Skolkovo Institute of Science and Technology

Curtin University

Application of LF-NMR measurements and supervised learning regression methods for improved characterization of heavy oils and bitumens

Doctoral Thesis

by

STRAHINJA MARKOVIC

JOINT DOCTORAL PROGRAM IN PETROLEUM ENGINEERING WITH SKOLKOVO INSTITUTE OF SCIENCE AND TECHNOLOGY AND CURTIN UNIVERSITY

Supervisors:

Professor Alexey Cheremisin

Professor Reza Rezaee

Co-supervisor: Professor Dmitry Koroteev

Moscow - 2022

I declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, and Curtin University, Perth. To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Candidate:
Strahinja Markovic

Supervisors:
Prof. Alexey Cheremisin, Skoltech University
Prof. Reza Rezaee, Curtin University
Co-supervisor: Professor Dmitry Koroteev

Abstract

Heavy oil and bitumen resources are the most abundant hydrocarbon energy source worldwide. However, thermal enhanced oil recovery (EOR) methods are frequently applied to enhance their mobility and production due to their high viscosity. In addition, owing 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 variability (e.g., mechanical systems). Information about the distribution of oil viscosity within the reservoir can be used to help manage the thermal EOR projects. Nuclear magnetic resonance (NMR) downhole tools provide a non-destructive way to determine the oil viscosity without recovering samples (core or produced oil) from the well.

A new analytical NMR viscosity model was developed and tested on a suite of 23 Canadian heavy oils recovered from various reservoirs. The model was based on two NMR parameters – T_2 logarithmic mean and relative hydrogen index. Subsequently, the model was tested on a single bitumen sample at the temperature range from 26-200 °C. Results were compared to nine well-known NMR viscosity models described in the literature, and in both cases, the proposed model got the most favorable statistical scores. Furthermore, a simple model optimization procedure was presented, employing nonlinear least squares (NLS) regression. Experiments were carried out at temperatures corresponding to those in the hot steam injection EOR treatments. The same methodology can be extended for use in cyclic solvent injection (CSI), where the NMR model can detect oil viscosity changes when the solvent vapor dissolves in oil.

To improve the viscosity forecast, we developed a framework that combines supervised learning algorithms with domain knowledge for synthesizing new features using only one NMR parameter – the T_2 logarithmic 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 the USA. 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. The proposed framework can also be applied to determine other physicochemical properties of oils by LF-NMR, where supervised learning is usually impractical due to the limited volume of data.

Finally, water saturation determination in heavy oil sands is one of the most important tasks in petrophysical well-logging, and it directly impacts the decision-making process in hydrocarbon exploration and production. However, quantifying oil and water volumes is problematic when their NMR signals are not distinct. We developed two machine learning frameworks to predict relative water content in oil-sand samples using LF-NMR spin-spin (T_2) relaxation and bulk density data to derive a model based on Extreme Gradient Boosting. The NMR T_2 distributions were obtained for 82 Canadian oil-sands samples at ambient and reservoir temperatures (164 data points). The actual water content was determined by Dean-Stark extraction. The results of the statistical analysis confirm the strong generalization ability of the feature engineering LF-NMR model and indicate that this approach can be extended for the improved *in-situ* water saturation evaluation by LF-NMR and bulk density measurements.

Publications

1. S. Markovic, J. L. Bryan, A. Turakhanov, A. Cheremisin, S. A. Raj Mehta, A. Kantzas. In-situ heavy oil viscosity prediction at high temperatures using low-field NMR relaxometry and nonlinear least squares. *Fuel, Volume 260, 2020.*
DOI: 10.1016/j.fuel.2019.116328

2. A. Askarova, A. Turakhanov, S. Markovic et al., Thermal enhanced oil recovery in deep heavy oil carbonates: an experimental and numerical study on a hot water injection performance, *Journal of Petroleum Science and Engineering, 2020.*
DOI: 10.1016/j.petrol.2020.107456

3. S. Markovic, J. L. Bryan, V. Ishimitsev, A. Turakhanov, R. Rezaee, A. Cheremisin, A. Kantzas, D. Koroteev, S. A. Mehta. Improved oil viscosity characterization by low-field NMR using feature engineering and supervised learning algorithms. *Energy & Fuels, 2020 34 (11), 13799-13813*
DOI: 10.1021/acs.energyfuels.0c02565

4. S. Markovic, J. L. Bryan, R. Rezaee, A. Turakhanov, A. Cheremisin, A. Kantzas, D. Koroteev. Application of XGBoost model for *in-situ* water saturation determination in Canadian oil-sands by LF-NMR and density data. *Scientific Reports, Nature, 2022.*

Conferences and conference proceedings

1. S. Markovic, A. Turakhanov, A. Cheremisin, J. L. Bryan, S. A. Raj Mehta, A. Kantzas. Evaluation and prediction of heavy oil viscosity at high temperatures - application of low-field NMR measurements. *World Heavy Oil Congress & Exhibition, Muscat, Oman, 2018.*

2. S. Markovic, J. L. Bryan, R. Rezaee, A. Cheremisin, A. Kantzas. In-situ water saturation by LF-NMR and supervised learning - application to Canadian oil sands. *EAGE Geotech, London, United Kingdom, 2022.*
DOI: 10.3997/2214-4609.20224021

Patents

1. S. Markovic, A. Cheremisin, S.A. Mehta et al., RU Patent 2021111887, Method for thermal enhanced oil recovery tests on whole core samples, 2021.

2. Method for *in-situ* water saturation oil-sands by LF-NMR and density data and XGBoost algorithm; (**under preparation).**

Non-thesis publications, relevant for the field

1. A. Ivanova, A. Orekhov, S. Markovic et al., Live Imaging of Micro and Macro Wettability Variations of Carbonate Oil Reservoirs for Enhanced Oil Recovery and CO2 Trapping/Storage, *Scientific Reports, Nature, 2022*
DOI: 10.1038/s41598-021-04661-2

2. E. Tokareva, I. Tkachev, G. Sansiev, G. Fedorchenko, A. Ivanova, P. Grishin, S. Markovic et al., Study of the process of hydrophobization of carbonate rock with organic acids (Russian), *Neftyanoe khozyaystvo-Oil Industry, OnePetro, 2022*
DOI: 10.24887/0028-2448-2022-3-73-76

Acknowledgments

I wish to express my deepest gratitude to my thesis jury members, supervisors, colleagues, as well as my friends and family, for their indispensable help, for sharing ideas, and for giving me the motivation to complete this thesis.

First, I would like to express my most profound appreciation to my thesis jury members for their valuable time and energy in helping me to improve this dissertation with their thorough and thought-provoking feedback and ideas for future research.

Most of all, I wish to thank my advisor, Alexey Cheremisin, for introducing me to NMR petrophysics during my master's studies. Since then, he has constantly encouraged me to explore new fields and develop new skills and has shown extreme patience throughout my PhD studies. I am particularly grateful to Alexey for endorsing me for the joint degree program with Curtin University and enabling me to visit and expand my network with the University of Calgary. His strong support of my ideas and research directions resulted in a fruitful collaboration between the three institutions.

I am incredibly grateful to my supervisor from Curtin University, Reza Rezaee, for always being available and for enormous understanding and support for the past two years of my remote online studies. Due to Covid-19, we haven't had an opportunity to meet in person at WASM, yet I feel like I was there for many years. Special thanks for helping me define the T_2 parameters for the last article and for reminding me about the T_2 peak.

I am very grateful to my co-supervisor, Dmitry Koroteev, for inspiring me to dive deeper into the machine learning field and for being supportive and open to sharing his ideas. I also thank Dmitry for appreciating my research strengths and patiently encouraging me to improve in my weaker areas. I greatly appreciate his advice on writing and communicating our findings better, especially when we were about to submit our first article.

I am also deeply indebted to Apostolos Kantzas, Jonathan Bryan, Sergey Kryuchkov, and Raj Mehta from the University of Calgary. Raj introduced me to the Prof. Apostolos group during my master's studies in 2017, and we have collaborated ever since. Without their support, none of my research would be possible. I am sincerely grateful to Apostolos for enabling me to join his group and providing the opportunity to learn more about the NMR petrophysics for heavy oils and use his lab, samples, and data for our studies.

I cannot begin to express my thanks to Jon, who spent countless hours reviewing my notes, methods, and manuscripts and making time to discuss my concerns even in the busiest times. Many thanks for fruitful and engaging discussions and for introducing me to the problems in heavy oil sands NMR petrophysics. I earnestly hope to have a chance to continue working on new projects in the future. I wish also to thank Sergey for practical advice, patience, and invaluable insight into NMR physics and CPMG decay curve inversion and regularization.

Special thanks to my lab and office mates (and desk neighbors) – Aman Turakhanov, Aliya Mukhametdinova, Aysylu Askarova, Anastasia Ivanova and Mohammad Ebadi. Many thanks to Aman for the long and entertaining discussions about NMR, data science, and EOR. Also, special thanks for introducing me to OriginPro. I am also grateful to Aliya for discussions on 2D NMR mapping, for teaching me how to work with Geospec 2, and always being there when I got stuck in details while writing. Thanks also to Aysylu for support and encouragement during IBA competitions and our research on hot water injection in carbonates. I am also grateful to Anastasia for introducing me to the wettability alteration of carbonates and for taking the time to answer my questions and share her thoughts about the topic and our experimental findings. Special thanks to Mohammad, for sharing his advice on scientific writing and publishing, and for the long discussions about data science in our field.

I thank my friends from Skoltech and back home in Serbia for the fun and support during my PhD journey. My sincere thanks go to my parents, Vladimir and Mirjana,

and my sisters Milica and Svetlana for their love, support, and understanding. You inspired me to pursue happiness and become a better person.

Finally, but not least, all my heartfelt thanks go to my wife, Kruna, without whom I would be lost. She gave me support and help and discussed my ideas. You were my emotional anchor every step of the way; without you, this work would not be possible.

Dedication

I dedicate this dissertation to my late father, who passed during my PhD studies. Thank you for inspiring me to dive into Geoscience.

To Vladimir Markovic

Contents

List of Symbols, Abbreviations

LF-NMR	Low-field nuclear magnetic resonance
OOIP	Original oil in place
EOR	Enhanced oil recovery
TE	Echo spacing
COD	Coefficient of determination
RMSE	Root mean square error
MAE	Mean absolute error
MSLE	Mean square logarithmic error
MAPE	Mean absolute percentage error
MaAE	Maximum absolute error
XGB	Extreme gradient boosting - Xgboost
FE	Feature engineering
FS	Full spectrum
XGB-FE	Extreme gradient boosting model based on feature engineering
XGB-FS	Extreme gradient boosting model based on full T_2 distribution
DS	Dean-Stark
DS-w	Water content by Dean-Stark
DS-o	Oil content by Dean-Stark
L1	Lasso regression
L2	Ridge regression
CPMG	Carr-Purcell-Meiboom-Gill pulse sequence
SNR	Signal-to-noise ratio
CANOVA	Continuous analysis of covariance
MI	Mutual information
BO	Bayesian optimization
LOOCV	Leave-one-out cross-validation
X-ray CT	X-ray computed tomography
BSS	Blind-source signal separation
ANN	Artificial neural network(s)

SL	Supervised learning
PLS	Partial least squares
HI	Hydrogen index
RHI	Relative hydrogen index (mass)
RHI_v	Relative hydrogen index (volume)
SRM	Structural risk minimization
ERM	Empirical risk minimization
LAD	Least absolute deviation
GS-CV	Grid-search cross-validation
RBF	Radial basis function
SV	Support vector
SVR	Support vector regression
SVM	Support vector machine(s)
GBRT	Gradient boosted regression trees
MLR	Multiple linear regression
K-NN	K nearest neighbors
DT	Decision trees
RF	Random forests
NLS	Non-linear least squares
ODR	Orthogonal distance regression
GOR	Gas oil ratio
VAPEX	Vapor extraction
SAGD	Steam-assisted gravity drainage
ISC	In-situ combustion
DTS	Distributed temperature sensing
CSS	Cyclic steam stimulation
FID	Free induction decay
NE	Number of echoes within CPMG sequence
NT	Number of trains within CPMG sequence
TW	Wait-time within CPMG sequence

ILT	Inverse Laplace transform
BBP	Bloembergen-Purcell-Pound NMR relaxation model
RHOB	Bulk density log
NPHI	Neutron porosity log
FFI	Free-fluid index
DSE	Debye-Stokes-Einstein model
ASTM	American Society for Testing and Materials
PEEK	Polyether ether ketone
M-L	Marquardt-Levenberg

List of Figures

List of Tables

INTRODUCTION

This work studied different types of unconventional hydrocarbons, namely oil sands, heavy oils, and bitumens. These hydrocarbons represent the most abundant unconventional hydrocarbon resources worldwide. Due to their high viscosity, thermally enhanced oil recovery (EOR) methods are often required for their economic production, followed by extensive refining processes for their downstream distribution. Therefore, the breakeven price for developing such resources is very high, which stimulates the major oil producers to constantly invest in developing new technologies to decrease their exploration and recovery prices. Petrophysical well-logging involves optimizing and developing new methods for the improved *in-situ* characterization of these hydrocarbons.

In light of the foregoing, I focused my research on advancing conventional and establishing new analytical and data-driven methods for *in-situ* characterization of heavy oil and bitumen resources, primarily using LF-NMR measurements. Special attention is paid to developing new workflows and models for predicting oil viscosity and water saturation, two critical factors for the recovery of hydrocarbons.

1.1. Heavy oil and bitumen resources

Bitumens and heavy oils usually form deposits in shallower geological settings, where due to the cold temperatures and a lack of the caprocks, they are being subjected to bacterial biodegradation. The depth of these deposits usually does not exceeds 4 km, while the temperatures are most often lower than 80 °C[1]. In such conditions, the biodegradation process spans over geological periods, where hydrocarbon-degrading bacteria use lighter oil fractions for the metabolic processes, which gradually reduces the viscosity, initial mass, and gas-to-oil ratio (GOR) of the oil.

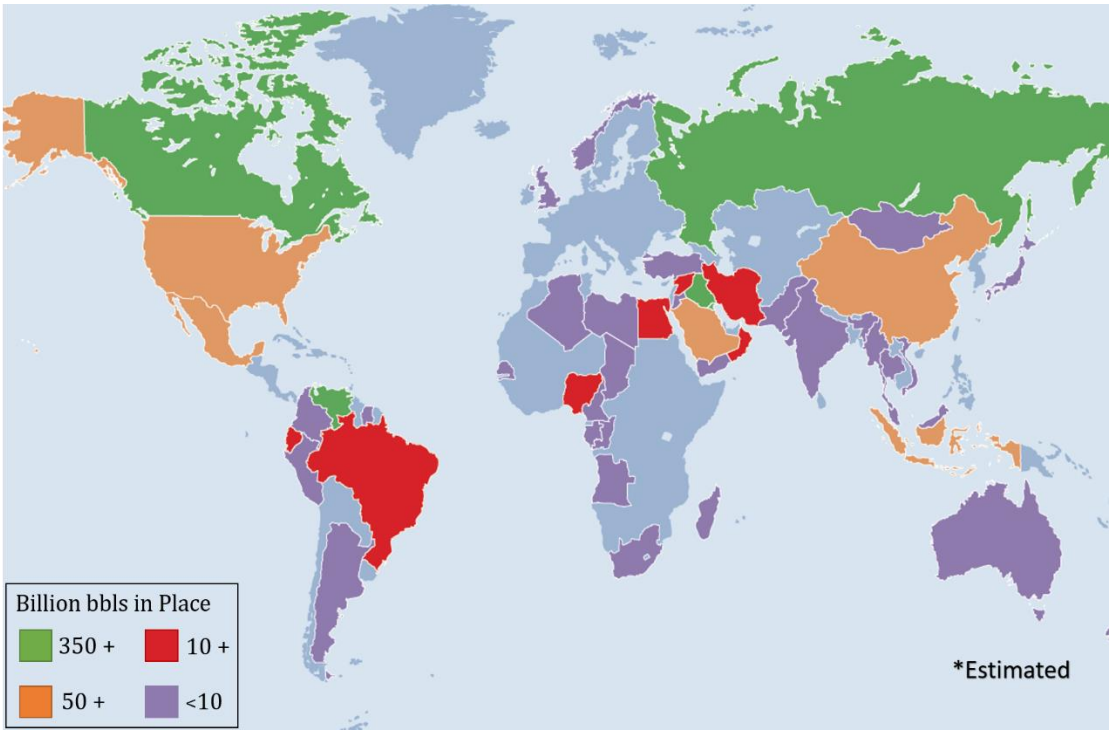


Figure 1: Estimated worldwide heavy oil reserves by Country (copyright Schlumberger).

Heavy oil and bitumen are among the most abundant hydrocarbon resources worldwide. They differ from other hydrocarbon resources because of their high viscosity, density, and increased concentration of heavy components such as asphaltenes, resins, and wax. Along with these components, heavy oils and bitumens often contain various heavy metals, sulfur, and nitrogen. Due to these factors, their economic value is considerably smaller than their lighter counterparts, as their recovery and refinement require the deployment of costly

technologies and time-demanding processing. However, their large in-place volumes and high oil market prices allow their profitable extraction and refinement, so major oil companies acquire licenses for their development as an alternative to conventional oil resources.

Studies estimate that heavy oil and bitumen reserves today amount to about 55% of the total world reserves (nearly 9 trillion barrels), while the most extensive deposits were found in Canada, Venezuela, and Russia[1].

1.2. Enhanced Oil Recovery (EOR) methods for heavy oil fields

Heavy oils and bitumens have distinctively high viscosities, so their recovery from the reservoir rocks is often performed using various thermal Enhanced Oil Recovery (EOR) techniques such as hot steam injection or in-situ combustion, where the heat exchange process reduces oil viscosity in the formation of interest[2]. In heavy oil reservoirs, the viscosity may vary up to a hundred times in vertical and horizontal directions. Therefore information about the spatial distribution of viscosity can affect not only the well placement and injection or production rates but mathematical reservoir simulations as well[3].

In literature, the EOR methods are generally divided into two groups: cold and thermal production. The cold production methods are economically and technically convenient recovery methods; however, they are limited to shallow deposits since they involve open-pit mining. Waterflooding was also occasionally used with some success, but in fields where oil viscosity was not more than 100 cP, the sweeping efficiency of the waterfront reduces with an increase of viscosity, most notably due to the viscous fingering[1]. Vapor-assisted extraction (VAPEX) is another cold production method that proved efficient for reservoirs with oil viscosities ranging from 100 cP up to 130,000 cP [4,5]. It is based on miscible solvent vapor injection into parallel stacked horizontal wells. The miscible solvent vapor is injected into the upper well, where the vapor chamber is produced around the well. This causes the dilution of solvent vapor into surrounding heavy oil and

bitumen, which leads to viscosity reduction and drainage of the hydrocarbons to the extraction well beneath, with the help of gravity. This process is energy efficient since it is not based on thermal exchange and does not require the infrastructure for generating hot steam or water.

In thermal methods, on the other hand, increasing the temperature in the well or reservoir causes viscosity to decrease, thus improving the heavy oil and bitumen mobility. For instance, in steam-assisted gravity drainage (SAGD), the placement of the injection and producer wells is almost identical to the one in the VAPEX method; however, instead of the solvent vapor, hot steam is injected. After the hydrodynamic linking between wells is obtained by the initial steam treatment of both wells, steam injection is continued in the injection well, where hydrocarbons are affected by the steam expansion. The temperatures of this treatment can be over 200 °C, thus initiating a highly mobile gravitational drainage of bitumen and heavy oil towards the underlying producing well. This method and its' variations were successfully utilized in numerous projects, most notably in Canadian heavy oil reservoirs where viscosities vary from a couple of hundreds up to $3 \cdot 10^6$ cP [6]. In addition to the SAGD, cyclic steam stimulation (CSS) is used, where the same well is used for injection and production. In this approach, hot steam is injected into the targeted formation and left to soak up to heat the oil, followed by the production cycle. Reports in the literature show that CSS is suitable for oil viscosities from 50 cP to 350,000 cP [7]. As a follow-up to CSS, steam flooding is usually performed, where steam is continuously injected into the targeted formation without a soaking period. Since this process is heavily influenced by horizontal and vertical sweep efficiency, much attention is given to monitoring steam fingering and channeling, which is why steam flooding has shown to be most successful for reservoirs with oil viscosity ranging from 20 – 20,000 cP [8]. Another well-known thermal method is in-situ combustion (ISC) or fire flooding, where an oxidizing gas (i.e., air) is injected into the formation, which causes the ignition of heavy oil in-situ. In this set-up, the portion of the hydrocarbon's heavier components is utilized as a fuel for further propagation of the combustion front,

where exothermal reaction heats the surrounding rocks and thus lowers the oil viscosity. An added benefit of ISC is that the heavier components are being consumed due to the thermal cracking, which results in upgraded oil. Although there were many pilot studies for ISC between the 1960s - 1980s in heavy oil reservoirs with oil viscosities from $40 \cdot 10^3$ to $1 \cdot 10^6$ cP, the commercially successful ISC projects were performed in reservoirs with oil viscosity between 20 – 8000 cP [9].

The success of the EOR project in heavy oil and bitumen deposits is dictated mainly by the initial viscosity of the hydrocarbons and their spatial distribution in the field. In steam EOR, for instance, monitoring of the steam front is often performed using periodic or permanent four-dimensional (4D) seismic surveying by collecting pressure and fluid saturation data within the field[10]. Monitoring based on logging is also commonly implemented by fiber-optic distributed temperature sensing (DTS), which provides constant temperature monitoring along the wellbore. NMR measurements have been considered for monitoring the fluid saturation changes and wettability alteration assessing[10,11]. If this technology is used for reservoir monitoring, NMR logging could also be used as a continuous, non-invasive technique for monitoring viscosity variations and changes caused by temperature variation in the reservoir. The NMR equipment can be used in-situ in observation wells, online, or inline for the heavy oil viscosity monitoring during a thermal EOR project with a 200 °C upper-temperature limit, which corresponds to steam injection temperatures, and other EOR methods such as solvent or miscible gas injection [12].

1.3. NMR theory review

1.3.1. Spin-lattice (T_1) and spin-spin (T_2) relaxation

The hydrogen atom in its core contains a single proton with a positive charge (H^+). Protons have spins and exhibit magnetic behavior, which is why an external

magnetic field can control the orientation of their spins. For instance, if a magnetic field (B_0) is introduced to the H^+ proton-containing system, the proton spins will tend to align along the direction of the B_0. The quantum theory posits that the protons will be distributed to a low or high energy state corresponding to spins -½ and ½. The difference between the number of protons in high and low energy states will generate the total or bulk magnetization M_0 that NMR tools can measure.

$$M_0 = N\frac{\gamma^2\hbar^2 I(1+I)}{3(4\pi^2)kT}B_0 \qquad (1)$$

where N is the number of protons in a unit volume, γ is a gyromagnetic ratio, I is the quantum spin number, T is the temperature in Kelvins, and k and $\hbar$ are Boltzmann's and Planck's constants, respectively. The polarization of protons is not immediate but grows by a specific time constant. This constant is called spin-lattice or longitudinal relaxation time (T_1). As the polarization is exponential, and under the assumption that the polarization orientation transpires along the z-axis in 3D space, then:

$$M_z(t) = M_0\left(1 - e^{\frac{-t}{T_1}}\right) \qquad (2)$$

where t is polarization time, and $M_z(t)$ is polarization magnitude along the z-axis at the time t. The T_1 time represents a moment at which ~63% of the magnetization is reached. At three T_1, about 95% of magnetization is reached (Figure 2).

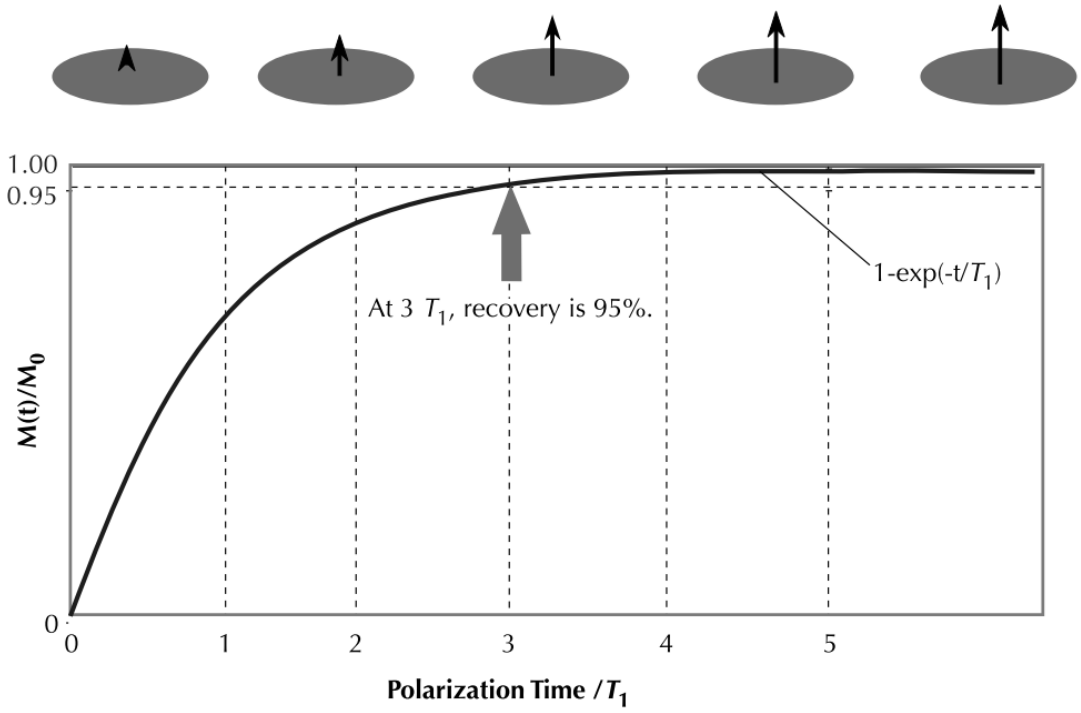


Figure 2: The proton polarization in a B_0 static magnetic field [13]. T_1 curve is presented as a function of time.

For nuclear magnetic resonance to occur, it is necessary to perform the pulse tipping of the protons (Figure 3). While protons are polarized along the B_0, an additional short radio-frequency oscillating pulse (B_1) is applied to the system. The B_1 Larmor frequency (f) must correspond to the Larmor frequency of the spins to achieve the resonance. The total magnetic moment or the tipping angle (θ) is defined as:

$$\theta = \gamma B_1 \tau \qquad (3)$$

where τ is the period when the B_1 is applied to the system. The tipping angle can be controlled by both B_1 and τ.

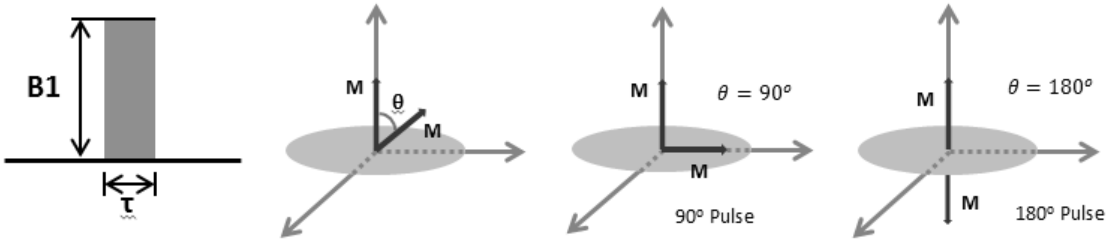


Figure 3: Tipping process by 90- and 180-degree pulse. The angle is controlled by B_1 field and τ time period of B_1 application [14].

In practice, the first B_1 pulse tips the proton system into the perpendicular plane (x, y) relative to the B_0 (z), thus $\theta = 90^o$. At this point, the protons are precessing about B_0 and are in phase, and the NMR device can detect their signal. However, immediately as the B_1 stops, the protons start to dephase due to the B_0 magnetic-field inhomogeneity and molecular tumbling, usually at an exponential rate. Therefore, their signal decays at the particular time constant (T_2^*) and is called free induction decay (FID).

Fortunately, the dephasing due to the B_0 inhomogeneity can be reversed by adding the B_1 pulse ($\theta = 180^o$) after τ time elapsed from the first pulse ($\theta = 90^o$). The second pulse flips the protons, which reverses the dephasing process, meaning that the protons will return to the same phase after τ time elapsed from the second pulse. This subsequent pulse is referred to as a 'refocusing pulse.' When spins return to the same phase, the 'spin echo' signal is produced. Generation of one spin echo is illustrated on Figure 4. The NMR device can be configured to produce a series of refocusing pulses, thus generating a series of spin echoes. This series or sequence of refocusing 180^o pulses is known as Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence, while the recorded sequence of echoes is called the 'spin echo train' [13].

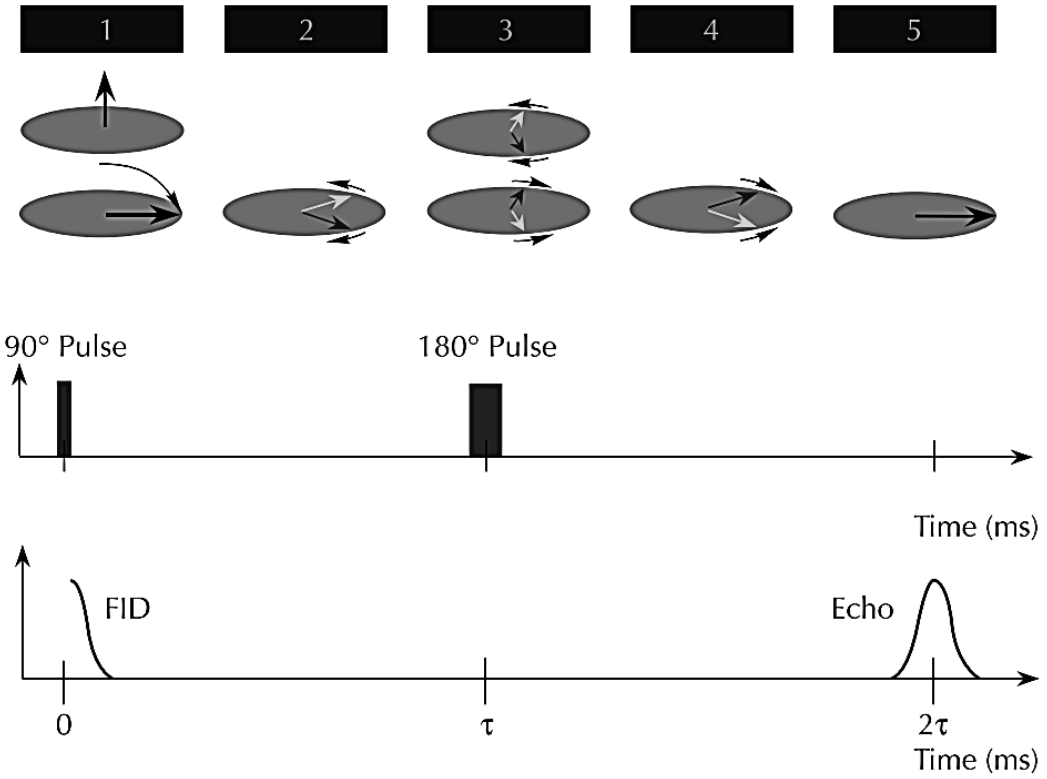


Figure 4: Spin echo generation steps. (1) A 90° B_1 pulse, (2) spin dephasing, (3) 180° pulse after τ time flips the spins, (4) rephrasing, (5) spins in phase generate an echo at 2τ [13].

Although the CPMG pulse sequence can reverse the effect of B_0 field inhomogeneity, the dephasing of spins due to molecular tumbling and diffusion is irreversible. Since the dephasing is mainly due to the interaction of spins, the decay of magnetization in the horizontal plane (x, y) is called spin-spin relaxation, or transversal relaxation, as the magnetization is in the transversal plane relative to B_0. The time constant associated with the decay rate is T_2 relaxation time.

$$M_{xy}(t) = M_{xy}(0)\, e^{\frac{-t}{T_2}} \qquad (4)$$

where $M_{xy}(0)$ is magnetization magnitude at $t=0$. According to Equation 4, for one T_2 time constant, the magnitude of M_{xy} will drop to ~37% of its initial value, and after three T_2 constants to ~95%.

1.3.2. Spin-spin (T_2) relaxation mechanisms of fluids in pore space

There are three separate T_2 relaxation mechanisms that fluids experience in pore space –T_2 bulk relaxation, T_2 surface relaxation, and T_2 diffusion-induced relaxation due to the gradients in the magnetic field:

$$\frac{1}{T_2} = \frac{1}{T_{2bulk}} + \frac{1}{T_{2surface}} + \frac{1}{T_{2diffusion}} \quad (5)$$

Here, T_2 is the spin-spin relaxation time of fluids in pores and results from a sum of T_{2bulk} and $T_{2surface}$ relaxation and relaxation due to $T_{2diffusion}$. The T_{2bulk} is an intrinsic spin-spin relaxation component mainly dependent on fluids' viscosity and chemical structure. To quantify it, one can place a fluid sample into a large vessel, expose it to a homogenous magnetic field, and perform a CPMG pulse sequence. In such a set-up, a vessel behaves like a large pore; therefore, surface relaxation has a negligible contribution to total T_2 relaxation. Some other factors which can affect the rate of T_{2bulk} in practice are pressure and temperature oscillations. It has been shown that T_{2bulk} relaxation of water and dead oil are generally[13,15]:

Water $$\frac{1}{T_{2bulk}} \cong \frac{298\eta}{3T_k} \quad (6)$$

Dead oil $$\frac{1}{T_{2bulk}} \cong \frac{\eta}{0.00713T_k} \quad (7)$$

where η is viscosity and T_k is the temperature in °K.

$T_{2surface}$ relaxation mechanism occurs in pore space, and it takes place at the interface of the pore wall and fluid. It can be calculated as:

$$\frac{1}{T_{2surface}} = \rho_2 \left(\frac{S}{V}\right) \quad (8)$$

where ρ_2 is the T_2 surface relaxivity of the pore wall, S is the pore surface, and V is the fluid volume. Since surface relaxivity is a property of a pore wall, its value depends on the mineralogy of the sample. These estimates are most often determined in laboratory experiments. The upside of $T_{2surface}$ relaxation is that it

is not affected by temperature and pressure, meaning that additional calibrations for reservoir conditions are not necessary after lab experiments under ambient conditions.

The $T_{2diffusion}$ relaxation is significant mainly for gas, water, and oils of low and medium viscosity when exposed to a magnetic field with gradients and CPMG pulse sequence with long echo spacing. If the gradient in the field is substantial, additional dephasing can occur due to molecular diffusion, resulting in faster T_2 relaxation. $T_{2diffusion}$ can be expressed as:

$$\frac{1}{T_{2diffusion}} = \frac{D(\gamma GTE)^2}{12} \quad (9)$$

where D is molecular diffusion, γ is proton gyromagnetic ratio, G is a field strength (gauss/cm), and TE is echo spacing.

1.3.3. CPMG pulse sequence configuration and NMR data processing

The CPMG pulse sequence consists of a 90° initial pulse that tips the polarized protons into an x, y-plane, followed by several refocusing 180° pulses. The quality of the obtained CPMG decay data, its inversion, and the quality of interpretation are strongly dependent on the configuration of the CPMG sequence. The principal controlling parameters include the time between two pulses or echo spacing (TE), the polarization time or waiting time (TW), the number of pulses or the number of echoes (NE), and a number of CPMG trains (NT).

Setting an NMR tool to a short echo spacing (TE) will influence the signal-to-noise ratio (SNR) twofold; first, the density of spin-echoes within a train will increase; second, the echoes will be recorded earlier. Consequently, this leads to increased SNR. However, the experiment time will increase proportionally when the number of echoes (NE) increases. The same is true for the number of trains (NT) and waiting time (TW). It should be noted that TW should be configured according to the sample or reservoir interval and the purpose. If the goal is to characterize heavy oil and clay-bound water, the TW can be decreased since protons will return to equilibrium much faster than pure water or light oil.

