

Thesis Changes Log

Name of Candidate: Strahinja Markovic

PhD Program: Petroleum Engineering

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

Supervisors: Prof. Alexey Cheremisin, Prof. Reza Rezaee

Co-supervisor: Prof. Dmitry Koroteev

The thesis document includes the following changes in answer to the external review process.

Amendments proposed by the review committee are below. I added line numbers to the manuscript for more convenient cross-referencing of changes. All changes are highlighted. Appendix A presents an extension to Chapter 4. It summarizes other machine learning models for water saturation prediction. It is attached to the end of manuscript.

Revisions per suggestions of Prof. Siddharth Misra

Comment 1: Thesis title seems very broad. The thesis title needs to be streamline to “heavy oil”, “supervised learning” and “regression”.

Reply 1: The thesis title has been modified to: “*Application of LF-NMR measurements and supervised learning regression methods for improved characterization of heavy oils and bitumens.*”

Comment 2: As suggested in Chapter 3 title has been improved: “IMPROVED OIL VISCOSITY PREDICTION BY LOW-FIELD NMR USING FEATURE ENGINEERING AND SUPERVISED LEARNING METHODS”

Comment 3: The Chapter 4 title has been modified to: “APPLICATION OF XGBOOST MODEL FOR IN-SITU WATER SATURATION DETERMINATION IN CANADIAN OIL-SANDS BY LF-NMR AND BULK DENSITY MEASUREMENTS”

Comment 4: Viscosity ranges for each EOR method were added along with relevant citations. (Pgs. 25-26)

Comment 5: Explanation about NMR logging monitoring for heavy oil EOR added. (Pg. 27, lines 20-25)

Comment 6: Explanation of other factors that influence T2 distribution and inversion (Pgs. 35-36)

Comment 7: Low-field is defined as magnetic fields below 550 gaussess (Pg. 38, lines 1-4)

Comment 8: (Pg. 44, lines 19-28) To improve the clarity of explanation, I modified and expanded a part on peak deconvolution and typical T2 responses of clay-bound, capillary-bound water, and bitumen.

Comment 9: I added a new subsection 1.3.2 that addresses relaxation mechanisms in pore space (Pg. 32-33)

Comment 10: Improved the clarity of the sentence (Pg. 45, lines 8-10)

Comment 11: I added an explanation for why the viscosity model performance degrades with temperature increase (Pg. 48, line 9)

Comment 12: Pg. 48, line 18, Why such a high value was chosen?

Reply 12: 70-21,000 cP is a range of viscosities of 23 heavy oil samples. I rewrote the sentence to make it more transparent.

Comment 13: All figures about model performance should clearly state whether it is training performance or testing performance.

Reply 13: I updated all figures in the manuscript to emphasize whether it is a training or testing score.

Comment 14: All input variables and a target are now clearly presented in Table 2 (Pg.87, lines 1-4).

Comment 15: The range is wide. Log-transformed viscosity should be used.

Reply 15: I did use log-transformed viscosity for prediction. Table 2 now shows that the target variable is $\text{Log}(\eta)$.

Comment 16: For SVR, what basis function was used, and what was the degree?

Reply 16: RBF kernel was used. To my understanding, the degree parameter is considered only for the polynomial kernel.

Comment 17: Chapter 4 - Please clearly mention the features used and the target and size of the dataset.

Reply 17: I added the size of training and test sets (Pg. 92, lines 13-16). The inputs and the target are in Table 2 (Pg.81, lines 1-4).

Comment 18: Chapter 4 - Please clearly mention the features used and the target and their size. Also, all figures about model performance should clearly state whether it is training or testing.

Reply 18: I added an explanation on Pg. 113, lines 2-3 about the dataset size (subsection 4.1.). Also, for both modeling approaches, I presented in tables input features and target variable (XGB-FE – Tables 10, 11; XGB-FS Table 12). For each figure with model prediction performance, the test set was used. I updated all figures to clarify this.

Comment 19: How can we know that deconvolution is reliable?

Reply 19: The procedure used in this study was successfully used in several other published works by Bryan et al. (citations are in the manuscript). However, in this study, we are claiming that deconvolution is not a reliable method when the overlapping of the water and oil signals is substantial, and we propose a data-driven approach instead.

Comment 20: Pgs. 114-116; Multiple comments related to XGBoost model derivation have been addressed, including the addition of equations that clarify what is f_i , an explanation about second order gradients, what leaves are, and their weight (score) etc.

Comment 21: Pg. 118, line 15; Proportion corrected.

Comment 22: Pg. 120, line 3; Corrected from 'frequency distribution' to 'distribution.'

Comment 23: Pg. 121, line 19; What is mutual information regression?

Reply 23: It is a feature selection method based on MI. In the scikit learn library, it is called mutual information regression.

Comment 24: Pg. 124, line 1; Why do some features have high MI, and what is the physical meaning?

Reply 24: This is a good question. In this work, our main objective was to develop a model with good forecasting of water saturation. When we discovered that density improves the model predictions, we assumed that water saturation might be negatively correlated with density (i.e., if the density of the rock is higher, fewer pores can be occupied by water. This, however, is not entirely true since the correlation coefficient between water and density is $r = -0.26$, which is too weak to make any conclusions. MI score between these two variables is $MI = 0.86$ nats (1 is max). This question remains to be inspected further by performing more experiments and deeper analysis.

