energy&fuels

pubs.acs.org/EF Article

Improved Oil Viscosity Characterization by Low-Field NMR Using Feature Engineering and Supervised Learning Algorithms

Strahinja Markovic,* Jonathan L. Bryan, Vladislav Ishimtsev, Aman Turakhanov, Reza Rezaee, Alexey Cheremisin, Apostolos Kantzas, Dmitry Koroteev, and Sudarshan A. Mehta

Cite This: https://dx.doi.org/10.1021/acs.energyfuels.0c02565 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Conventional methods for determining and monitoring the viscosity of oils are time-consuming, expensive, and in some instances, technically unfeasible. These limitations can be avoided using low-field nuclear magnetic resonance (LF-NMR) relaxometry. However, due to the chemical dissimilarity of oils and various temperatures these oils are exposed to, as well as LF-NMR equipment limitations, the commonly used models fail to perform at a satisfactory level, making them impractical for use in heavy oil and bitumen reservoirs and in environments with large temperature oscillations (e.g., mechanical systems). We present a framework that combines supervised learning algorithms with domain knowledge for synthesizing new features to improve model forecasts using only one NMR parameter—T_2 geometric mean. Two principal methods were considered, support vector regression (SVR) and gradient boosted trees (GBRT). Models were trained using the experimental data from our previous studies and literature data combining conventional oils, heavy oils, and bitumens from various reservoirs in Canada and United States. The models' performance was compared against four other intelligent algorithms and four well-known empirical NMR models against which the SVR- and GBRT-based models achieved the highest statistical scores. These two models can be used for oil viscosity prediction in conventional and heavy oil reservoirs with a wide range of oil viscosities and in situations where high precision is needed, such as in the determination of viscosity of petroleum distillates or for monitoring of oil viscosity in mechanical systems. The proposed framework can also be applied to determine other physicochemical properties of oils by LF-NMR, where the application of supervised learning is usually impractical due to the limited volume of experimental data.

1. INTRODUCTION

Viscosity is among the most important physicochemical properties of petroleum products and crude oils. Understanding the behavior of their viscosity in various conditions is essential for the interpretation and modeling of numerous processes in fields of chemical, petroleum, and mechanical engineering. The oil is a blend of a diverse range of liquid hydrocarbons with inconsistent molecular structures.[1] When it has a higher proportion of complex high molecular weight compounds, such as asphaltenes and resins, oil viscosity will tend to be higher, signifying that viscosity is a reflection of oil's chemical complexity.[2] This natural inconsistency of oil compositions elicits a constant demand for the development of new techniques for their efficient characterization. In recent years, the wave of innovation has led to the application of low-field nuclear magnetic resonance (LF-NMR) tools for the characterization of hydrogen-bearing liquids due to their ability to convey series of contactless, noninvasive experiments rapidly. Although this technology has been proved to be a viable tool for observing differences in variable viscosity oils, numerous constraints arise as a consequence of not only the embedded chemical complexity of oils and limitations of LF-NMR devices but also from analytical tools and models used for the interpretation of experimental results.[3−8] Since the former two are technologically challenging to change, one can attempt to improve the analytical tools and frameworks using new mathematical approaches. In such circumstances, the supervised learning (SL) methods have been proven to be useful in developing more reliable mathematical models in many relevant fields such as fuel processing, petrophysical studies of porous mediums, and oil viscosity monitoring equipment in mechanical systems.[9−13]

One potential application is in fuel processing, where there has been a surge for the last few years in the development of fast methods for the characterization of petroleum fractions by LF-NMR.[14] Among many studied physicochemical properties, the oil viscosity was found to be of the principal importance in determining the rate of interaction with fuel during combustion processes in internal combustion engines.[15,16] In these studies, the LF-NMR predictive models were typically derived using multivariate calibration with partial least squares (PLS) regression or artificial neural networks (ANN), which proved to be the right approach in most cases. However, in nonlinear datasets, the reports in the literature show that PLS did not provide satisfactory accuracy, whereas ANN tended to overfit the data, thus leading to poor model generalization.[17,18] In the LF-NMR examination of petroleum fractions, this nonlinearity

Received: July 30, 2020
Revised: October 2, 2020

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

can occur due to their chemical intricacy, which ultimately leads to the degradation of model forecasting performance.[19]

Conversely, in petrophysical studies of unconventional oil reservoirs, the viscosity of the oil may vary up to 3 orders of magnitude, meaning that flow dynamics through the pore space will vary drastically.[3,20] In this sense, the spatial variation of oil viscosity influences not only the calculation of recoverable reserves but also the technical aspects of the recovery scheme, selection of enhanced oil recovery (EOR) methods, and the way the numerical simulation of the reservoirs are performed. Studies have shown that LF-NMR tools can help resolve this problem using the empirical NMR viscosity models, which can account for chemical complexity and a wide span of temperatures with the help of NMR-derived parameters such as the relative hydrogen index (RHI).[7,21] However, the determination of RHI requires a recovery of the representative oil sample from the given formation, preferably with the preserved gas content, for subsequent laboratory measurements, which is often a technically challenging and expensive task.[22,23] Moreover, oil saturation volumes must be determined independently, which is necessary for normalization of measured oil NMR response by the amplitude of an equal quantity of water.[24] Unfortunately, in circumstances like these, the empirical NMR models without RHI do not perform at a satisfactory level for predicting accurate viscosities in heavy oil and bitumen systems.[25−27]

Finally, in mechanical systems (tribosystems), the magnitude of viscosity reflects the oil's capacity to render the sufficient thickness of the lubricating film between the surfaces exposed to friction. To efficiently buffer the rate of machinery wear, the oil selection is made under the speed-load and temperature conditions of the system.[28] The prevention of malfunctions in tribosystems is usually performed by monitoring oil viscosity, where its relative increase may indicate excessive oxidation or contamination of the oil by other fluids. In contrast, its decrease may indicate the beginning of a thermal cracking process, which occurs at high temperatures.[29] In earlier studies, LF-NMR measurements were proposed as an alternative to conventional monitoring approaches that involve direct-contact instruments based on vibration, acoustic, and microdisplacement methods.[13] In circumstances where these instruments would be difficult to utilize, the LF-NMR tools could be used instead for noninvasive, real-time viscosity monitoring.[13,30,31] As the operating conditions of these systems may lead to significant oil viscosity fluctuations, the robust data-driven NMR model could be used to measure the fluctuations accurately, and in this manner, help in early detection of the equipment failure.

In this work, a supervised learning framework was introduced to improve the oil viscosity characterization by LF-NMR relaxometry, using gradient boosting regression trees (GBRT) and support vector regression (SVR).[32,33] A feature engineering (FE) approach was integrated to maximize the forecasting capacity of the models by deriving new features using the knowledge from the NMR oil characterization domain.[34] The study results indicate that this strategy can be successfully applied even to smaller datasets. As the underlying mathematical principals of GBRT and SVR techniques are substantially different, we could observe the study task from different perspectives. The database used for calibration of models in the study was formed out of the previously published LF-NMR crude oil data, which contain over 130 light and heavy oil samples recovered from various reservoirs in Canada and United States.[25,35,36] The study was segmented into two stages. In the first stage, the preprocessing of data was performed together with feature engineering, which enabled the appropriate training of GBRT and SVR models. The generalization ability of the models was assessed by the *K*-fold cross-validation, while model performance was recorded using several statistical metrics. In the second stage, the performance of GBRT and SVR models was compared against other four popular supervised learning algorithms and four well-known empirical NMR viscosity models that were trained using the same framework. The code and the data have been uploaded to the GitHub repository and are available for use.

2. METHODOLOGY

2.1. NMR Theory. Crude oils and their derivatives contain large quantities of hydrogen (H), and modern LF-NMR instruments can detect the response of their H protons within static magnetic fields. Two responses are usually measured: the spin−lattice or T_1 relaxation, and spin−spin or T_2 relaxation. The theoretical and empirical evidence shows that T_2 relaxation has a strong correlation with oil viscosity.[37,38] However, in heavier, more viscous oils, the T_2 relaxation behavior deviates from conventional models, which in turn changes T_2 correlation with viscosity. Although studies are being conducted to better understand the underlying physics of H proton relaxation behavior in heavy oils,[3,4] there is enough scientific evidence to confirm that these variations are associated with the presence of heavy components and their complex molecular structures (e.g., asphaltenes and resins).[19] As a result, the conventional NMR viscosity models for lighter and chemically "simpler" oils cannot correctly describe this change in a relationship, which leads to poor prediction performance. Additional hindrance comes from the rapid relaxation of H protons in heavier oils, where LF-NMR tools cannot detect the actual amount of hydrogen. Numerous heavy oil models were developed in the last 20 years, demonstrating poor stability and generalization, especially when NMR data from "unseen" types of oil is introduced.[5,25] Besides, their performance deteriorates further when the NMR measurements are performed at higher temperatures, where due to the Curie effect, a portion of the NMR signal is lost. As the NMR signal represents the response of H protons detected by the NMR device, this effect is illustrated in Figure 1, where the slope of the NMR signal amplitude of the bitumen sample becomes negative at >380 K, instead of an anticipated increase of the amplitude. Some empirical models, which use the

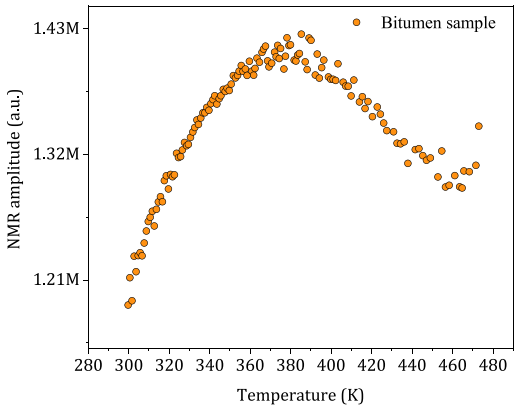

Figure 1. Amplitude of the NMR signal is a function of temperature for a bitumen oil sample. Note the decrease of the amplitude at temperatures above 380 K due to the Curie effect (loss of the H proton magnetization[21]). The plotted data can be found elsewhere.[25]

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

measured hydrogen index (HI) detected by the NMR device or RHI alongside T_2 relaxation data, have proven to be successful even in such conditions since the change in water amplitude and temperature is well known.[7,25] However, for in situ petrophysical surveying by LF-NMR, the values of RHI may not always be known, and for their estimation, independent validation of fluid saturation would be necessary.[24] To avoid these issues, the proposed strategy in this study relies solely on one NMR parameter—T_2 relaxation time. Since spin-echo decay (or NMR signal) after inversion is represented in the T_2 relaxation time-domain spectrum, a single parameter can be derived to characterize the whole spectrum, known as a T_2 geometric mean

$$T_{2\mathrm{gm}} = \exp\left[\sum \frac{A_i}{A} \ln(T_{2i})\right] \tag{1}$$

where A is the total NMR signal amplitude and A_i is the ith amplitude component corresponding to T_{2i} response. The echo-spacing (TE) is the second important parameter that needs to be considered. It relates to the "speed" of NMR signal detection. The value of TE depends on the type of the LF-NMR device, and it usually ranges between 0.1 and 1.2 ms. When dealing with highly viscous fluids, the lower values are preferable. Finally, the temperature (T) of the oil sample is the third recorded parameter, expressed in kelvin.