The representation of the decaying echo train in the time domain is usually obtained by the Laplace Inverse Transform (ILT)[16], whereas the output of a T_2 distribution is obtained (Figure 3). However, the inversion of the NMR signal represents the ill-posed problem since minor perturbations (i.e., noise) in the measurements can substantially impact the T_2 distribution form, that is, the stability of the solution. A stable and unique solution can be obtained if the inversion of the signal is performed numerically. In such a case, the representation of the T_2 decay is achieved from echo-fitting. To simplify the echo-fitting, the discretization of the T_2 decay signal can be performed, where the predefined number of discrete T_{2i} relaxation times correspond to individual exponential decay. Then, the set of echo trains can be expressed as a system of linear equations where each equation corresponds to an individual echo train. Since the fitting is performed to the sum of multi-exponentials, the stable solution can be obtained by solving non-negative least squares. The standard approaches are Lawson-Hanson and Tikhonov regularization[17].

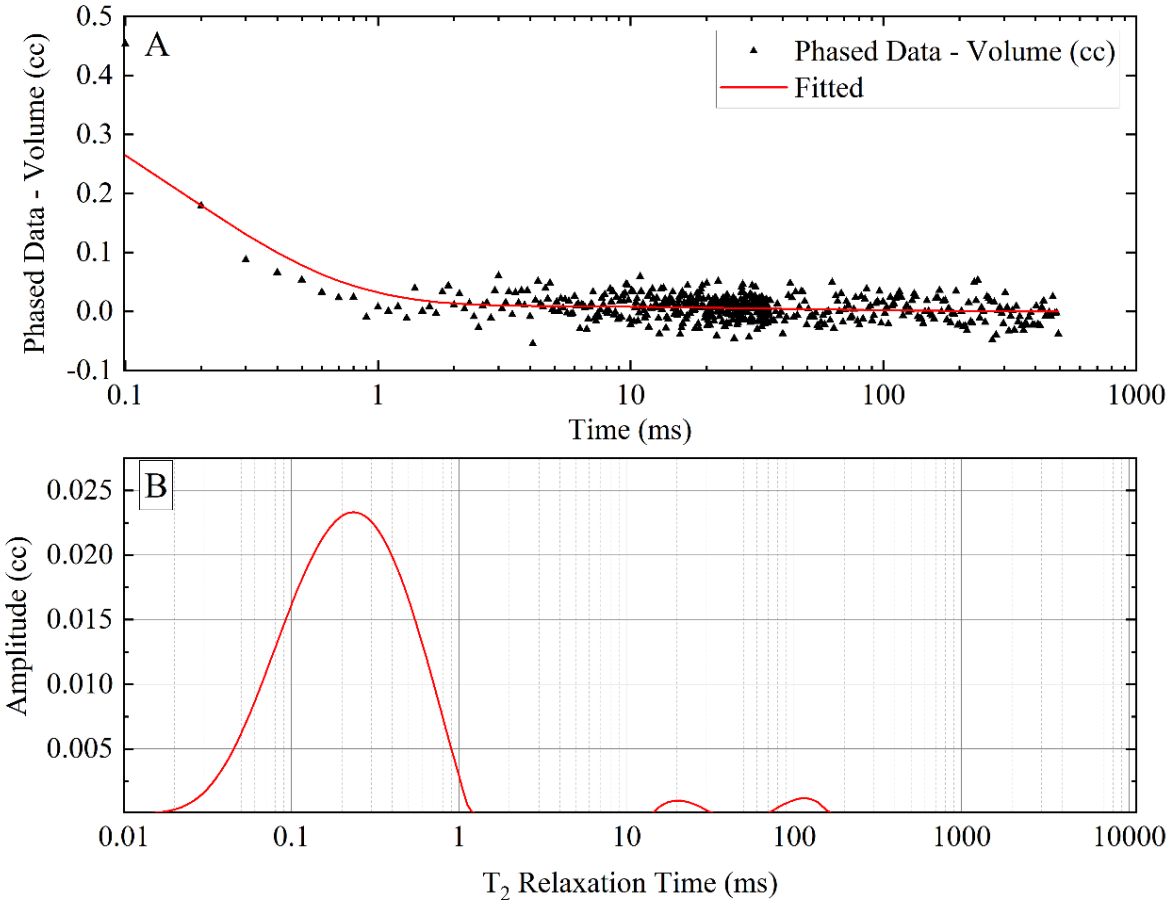


Figure 5: (A) Raw T_2 NMR relaxation data and (B) T_2 data after inversion.

In NMR petrophysical practice, the essential consideration is tuning the tool's configuration for the specific reservoir and hydrocarbon type and choosing the adequate inversion procedure; the processes which significantly impacts the form of T_2 relaxation distribution. Reports in the literature show that this is particularly important for unconventional reservoirs such as shales and heavy oil sands. For instance, the non-negative least squares inversion by the Lawson-Hanson approach has been shown to generally converge to very smooth T_2 distributions, which may lead to the merging of amplitudes, and a loss of valuable information [17]. In addition, shales and heavy oils produce a considerable amount of T_2 signal at fast relaxing parts of T_2 distribution where measurability is low. That is why parameters such as echo-spacing, waiting time, number of pulses, and trains must

be carefully tuned to achieve a balanced tradeoff between signal-to-noise ratio, the shape of T_2 distribution, and measurement time.

1.4. Oil viscosity

1.4.1. Conventional measurements of oil viscosity

Viscosity is a physical fluid property that reflects the amount of internal friction of a fluid. In other words, the fluid's viscosity indicates the magnitude of the resistance to flow. Mobilizing a fluid requires a certain amount of force, and the rate of change of the induced deformation can be measured in function of time. Since viscous forces control the flow velocity in fluids, we measure the flow velocity for the applied amount of force or applied amount of shear stress. Bryan et al., in their work, link the relationship of rheological viscosity measurements with Eyring's theory of viscosity, which is also consistent with the Bloembergen-Purcell-Pound (BPP) NMR relaxation model[18,19]. As the fluid undergoes shearing deformation, the rate of shearing is proportional to the applied shear stress (Equation 10).

$$\tau_{ss} = \eta \cdot \gamma_{sr} \tag{10}$$

where τ_{ss} presents the amount of shear stress applied to the fluid, γ_{sr} stands for the shear rate, and η is the fluid viscosity. Based on the Eyring's theory of viscosity, molecules in fluids are structured in lattices, while intermolecular space remains vacant but not spacious enough for other molecules to shift through readily[20]. However, if force is applied, the molecules will reconfigure their positions until the vacant space becomes large enough for another molecule to enter. Eyring proposed an analytical model for such behavior (Equation 11)

$$\eta = \left(\frac{\delta}{a}\right)^2 \cdot \frac{N\hbar}{V} \cdot e^{\left(\frac{\Delta G_0}{RT}\right)} \tag{11}$$

where δ is intermolecular layer distance, a is the distance between the vacant space and a molecule, N is Avogadro's number, $\hbar$ is Planck's constant, and V is a fluid's molar volume G_0 activation energy, R is the universal gas constant, and T is the absolute temperature. Heavy oils and bitumens comprise long chain-like

molecules, cyclic paraffin, and branches from heavy components. To mobilize these molecules, more activation energy is required (G_0) since the attractive forces of surrounding molecules hinder the movement of molecules attempting to occupy the vacant space. In rheological measurements of heavy oils by a cone and plate viscometers, this resistance to flow translates to high shear stresses. From Equation 11, it can be observed that the temperature is in an exponential relationship with viscosity. If we apply heat to the heavy oil, the heavy components will gain more energy while the intermolecular distance will increase, enabling the less restricted motion of molecules. This is also consistent with rheological viscosity measurements, where for fixed shear stress and with an increase in temperature, we observe increased shear rate[18]. Equation 6 can be re-written by substituting constants (Equation 12).

$$\eta = A \cdot e^{\left(\frac{E_a}{RT}\right)} \tag{12}$$

where A is a constant and E_a is the viscous free energy of activation. In this Arrhenius-type equation, the values of E_a and A vary for different oil samples. These variations are associated with molecular weight variation of different oil components and their chemical composition and structure[20]. These variations may be substantial since the composition of heavy oils and bitumen, in particular, can be significantly different, indicating that derivation of the general viscosity model is challenging.

1.4.2. Oil viscosity by LF-NMR measurements

The oil is a blend of a diverse range of liquid hydrocarbons with inconsistent molecular structure [21]. When it has a higher proportion of complex high molecular weight compounds such as asphaltenes and resins, oil viscosity will be higher, signifying that viscosity reflects oil's chemical complexity[22]. 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 to characterize hydrogen-bearing liquids due to their ability to rapidly

convey a series of contactless, non-invasive experiments. In NMR petrophysical logging, the pulsed NMR tools typically generate magnetic fields between 120 to 550 gauss (0.012 to 0.055 tesla) [23]. In this work, low-field NMR is defined by fields below 550 gauss.

To relate a fluid viscosity to NMR relaxation, it is necessary to comprehend and model the molecular interactions. Equations 11 and 12 demonstrate that viscosity can be expressed without macroscopic flow or shearing (Equation 10). These findings are consistent with a theoretical Bloembergen-Purcell-Pound (BPP) NMR relaxation model, associated Debye-Einstein-Stokes spherical molecules model, which anticipates different rotational correlation times (τ_c) for various molecule sizes [19,24,25]. The correlation time represents the mean time required for a molecule to rotate one radian and is a crucial parameter for determining microviscosity. It is also a fundamental component of the BPP relaxation model. Based on this relationship, one can study the association of T_2 relaxation with viscosity[25]. The physics of this relationship is discussed in section 2.2.

As previously mentioned, the T_2 relaxation distribution after mathematical inversion can be represented in a time domain. Distributions of T_2 relaxation between different fluids can be compared using a mean T_2 distribution time, such as T_2 logarithmic mean (T_{2lm}). When fluids are measured in a bulk state, the primary relaxation mechanism will be bulk relaxation (T_{2B}), which is due to the energy exchange of the H spins and diffusion. Straley et al.[23] and Coates et al.[13] have experimentally shown that the T_{2B} is proportional to the ratio of temperature (T) and viscosity (η):

$$\frac{1}{T_{2B}} \propto \frac{\eta}{T} \tag{13}$$

Since the relaxation times of light oils, water and gases are long, and high viscosity fluids such as heavy oils have short relaxation times, the T_{2B} and T_{2lm} can be correlated with oil viscosity[23,26,27]. Although this relationship is used as a foundation for nearly all existing NMR viscosity models, it only works well for the

light and medium viscous oils (~1-800 cP)[28] composed of lighter hydrocarbons with a less complex chemical structure. In the case of heavier oils, the T_2 relaxation deviates from the classical BPP model, and the relationship described in Equation 13 alters significantly.

In the past 30 years, many analytical NMR viscosity models have been proposed for characterizing crude oil. However, the reports in the literature show inconsistency in the prediction accuracy of these models due to three main reasons: use of the light oil NMR models for the prediction of heavy oil viscosity[15,23,27,29], use of models (including heavy oil models) without prior tuning for a given reservoir or a suit of oils, and due to use of ambiguous mathematical procedures for model tuning[3,30,31]. Moreover, there were several attempts to develop a "universal model" for *in-situ* heavy oil viscosity prediction aiming to estimate viscosity in the formations with weak prior knowledge about the oil properties[30–33]. These models have default parameters derived for heavy oils from a particular oil field. However, when applied to different heavy oils, they generate significant prediction errors, in some cases over a factor of three[31]. To develop a universal analytical NMR viscosity model for systems with oils of various compositions would be contradictory to Debye-Stokes-Einstein's findings stating that different correlation times are expected for different molecule sizes [24]. However, for the order-of-magnitude *in-situ* viscosity evaluation, existing analytical models can be improved to a degree where more reliable estimates can be utilized to optimize the decision-making process for viscosity variation monitoring during EOR projects.

Although LF-NMR 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 [31,33–37]. 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 helpful 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 [38–42].

1.4.3. LF-NMR oil viscosity measurements in other industrial fields

The application of NMR viscosity models is not relevant only for petrophysical well-logging. 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[43]. Among many studied physicochemical properties, the oil viscosity showed to be of the principal importance in determining the rate of interaction with fuel during combustion processes in internal combustion engines [44,45]. 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 efficient. 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[46,47]. In the LF-NMR examination of petroleum fractions, this nonlinearity can occur due to their chemical intricacy, leading to the degradation of model forecasting performance[24].

Moreover, in mechanical systems (tribosystems), viscosity reflects the oil's capacity to render the sufficient thickness of the lubricating film between the surfaces exposed to friction. In order to efficiently buffer the rate of machinery wear, the oil selection is made under the speed-load and temperature conditions of the system[48]. 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, occurring at high temperatures[49]. 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 micro-displacement methods [42]. In circumstances where these instruments would be difficult to utilize, the LF-NMR tools could be used instead for non-invasive, real-time viscosity monitoring[42,50,51]. As the operating conditions of these systems may lead to significant oil viscosity fluctuations, the robust data-driven or analytical NMR model could be used to measure the fluctuations accurately and, in that manner, help in the early detection of equipment failure.

1.5. Water saturation

1.5.1. Water saturation by resistivity measurements

One of the primary purposes of petrophysical formation evaluation is the quantification of hydrocarbon and water saturations. To properly evaluate the volume of hydrocarbons, it is necessary to determine water saturation beforehand, since generally, for pressures above the bubble point:

$$S_o = 1 - S_w \tag{14}$$

where S_o is oil saturation and S_w is water saturation. Conventionally, in well-logging practice, the water saturation was determined using resistivity logs, and depending on the reservoir type, resistivity data would be used in combination with other standard logs such as density (RHOB) and neutron (NPHI)[52]. The widely used empirical model for water saturation estimation in conventional hydrocarbon reservoirs was developed by Archie[53] (Equation 17).

$$\mathrm{F} = \frac{\mathrm{R_0}}{\mathrm{R_w}} = \frac{\mathrm{a}}{\varphi^{\mathrm{m}}} \tag{15}$$

$$\mathrm{I_r} = \frac{\mathrm{R_t}}{\mathrm{R_0}} = \frac{1}{\mathrm{S_w^n}} \tag{16}$$

$$\mathrm{S_w} = \sqrt[\mathrm{n}]{\frac{\mathrm{aR_w}}{\mathrm{R_t}\varphi^{\mathrm{m}}}} \tag{17}$$

In Equation 15, F is a formation factor, R_0 fully water-saturated rock resistivity, R_w brine resistivity, φ fractional porosity, a and m are tortuosity coefficient and cementation exponent, respectively. In Equation 16, I_r is the resistivity index, R_t is rock resistivity, R_0 is the resistivity of fully water-saturated rock resistivity, S_w is fractional water saturation of the formation, and n is the saturation exponent. Finally, Equation 17 presents Archie's water saturation model, which is obtained by combining Equations 15 and 16. The *m, n,* and *a* are known as rock resistivity parameters. While most of the parameters can be obtained from conventional logs, particular attention is required to calibrate *a, m, n.* These are obtained from laboratory-controlled resistivity tests and subsequent least squares regression.

Archie's equation was successfully used in systems with simple, uniform pore space saturated by water[54]. However, issues arise in reservoirs with large amounts of clay-bound and capillary-bound water, strong variations of salinity with depth, and formations containing clays or conductive minerals such as pyrite[55,56]. This is also true for Canadian oil-sands[57]. In these terms, the presence of bound water and clays will cause the underprediction of OOIP, while the variable salinity can cause either overprediction or underprediction of OOIP.

1.5.2. Water saturation by LF-NMR measurements – T_2 cutoff approach

Since the LF-NMR tools measure the response H^+ protons of the fluids, many of the issues relevant for resistivity logging can be avoided. Another advantage of NMR measurements is differentiating between irreducible water saturation (capillary and clay bound water) and producible fluids. The primary assumption is that larger pores are saturated by producible fluids, where the flow can occur in the presence of pressure gradient, while smaller pores contain fluids trapped by capillary forces or are bound within the lattice of clay minerals. If this condition is true, a T_2 cutoff value (location in T_2 distribution) can be defined, separating the T_2 distribution to signals corresponding to clay-bound, capillary-bound and

producible fluids. Integration of the separated regions can be performed and related to producible and bound fluid volumes. Straley et al.,[23] were the first to empirically identify the universal T_2 cutoff for clay-bound water at 3 ms in conventional sandstones. The cutoff value was determined by comparing the clay-bound water calculated from the ratio of cation exchange capacity and pore volume (Q_{ve}) with cumulative T_2 distribution porosity, using sandstone core plugs from 45 American and European oilfields. To evaluate the producible porosity or free-fluid index (FFI), it was necessary to perform NMR T_2 measurements on core samples in two states – cleaned and fully saturated state (S_w 100%) and after centrifuging to the irreducible water saturation (S_{wirr}). These experiments were performed for the suit of 86 sandstone samples, and the universal cutoff was found to be at 33 ms[23].

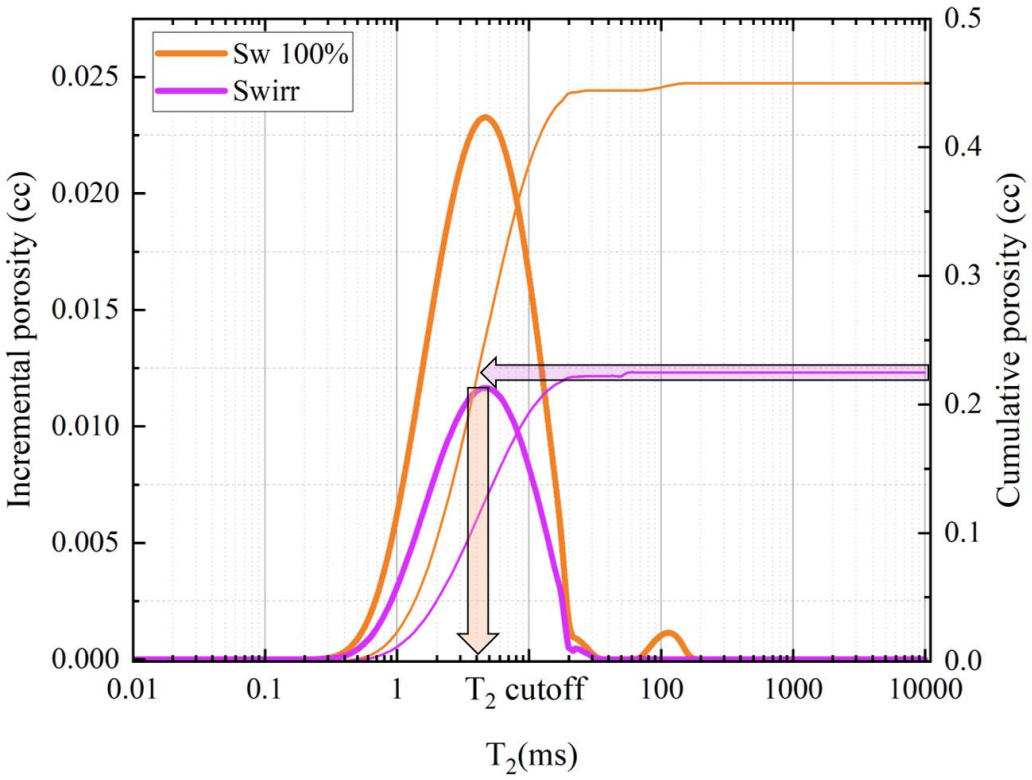


Figure 6: An example of the determination of T_2 cutoff value by LF-NMR and centrifuge. The orange curve presents the T_2 distribution of the 100% water-saturated sample (Sw 100%). The purple curve presents the T_2 distribution of the sample centrifuged to irreducible water saturation (S_{wirr}).

Although these universal cutoff times work for conventional sandstone reservoirs where porosity and permeability are generally uniform, many recent studies have shown that T_2 cutoff values vary dramatically for reservoirs of other lithologies, such as shales, carbonates, oil-sands, coals, and tight sandstones[58]. In addition, even if the T_2 cutoffs are determined experimentally, it is not recommended to use a single T_2 cutoff for the same well or oilfield since the T_2 distribution can vary drastically both vertically along the well and laterally, which would potentially cause erroneous estimation of OOIP[58].

1.5.3. Water saturation by LF-NMR measurements in oil-sands

The T_2 cutoff determination for oil-sands is even more problematic for two principal reasons. First, the centrifuging to S_{wirr} cannot be adequately performed since the sand exposed to severe centrifugal forces will lose its structure and native filtration properties, rendering the subsequent NMR experiments inadequate. Second, the heavy oil and bitumen in oil-sands have a fast T_2 relaxation time and produce a signal in the same region as capillary and clay-bound water, causing a significant overlap. One of the well-known approaches that had considerable success in addressing this effect is based on NMR spin-spin relaxation (T_2) distribution peak deconvolution[59]. The principal assumption behind this approach is that bitumen relaxes faster in an NMR distribution than surface-bound water, so early T_2 signals are attributed to bitumen, and later T_2 signals correspond to the capillary bound and free water saturation in the rock. For reference, Canadian bitumens produce a T_2 signal at approximately 0.1 – 4 ms range [60,61], the same as clay bound water in sandstones, while capillary bound water produces a signal roughly from 3 – 33 ms in sandstones [23]. Consequently, deconvolution approach works best for oil sands with low clay and capillary bound water content, where the overlapping of bitumen and water signals is not extensive.

Assuming that the oil-sands are largely water-wet, water will generally be found in the corners of connected sand grains and potentially as a thin film over the grain surface. The principal relaxation mechanism of hydrogen protons in high viscosity oils and bitumen would be bulk relaxation, while water would strongly influence surface relaxation, with bulk relaxation playing a minor role in the water T_2 values. Bulk relaxation and surface relaxation times of water and oils are unique for the most part, that is, the heavy oil molecules generally relax quicker relative to the water molecules. When their relaxation rates differ, the NMR T_2 distribution will display distinct oil and water responses (Figure 7A), and a simple visual cutoff method can be applied to separate their amplitudes and quantify their volumes.

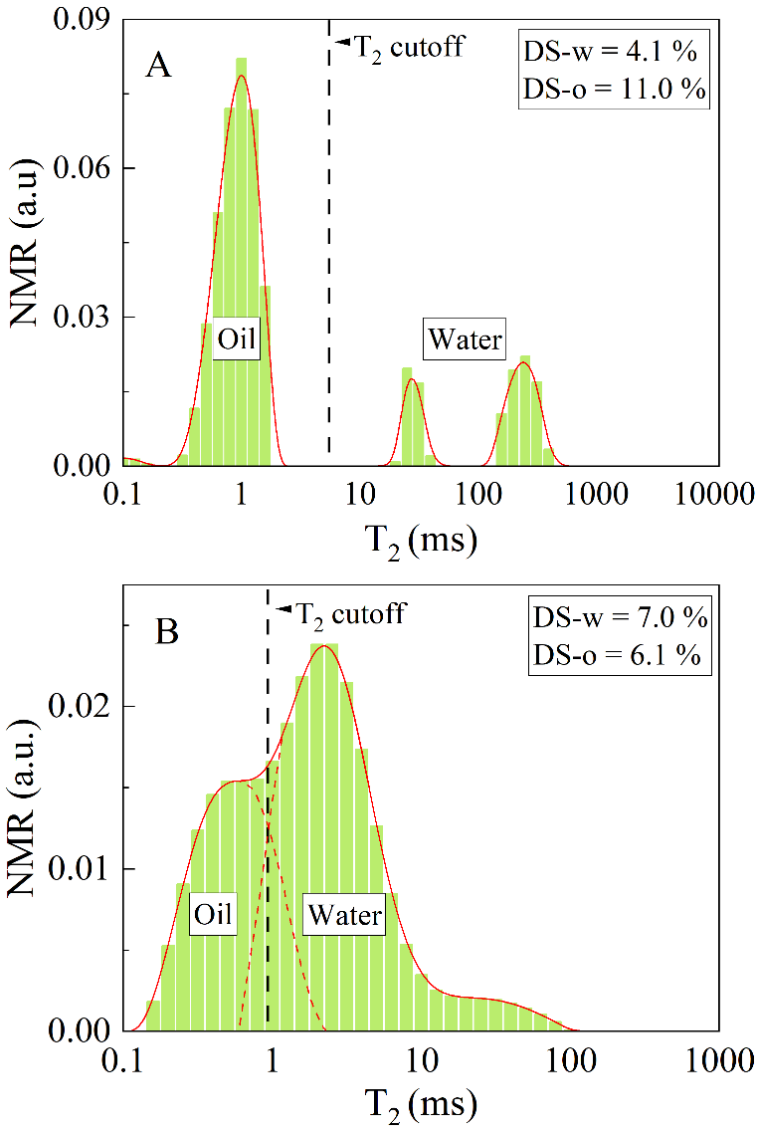


Figure 7: Representative NMR T_2 distributions of two oil-sand samples. (A) An example of distinct oil and water signals where a simple cutoff method can be used for oil-water separation. (B) An example of NMR T_2 distribution with overlapped oil and water signals where deconvolution with T_2 cutoff cannot provide a satisfactory solution. Black vertical dashed lines present potential cutoff times. DS-w and DS-o are percentages of water and oil by Dean-Stark, respectively, relative to solids.

However, in fines and clays, where pores are tiny, the water protons relax faster due to the surface relaxation at the water-rock interface, thus generating the signal in the fast-relaxing part of distribution where it can overlap with the signal originating from heavy oil and bitumen (Figure 7B). In addition to that, the diffusion coupling effect may further decrease the interpretability of the oil and water signals. This effect occurs in saturated and connected micro- and macropores when water is in diffusional exchange, causing the change in the relationship between T_2 relaxation and pore size distribution[62]. In strong diffusive-coupling conditions, macro- and micropore water signals will merge into a single peak, rendering the single T_2 cutoff and deconvolution approach inaccurate[63]. Another limitation of this approach is that it requires the separate determination of water and oil NMR amplitudes and the independent measurement of their volume or mass.

Alternative methods involve 2D LF-NMR measurements, where instead of using one NMR parameter (i.e., T_2 relaxation), additional parameters are employed (i.e., T_1 relaxation or diffusion) to obtain so-called 2D NMR maps[64–66], which can theoretically help to separate these overlapping bitumen and water signals. Application of 2D maps showed considerable success in fluid saturation evaluation, compared to 1D T_2 relaxation distribution analysis, since T_1 relaxation or diffusion of reservoir fluids can be sufficiently different, thus enabling relatively simple separation of their signals. Unfortunately, 2D NMR is slower and more expensive to run, and there can still be instances where these signals are not distinct, in which case estimation of fluid types and fluid volumes can be challenging and require advanced analysis involving blind-source signal

separation (BSS), clustering algorithms, and a certain degree of knowledge in 2D NMR maps interpretation[67].

Chapter 2 Heavy oil viscosity prediction at high temperatures by low-field NMR relaxometry and nonlinear least squares

2.1. MOTIVATION

Evaluation of crude oil viscosity from LF-NMR data has proven to be a viable alternative to laborious and time-consuming conventional measurements requiring sample recovery. However, this work shows that for most heavy oil correlations[10,12–14], accuracy degrades dramatically with the temperature increase due to the loss of magnetization in protons (Curie effect), making them unreliable for continuous viscosity monitoring in oil wells. Another problem commonly seen in practice is vertical and horizontal anisotropy of the oil viscosity within the same heavy oil reservoir and sometimes the same well[3,69], meaning that the NMR model must be robust enough to provide satisfactory predictions within the group of chemically different heavy oil samples.

In this work, we derived a new enhanced NMR viscosity model and three key improvements:

- Enhanced prediction of viscosity for the suite of 23 heavy oils with viscosities ranging from 70–21,600 cP;
- Enhanced prediction of viscosity for the bitumen sample at elevated and high temperatures (26-200 ℃), for viscosity range from 10 to 170,000 cP;
- Nonlinear least squares (NLS) regression procedure for obtaining the optimal fitting parameters of NMR viscosity models (tuning);

Two separate datasets were used in the study. The first consisted of measurements on 23 heavy oil samples recovered from a heavy oil sand reservoir. The second one consisted of measurements made on a single bitumen sample, which is referred to as 'JC bitumen' further in the text. The dynamic viscosities (η) were determined by rheological experiments. The spin-spin relaxation time (T_2-

relaxation) and relative hydrogen index (RHI_v) were measured for all samples and were used as model inputs, while model tuning was achieved by NLS regression. To quantify the effect of tuning on the reduction of prediction errors, we evaluated the performance of models both in their default form (reported fitting parameters) and after tuning by NLS regression (fitting parameters used in this work).

Model performance was evaluated using root mean square error (RMSE), maximum absolute error (MaAE), and adjusted coefficient of determination (adj. R^2).

2.2. THEORY AND EXPERIMENTS

2.2.1. SPIN-SPIN RELAXATION (T_2)

As mentioned before, the oil reservoirs contain various fluids rich in hydrogen. Modern LF-NMR logging devices measure the response of H^+ protons in fluids and provide information about petrophysical properties of rocks and physiochemical properties of fluids *in-situ*. In the case of hydrocarbons, the rate of T_2 relaxation shows a strong correlation with viscosity, which is why T_2 relaxation was used as a theoretical foundation for nearly all NMR viscosity models. Since the relaxation times of oil vary significantly with its chemical composition and temperature, a T_2 logarithmic mean relaxation time is calculated to characterize the whole NMR spectra:

$$T_{2lm}=\mathrm{Exp}\left[\sum \frac{A_i}{A}\cdot \ln(T_{2i})\right] \tag{18}$$

where A_i is the amplitude from i-th corresponding T_{2i} response.

2.2.2. MOLECULAR SIZE AND INTRAMOLECULAR DISTANCE

Dynamic viscosity correlation with T_2-relaxation time can be inferred from Bloembergen-Purcell-Pound's (BPP) model, describing T_1 and T_2-relaxation rate dependency with dipole-dipolar interaction[19]:

$$\frac{1}{T_1} = C\left(\frac{\tau_c}{1+\omega_0^2\,\tau_c^2} + \frac{4\tau_c}{1+4\omega_0^2\,\tau_c^2}\right) \tag{19}$$

$$\frac{1}{T_2} = C\left(\frac{3}{2}\tau_c + \frac{5}{2}\,\frac{\tau_c}{1+\omega_0^2\,\tau_c^2} + \frac{4\tau_c}{1+4\omega_0^2\,\tau_c^2}\right) \tag{20}$$

$$C = \frac{3}{10}\left(\frac{\mu_0}{4\pi}\right)^2 \frac{\hbar^2\gamma^4}{b^6} \tag{21}$$

where τ_c is molecule correlation time, and ω_0 is a Larmor frequency. The parameter C is defined for the ½ spin by gyromagnetic ratio γ, the magnetic permittivity of space μ_0, reduced Planck constant $\hbar$, and interproton distance b. Molecular collisions lead to a time-dependent change in molecule orientation and interproton distances. From Equation 19 and Equation 20, it is evident that proton relaxation rates are primarily influenced by correlation time for the liquid substance. Random change of the molecule orientation can be described by a rotational diffusivity D_r, a function of viscosity, temperature, and molecular size. By employing a Debye-Stokes-Einstein (DSE) model for spherical molecules, we can express τ_c as

$$\tau_c = \frac{1}{6D_r} \tag{22}$$

$$D_r = \frac{kT}{8\pi\eta a^3} \tag{23}$$

where k is Boltzmann constant, a is the radius of the spherical particle and η is the dynamic viscosity of the medium. The BPP model and DSE equations were developed for pure homogeneous substances, while crude and heavy oils have a complex chemical composition and molecular structures that contain multiple bonds, chains, solid asphaltene agglomerates, and clusters. Consequently, we can anticipate variability in molecule sizes and interproton distances, which causes fluctuation of parameters a and b (Equation 23 and Equation 21). In that sense, for

any universal NMR viscosity model, discrepancies in predictive ability will grow with the complexity of the chemical composition [24,32,70].

2.2.3. T_2-RELAXATION MECHANISMS IN HEAVY OILS

Recall that in the porous medium the total T_2 relaxation represents the sum of three relaxation components (Equation 5). The total spin-spin relaxation is governed by a sum of bulk relaxation T_{2bulk}, relaxation influenced by the pore surface $T_{2surface,}$ and relaxation caused by the gradients of magnetic field $T_{2diffusion}$. In the case of water-wet porous media, surface relaxation is dominant for water and bulk relaxation for oil. Since the bulk relaxation remains a primary mechanism in such scenario, it can be considered that for heavy oil T2 ≅ T2bulk.

According to the nuclear spin relaxation theory[71], there are two extreme cases:

1. Fast motion or extreme narrowing case ($\omega_0\tau_c \ll 1$), characteristic for small molecules, low viscosities, or high temperatures. In such cases, $T_1 \approx T_2$ (Figure 8).
2. Slow-motion case ($\omega_0\tau_c \gg 1$) characteristic for relaxation of large molecules in high viscous substances or at low temperatures.

In both cases, the $1/T_2$ relaxation rate is proportional to correlation time τ_c and η/T ratio, respectively.

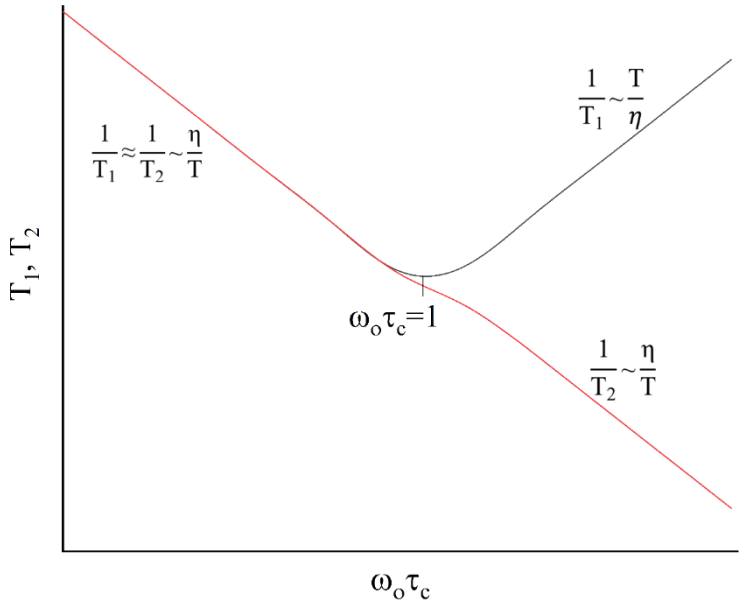


Figure 8: T_1 (black) and T_2 (red) dependence on correlation time (τ_c), according to the BPP relaxation model.

Equations 19-23 describe the molecular relaxation processes governed by a single exponential function which explains why the viscosity models for short correlation times (Figure 8) provide sufficiently accurate predictions for light oils[15,23,27,29]. Solid-like components in heavy oils induce various relaxation rates as opposed to light oils, and the cumulative spin-echo decay can exhibit non-exponential behavior. This behavior can be approximated by the stretched-exponential function, also known as the Kohlrusch-Williams-Watts function[33]:

$$G(\tau)=\langle F(0)^2\rangle e^{-(\tau/\tau_c)^{\gamma}} \quad (24)$$

Function *F(t)* is a time-dependent function of molecule location and orientation. *G(τ)* describes a relationship between function F(t) in different time steps, and $(\tau/\tau_c)^{\gamma}$ is a stretch parameter. Equation 24 does not have an analytical Fourier transform, but a modified Cole-Davidson function approach[31,72] can be used instead.

$$T_2 \sim \left(\frac{\eta}{T}\right)^{-\beta} \quad (25)$$

where $\beta \sim (\tau/\tau_c)^{\gamma}$ and is $0 \leq \beta \leq 1$. The detailed derivation of Equation 25 is described in detail by Cheng et al [33]. This approach was confirmed to be effective by several

authors[3,25,31,73]. It should be noted that T_1 dependence on correlation time also does not follow the classical BPP model for heavy oils and bitumen in the slow-motion case, which was experimentally proved by several authors[25,60], most recently by Singer et al.[35]. In Figure 8, instead of the anticipated increase, T_1 plateaus to 3 ms value on a frequency-normalized scale for various viscosity samples. The authors explained the observed plateau effect by combining the dipole-dipole interaction model for intramolecular interactions and the modified Lipari-Szabo model for internal motions of the non-rigid structure.

2.2.4. ECHO SPACING (TE) AND RELATIVE HYDROGEN INDEX (RHI_V)

Due to the presence of solid-like components in heavy oils (e.g., paraffin and asphaltene), the T_2-relaxation is often so fast that many LF-NMR logging tools cannot measure the whole relaxation spectrum the sample[25,31,74]. The parameter that expresses the NMR tool's signal sampling rate is known as echo spacing (TE), where TE ≥ 0.1 ms. Consequently, for heavy oils with a very short mean T_2 relaxation time (T_{2lm}), the logging devices cannot capture the fast-relaxing part of the NMR T_2 distribution. The result is that the tools fail to accurately reflect the actual number of hydrogen atoms (HI) and output HI that is too small since part of the fast-relaxing signal is not measured. Many authors tried to address this issue by adjusting TE using correction coefficient or integrating some form of hydrogen index into the model [18,30,31,68,74,75].