Comment 25: Pgs. 119-121; The dimensionality is too high for this approach to work. XGBoost will not be reliable.

Reply 25: I agree. This is what I wanted to demonstrate and to show that it is better to use feature engineering and extraction. Dimensionality reduction techniques like PCA were not tested, but I do plan perform a follow-up study. Thank you for the excellent suggestion. I added an explanation for future work on Pg. 137, lines 8-13 and Pg. 140, lines 19-22.

Comment 26: Pg. 122, Table 12; This is a very complex model. It should have fewer estimators and smaller depth.

Reply 26: After reading more about XGBoost documentation and considering the dataset size and the task, I agree. However, these are hyperparameters obtained from Bayesian Optimization. Now, I tried to restrict the search space to a smaller range ($n_estimators$, max_depth). The results show a slight drop in both training and testing scores, but are still quite similar to what I reported before. The details are in Appendix A (Pg. 150).

Comment 27: Pg. 129, lines 1-4; Are you sure RMSE and MAE can be the same? Explain what is 'negatively oriented score'.

Reply 27: Yes. RMSE and MAE can be the same if and only if the absolute values of all prediction errors are identical. Since this is rarely the case, I deleted the relevant part to avoid the confusion. Also, negatively oriented scores are those for which smaller values are preferred. I added an explanation to Pg. 123, lines 3-4.

Comment 28: Feature scaling is not mentioned. It does not affect tree-based methods but enables the application of other methods. Also, it allows use of simpler models like SVR and Elastic Net.

Reply 28: We did test many different methods, including SVR, elastic net, random forests, gaussian process. However, to fit within the journal requirements, we had to reduce to one particular method. XGBoost showed the best scores, so we went with it. Therefore, scaling was done, but the article did not need to report it. Appendix A contains a brief feature-scaling section (Pg., 141), and a summary of other machine learning methods.

Comment 29: Pg. 136, lines 26-30; Reduce the SNR by adding noise to the lab measurements and study the performance of models with low SNR inputs.

Reply 29: The noise should be added to the raw data (spin-echo decay curves) before the inversion. The experiments were performed in the mid-2000s at the University of Calgary, and unfortunately, the raw data were not preserved. Presumably, by adding noise and performing inversion, broader amplitudes and smoother curves are anticipated, in which case T2 cutoff and deconvolution methods become even less reliable. Since, in the data-driven approach, the general behavior of the T2 distribution with the water saturation change is implicitly used during model training, it is expected that the XGBoost will still prove as a more reliable solution. Studies in literature also confirm this (citation is in the manuscript). This remains to be confirmed in future work.

Comment 30: Pg. 138, line 6; Feature engineering is a very broad term. It may not be the right word for what you did.

Reply 30: 'Feature engineering' replaced with 'Feature extraction.

Comment 31: Pg. 140, lines 19-22; Minor changes in terminology.

Revisions per suggestions of Prof. Christoph Arns

Comment 32: Pg. 28, line 8; Gyromagnetic ratio defined.

Comment 33: Pg. 36, line 15; I added a suffix 'ss' to tau.

Comment 34: Pg. 42, line 23; Lithology-dependent NMR responses: not so sure about that – the initial amplitude is not, but the decay is influenced by internal gradients, diffusion coupling.

Reply 34: I agree. I removed the relevant part, and modified the sentence.

Comment 35: 3ms for clay-bound water; should this be 33ms?

Reply 35: Not quite. According to Straley, in sandstones, responses < 3ms belong to clay-bound water, < 33ms to capillary-bound water, and >33ms to free fluids.

Comment 36: JC bitumen – define or provide reference

Reply 36: JC bitumen was the name of the sample. I defined it at Pg. 48, line 28.

Comment 37: I fixed omega and omega_0 and referenced *b* term correctly (Pgs. 50-51)

Comment 38: eq. (19) is for porous media; but afterwards heavy oil seems to be treated as fluids. This can cause confusion, since solid components in heavy oils may act as surface relaxation centers. Careful with the terms T₂ vs T₂bulk here. Is Fig. 6 referring to T₂bulk?

Reply 38: This is true, but, under assumption that porous system is water wet, water will form a thin film over the grain surface, which presents a barrier for oil to get in direct contact with the pore wall. Therefore, T₂ bulk relaxation is the main mechanism for oil which is why its $T_2 \cong T_{2\text{bulk}}$. I added a brief explanation in Pg. 51, line 10. I also rewrote T₂bulk to T₂ in Equation 25 to avoid confusion.

As for the solids acting as surface relaxation centers, the recent research by Singer et al. ([10.1021/acs.energyfuels.7b03603](https://doi.org/10.1021/acs.energyfuels.7b03603)) suggests that in high-viscosity regime, the T₂ relaxation for bitumens and heavy oils can be explained solely by dipole-dipole interactions, without a need to employ surface paramagnetism. But in polymer-heptane mixes for instance, surface paramagnetism must be considered (polymer acts as a surface relaxation center for heptane).