2.2. Gradient Boosted Regression Trees. In supervised learning, the gradient boosting represents an ensemble (additive) model that can be used for solving supervised regression and classification problems. The main idea behind it is to derive a model from a set of weak learners, typically decision trees (DTs) or their simplified versions known as decision tree stumps. The construction of the viscosity model $\hat{\eta} = F(x)$, evolves in sequences or boosting iterations (m). For each iteration, a new decision tree (h) is added to the existing model, to minimize the loss function further. This way, an updated and improved version of the model is obtained $F_{m+1}(x)$. This process is repeated until the specified number of boosting iterations is reached.[39−41]

As the goal is to estimate the vector of viscosities η from the training set (x), which consists of input features from Tables 1 and 2, the model can be expressed in the forward stage-wise form as

$$F_m(x_i) = F_{m-1}(x_i) + h_m(x_i) = \eta_i \tag{2}$$

where $h_m(x_i)$ is the underlying model at mth iteration for ith observation. This equation can be rewritten as

$$h_m(x_i) = \eta_i - F_m(x_i) \tag{3}$$

From eq 3, it can be observed that each added h is fitted to prediction residuals. In gradient boosting regression, the residuals are integrated into the concept of negative gradients, enabling the use of other loss functions such as absolute loss and Huber loss.[39] When dealing with datasets with a large number of outliers, the commonly used squared error loss function $L = \sum(y_i - F(x_i))^2$ will emphasize the larger residuals. The absolute loss function is not squaring the errors $L = \sum|(y_i - F(x_i))|$, making it therefore more resistant to outliers. The negative gradient with an absolute loss function can be denoted as

$$-\frac{\partial L(\eta_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} = \mathrm{sign}(\eta_i - F_{m-1}(x_i)) \tag{4}$$

Since the loss function is minimized by adding a new DT and fitting it to F_{m-1}. The number of DTs can become excessively large, which can result in overfitting the training data. To prevent this, a shrinkage coefficient (ν) is introduced in the calculation of $F_m(x)$, which gauges the contribution of each tree $h_m(x_i)$.

$$F_m(x_i) = F_{m-1}(x_i) + \nu h_m(x_i) \tag{5}$$

This coefficient is also known as the "learning rate," and its optimal value can be estimated using some of the parameter search techniques.[42] It should be noted that learning rate ν is in the strong inverse relationship with the number of DTs, which is the number of boosting iterations (M). Usually, lower values of ν lead to a smoother convergence if used with larger values of M.[39] A more detailed explanation of gradient boosting concepts can be found elsewhere.[32,40,41,43]

2.3. Support Vector Machines for Regression (SVR). The SVR is a sophisticated and straightforward supervised learning (SL) algorithm used in classification and regression tasks. The SVR is based on the structural risk minimization (SRM) principle, which was confirmed to have better performance compared to empirical risk minimization (ERM) used, for instance, in neural networks. In simpler terms, SRM prevents the overfitting of the model by balancing two inversely related hyperparameters and consequently making a gap between the training set errors and test set errors smaller, while reducing model complexity. In contrast, in ERM, a single objective is the minimization of the training error. What made support vector machines so famous was the introduction of kernels—the arbitrary functions whose purpose is to map the dot product of input features into the higher-dimension feature space. This functionality enables the utilization of hyperplanes, which are particularly useful in nonlinear classification problems. Fortunately, the same concept was generalized for regression tasks.[44] In addition, SVR has been proven to be an effective method even in application to small datasets, which is a necessary implication for the task at hand.

In terms of viscosity prediction by NMR parameters, SVR has to be associated with our input features ($T_{2\mathrm{gm}}$, TE, and T) and output vector η (Tables 1 and 2). Suppose we arrange all of the preprocessed input features in a matrix form as $x = [x_1, x_2, x_3, ..., x_n]$, where x_n are column vectors of inputs. The measured

Table 1. Descriptive Statistics of Input Variables (Features) and Observations of Oil Viscosity η

input features	units	range/values	mean	standard deviation
dynamic viscosity (η)	mPa·s	0.87−867 634	12 978	61 373
T_2 geometric mean ($T_{2\mathrm{gm}}$)	ms	0.23−1239.9	59.4	165.6
echo-spacing (TE)	ms	[0.1, 0.24, 0.3]		
temperature (T)	K	299.15−468.15	337.4	45.1

Table 2. Descriptive Statistics of Engineered Features Used for Training of SL Models

engineered features	range	mean	standard deviation
$\log(\eta)^a$	−0.1372 to 13.6735	6.0015	2.9336
$\log(T_{2\mathrm{gm}})$	−1.4696 to 7.1227	2.0813	1.9129
$\log(T)$	5.7009−6.1487	5.8132	0.1248
$\log(T)$/TE	19.04−57.14	34.55	16.25
$\log(T_{2\mathrm{gm}})$/TE	−14.69 to 71.22	14.74	19.23

[a]Target output.

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

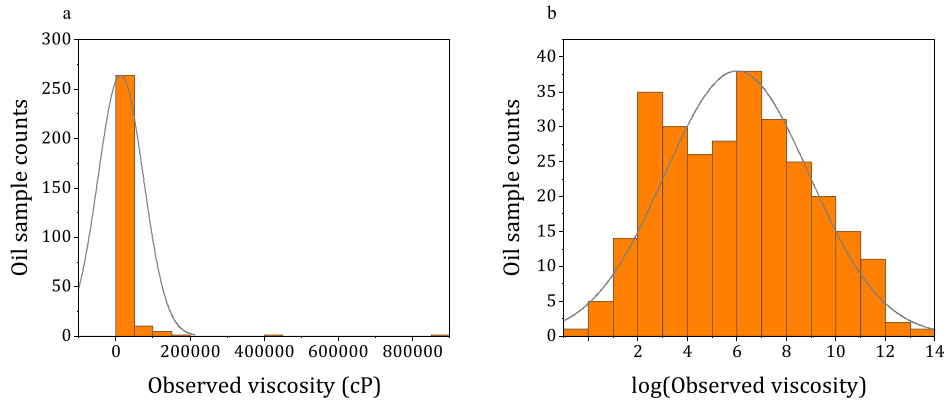

Figure 2. Distribution of oil viscosity η before (a) and after the log transformation (b).

viscosity instances can be rewritten into a response vector $\eta = [\eta_1, \eta_2, \eta_3, ..., \eta_n]$. Thus, the dataset can be defined then as $\{(x_i, \eta_i)\}_{i=1}^{n}$, where n is a number of oil samples. The support vector machine regression between the input and response vector can be written as

$$\eta: f(x) = W \cdot \phi(x) + b \tag{6}$$

Here, $\phi(x)$ is the interpretation of an input matrix x in the higher-dimension space, while W and b are weight vector and bias term, respectively. The latter two are obtained by minimizing the risk function

$$\min: \frac{\|W^2\|}{2} + C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(\eta_i, f(x_i)) \tag{7}$$

$$L_\varepsilon(\eta_i, f(x_i)) \begin{cases} 0, & \text{if } |\eta_i - f(x_i)| \le \varepsilon \\ |\eta_i - f(x_i)| - \varepsilon, & \text{otherwise} \end{cases} \tag{8}$$

where the $\|W\|$ term is a magnitude of a vector of feature weights, which reduces the function's sensitivity to the perturbations in input x (i.e., flatness), thus gauging the robustness of a model. The right-hand side term quantifies the prediction error, measured by the L_ε loss function (eq 8). The magnitude of residuals $|\eta_i - f(x_i)|$ is compared with the predefined value of ε so that the residuals smaller than ε are ignored, but residuals larger than ε are penalized. Since any ε can be defined, the C parameter is introduced to regulate the tradeoff between the flatness of the $f(x_i)$ and penalty size for residuals larger than ε.[44] The optimization of eqs 7 and 8 can be simplified by introducing slack variables (ξ_i, ξ_i^*) instead of prediction residuals[33]

$$\min: \frac{\|W^2\|}{2} + C\frac{1}{n}\sum_{i=1}^{n} (\xi + \xi^*) \tag{9}$$

$$\text{subject to: } \begin{cases} \eta_i - W \cdot \phi(x_i) - b \le \varepsilon + \xi_i \\ W \cdot \phi(x_i) + b - Y_i \le \varepsilon + \xi_i^*, \quad i = 1, \ ..., \ n \\ \xi_i \ge 0, \qquad \xi_i^* \ge 0 \end{cases} \tag{10}$$

To find the local minimum with respect to the given constraints, one can introduce Lagrange multipliers, in which case eq 6 is transformed into

$$\eta: f(x) = \sum_{i=i}^{n} (\alpha - \alpha_i^*) \cdot K(x_i, x_j) + b \tag{11}$$

where α and α_i^* are Lagrange multipliers and $K(x_i, x_j)$ is the kernel function, which maps the input features into the higher-dimensional space. Further details about support vector machine regression can be found elsewhere.[33,44]

2.4. Database of Rheological and NMR Measurements. The oil data were collected from our previous research and other published works.[25,35,36] In all studies, the procedure for experiments was similar: the dynamic viscosity of oils was determined using conventional laboratory instruments (i.e., cone and plate rheometers), whereas the T_2 relaxation spectra of the samples were obtained after raw data mathematical inversion from the measurements made by LF-NMR relaxometers. For this study, a total of 282 data points was used for model development. The dataset is available as the Supporting Information in Appendix A. The statistical description of these parameters is shown in Section 2.6 in Table 1.

2.5. Preprocessing and Analysis of the Dataset. The preprocessing and analysis of all rheological and NMR data were performed using Python environment version 3.7.2 with the scikit-learn package and OriginPro 2019b.[42] The feature dataset consists of T_{2gm}, TE, and T, while viscosity observations were stored as the output vector. The data was divided into the training set and a test set in the 3:1 proportion, respectively. In this way, we obtained a training set of 211 data points and a test set (unseen data) of 71 data points, which were used for the estimation of model accuracy only.