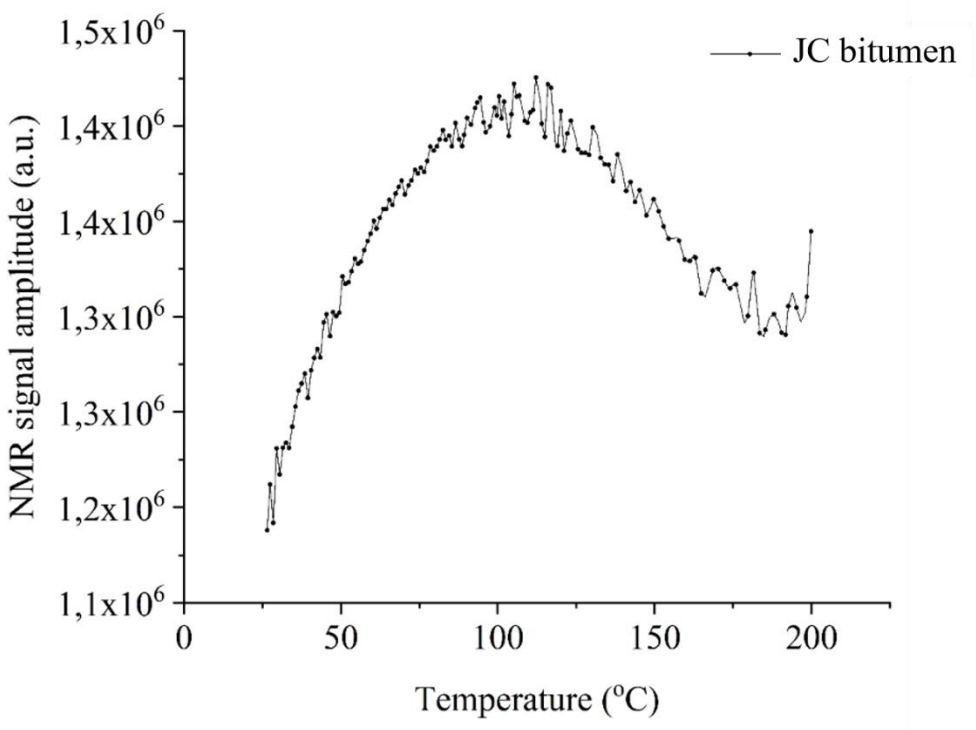


Figure 9: NMR signal amplitude of a single bitumen sample (JC bitumen) in the function of the temperature. The slope of the NMR signal decreases with temperature rise, and approximately at >100 °C, the slope becomes negative due to the Curie effect.

From Figure 9, it can be observed that at temperatures above 100 °C, the NMR amplitude decreases. This is known as a Curie effect[60,76], where magnetization loss (NMR signal loss) occurs due to the high temperature of the sample. One of the means to account for this loss is the implementation of the relative hydrogen index (RHI). The RHI represents the relative amount of measured (detectable) hydrogen protons by the NMR device in the oil sample. It is expressed as the ratio of oil and water NMR signal amplitudes per unit mass[77]. In the case of using NMR tools at elevated temperatures, it is compulsory to implement temperature correction for RHI to account for the magnetization loss:

$$\mathrm{RHI}=\left(\frac{\mathrm{A_0 m_w}}{\mathrm{A_w m_0}}\right)\left(\frac{\mathrm{T(°K)}}{\mathrm{T_{amb}(°K)}}\right) \tag{26}$$

where A_o and A_w are the amplitudes of the oil and water signal respectively, m_o and m_w are masses of oil and water respectively, and T/T_{amb} is a temperature correction term in kelvins. Also, note that T is a sample temperature while T_{amb} is

the temperature at which water sample was measured (ambient). If the RHI is normalized to the sample volume, a relative hydrogen index for a defined sample volume can be obtained (RHI_v). This normalization is consistent with Curie's expression for magnetic susceptibility, where magnetization is expressed per unit volume[75]. Also, RHI_v is more suitable for application to well logs because the NMR tool detects a defined volume of the formation[68]. Burcaw et al. proposed a simple approach for conversion of RHI to HI[68]:

$$RHI_v = \frac{\rho_o}{\rho_w} \cdot RHI \tag{27}$$

where ρ_o and ρ_w are densities of oil and water, respectively. Note that Equation 27 should be valid under the assumption that sample volume change due to the temperature change is negligible. It should also be noted that Equation 27 represents the hydrogen proton response detected by the NMR device, and it is not to be associated with a true HI of the sample.

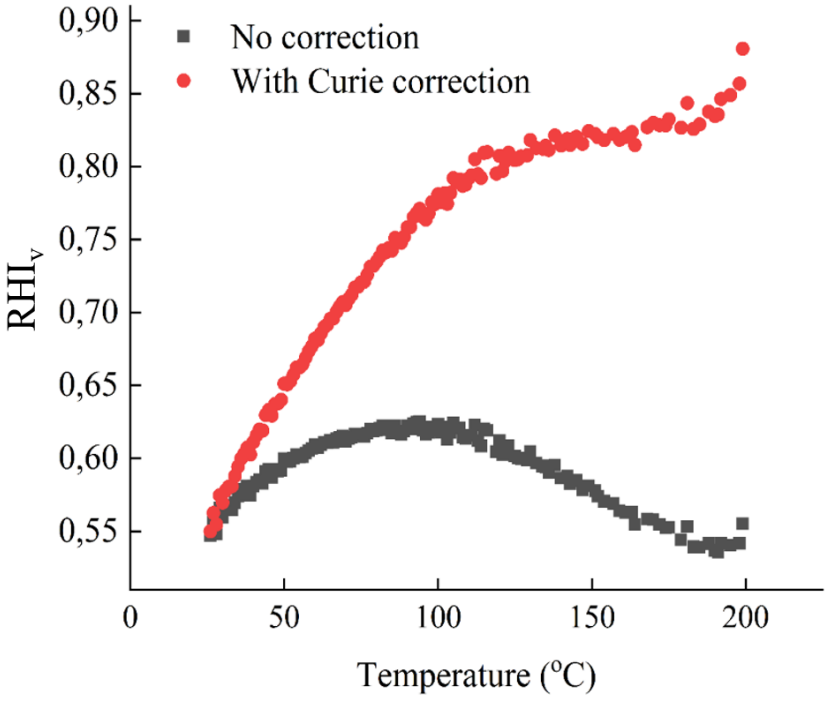


Figure 10: Relative hydrogen index for a defined volume of JC bitumen sample in the function of temperature with implemented correction (red) and without correction for the Curie effect (black).

HI reflects the number of H protons in a liquid, which is a finite value. However, in Figure 10, it is evident that even with Curie correction, RHI_v for JC bitumen changes with the temperature. This is due to the hardware limitation of the LF-NMR tools and in the presence of solid-like components (e.g., asphaltenes), which relax faster than an echo spacing (TE), which essentially means that for heavy oils and bitumens, a significant part of oil signal (i.e., H protons) remains invisible in the T_2 distribution[28,76,78].

2.2.5. ENHANCED NMR VISCOSITY MODEL

The RHI_v works as a correction factor by compensating the magnetization loss (Curie effect) at high temperatures. The RHI_v also accounts to a degree for the long TE, meaning that the correction coefficient for the TE term is redundant. For longer correlation times characteristic for very viscous oils ($\omega_o\tau_c>>1$), the right addend in Equation 28 contains the stretching parameter *d* (or β in Equation 25), which accounts for non-exponential relaxation, i.e., the power-law relationship with T_{2lm}. Lastly, the enhanced model contains a T_{2lm} term, inversely proportional to viscosity, which properly works for light oils. This model is expected to mitigate the following known pitfalls in NMR viscosity prediction:

- Magnetization loss due to the high temperature (Curie effect).
- Long echo times (TE) which hinder the detection of solid-like components in heavy oils.
- Non-exponential relaxation (i.e., glass transition) in viscous oils where $\omega_o\tau_c>>1$.

Taking into account Equations 19-27, an analytical form of the enhanced NMR viscosity prediction model can be derived as:

$$\eta = \frac{a}{RHI_v^{\,b}\, T_{2lm}} + c\, T_{2lm}^{-d} \qquad (28)$$

where *a, b, c*, and *d* are obtained from the nonlinear least squares (NLS) regression. The left-hand addend of Equation 28 is adapted from the Bryan et al. model[18],

which correlates measured hydrogen content and T_2 logarithmic mean with oil viscosity.

2.2.6. PREPARATION OF OIL SAMPLES

Twenty-three heavy oils were analyzed in this study. All the samples were recovered from different oil formations and wells from the heavy oil reservoirs in Alberta, Canada. All the samples except one were previously used in another NMR study by Bryan et al.,[79], and the same methodology was used for sample preparation as described previously. Oil samples were extracted from the core samples by spinning in the Ultracentrifuge Beckman 18-M at 15,000 rpm at 40 °C for approximately 60 minutes. After extraction, the water content levels were reduced below 1.0 wt% using decantation with gravitation for one hour.

The JC bitumen selected for the high-temperature tests had a residual emulsified water content of the oil determined using the Dean-Stark distillation method. The emulsified water and solid impurities were removed from the oil through a proprietary oil-cleaning system developed by the *In-Situ* Combustion Research group at the University of Calgary. Following the cleaning procedure, the emulsified water content was 0.78 wt%, and there was no dissolved gas (dead oil).

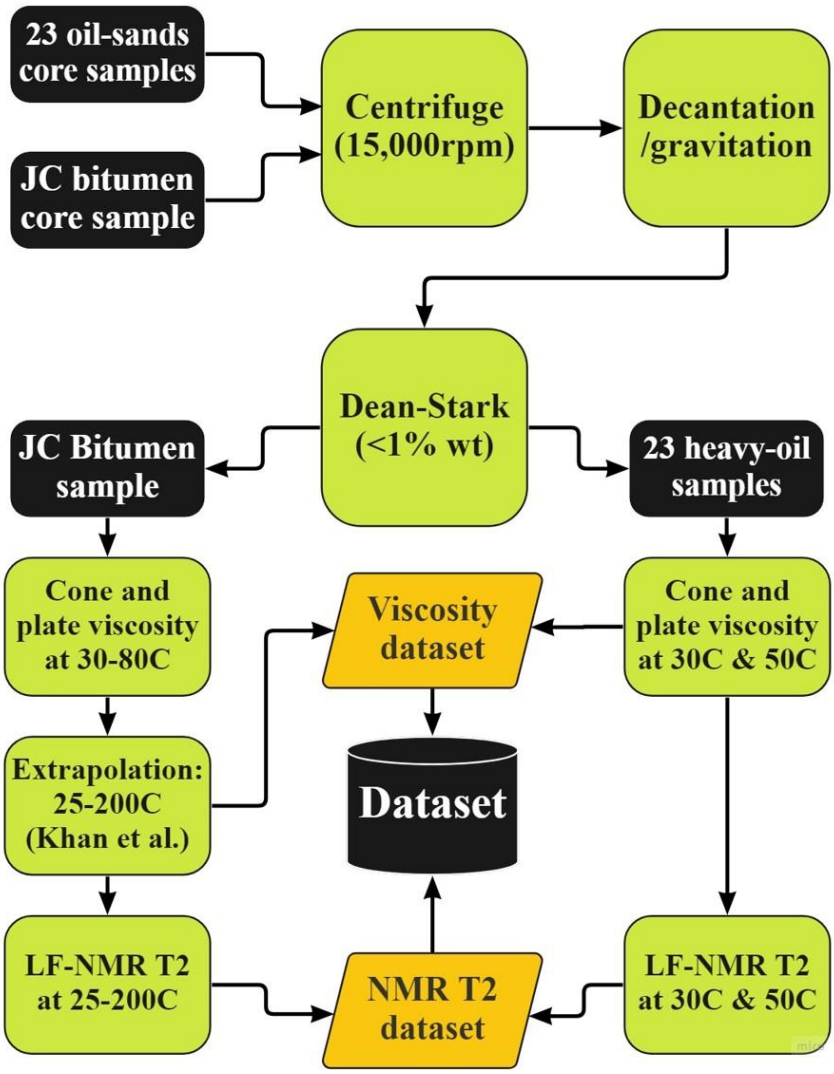


Figure 11: Flowchart representation of experimental program for 23 heavy oil samples and JC bitumen sample

2.2.7. RHEOLOGICAL MEASUREMENTS – 23 HEAVY OIL SAMPLES

Rheological measurements were executed on 23 heavy oils to obtain a reference dataset compared with predictions from NMR viscosity models. Previous studies have shown that cone and plate rotational viscometers provide higher accuracy in measuring viscous fluids than glass-capillary and oscillating-piston viscometers[79]. The Brookfield DV-II-Pro cone and plate viscometer was used, which meets ASTM D4287 industry standard for oil viscosity measurement[80]. As previously described by Bryan et al.[32], the cone diameter was 12 mm with an angle of 1.5°. Shear rate

accelerated from 0.8 s^{-1} to 100 s^{-1} while shear stress was continuously logged. Three milliliters (3 ml) of oil were used for each experiment at two fixed temperatures – 30 °C and 50 °C. By measuring at two temperatures rather than one, we increased the number of data points. However, four heavy oils had a limited supply, and their measurements were taken only at 50 °C, making 42 data points in total. The viscosity was expressed as a ratio between shear stress and shear rate.

2.2.8. RHEOLOGICAL MEASUREMENTS AT HIGH TEMPERATURE– JC BITUMEN

The most viscous heavy oil sample (JC bitumen) was selected for assembling the high-temperature viscosity dataset. Since Brookfield DV-II-Pro viscometer is equipped with a thermal bath, measurements were made from 30 °C to 80 °C on every 5 °C making 11 data points. For the reliability of the measurements, the experiment was repeated three times at each temperature. Extrapolation of viscosities to from 26 ℃ to 200 ℃ was carried out using the dynamic viscosity model for gas-free Athabasca bitumens defined by Khan et al.[20]:

$$\ln \ln (\eta) = A \cdot \ln(T_{abs}) + B \qquad (29)$$

where A and B are empirical constants calculated as a slope and intercept of absolute temperature and measured dynamic viscosity, respectively. The relationship between JC bitumen viscosity and temperature is depicted in Figure 12.

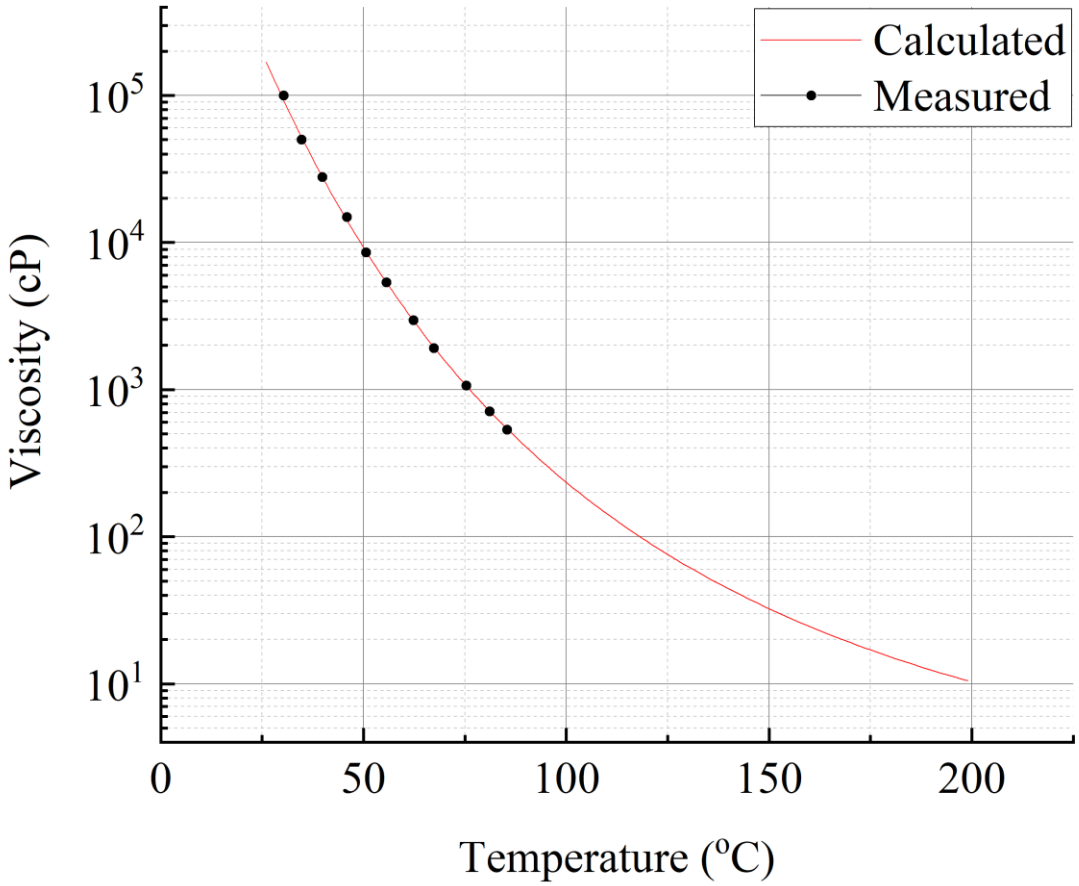


Figure 12: Dynamic viscosity of JC bitumen in 26 °C – 200 °C temperature range. Extrapolation and interpolation was performed using a model by Khan et al.[20]

2.2.9. NMR EXPERIMENTS – 23 HEAVY OIL SAMPLES

The suite of 23 samples was tested as a part of the previous two studies by Bryan et al.,[32,77]. The NMR experiments were carried out at 30 °C and 50 °C using a 1.1-MHz LF-NMR Corespec 1000™ relaxometer. The Carr-Meiboom-Purcell-Gill (CPMG) pulse sequence parameters were tuned to reduce the effect of the temperature decrease within a single experiment – TE was 0.3 ms, with 2,600 pulses and a wait time 2,400 ms. Measured data were transformed into T_2 relaxation distribution in the time domain using NNLS inversion software ExpFit[77]. To be consistent with the rheological viscosity dataset, the two samples left out from viscosity measurements at 30 °C were also left out from NMR measurements at 30 °C. In total, 42 data points were obtained.

2.2.10. NMR EXPERIMENTS AT HIGH TEMPERATURE – JC BITUMEN

NMR experiments on the JC bitumen were carried out at the temperature range from 26 °C to 200 °C using a 2.66-MHz LF-NMR PERM Labmeter. The sample was stored in a polyether-ether-ketone (PEEK) thermoplastic polymer vessel with an integrated non-magnetic thermocouple for continuous temperature logging. The vessel with the oil was heated in the oven up to 200 °C and then inserted into the NMR device. The highest cooling temperature gradient occurred at 170-200 ^{o}C (~0.7 ^{o}C/min). Since heavy oils in this study had very fast relaxation at ambient conditions (1-2 ms), the CPMG sequence had to be configured to capture the earliest signals at ambient temperatures, while capturing the oil signal at 200 °C with high SNR. To achieve this, the shortest echo spacing limited by the equipment was selected (TE=0.24 ms). The number of pulses was 5,000, and the wait time was 5,000 ms, which enabled sufficiently quick experiments (<1 minute per experiment) with SNR>100. Experiments were performed consecutively until the sample reached ambient temperature. A total of 136 data points were used for the regression analysis.

It should be noted that PEEK plastic can produce an NMR signal in some instances, and its contribution depends on the shape of the vessel, whether it was extruded or molded, and whether PEEK contains impurities. In this study, the PEEK signal can be observed on NMR spectra, between T_2 ~ 10-20 ms, representing <1% of the total signal produced by the bitumen sample, which is negligible. Similar reports can be found in the literature[81].

2.2.11. NONLINEAR LEAST SQUARES (NLS) REGRESSION – MODEL TUNING

The NLS regression was performed in Origin Pro software version 2018b to tune the NMR models, that is, to obtain optimal values for their empirical constants (parameters). Where it was possible, the data population were split into the calibration set and the prediction set in the proportion of 70-30%, respectively, minimizing the overfitting. A calibration set was used to tune the model

parameters and subsequently applied to the prediction set. The model calibration was evaluated by comparing predicted NMR viscosity with rheological viscosity.

The NLS regression was performed using Levenberg – Marquardt (M-L) iteration algorithm[82].

$$\hat{\beta}\ argmin_{\beta}\ S(\beta) \equiv\ argmin_{\beta} \sum_{i=1}^{m} [y_i - f(x_i, \beta)^2] \qquad (30)$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_i)$ are fitting parameters to be obtained from the minimization of the sum of the squared residuals $S(\beta)$ from fitted model $f(x_i, \beta)$ for the given set of independent variables (x_i) and target output (y_i). This is a step-wise (iterative) approach, where initial parameter values are set manually. To avoid convergence to a local minimum, different initial parameters were used and constrained to a specific range in some instances. For each new iteration, the parameter vector β is updated by the new estimate $\beta + \delta$, where δ can be linearly approximated from function $f(x_i, \beta + \delta)$ as:

$$f(x_i, \beta + \delta) \approx f(x_i, \beta) + J_i\delta \qquad (31)$$

where J_i is a gradient of function f with respect to parameter vector β:

$$J_i = \frac{\partial f(x_i, \beta)}{\partial \beta} \qquad (32)$$

Like in Tikhonov regularization, a damping factor λ is added to regulate the reduction rate of S and for more efficient discovery of a gradient direction. For the initialization of L-M, the λ was set to 0.001, and after each successive iteration was automatically increased or decreased by a factor of 10 relative to the gradient direction, depending on whether the squared residuals were reduced or increased. The L-M algorithm converges when the sum of squared residuals remains unchanged relative to the set tolerance or is equal to zero. The tolerance parameter was reduced chi-square (χ^2), which can be calculated by dividing a sum of squared residuals (RSS) by degrees of freedom. The tolerance was set to $\chi^2 < 1 \cdot 10^{-9}$.

The results of NLS tuning were assessed using root mean square error (RMSE). (Equation 33). RMSE is a useful statistic for evaluating model prediction accuracy based on the new data.

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(P_i-O_i)^2}{n}} \tag{33}$$

where P_i is the predicted value, O_i is the observed value, and n is the number of samples.

In the existing literature, the NMR model accuracy is usually assessed visually on cross-plots relative to the prediction bands and by comparing coefficients of determination - R^2 [3,30,33,73,74,77]. In this study, an adjusted coefficient of determination and standard coefficient of determination is used for the evaluation of prediction variation captured by the model

$$\bar{R}^2=1-(1-R^2)\cdot\frac{n-1}{n-p-1} \tag{34}$$

$$R^2=1-\frac{SS_{res}}{SS_{tot}} \tag{35}$$

where $\bar{R}^2$ is the adjusted coefficient of determination (COD), n is the number of observations, and p is the number of independent variables (inputs). The standard COD (R^2) is calculated as a difference between the unity and ratio of the sum of squared prediction residuals (SS_{res}) and the total sum of squared residuals (SS_{tot}). Since viscosity may vary up to six orders of magnitude in thermal EOR projects, besides using RMSE and R^2, the maximum absolute error (MaAE) of the predictions was calculated as an additional statistical metric. The MaAE represents the maximum absolute difference between predicted and observed viscosity values. RMSE and MaAE are negatively-oriented statistics expressed in source units (i.e., centipoises.)

2.3. RESULTS AND DISCUSSION

2.3.1. NMR VISCOSITY PREDICTION – 23 VARIOUS HEAVY OIL SAMPLES

The models tested in this study are listed in Table 1, describing input parameters and the number of fitting (free) parameters.

Table 1: Tested literature NMR viscosity correlations

Model	Input data	Fitting parameters
Straley, 1997	T_{2lm}	2
LaTorraca, 1999	T_{2lm}, TE, T	2
Bryan, 2003	T_{2lm}, RHI	2
Nicot, 2007	T_{2lm}, proton radius (a), inter proton distance (b)	1
Burcaw, 2008	T_{2lm}, HI	3
Cheng, 2009	T_{2lm}	3
Ahmed, 2014	T_{2lm}, TE, T	2
Musin, 2016	T_{2lm}	3
Sandor, 2016	T_{2lm}, TE, T	2
Markovic, 2019	T_{2lm}, RHI_v	4

Figure 13 shows NMR viscosity predictions and observed viscosity for a suite of 23 heavy oil samples at 30 °C and 50 °C with viscosities ranging from 70 – 21,600 cP. The NMR viscosity predictions generated by nine well-known literature models by Ahmed et al.[30], LaTorraca et al.[74], Sandor et al.[31], Bryan et al.[18], Burcaw et al.[68], Cheng et al.[33], Nicot et al.[25], Straley et al.[23], and Musin et al.[3]. Predictions made by the enhanced model are shown in Figure 13j. To emphasize the effect of NLS regression, predictions were produced with NLS tuned parameters (red) and with default model form, where used parameter values were reported by the authors (black). The analytical form of models with fitting parameters obtained by nonlinear least squares (NLS) regression is depicted in the lower right corners.

The NLS regression was applied without splitting the data into the calibration and prediction sets due to the small number of data points (42 in total). Since the oils were sampled from different locations, independent variables (T_{2lm} and RHI_v) show high variability compared to JC bitumen dataset. This is also reflected in varied adj. R^2 scores (Figure 14a), indicating that models captured different amounts of variability related to the response variable (i.e., viscosity). Compared statistical scores are shown in Figure 14.

Figure 13 shows that models by Burcaw et al., and Musin et al., do not contain predictions from the default configuration (black points) because their authors did not propose parameter values explicitly. Moreover, Figure 13 shows that generally, model predictions improve after NLS regression compared to the predictions generated by general model forms to various extent. However, in Figure 14, the effect of tuning and variation in prediction accuracy between models is clearly illustrated. For example, Bryan et al. show RMSE and MaAE scores to be 4.5 and 8 times lower after tuning, respectively, while adj. R^2 score is marginally increased.

Heavy oil models by LaTorraca et al., Ahmed et al., and Sandor et al. show a similar performance since they are based on TE correction and temperature. Sandor et al., and Ahmed et al., proposed multiple models in their work, but for this work were selected the ones with the highest reported score. After NLS regression, the model by Sandor et al. (Figure 13c) shows that most predictions fall within a factor of one and two at the viscosity range between 30 and 3,000 cP. However, at viscosities >3,000 cP, predictions scatter into a factor of two and three, causing the inflation of RMSE and MaAE scores. (Figure 14). The models by LaTorraca et al. and Ahmed et al. have less variance in >3,000 cP domain after NLS regression but tend to overpredict viscosity in the <6,000 cP domain, with most predictions falling within a factor of two and three. This is due to the cost function minimization, where the iteration algorithm minimizes larger squared errors in the higher viscosity domain at the expense of accuracy in the <6,000 cP domain.

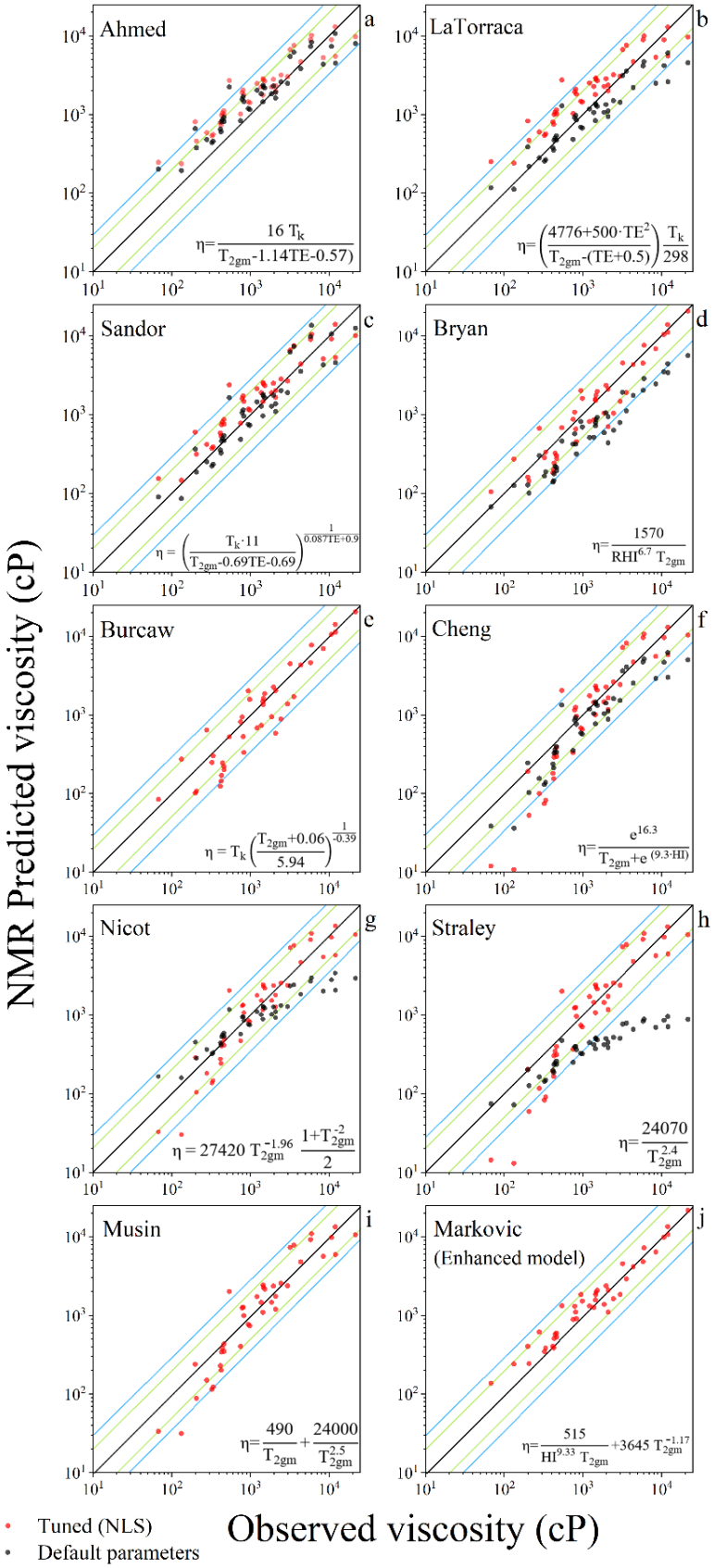

Ahmed
a
$\eta=\frac{16\ T_k}{T_{2gm}-1.14TE-0.57)}$
LaTorraca
b
$\eta=\left(\frac{4776+500\cdot TE^2}{T_{2gm}-(TE+0.5)}\right)\frac{T_k}{298}$
Sandor
c
$\eta=\left(\frac{T_k\cdot 11}{T_{2gm}-0.69TE-0.69}\right)^{\frac{1}{0.087TE+0.9}}$
Bryan
d
$\eta=\frac{1570}{RHI^{6.7}\ T_{2gm}}$
Burcaw
e
$\eta=T_k\left(\frac{T_{2gm}+0.06}{5.94}\right)^{\frac{1}{-0.39}}$
Cheng
f
$\eta=\frac{e^{16.3}}{T_{2gm}+e^{(9.3\cdot HI)}}$
Nicot
g
$\eta=27420\ T_{2gm}^{-1.96}\ \frac{1+T_{2gm}^{-2}}{2}$
Straley
h
$\eta=\frac{24070}{T_{2gm}^{2.4}}$
Musin
i
$\eta=\frac{490}{T_{2gm}}+\frac{24000}{T_{2gm}^{2.5}}$
Markovic
(Enhanced model)
j
$\eta=\frac{515}{HI^{9.33}\ T_{2gm}}+3645\ T_{2gm}^{-1.17}$
NMR Predicted viscosity (cP)
Observed viscosity (cP)
Tuned (NLS)
Default parameters

Figure 13: Rheological viscosities compared to NMR viscosities of 23 heavy oils. The Markovic et al. (j) model demonstrates the highest accuracy. A solid black line (x=y) presents a perfect prediction.

The models by Nicot et al., and Cheng et al. are heavy oil models, while Straley et al. is a light oil model. All three are based on the power-law T_{2lm}, which accounts for long TE (echo spacing) and have similar performance. Figure 13 and Figure 14 show that the models are affected by NLS to a different extent but generally achieve considerably improved scores after NLS. Most of the predictions in the >600 cP domain fall within a factor of one and two. Recall that Nicot et al., Cheng et al., and Straley et al. models produce large prediction residuals in the domain <600 cP after NLS regression. Again, this is caused by minimizing squared errors in the higher viscosity domain at the expense of accuracy in the lower viscosity domain. This issue can be solved either by constraining fitting parameters to a specific range, at the expense of the accuracy at higher viscosities, or by using these models for oilfields or wells where the viscosity varies no more than three orders of magnitude.

Figure 14 shows the segregation of the three models (Markovic et al., Bryan et al., and Burcaw et al.), which perform substantially better than the remaining seven based on all three statistics. Predictions by Markovic et al. fall within a factor of one and two along with the whole range, with only four exceptions in the <600 cP domain. The most accurate predictions are distributed in the >1,000 cP domain. Also, Bryan et al. and Burcaw et al. perform almost on par with the enhanced model. The new model has a marginally better score than the latter two. This marginal improvement is due to the power-law parameter, which was not considered in the models by Bryan et al., and Burcaw et al.

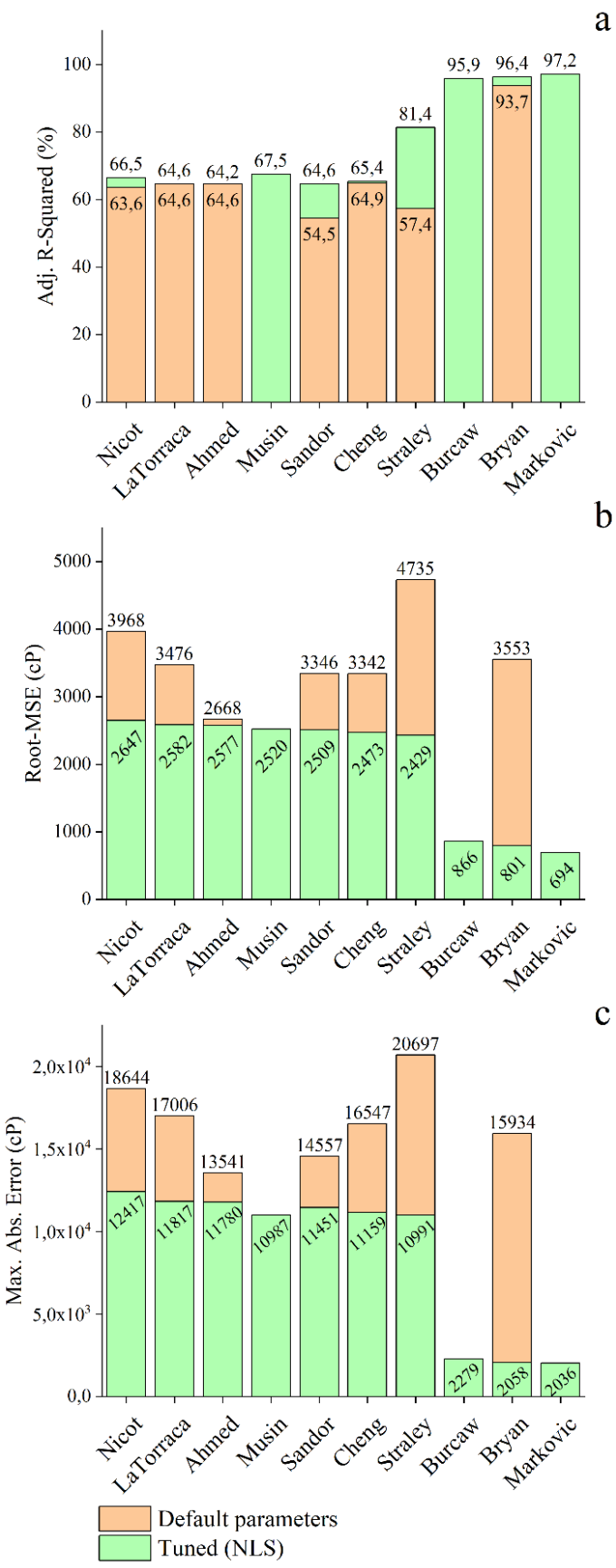

a
Adj. R-Squared (%)
100
80
60
40
20
0
66,5
63,6
64,6
64,6
64,2
64,6
67,5
64,6
54,5
65,4
64,9
81,4
57,4
95,9
96,4
93,7
97,2
Nicot
LaTorraca
Ahmed
Musin
Sandor
Cheng
Straley
Burcaw
Bryan
Markovic
b
Root-MSE (cP)
5000
4000
3000
2000
1000
0
3968
2647
3476
2582
2668
2577
2520
3346
2509
3342
2473
4735
2429
866
3553
801
694
Nicot
LaTorraca
Ahmed
Musin
Sandor
Cheng
Straley
Burcaw
Bryan
Markovic
c
Max. Abs. Error (cP)
2,0x10^4
1,5x10^4
1,0x10^4
5,0x10^3
0,0
18644
12417
17006
11817
13541
11780
10987
14557
11451
16547
11159
20697
10991
2279
15934
2058
2036
Nicot
LaTorraca
Ahmed
Musin
Sandor
Cheng
Straley
Burcaw
Bryan
Markovic
Default parameters
Tuned (NLS)

Figure 14: Compared bar chart of adjusted R^2 (a), Root-MSE (b), and MaAE (c) for tuned and default NMR model predictions of 23 heavy oils. Markovic et al. model demonstrate the highest accuracy.