Also, equation 19 (now equation 5) was transferred to subsection 1.3.2 where I expanded the theory for T₂ relaxation in porous medium.

Comment 39: p.52, eq 24 & 25 use different exponents, yet beta seems to be introduced for Eq. 25 in the text.

Reply 39: Thank you for pointing out. I skipped a couple of derivation steps from the original source. Now I attempted to make it a bit clearer Pg. 52, lines 12-17.

Comment 40: Do you need to consider oxygen effects or boiling points of different components in your model? Also, correction term has not been defined;

Reply 40: Thank you for the excellent question. Boiling point of a certain component is related to its' molar weight and density. In NMR relaxometry boiling point can be predicted by calculating RHI measured by NMR device. We can convert RHI to hydrogen number. Studies showed that hydrogen number has a strong linear positive correlation with boiling point. My understanding is that the boiling point is related to RHI, which is used in our model, and therefore does not have to be considered explicitly. However, it would be interesting to try to use boiling point in a data-driven approach to predict viscosity or other oil properties.

As for the dissolved oxygen which is paramagnetic, we know that it shortens the T₂ relaxation of fluids. Most of the samples I worked with are dead heavy and extra heavy oils, which naturally contain smaller amount of oxygen and high concentrations of asphaltenes and resins. In addition, studies show that dissolved oxygen has less affinity to these oils, so the paramagnetic effects are negligible as confirmed by Hürlimann et al (doi: 10.1007/BF03167180).

Also, I defined the temperature correction term at Pgs. 54, lines 15-16.

Comment 41: What was the SNR of the experiments?

Reply 41: For 23 heavy oil samples. The dataset was obtained from another study done in 2003 (doi: 10.2118/03-07-02) and unfortunately the SNR was not disclosed. For bitumen sample it was >100 for each measurement (Pg. 61, line 14).

Comment 42: Peek at 200 degC is past the glass transition. Does the NMR signal of peek change with temperature?

Reply 42: Yes, but the signal is negligible as it represents less than 1% of the total signal. The explanation is at Pg. 61, lines 18-23. This also corresponds to findings in other studies (doi:10.1038/s41598-019-47634-2).

Comment 43: Pg 62, lines 21-24; Xsi defined as reduced chi-squared.

Comment 44: How the test set was selected?

Reply 44: The dataset consisted of 282 observations. After feature engineering, the dataset was randomly shuffled and split into a train and test sets. The split seed was selected randomly (random_state=42) for reproducibility of the results. I added an explanation at Pg. 78, lined 12-18.

Comment 45: 'T2-related' features – should this be T?

Reply 45: No. T is temperature in kelvins T(°K). I corrected this at Pgs. 82-85.

Comment 46: Is your training data split to carry out CV?

Reply 46: Yes. We used training data to perform CV. In CV a user defines a number of folds. I defined 5-folds. Then, within the training set, the algorithm uses for each fold 4/5ths of the training set for training and remaining 1/5th is used for validation. Then a different fold is selected, and the procedure repeats until all 5-folds are tested. The average score is calculated for 5-folds.

Comment 47: Figure 20 - Should this be lambda?

Reply 47: Yes, thank you. Eta changed to lambda in Figure 20.

Comment 48: Check last paragraph Pg. 111.

Reply 48: I fixed the sentence at Pg. 111, line 8.

Revisions per suggestions of Prof. Dmitry Eskin

Comment 49: I fixed the equation 13.

Comment 50: The correlation for oil viscosity (Eq.28). The author employed more parameters than the other investigators mentioned in the thesis. Can the author be sure that the model is not overfitted? Has the author verified model performance using not his experimental data but those obtained by other researchers? This issue has to be clearly explained in the thesis.

Reply 50: The model has been tested only using the data from the study (23 heavy oil samples in one dataset, and separately 1 bitumen sample at a broad range of temperatures). Indeed, testing the model on a set of new samples would enable additional validation of the model. I added an explanation for this limitation at Pgs. 75-76.

Comment 51: For obtaining the viscosity model parameters, the author used the approach based on the least square mean difference technique. Such an approach is questionable because it is claimed that the correlation obtained is applicable for a wide range of viscosities, whereas the least square technique to a significant extent ignores a contribution of small viscosities into the model. This issue has to be explained in the thesis.

Reply 51: I agree. At the time I worked on this problem, I used OriginLab software, which does not allow the use of different loss functions. However, in the following research I did try to use orthogonal distance regression instead non-linear least squares, which seems to provide less biased optimization. The main choice between optimization method will depend on the users' goals and the range of viscosities of the expected viscosities in the well. I provided an explanation about this limitation at Pg. 75-76.

Revisions per suggestions of Prof. Dmitriy Potapenko

Comment 52: Pg. 23, I rephrased the sentence to make it more transparent.

Comment 53: The Figure 2 is a bit confusing - one may conclude that application of B1 causes oscillation of individual spins along the direction of B1. Suggesting to add the reference to the picture if it was taken from some literature source or to re-draw it.