2.6. Feature Engineering and Transformation. Feature engineering (FE) is a process in which domain knowledge is applied to perform appropriate transformations of the inputs and to extract new information from their known empirical relationship. This strategy proved to be effective in reducing the complexity of SL models, which in turn leads to an increase in prediction performance.[34] In our case, this entailed: (1) the transformation of inputs T_{2gm}, TE, T, and target output η, and (2) deriving new inputs from the empirical relationship of T_{2gm}, TE, and T with target output η.

Table 1 shows that the ranges of inputs and outputs are out of scale, which implies that a certain transformation should be applied to normalize the data. Also, the observed viscosity data has a long-tailed distribution as it is skewed to the right-hand side of Figure 1a, with over 95% of samples distributed between the 0.8 and 100 000 mPa·s range. In the field of statistics, the observations outside three standard deviations (outliers) typically degrade the forecasting performance of the models and can be, therefore, omitted.[45] In our case, however, the outliers correspond to extra-heavy oils and bitumens (e.g., >180 000 mPa·s). In practice, the natural reservoirs in which

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

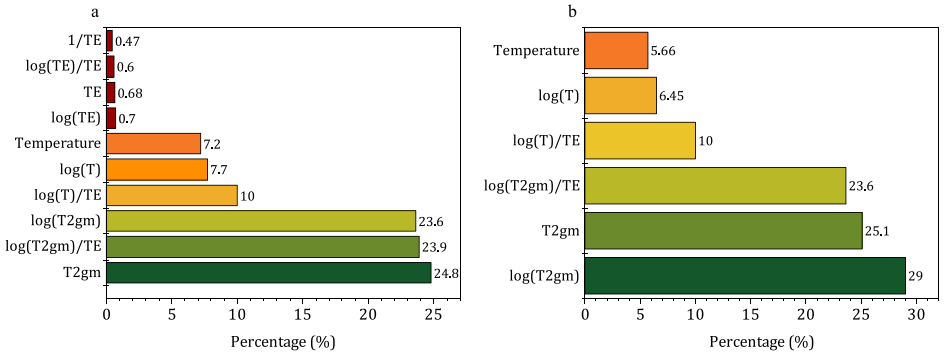

Figure 3. Relative feature importance (ranking) by the GBRT model of all input features, (a) before and (b) after the removal of redundant TE-derived features with <1% relative contribution.

these oils reside are often thermally treated to facilitate their recovery.[46] Therefore, if these samples were omitted from the training data, the valuable information about their T_2 relaxation behavior at high temperatures would be lost. This information was preserved by applying a simple logarithmic transformation to all features, which normalized the distribution of the data. The effect of log transformation is illustrated in Figure 1b, on the example of target output η. Also, the log transformation reduced nonlinearity of the dataset, which, in theory, should improve the performance of the SL regression models, which are efficient in solving linear problems (i.e., multiple linear regression and support vector regression) (Figure 2).

In the second stage, we derived new features by employing the findings from previous studies.[5,6,47] To evaluate the importance of newly derived features, we employed the GBRT algorithm. One of the benefits of ensemble models such as GBRT is their capability of feature ranking by their relative contribution to the prediction accuracy, thus making the interpretation and selection of new features more convenient. It should be noted that there are some downsides to feature ranking. For instance, two or more features may have a comparable correlation with the output. During feature ranking, one feature will be assigned a higher rank, which will lead others to get lower rank, thus potentially leaving out a strong predictor.[48] Figure 3a shows the ranking of the seven new features, alongside with T_{2gm}, TE, and T, with the bottom ones being the most relevant. The ranking is achieved by assessing the reduction of the training error generated from splitting the nodes of the DTs. Therefore, the features which reduce the training error more frequently during splitting will be ranked higher. Note in Figure 3a that T_{2gm}-related features ($\log(T_{2gm})$/TE, $\log(T_{2gm})$, and T_{2gm}) capture most of the variability (~72%), while T-derived features ($\log(T)$/TE, $\log(T)$, and T) capture about 25% of the variability. This variability distribution was expected, considering that T_{2gm} has a strong correlation with η, whereas T-related features become more important at high temperatures when severe NMR signal loss occurs (Figure 1).

In contrast, the TE-related features (1/TE, log(TE)/TE, TE, and log(TE)) affect prediction accuracy negligibly, with each being less than 1%. This is because the NMR measurements used as inputs for this study were all acquired with an attempt to optimize the signal of fast relaxing fluids, i.e., through the use of small TE values. If this dataset was to be expanded to systems with larger TE values (0.6−1.2 ms), the impact of TE would be expected to be higher. Within this dataset, the impact of TE was removed from further consideration. However, even with the perceived insignificance of TE, it should be noted that features that include the TE in the denominator demonstrate higher relevance (e.g., $\log(T)$/TE and $\log(T_{2gm})$/TE). Figure 3b illustrates the relative importance of the remaining six features, which were used for the training of the SL models. Finally, Table 2 summarizes the statistical description of log-transformed viscosity (i.e., the target output) and engineered inputs, which were used for SL viscosity forecasting alongside with original features (T_{2gm}, TE, and T).

2.7. Evaluation Metrics. Five statistical metrics were chosen for the evaluation of prediction performance of the models, including root mean square error (RMSE), mean absolute error (MAE), mean square logarithmic error (MSLE), mean absolute percentage error (MAPE), and adjusted coefficient of determination ($\overline{R}^2$). All metrics are negatively oriented statistical measures (i.e., smaller values are favorable), except for $\overline{R}^2$ for which is positively oriented.

The RMSE is regularly employed in scientific studies for the evaluation of model performance.[15,49] In this study, the RMSE is the square root of the average of squared differences between the predicted viscosity and observed viscosity, and is expressed in the centipoises

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\eta_i - \hat{\eta}_i)^2} \tag{12}$$

where n is the number of samples, η_i is the predicted viscosity, and $\hat{\eta}_i$ is the observed viscosity. However, this metric can be sensitive to the outliers, which can inflate the value of RMSE.[50] To address this issue, MAE is introduced for the calculation of averaged prediction errors of the models, in centipoises

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\eta_i - \hat{\eta}_i| \tag{13}$$

In contrast to RMSE, MAE does not square the differences between the predicted and observed viscosity, making MAE less sensitive on outliers.[51] In this manner, the MAE score gives less weight to the large prediction residuals and, therefore, can be used as a control measure in RMSE interpretation. The shared disadvantage of RMSE and MAE is that both metrics do not provide any information about percentual differences between predictions and observations. MSLE accounts for this by associating squared differences between the log-scaled predictions and observations

$$\mathrm{MSLE} = \frac{1}{n}\sum_{i=1}^{n}(\log(\eta_i + 1) - \log(\hat{\eta}_i + 1))^2 \tag{14}$$

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

In this manner, the MSLE avoids the heavy penalization of prediction errors in the high viscosity domain, as it is the case with RMSE and MAE. Instead, it considers the relative percentual difference between the observation and prediction rather than the size of their residual.[52] In addition to MSLE, MAPE was used to illustrate the relative percentual difference between the sum of errors. MAPE represents the mean of the sum of absolute percentage errors of viscosity predictions. This metric enabled a more intuitive interpretation of the model forecasts since the errors are expressed in percentages[53]

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\eta_i - \hat{\eta}_i}{\hat{\eta}}\right| \times 100\% \tag{15}$$

Finally, the proportion of model variance is typically expressed by the coefficient of determination (R^2), which is a standard measure of goodness-of-fit for the regression models

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\eta_i - \hat{\eta}_i)^2}{\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2} \tag{16}$$

Although this metric provides a fast and straightforward evaluation, it might get inflated due to the addition of new variables obtained from feature engineering. This inflation is a well-known problem, which can be addressed by adding a term that penalizes the score with each additional predictor[54]

$$R^2 = 1 - (1 - R^2)\cdot\frac{n-1}{n-p-1} \tag{17}$$

where p is the number of features. Note that $\bar{R}^2 < R^2$.

2.8. Model Selection. *2.8.1. GBRT Optimization.* As mentioned in Section 2.2, a choice of an arbitrary differentiable loss function ("loss" parameter) can be made according to the statistical properties of the dataset. The NMR viscosity dataset contains a substantial amount of outliers, which implied the use of an outlier-resistant loss function, such as the least absolute deviation (LAD).[45] To test this premise, 5-fold cross-validation was executed for four commonly used loss functions: Huber loss, least squares, least absolute deviation, and quantile loss. The rest of the parameters and hyperparameters were fixed to default values. Based on the lowest mean validation error, it was found that the GBRT configuration with LAD loss function generated the most stable predictions in terms of all error metrics (Table 3).

The "criterion" parameter can be determined in the same manner. This parameter allows a user to select the function, which will estimate the quality of the DT node split. Usually, in regression tasks with DTs, the difference between the observed and predicted value is quantified by mean squared error (MSE). Subsequently, the node splitting for a particular DT will be achieved so that the lowest MSE value is obtained. Since MSE heavily penalizes outliers, MAE was expected to perform the splitting task more efficiently (Table 4).

Table 3. Results of Four Loss Functions Used for Optimizing the Model Performance after 5-Fold Cross-Validation[a]

	loss parameter			
test scores	LAD	Huber	LS	quantile
$\mathrm{MAE_{cv}}$	6332	7319	9072	6969
$\mathrm{RMSE_{cv}}$	17 714	22 289	34 305	29 480
$\mathrm{MSLE_{cv}}$	0.198	0.251	0.26	0.464
$\bar{R}_{\mathrm{cv}}^2$	0.58	0.348	−0.54	−0.14

[a]The LAD function exhibits the best performance based on $\mathrm{MAE_{cv}}$, $\mathrm{RMSE_{cv}}$, $\mathrm{MSLE_{cv}}$, and $\bar{R}_{\mathrm{cv}}^2$ cross-validation (CV) scores.

Table 4. Results of Three Commonly Used Loss Functions for Estimating the Node Splitting Quality after 5-Fold Cross-Validation[a]

	criterion parameter		
test scores	MAE	MSE	Friedman-MSE
$\mathrm{MAE_{cv}}$	3666	5756	5757
$\mathrm{RMSE_{cv}}$	9925	16 286	16 287
$\mathrm{MSLE_{cv}}$	0.149	0.183	0.183
$\bar{R}_{\mathrm{cv}}^2$	0.86	0.66	0.64

[a]The MAE function performs optimal splitting based on $\mathrm{MAE_{cv}}$, $\mathrm{RMSE_{cv}}$, $\mathrm{MSLE_{cv}}$, and $\bar{R}_{\mathrm{cv}}^2$ cross-validation (CV) scores.