These results demonstrate that the integration of RHI_v, into the correlation, substantially improves the prediction accuracy of heavy oil models in the 70–21,600 cP viscosity range at 30 °C and 50 °C, especially in the >3,000 cP domain, as demonstrated by correlations from Bryan et al., Burcaw et al., and the newly proposed model from Markovic et al. (Figure 13d, Figure 13e, Figure 13j). This finding complements the adj. R^2 scores in Figure 14a show that these three models have the highest explained variability (>96 %). In addition to the RHI_v, the enhanced model contains the power-law term in T_{2lm}, which marginally improves prediction capacity by rectifying the non-exponential relaxation effect of heavy components, for which T_{2lm} and measured RHI_v cannot account.

In conclusion, the new model (Markovic et al.) achieved the most favorable statistical scores, while the models by Sandor et al., Bryan et al., Burcaw et al., and Musin et al. have satisfactory performance only after the NLS regression. The prediction capacity of the models in high temperatures is discussed in the following section.

2.3.2. NMR VISCOSITY PREDICTION AT HIGH TEMPERATURES – JC BITUMEN

To validate the enhanced NMR viscosity model for use in steam EOR projects, it was necessary to examine how its prediction capacity is affected by the temperature increase, by testing it on a JC bitumen sample, with a viscosity range of 10–170,000 cP, for the temperature range 26–200 °C. Apart from the enhanced model, five literature correlations were selected to compare based on their performance in the previous section. These are Straley et al., Cheng et al., Bryan et al., Burcaw et al., and Sandor et al. To avoid overfitting, the JC bitumen dataset was divided into the training set the test set in proportions of 70–30%, respectively.

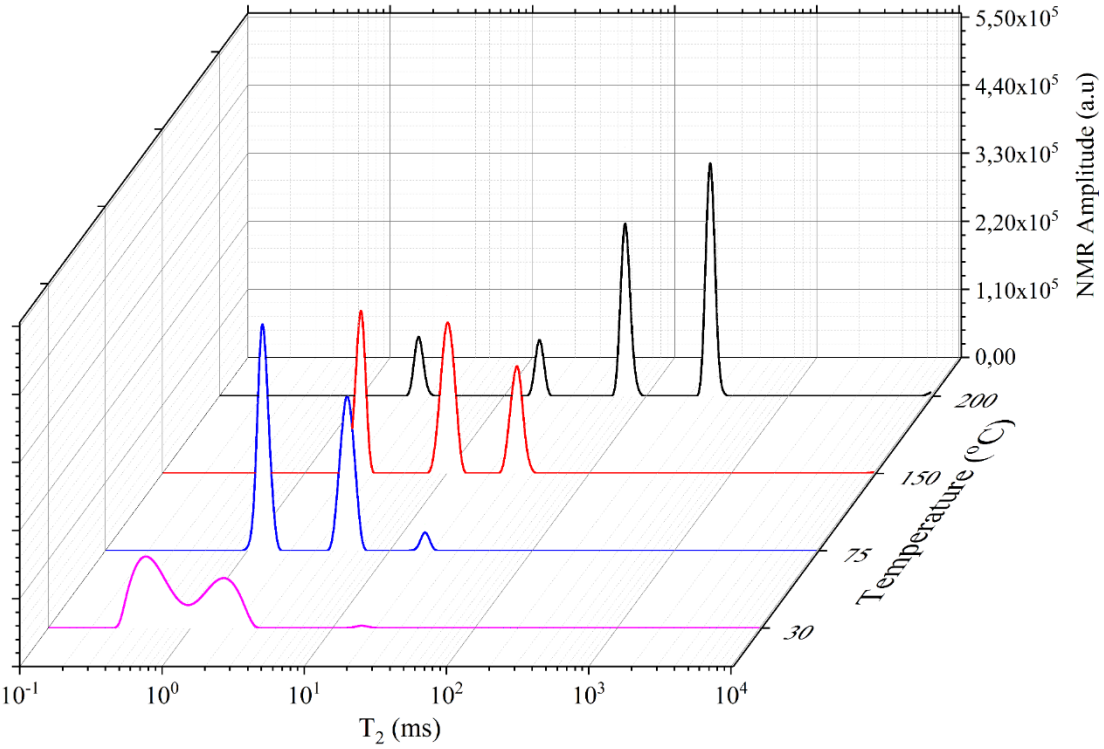


Figure 15: T_2 distribution curves of a single bitumen sample (JC bitumen) in the function of temperature.

Figure 15 presents the T_2 spectrum of JC bitumen, in the time domain, obtained using Tikhonov regularization[18]. The NMR relaxation spectra shift to the right-hand side (slow-relaxing part) with increasing temperature, and the NMR signal amplitude varies with temperature. The distribution curve at 200 °C shows four distinct peaks, which might indicate oil separation to several relaxometry components, possibly heavy and light fractions of the bitumen[83].

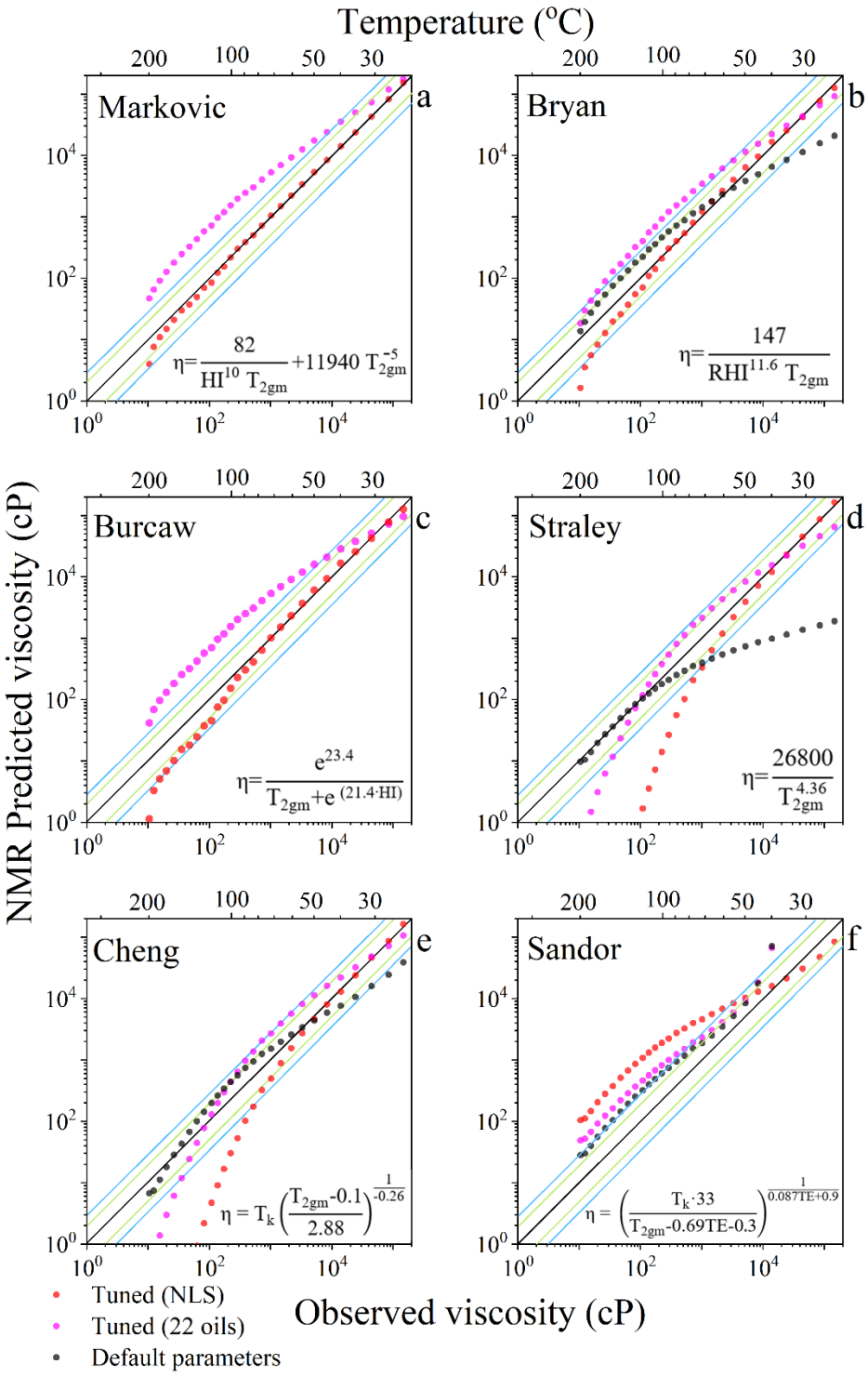


Figure 16: Rheological viscosities compared to NMR viscosities for the JC bitumen dataset. The temperature scale (top axis) is shown for clarity. A solid black line (x=y) presents a perfect prediction.

Figure 16 compares predicted and observed viscosity over the 26–200 °C temperature range. The predictions were generated from three model configurations: using default parameter values (black), using parameters from the previous section (magenta), and applying the NLS regression to the JC bitumen training set to calculate the new parameter values (red). The models in the analytical form after NLS regression are depicted in the lower right-hand corners of the plots.

Figure 17 shows three statistical scores used to compare the model forecasting performance. Note that the y-axis in Figure 17a is truncated for convenience. The first observation comes from Figure 17a, where high adj. R^2 scores indicate a low variability of response data (i.e., viscosity predictions), which was expected since only one sample was analyzed. However, the RMSE scores in Figure 17b and MaAE scores in Figure 17c show that NMR models achieve substantially different scores. In this way, using the adj. R^2 alone for the model performance evaluation is not sufficient.

The majority of the predictions by the Sandor et al. model (Figure 16f) fall outside the prediction bands for all three configurations. After NLS regression, the model improves accuracy in the >10,000 cP domain. It should be noted that Sandor et al., for the given dataset, cannot be used with default fitting parameters because the denominator in this correlation becomes negative for the high values of the T_{2lm}, (i.e., high viscosities). This limitation was addressed by changing the constant in the denominator from -0.69 to -0.3 for NLS tuned predictions. Due to the absence of predictions for the mentioned interval, Figure 17 only contains statistics of Sandor et al. correlation for the predictions generated after NLS regression.

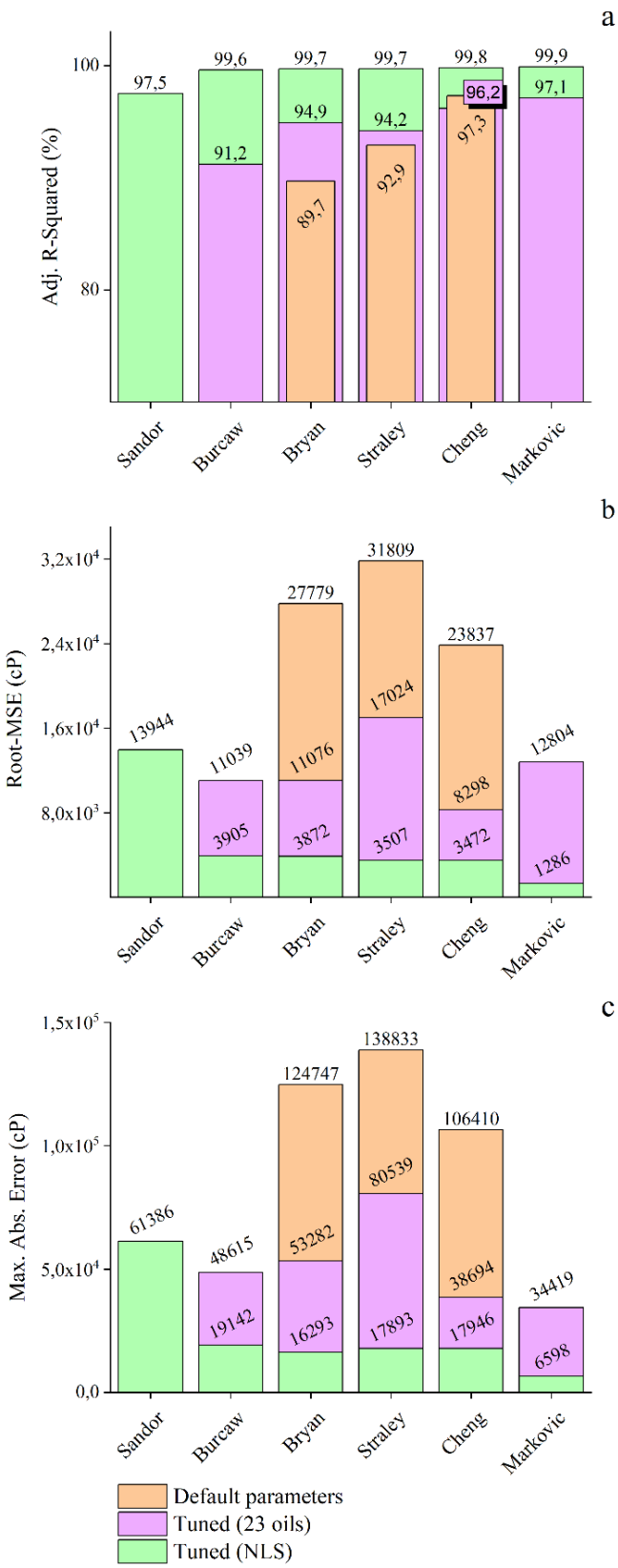

a
Adj. R-Squared (%)
100
80
97,5
99,6
91,2
99,7
94,9
89,7
99,7
94,2
92,9
99,8
96,2
97,3
99,9
97,1
Sandor
Burcaw
Bryan
Straley
Cheng
Markovic
b
Root-MSE (cP)
3,2x10⁴
2,4x10⁴
1,6x10⁴
8,0x10³
13944
11039
3905
27779
11076
3872
31809
17024
3507
23837
8298
3472
12804
1286
Sandor
Burcaw
Bryan
Straley
Cheng
Markovic
c
Max. Abs. Error (cP)
1,5x10⁵
1,0x10⁵
5,0x10⁴
0,0
61386
48615
19142
124747
53282
16293
138833
80539
17893
106410
38694
17946
34419
6598
Sandor
Burcaw
Bryan
Straley
Cheng
Markovic
Default parameters
Tuned (23 oils)
Tuned (NLS)

Figure 17: Compared bar chart of adjusted R^2 (a), Root-MSE (b), and MaAE (c) of NMR model predictions for JC bitumen using three model configurations. The model by Markovic et al. demonstrates the highest accuracy after NLS regression.

The Cheng et al. model (Figure 16e) with default fitting parameters reproduces the most accurate predictions compared to the other two correlations with default parameters by Bryan et al. and Straley et al., whose predictions fall within a factor one and two in the 10–10,000 cP interval. Accuracy deteriorates at >10,000 cP, which induces high RMSE and MaAE scores (Figure 17b, 17c, Cheng). In comparison, Cheng et al. underpredicts the viscosity <1,000 cP domain after NLS regression, especially in the 10–60 cP interval where predictions approach 0 cP, while accuracy is improved in the >10,000 cP domain. As expected, similar prediction behavior is exhibited by the model from Straley et al. This behavior can be attributed to the Curie effect and NLS regression.

The Curie effect manifests through NMR signal loss with temperature increase. Figure 9 shows that the slope of the JC bitumen NMR signal gradually decreases with rising temperature until the inflection point at 100 ℃, after which the slope becomes negative. This effect is illustrated in a varying degree for all models in Figure 16 (particularly in Figure 16d, 16e in the <1,000 cP, or >70 ℃ domains); Sandor et al. overpredict the viscosity in this domain due to the TE coefficient, which overcompensates for Curie effect. However, the NLS regression process affects this further. During NLS regression, the iteration algorithm minimizes the squared residuals in the domain where the highest errors occur (i.e., >10,000 cP), causing the prediction accuracy to decline in the 10–1,000 cP range. After NLS regression, this shift in accuracy explains why all models work better in the high viscosity domain.

The combination of heavy oil chemical complexity, signal loss due to long TE (i.e., echo spacing), and the Curie effect represent the main challenge for developing a single NMR model for predicting viscosity in both low and high viscosity systems and for the same oil (JC bitumen) at various temperatures. Each heavy oil and

bitumen component relaxes exponentially, resulting in a complex total echo decay depending on temperature and each component's phase state. The proposed enhanced model addresses these problems by integrating the RHI_v and T_{2lm} power-law term into the correlation. Stretched-exponential derived power-law term considers a smooth distribution of relaxation times for fast-relaxing components. The RHI_v compensates for signal loss since it is in the corrected form while simultaneously accounting for the long echo times (TE). The power-law term in these conditions has two functions: it improves the accuracy in the >1,000 cP domain, and it supplements the measured RHI_v at high temperatures (>100 °C) in correcting the predictions in the <1,000 cP domain (e.g., Bryan et al., vs. Markovic et al. in >100 °C domain in Figure 16, red points).

Alternatively, the accuracy of any model presented in this paper can be improved for datasets with smaller viscosity and temperature ranges using the NLS regression or by splitting the calibration set into two or three subsets (e.g., low, medium, and high viscosity subsets), and performing NLS regression individually for each subset.

Although the newly proposed model generally produces better forecasts in the case of a suite of heavy oil samples and a single bitumen sample at a wide temperature range, certain aspects could be improved in further work.

- Evaluation of the model performance on the set of different heavy oil and bitumen samples from other heavy oil reservoirs would serve as an additional validation, further reducing the forecasts' uncertainty and enabling us to verify whether the proposed model is overfitted.
- In this work, we used non-linear least squares regression which strongly penalizes squared residuals in the high viscosity range, giving less weight to lower viscosity samples. Depending on a task, other cost functions which give less weight to large residuals (mean absolute error) could be minimized. Another option for model optimization could be orthogonal

distance regression (ODR) which proved to work well in instances with asymmetric distribution of observations by providing a less biased fit[84]. Note that ODR was also tested in Chapter 3.

2.4. SUMMARY

This study demonstrates that LF-NMR relaxometry can be applied for viscosity prediction in a broad viscosity range and at a broad range of temperatures (26-200 °C). The results show that published NMR viscosity models cannot accurately predict heavy oil viscosity at this range of temperatures. The enhanced NMR model was associated with an NLS regression (parameter tuning) and used to predict the viscosity of two distinct datasets: 23 heavy oils at 30°C and 50°C from several wells and reservoirs in Alberta and a single bitumen sample (JC bitumen dataset) at 26–200 °C. The prediction quality was evidenced by the root mean square error (RMSE), maximum absolute error (MaAE), and adjusted coefficient of determination (adj. R^2). The new model scored an RMSE of 1,286 cP for the JC bitumen sample compared to the RMSE of 23,837 cP generated by the first runner-up model in default calibration from the literature. For the suite of 23 heavy oils, the enhanced model scored an RMSE of 2,036 cP compared to the RMSE of 15,934 cP generated by the first runner-up literature model. The results also indicate that the new heavy oil NMR viscosity model can be configured for monitoring purposes in high-temperature conditions for order-of-magnitude viscosity monitoring.

Chapter 3 IMPROVED OIL VISCOSITY PREDICTION BY LOW-FIELD NMR USING FEATURE ENGINEERING AND SUPERVISED LEARNING METHODS

3.1. MOTIVATION

In the previous section, it was shown that LF-NMR data could be used for viscosity evaluation of various crude oils by using the enhanced NMR viscosity model, which can account for chemical complexity and a wide span of temperatures with the help of NMR derived parameters such as relative hydrogen index (RHI_v) and T_2 logarithmic mean. However, the determination of RHI or RHI_v requires a recovery of the representative oil sample from the given formation, preferably with preserved gas content, for subsequent laboratory measurements, which is often a technically challenging and expensive task[61,85]. Moreover, oil saturation volumes must be determined independently, which is necessary to normalize the measured oil NMR response by the amplitude of an equal quantity of water [32]. Unfortunately, in the circumstances like these, the empirical NMR models without RHI do not perform satisfactorily for predicting accurate viscosities in heavy oil and bitumen systems[86,87,88]. The theoretical and empirical evidence presented in previous sections shows that T_2-relaxation strongly correlates with oil viscosity [23,29]. However, in heavier, more viscous oils, the T_2 relaxation behavior deviates from conventional models, changing the T_2 correlation with viscosity. Although studies are being conducted to understand better the underlying physics of H proton relaxation behavior in heavy oils[34,35], 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)[24].

This work introduced a supervised learning framework to improve the oil viscosity characterization by LF-NMR relaxometry, using only a single NMR parameter – T_2 logarithmic mean. Although the emphasis has been made on gradient boosting regression trees (GBRT)[89] and support vector regression

(SVR)[90], several other machine learning algorithms were tested as well, including decision trees (DT)[91], random forests (RF)[92], and k-nearest neighbors (KNN)[93]. Multiple linear regression (MLR)[94] was also used. A feature engineering (FE) approach was integrated to maximize the forecasting capacity of the models by deriving new features using the empirical findings from the NMR oil characterization domain [95]. The study results indicate that this strategy can be successfully applied even to small datasets. As most of the underlying mathematical principles of tested algorithms are substantially different, we could observe the study task from different perspectives. The database used for calibration of models in the study was formed from the previously published LF-NMR crude oil data, containing over 130 light and heavy oil samples recovered from various reservoirs in Canada and the USA[18,60,86]. The database consisted of 282 observations in total. The study was segmented into two stages. In the first stage, the feature engineering was performed, and the dataset was randomly shuffled and split into training and testing sets in the 0.75:0.25 proportion. A split seed was selected randomly (random_state=42) and fixed for the reproducibility of the results. Therefore, the training set consisted of 211 observations and a test set of 71. The generalization ability of the models was assessed by the K-fold cross-validation, while model performance was recorded using root mean squared error (RMSE), mean absolute error (MAE), mean square log error (MSLE) mean absolute percentage error (MAPE) and coefficient of determination (R^2). In the second stage, the performance of models was compared against another 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.

3.2. METHODOLOGY

3.2.1. GRADIENT BOOSTED REGRESSION TREES

In supervised learning, gradient boosting represents an ensemble (additive) model that can be used for solving supervised regression and classification problems. The main idea 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 [91,96,97].

As the goal is to estimate the vector of viscosities η from the training set (x), which consists of input features from the Table 2, and Table 3, the model can be expressed in the forward stage-wise form as:

$$F_m(x_i) = F_{m-1}(x_i) + h_m(x_i) = \eta_i \tag{36}$$

where $h_m(x_i)$ is the underlying model at m-th iteration for i-th observation. This equation can be rewritten as:

$$h_m(x_i) = \eta_i - F_m(x_i) \tag{37}$$

From Equation 37, 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[91]. When dealing with datasets with a large number of outliers, the commonly used squared error loss function $L = \Sigma(y_i - F(x_i))^2$ will emphasize the larger residuals. Absolute loss function is not squaring the errors $L = \Sigma|(y_i - F(x_i)|$, making it therefore more resistant to outliers. The negative gradient with an absolute loss function can be denoted as:

$$-\frac{\partial L\left(\eta_i, F_{m-1}(x_i)\right)}{\partial F_{m-1}(x_i)} = sign\left(\eta_i - F_{m-1}(x_i)\right) \tag{38}$$

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 it, a shrinkage coefficient (ν) is introduced in the calculation of $F_m(x)$, which gauges the contribution of each tree $h_m(x_i)$.

$$F_m(x_i) = F_{m-1}(x_i) + \nu h_m(x_i) \quad (39)$$

This coefficient is also known as the "learning rate," and its optimal value can be estimated using some of the parameter search techniques [98]. It should be noted that learning rate ν is in the strong inverse relationship with number of DTs, that is the number of boosting iterations (M). Usually, lower values of ν lead to a smoother convergence if used with larger values of M [91]. A more detailed explanation of gradient boosting concepts can be found elsewhere [89,96,97,99].

3.2.2. 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 minimizes 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 non-linear classification problems. Fortunately, the same concept was generalized for regression tasks [100]. 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_{2lm}, TE, and T(°K)) and output vector η (Tables 2 and 3). Suppose we arrange all the preprocessed input features in a matrix form as $x = [x_1, x_2, x_3 \dots x_n]$, where x_n are column vectors of inputs. The measured viscosity instances can be rewritten into a response vector $\eta = [\eta_1, \eta_2, \eta_3, \dots, \eta_n]$. Thus the

dataset can be defined then as $\{(x_i, \eta_i)\}_{i=1}^{n}$. Where n is the number of oil samples. The support vector machine regression between input and response vector can be written as:

$$\eta: f(x) = W \cdot \phi(x) + b \tag{40}$$

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 terms. The latter two are obtained by minimizing the risk function:

$$Min: \frac{\|W\|^2}{2} + C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(\eta_i, f(x_i)) \tag{41}$$

$$L_\varepsilon(\eta_i, f(x_i)) \begin{cases} 0 & if\ |\eta_i - f(x_i)| \le \varepsilon \\ |\eta_i - f(x_i)| - \varepsilon & otherwise \end{cases} \tag{42}$$

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 (Equation 42). 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 [100]. The optimization of Equations 41 and 42 can be simplified by introducing slack variables (ξ_i, ξ_i^*) instead of prediction residuals [90]:

$$Min: \frac{\|W\|^2}{2} + C\frac{1}{n}\sum_{i=1}^{n} (\xi + \xi^*) \tag{43}$$

$$Subject\ to: \begin{cases} \eta_i - W \cdot \phi(x_i) - b \le \varepsilon + \xi_i \\ W \cdot \phi(x_i) + b - Y_i \le \varepsilon + \xi_i^*, \quad i = 1, \dots, n \\ \xi_i \ge 0 \qquad\qquad \xi_i^* \ge 0 \end{cases} \tag{44}$$

To find the local minimum with respect to the given constraints, one can introduce Lagrange multipliers, in which case Equation 40 is transformed into:

$$\eta: f(x) = \sum_{i=i}^{n} (\alpha - \alpha_i^*) \cdot K(x_i, x_j) + b \quad (45)$$

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-dimension space. Further details about support vector machine regression can be found elsewhere [90,100].

3.2.3. DATABASE OF RHEOLOGICAL AND NMR MEASUREMENTS

The oil data were collected from our previous research and other published works [18,60,86]. In all studies, the experimental procedure 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 materials data mathematical inversion from the measurements made by LF-NMR relaxometers. For this study, 282 data points were used for model development.

3.2.4. 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 [98]. The feature dataset consists of T_{2lm}, TE, and T(°K), while viscosity observations were stored as output vectors. The data was divided into the training set and a test set in the 3:1 proportion, respectively. This way, we obtained a training set of 211 data points and a test set (unseen data) of 71 data points, which was used to estimate model accuracy only.

3.2.5. 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 relationships. This strategy proved to be effective in

reducing the complexity of SL models, which in turn led to an increase in prediction performance[95]. In our case, this entailed: (1) the transformation of inputs T_{2lm}, TE, T(°K), and target output η, and (2) deriving new inputs from empirical relationships of T_{2lm}, TE, and T(°K) with target output η.

Table 2 shows that the ranges of inputs and outputs are out of scale, which implies that a particular 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 18a, with over 95 % of samples distributed between 0.8 – 100,000 cP 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 [101]. In our case, however, the outliers correspond to extra-heavy oils and bitumens (e.g., > 180,000 cP). In practice, the natural reservoirs in which these oils reside are often thermally treated in order to facilitate their recovery [102]. 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 18b, 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).

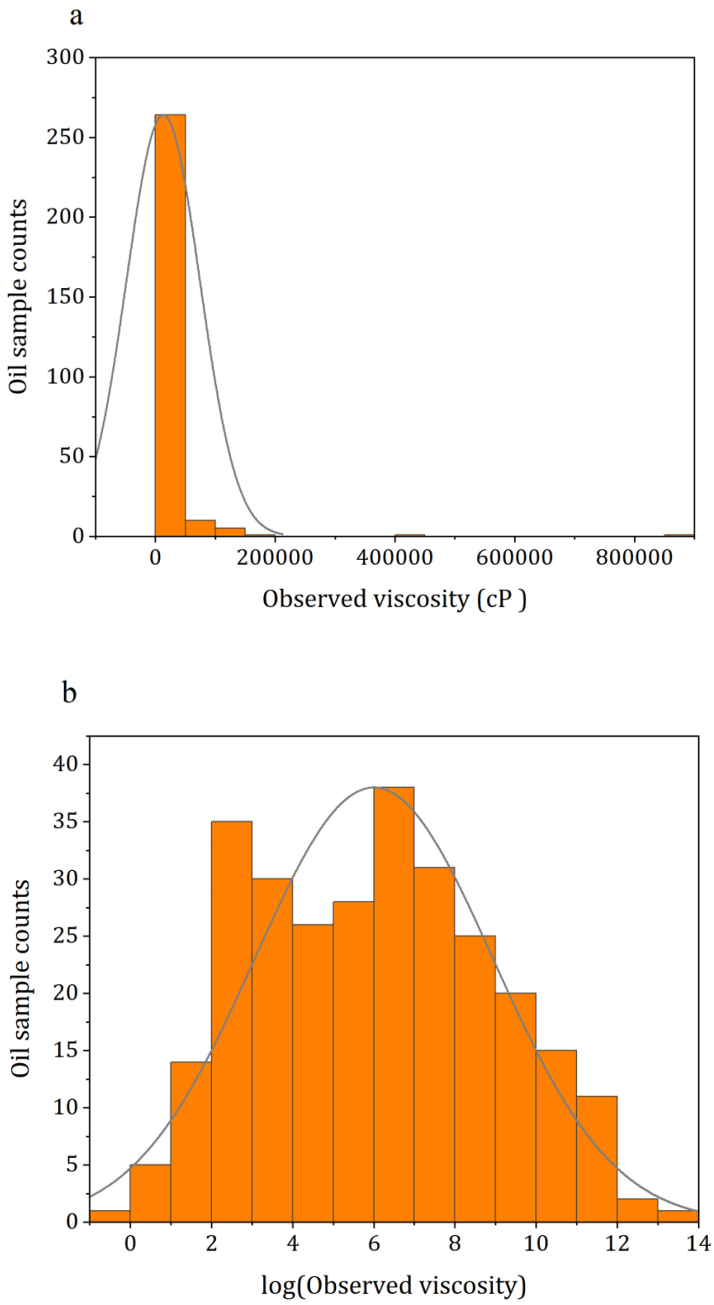


Figure 18: Distribution of oil viscosity η before (a), and after the log transformation (b). Note that η is a target variable.

In the second stage, we derived the new features by employing the findings from previous studies[31,33,103]. 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 a lower rank, thus potentially leaving out a strong predictor [104]. Figure 19a shows the ranking of the seven new features, alongside with T_{2lm}, TE, and T(°K), 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 19a that T_{2lm}-related features (*log(T_{2lm})/TE, log(T_{2lm})*, and *T_{2lm}*) capture most of the variability (~72 %), while T(°K)-derived features (*log(T)/TE, log(T)*, and *T*) capture about 25% of the variability. This variability distribution was expected, considering that the T_{2lm} strongly correlates with η, whereas T_2-related features become more important at high temperatures when severe NMR signal loss occurs.

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 to optimize the signal of fast relaxing fluids, i.e., through the use of small TE values. If this dataset were to be expanded to systems with larger TE values (0.6 – 1.2 ms), TE's impact would 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_{2lm})/TE*). Figure 19b illustrates the relative importance of the remaining six features used for the training of the SL models. Finally, Table 3 summarizes the statistical description of log-transformed viscosity (i.e., target output) and engineered inputs used for SL viscosity forecasting alongside original features (T_{2lm}, TE, and T(°K)).

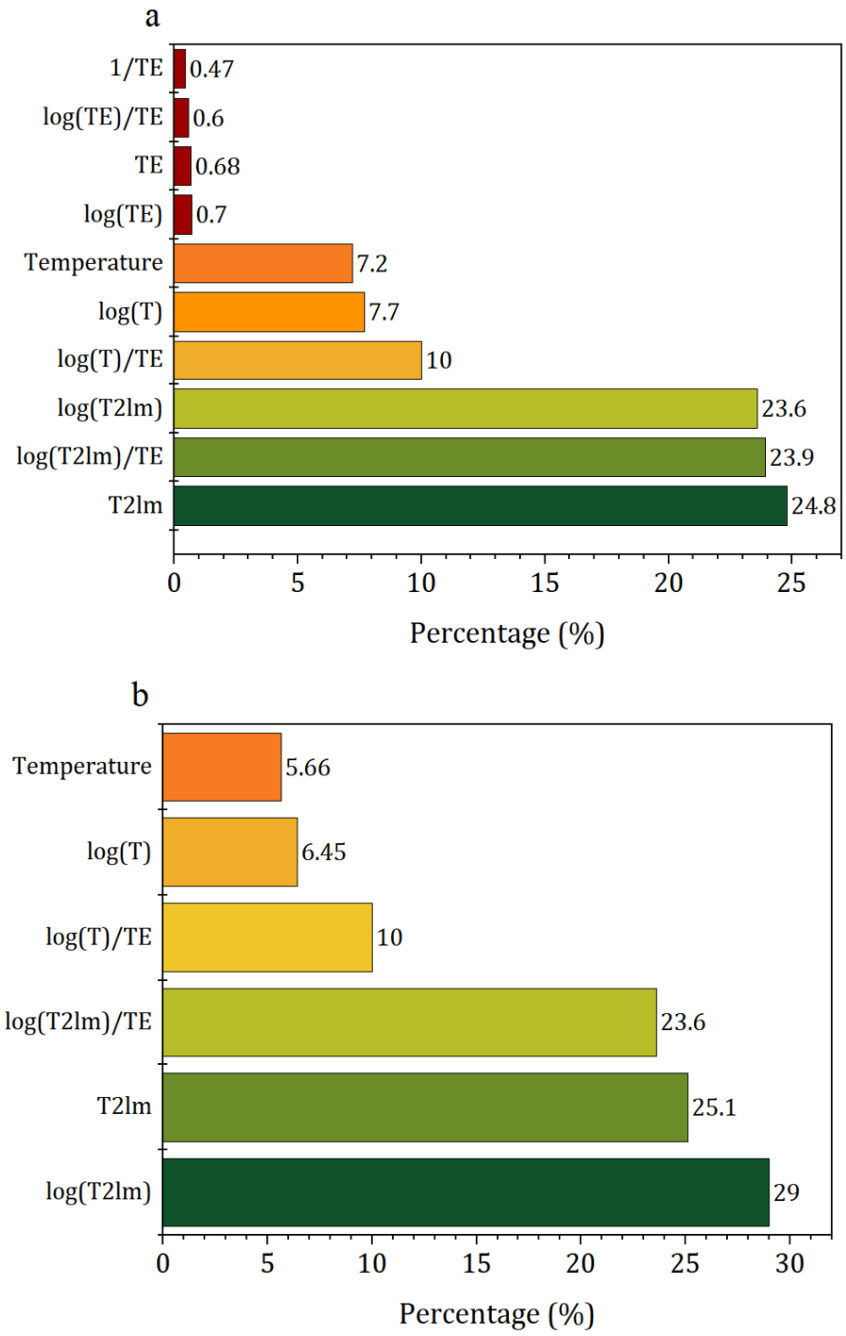


Figure 19: Training set relative feature importance (ranking) by GBRT model of all input features, (a) before, and (b) after removal of redundant TE-derived features with less than <1% relative contribution.

Table 2: Descriptive statistics of six input features (input variables) used for training of SL models, and a log-transformed viscosity (Log(η)) which is a target variable.