Reply 53: I removed the Figure 2 and added three new Figures (2,3,4 in updated manuscript) which illustrate (1) T1 polarization, (2) B1 90 and 180 pulses and (3) generation of spin-echo. Relevant citations are provided (Coates, G., NMR logging: Principles and Applications, 1999). In addition, I added a new section about T2 relaxation in porous media where I explained in a greater detail the relaxation mechanisms. Pgs. 32-33.

Comment 54: Pg. 38, I suggest to define what low-field means.

Reply 54: I defined low-field as magnetic fields below 550 gauss (Pg. 38, lines 1-4).

Comment 55: Pg. 41, line 15; The equation is for pressures above the bubble point.

Reply 55: Thank you for pointing out to the omission. I added the explanation.

Comment 56: Pg. 56, line 12; Where the samples de-gassed?

Reply 56: Yes, dead oil samples. I added the explanation to Pg. 57, line 17.

Comment 57: Pg. 61., lines 8-14; Any particular reason for selecting these parameter values.

Reply 57: Thank you for the comment. Yes, we wanted to capture the fast-relaxing part of the distribution with better resolution at ambient conditions, but also to have good SNR at high temperatures. Since the temperature gradient was large between sample at 200 degC and environment (25 degC), these experiments had to be fast, so we can perform at least 1 measurement per 1 degC. I expanded the discussion in the relevant section Pg. 58, lines 8-14.

Comment 58: I applied most of the minor textual corrections suggested.

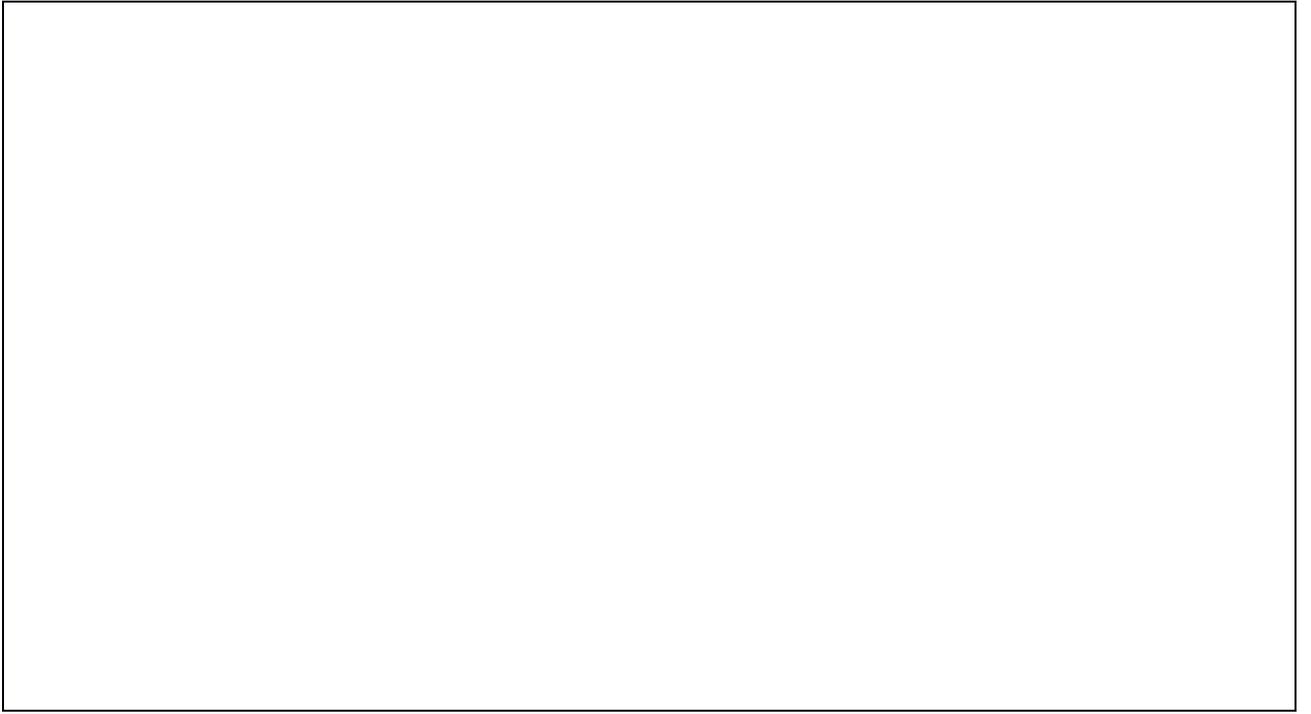
Revisions per suggestions of Prof. Alexei Tchistiakov

Comment 59: There is a disruption in automatic numbering of subtitles of chapter 3 in the table of contents (between subtitles 3.2.3 and 3.2.4).

Reply 59: Thank you for pointing out this omission. Table of content corrected.

Comment 60: Pg. 4, line 11; This statement seems to be too straight forward as saturation determination is a classical task of well logging. It is recommended to either rephrase this statement or put it better in the context.

Reply 60: I rephrased the sentence accordingly.