The next step was to find the optimal hyperparameter values for the GBRT model. According to refs 39 and 43, five hyperparameters have a considerable impact on GBRT model performance

- Number of trees (M): maximum number of estimators or boosting iterations ($n_$estimators).
- Learning rate (ν): shrinkage coefficient, which regulates the individual tree prediction contribution, where each tree is being scaled by $0 < \nu < 1$.
- Subsample (λ): the proportion of the data for fitting to the individual trees.
- Max depth (J): maximum depth (size) of a tree. This value constrains the number of nodes in the tree.
- Max features (ψ): number of features used in the search for the optimal split of a tree node.

Mathematically speaking, these hyperparameters are mutually dependent (also observable from eq 5), which is why it was required to use a more sophisticated optimization technique other than pure trial-and-error. Therefore, these were evaluated simultaneously with the help of GridSearchCV (GS-CV), an exhaustive search cross-validation algorithm available in a scikit-learn package.[42] From the computer science standpoint, this metaheuristic approach iteratively optimizes an algorithm by searching for an appropriate combination of hyperparameters in multidimensional real-valued parameter space (grid), relative to some measure of accuracy (e.g., R^2). This approach captures the interaction between the hyperparameters, therefore significantly reducing the optimization time. However, due to the presence of discrete data (i.e., TE) and other distributions in the input data, the grid-search optimization may fail to discover the best hyperparameter configuration, even with the appropriate transformations applied to the dataset. Also, with a growing number of hyperparameters, its' utilization becomes computationally intensive. Hence, optimization was assessed further using error curves. Table 5 shows the hyperparameters and their value range, which were optimized using the GS-CV approach. Note that the loss and criterion parameters were fixed according to results in Tables 3 and 4.

Figure 4a shows how the GBRT model deviance evolves with different learning rates (ν) as a function of a number of trees M. The GBRT loss is expressed in the least absolute deviations. Larger values of ν (e.g., $\nu = 1$) lead to faster convergence, that is, smaller values of M are needed for the deviance to converge. However, when a value is decreased ($\nu = 0.01$), the contribution of every additional estimator is reduced further, leading M to

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

Table 5. GBRT Hyperparameter Optimization by Grid-Search Based on 5-Fold Cross-Validation

GBRT hyperparameters	value range/method	optimal values	score
n_estimators (M)	[1−500]	[220]	RMSE: 8704
learning_rate (ν)	[0.01−1.0]	[0.03]	MAE: 3377
subsample (λ)	[0.1−1.0]	[1.0]	MSLE: 0.136
max_depth (J)	[1−8]	[4]	MAPE: 29
max_features (ψ)	[auto, sqrt, log 2]	[log 2]	$\bar{R}^2$: 0.91

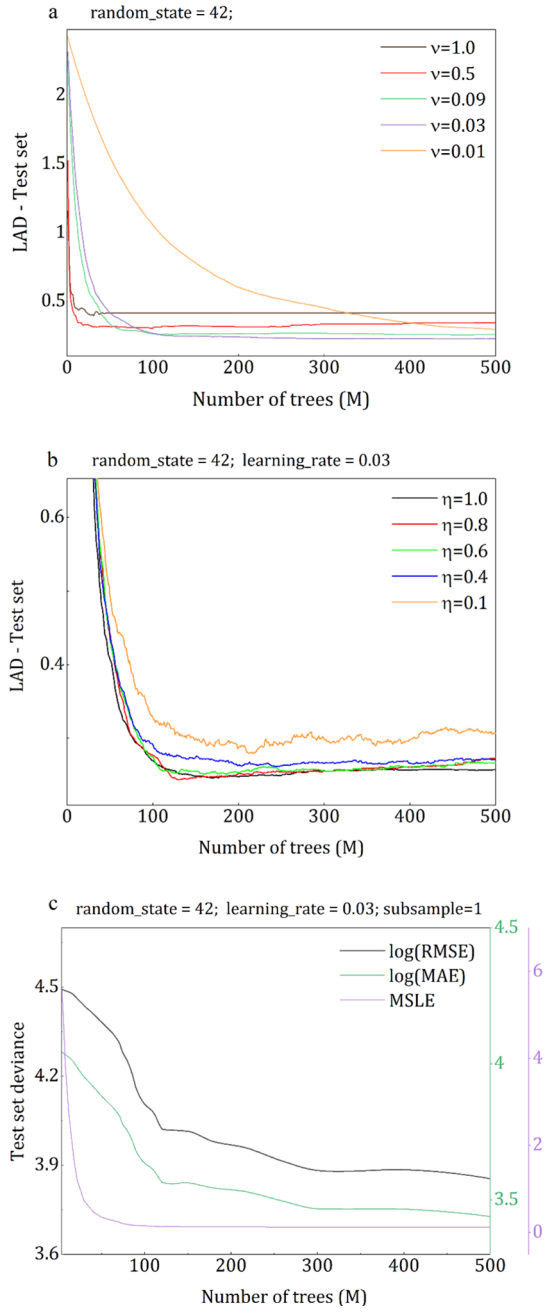

Figure 4. Test set GBRT model performance in terms of least absolute deviations (LAD) for various learning rates (a) and subsample sizes (b) relative to the number of trees M. Bottom plot (c) illustrates the model accuracy evaluation as a function of M, in terms of MAE, RMSE, and MSLE.

increase for ensuring the smooth convergence, which also means an increase in the computational cost. Since the apparent tradeoff exists between these two parameters, the parameter grid-search cross-validation was used as a strategy for obtaining their appropriate values.

Additionally, subsampling λ is a parameter that enforces the variance reduction of the sample population. In GBRT applications for large datasets, this technique proved to be very useful for improving computing performance and accuracy.[39] Figure 4b, however, shows that no subsampling ($\lambda = 1$) leads to smoothest and lowest deviance for the given input, possibly due to the small number of data points. Also, alternating the λ parameter has only a minor effect on deviance magnitude. In fact, the variations are so small that one must zoom in the y-axis to observe this behavior (note y-axis scales of Figure 4a,b).

Finally, the maximum depth of all trees was restricted to the same size ($J = 4$), as determined by the GS-CV, which is in agreement with recommendations in the literature.[39] The optimal number of features for the best split of the tree node was found to be $\psi = 2$ (i.e., max_features = "log 2"). It should be noted that the value of the latter has the least impact on the prediction performance of the GBRT model and, therefore, using the default value (i.e., "auto") is also acceptable.

2.8.2. SVR Optimization. In eq 12, the term $K(x_i, x_j)$ represents the kernel function. The standard kernel functions used in SVR are linear, polynomial, sigmoid, and Gaussian. The NMR input parameters are in the nonlinear relationship with the oil viscosity; therefore, the kernel must capture this relationship once the input features are mapped into a higher-dimension space. In these circumstances, the Gaussian or radial basis function (RBF) kernel has proven to be effective.[55] From eq 10, the kernel function can be therefore expressed as

$$K(x_i, x_j) = \exp(\gamma \cdot x_i - x_j^2) \tag{18}$$

where γ is the width hyperparameter of the RBF kernel. Hence, there are three main hyperparameters which needed to be optimized:

- γ: RBF kernel-specific parameter, which defines support vector's radius of impact.
- ε: the insensitivity radius-ε within which the prediction residuals are ignored (loss = 0). This value controls the number of support vectors (SVs) and the smoothness of the function.
- Regularization parameter (C): hyperparameter, which affects the size of the penalty applied to model predictions. If too large, the model may store an excessively large number of SVs and cause overfitting.

Literature findings show that the behavior of these hyperparameters is interrelated, which should be considered during their optimization.[44] Thus, the GS-CV approach was used for simultaneous approximation of C, ε, and RBF kernel parameter γ (Table 6). Also, their in-depth assessment was performed from the analysis of the error curves (Figure 5).

Figure 5a presents how SVR model prediction accuracy behaves for various values of radius-ε in the function of C. For the $\varepsilon = 10^{-4}$ obtained by GS-CV, it was found that the SVR model utilized over 70% of the data samples as support vectors, which indicates overfitting.[56] For values of $\varepsilon = 10^{-3}$ and 10^{-2}, the deviance converged smoothly at $C = 30$. The number of SVs was reduced by increasing ε to 10^{-3} (Figure 5a, black) while preserving nearly identical accuracy.

Figure 5b shows deviance for the fixed ε and various radii of individual SV impact γ. The smoothest convergence and lowest deviation are achieved when γ = scale, which is the value when an

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

Table 6. SVR Hyperparameter Optimization by Grid-Search Based on 5-Fold Cross-Validation

SVR hyperparameters	range/method	optimal values	score
γ	["scale", 0.1×10^{0}–1×10^{-5}]	[5×10^{-4}]	RMSE: 8704 MAE: 3377
ϵ	[1×10^{0}–1×10^{-5}]	[1×10^{-4}]	MSLE: 0.136 MAPE: 29
regularization (C)	[1–500]	[25]	$\bar{R}^2$: 0.91

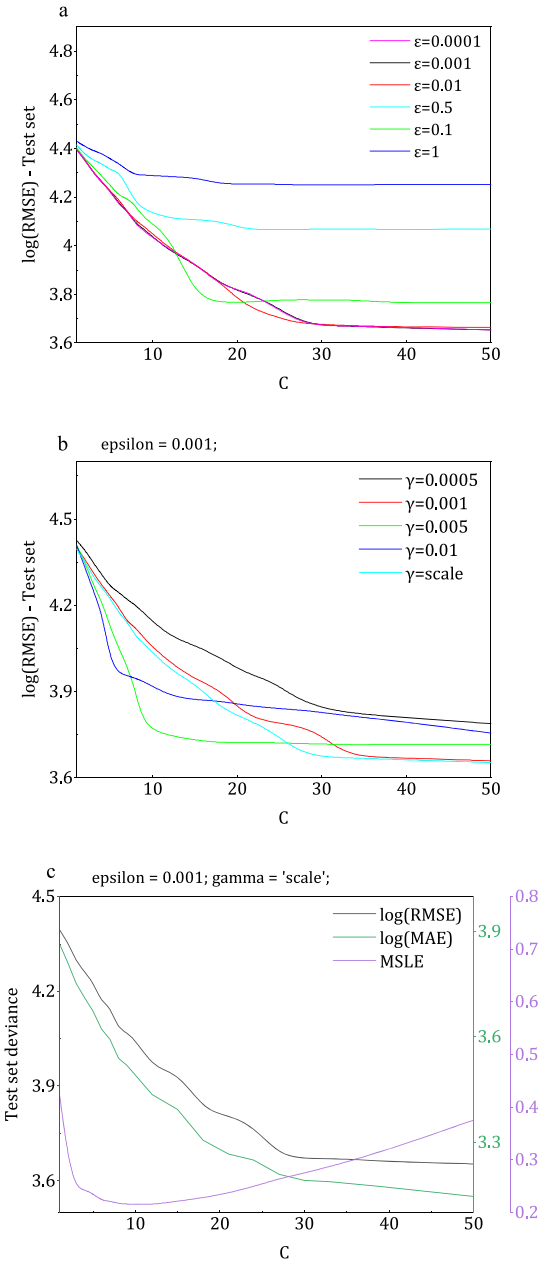

Figure 5. Test set SVR model performance in terms of log-normalized RMSE for various values of ϵ (a) and γ (b) with respect to regularization C. The bottom plot (c) illustrates the accuracy of the optimized SVR model as a function of C, in terms of three error metrics: log(RMSE), log(MAE), and MSLE.

inverse of the number of features is scaled by their standard deviation. Interestingly, the GS-CV obtained $\gamma = 5 \times 10^{-4}$, but according to its' plot (black curve), the deviance converges when $C > 50$, at which point the SVR model attempts to perfectly predict each entry from the training set (hard-margin SVM behavior). Since this might lead to overfitting and increased model complexity, γ was set to scale.