Features	Range	Mean	Standard deviation	Count
Log(η)*	-0.13 - 13.67	6.00	2.93	282
Log(T_{2lm})	-1.46 - 7.12	2.08	1.91	282
Log(T)	5.70 - 6.15	5.81	0.12	282
Log(T)/TE	19.04 - 57.14	34.55	16.25	282
Log(T_{2lm})/TE	-14.69 - 71.22	14.74	19.23	282
T_{2lm} (ms)	0.23 – 1239.90	59.38	165.58	282
T (°K)	299.15 – 468.15	337.43	45.15	282

*target variable

3.2.6. EVALUATION METRICS

Five statistical metrics were chosen for the evaluation of the 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 ($\bar{R}^2$). All metrics are negatively oriented statistical measures (i.e., smaller values are favorable), except $\bar{R}^2$, which is positively oriented.

The RMSE is regularly employed in scientific studies to evaluate model performance [44,105]. In this study, the RMSE is the square root of the average of squared differences between predicted viscosity and observed viscosity and is expressed in the centipoises:

$$\text{RMSE}= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\eta_i - \hat{\eta}_i)^2} \qquad (46)$$

where n is a number of samples, η_i is a predicted and $\hat{\eta}_i$ is the observed viscosity. However, this metric can be sensitive to outliers, which can inflate the value of RMSE [106]. To address this issue, MAE is introduced for the calculation of averaged prediction errors of the models, in centipoises:

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\eta_i-\hat{\eta}_i| \tag{47}$$

In contrast to RMSE, MAE does not square the differences between predicted and observed viscosity, making MAE less sensitive to outliers [107]. 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 *log-scaled* predictions and observations:

$$\text{MSLE}=\frac{1}{n}\sum_{i=1}^{n}(log(\eta_i+1)-log(\hat{\eta}_i+1))^2 \tag{48}$$

In this manner, the MSLE avoids the heavy penalization of prediction errors in the high viscosity domain, as is the case with RMSE and MAE. Instead, it considers the relative percentual difference between observation and prediction rather than the size of their residual[108]. In addition to MSLE, MAPE illustrated the relative percentual difference between sums of errors. MAPE represents the mean of the sums 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[109]:

$$\text{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\eta_i-\hat{\eta}_i}{\hat{\eta}}\right|\cdot 100\% \tag{49}$$

Lastly, 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:

$$R^2=1-\frac{\sum_{i=1}^{n}(\eta_i-\hat{\eta}_i)^2}{\sum_{i=1}^{n}(\eta_i-\bar{\eta})^2}=1-\frac{SS_{res}}{SS_{tot}} \tag{50}$$

Although this metric provides a fast and straightforward evaluation, it might inflate due to the addition of new variables obtained from feature engineering.

This inflation is a well-known problem that can be addressed by adding a term that penalizes the score with each additional predictor [110]:

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1} \quad (51)$$

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

3.2.7. GBRT OPTIMIZATION

As mentioned in section 3.2.1, 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 number of outliers, which implied using an outlier-resistant loss function, such as the least absolute deviation (LAD)[101]. 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 4).

Table 3: The five-fold cross validation training set scores for different GBRT loss functions. The LAD loss exhibits the best performance based on MAE_{cv}, $RMSE_{cv}$, $MSLE_{cv}$, $\bar{R}^2_{cv}$ cross-validation (CV) scores.

Test scores	**'Loss' parameter**			
	LAD	**Huber**	**LS**	**Quantile**
MAE_{cv}	**6332**	7319	9072	6969
$RMSE_{cv}$	**17,714**	22,289	34,305	29,480
$MSLE_{cv}$	**0.198**	0.251	0.26	0.464
$\bar{R}^2_{cv}$	**0.58**	0.348	-0.54	-0.14

The 'criterion' parameter can be determined in the same manner. This parameter allows a user to select the function that will estimate the DT node split quality. 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 5).

Table 4: The five-fold cross validation training set scores for different GBRT criterion hyperparameters. Node splits by MAE criterion achive best scores in terms of MAE_{cv}, $RMSE_{cv}$, $MSLE_{cv}$, $\bar{R}^2_{cv}$ cross-validation (CV) scores.

Test scores	**'Criterion' parameter**		
	MAE	**MSE**	**Friedman-MSE**
MAE_{cv}	**3666**	5756	5757
$RMSE_{cv}$	**9925**	16286	16287
$MSLE_{cv}$	**0.149**	0.183	0.183
$\bar{R}^2_{cv}$	**0.86**	0.66	0.64

The next step was to find the optimal hyperparameter values for the GBRT model. According to [91,99], five hyperparameters have a considerable impact on GBRT model performance:

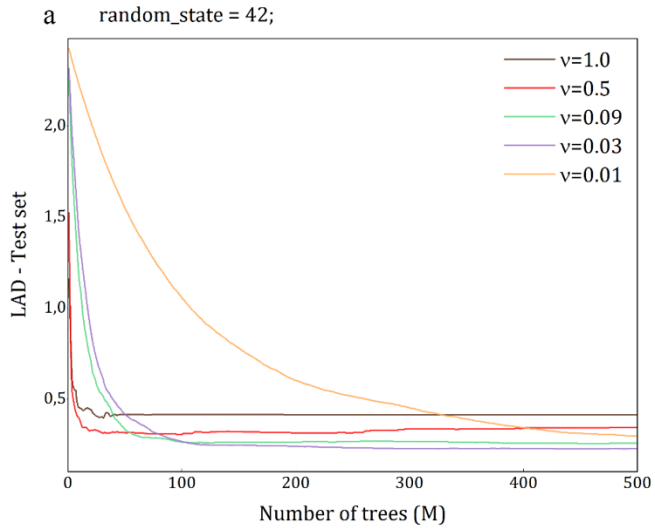
- Number of trees (M): maximum number of estimators or boosting iterations (n_estimators).
- Learning rate (ν): shrinkage coefficient, which regulates the individual tree prediction contribution, where each tree is being scaled by $0 < \nu < 1$.
- Subsample (λ): the proportion of the data for fitting 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.

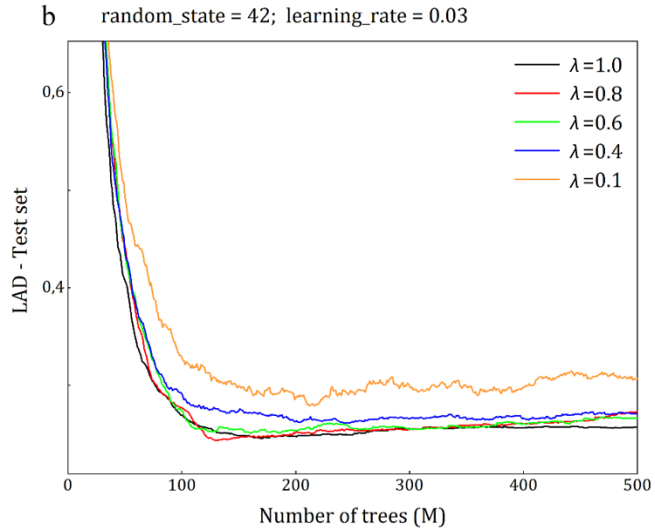
Mathematically speaking, these hyperparameters are mutually dependent (also observable from Equation 39), which is why it was required to use a more sophisticated optimization technique other than pure trial-and-error. Therefore,

these were evaluated simultaneously with the help of GridSearchCV (GS-CV), an exhaustive search cross-validation algorithm available in a scikit-learn package [98]. 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 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 6 shows the hyperparameters and their value range, which were optimized using the GS-CV approach. Recall that 'loss' and 'criterion' parameters were fixed according to results in Tables 4 and 5.

Table 5: Training set 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, log2]	[log2]	$\bar{R}^2$: 0.91


a
random_state = 42;
ν=1.0
ν=0.5
ν=0.09
ν=0.03
ν=0.01
LAD - Test set
2,0
1,5
1,0
0,5
0
100
200
300
400
500
Number of trees (M)


b
random_state = 42; learning_rate = 0.03
λ=1.0
λ=0.8
λ=0.6
λ=0.4
λ=0.1
LAD - Test set
0,6
0,4
0
100
200
300
400
500
Number of trees (M)

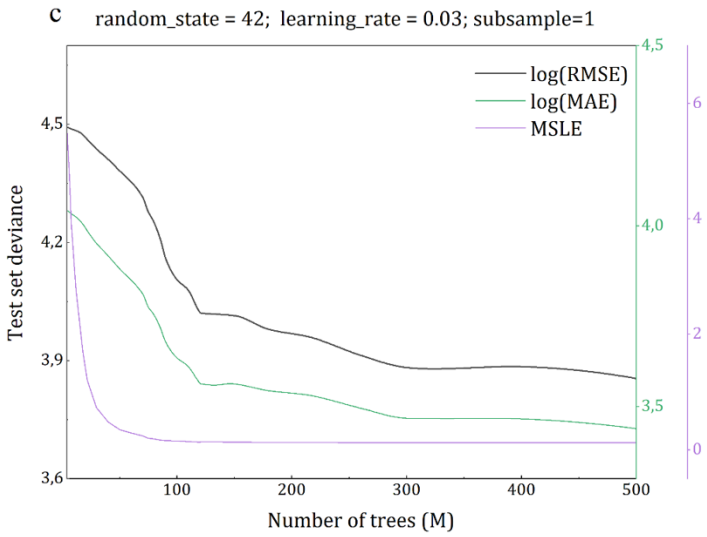

c
random_state = 42; learning_rate = 0.03; subsample=1
log(RMSE)
log(MAE)
MSLE
Test set deviance
4,5
4,2
3,9
3,6
4,5
4,0
3,5
6
4
2
0
100
200
300
400
500
Number of trees (M)

Figure 20: The 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.

Figure 20a 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 increase to ensure smooth convergence, which also means an increase in 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 helpful for improving computing performance and accuracy [91]. Figure 20b, however, shows that no subsampling ($\lambda = 1$) leads to the 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 minor that one must zoom in on the y-axis to observe this behavior (note y-axis scales of Figure 20b).

Finally, the maximum depth of all trees was restricted to the same size ($J=4$), as determined by the GS-CV, which agrees with recommendations in literature [91]. The optimal number of features for the best split of the tree node was found to be ψ=2 (i.e., *max_features = "log2"*). 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.

3.2.8. SVR OPTIMIZATION

In Equation 45, 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 [111]. From Equation 45, the kernel function can be expressed as:

$$K(x_i, x_j) = exp\left(\gamma \cdot \left\|x_i - x_j\right\|^2\right) \quad (52)$$

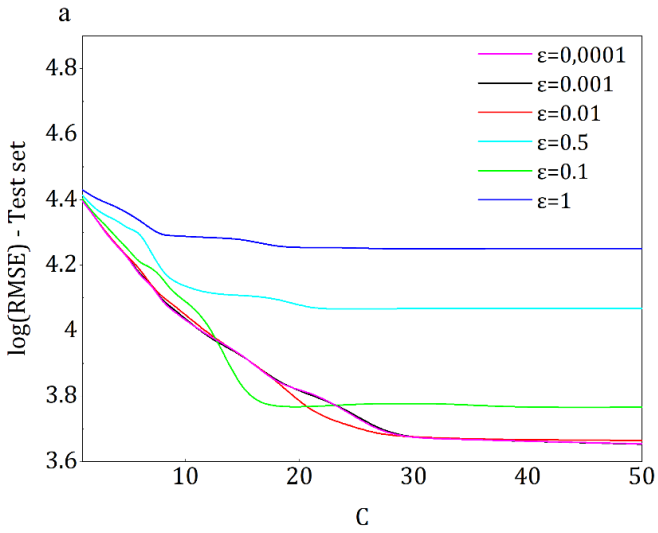
where γ is the width hyperparameter of the RBF kernel. Hence, there are three main hyperparameters which need to be optimized:

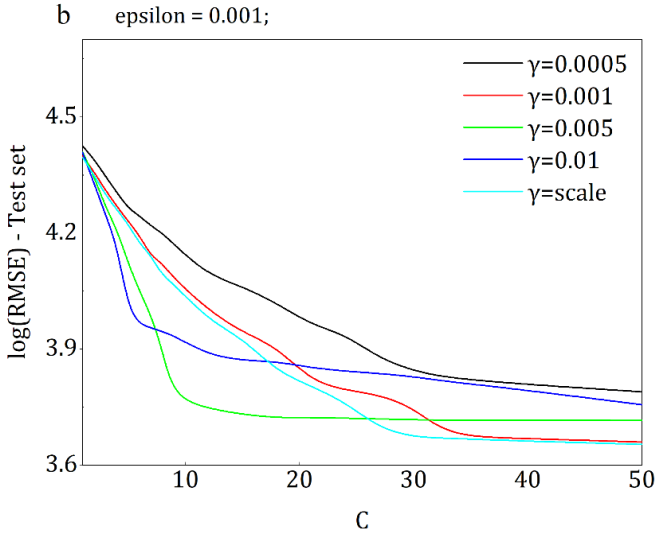
- Gamma (γ): RBF kernel specific parameter which defines the support vector's radius of impact.
- Epsilon (ε): the insensitivity radius-ε within which the prediction residuals are ignored (*loss=0*). This value controls the number of support vectors (SVs) and the smoothness of the function.
- Regularization parameter (C): hyperparameters, 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[100]. Thus, the GS-CV approach was used to simultaneously approximate *C,* ε, and RBF kernel parameter γ (Table 7). Also, their in-depth assessment was performed from the analysis of the error curves (Figure 21).

Table 6: Training set SVR hyperparameter optimization by grid-search based on 5-fold cross-validation. Note that radial basis function (RBF) kernel was used.

SVR hyperparameters	**Range/method**	**Optimal values**	**Score**
Gamma (γ)	['scale', 0.1-1·10^{-5}]	[5·10^{-4}]	RMSE: 8704
			MAE: 3377
Epsilon (ε)	[1-1·10^{-5}]	[1·10^{-4}]	MSLE: 0.136
			MAPE: 29
Regularization (C)	[1-500]	[25]	$\bar{R}^2$: 0.91


a
log(RMSE) - Test set
C
ε=0,0001
ε=0.001
ε=0.01
ε=0.5
ε=0.1
ε=1


b
epsilon = 0.001;
log(RMSE) - Test set
C
γ=0.0005
γ=0.001
γ=0.005
γ=0.01
γ=scale

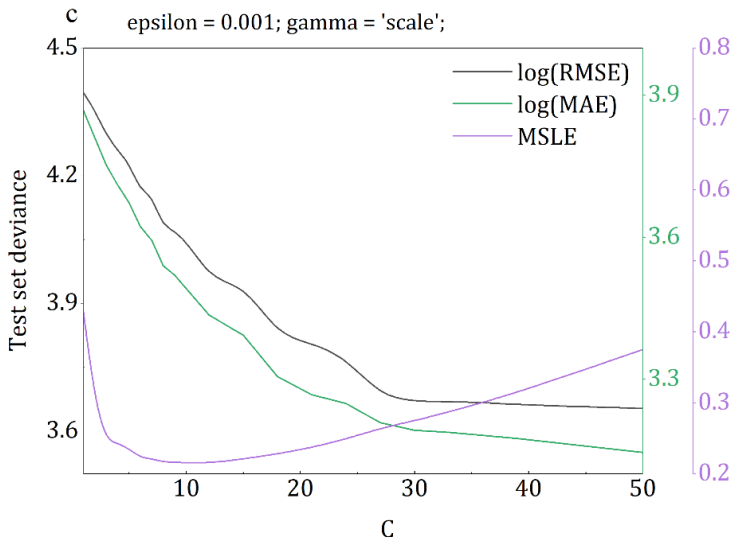

c
epsilon = 0.001; gamma = 'scale';
Test set deviance
C
log(RMSE)
log(MAE)
MSLE

Figure 21: 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. Bottom plot (c) illustrates the accuracy of optimized SVR model as a function of C, in terms of three error metrics; log(RMSE), log(MAE) and MSLE.

Figure 21a 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[112]. For values of $\varepsilon=10^{-3}$ and $\varepsilon=10^{-2}$, the deviance converged smoothly at *C=30*. The number of SVs was reduced by increasing ε to 10^{-3} (Figure 21a, black) while preserving nearly identical accuracy.

Figure 21b shows deviance for the fixed ε and various radii of individual SV impact γ. The smoothest convergence and lowest deviation are achieved when $\gamma=$*' scale,'* which is the value when an inverse of the number of features is scaled by their standard deviation. Interestingly, the GS-CV obtained $\gamma= 5\cdot10^{-4}$, 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, the γ was set to *'0.001.'*

As a final step, the model with fixed ε and γ hyperparameters was evaluated in Figure 21c, where three metrics were utilized to evaluate 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 7).

3.3. 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 chapter 3.2.6. Also, the cross-plots with predicted and observed viscosities are provided for the in-depth analysis.

3.3.1. SUPERVISED LEARNING MODELS

The performance of the GBRT and SVR was evaluated against an additional four SL algorithms, including multiple linear regression (MLR) [94], K-nearest neighbors (K-NN) [93], decision trees (DTs) [91], and random forests (RF) [92]. Their optimization was performed with GS-CV, similarly as for GBRT and SVR. Recall that the dataset contains 282 observations which were shuffled and split into train and test set in 0.75:025 proportion, making a training and test set of 211 and 71 observations, respectively.

Table 7: Training set GS-CV hyperparameter optimization results for all supervised learning algorithms which were tested in this work.

Model	Hyperparameters	Range/method	Optimal values
Decision trees (DTs) [91]	criterion	['mse', 'mae']	['mse']
	max_features	[1, 2, 3, 'sqrt', 'log2', 'auto']	[3]
	max_depth	[1-4]	[4]
	min_samp_leaf	[1-4]	[3]
	min_samp_split	[1-4]	[2]
	splitter	['best', 'random']	['best']
K-Nearest Neighbors (KNN) [93]	n_neighbors	[1-50]	[3]
	weights	['uniform', 'distance']	['uniform']
	algorithm	['ball_tree', 'kd_tree', 'brute']	['ball_tree']
	p	[1, 2]	[1]
Random forests (RF) [92]	n_estimators	[1-80]	[7]
	criterion	['mse', 'mae']	['mae']
Support vector machines for regression (SVR) [100]	gamma	['scale', 0.0005-0.1]	[0.01]
	epsilon	[1-0.0001]	[0.001]
	C	[1-500]	[25]
Gradient boosted regression trees (GBRT) [91]	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, log2]	[log2]
	max_depth	[1-8]	[4]
	subsample	[0.1-1.0]	[1.0]

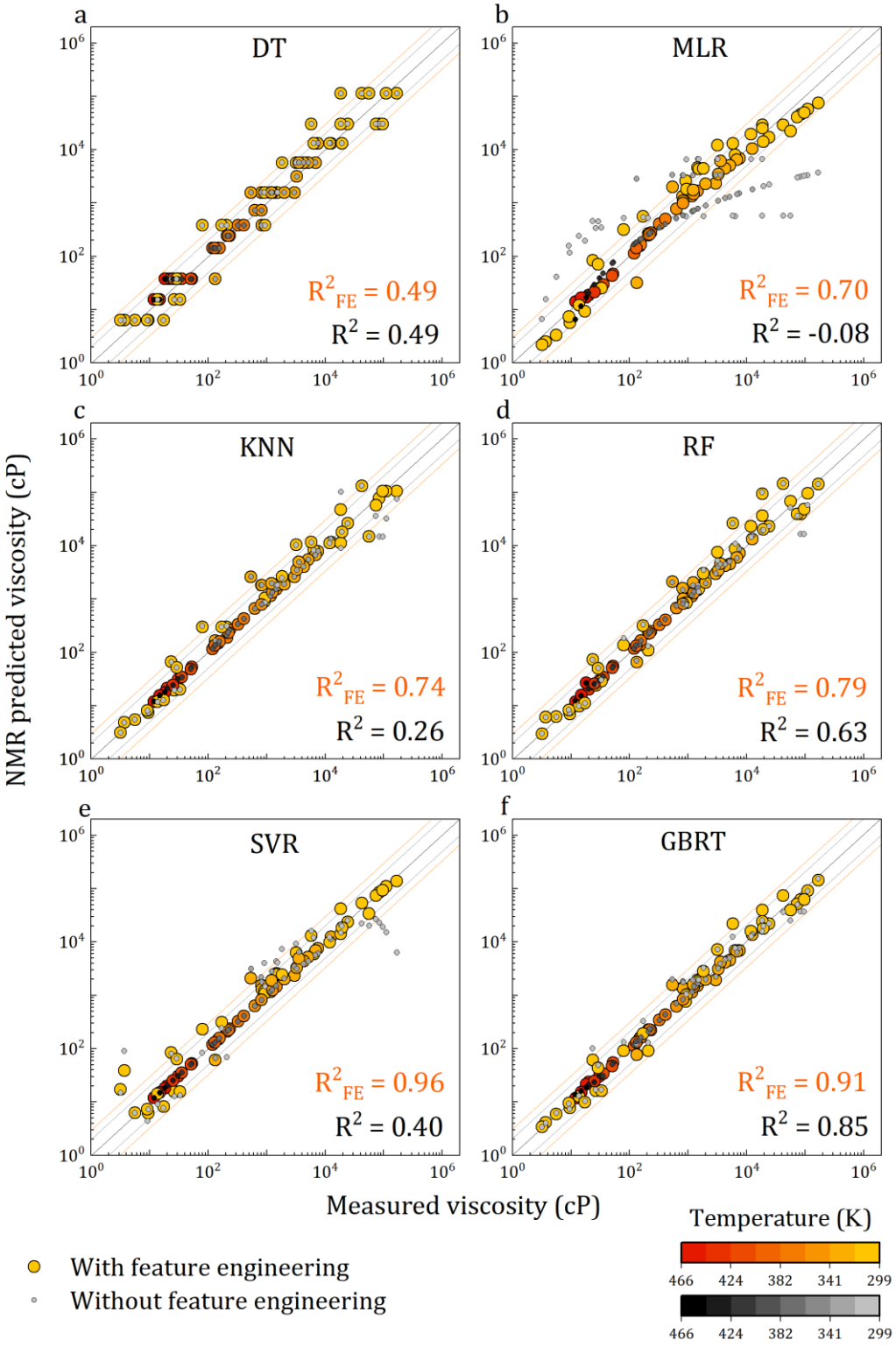


Figure 22: Comparison of NMR SL viscosity model test set predictions and observations. Note that the grayscale points are predictions of models generated without FE, while warm color points are predictions with FE. Lighter colors indicate lower temperatures

(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.

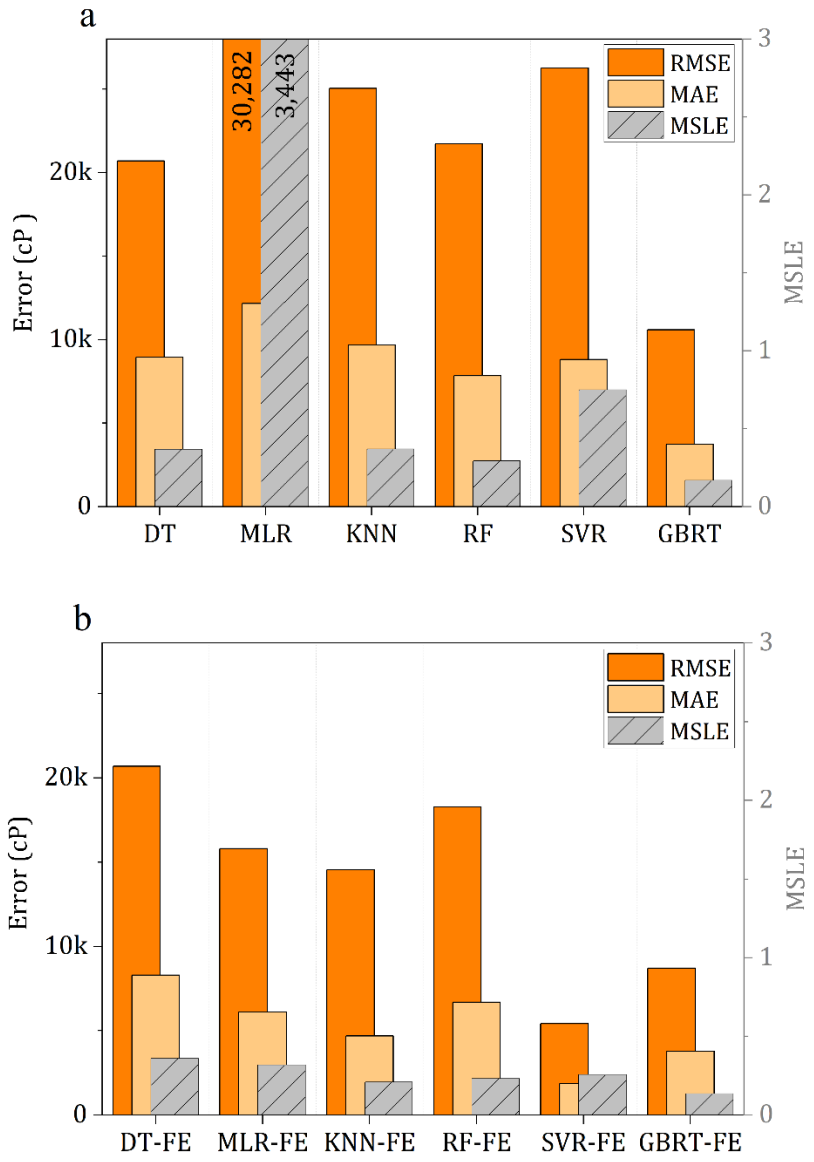


Figure 23: Compared test set statistical scores of SL models without FE (a) and SL models with integrated FE (b). SVR-FE and GBRT-FE demonstrate the best statistical performance.

When Figures 19 and 20 are examined together, one can note that the overall performance of each model improves after the integration of FE. This effect is, however, not proportionally pronounced for all models. For instance, the MLR-FE

model's prediction variance has reduced dramatically after employing FE (Figure 22b). Interestingly, for GBRT and RF models, the variance-reducing effect from FE is much smaller than in the latter's case (Figures 22e and 22c). The same observation applies to the DT model, the base estimator of GBRT, and RF models .

This difference in performance comes from the difference in the underlying mathematical principles of these models. As the MLR model is *linear*, the nonlinearity reduction 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. A similar is valid for the SVR model, though to a smaller extent (Figure 22f) [113]. 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 remained the same after transformations [91]. 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 23, with those by GBRT, we observe a massive gap in performance, which perfectly illustrates the advantage of ensembles of DTs over a single DT. One of the reasons for the poor performance of single DT models is their 'habit' of overfitting the training data, making them unstable with unseen data. Therefore, ensembles of DTs generate a variance that minimizes the overfitting [91].

Lastly, the KNN 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, defined by the user [93]. Feature engineering improves KNN scores almost proportionally to RF and GBRT models, however not enough to minimize the large residuals in the high viscosity domain, which causes the RMSE and MAE scores to rise (Figures 20a and 20b).

Another remark is that the significant temperature variations 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 observed oil viscosity and NMR signal loss that occurs at high temperatures. In comparison, empirical models tested in this work[23,31,33,103], exhibit poor performance in these conditions [86], as seen in the following chapter.

3.3.2. 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_{2lm}, TE, and T. These models were developed by Straley et al. [23]., Nicot et al. [103], Cheng et al. [33], and Sandor et al. [31]. Previous research showed that tuning by non-linear least squares (NLS) improves the performance of empirical models [86]. However, the viscosity 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 [114,115]. Hence, the model fitting was performed using the orthogonal distance regression (ODR), proved to be a successful technique for dealing with outliers [116]. Figure 24 and Figure 25 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 24 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.

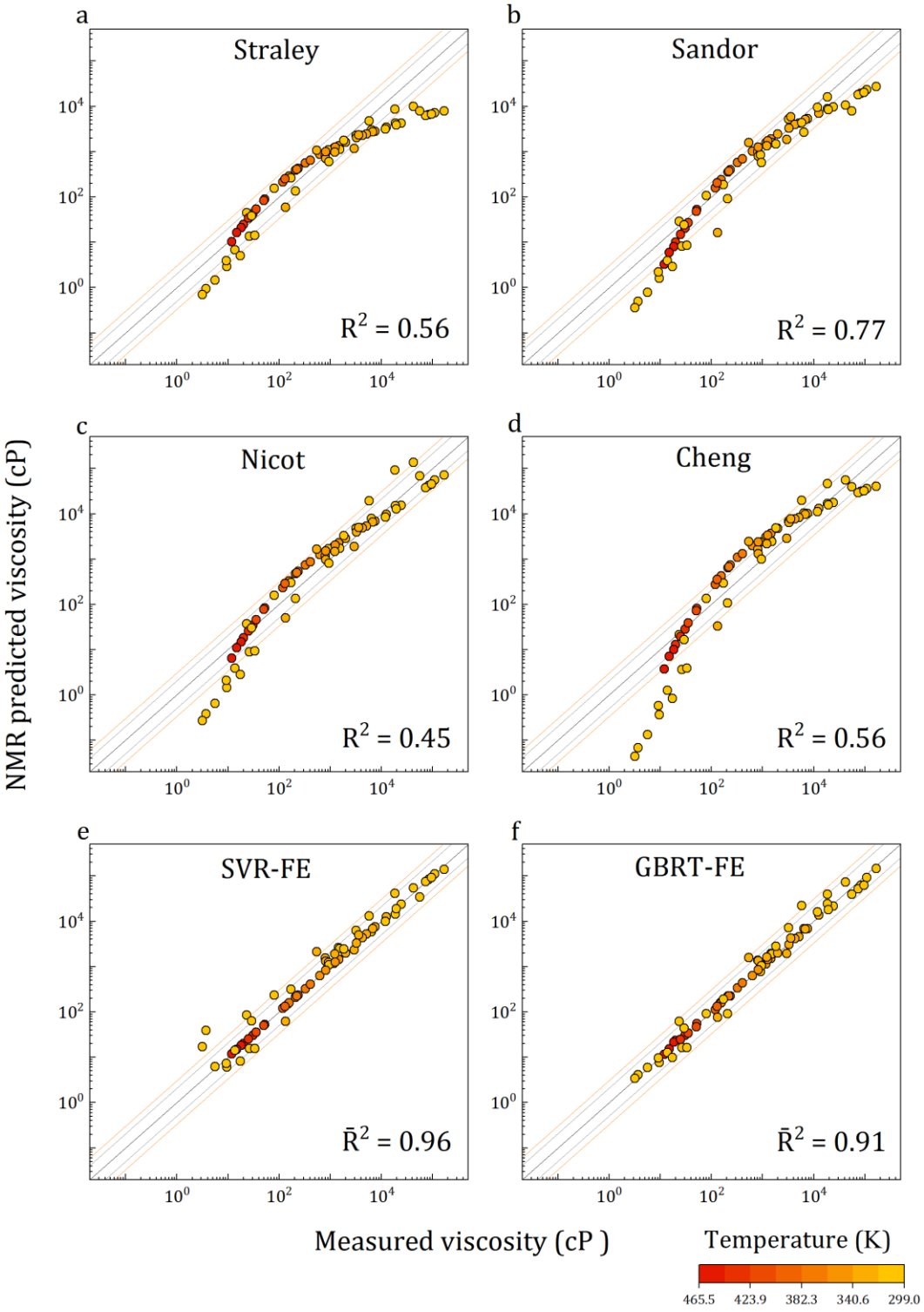


Figure 24: Performance comparison of empirical NMR viscosity model test set predictions (a, b, c, and d) with SVR-FE (e) and GBRT-FE (f) models. GBRT-FE and SVR-FE demonstrate significantly better statistical performance.

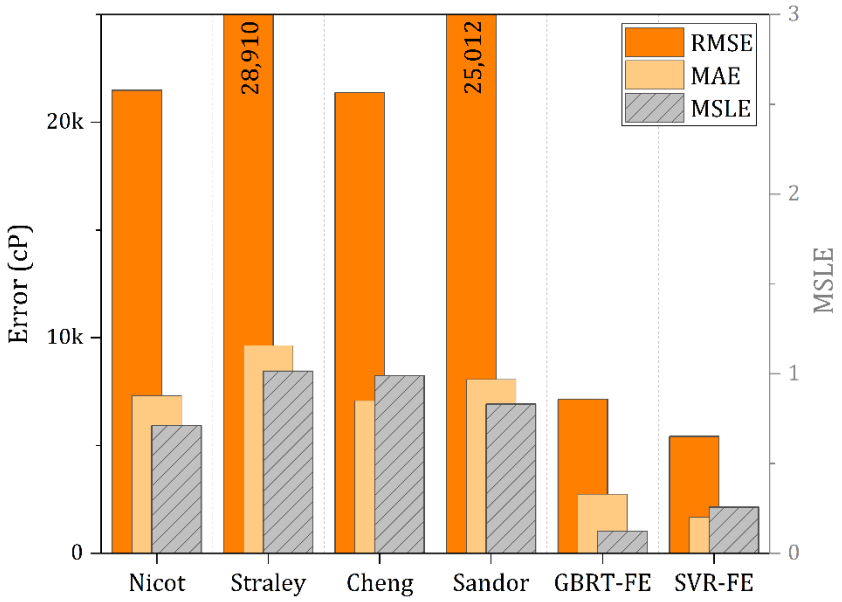


Figure 25: Compared test set statistical scores of four empirical NMR viscosity models and SVR-FE and GBRT-FE supervised learning models in terms of RMSE, MAE, and MSLE. SVR-FE and GBRT-FE demonstrate significantly better statistical performance.

3.3.3. SVR-FE vs. GBRT-FE

When cross-plots from Figure 22 and Figure 24 are analyzed along with scores in Figure 23, Figure 25, one can conclude that SVR-FE and GBRT-FE models have superior test performance than any other SL model. For instance, in the case of MLR-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 than empirical models (~4.5 times lower) but compared to SL models, KNN-FE and RF-FE marginally outperform SVR-FE. The same is true for MAPE scores and percentage error box plots when further examined in Figure 26. 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, KNN-FE, and RF-FE models. For more convenience, all evaluation scores are summarized in Table 9.

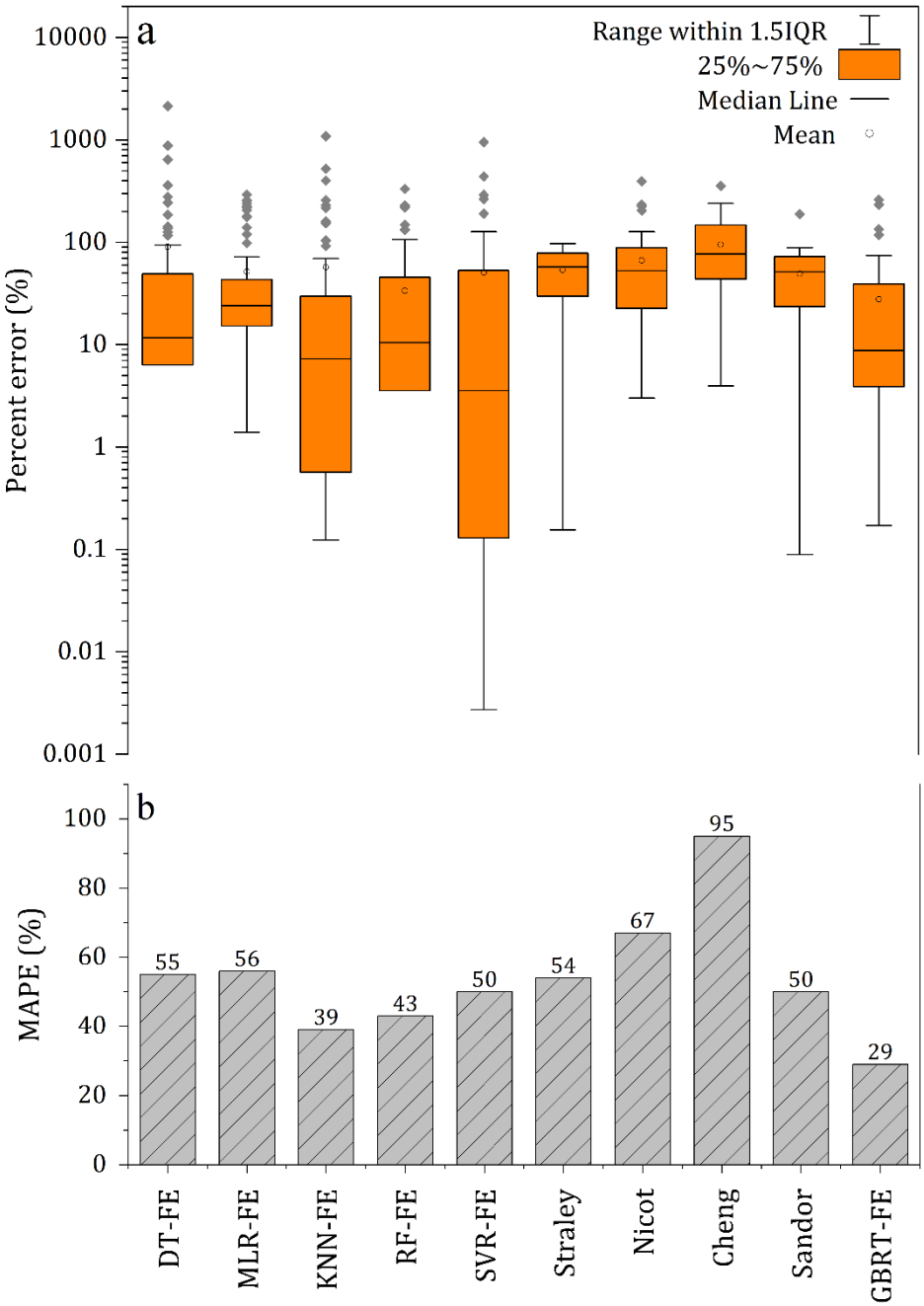


Figure 26: Test set percent error box plots (a) and MAPE scores (b) for six supervised learning models with feature engineering, and four empirical models. Note that in the plot (a) the y-axis is in log-scale. GBRT-FE model demonstrates the best performance in terms of MAPE.