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

Skoltech

Skolkovo Institute of Science and Technology

Skolkovo Institute of Science and Technology

**Curtin
University**

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

© Strahinja Markovic 2022

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

11

12

13 Candidate:

14 Strahinja Markovic



15

16

17 Supervisors:

18 Prof. Alexey Cheremisin, Skoltech University

19 Prof. Reza Rezaee, Curtin University

20

21

1 **Abstract**

2

3 Heavy oil and bitumen resources are the most abundant hydrocarbon energy
4 source worldwide. However, thermal enhanced oil recovery (EOR) methods are
5 frequently applied to enhance their mobility and production due to their high
6 viscosity. In addition, owing to the chemical dissimilarity of oils and various
7 temperatures these oils are exposed to, as well as LF-NMR equipment limitations,
8 the commonly used models fail to perform at a satisfactory level, making them
9 impractical for use in heavy oil and bitumen reservoirs, and in environments with
10 large temperature **variability** (e.g., mechanical systems). Information about the
11 distribution of oil viscosity within the reservoir can be used to help **manage** the
12 thermal EOR projects. Nuclear magnetic resonance (NMR) downhole tools provide
13 a non-destructive way to determine the oil viscosity without recovering samples
14 (core or produced oil) from the well.

15

16 A new analytical NMR viscosity model was developed and tested on a suite of 23
17 Canadian heavy oils recovered from **various** reservoirs. The model was based on
18 two NMR parameters – T₂ logarithmic mean and relative hydrogen index.
19 Subsequently, the model was tested on a single bitumen sample at **the**
20 temperature range from 26-200 °C. Results were compared to nine well-known
21 NMR viscosity models described in the literature, and in both cases, the **proposed**
22 model got the most favorable statistical scores. Furthermore, a simple model
23 optimization procedure was presented, employing nonlinear least squares (NLS)
24 regression. Experiments were carried out at temperatures corresponding to those
25 in the hot steam injection EOR treatments. The same methodology can be
26 extended for use in cyclic solvent injection (CSI), where the NMR model can detect
27 oil viscosity changes when the solvent vapor dissolves in oil.

28

29 To **improve the viscosity forecast**, we developed a framework that combines
30 supervised learning algorithms with domain knowledge for synthesizing new
31 features using only one NMR parameter – the T₂ logarithmic mean. Two principal

1 methods were considered, support vector regression (SVR) and gradient boosted
2 trees (GBRT). Models were trained using the experimental data from our previous
3 studies and literature data combining conventional oils, heavy oils, and bitumens
4 from various reservoirs in Canada and the USA. The models' performance was
5 compared against four other intelligent algorithms and four well-known empirical
6 NMR models against which the SVR and GBRT-based models achieved the highest
7 statistical scores. The proposed framework can also be applied to determine other
8 physicochemical properties of oils by LF-NMR, where supervised learning is
9 usually impractical due to the limited volume of data.

10

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

24

25

26

1 **Publications**

2 1. S. Markovic, J. L. Bryan, A. Turakhanov, A. Cheremisin, S. A. Raj Mehta, A. Kantzas.
3 In-situ heavy oil viscosity prediction at high temperatures using low-field NMR
4 relaxometry and nonlinear least squares. *Fuel, Volume 260, 2020.*

5 DOI: 10.1016/j.fuel.2019.116328

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

9 DOI: 10.1016/j.petrol.2020.107456

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

14 DOI: 10.1021/acs.energyfuels.0c02565

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

19

20 **Conferences and conference proceedings**

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

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

28 DOI: 10.3997/2214-4609.20224021

29

30

31

1 **Patents**

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

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

6

7 **Non-thesis publications, relevant for the field**

8 1. A. Ivanova, A. Orekhov, S. Markovic et al., Live Imaging of Micro and Macro
9 Wettability Variations of Carbonate Oil Reservoirs for Enhanced Oil Recovery and
10 CO₂ Trapping/Storage, *Scientific Reports, Nature, 2022*

11 DOI: 10.1038/s41598-021-04661-2

12 2. E. Tokareva, I. Tkachev, G. Sansiev, G. Fedorchenko, A. Ivanova, P. Grishin, S.
13 Markovic et al., Study of the process of hydrophobization of carbonate rock with
14 organic acids (Russian), *Neftyanoe khozyaystvo-Oil Industry, OnePetro, 2022*

15 DOI: 10.24887/0028-2448-2022-3-73-76

16

17

1 **Acknowledgments**

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

5

6 First, I would like to express my most profound appreciation to my thesis jury
7 members for their valuable time and energy in helping me to improve this
8 dissertation: Alexei Tchistiakov, Christoph Arns, Dmitry Eskin, Dmitry Potapenko,
9 and Siddharth Misra. In addition, I am incredibly grateful to them for managing to
10 dive deep into the thesis topic on such short notice and still providing thorough
11 and thought-provoking feedback and ideas for future research.

12

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

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

27 I am very grateful to my co-supervisor, Dmitry Koroteev, for inspiring me to dive
28 deeper into the machine learning field and for being supportive and open to
29 sharing his ideas. I also thank Dmitry for appreciating my research strengths and
30 patiently encouraging me to improve in my weaker areas. I greatly appreciate his

1 advice on writing and communicating our findings better, especially when we
2 were about to submit our first article.

3

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

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

18

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

1 I thank my friends from Skoltech and back home in Serbia for the fun and support
2 during my PhD journey. My sincere thanks go to my parents, Vladimir and Mirjana,
3 and my sisters Milica and Svetlana for their love, support, and understanding. You
4 inspired me to pursue happiness and become a better person.