As a final step, the model with fixed ϵ and γ hyperparameters was evaluated from Figure 5c, where three metrics were utilized for the evaluation of the tuned SVR model. While both RMSE and MAE follow the same decreasing trend, the MSLE error decreases until the $C = 12$ inflection point, after which it starts increasing. This behavior is due to the inflation of residuals in the low viscosity domain. To restrict the further growth of residuals and to preserve the overall model performance, the regularization was set to $C = 25$, in line with the grid-search results (see Table 5).

3. RESULTS AND DISCUSSION

This section is divided into two parts. In the first part, we compare SVR and GBRT model performance against four other popular regression models, whereas in the second part, a performance of four well-known empirical NMR viscosity models was considered. The models are compared using the five error metrics introduced in Section 2.7. Also, the cross-plots with predicted and observed viscosities are provided for the in-depth analysis.

3.1. Supervised Learning Models. The performance of the GBRT and SVR was evaluated against an additional four SL algorithms, including multiple linear regression (SLR),[57] K-nearest neighbors (K-NN),[58] decision trees (DTs),[39] and random forests (RF).[59] Their optimization was performed with GS-CV, similarly as for GBRT and SVR (Table 7).

When Figures 6 and 7 are examined together, one can note that the overall performance of each model improves after the

Table 7. GS-CV Hyperparameter Optimization Results for All Supervised Learning Algorithms That Were Tested in This Work

model	hyperparameters	range/method	optimal values
decision trees (DTs)[39]	criterion	["mse", "mae"]	["mse"]
	max_features	[1, 2, 3, "sqrt", "log2", "auto"]	[1]
K-nearest neighbors (K-NN)[58]	n_neighbors	[1–50]	[3]
	weights	["uniform", "distance"]	["uniform"]
	algorithm	["ball_tree", "kd_tree", "brute"]	["ball_tree"]
	p	[1, 2]	[1]
random forests (RF)[59]	n_estimators	[1–80]	[7]
	criterion	["mse", "mae"]	["mae"]
support vector machines for regression (SVR)[44]	γ	["scale", 0.0005–0.1]	["scale"]
	ϵ	[1–0.0001]	[0.001]
	C	[1–500]	[25]
gradient boosted regression trees (GBRT)[39]	loss	["ls", "lad", "huber", "quantile"]	["lad"]
	n_estimators	[1–500]	[220]
	criterion	["mse", "mae", "friedman_mse"]	["mae"]
	learning_rate	[0.01–0.1]	[0.03]
	max_features	[auto, sqrt, log 2]	[log 2]
	max_depth	[1–8]	[4]
	subsample	[0.1–1.0]	[1.0]

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX–XXX

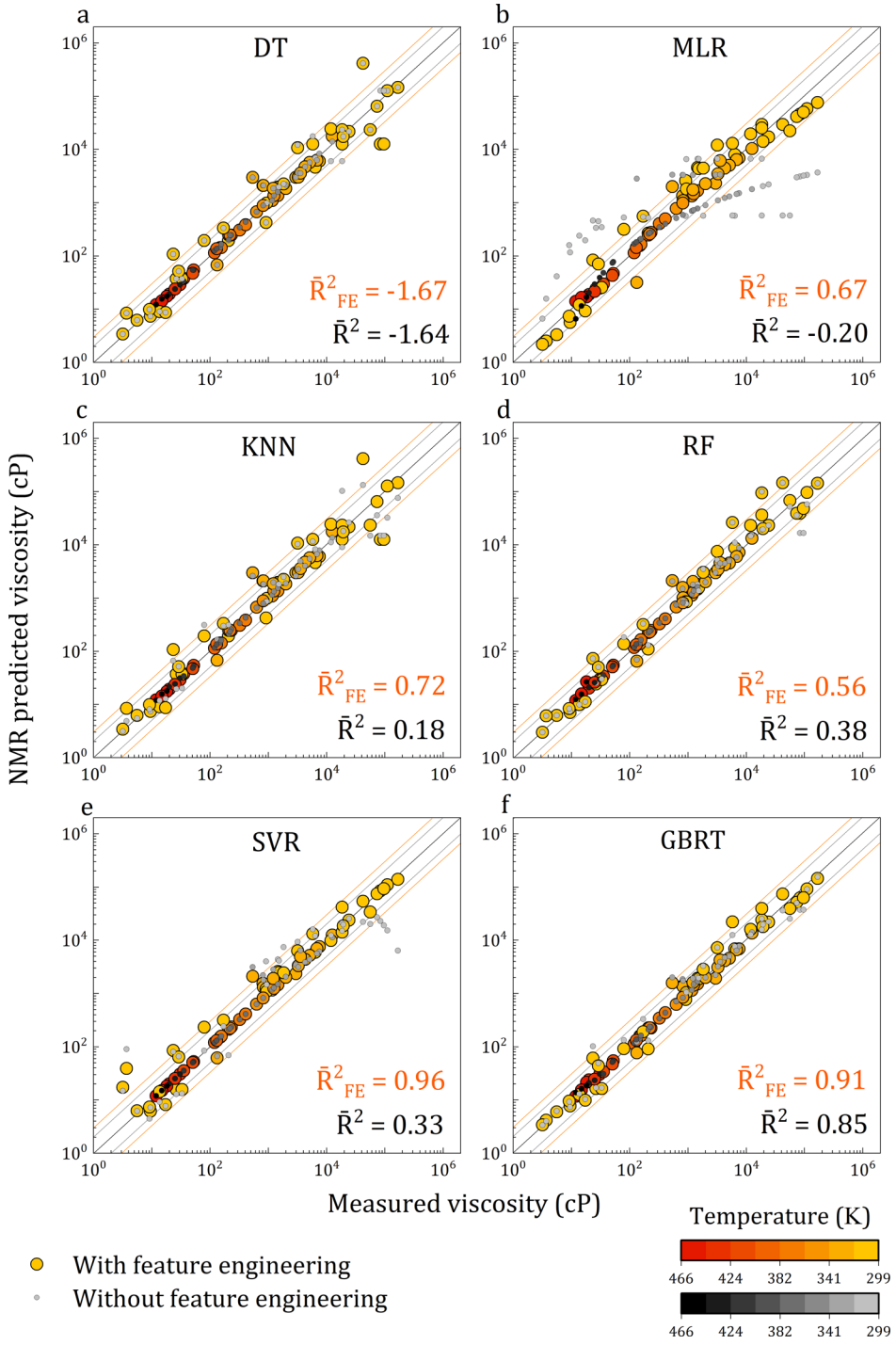

Figure 6. Comparison of NMR SL viscosity model predictions and observations. Note that the grayscale points are predictions of models generated without FE, while warm color points are predictions with FE. Lighter colors indicate lower temperatures (from 25 °C/299 K) and more intense, darker colors indicate higher temperatures (up to 200 °C/466 K). GBRT and SVR models with integrated FE demonstrate the best performance.

integration of FE. This effect is, however, not proportionally pronounced for all models. For instance, the SLR-FE model's prediction variance has reduced dramatically after employing FE (Figure 6b). Interestingly, for GBRT and RF models, the variance-reducing effect from FE is much smaller compared to the case of the latter (Figure 6e,c). The same observation applies to the DT model, the base estimator of GBRT, and RF models (Figure 6a).

This difference in performance comes from the difference in the underlying mathematical principals of these models. As the MLR model is linear, the nonlinearity reduction that came from the log transformation naturally improved the model's generalization and stability. Additionally, the integration of new features reduced the variance of the predictions, which resulted in a further shrinking of residuals. The similar is valid for the SVR model, though to a smaller extent (Figure 6f).[60] However, in the case of RF, GBRT, and DT models, the log transformation did not impact the performance because the background tree-branching process does not rely on numerical values of the features, but instead uses the rank of the features, which

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

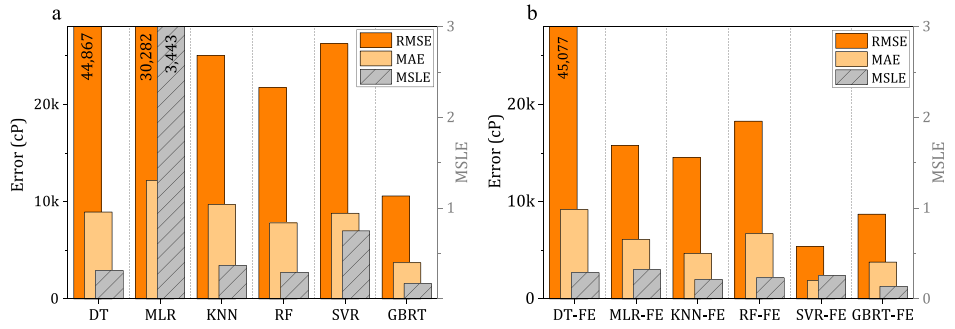

Figure 7. Compared statistical scores of SL models without FE (a) and SL models with integrated FE (b). SVR-FE and GBRT-FE demonstrate the best statistical performance.