Table 8: Compared view of test set statistical scores for all SL and empirical models. Bolded values correspond to the best score.

Model	Test set scores				
	RMSE (cP)	MAE (cP)	MSLE	MAPE (%)	R^2
DT	20,712	7968	0.368	55	0.49
MLR	30,282	10,858	3.443	282	-0.08
KNN	25,044	8642	0.369	47	0.26
RF	21,725	6989	0.293	44	0.63
SVR	26,266	7858	0.749	93	0.40
GBRT	10,587	3331	0.168	32	0.85
DT-FE	20688	7409	0.359	55	0.49
MLR-FE	15,808	5447	0.319	56	0.70
KNN-FE	14,559	4182	0.210	39	0.74
RF-FE	18,285	5979	0.232	43	0.79
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

3.3.4. PHYSICAL IMPLICATIONS OF SVR-FE AND GBRT-FE PERFORMANCE

From the physical point of view, meaningful insight can be extracted from the models' fundamental understanding. Although SVR-FE is moderately stable and achieves good accuracy, it struggles with capturing additional variability from the oil samples' diverse chemical compositions. This behavior can be observed in Figure 24f, where SVR-FE predictions favor the high-temperature heavy oil sample over the other samples. This occurs due to the structural risk minimization 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; firstly, 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 by 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 chemically diverse set 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 0.1-0.3 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 might 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 [31,74,103]. 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, 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 the chemical diversity of oils and various temperature conditions.

3.4. SUMMARY

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 another 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 $\bar{R}^2$.

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 a miscible gas injection [12]. 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 non-invasive 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 where researchers work with relatively smaller datasets.

Chapter 4 APPLICATION OF XGBOOST MODEL FOR *IN-SITU* WATER SATURATION DETERMINATION IN CANADIAN OIL-SANDS BY LF-NMR AND BULK DENSITY MEASUREMENTS

4.1. MOTIVATION

As discussed in section 1.5.3, the main issues related to the determination of water content in oil-sands by LF-NMR are:

- Insufficient understanding of dominating T_2 relaxation mechanism in fine pore-space saturated by fluid (T_2 bulk + T_2 surface);
- The diffusive-coupling phenomenon associated with the water relaxation between macro- and micro-pores, and;
- Resultant overlapping of water and oil signals in T_2 distribution.

As these issues are intertwined, the determination of T_2 cutoffs for splitting the T_2 distribution to producible and bound fluids and interpretation of fluid types becomes a laborious task in which seemingly minor errors can lead to erroneous predictions of water saturation and, therefore of OOIP.

In this work, we postulated that the combination of LF-NMR T_2 data and bulk density data could be used to effectively separate the contributions of oil and water signals to a degree at which an accurate determination of relative water fraction is possible. For model derivation, two machine learning approaches based on Extreme Gradient Boosting (XGB) were employed. The first modeling approach is based on a feature engineering process that reduces the number of inputs while maximizing model generalization capacity. This was achieved by deriving new features using empirical knowledge from the T_2 distribution analysis domain and a feature extraction technique based on information theory. In contrast, the second approach considers as input the whole NMR T_2 distribution of the sample, aiming to preserve all available information originating from fluids residing in the sample pore space. The dataset comprised 82 oil-sands core samples recovered

from northern Alberta in Canada. The NMR T_2 relaxation distribution of samples was obtained at ambient and reservoir conditions, comprising dataset of 164 observations. Water content percentage relative to of the total mass of the sample was determined by Dean-Stark extraction (%DS-w). The model training and prediction test scores of the models were evaluated using three statistical metrics and a leave-one-out cross-validation (LOOCV). These scores were compared with water content predictions based on the previously published deconvolution method. Deconvolution was performed according to Bryan et al [32,61].

4.2. THEORY

4.2.1. LF-NMR MEASUREMENTS FOR WATER SATURATION DETERMINATION

Three main processes comprise the total T_2 relaxation; bulk relaxation, surface relaxation, and diffusion relaxation due to the gradient in a magnetic field. In this work, the benchtop LF-NMR relaxometer was used in which the gradient is absent, thus the diffusion term can be neglected. In such case, in the Equation 5 the $T_{2diffusion}$ term can be omitted:

$$\frac{1}{T_2} = \frac{1}{T_{2Bulk}} + \frac{1}{T_{2Surface}} \tag{53}$$

$$\frac{1}{T_{2surface}} = \rho_2 \left(\frac{S}{V}\right) \tag{54}$$

Recall that T_{2Bulk} represents the relaxation occurring in bulk fluids or fluids in large pores, and $T_{2Surface}$ generally quantifies the relaxation of fluids in smaller pores. Also recall that ρ_2 is T_2 surface relaxivity, and S/V is a ratio of the fluid volume and surface of the pore. Each of these mechanisms will contribute to the total relaxation in varied proportions depending on reservoir rock properties and physicochemical properties of the fluids, such as rock wettability, pore size and pore surface area, fluid viscosity and chemical composition.

4.2.2. XGBOOST PRINCIPLES

XGBoost stands for Extreme Gradient Boosting (XGB), and it presents an implementation of the gradient boosting decision trees[117]. The main principle of gradient boosting is to utilize the individual weak learner, such as a decision tree,

and in a stage-wise manner, add iteratively new trees, to minimize further the objective function. This process continues for the specified number of boosting iterations, after which the prediction model is obtained in a final form. The algorithm uses gradient descent to minimize the loss function by finding the direction of the "steepest" descent. The loss function is minimized by finding a new best model (tree), and fitting it to the prediction residual. As this is performed in the step-wise manner, the size of the step is controlled by learning rate. If the step is too small the gradient descent will be slow, or may just converge to the local minimum. If it is too large, it will diverge. To find a minimum, the gradient descent uses a first order approximation which assumes that the loss surface is planar in the direction of the descent (i.e., in 3D, a plane tangent to the error surface). If the loss surface is not planar but convex, the gradient descent will ignore information about convexity, which may lead to a slow convergence or convergence to a local minimum. XGBoost, on the other hand, behaves much like Newton's method[118], which uses a second order approximation (i.e., in 3D, a convex quadratic surface with greater overlap with error surface) which assumes that the loss function is twice differentiable. The second order approximation allows considering the curvature of the loss function, and therefore ensures a faster convergence to the loss function minimum compared to the gradient descent. Although, Newton's method requires calculating the Hessian matrix for storing the quadratic term coefficients of local second order function, the XGBoost instead computes the second partial derivative of the loss function (element-wise), which is less computationally demanding.

Another advantage of XGBoost is implementation of L1 and L2 regularization in the penalty function (Equation 57) which reduce the model complexity and overfitting. The XGBoost model which uses K additive functions (trees) can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k\,(x_i) \tag{55}$$

where $\hat{y}_i$ is predicted Dean-Stark water content (%DS-w), x_i is a vector of input features and f_k is an independent tree of the k-th instance. In contrast to decision

trees, in XGBoost, trees contain a continuous prediction residual (score) in each i-th leaf *(w_i)*. By summing up the score in corresponding leaves we can calculate the final prediction. In order for trees to learn, a regularized loss function (L) has to be minimized.

$$L = \sum_{i=1}^{n} l(\hat{y}_i , y_i) + \sum_{k} \Omega(f_k) \tag{56}$$

where l is convex loss function which represents the difference between prediction $\hat{y}_i$ and a target y_i. The right-hand side term Ω penalizes the complexity of the tree, and can be donated as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{57}$$

Here, γ is L1 and λ is L2 regularization parameters, T is the number of leaf nodes in a tree, and w_j^2 are the squared scores of leaves. Equation 57 shows that γ penalizes the growing complexity of the model (large T), while λ serves as a smoothing parameter for the learnt scores w_i to prevent overfitting. Like in gradient boosting regression trees, the model is trained by adding new trees in a stage-wise manner. For the prediction $\hat{y}_i^{(t)}$ at i-th instance and t-th stage:

$$\hat{y}_i^{(0)} = 0 \tag{58}$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \tag{59}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \tag{60}$$

$$\dots$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{n} f_k\,(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{61}$$

where f_t represents a new tree at t-th stage. The Equation 56 can be then expanded such that the minimization of the loss function is performed with respect to penalty term Ω:

$$L^{(t)} = \sum_{i=1}^{n} l\Big(y_i , \hat{y}_i^{(t-1)} + f_t(x_i)\Big) + \Omega(f_t) \tag{62}$$

To enable a quicker minimization of the Equation 62, we can expand it to a second order approximation (Taylor expansion) in which case it will become:

$$L^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i , \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \qquad (63)$$

where g_i is the first order gradient and h_i is a second order gradient. The Equation 63 can be further improved by removing constant terms and expanding the penalty term. These details can be found in the work of Chen et al. [117].

4.3. METHODOLOGY

4.3.1. EXPERIMENTAL PROCEDURE AND DATA PREPROCESSING

Oil-sand samples were collected in northern Alberta in Canada from a single delineation well. Two sets of 82 whole core samples were recovered. The first set was used for laboratory LF-NMR measurements, and a second set represented sister samples used in Dean-Stark extraction for determining the relative fraction of water, oil, and solids. Samples for NMR experiments were stored in glass vials and measured using a Corespec 1000™ benchtop LF-NMR relaxometer at reservoir temperature (6 ℃) and ambient temperature (25 ℃). The Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence was used for obtaining T_2-relaxation distribution. The CPMG parameters were predetermined after a series of test NMR experiments on different oil-sand samples. There were two aspects which had to be taken into account. The first was to tune the CPMG parameters to detect the fast relaxing heavy oil and clay-bound water signals. This was achieved by setting the shortest echo time TE that the equipment allowed (0.2 ms). The second aspect was achieving a lower signal-to-noise ratio (SNR) to simulate the well-logging *in-situ* NMR tool output by reducing the number of trains, which in turn resulted in a noisier signal. After trial rounds of measurements, the following parameters produced optimal T_2 distribution and SNR (Table 10).

Table 9: Optimal CPMG pulse sequence parameters for detection of fast relaxing clay-bound water and heavy oil signals.

CPMG pulse parameters	Values
Echo time, TE (ms)	0.2
Number of pulses, Np	5000
Wait time/post train delay (ms)	6500
Number of trains, NT	10

For the dataset, the range of SNR varied from 5 to 56, with an average of 23. The ExpFit in-house software for multi-exponential analysis of the NMR signal was used. The representation of the signal after Inverse Laplace Transform (ILT) was obtained using Tikhonov regularization[119]. The practice has shown that the regularization parameter helps avoid oscillations in solution associated with noise and provides smooth T_2 distributions[17]. The regularization parameter can be determined by direct and indirect methods such as Butler-Reed-Dawson, L-curve, or generalized cross-validation[17]. In the case of oil-sands, after initial analysis, the regularization parameter was determined directly and α=0.05 was found to provide the most stable solution for most samples. The density values of these samples were measured beforehand by X-ray Computed Tomography (X-ray CT) using GE 9800 CT scanner as a substitute for the density logging data.

The experimental program for Dean-Stark extraction, LF-NMR measurements, and X-ray CT density measurements is illustrated in the flowchart (Figure 27).

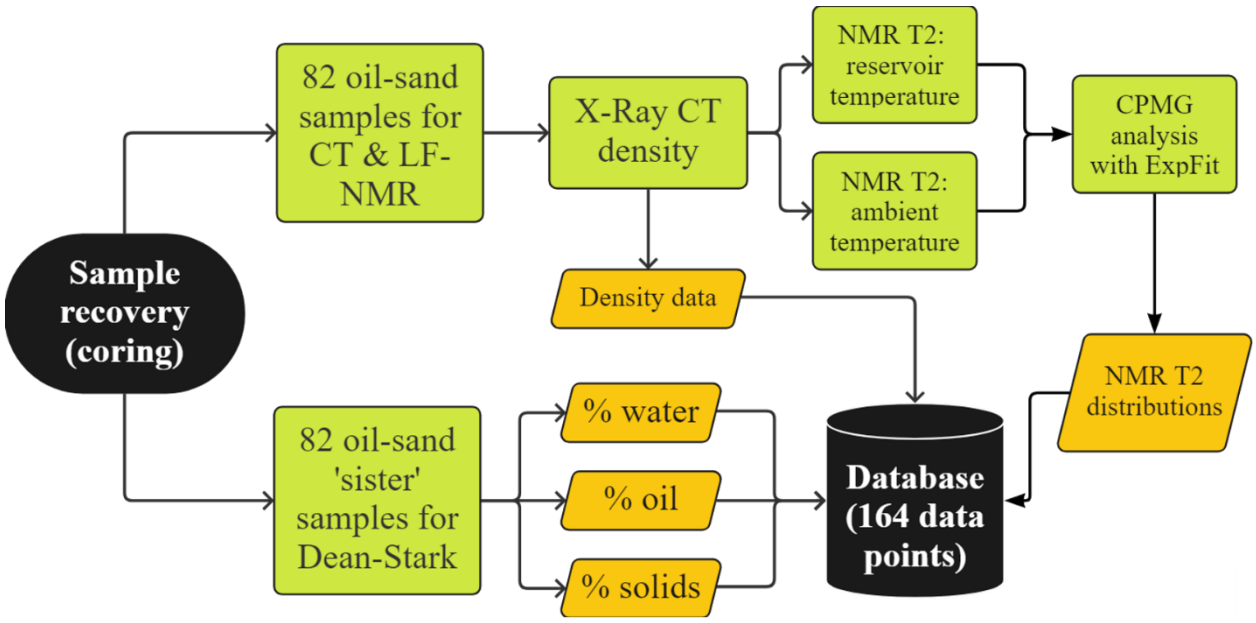


Figure 27: Flowchart representing the experimental program for oil-sands samples by X-Ray CT, LF-NMR T_2 measurements, and Dean-Stark extraction.

The final size of the dataset comprised 164 data points – 82 T_2 distributions at ambient temperature and 82 T_2 distributions at reservoir temperature, with corresponding density data and Dean-Stark sample composition. To compare the performance of machine learning models with the well-known peak deconvolution approach, the prediction of water content by LF-NMR measurements was also performed using the T_2 cutoff approach developed by Bryan et al. [59].

The data processing and model training was performed in Python 3.9 environment, while figures were produced using OriginPro 2019b software. For XGBoost model development and training, the dataset was randomly split into a training set and a test set in 0.75:0.25 proportion, respectively. To ensure the reproducible split of the data, a random split seed was fixed to *random_state = 2*. The XGBoost models were optimized using Bayesian Optimization (BO), while the training quality was evaluated by leave-one-out cross-validation (LOOCV). The forecasting performance of the models was evaluated using three error metrics

and residual distribution analysis. These steps will be discussed in detail in the following sections.

4.3.2. XGBoost model based on feature engineering (XGB-FE)

Feature engineering (FE) is a process in a part of a machine learning pipeline where domain knowledge is utilized to extract the most relevant information from the raw data. In this work, we used feature engineering to extract information from the NMR T_2-relaxation distribution. The complete FE model derivation procedure is illustrated in Figure 28.

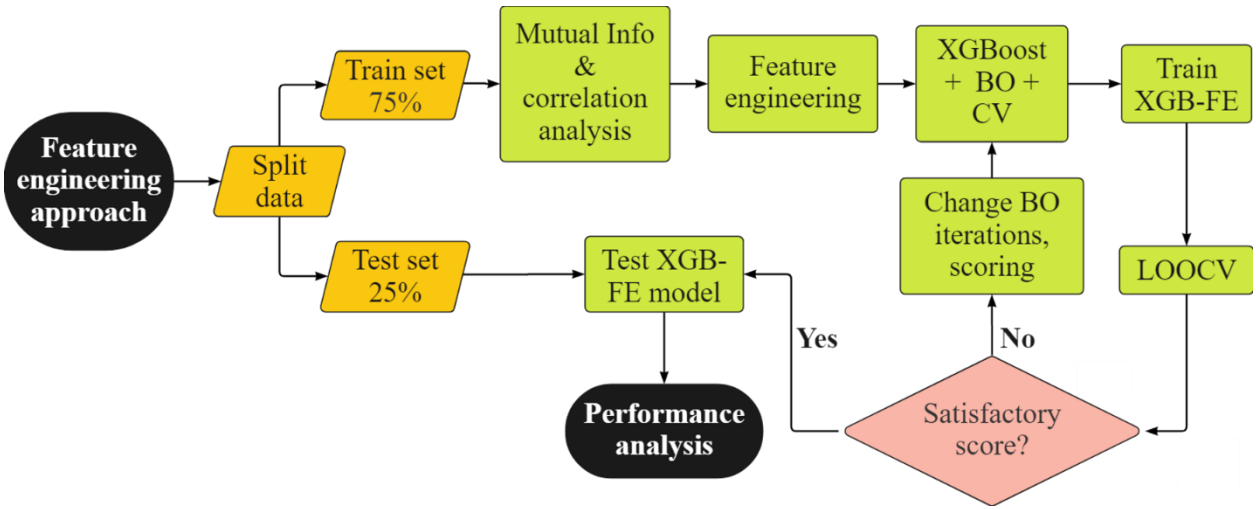


Figure 28: Flowchart for XGB-FE model development.

In petrophysics, the T_2-relaxation is regularly analyzed by geoscientists to determine fluid saturations in reservoirs, differentiate between different types of fluids, study pore size distribution, and evaluate the physicochemical properties of fluids. However, depending on the task, some parts of the T_2 distribution may have more relevance than others. In the context of studying the water content in oil-sands by NMR, we use feature engineering to reduce the amount of unnecessary information while preserving the data carrying the most information about the water in samples. A time-domain distribution of the T_2-relaxation was obtained by processing the spin-echo signal decay using a mathematical inversion.

As the T_2 distribution has a form of a continuous function, the discretization was performed for data binning which simplifies the input of data into the machine learning model. After the discretization, the T_2 data was presented as a distribution of 52 bins, with each bin corresponding to a particular T_{2i} relaxation time in milliseconds. We defined five new NMR T_2 features to limit the number of inputs.

As the T_2 distribution of relaxation times is represented on the semi-logarithmic scale, the standard parameter for representing the average T_2 relaxation is T_2 logarithmic mean (T_{2lm}):

$$T_{2lm} = exp\left[\sum \frac{A_i}{A} \cdot ln(T_{2i})\right] \tag{64}$$

where A_i is an amplitude at the corresponding T_{2i} bin, and A is a total NMR amplitude. Empirical evidence shows the strong relationship between viscosity of fluids and T_{2lm}, implying that in a water-oil system where distribution tends to be multimodal due to their different relaxation properties, the T_{2lm} provides a better measure of central tendency favoring both fast and slow relaxing parts of the distribution.

To account for the variation in T_2 distribution (i.e. narrow vs. wide peaks), the T_2 standard deviation was defined as:

$$T_{2std} = \sqrt{\frac{\sum(A_i - \mu)}{N}} \tag{65}$$

where μ is the T_2 distribution mean, and N is the number of the T_2 bins.

The T_{2p} was defined as a location of a maximum value (peak) of the T_2 amplitude on T_{2i} axis. This parameter is used in the petrophysical practice for the separation of bound and producible fluids and fluid typing since T_{2p} gives an indication of whether the largest amplitude portion of the signal corresponds to low or high T_2 values.

$$T_{2p} = \max\left(f\left(T_{2_1}\right), \ldots, f\left(T_{2_n}\right)\right) \quad (66)$$

Intelligent algorithms like XGBoost have gained popularity due to their ability to generalize complex data dependencies in large datasets and achieve state-of-the-art forecasting results. However, a small dataset is used in this study, where the overlapping of water and oil T_2 signals are likely to remain hidden or poorly represented. So, instead of allowing the algorithm to search through the whole NMR T_2 distribution, we can 'show' it where to look for the patterns and changes in the amplitude. One of the essential parts of the T_2 distribution in sandstones is the empirical clay-bound water T_2 cutoff located at 3 ms, which presents the boundary between capillary-bound and clay-bound fluids in a water-saturated core[13,120]. In order to capture the possible T_2 response of clay-bound water and monitor its signal variation with different training samples, we defined a T_2 bound fluid (T_{2bf}) interval as:

$$T_{2bf} = \sum_{0.1(ms)}^{3.0(ms)} A_i \quad (67)$$

However, this parameter cannot be used on its own to describe the changes in water content since the oil signal may also be located in the relevant interval. The true T_2 cutoff value in petrophysical practice is usually determined by performing lab tests on the saturated core samples (i.e., centrifuging), and even then, the use of a fixed or averaged T_2 cutoff value leads to the erroneous prediction of producible fluids. Instead, we attempt to obtain insights into the true T_2 cutoffs using a feature extraction technique called Mutual Information (MI) regression, based on the information entropy between variables. In classical regression analysis, statistical tests like F-test are carried out to study the degree of the linear association or continuous analysis of covariance (CANOVA) for the non-linear association between variables. However, mutual information is not 'concerned' whether the variables have apparent linear correlation or covariance of zero, and they may still be stochastically dependent. This is the case in studying the changes in the conditional probability of one variable when another is modified[121]. In other words, by using MI regression, one can measure the level of association of the

specific parts of T_2 distribution with the target output (i.e., water content by Dean-Stark), regardless of their correlation or covariance. The score is measured in natural units of information or 'nats' based on natural logarithms and powers of *e*. The MI regression was performed on the training set using a Python library sklearn.feature_selection class mutual_info_regression.

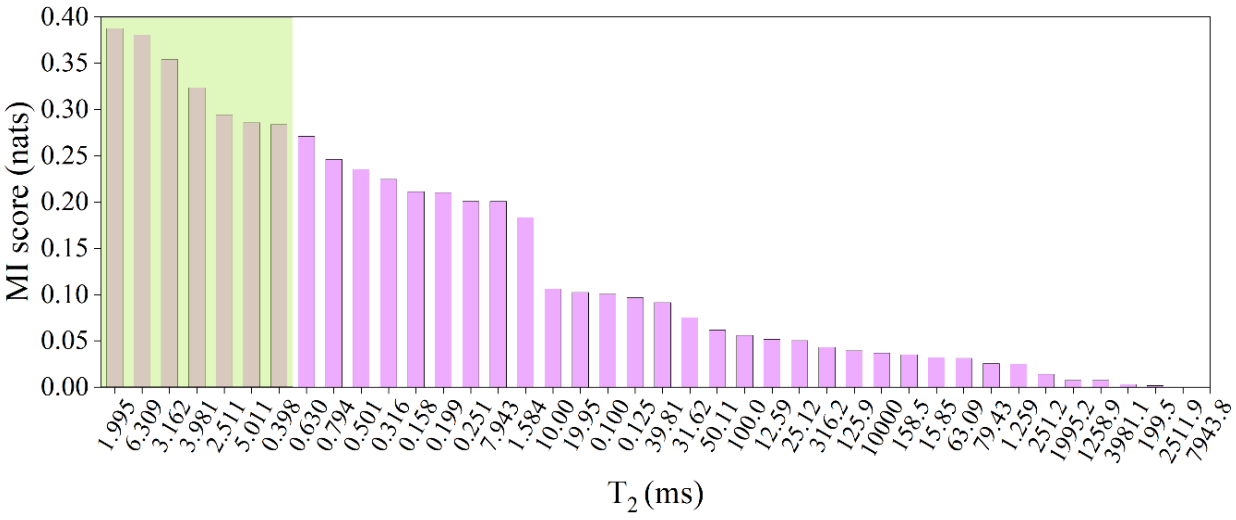


Figure 29: The mutual information regression results applied to the training set T_2 distributions of the oil-sand samples relative to the Dean-Stark water content (DS-w). The shaded area presents the continuous cluster of T_2 responses with a strong mutual association with DS-w, which were used to calculate the T_2 cutoff range parameter – T_{2cr}.

Figure 29 shows the relative mutual information scores of T_2 responses, where higher values indicate a stronger association with water content by Dean-Stark. For this dataset, the responses from 1.99-6.30 ms have the highest association with the water signal and form a continuous cluster between 10^0 and 10^1 decades along the T_2 semi-log scale, suggesting that most theoretical T_2 cutoff values lie in this interval. Therefore, the T_2 cutoff range (T_{2cr}), was defined as:

$$T_{2cr} = \sum_{1.99\,(ms)}^{6.30(ms)} A_i \qquad (68)$$

As previously mentioned, the T_2 surface relaxation and diffusive coupling play a vital role in identifying clay-bound water, which causes the overlapping of the water and oil signals. Unfortunately, to determine their contribution, a sample

recovery for the subsequent lab experiments is required. However, a common practice in well-logging is to combine NMR and bulk density logs to improve interpretation. Therefore, the bulk density was used as an additional parameter which we postulate is associated with T_2 surface relaxation and diffusive coupling.

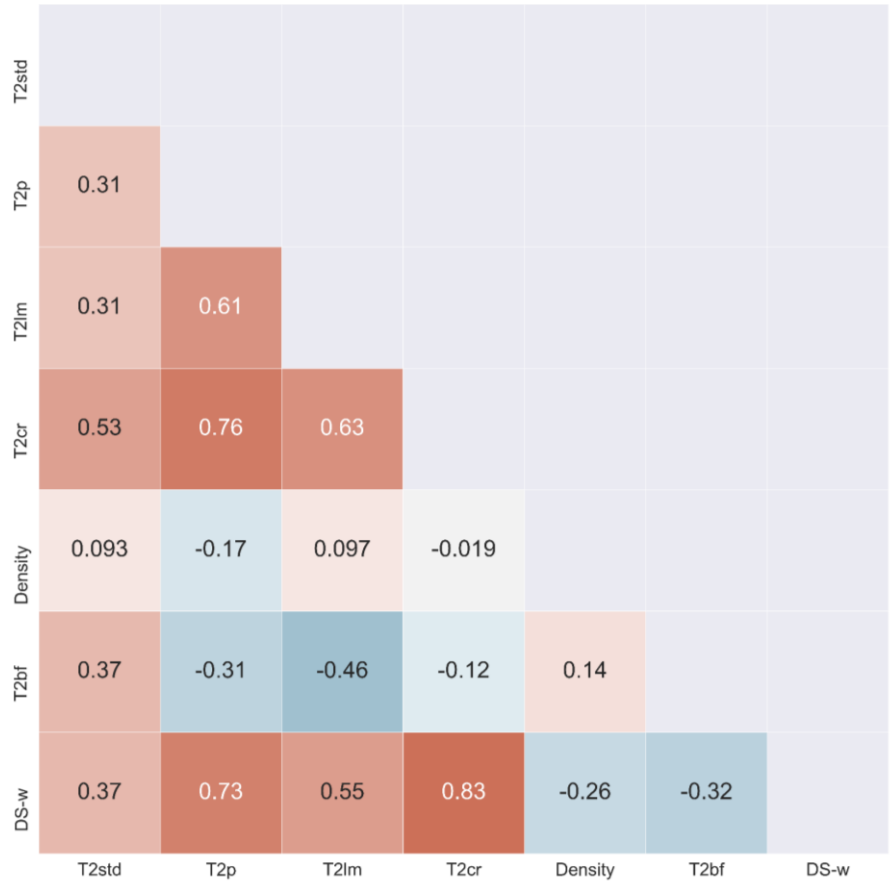


Figure 30: The diagonal correlation matrix shows the linear dependence between six input features with Dean-Stark water content (DS-w) in the training set. Scores represent Pearson's correlation coefficient and are color-coded (heatmap).

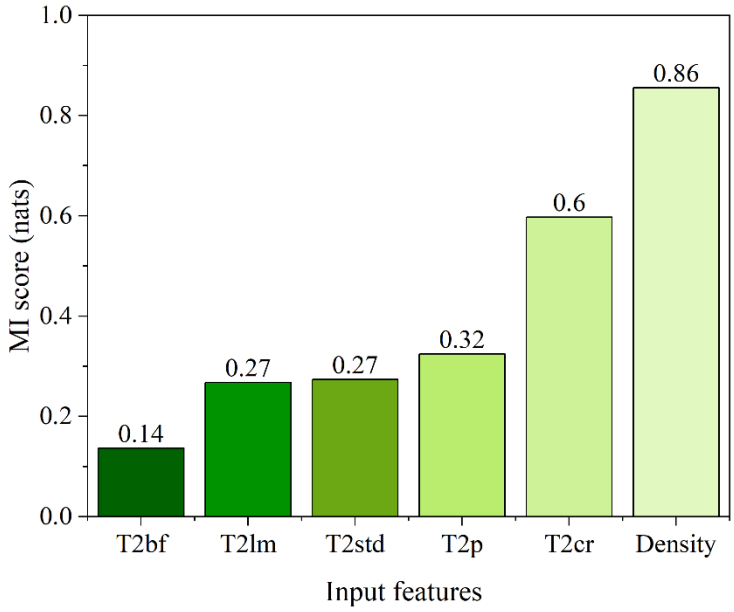


Figure 31: Mutual information regression train set scores for five NMR parameters and bulk density (input features) relative to the Dean-Stark water content (DS-w).

The correlation matrix (Figure 30) shows the linear dependence between the input features and target output. According to the Pearson score, T_2 cutoff range and T_2 peak, and T_2 logarithmic mean features exhibit the strongest positive correlation with the water content by Dean-Stark (DS-w). The T_2 standard deviation shows a moderate degree of positive correlation, while T_2 bound fluid and density features show moderate to low negative correlation with DS-w. Interestingly, when compared with mutual information scores from Figure 31, it can be observed that all features are ranked by score accordingly to Pearson's scores except for density which has the highest MI score (0.86 nats), indicating its strong stochastic (nonlinear) dependence with DS-w, thus justifying integration of density measurements into the model. Therefore, the XGB-FE model was trained using the six features presented in Table 10. The target variable is presented in Table 11 (DS-Sw).

Table 10: Descriptive statistics of six input features used for XGB-FE model training.

Statistic	T_{2std} (a.u.)	T_{2p} (ms)	T_{2lm} (ms)	T_{2cr} (a.u.)	T_{2bf} (a.u.)	ρ (kg/m^3)
Count	164	164	164	164	164	164
Mean	0.016	11.57	1.84	0.18	0.31	1626
Std	0.005	4.54	1.22	0.13	0.10	80
Min	0.006	1.00	0.32	0.00	0.10	1442
25%	0.013	8.00	0.90	0.07	0.23	1581
50%	0.016	13.00	1.61	0.18	0.30	1634
75%	0.019	15.25	2.49	0.29	0.38	1677
Max	0.032	20.00	8.59	0.50	0.54	1842

Table 11: Descriptive statistics of the target variable Dean-Stark water saturation (DS-Sw).

Statistic	DS-Sw (%)
Count	164.0
Mean	6.20
Std	2.16
Min	2.50
25%	4.20
50%	6.40
75%	8.10
Max	10.0

4.3.3. XGBoost model based on the full T_2 relaxation distribution (XGB-FS)

The second modeling method facilitates the complete sample T_2 distribution. There are two main incentives for this approach. First, the T_2 relaxation distribution contains a large amount of information about the fluids residing in the pore space, indicating that using a single or even a few features to characterize

the whole distribution may lead to significant information loss and, therefore to poor model forecasting performance[73]. By using the entire T_2 distribution, variations such as changes in slope or local minima can implicitly be used to help separate oil and water signals. Secondly, predictions generated by the full-T_2 distribution model provide a good baseline for comparison with the feature engineering and conventional deconvolution approaches. Therefore, the input features were arranged as $X = [A_1, A_2, A_3,\ldots, A_{52}, \rho_i]$, where A_i is the i-th column vector of the amplitudes at the corresponding T_{2i} bin, and ρ_i is a column vector of density measurements. The water content by Dean-Stark (DS-w) was arranged as $Y = [DS\text{-}w_1, DS\text{-}w_2,\ldots, DS\text{-}w_n]$, thus defining the dataset as $\{(X_i, Y_i)\}_{i=1}^{n}$ where n is the number of oil-sands samples. The complete XGB-FS model derivation the procedure is illustrated in Figure 32. Statistical description of input features is summarized in Table 12, and the target variable description is in Table 11.

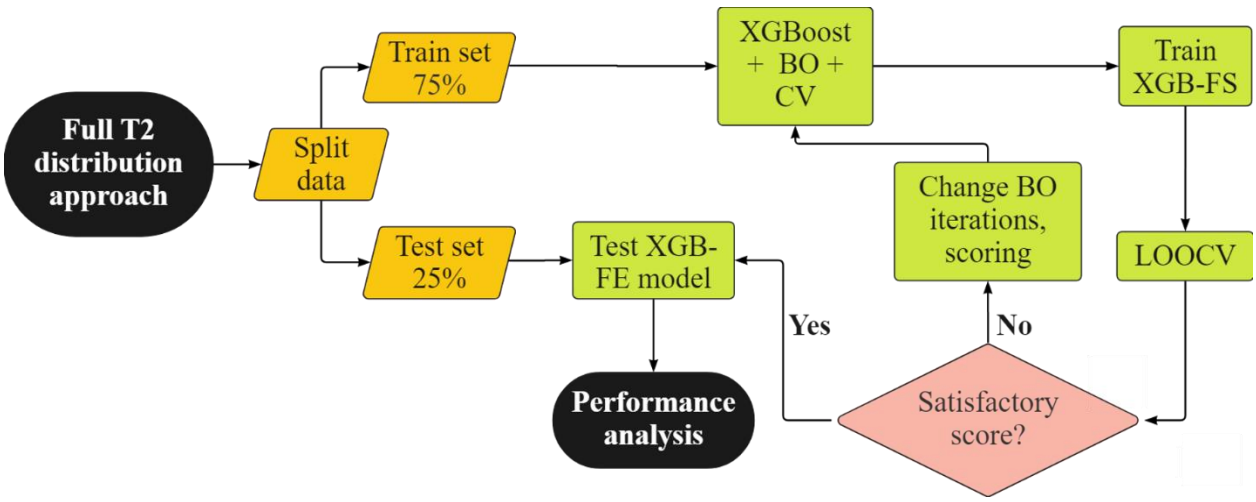


Figure 32: Flowchart for XGB-FS model development.

Table 12: Descriptive statistics of 52 discrete T_2 responses used as an input features for XGB-FS model training. Table is truncated for convenience.

	1	**2**	**...**	**51**	**52**
Statistic	$T_{2(0.10ms)}$ (a.u.)	$T_{2(0.12ms)}$ (a.u.)	**...**	$T_{2(7943.3)}$ (a.u.)	$T_{2(10,000)}$ (a.u.)
Count	164	164	**...**	164	164
Mean	0.0074	0.0103	**...**	0.0005	0.0014
Std	0.0091	0.0122	**...**	0.0011	0.0036
Min	0.0000	0.0000	**...**	0.0000	0.0000
25%	0.0000	0.0000	**...**	0.0000	0.0000
50%	0.0036	0.0068	**...**	0.0000	0.0000
75%	0.0121	0.0158	**...**	0.0002	0.0003
Max	0.0395	0.0567	**...**	0.0059	0.0192

4.3.4. MODEL OPTIMIZATION

The XGBoost algorithm contains many hyperparameters which enable fine model tuning. From the standpoint of statistical learning, the tuning usually involves the use of iterative algorithms which search for a suitable combination of hyperparameters in real-valued parameter space relative to the specified measure of model forecasting performance (e.g., mean squared error). However, as the number of parameters grows, the optimization becomes computationally expensive due to the combinatorial explosion, making the manual optimization or exhaustive grid searching techniques inefficient. In contrast, Bayesian Optimization (BO) sets a probabilistic approach where each successive combination of hyperparameters is selected based on the information obtained in the previous optimization step, thus avoiding the redundant calculations for unlikely parameter combinations and reducing the number of required iterations to reach the global minimum of the objective function.