5

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

10

11

12

1 **Dedication**

2

3 I dedicate this dissertation to my late father, who passed during my PhD studies.

4 Thank you for inspiring me to dive into Geoscience.

5

6

7 *To Vladimir Markovic*

8

9

10

11

12

1 **Contents**

2

3 Introduction23

4 1.1. Heavy oil and bitumen resources.....24

5 1.2. Enhanced Oil Recovery (EOR) methods for heavy oil fields.....25

6 1.3. NMR theory review.....27

7 1.3.1. Spin-lattice (T_1) and spin-spin (T_2) relaxation.....27

8 **1.3.2. Spin-spin (T_2) relaxation mechanisms of fluids in pore space.....32**

9 1.3.3. CPMG pulse sequence configuration and NMR data processing33

10 1.4. Oil viscosity.....36

11 1.4.1. Conventional measurements of oil viscosity.....36

12 1.4.2. Oil viscosity by LF-NMR measurements37

13 1.4.3. LF-NMR oil viscosity measurements in other industrial fields.....40

14 1.5. Water saturation.....41

15 1.5.1. Water saturation by resistivity measurements.....41

16 1.5.2. Water saturation by LF-NMR measurements – T_2 cutoff approach42

17 1.5.3. Water saturation by LF-NMR measurements in oil-sands.....44

18 Chapter 2 Heavy oil viscosity prediction at high temperatures by low-field

19 NMR relaxometry and nonlinear least squares.....48

20 2.1. Motivation.....48

21 2.2. Theory and experiments.....49

22 2.2.1. Spin-spin relaxation (T_2)49

23 2.2.2. Molecular size and intramolecular distance50

24 2.2.3. T_2 -relaxation mechanisms in heavy oils.....51

25 2.2.4. Echo spacing (TE) and relative hydrogen index (RHI_v)53

26 2.2.5. Enhanced NMR viscosity model.....56

27 2.2.6. Preparation of oil samples57

28 2.2.7. Rheological measurements – 23 heavy oil samples.....58

29 2.2.8. Rheological measurements at high temperature– JC bitumen59

30 2.2.9. NMR experiments – 23 heavy oil samples.....60

31 2.2.10. NMR experiments at high temperature – JC bitumen.....61

1	2.2.11.	Nonlinear least squares (NLS) regression – model tuning	61
2	2.3.	Results and discussion.....	64
3	2.3.1.	NMR viscosity prediction – 23 various heavy oil samples	64
4	2.3.2.	NMR viscosity prediction at high temperatures – JC bitumen.....	69
5	2.4.	Summary	76
6	Chapter 3	Improved oil viscosity prediction by low-field NMR using feature	
7		engineering and supervised learning methods	77
8	3.1.	Motivation.....	77
9	3.2.	Methodology.....	79
10	3.2.1.	Gradient boosted regression trees	79
11	3.2.2.	Support vector machines for regression (SVR)	80
12	3.2.3.	Database of rheological and NMR measurements.....	82
13	3.2.4.	Preprocessing and analysis of the dataset.....	82
14	3.2.5.	Feature engineering and transformation.....	82
15	3.2.6.	Evaluation metrics	87
16	3.2.7.	GBRT optimization.....	89
17	3.2.8.	SVR optimization	94
18	3.3.	Results and discussion.....	98
19	3.3.1.	Supervised learning models.....	98
20	3.3.2.	Empirical NMR models.....	103
21	3.3.3.	SVR-FE vs. GBRT-FE	106
22	3.3.4.	Physical implications of SVR-FE and GBRT-FE performance.....	108
23	3.4.	Summary	110
24	Chapter 4	Application of XGBoost model for <i>in-situ</i> water saturation	
25		determination in Canadian oil-sands by LF-NMR and bulk density measurements	
26		112	
27	4.1.	Motivation.....	112
28	4.2.	Theory.....	113
29	4.2.1.	LF-NMR measurements for water saturation determination	113
30	4.2.2.	XGBoost principles.....	113
31	4.3.	Methodology	116

1	4.3.1.	Experimental procedure and data preprocessing	116
2	4.3.2.	XGBoost model based on feature engineering (XGB-FE)	119
3	4.3.3.	XGBoost model based on the full T ₂ relaxation distribution (XGB-FS)	125
4	4.3.4.	Model optimization.....	127
5	4.3.5.	Performance metrics and model validation.....	128
6	4.4.	Results.....	129
7	4.5.	Discussion.....	133
8	4.6.	Summary	138
9	Chapter 5	Conclusions	139
10	A.1	Appendix to Chapter 4 - Prediction performance of other machine learning	
11		models.....	141
12	A 1.1	Feature scaling.....	141
13	A 1.2	Random forests.....	144
14	A 1.3	Gradient Boosting Regression Trees.....	145
15	A 1.4	Gaussian Process Regression.....	146
16	A 1.5	Elastic Net.....	147
17	A 1.6	Support Vector Regression.....	148
18	A 1.7	XGBoost (constrained).....	149
19	A 1.8	Results summary.....	150
20			
21			
22			
23			