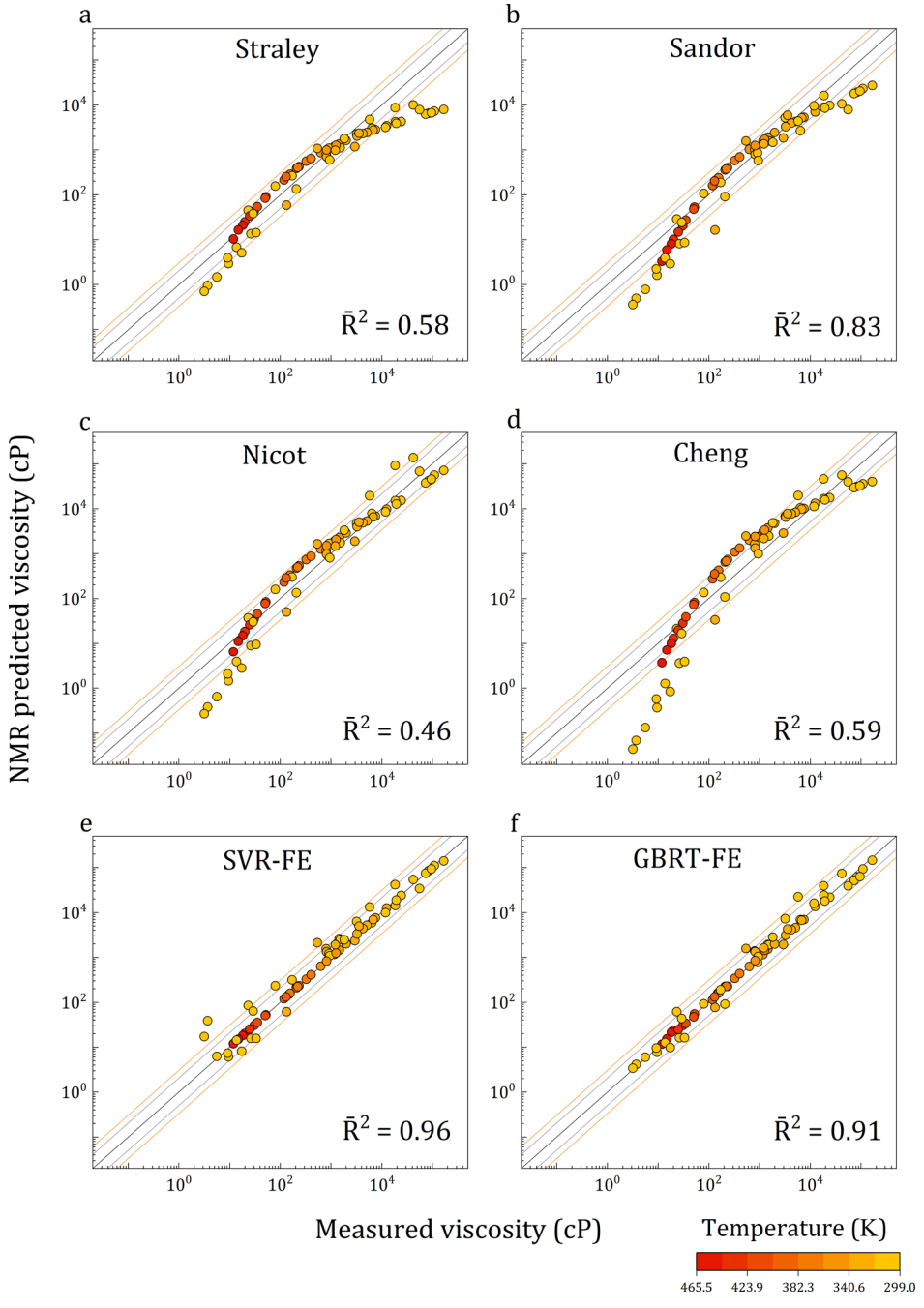

Figure 8. Performance comparison of empirical NMR viscosity models (a–d) with SVR-FE (e) and GBRT-FE (f) models. GBRT-FE and SVR-FE demonstrate significantly better performance.

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

remained the same after transformations.[39] Thus, the variance reduction came solely from capturing more variability from the new features derived in the second step of FE. Furthermore, when we compare the DT scores in Figure 7, with those by RF and GBRT, we observe a massive gap in performance, which perfectly illustrates the advantage of ensembles of DTs over the single DT. One of the reasons for the poor performance of single DT models is their "habit" to overfit the training data, making them unstable with unseen data. Therefore, ensembles of DTs generate a variance that minimizes the overfitting.[39]

Finally, the *K*-NN is a simple algorithm, where the output values are forecasted based on the similarity between the input features. This similarity is calculated as a distance (e.g., Euclidian, Manhattan, etc.) from *k*-instances, which are defined by the user.[58] Feature engineering improves *K*-NN scores almost proportionally to RF and GBRT models; however, it is not enough to minimize the large residuals in high viscosity domain, which causes the RMSE and MAE scores to remain inflated (Figure 7a,b).

Another remark is that the significant variations in temperature seem to have a negligible effect on the performance of all supervised learning algorithms. The predictions in the highest temperature domain overlap with the $x = y$ line in all six cases, demonstrating that each algorithm has appropriately captured the relationship between the observed oil viscosity and NMR signal loss that occurs at high temperatures. In comparison, empirical models that were tested in this work[5,6,37,47] exhibit poor performance in these conditions,[25] as seen in the following section.

3.2. Empirical NMR Models. The performances of GBRT-FE and SVR-FE models are compared with four well-known empirical NMR viscosity models based on T_{2gm}, TE, and T. These models were developed by Straley et al.,[37] Nicot et al.,[47] Cheng et al.,[6] and Sandor et al.[5] Previous research showed that tuning by nonlinear least squares (NLS) improves the performance of empirical models.[25] However, the dataset in the present study has long-tailed distribution with many outliers at higher viscosities, which dominate the sum of squares minimization, thus ultimately leading to erroneous model fit and misleading statistical scores.[61,62] Hence, the model fitting was performed using the orthogonal distance regression (ODR), which has been proved as a successful technique for dealing with outliers.[63] Figures 8 and 9 demonstrate the superior performance of both SVR-FE and GBRT-FE models in terms of all statistical metrics. The curvature of the viscosity forecasts by empirical models in Figure 8 reflects the combined influence of NMR signal loss due to the Curie effect, which occurs at high temperatures, and fast relaxation by solid-like components in heavy oil, which led to poor model generalization.

3.3. SVR-FE vs GBRT-FE. Overall, when cross-plots from Figures 6 and 8 are analyzed along with scores in Figures 7 and 9, one can conclude that SVR-FE and GBRT-FE models have superior test performance compared to any other SL model. For instance, in the case of SLR-FE, *K*-NN-FE, and RF-FE, the SVR-FE model, on average, scores 2.5 times lower RMSE and MAE, while GBRT-FE achieves nearly two times lower scores. When their performance is compared to empirical models, the difference in performance is even more substantial; the SVR-FE model has about 4.5 times lower RMSE and MAE scores, whereas GBRT-FE achieves nearly 3.5 times lower scores.

The principal difference in the performance of these two models is related to their precision (i.e., variance), which is evidenced by their different MSLE and MAPE scores. For instance, the SVR-FE model has a better MSLE score relative to empirical models (~4.5 times lower) but compared to SL models, SVR-FE is marginally outperformed by *K*-NN-FE and RF-FE. The same is true for MAPE scores and percentage error box plots when further examined in Figure 10. The GBRT-FE model, on the other hand, scored the best MSLE and MAPE scores in this work. These results imply that SVR-FE has the highest accuracy but somewhat lower precision (i.e., variance), relative to GBRT-FE, *K*-NN-FE, and RF-FE models. For more convenience, all evaluation scores are summarized in Table 8.

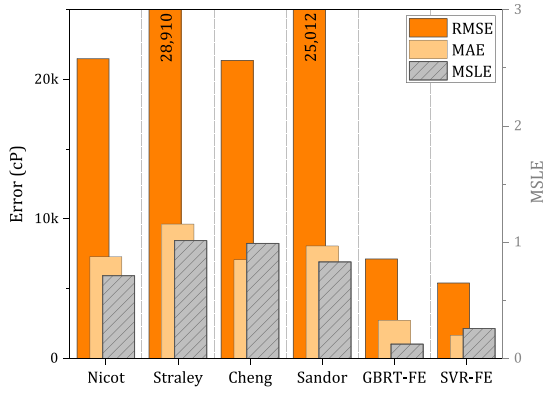

Figure 9. Compared statistical scores of four empirical NMR viscosity models and SVR-FE and GBRT-FE supervised learning models in terms of RMSE, MAE, and MSLE. SVR-FE and GBRT-FE demonstrate significantly better statistical performance.

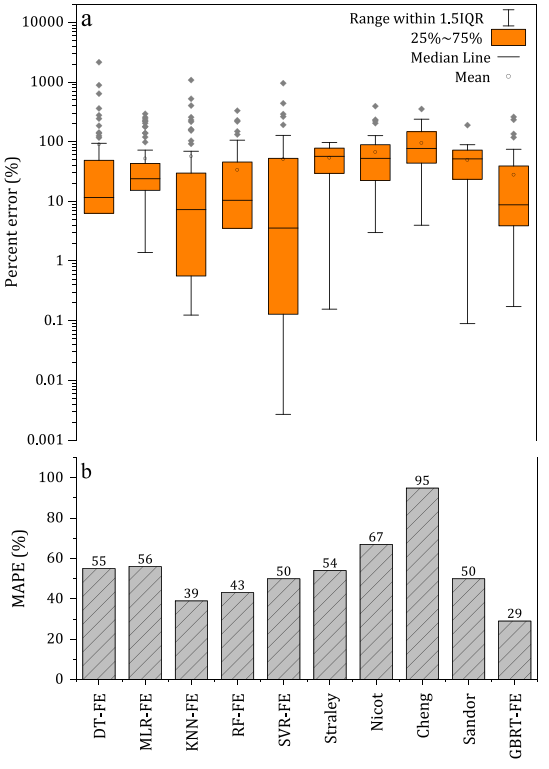

Figure 10. Percent error box plots (a) and MAPE scores (b) for six supervised learning models with feature engineering and four empirical models. Note that in the plot, (a) the *y*-axis is in the log scale. GBRT-FE model demonstrates the best performance in terms of MAPE.