The BO was performed in Python using scikit-optimize package class skopt.BayesSearchCV. The hyperparameters and their optimal values are presented in Table 12.

Table 12: Results of Bayesian Optimization on training set with 5-fold cross-validation, for XGB-FS and XGB-FE models.

XGBoost hyperparameters	**Search range**	**XGB-FS optimal**	**XGB-FE optimal**
n_estimators	[50-1000]	[650]	[300]
learning_rate	[0.004-0.1]	[0.008]	[0.053]
subsample	[0.7-1.0]	[0.7]	[0.6]
max_depth	[6-12]	[8]	[7]
objective	['squared_error', 'pseudo_huber']	['pseudo_huber']	['squared_error']
grow_policy	['depthwise', 'lossguide']	['lossguide']	['lossguide']
booster	['gbtree', 'dart']	['gbtree']	['gbtree']

4.3.5. PERFORMANCE METRICS AND MODEL VALIDATION

The forecasting performance of the models was evaluated using three performance metrics, including coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The R^2 is the positively oriented metric used in regression for representing the amount of model variance, and how well the model predictions generalize the observations. However, R^2 alone does not provide information on prediction errors. The RMSE is another regularly employed error metric, used alongside R^2, but under the assumption that residuals follow the normal distribution[122]. As a result of the heavy penalization of larger residuals, the RMSE is a convenient metric for revealing the differences in performance between multiple models with normally distributed residuals. At the same time, large residuals can cause the inflation of the RMSE score, which is why MAE can be used for additional evaluation. The MAE measures the mean magnitude of model prediction errors, but in contrast to RMSE, the errors are not

squared. Therefore, RMSE scores are generally higher than MAE scores. These two metrics can be used together to estimate the variation in errors. Recall that both RMSE and MAE are negatively oriented scores (smaller values are preferred) expressed in %DS-w.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (69)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (70)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (71)$$

where y_i is predicted %DS-w, $\hat{y}_i$ is observed %DS-w, $\bar{y}$ the sample mean, and n presents the number of samples.

The further model performance evaluation and validation were performed using leave-one-out cross-validation (LOOCV) due to its convenience for use on small datasets (Figure 33). Cross-validation is a resampling method in which the sample subsets are drawn repeatedly from the training set, followed by model refitting for each subset, thus providing information on model fitting variability. In LOOCV, the samples are drawn for one observation at a time, while the rest of the data is used for model training. Therefore, this process has a number of iterations equal to the number of samples, making it computationally expensive for large datasets. In addition to LOOCV, the permutation tests were conducted to assess the significance of 5-fold cross-validated model prediction scores with 150 random permutations. This enabled the evaluation of the statistical significance of model predictions and their inputs by a permutation test P-value.

4.4. RESULTS

In this section, the performance of three models is presented, including the XGB-FE model based on the XGBoost algorithm with feature engineering, the XGB-FS

model based on the XGBoost algorithm using the whole sample T_2 distribution, and a peak deconvolution approach (Bryan et al[7]). To assess the model performance in more detail, residual plots (Figures 33-2, 33-5, and 33-8) and quantile-quantile plots (Figures 33-3, 33-6, and 33-9) are used for the analysis of the residual normality, and model variance and bias. All results are summarized in Figures 33-34.

Analysis and comparison of error statistics, cross-plots, and distribution of residuals indicate that the XGB-FE model achieves the highest accuracy and generalization ability in the prediction of water content in oil-sand samples. Figure 33-1 shows that apart from slight overprediction in the 3-5% DS-w range, all XGB-FE predictions spread along the x=y line with low variance, achieving the highest R^2 score in the study (R^2=0.90). Figure 33-2 shows the constant low variance of the residuals, indicating that the model inputs capture variation in the data correctly. Finally, the Q-Q plot (Figure 33-3) confirms the residual normality and thereby the underlying assumption that XGB-FE model residuals follow the normal distribution (low bias, low variance). Finally, the XGB-FE model achieves 1.5-3 times lower RMSE and MAE scores compared to the XGB-FS and Bryan et al.[7] models indicating the best generalization ability of the three.

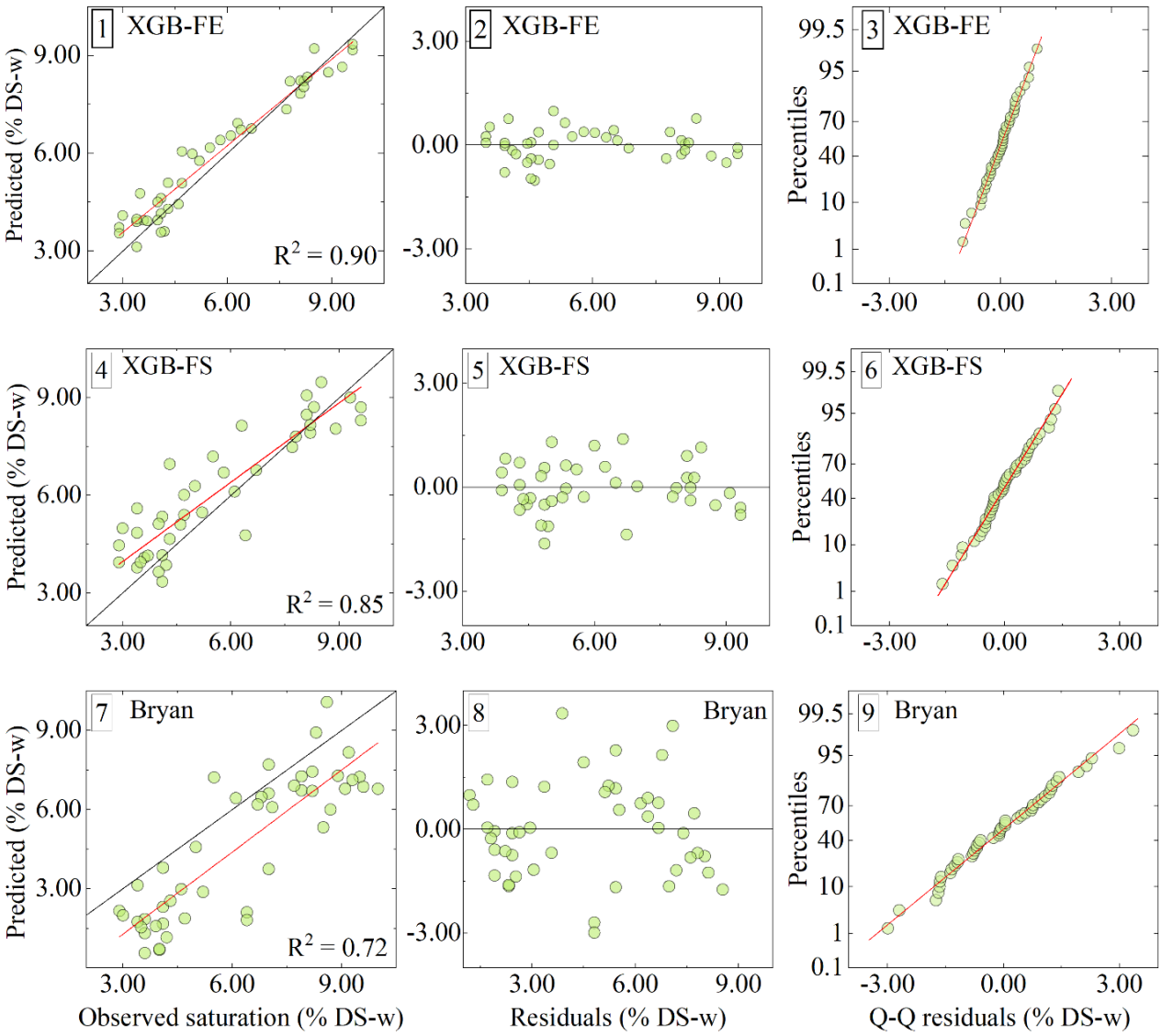


Figure 33: Residual analysis from test set predictions by XGB-FE, XGB-FS, and Bryan et al.[7]. Cross-plots between the model test set predictions and observed saturation in %DS-w (1, 4, 7), distribution of regular residuals (2, 5, 8), and quantile-quantile plots for comparing distributions between test predictions and observations and evaluating normality of residuals (3, 6, 9).

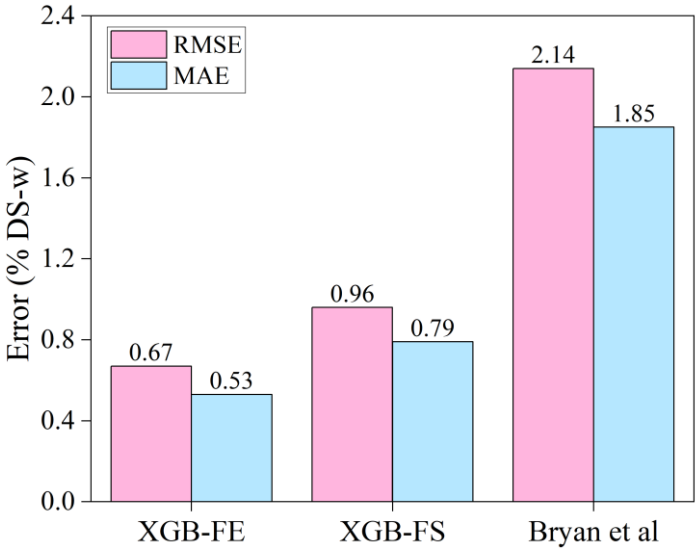


Figure 34: Comparison of RMSE and MAE test prediction scores for the three models (*'random_state=2'*).

As for the XGB-FS model predictions, Figures 33-4, 33-5, and 33-6 show a similar residual distribution to XGB-FE (normality and bias). However, the residual variance is increased but constant, therefore achieving a somewhat lower R^2 score (R^2=0.85) and 1.5 times higher RMSE and MAE than XGB-FE. From Figure 33-7, it can be observed that the Bryan et al.[7] model generally tends to underpredict the water content in samples. In addition, Figures 33-8 and 33-9 show inflated but the constant variance in the distribution of residuals, while residual normality still holds with some local perturbing. As a result, Bryan et al.[7] model RMSE and MAE scores are the highest (Figure 34).

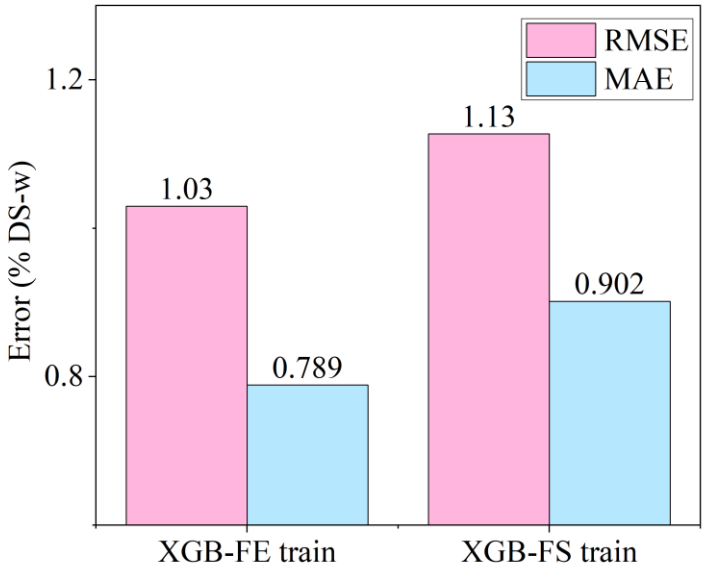


Figure 35: Leave-one-out cross-validation (LOOCV) scores for XGB-FS and XGB-FE machine learning models for the training set with fixed random split seed *'random_state=2'.* Note y-axis was truncated for convenience.

4.5. Discussion

The two machine learning models in this study were designed to test two principal hypotheses. First, to confirm that the integration of density measurements into the machine learning models can help separate the contribution from overlapping oil and water signals. Second, to show that the derivation of new LF-NMR T_2 features can improve the generalization ability of the machine learning model to the degree that can enable the accurate forecasting of water content by LF-NMR in oil wells *(in-situ).*

Table 13: Comparison of XGB-FS model test set prediction scores with and without *bulk density* parameter.

XGB-FS		
Statistic	wo/ Density	w/ Density
RMSE (%DS-w)	1.08	**0.96**
MAE (%DS-w)	0.86	**0.79**
R^2	0.72	**0.85**

Table 14: Comparison of XGB-FE model test set prediction scores with and without *bulk density* parameter.

XGB-FE		
Statistic	wo/ Density	w/ Density
RMSE (%DS-w)	0.91	**0.67**
MAE (%DS-w)	0.73	**0.53**
R^2	0.81	**0.90**

Bulk density measurements are used together with LF-NMR measurements in petrophysical practice to improve the interpretation of well logs[28,123]. LF-NMR measures the response of the fluids in the rock pore space, therefore carrying information about the fluids and pore size distribution of the rock. On the other hand, density logging equipment measures the response of the solids (rock matrix) together with fluids. The two are related in terms of T_2 surface relaxation, which depends on the rock pore to surface ratio with the fluid volume and the diffusive-coupling effect. This dependence can also be observed from the prediction test scores of the XGB-FS and XGB-FE models with and without bulk density as one of the model inputs. Prediction scores from Tables 13 and 14 indicate that models achieve better scores with the integration of bulk density,

therefore confirming the relationship between Dean-Stark water content and density discovered by mutual information regression.

To affirm the second hypothesis: when XGB-FS and XGB-FE are compared (Figure 33), it can be observed that XGB-FE achieves better performance, especially in terms of prediction variance. The XGB-FE variance reduction supports two premises. First, the engineered features properly capture all the relevant information from the T_2 distribution, indicating negligible information loss. Secondly, in the feature engineering case, the XGBoost algorithm generalized the variability in the data with the output more effectively, suggesting that for smaller datasets, the appropriate feature engineering enables the XGBoost algorithm to discover dependencies within the data more effectively than for a large number of raw information (53 features in case of XGB-FS), due to the high dimensionality. In other words, the new features contain all the relevant parts of the T_2 distribution compressed into a few values, which reduces the XGB-FE model complexity and enables better generalization of the relationship between inputs and a target variable (DS-Sw).

According to Figure 31, along with a bulk density (MI=0.86 nats), the T_2 cutoff range feature ranks second by MI score (0.60 nats), indicating the variability of the sum of T_2 responses between 1.99 – 6.30 ms has a strong relationship with water signal. The location of the T_2 peak (T_{2p}), T_2 standard deviation of the spectrum (T_{2std}), and T_2 logarithmic mean (T_{2lm}) achieve similar MI scores (0.27 and 0.30 nats, respectively), signifying that these features alone do not capture enough information about the water content. Finally, the sum of T_2 responses representative of the empirical clay-bound water part of the T_2 distribution (0.1-3.0 ms) shows the least association with the target (DS-w). Although these features alone cannot explain variance in data effectively, their mutual interaction can improve it. Since MI does not consider this mutual interaction between features relative to the target output, the correlation matrix can be used. For instance, Figure 30 shows that the T_{2cr} vs. T_{2p} and T_{2cr} vs. T_{2lm} have a strong positive

correlation (0.76 and 0.63, respectively), while T_{2bf} vs. T_{2lm} have a moderate negative correlation. These interactions are likely to be generalized in the XGB-FE model training process, thus explaining improved performance. Furthermore, the permutation test score of XGB-FE using 150 permutations generated a P-value of 0.001, compared to the XGB-FS P-value of 0.007. In both cases, the P-value is well below 0.05, showing a very low likelihood of obtaining such model performance purely by chance.

As for the deconvolution approach (Bryan et al.[7]), the main challenge lies in separating overlapping fluid contributions in T_2 distribution. Even under the assumption that T_2 cutoff and deconvolution are performed such that a precise distinction between fluid signals is possible, the issue of how to associate the amplitudes with respect to mass persists. This approach leads to underprediction of water content for the given dataset, indicating that the oil and water signals are not sufficiently separated. The machine learning-based approach is more robust because it removes the necessity to manually identify peak separation and the errors associated with visually separating oil and water signals, especially in the case of NMR measurements acquired at low SNR.

It is essential to point out the limitations of these models, which are related to reservoir lithology (a) and SNR of the measurements (b):

(a) The models presented in this study were derived for the oil-sands reservoir, so their application is limited only to similar reservoir types. However, the presented approach can be extended for use in other types of oil reservoirs under the assumption that a sufficiently large amount of observations is available.

(b) The SNR achieved by the benchtop LF-NMR relaxometers can be up to 30 times higher than the SNR values obtained using well-logging tools. In this study, the NMR signal-to-noise ratio was, on average 20, which can still be considered high relative to the logging tools where the SNR of 3-5 is considered satisfactory[124]. Although the recent research

demonstrated that the XGBoost algorithm is sufficiently robust even with noisy data[125], an additional validation using the data obtained by the LF-NMR logging tools would be desirable. It is worth noting that, in lower SNR samples, the deconvolution approach will be even more challenging, and the value of using just the general properties of the T_2 distribution and XGBoost may be even further enhanced.

The XGB-FS method which uses all T_2 discrete responses as an input, is not optimal approach as it has been shown, due to the high dimensionality of inputs. It is worth noting that dimensionality reduction techniques could be used such as principal component analysis, non-negative matrix factorization, linear discriminant analysis and other methods, which would enable construction of less complex models. This is planned to be done as a follow-up study. In addition, the procedures for NMR measurements with a controlled saturation and desaturation of samples, similar to those reported in recent literature[66], would enable deeper sensitivity analysis of the features derived in this work and further improvement XGB-FE model. In such a setup, the Dean-Stark measurements could be replaced by the more cost and time-effective mass-volume measurements, ultimately allowing the collection of a larger database, at which point the application of artificial neural networks (ANNs) would be possible.

It is also worth noting that logging equipment configuration can be substantially different from desktop NMR relaxometers, which may cause inconsistencies between NMR T_2 distributions obtained in the lab and the field. This can cause the variable performance of proposed NMR data-driven model, which is why the parameters of the NMR logging device, such as TW, TE and number of trains, should be relatively consistent to the values reported in this study.

4.6. SUMMARY

This study presents the approach which integrates extreme gradient boosting with LF-NMR measurements and bulk density data for the water saturation determination in oil-sands. Two models were developed using full NMR T_2 distribution (XGB-FS), and feature engineering (XGB-FE). It is concluded that;

- Feature extraction methods such as mutual information regression can effectively select the most relevant information from NMR T_2 distribution.
- Integrating bulk density data as a model input notably improves the XGB-FS and XGB-FE forecasting performance.
- XGB-FE achieved RMSE = 0.67%, MAE = 0.53% and R^2 = 0.90 in predicting relative water content by Dean-Stark, a substantial improvement compared to deconvolution method.

These results suggest that the XGB-FE model can be extended for the improved *in-situ* water saturation determination.

Chapter 5 CONCLUSIONS

This thesis focuses on nonlinear problems in petrophysical logging, such as characterization of bitumen and heavy oil viscosity and water saturation quantification, on the example of Canadian oil-sands by LF-NMR relaxometry data. We employed various statistical and machine learning tools to model the relationship between the NMR outputs and experimental data, which reduced uncertainties associated with data interpretation and enabled us to capitalize on new, previously unknown relationships. In conclusion, we found that:

- Heavy oil and bitumen viscosity can be analytically approximated from T_2 relaxation data, even in high-temperature conditions, by integrating T_2 logarithmic mean, relative hydrogen index per unit volume (RHI_v), and power-law corrected T_{2lm}. The statistical scores of the new analytical model demonstrate a better generalization ability compared to all approaches in the literature to date.
- Machine learning-based predictive modeling of oil viscosity (gradient boosting and support vectors in particular) using only one NMR parameter (T_{2lm}) along with suitable feature engineering can provide highly accurate predictions of oil viscosity. These models work well even for a set of chemically diverse light, heavy and extra-heavy oils. It also overcomes issues related to NMR hardware limitation (finite echo spacing), magnetization loss at high temperatures (Curie effect), and additional costs and uncertainties associated with determining RHI_v.
- The Extreme Gradient Boosting algorithm can be utilized for improved water and oil saturation evaluation with only T_2 relaxation and bulk density data required as an input. This approach bypasses issues stemming from a poor understanding of dominating T_2 relaxation processes in micro and macro-pores, diffusive coupling, and consequential oil and water signal overlapping. It also does not require determining the T_2 cutoff value and associated laboratory tests.

Even though a large number of experimental data was used for this work (hundreds of observations), the obtained datasets are considered small from the machine learning point of view (tens of thousands of observations). This work demonstrates that even relatively small datasets such as those characteristic for petrophysical laboratory tests can be used for machine learning modeling by performing feature engineering, given that understanding the problem and associated processes is sufficiently understood.

This work also lays the foundations for further research. The following recommendations can be made:

1. Integrating bulk density measurements into the machine learning models for viscosity determination to perform additional validation for *in-situ* applications.
2. Additional LF-NMR and bulk density measurements (dataset expansion) would further improve the understanding of T_2 surface relaxation and diffusive coupling in connected micro- and macropores. This would also enable us to use artificial neural networks with more success.
3. Performing a deeper study on the effect of bulk density data on the water saturation prediction by LF-NMR measurements and studying the effects of noise. Another interesting study would be the application of dimensionality reduction methods on NMR T_2 distribution.
4. The combination of other conventional logging data with NMR data for machine learning modeling shows excellent potential for other *in-situ* applications such as rock wettability characterization or quantification of saturates, aromatics, resins, and asphaltenes in hydrocarbons.

Appendix A

A.1 Appendix to Chapter 4 – Prediction performance of other machine learning models

A 1.1 Summary

Machine learning models tested in this section include Random Forests (RF), Gradient Boosted Regression Trees (GBRT), Gaussian Process (GP), Support Vector Regression (SVR), and Elastic Net (EL). Also, XGBoost was retrained by constraining 'n_estimators' and 'max_depth' hyperparameters to a smaller range during optimization to reduce the complexity of the model. It should be noted that all machine learning models were optimized using the Bayesian Optimization approach (scikit-optimize package class skopt.BayesSearchCV).

Lastly, as discussed previously, using full T_2 distribution for model training is not viable due to the high dimensionality. Therefore, only the approach with engineered features and feature extraction was applied.

A 1.2 Feature scaling

Feature scaling is a regular step in the machine learning pipeline used to normalize the distribution of input features to a common scale. Although tree-based methods such as XGBoost, do not require normalization, other intelligent algorithms may not work correctly without previous normalization. To predict water saturation by LF-NMR and bulk density data, I selected the Robust Scaler method from sklearn.preprocessing, which considers the feature quantile range to remove the scale and median of the data. Given the set of inputs, the algorithm calculates statistics for each input. These statistics are applied independently to each input for centering and scaling. An added benefit of Robust Scaler is that it effectively works on datasets with outliers, which is particularly useful when

working with small datasets. Descriptive statistics of scaled inputs and targets are presented in Tables 15-17.

Table 15: Training set descriptive statistics of six input features after scaling.

Statistic	$\mathbf{T_{2std}}$	$\mathbf{T_{2p}}$	$\mathbf{T_{2lm}}$	$\mathbf{T_{2cr}}$	$\mathbf{T_{2bf}}$	**ρ**
Count	123	123	123	123	123	123
Mean	0,05	-0,14	0,11	0,03	-0,07	0,07
Std	0,75	0,55	0,79	0,57	0,79	0,67
Min	-1,47	-1,50	-0,88	-0,74	-1,95	-1,38
25%	-0,47	-0,63	-0,49	-0,47	-0,54	-0,45
50%	0,00	0,00	0,00	0,00	0,00	0,00
75%	0,53	0,38	0,51	0,53	0,46	0,55
Max	2,31	0,88	4,44	1,56	2,07	1,80

Table 16: Test set descriptive statistics of six input features after scaling.

Statistic	$\mathbf{T_{2std}}$	$\mathbf{T_{2p}}$	$\mathbf{T_{2lm}}$	$\mathbf{T_{2cr}}$	$\mathbf{T_{2bf}}$	**ρ**
Count	41,00	41,00	41,00	41,00	41,00	41,00
Mean	-0,11	-0,30	0,06	-0,10	-0,19	0,26
Std	0,58	0,59	0,78	0,46	0,85	0,77
Min	-1,22	-1,38	-0,82	-0,74	-1,91	-1,36
25%	-0,47	-0,75	-0,59	-0,49	-0,75	-0,41
50%	-0,11	-0,25	-0,11	-0,14	-0,03	0,34
75%	0,16	0,25	0,62	0,29	0,23	0,86
Max	1,40	0,75	2,48	0,98	1,84	1,48

Table 17: Train and test set descriptive statistics of the target feature (Dean-Stark water saturation) after scaling.

Target-Train set	
Statistic	DS-Sw
Count	123,00
Mean	-0,08
Std	0,54
Min	-1,05
25%	-0,63
50%	0,00
75%	0,38
Max	0,83

Target-Test set	
Statistic	DS-Sw
Count	41,00
Mean	-0,25
Std	0,53
Min	-0,95
25%	-0,68
50%	-0,43
75%	0,28
Max	0,73

A 1.3 Random forests

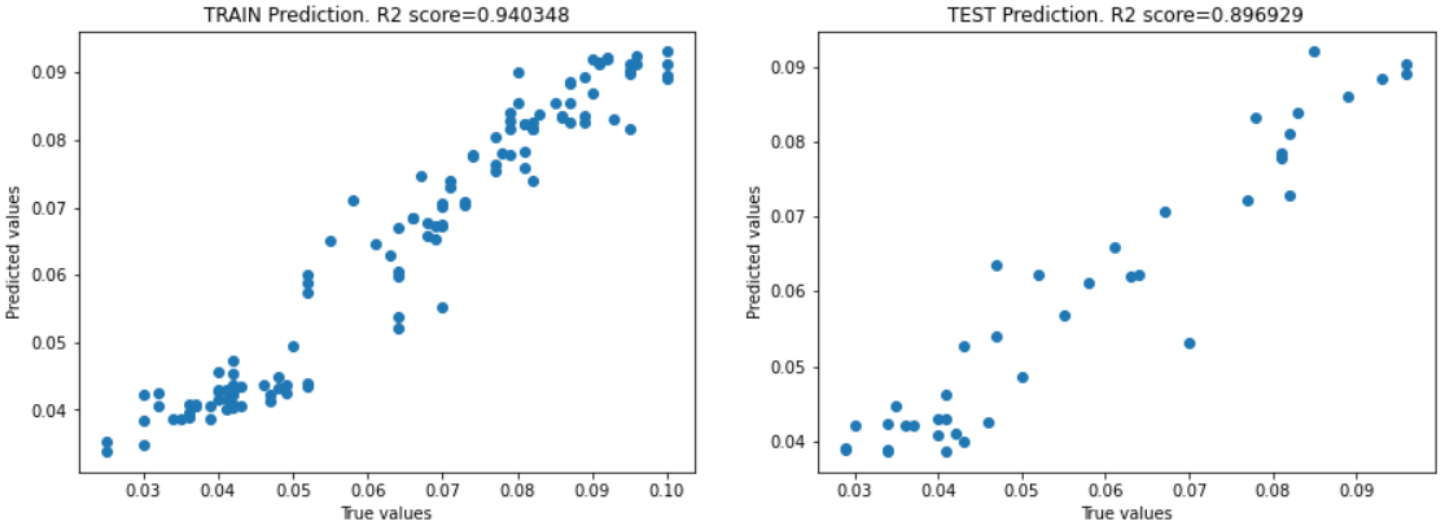


Figure 36: Cross-plot of Random Forest model train and test predictions

Table 18: Hyperparameters for Random Forest model by Bayesian Optimization

Random forest hyperparameters	Search range	BO search optimal
n_estimators	[20-200]	[133]
criterion	['mse', 'mae']	['mse']
max_depth	[2-7]	[5]
max_features	['sqrt', 'log2', 'auto']	['auto']
min_samples_leaf	[1-4]	[1]
min_samples_split	[2-8]	[2]

A 1.4 Gradient Boosting Regression Trees

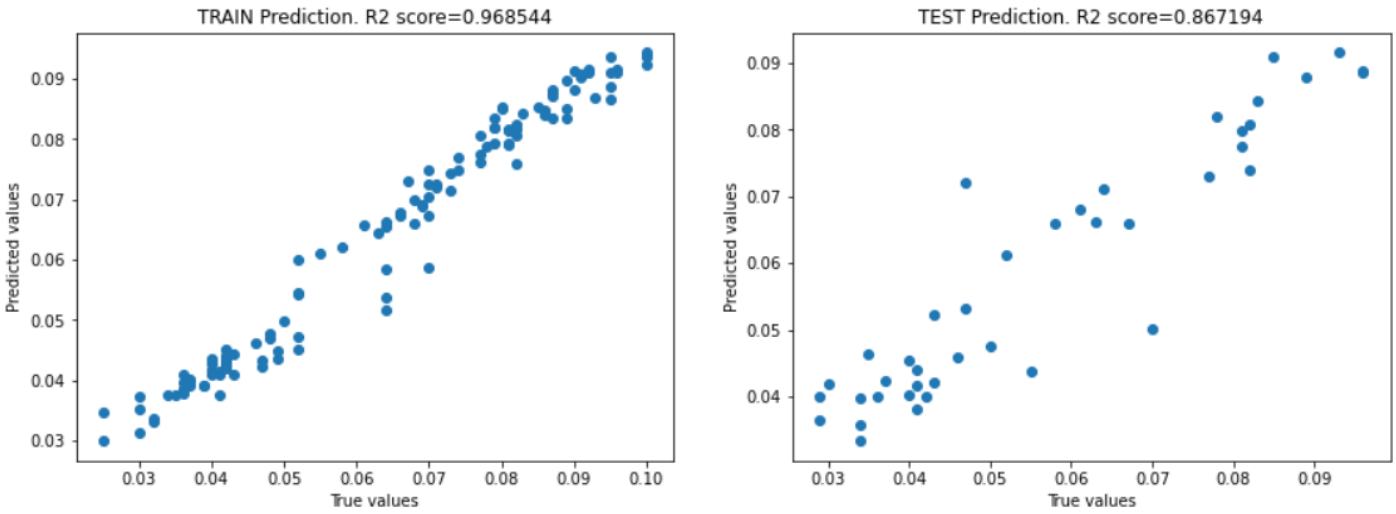


Figure 37: Cross-plot of Gradient Boosting Regression model train and test predictions

Table 19: Hyperparameters for Gradient Boosting Regression model by Bayesian Optimization

GBRT hyperparameters	**Search range**	**BO search optimal**
n_estimators	[20-200]	[128]
criterion	['mse', 'mae']	[mse]
max_depth	[2-6]	[3]
loss	['mse', 'ls', 'mae', 'lad']	['ls']
subsample	[0.6-1]	[1]
min_samples_leaf	[1-4]	[2]
min_samples_split	[2-6]	[5]

A 1.5 Gaussian Process Regression

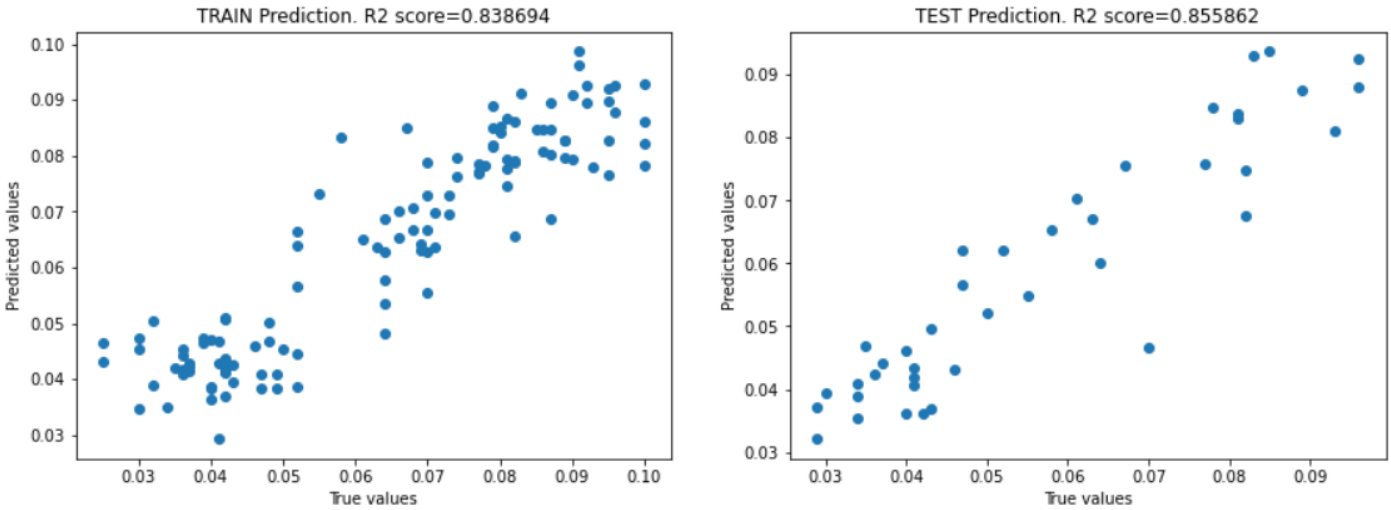


Figure 38: Cross-plot of Gaussian Process model train and test predictions

Table 20: Hyperparameters for Gaussian Process model by Bayesian Optimization

Gaussian Process hyperparameters	**Search range**	**BO search optimal**
kernel*	['RBF', 'DotProduct', 'Matern', 'RationalQuadratic', 'ConstantKernel']	['RBF']
alpha	[$1e^{-5}$ - $1e^{-9}$]	[$1e^{-7}$]

*each kernel was run with the addition of WhiteKernel (noise_level=0.1)

A 1.6 Elastic Net

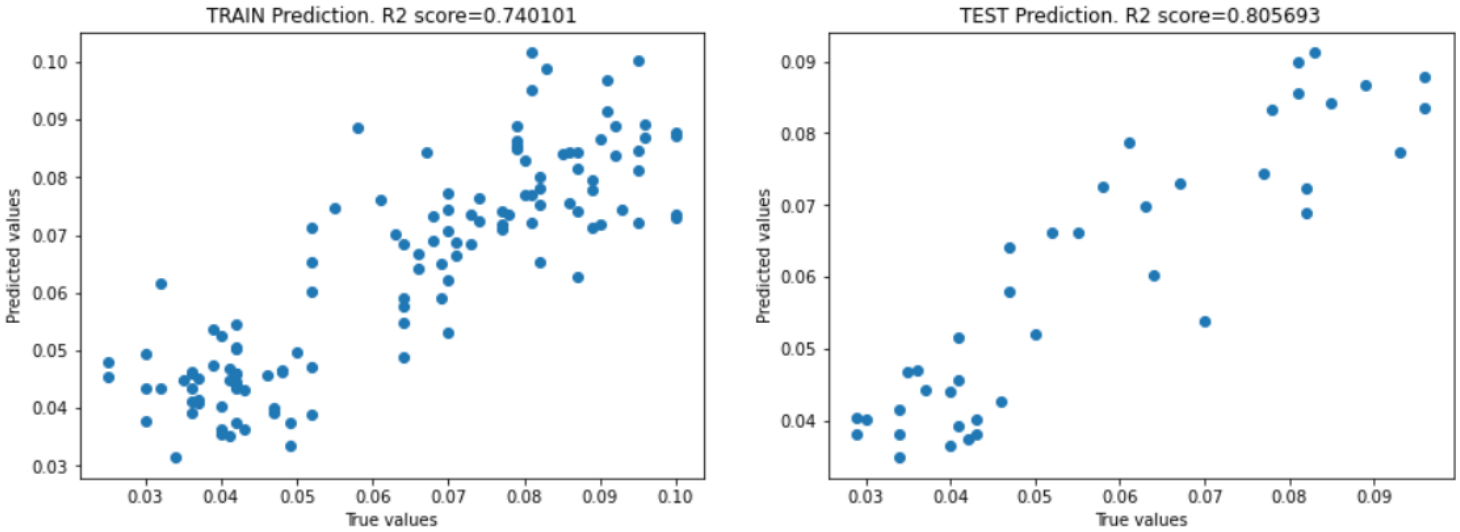


Figure 39: Cross-plot of Elastic Net model train and test predictions

Table 20: Hyperparameters for Elastic Net model by Bayesian Optimization

Elastic Net hyperparameters	**Search range**	**BO search optimal**
alpha	[0-1]	[0.003]
l1_ratio	[0-1]	[1]