1 List of Symbols, Abbreviations

2

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

1

2

1 **List of Figures**

2 Figure 1: Estimated worldwide heavy oil reserves by Country (copyright
3 Schlumberger).24

4 Figure 2: The proton polarization in a B_0 static magnetic field ¹³. T_1 curve is
5 presented as a function of time..... 29

6 Figure 3: Tipping process by 90- and 180-degree pulse. The angle is controlled
7 by B_1 field and τ time period of B_1 application
830

9 Figure 4: Spin echo generation steps. (1) A $90^\circ B_1$ pulse, (2) spin dephasing, (3)
10 180° pulse after τ time flips the spins, (4) rephrasing, (5) spins in phase generate
11 an echo at 2τ31

12 Figure 5: (A) Raw T_2 NMR relaxation data and (B) T_2 data after inversion.....35

13 Figure 6: An example for determination of T_2 cutoff value by LF-NMR and
14 centrifuge. Orange curve presents the T_2 distribution of 100% water saturated
15 sample (S_w 100%). Purple curve presents the T_2 distribution of sample
16 centrifuged to irreducible water saturation (S_{wirr})......43

17 Figure 7: Representative NMR T_2 distributions of two oil-sand samples. (A) An
18 example of distinct oil and water signals where a simple cutoff method can be used
19 for oil-water separation. (B) An example of NMR T_2 distribution with overlapped
20 oil and water signals where deconvolution with T_2 cutoff cannot provide a
21 satisfactory solution. Black vertical dashed lines present potential cutoff times.
22 DS-w and DS-o are percentages of water and oil by Dean-Stark, respectively
23 relative to solids.....45

24 Figure 8: T_1 (black) and T_2 (red) dependence on correlation time (τ_c), according
25 to BPP relaxation model.....52

26 Figure 9: NMR signal amplitude of a single bitumen sample (JC bitumen) in the
27 function of the temperature. The slope of the NMR signal decreases with
28 temperature rise and approximately at $>100^\circ C$, the slope becomes negative due
29 to the Curie effect.....54

1 Figure 10: Relative hydrogen index for a defined volume of JC bitumen sample in
2 function of temperature with implemented correction (red) and without
3 correction for the Curie effect (black).55
4 Figure 11: Flowchart representation of experimental program for 23 heavy oil
5 samples and JC bitumen sample58
6 Figure 12: Dynamic viscosity of JC bitumen in 26 °C – 200 °C temperature range.
7 Extrapolation and interpolation was performed using a model by Khan et al.²⁰
8 ($R^2=0.99$).60
9 Figure 13: Rheological viscosities compared to NMR viscosities of 23 heavy oils.
10 The Markovic et al., (j) model demonstrates the highest accuracy. Solid black line
11 ($x=y$) presents a perfect prediction.66
12 Figure 14: Compared bar chart of adjusted R^2 (a), Root-MSE (b) and MaAE (c) for
13 tuned and default NMR model predictions of 23 heavy oils. Markovic et al. model
14 demonstrates the highest accuracy.68
15 Figure 15: T_2 distribution curves of a single bitumen sample (JC bitumen) in the
16 function of temperature.70
17 Figure 16: Rheological viscosities compared to NMR viscosities for JC bitumen
18 dataset. The temperature scale (top axis) is shown for clarity. Solid black line ($x=y$)
19 presents a perfect prediction.71
20 Figure 17: Compared bar chart of adjusted R^2 (a), Root-MSE (b) and MaAE (c) of
21 NMR model predictions for JC bitumen using three model configurations. Model
22 by Markovic et al. demonstrates the highest accuracy after NLS regression.73
23 Figure 18: Distribution of oil viscosity η before (a), and after the log
24 transformation (b).84
25 Figure 19: Relative feature importance (ranking) by GBRT model of all input
26 features, (a) before, and (b) after removal of redundant TE-derived features with
27 less than <1% relative contribution.86
28 Figure 20: The test set GBRT model performance in terms of least absolute
29 deviations (LAD) for various learning rates (a), and subsample sizes (b) relative
30 to the number of trees M . Bottom plot (c) illustrates the model accuracy evaluation
31 as a function of M , in terms of MAE, RMSE, and MSLE.92

