almeida, J.M., et al., 2020. Brine-oil interfacial tension modeling: assessment of machine learning techniques combined with molecular dynamics. ACS Appl. Mater. Interfaces 12 (13), 15837–15843. https://doi.org/10.1021/acsami.9b22189.

Konstantinov, A.V., Utkin, L.V., 2021. Interpretable machine learning with an ensemble of gradient boosting machines. Knowl. Base Syst. 222. https://doi.org/10.1016/j.knosys.2021.106993.

Larter, S.R., Adams, J., Gates, I.D., et al., 2008. The origin, prediction and impact of oil viscosity heterogeneity on the production characteristics of tar sand and heavy oil reservoirs. J. Can. Pet. Technol. 47, 1–16. https://doi.org/10.2118/08-01-52.

Li, C., Chen, Y., Hou, J., et al., 2018a. A mechanism study on the viscosity evolution of heavy oil upon peroxide oxidation and pyrolysis. Fuel 2018 (214), 123–126. https://doi.org/10.1016/j.fuel.2017.10.125.

Li, H., Cui, K., Jin, L., et al., 2018b. Experimental study on the viscosity reduction of heavy oil with nano-catalyst by microwave heating under low reaction temperature. J. Petrol. Sci. Eng. 170, 374–382. https://doi.org/10.1016/j.petrol.2018.06.078.

Li, H., Zhang, Y., Xu, C., et al., 2020. Quantitative molecular composition of heavy petroleum fractions: a case study of fluid catalytic cracking decant oil. Energy Fuels 34 (5), 5307–5316. https://doi.org/10.1021/acs.energyfuels.9b03425.

Li, S., Wu, J., Wang, Y., et al., 2023. Semi-quantitative analysis of molecular composition for petroleum fractions using electrospray ionization high-resolution mass spectrometry. Fuel 2023 (335), 127049. https://doi.org/10.1016/j.fuel.2022.127049.

Liebal, U.W., Phan, A.N.T., Sudhakar, M., et al., 2020. Machine learning applications for mass spectrometry-based metabolomics. Metabolites 10 (6). https://doi.org/10.3390/metabo10060243.

Luo, P., Gu, Y., 2007. Effects of asphaltene content on the heavy oil viscosity at different temperatures. Fuel 2007 (86), 1069–1078. https://doi.org/10.1016/j.fuel.2006.10.017.

Mahinpey, N., Ambalae, A., Asghari, K., 2007. In situ combustion in enhanced oil recovery (EOR): a review. Chem. Eng. Commun. 194 (8), 995–1021. https://doi.org/10.1080/00986440701242808.

McKenna, A.M., Chacón-Patiño, M.L., Weisbrod, C.R., et al., 2019. Molecular-level characterization of asphaltenes isolated from distillation cuts. Energy Fuels 33 (3), 2018–2029. https://doi.org/10.1021/acs.energyfuels.8b04219.

Mortier, T., Wieme, A.D., Vandamme, P., et al., 2021. Bacterial species identification using MALDI-TOF mass spectrometry and machine learning techniques: a large-scale benchmarking study. Comput. Struct. Biotechnol. J. 19, 6157–6168. https://doi.org/10.1016/j.csbj.2021.11.004.

Mowbray, M., Savage, T., Wu, C., et al., 2021. Machine learning for biochemical engineering: a review. Biochem. Eng. J. 172. https://doi.org/10.1016/j.bej.2021.108054.

Muller, H., Andersson, J.T., 2005. Characterization of high-molecular-weight sulfur-containing aromatics in vacuum residues using Fourier transform ion cyclotron resonance mass spectrometry. Anal. Chem. 382 (3), 735–741. https://doi.org/10.1007/s00216-004-3026-y.

Muraza, O., 2015. Hydrous pyrolysis of heavy oil using solid acid minerals for viscosity reduction. J. Anal. Appl. Pyrol. 114, 1–10. https://doi.org/10.1016/j.jaap.2015.04.005.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Qian, K., Robbins, W.K., Hughey, C.A., et al., 2001. Resolution and identification of elemental compositions for more than 3000 crude acids in heavy petroleum by negative-ion microelectrospray high-field Fourier transform ion cyclotron resonance mass spectrometry. Energy Fuels 15 (6), 1505–1511. https://doi.org/10.1021/ef010111z.