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

Table 8. Compared View of Statistical Scores for All SL and Empirical Models[a]

	test scores				
model	RMSE (mPa·s)	MAE (mPa·s)	MSLE	MAPE (%)	$\overline{R}^2$
DT	44 867	7968	0.313	55	−1.64
MLR	30 282	10 858	3.443	282	−0.20
K-NN	25 044	8642	0.369	47	0.18
RF	21 725	6989	0.293	44	0.38
SVR	26 266	7858	0.749	93	0.09
GBRT	10 587	3331	0.168	32	0.85
DT-FE	45 077	8215	0.289	55	−1.67
SLR-FE	15 808	5447	0.319	56	0.67
K-NN-FE	14 559	4182	0.210	39	0.72
RF-FE	18 285	5979	0.232	43	0.56
SVR-FE	**5418**	**1671**	0.257	50	**0.96**
GBRT-FE	8704	3377	**0.136**	**29**	0.91
Straley	28 910	9638	1.014	54	0.58
Sandor	25 012	8066	0.831	50	0.83
Nicot	21 489	7306	0.712	67	0.46
Cheng	21 371	7085	0.990	95	0.59

[a]Values in bold correspond to the best score.

3.4. Physical Implications of SVR-FE and GBRT-FE Performance. From the physical point of view, meaningful insight can be obtained from the fundamental understanding of the models. Although SVR-FE is moderately stable and achieves good accuracy, it seems to struggle with capturing additional variability originating from the oils' diverse chemical composition. This behavior can be observed in Figure 6f, where SVR-FE predictions favor the high-temperature heavy oil sample over the other samples (Figure 6f). This occurs due to the SRM principle, which balances model complexity to avoid overfitting the training data and ensures the best possible generalization of new data. In other words, the model can be adapted to perform with more precision, but that would likely deteriorate its generalization ability. However, two possible strategies could rectify this; first, SVR heavily relies on feature engineering, which implies that the SVR training on a set of lighter or more chemically alike oils, would improve the forecasts' precision preserving good generalization. The second strategy would be to expand the database by adding more NMR data from new samples.

On the other hand, GBRT-FE effectively handles the discrepancies from a chemically diverse set of oil samples and a wide span of temperature and viscosity. This is due to its stage-wise estimator addition principle, where overfitting is controlled by tuning the learning rate and restriction of tree sizes. In this way, the GBRT hyperparameters limit individual trees' contribution, but by adding many estimators, the model manages to "learn" nuanced relationships that stem from mixed oil chemistry, thus outperforming the SVR approach. As a result, GBRT-FE achieves the best tradeoff between variance and bias for the task at hand, at a negligible increase in computational costs.

On another note, models presented in this study have certain limitations originating from (a) NMR hardware configuration and (b) data availability.

(a) One of the limitations of the presented SL models is that they were trained on NMR oil data acquired with echo-spacing (TE) ranging from 0.1 to 0.3 ms. Thus, the NMR data acquired using older NMR tools where echo-spacing (TE) values are hardware-limited to longer TE (0.3−1.2 ms), might have less reliable predictions. Reliability could be particularly problematic for heavy oils and bitumens, where due to the fast relaxation of solid-like constituents, the NMR device fails to measure the whole T_2 relaxation spectra, which would cause the models to underpredict the real viscosity.[5,47,64] However, by adding new NMR data to the dataset acquired using longer TE, preferably from heavy oil samples, the SL algorithm could capture the relationship between long TE and viscosity, therefore compensating for the undetected part of the T_2 spectra.

(b) Small datasets are very common in petrophysics, especially NMR data, due to the confidentiality regulations of oil companies and high well-logging costs, making the application of artificial intelligence challenging. Additional data, acquired from heavier oils and at various temperatures, would make these models more robust to both chemical diversity of oils and various temperature conditions.

4. CONCLUSIONS

In this study, we used SVR and GBRT algorithms to develop NMR models for oil viscosity prediction using NMR T_2 relaxation time, echo-spacing and temperature as an input, and dynamic oil viscosity as the target output. Also, a strategy to reduce the variance of the forecasts was introduced, where domain knowledge was used to implement feature engineering. Model performance was assessed against four other popular SL algorithms and other four analytical models from the literature. The SVR-FE and GBRT-FE have achieved statistically most favorable scores in the study in terms of five error metrics: RMSE, MAE, MSLE, MAPE, and $\overline{R}^2$.

In summary, GBRT-FE demonstrated the best overall generalization ability, thus producing predictions with a well-balanced variance-bias tradeoff. Consequently, the use of GBRT-FE might prove as a viable solution in circumstances where a wide span of oil types (light oils, heavy oils, and bitumens) is being tested at various temperatures. Environments like these correspond to heavy oil reservoirs undergoing or being screened for thermal treatment and other EOR approaches such as solvent injection or miscible gas injection.[65] On the other hand, the SVR-FE model exhibited a high accuracy but could not account for the variability originating from the diverse chemical composition of the oils at the level that the GBRT-FE model did. These findings indicate that SVR-FE would be a better choice when sets of chemically more similar oils are being studied (e.g., only light or only heavy oils) at various temperatures. In such conditions, the variance of SVR-FE predictions would reduce to the degree where high accuracy and precision come into play, such as in laboratory NMR characterization of petroleum fractions or contactless noninvasive oil viscosity monitoring in mechanical systems.

Finally, the proposed strategy for supervised learning application proved to be effective even for a small dataset, suggesting that this approach can be extended to characterize other physicochemical properties of oils, fuels, and petroleum distillates researchers work with relatively smaller datasets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.energyfuels.0c02565.

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

Appendix A (XLSX)

AUTHOR INFORMATION

Corresponding Author

Strahinja Markovic − *Center for Hydrocarbon Recovery, Skolkovo Institute of Science and Technology, Moscow 121205, Russian Federation; Curtin University, Perth, WA 6845, Australia;* orcid.org/0000-0002-6143-9370; Phone: +7 (915) 118-15-62; Email: strahinja.markovic@postgrad.curtin.edu.au

Authors

Jonathan L. Bryan − *University of Calgary, Calgary, Alberta T2N 1N4, Canada*

Vladislav Ishimtsev − *Center for Hydrocarbon Recovery, Skolkovo Institute of Science and Technology, Moscow 121205, Russian Federation*

Aman Turakhanov − *Center for Hydrocarbon Recovery, Skolkovo Institute of Science and Technology, Moscow 121205, Russian Federation*

Reza Rezaee − *Curtin University, Perth, WA 6845, Australia*

Alexey Cheremisin − *Center for Hydrocarbon Recovery, Skolkovo Institute of Science and Technology, Moscow 121205, Russian Federation*

Apostolos Kantzas − *University of Calgary, Calgary, Alberta T2N 1N4, Canada*

Dmitry Koroteev − *Center for Hydrocarbon Recovery, Skolkovo Institute of Science and Technology, Moscow 121205, Russian Federation*

Sudarshan A. Mehta − *University of Calgary, Calgary, Alberta T2N 1N4, Canada*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.energyfuels.0c02565

Notes

The authors declare no competing financial interest.

The code was developed by Vladislav Ishimtsev and Strahinja Markovic using Python 3.6, and can be run using Google Colab (embedded in README file in GitHub repository) or using Python 3.6 and higher versions (https://github.com/markovicstrahinja/ML_NMR).

ACKNOWLEDGMENTS

The authors would like to acknowledge the Integrated Center for Hydrocarbon Recovery (CHR) and Center for Computational and Data-Intensive Science and Engineering (CDISE) from Skolkovo Institute of Science and Technology, and the Fundamentals of Unconventional Resources (FUR) and In situ Combustion (ISC) groups from the University of Calgary for technical support. We also gratefully acknowledge contributions from the WA School of Mines from Curtin University.

ABBREVIATIONS

LF-NMR = low-field nuclear magnetic resonance
SL = supervised learning
PLS = partial least squares
ANN = artificial neural networks
HI = hydrogen index
RHI = relative hydrogen index
TE = echo spacing
SRM = structural risk minimization
ERM = empirical risk minimization
RMSE = root mean square error
MAE = mean absolute error
MSLE = mean square logarithmic error
MAPE = mean absolute percentage error
LAD = least absolute deviation
GS-CV = grid-search cross-validation
RBF = radial basis function
SV = support vector
SVR = support vector regression
GBRT = gradient boosted regression trees
MLR = multiple linear regression
K-NN = *K*-nearest neighbors
DT = decision trees
RF = random forests
FE = feature engineering
NLS = nonlinear least squares
ODR = orthogonal distance regression
EOR = enhanced oil recovery

REFERENCES

(1) Meyer, R.; Attanasi, E.; Freeman, P. *Heavy Oil and Natural Bitumen Resources in Geological Basins of the World*, Open-File Report 2007-1084; USGS, 2007; p 36.

(2) Luo, P.; Gu, Y. Effects of Asphaltene Content on the Heavy Oil Viscosity at Different Temperatures. *Fuel* **2007**, *86*, 1069−1078.

(3) Singer, P. M.; Parambathu, A. V.; Wang, X.; Chapman, W. G.; Hirasaki, G. J.; Fleury, M. Elucidating the 1 H NMR Relaxation Mechanism in Polydisperse Polymers and Bitumen Using Measurements, MD Simulations, and Models. *J. Phys. Chem. B* **2020**, 4222.

(4) Singer, P. M.; Chen, Z.; Alemany, L. B.; Hirasaki, G. J.; Zhu, K.; Xie, H.; Vo, T. D. Interpretation of NMR Relaxation in Bitumen and Organic Shale Using Polymer-Heptane Mixes. *Energy Fuels* **2018**, 1534.

(5) Sandor, M.; Cheng, Y.; Chen, S. Improved Correlations for Heavy-Oil Viscosity Prediction with NMR. *J. Pet. Sci. Eng.* **2016**, *147*, 416−426.

(6) Cheng, Y.; Kharrat, A. M.; Badry, R.; Kleinberg, R. L. In *Power-Law Relationship between the Viscosity of Heavy Oils and NMR Relaxation*, SPWLA 50th Annual Logging Symposium; Society of Petrophysicists and Well-Log Analysts, 2009.

(7) Kantzas, A. Advances in Magnetic Resonance Relaxometry for Heavy Oil and Bitumen Characterization. *J. Can. Pet. Technol.* **2009**, *48*, 15−23.

(8) Yang, Z.; Hirasaki, G. J.; Appel, M.; Reed, D. A. In *Viscosity Evaluation for NMR Well Logging of Live Heavy Oils*, Society of Petrophysicists and Well-Log Analysts, 2012; Vol. *53*, pp 22−37.

(9) Joss, L.; Mu, E. A. Machine Learning for Fluid Property Correlations: Classroom Examples with MATLAB. *J. Chem. Educ.* **2018**, 697.

(10) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **2019**, *5*, No. 83.

(11) Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A. K. Applications of Machine Learning to Machine Fault Diagnosis: A Review and Roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, No. 106587.

(12) Gao, Z.; Zou, X.; Huang, Z.; Zhu, L. Predicting Sooting Tendencies of Oxygenated Hydrocarbon Fuels with Machine Learning Algorithms. *Fuel* **2019**, *242*, 438−446.