A 1.7 Support Vector Regression

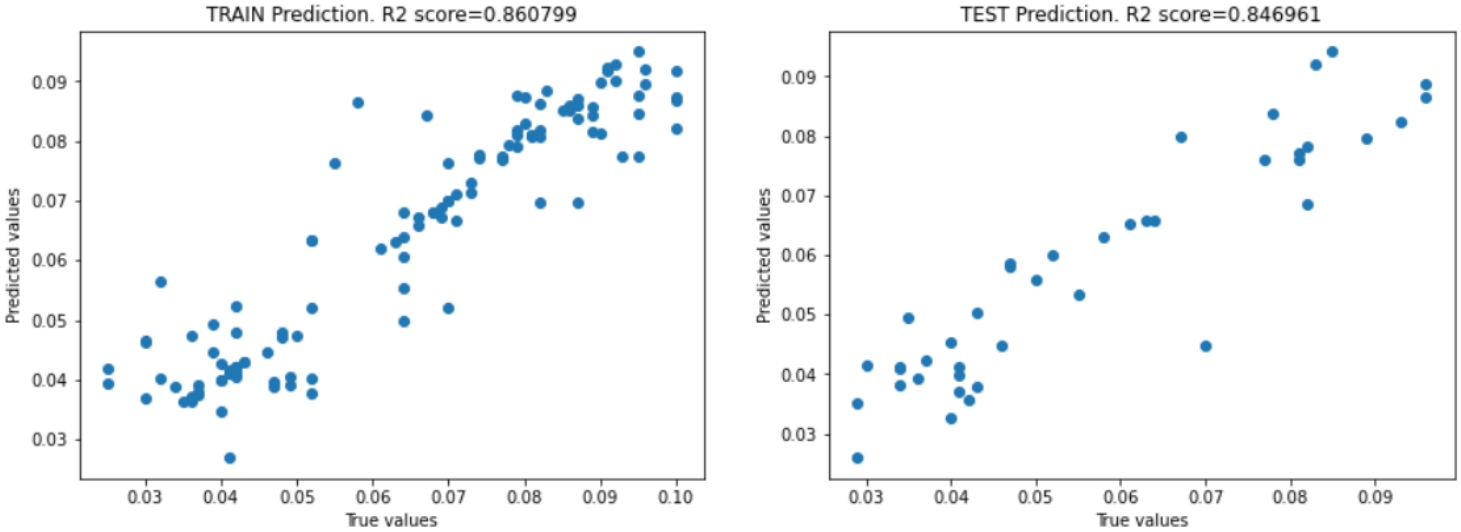


Figure 40: Cross-plot of Support Vector Regression model train and test predictions

Table 21: Hyperparameters for Support Vector Regression model by Bayesian Optimization

SVR hyperparameters	Search range	BO search optimal
alpha	[0-1]	[0.003]
l1_ratio	[0-1]	[1]

A 1.8 XGBoost (constrained)

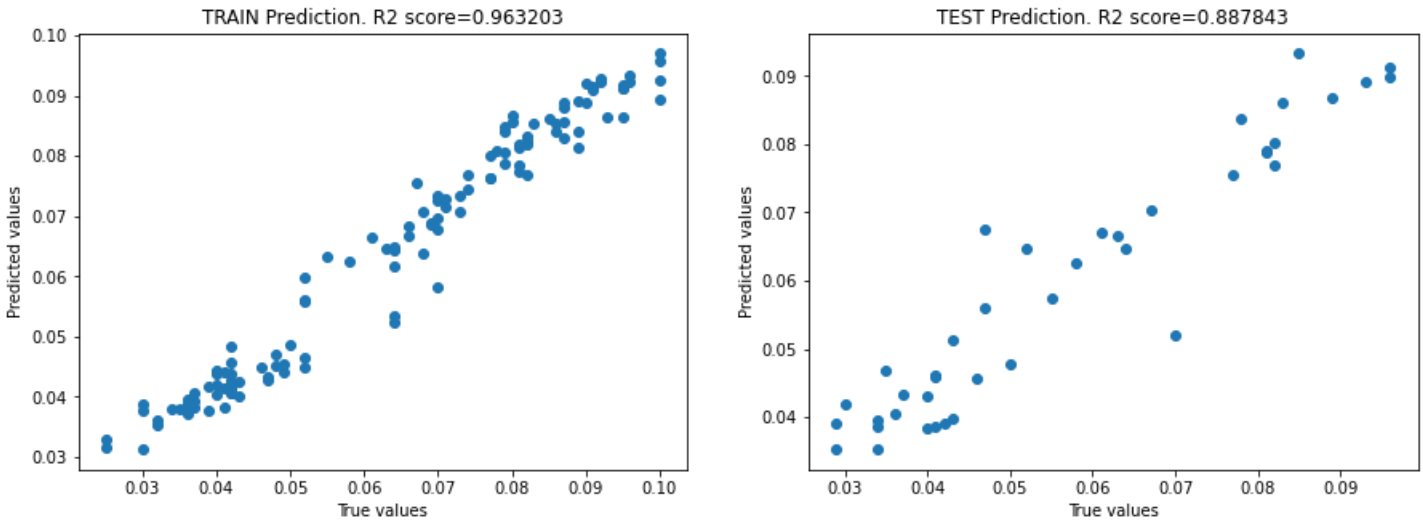


Figure 41: Cross-plot of Support Vector Regression model train and test predictions

Table 22: Hyperparameters for XGBoost model by Bayesian Optimization

XGBoost hyperparameters	**Search range**	**XGB optimal**
n_estimators	[50-250]	[**132**]
learning_rate	[0.02-0.1]	[0.045]
subsample	[0.6-1.0]	[0.6]
max_depth	[2-6]	[**3**]
objective	['squared_error', 'pseudo_huber']	['squared_error']
grow_policy	['depthwise', 'lossguide']	['lossguide']
booster	['gbtree', 'dart']	['dart']

A 1.9 Results summary

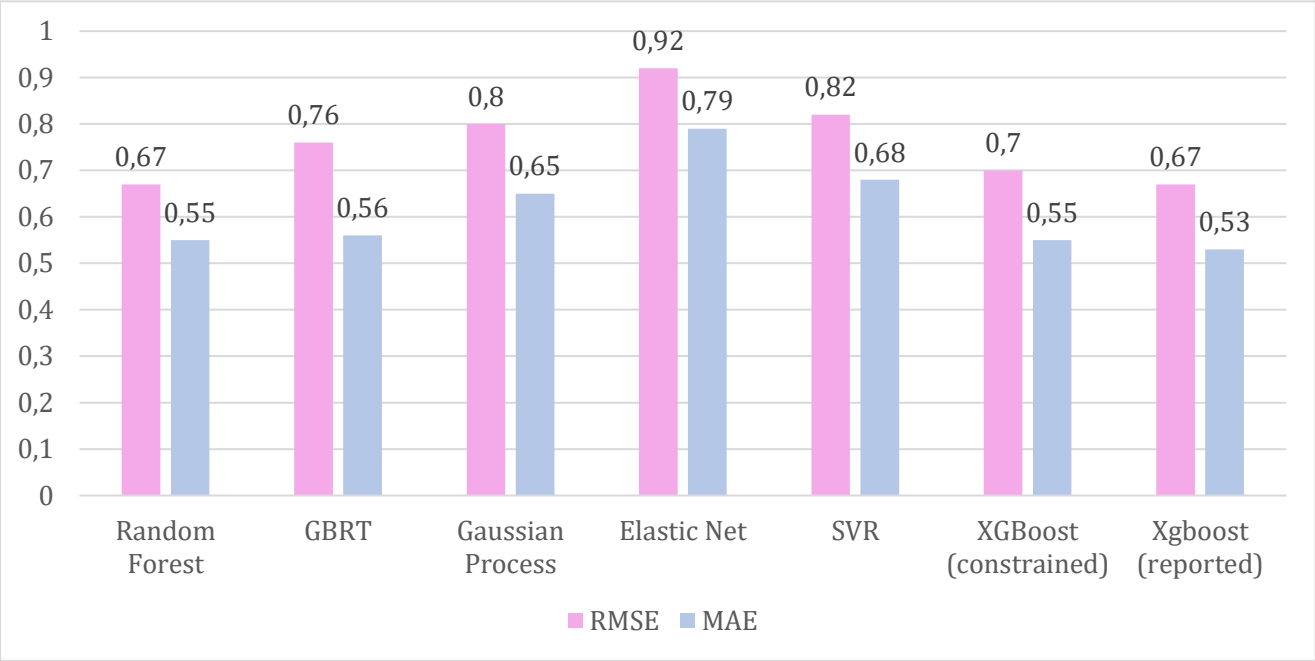


Figure 42: Comparison of RMSE and MAE test set prediction scores for seven models.

Table 23: Test set model prediction scores

Model	Metric		
	RMSE	**MAE**	**R^2**
Random Forest	**0,67**	0,55	**0.90**
GBRT	0,76	0,56	0.87
Gaussian Process	0,80	0,65	0.86
Elastic Net	0,92	0,79	0.81
SVR	0,82	0,68	0.85
XGBoost (constrained)	0,70	0,55	0.89
XGBoost (reported)	**0,67**	**0,53**	**0.90**

Appendix B

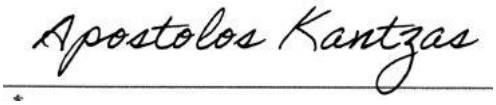
Attributions

In-situ heavy oil viscosity prediction at high temperatures using low-field NMR relaxometry and nonlinear least squares

Authors: Strahinja Markovic[a*], Jonathan L. Bryan[b], Aman Turakhanov[a], Alexey Cheremisin[a], Apostolos Kantzas[b], Sudarshan A. Mehta[b].

[a] Skolkovo Institute of Science and Technology, Moscow, Russian Federation
[b] University of Calgary, Calgary, Alberta, Canada

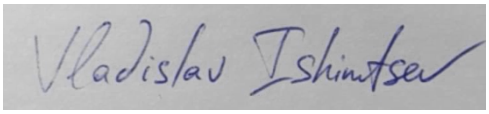
Name of Co-authors	Conception and design	Acquisition of data & method	Data conditioning & manipulation	Analysis & statistical method	Interpretation & discussion & editing	Final approval
Strahinja Markovic	☒	☒	☒	☒	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Jonathan L. Bryan	☐	☒	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Aman Turakhanov	☐	☐	☒	☒	☐	☐
I acknowledge that these represent my contribution to the above research output. 						

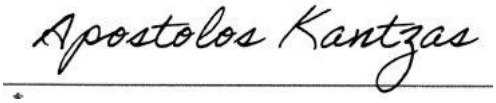
Alexey Cheremisin	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Apostolos Kantzas	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Sudarshan A. Mehta	☐	☒	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						

Improved oil viscosity characterization by low-field NMR using feature engineering and supervised learning algorithms. *Energy & Fuels, 2020 34 (11), 13799-13813*.

Authors: Strahinja Markovic[a,c*], Jonathan L. Bryan[b], Vladislav Ishimtsev[a], Aman Turakhanov[a], Reza Rezaee[c], Alexey Cheremisin[a], Apostolos Kantzas[b], Dmitry Koroteev[a], Sudarshan A. Mehta[b].

[a] Skolkovo Institute of Science and Technology, Moscow, Russian Federation
[b] University of Calgary, Calgary, Alberta, Canada
[c] Curtin University, Perth, Australia

Name of Co-authors	Conception and design	Acquisition of data & method	Data conditioning & manipulation	Analysis & statistical method	Interpretation & discussion & editing	Final approval
Strahinja Markovic	☒	☒	☒	☒	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Jonathan L. Bryan	☐	☒	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Vladislav Ishimtsev	☐	☐	☒	☒	☐	☐
I acknowledge that these represent my contribution to the above research output. 						

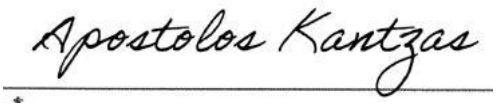
Aman Turakhanov	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Reza Rezaee	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Alexey Cheremisin	☐	☒	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Apostolos Kantzas	☐	☒	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Dmitry Koroteev	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Sudarshan A. Mehta	☐	☒	☐	☐	☐	☒
I acknowledge that these represent my contribution to the above research output. 						

Application of XGBoost model for *in-situ* water saturation determination in Canadian oil-sands by LF-NMR and density data

Scientific Reports, Nature, 2022, https://doi.org/10.1038/s41598-022-17886-6

S. Markovic[a,c], J. L. Bryan[d], R. Rezaee[c], A. Turakhanov[a], A. Cheremisin[a], A. Kantzas[b], D. Koroteev[a]

[a] Skolkovo Institute of Science and Technology, Moscow, Russian Federation
[b] University of Calgary, Calgary, Alberta, Canada
[c] Curtin University, Perth, Australia
[d] PERM Inc., Calgary, AB, Canada

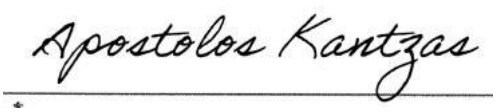
Name of Co-authors	Conception and design	Acquisition of data & method	Data conditioning & manipulation	Analysis & statistical method	Interpretation & discussion & editing	Final approval
Strahinja Markovic	☒	☐	☒	☒	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
J. L. Bryan	☒	☒	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
A. Kantzas	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
A. Turakhanov	☐	☐	☐	☐	☒	☐
I acknowledge that these represent my contribution to the above research output. 						

A. Cheremisin	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Reza Rezaee	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
D. Koroteev	☐	☐	☐	☐	☒	☐
I acknowledge that these represent my contribution to the above research output. 						

In-situ water saturation by LF-NMR and supervised learning - application to Canadian oil sands. *EAGE Geotech, London, United Kingdom, 2022.* DOI: 10.3997/2214-4609.20224021

S. Markovic[a,c], J. L. Bryan[b], R. Rezaee[c], A. Cheremisin[a], A. Kantzas[b]

[a] Skolkovo Institute of Science and Technology, Moscow, Russian Federation
[b] University of Calgary, Calgary, Alberta, Canada
[c] Curtin University, Perth, Australia

Name of Co-authors	Conception and design	Acquisition of data & method	Data conditioning & manipulation	Analysis & statistical method	Interpretation & discussion & editing	Final approval
Strahinja Markovic	☒	☐	☒	☒	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
J. L. Bryan	☒	☒	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
A. Cheremisin	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
A. Kantzas	☐	☐	☐	☐	☒	☒
I acknowledge that these represent my contribution to the above research output. 						
Reza Rezaee	☐	☐	☐	☐	☒	☒

I acknowledge that these represent my contribution to the above research output.

Bibliography

1. Alboudwarej, H. *et al.* Highlighting Heavy Oil. *Oilf. Rev.* **18**, 34–53 (2006).
2. Hascakir, B. Introduction to Thermal Enhanced Oil Recovery (EOR) special issue. *J. Pet. Sci. Eng.* **154**, 438–441 (2017).
3. Musin, K., Abdullin, T., Shipunov, T. & Khisamov, R. Estimation Viscosity and its Heterogeneity by NMR Logging Tool in Reservoir Conditions in Oilfield with Heavy Oil: Practical Results. *SPE Russ. Pet. Technol. Conf. Exhib.* 1–19 (2016) doi:10.2118/182066-RU.
4. James, L. A., Rezaei, N. & Chatzis, I. VAPEX, warm VAPEX and hybrid VAPEX - The state of enhanced oil recovery for in situ heavy oils in Canada. *J. Can. Pet. Technol.* **47**, 12–18 (2008).
5. Das, S. K. Vapex: An Efficient Process for the Recovery of Heavy Oil and Bitumen. *SPE J.* **3**, 232–237 (1998).
6. Cui, G., Liu, T., Xie, J., Rong, G. & Yang, L. A review of SAGD technology development and its possible application potential on thin-layer super-heavy oil reservoirs. *Geosci. Front.* **13**, 101382 (2022).
7. Sheng, J. J. *Cyclic Steam Stimulation*. *Enhanced Oil Recovery Field Case Studies* (Elsevier Inc., 2013). doi:10.1016/B978-0-12-386545-8.00016-6.
8. Sheng, J. J. *Steam Flooding*. *Enhanced Oil Recovery Field Case Studies* (Elsevier Inc., 2013). doi:10.1016/B978-0-12-386545-8.00015-4.
9. Turta, A. *In Situ Combustion*. *Enhanced Oil Recovery Field Case Studies* (Elsevier Inc., 2013). doi:10.1016/B978-0-12-386545-8.00018-X.
10. Djuraev, U. A review on conceptual and practical oil and gas reservoir monitoring methods. *J. Pet. Sci. Eng.* **152**, 586–601 (2017).
11. Shikhov, I., Li, R. & Arns, C. H. Relaxation and relaxation exchange NMR to characterise asphaltene adsorption and wettability dynamics in siliceous systems. *Fuel* **220**, 692–705 (2018).
12. Fayazi, A., Kryuchkov, S. & Kantzas, A. Evaluating Diffusivity of Toluene in Heavy Oil Using Nuclear Magnetic Resonance Imaging. *Energy and Fuels* **31**, 1226–1234 (2017).
13. Coates, G. R., Xiao, L. & Prammer, M. G. NMR logging. *Ebooks* 253 (1999).

14. Kantzas, A. & Bryan, J. L. Tipping process for NMR: Fundamentals of Fluid Flow in porous media. https://perminc.com/resources/fundamentals-of-fluid-flow-in-porous-media/chapter-3-molecular-diffusion/diffusion-coefficient/measurement-techniques/nmr-method/principles-nmr-processing/tipping-process/ (2016).
15. Zhang, Q., Lo, S. & Huang, C. Some exceptions to default NMR rock and fluid properties. *SPWLA 39th Annu. …* FF (1998).
16. Testamanti, M. N. & Rezaee, R. Determination of NMR T 2 cut-off for clay bound water in shales: A case study of Carynginia Formation, Perth Basin, Western Australia. *J. Pet. Sci. Eng.* **149**, 497–503 (2017).
17. Testamanti, M. N. & Rezaee, R. Considerations for the acquisition and inversion of NMR T2 data in shales. *J. Pet. Sci. Eng.* **174**, 177–188 (2019).
18. Bryan, J., Kantzas, A. & Bellehumeur, C. Oil-Viscosity Predictions From Low-Field NMR Measurements. *SPE Reserv. Eval. Eng.* **8**, 44–52 (2005).
19. Bloembergen, N., Purcell, E. M. & Pound, R. V. Relaxation Effects in Nuclear Magnetic Resonance Absorption. *Phys. Rev.* **73**, 679–712 (1948).
20. Khan, M. A. B., Mehrotra, A. K. & Svrcek, W. Y. Viscosity Models for Gas-Free Athabasca Bitumen. *J. Can. Pet. Technol.* **23**, 47–53 (1984).
21. Meyer, R., Attanasi, E. & Freeman, P. Heavy Oil and Natural Bitumen Resources in Geological Basins of the World. *Usgs* **1084**, 36 (2007).
22. Luo, P. & Gu, Y. Effects of asphaltene content on the heavy oil viscosity at different temperatures. *Fuel* **86**, 1069–1078 (2007).
23. Straley, C., Rossini, D., Vinegar, H., Tutunjian, P. & Morriss, C. Core Analysis by Low Field NMR. *Log Anal.* **38**, 84–94 (1997).
24. Straley, C. Reassessment of Correlations of between Viscosity and NMR Measurements. *Spwla 47* AA (2006).
25. Nicot, B., Fleury, M. & Leblond, J. Improvement of viscosity prediction using NMR relaxation. *48th Annu. Logging Symp.* 1–7 (2007).
26. Morriss, C. E. *et al.* Hydrocarbon saturation and viscosity estimation from NMR logging in the belridge diatomite. *SPWLA 35th Annu. Logging Symp. 1994* (1994).

27. Zega, J. A., House, W. V & Kobayashi, R. Spin-Lattice Relaxation and Viscosity in Mixtures of n -Hexane and n -Hexadecane. **1**, 909–912 (1990).
28. Sun, B., Dunn, K. J., Latorraca, G. A., Liu, C. & Menard, G. Apparent hydrogen index and its correlation with heavy oil viscosity. *48th Annu. Logging Symp. 2007* **298**, 1–14 (2007).
29. Kleinberg, R. I. & Vinegar, H. J. NMR properties of reservoir fluids. *Log Anal.* 20–32 (1996) doi:10.1016/S0730-725X(03)00135-8.
30. Ahmed, K. *et al.* Viscosity Predictions of Viscous Oil from a Kuwait Oil Field by Low-field. 1–10 (2014).
31. Sandor, M., Cheng, Y. & Chen, S. Improved Correlations for Heavy-Oil Viscosity Prediction with NMR. *J. Pet. Sci. Eng.* **147**, 416–426 (2016).
32. Bryan J. & Kantzas A. Badry R. Emmerson J. Hancsicsak T. In-situ Viscosity of Heavy Oil : Core and Log Calibrations. *J. Can. Pet. Technol. 46(11) 47-55* (2006) doi:10.2118/07-11-04.
33. Cheng, Y., Kharrat, A. M., Badry, R. & Kleinberg, R. L. Power-law relationship between the viscosity of heavy oils and NMR relaxation. *SPWLA 50th Annu. Logging Symp. 2009* 1–7 (2009).
34. Singer, P. M. *et al.* Elucidating the 1 H NMR Relaxation Mechanism in Polydisperse Polymers and Bitumen using Measurements , MD Simulations , and Models. (2020) doi:10.1021/acs.jpcb.0c01941.
35. Singer, P. M. *et al.* Interpretation of NMR Relaxation in Bitumen and Organic Shale using Polymer-Heptane Mixes. (2018) doi:10.1021/acs.energyfuels.7b03603.
36. Kantzas, A. Advances in magnetic resonance relaxometry for heavy oil and bitumen characterization. *J. Can. Pet. Technol.* **48**, 15–23 (2009).
37. Yang, Z., Hirasaki, G. J., Appel, M. & Reed, D. A. Viscosity Evaluation for NMR Well Logging of Live Heavy Oils. *Petrophysics* **53**, 22–37 (2012).
38. Joss, L. & Mu, E. A. Machine Learning for Fluid Property Correlations: Classroom Examples with MATLAB. (2018) doi:10.1021/acs.jchemed.8b00692.
39. 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.* **5**, 83 (2019).

40. Lei, Y. *et al.* Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **138**, 106587 (2020).
41. Gao, Z., Zou, X., Huang, Z. & Zhu, L. Predicting sooting tendencies of oxygenated hydrocarbon fuels with machine learning algorithms. *Fuel* **242**, 438–446 (2019).
42. Myshkin, N. K. *On-line Condition Monitoring in Industrial Lubrication and Tribology*.
43. Barbosa, L. L. *et al.* Relative hydrogen index as a fast method for the simultaneous determination of physicochemical properties of petroleum fractions. *Fuel* **210**, 41–48 (2017).
44. Constantino, A. F. *et al.* Determination of physicochemical properties of biodiesel and blends using low-field NMR and multivariate calibration. *Fuel* **237**, 745–752 (2019).
45. Alptekin, E. & Canakci, M. Determination of the density and the viscosities of biodiesel–diesel fuel blends. *Renew. Energy* **33**, 2623–2630 (2008).
46. 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.* **18**, 813–820 (2009).
47. Li, H. *et al.* OPEN A novel multi-target regression framework for time-series prediction of drug efficacy. *Nat. Publ. Gr.* 1–9 (2017) doi:10.1038/srep40652.
48. 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* **259**, 1048–1055 (2005).
49. Khakimova, L. *et al.* High-pressure air injection laboratory-scale numerical models of oxidation experiments for Kirsanovskoye oil field. *J. Pet. Sci. Eng.* **188**, 106796 (2020).
50. US6794865 Method and apparatus for monitoring the health of a fluid system , particularly a gas turbine oil system. (2004).

51. 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.* **52**, 2049–2057 (1997).
52. Liang, X., Chang-chun, Z., Zhi-qiang, M., Yu-jiang, S. & Yan, J. Estimation of water saturation from nuclear magnetic resonance (NMR) and conventional logs in low permeability sandstone reservoirs. *J. Pet. Sci. Eng.* **108**, 40–51 (2013).
53. Donaldson, E. C. *Well logging for earth scientists*. *Journal of Petroleum Science and Engineering* vol. 2 (1989).
54. Tiab, D. Formation Resistivity and Water Saturation. in *Petrophysics* (Elsevier, 2012). doi:10.1016/B978-0-12-383848-3.00004-9.
55. Wang, W. & Wei, W. Chapter 3 - Water chemistry. in *Fluid Chemistry, Drilling and Completion* (ed. Wang, Q.) 95–114 (Gulf Professional Publishing, 2022). doi:https://doi.org/10.1016/B978-0-12-822721-3.00005-8.
56. Mohammadlou, M. H. & Langeland, H. Use of the NMR and Resistivity Logs to Quantify Movable Hydrocarbon ; Solution for the Tight and Low-Resistivity Carbonate Reservoirs. **SPE 141047**, (2011).
57. Cowie, B. R., James, B. & Mayer, B. Distribution of total dissolved solids in McMurray Formation water in the Athabasca oil sands region, Alberta, Canada: Implications for regional hydrogeology and resource development. *Am. Assoc. Pet. Geol. Bull.* **99**, 77–90 (2015).
58. Zheng, S. *et al.* Nuclear magnetic resonance T2 cutoffs of coals: A novel method by multifractal analysis theory. *Fuel* **241**, 715–724 (2019).
59. Bryan, J., Mai, A., Hum, F. M. & Kantzas, A. Oil- and water-content measurements in bitumen ore and froth samples using low-field NMR. *SPE Reserv. Eval. Eng.* **9**, 654–663 (2006).
60. Yang, Z. & Hirasaki, G. J. NMR measurement of bitumen at different temperatures. *J. Magn. Reson.* **192**, 280–293 (2008).
61. Bryan, J., Moon, D. & Kantzas, A. In situ viscosity of oil sands using low field NMR. *J. Can. Pet. Technol.* **44**, 23–29 (2005).
62. Anand, V. & Hirasaki, G. J. Diffusional coupling between micro and

macroporosity for NMR relaxation in sandstones and grainstones. *SPWLA 46th Annu. Logging Symp. 2005* **48**, 289–307 (2005).

63. Singer, P. M., Chen, Z., Wang, X. & Hirasaki, G. J. Diffusive coupling in heptane-saturated kerogen isolates evidenced by NMR T1-T2 and T2-T2 maps. *Fuel* **280**, 118626 (2020).
64. Mukhametdinova, A., Habina-Skrzyniarz, I., Kazak, A. & Krzyżak, A. NMR relaxometry interpretation of source rock liquid saturation — A holistic approach. *Mar. Pet. Geol.* **132**, (2021).
65. Newgord, C., Tandon, S. & Heidari, Z. Simultaneous assessment of wettability and water saturation using 2D NMR measurements. *Fuel* **270**, 117431 (2020).
66. Krzyżak, A. T., Habina-Skrzyniarz, I., Machowski, G. & Mazur, W. Overcoming the barriers to the exploration of nanoporous shales porosity. *Microporous Mesoporous Mater.* **298**, (2020).
67. Venkataramanan, L. *et al.* An unsupervised learning algorithm to compute fluid volumes from NMR T1-T2 logs in unconventional reservoirs. *Petrophysics* **59**, 617–632 (2018).
68. Burcaw, L. *et al.* Improved Methods for Estimating the Viscosity of Heavy Oils From Nmr Data. *Spwla* **49**, 1–14 (2008).
69. Miller, K. A. Should You Trust Your Heavy Oil Viscosity Measurement ?
70. Nicot, B., Fleury, M. & Leblond, J. A New Methodology For Better Viscosity Prediction Using Nmr Relaxation. *SPWLA 47th Annu. Logging Symp.* 1–12 (2006).
71. Bloch, F. Nuclear Induction. *Phys. Rev.* **70**, 460–474 (1946).
72. P. Lindsey, C. & Patterson, G. *Detailed Comparison of the Williams-Watts and Cole-Davidson Functions*. *The Journal of Chemical Physics* vol. 73 (1980).
73. Ahmad, K. *et al.* Radial-basis-function-based nuclear magnetic resonance heavy oil viscosity prediction model for a Kuwait viscous oil field. *Interpretation* **4**, SF81–SF92 (2016).
74. LaTorraca, G. A., Stonard, S. W., Webber, P. R., Carison, R. M. & Dunn, K. J. Heavy oil viscosity determination using NMR logs. *SPWLA 40th Annu.*

Logging Symp. **3**, 11 (1999).

75. Galford, J. E. & Marschall, D. M. Combining NMR and Conventional Logs to Determine Fluid Volumes and Oil Viscosity in Heavy-Oil Reservoirs. *SPE Annu. Tech. Conf. Exhib.* 12 (2007) doi:10.2118/63257-ms.
76. Mirotchnik, K. D. & Allsopp, K. Low-Field NMR Method for Bitumen Sands Characterization : A New Approach. *SPE J.* 88–96 (2001).
77. Bryan, J., Mirotchnik, K. & Kantzas, A. Viscosity determination of heavy oil and bitumen using NMR relaxometry. *J. Can. Pet. Technol.* **42**, 29–34 (2003).
78. Yang, Z. & Hirasaki, G. J. NMR measurement of bitumen at different temperatures. *J. Magn. Reson.* **192**, 280–293 (2008).
79. Bryan, J., Mirotchnik, K. & Kantzas, A. Viscosity Determination of Heavy Oil and Bitumen Using NMR Relaxometry. **42**, (2003).
80. American Society for Testing Materials, (ASTM D4287). Standard Test Method for High-Shear Viscosity Using a Cone/Plate Viscometer. *B. Stand.* **06.01**,.
81. Tang, Y., Mccowan, D. & Song, Y. OPEN A miniaturized spectrometer for NMR relaxometry under extreme conditions. *Sci. Rep.* 1–9 (2019) doi:10.1038/s41598-019-47634-2.
82. Madsen, K. & H. B. Nielsen, O. T. Methods for non-linear least squares problems. at (2004).
83. Barbosa, L. L., Kock, F. V. C., Almeida, V. M. D. L., Menezes, S. M. C. & Castro, E. V. R. Low-field nuclear magnetic resonance for petroleum distillate characterization. *Fuel Process. Technol.* **138**, 202–209 (2015).
84. P. T. Boggs and J. E. Rogers. "Orthogonal Distance Regression," in "Statistical analysis of measurement error models and applications: proceedings of the AMS-IMS-SIAM joint summer research conference held June 10-16, 1989,". *Contemp. Math.* **112**, 186 (1990).
85. Hirasaki, G. J., Lo, S. W. & Zhang, Y. NMR properties of petroleum reservoir fluids. in *Magnetic Resonance Imaging* vol. 21 269–277 (2003).
86. Markovic, S. *et al.* In-situ heavy oil viscosity prediction at high temperatures using low-field NMR relaxometry and nonlinear least squares. *Fuel* **260**,

(2020).

87. Li, H. & Misra, S. Prediction of subsurface NMR T2 distribution from formation-mineral composition using variational autoencoder. in *SEG Technical Program Expanded Abstracts 2017* 3350–3354 (2017). doi:10.1190/segam2017-17798488.1.
88. 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.* **32**, 3873–3885 (2020).
89. Bühlmann, P. & Hothorn, T. Boosting algorithms: Regularization, prediction and model fitting. *Stat. Sci.* **22**, 477–505 (2007).
90. Cortes, C., Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).
91. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*.
92. Ho, T. K. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* vol. 1 278–282 vol.1 (1995).
93. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
94. Cohen, P., West, S. G. & Aiken, L. S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. (Taylor & Francis, 2014).
95. Zheng, A. & Casari, A. *Feature engineering for machine learning*. *O'Reilly Media* (O'Reilly Media, Inc., 2018). doi:10.13140/RG.2.1.3564.3367.
96. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
97. Schapire, R. E. *Boosting: Foundations and Algorithms*. *Kybernetes* vol. 42 (2013).
98. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
99. Ridgeway, G. Generalized Boosted Models : A guide to the gbm package. 1–15 (2019).
100. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat.*

Comput. **14**, 199–222 (2004).

101. Dodge, Y. LAD Regression for Detecting Outliers in Response and Explanatory Variables. *J. Multivar. Anal.* **61**, 144–158 (1997).
102. Askarova, A. *et al.* Thermal enhanced oil recovery in deep heavy oil carbonates: Experimental and numerical study on a hot water injection performance. *J. Pet. Sci. Eng.* **194**, 107456 (2020).
103. Nicot, B., Fleury, M. & Leblond, J. Improvement Of Viscosity Prediction Using NMR Relaxation. *SPWLA 48th Annual Logging Symposium* 7 at https://doi.org/ (2007).
104. Misra, S. & Wu, Y. Chapter 10 - Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. in (eds. Misra, S., Li, H. & He, J. B. T.-M. L. for S. C.) 289–314 (Gulf Professional Publishing, 2020). doi:https://doi.org/10.1016/B978-0-12-817736-5.00010-7.
105. Valentín, M. B. *et al.* Estimation of permeability and effective porosity logs using deep autoencoders in borehole image logs from the brazilian pre-salt carbonate. *J. Pet. Sci. Eng.* **170**, 315–330 (2018).
106. Wherity, S., Sidley, T., Cowling, M., Ismayilov, A. & Olie, M. Observation and Monitoring Well : in Situ Window to Assess Recovery Introduction to the Halfdan Field. (2014).
107. 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.* **30**, 79–82 (2005).
108. Chollet, F. & others. Keras. at (2015).
109. de Myttenaere, A., Golden, B., Le Grand, B. & Rossi, F. Mean Absolute Percentage Error for regression models. *Neurocomputing* **192**, 38–48 (2016).
110. Theil, H., Cramer, J. S., Moerman, H. & Russchen, A. *Economic Forecasts and Policy*. (North-Holland Publishing Company, 1961).
111. Biancolini, M. E. *Radial Basis Functions for Engineering Applications*. (Springer US, 2017).

112. Cherkassky, V. & Ma, Y. Selection of meta-parameters for support vector regression. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2415 LNCS**, 687–693 (2002).
113. Sarle, W.S. *Neural Network FAQ, part 1 of 7: Introduction, periodic posting to the Usenet newsgroup comp.ai.neural-nets*. url: ftp://ftp.sas.com/pub/neural/FAQ.html (1997).
114. Riazoshams, H., Midi, H. & Ghilagaber, G. Outlier Detection in Nonlinear Regression. *Robust Nonlinear Regression* 107–141 at https://doi.org/10.1002/9781119010463.ch6 (2018).
115. Sohn, B. Y. & Kim, G. B. Detection of outliers in weighted least squares regression. *Korean J. Comput. Appl. Math.* **4**, 441–452 (1997).
116. Jones, E., Oliphant, T., Peterson, P. SciPy: Open source scientific tools for Python. Retrieved from 'http://www.scipy.org/'. http://www.scipy.org/ (2001).
117. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* vols 13-17-Augu 785–794 (2016).
118. Nocedal, J. & Wright, S. J. Numerical optimization. *Springer Ser. Oper. Res. Financ. Eng.* 1–664 (2006) doi:10.1201/b19115-11.
119. Tikhonov, A. N. & Arsenin, V. Y. *Solutions of ill-posed problems*. (V. H. Winston & Sons, 1977).
120. Prammer, M. G., Drack, E. D., Bouton, J. C. & Gardner, J. S. Measurements of clay-bound water and total porosity by magnetic resonance logging. *Log Anal.* **37**, 61–69 (1996).
121. Ross, B. C. Mutual information between discrete and continuous data sets. *PLoS One* **9**, (2014).
122. Chai, T. & Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **7**, 1247–1250 (2014).
123. Chen, J. & Bryan, J. In situ bitumen viscosity and saturation estimation from core log integration for Canadian oil sands. *Soc. Pet. Eng. - SPE Heavy Oil*

Conf. Canada 2013 **3**, 1686–1693 (2013).

124. Jin, G., Xie, R., Liu, M. & Guo, J. Petrophysical Parameter Calculation Based on NMR Echo Data in Tight Sandstone. *IEEE Trans. Geosci. Remote Sens.* **57**, 5618–5625 (2019).

125. Gómez-Ríos, A., Luengo, J. & Herrera, F. A study on the noise label influence in boosting algorithms: Adaboost, GBM and XGBoost. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **10334 LNCS**, 268–280 (2017).