Raljević, D., Parlov Vuković, J., Smrečki, V., et al., 2021. Machine learning approach for predicting crude oil stability based on NMR spectroscopy. Fuel 2021 (305), 121561. https://doi.org/10.1016/j.fuel.2021.121561.

Santos, R.G., Loh, W., Bannwart, A.C., et al., 2014. An overview of heavy oil properties and its recovery and transportation methods. Braz. J. Chem. Eng. 31 (3), 571–590. https://doi.org/10.1590/0104-6632.20140313s00001853.

Schmidt, J., Marques, M.R.G., Botti, S., et al., 2019. Recent advances and applications of machine learning in solid-state materials science. npj Comput. Mater. 5 (83), 1–36. https://doi.org/10.1038/s41524-019-0221-0.

Shi, Q., Pan, N., Liu, P., et al., 2010. Characterization of sulfur compounds in oil sands bitumen by methylation followed by positive-ion electrospray ionization and Fourier Transform Ion Cyclotron Resonance mass spectrometry. Energy Fuels 24 (5), 3014–3019. https://doi.org/10.1021/ef9016174.

Sun, G., Li, C., Wei, G., et al., 2017. Characterization of the viscosity reducing efficiency of $CO_2$ on heavy oil by a newly developed pressurized stirring-viscometric apparatus. J. Petrol. Sci. Eng. 156, 299–306. https://doi.org/10.1016/j.petrol.2017.06.009.

Tang, X., Zhou, T., Li, J., et al., 2019. Experimental study on a biomass-based catalyst for catalytic upgrading and viscosity reduction of heavy oil. J. Anal. Appl. Pyrol. 143. https://doi.org/10.1016/j.jaap.2019.104684.

van Oosten, L.N., Klein, C.D., 2020. Machine learning in mass spectrometry: a MALDI-TOF MS approach to phenotypic antibacterial screening. J. Med. Chem. 63 (16), 8849–8856. https://doi.org/10.1021/acs.jmedchem.0c00040.

Wang, D., Lai, N., 2019. Development and application of polymetric surfactant emulsification and viscosity reduction system. Petroleum 5 (4), 402–406. https://doi.org/10.1016/j.petlm.2018.12.006.

Zhang, F., Liu, Y., Wang, Q., et al., 2021. Fabricating a heavy oil viscosity reducer with weak interaction effect: synthesis and viscosity reduction mechanism. Colloid and Interface Science Communications 42. https://doi.org/10.1016/j.colcom.2021.100426.

Zhang, H., Chen, G., Dong, M., et al., 2016. Evaluation of different factors on enhanced oil recovery of heavy oil using different alkali solutions. Energy Fuels 30 (5), 3860–3869. https://doi.org/10.1021/acs.energyfuels.6b00196.

Zhao, D.W., Wang, J., Gates, I.D., 2013. Thermal recovery strategies for thin heavy oil reservoirs. Fuel 2013 (117), 431–441. https://doi.org/10.1016/j.fuel.2013.09.023.

Zhao, D.W., Wang, J., Gates, I.D., 2015. An evaluation of enhanced oil recovery strategies for a heavy oil reservoir after cold production with sand. Int. J. Energy Res. 39 (10), 1355–1365. https://doi.org/10.1002/er.3337.

Zhao, Q.H., Ma, S., Wu, J.X., et al., 2023. Molecular composition of naphthenic acids in a Chinese heavy crude oil and their impacts on oil viscosity. Petrol. Sci. 20 (2), 1225–1230. https://doi.org/10.1016/j.petsci.2022.09.016.

Zhou, M., Sun, W., Li, K., 2017. Experimental research of nano catalyst assisted oxidization upgrading of super heavy oil. SCIENTIA SINICA Technologica 47 (2), 197–203. https://doi.org/10.1360/n092016-00307.

Zhou, X., Shi, Q., Zhang, Y., et al., 2012. Analysis of saturated hydrocarbons by redox reaction with negative-ion electrospray Fourier transform ion cyclotron resonance mass spectrometry. Anal. Chem. 84 (7), 3192–31999. https://doi.org/10.1021/ac203035k.

Zhu, Z., Lin, R., Wang, S., 2004. The influence of heavy oil complosition on its viscosity. Xinjing Pet. Geol. 25 (5), 512–513. https://doi.org/10.3969/j.issn.1001-3873.2004.05.016 (in Chinese).

Zien, A., Krmer, N., Sonnenburg, So, et al., 2009. The feature importance ranking measure. Machine Learning and Knowledge Discovery in Databases 2009, 694–709.