(13) Myshkin, N. K.; Markova, L. V. *On-Line Condition Monitoring in Industrial Lubrication and Tribology*; Springer, 2018.

(14) Barbosa, L. L.; Montes, L. F.; Kock, F. V. C.; Morgan, V. G.; Souza, A.; Song, Y.; Castro, E. R. V. Relative Hydrogen Index as a Fast Method for the Simultaneous Determination of Physicochemical Properties of Petroleum Fractions. *Fuel* **2017**, *210*, 41−48.

(15) Constantino, A. F.; Cubides-román, D. C.; Reginaldo, B.; Queiroz, L. H. K.; Colnago, L. A.; Neto, Á. C.; Barbosa, L. L.; Romão, W.; de Castro, E. V. R.; Filgueiras, P. R.; et al. Determination of

Physicochemical Properties of Biodiesel and Blends Using Low-Field NMR and Multivariate Calibration. *Fuel* **2019**, *237*, 745−752.
(16) Alptekin, E.; Canakci, M. Determination of the Density and the Viscosities of Biodiesel−Diesel Fuel Blends. *Renewable Energy* **2008**, *33*, 2623−2630.
(17) Li, G. Z.; Meng, H. H.; Yang, M. Q.; Yang, J. Y. Combining Support Vector Regression with Feature Selection for Multivariate Calibration. *Neural Comput. Appl.* **2009**, *18*, 813−820.
(18) Li, H.; Zhang, W.; Chen, Y.; Guo, Y.; Li, G.; Zhu, X. A Novel Multi-Target Regression Framework for Time-Series Prediction of Drug Efficacy. *Sci. Rep.* **2017**, *7*, No. 40652.
(19) Straley, C. A. In *Reassessment of Correlations Between Viscosity and NMR Measurements*, SPWLA 47th Annual Logging Symposium; Society of Petrophysicists and Well-Log Analysts, 2006.
(20) Mirotchnik, K. D.; Allsopp, K.; et al. Low-Field NMR Method for Bitumen Sands Characterization: A New Approach. *SPE Reservoir Eval. Eng.* **2001**, 88−96.
(21) Yang, Z.; Hirasaki, G. J. NMR Measurement of Bitumen at Different Temperatures. *J. Magn. Reson.* **2008**, *192*, 280−293.
(22) Bryan, J.; Moon, D.; Kantzas, A. In Situ Viscosity of Oil Sands Using Low Field NMR. *J. Can. Pet. Technol.* **2005**, *44*, 23−29.
(23) Hirasaki, G. J.; Lo, S. W.; Zhang, Y. NMR Properties of Petroleum Reservoir Fluids. *Magn. Reson. Imaging* **2003**, *21*, 269−277.
(24) Bryan, J.; Kantzas, A.; Badry, R.; Emmerson, J.; Hancsicsak, T. In-Situ Viscosity of Heavy Oil: Core and Log Calibrations. *J. Can. Pet. Technol.* **2006**, *46*, 47−55.
(25) Markovic, S.; Bryan, J. L.; Turakhanov, A.; Cheremisin, A.; Mehta, S. A.; Kantzas, A. In-Situ Heavy Oil Viscosity Prediction at High Temperatures Using Low-Field NMR Relaxometry and Nonlinear Least Squares. *Fuel* **2020**, *260*, No. 116328.
(26) Li, H.; Misra, S. Prediction of Subsurface NMR T2 Distribution from Formation-Mineral Composition Using Variational Autoencoder. *SEG Tech. Program Expanded Abstr.* **2017**, 3350−3354.
(27) Li, H.; Misra, S.; He, J. Neural Network Modeling of in Situ Fluid-Filled Pore Size Distributions in Subsurface Shale Reservoirs under Data Constraints. *Neural Comput. Appl.* **2020**, *32*, 3873−3885.
(28) Truhan, J. J.; Qu, J.; Blau, P. J. The Effect of Lubricating Oil Condition on the Friction and Wear of Piston Ring and Cylinder Liner Materials in a Reciprocating Bench Test. *Wear* **2005**, *259*, 1048−1055.
(29) Khakimova, L.; Askarova, A.; Popov, E.; Moore, R. G.; Solovyev, A.; Simakov, Y.; Afanasiev, I.; Belgrave, J.; Cheremisin, A. High-Pressure Air Injection Laboratory-Scale Numerical Models of Oxidation Experiments for Kirsanovskoye Oil Field. *J. Pet. Sci. Eng.* **2020**, *188*, No. 106796.
(30) Astley, K. R. et al. Monitoring the Health of a Fluid System. US6,794,865B2, 2004.
(31) Arola, D. F.; Barrall, G. A.; Powell, R. L.; McCarthy, K. L.; McCarthy, M. J. Use of Nuclear Magnetic Resonance Imaging as a Viscometer for Process Monitoring. *Chem. Eng. Sci.* **1997**, *52*, 2049−2057.
(32) Bühlmann, P.; Hothorn, T. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Stat. Sci.* **2007**, *22*, 477−505.
(33) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.
(34) Zheng, A.; Casari, A. *Feature Engineering for Machine Learning*; O'Reilly Media, Inc., 2018.
(35) Bryan, J.; Kantzas, A.; Bellehumeur, C. Oil-Viscosity Predictions From Low-Field NMR Measurements. *SPE Reservoir Eval. Eng.* **2005**, *8*, 44−52.
(36) Yang, Z.; Hirasaki, G. J. NMR Measurement of Bitumen at Different Temperatures. *J. Magn. Reson.* **2008**, *192*, 280−293.
(37) Straley, C.; Rossini, D.; Vinegar, H.; Tutunjian, P.; Morriss, C. Core Analysis by Low Field NMR. *Log Analyst*; Society of Petrophysicists and Well-Log Analysts, 1997; Vol. *38*, pp 84−94.
(38) Kleinberg, R. I.; Vinegar, H. J. NMR Properties of Reservoir Fluids. *Log Analyst*; Society of Petrophysicists and Well-Log Analysts, 1996; pp 20−32.
(39) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer, 2009.
(40) Friedman, J. H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367−378.
(41) Schapire, R. E.; Freund, Y. Boosting: Foundations and Algorithms. *Kybernetes* **2013**, *42*, 164.
(42) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(43) Ridgeway, G. Generalized Boosted Models: A Guide to the Gbm Package, 2019, 4, pp 1−15.
(44) Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199−222.
(45) Dodge, Y. LAD Regression for Detecting Outliers in Response and Explanatory Variables. *J. Multivar. Anal.* **1997**, *61*, 144−158.
(46) Askarova, A.; Turakhanov, A.; Markovic, S.; Popov, E.; Maksakov, K.; Usachev, G.; Karpov, V.; Cheremisin, A. Thermal Enhanced Oil Recovery in Deep Heavy Oil Carbonates: Experimental and Numerical Study on a Hot Water Injection Performance. *J. Pet. Sci. Eng.* **2020**, *194*, No. 107456.
(47) Nicot, B.; Fleury, M.; Leblond, J. In *Improvement of Viscosity Prediction Using NMR Relaxation*, SPWLA 48th Annual Logging Symposium; Society of Petrophysicists and Well-Log Analysts: Austin, Texas, 2007; p 7.
(48) 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*; Misra, S.; Li, H.; He, J., Eds.; Gulf Professional Publishing, 2020; Chapter 10, pp 289−314.
(49) Valentín, M. B.; Bom, C. R.; Martins Compan, A. L.; Correia, M. D.; Menezes de Jesus, C.; de Lima Souza, A.; de Albuquerque, M. P.; de Albuquerque, M. P.; Faria, E. L. Estimation of Permeability and Effective Porosity Logs Using Deep Autoencoders in Borehole Image Logs from the Brazilian Pre-Salt Carbonate. *J. Pet. Sci. Eng.* **2018**, *170*, 315−330.
(50) Wherity, S.; Sidley, T.; Cowling, M.; Ismayilov, A.; Noe-Nygaard, J. In *Observation and Monitoring Well: In Situ Window to Assess Recovery Efficiency*, SPE Annual Technical Conference and Exhibition; Society of Petroleum Engineers, 2014.
(51) Willmott, C. J.; Matsuura, K. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Clim. Res.* **2005**, *30*, 79−82.
(52) Chollet, F. et al. *Keras*; GitHub, 2015.
(53) de Myttenaere, A.; Golden, B.; Le Grand, B.; Rossi, F. Mean Absolute Percentage Error for Regression Models. *Neurocomputing* **2016**, *192*, 38−48.
(54) Theil, H.; Cramer, J. S.; Moerman, H.; Russchen, A. *Economic Forecasts and Policy; Contributions to Economic Analysis*; North-Holland Publishing Company, 1961.
(55) Biancolini, M. E. *Radial Basis Functions for Engineering Applications*; Springer US, 2017.
(56) Cherkassky, V.; Ma, Y. Selection of Meta-Parameters for Support Vector Regression. In *Artificial Neural Networks—ICANN 2002*; Dorronsoro, J. R., Eds.; Lecture Notes in Computer Science; Springer, 2002; Vol. *2415*, pp 687−693.
(57) Cohen, P.; West, S. G.; Aiken, L. S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*; Taylor & Francis, 2014.
(58) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175−185.
(59) Ho, T. K. In *Random Decision Forests*, Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, Vol. *1*, pp 278−282.
(60) Sarle, W. S. *Neural Network FAQ, Part 1 of 7: Introduction*, Periodic Posting to the Usenet Newsgroup comp.ai.neural-nets, 1997.
(61) Riazoshams, H.; Midi, H.; Ghilagaber, G. Outlier Detection in Nonlinear Regression. *Robust Nonlinear Regression*; John Wiley & Sons, Ltd., 2018; pp 107−141.
(62) Sohn, B. Y.; Kim, G. B. Detection of Outliers in Weighted Least Squares Regression. *Korean J. Comput. Appl. Math.* **1997**, *4*, 441−452.

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX

(63) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. Retrieved from http://www.scipy.org/.
(64) LaTorraca, G. A.; Stonard, S. W.; Webber, P. R.; Carison, R. M.; Dunn, K. J. In *Heavy Oil Viscosity Determination Using NMR Logs*, SPWLA 40th Annual Logging Symposium, 1999; p 11.
(65) Fayazi, A.; Kryuchkov, S.; Kantzas, A. Evaluating Diffusivity of Toluene in Heavy Oil Using Nuclear Magnetic Resonance Imaging. *Energy Fuels* **2017**, *31*, 1226−1234.

https://dx.doi.org/10.1021/acs.energyfuels.0c02565
Energy Fuels XXXX, XXX, XXX−XXX