1 Figure 21: Test set SVR model performance in terms of log-normalized RMSE for
2 various values of ϵ (a), and γ (b) with respect to regularization C. Bottom plot (c)
3 illustrates the accuracy of optimized SVR model as a function of C, in terms of three
4 error metrics; log(RMSE), log(MAE) and MSLE.96
5 Figure 22: Comparison of NMR SL viscosity model predictions and observations.
6 Note that the grayscale points are predictions of models generated without FE,
7 while warm color points are predictions with FE. Lighter colors indicate lower
8 temperatures (from 25 °C/299 K), and more intense, darker colors indicate higher
9 temperatures (up to 200 °C/466 K). GBRT and SVR models with integrated FE
10 demonstrate the best performance. 100
11 Figure 23: Compared statistical scores of SL models without FE (a) and SL models
12 with integrated FE (b). SVR-FE and GBRT-FE demonstrate the best statistical
13 performance..... 101
14 Figure 24: Performance comparison of empirical NMR viscosity models (a, b, c,
15 and d) with SVR-FE (e) and GBRT-FE (f) models. GBRT-FE and SVR-FE
16 demonstrate significantly better performance. 104
17 Figure 25: Compared statistical scores of four empirical NMR viscosity models
18 and SVR-FE and GBRT-FE supervised learning models in terms of RMSE, MAE, and
19 MSLE. SVR-FE and GBRT-FE demonstrate significantly better statistical
20 performance..... 105
21 Figure 26: Percent error box plots (a) and MAPE scores (b) for six supervised
22 learning models with feature engineering, and four empirical models. Note that in
23 the plot (a) the y-axis is in log-scale. GBRT-FE model demonstrates the best
24 performance in terms of MAPE. 107
25 Figure 27: Flowchart representing the experimental program for oil-sands
26 samples, by X-Ray CT, LF-NMR T₂ measurements and Dean-Stark extraction... 118
27 Figure 28: Flowchart for XGB-FE model development. 119
28 Figure 29: Results of the mutual information regression applied to the training set
29 T₂ distributions of the oil-sand samples relative to the Dean-Stark water content
30 (DS-w). The shaded area presents the continuous cluster of T₂ responses with a

1	strong mutual association with DS-w, which were used for the calculation of the	
2	T ₂ cutoff range parameter – T _{2cr}	122
3	Figure 30: The diagonal correlation matrix showing the amount of linear	
4	dependence between six input features with Dean-Stark water content (DS-w).	
5	Scores represent the Pearson’s correlation coefficient and are color coded	
6	(heatmap).	123
7	Figure 31: Mutual information regression scores for five NMR parameters and	
8	bulk density (input features) relative to the Dean-Stark water content (DS-w).	
9	124
10	Figure 32: Flowchart for XGB-FS model development.....	126
11	Figure 33: Evaluation of XGB-FE, XGB-FS and Bryan et al. ⁷ , performance by cross-	
12	plots between the model predictions and observed saturation in %DS-w (1, 4, 7),	
13	distribution of regular residuals (2, 5, 8), and quantile-quantile plots for	
14	comparing distributions between predictions and observations and evaluating	
15	normality of residuals (3, 6, 9).	131
16	Figure 34: Comparison of RMSE and MAE test prediction scores for the three	
17	models (<i>‘random_state=2’</i>).	132
18	Figure 35: Leave-one-out cross-validation (LOOCV) scores for XGB-FS and XGB-	
19	FE machine learning models for the training set with fixed random split seed	
20	<i>‘random_state=2’</i> . Note y-axis was truncated for convenience.	133
21	Figure 36: Cross-plot of Random Forest model train and test	
22	predictions.....	144
23	Figure 37: Cross-plot of Gradient Boosting Regression model train and test	
24	predictions	145
25	Figure 38: Cross-plot of Gaussian Process model train and test	
26	prediction.....	146
27	Figure 39: Cross-plot of Elastic Net model train and test	
28	predictions.....	147
29	Figure 40: Cross-plot of Support Vector Regression model train and test	
30	predictions	148

1	Figure 41: Cross-plot of constrained XGBoost model train and test	
2	predictions.....	149
3	Figure 42: Comparison of RMSE and MAE test set prediction scores for seven	
4	models.....	150
5		
6		
7		
8		

1 **List of Tables**

2 Table 1: Tested literature NMR viscosity correlations.....64

3 Table 2: Descriptive statistics of six input features (input variables) used for

4 training of SL models, and a target variable which is log-transformed viscosity

5 (Log(η)).....87

6 Table 3: Results of four loss functions used for optimizing the model performance

7 after 5-fold cross-validation. The LAD function exhibits the best performance

8 based on MAE_{cv}, RMSE_{cv}, MSLE_{cv}, \bar{R}^2_{cv} cross-validation (CV) scores.....89

9 Table 4: Results of three commonly used loss functions for estimating the node

10 splitting quality after 5-fold cross-validation. The MAE function performs optimal

11 splitting based on MAE_{cv}, RMSE_{cv}, MSLE_{cv}, \bar{R}^2_{cv} cross-validation (CV) scores.....90

12 Table 5: GBRT hyperparameters optimization by grid-search based on 5-fold

13 cross-validation91

14 Table 6: SVR hyperparameter optimization by grid-search based on 5-fold cross-

15 validation95

16 Table 7: GS-CV hyperparameter optimization results for all supervised learning

17 algorithms which were tested in this work.....99

18 Table 8: Compared view of statistical scores for all SL and empirical models.

19 Bolded values correspond to the best score..... 108

20 Table 9: Optimal CPMG pulse sequence parameters for detection of fast relaxing

21 clay-bound water and heavy oil signals 117

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

23 development..... 125

24 Table 11: Descriptive statistics of the target variable Dean-Stark water saturation

25 (DS-Sw)..... 128

26 Table 12: Results of Bayesian Optimization with 5-fold cross-validation, for XGB-

27 FS and XGB-FE models..... 134

28 Table 13: Comparison of XGB-FS model performance with and without bulk

29 density parameter. 134

30 Table 14: Comparison of XGB-FE model performance with and without bulk

31 density parameter.....126

1 **INTRODUCTION**

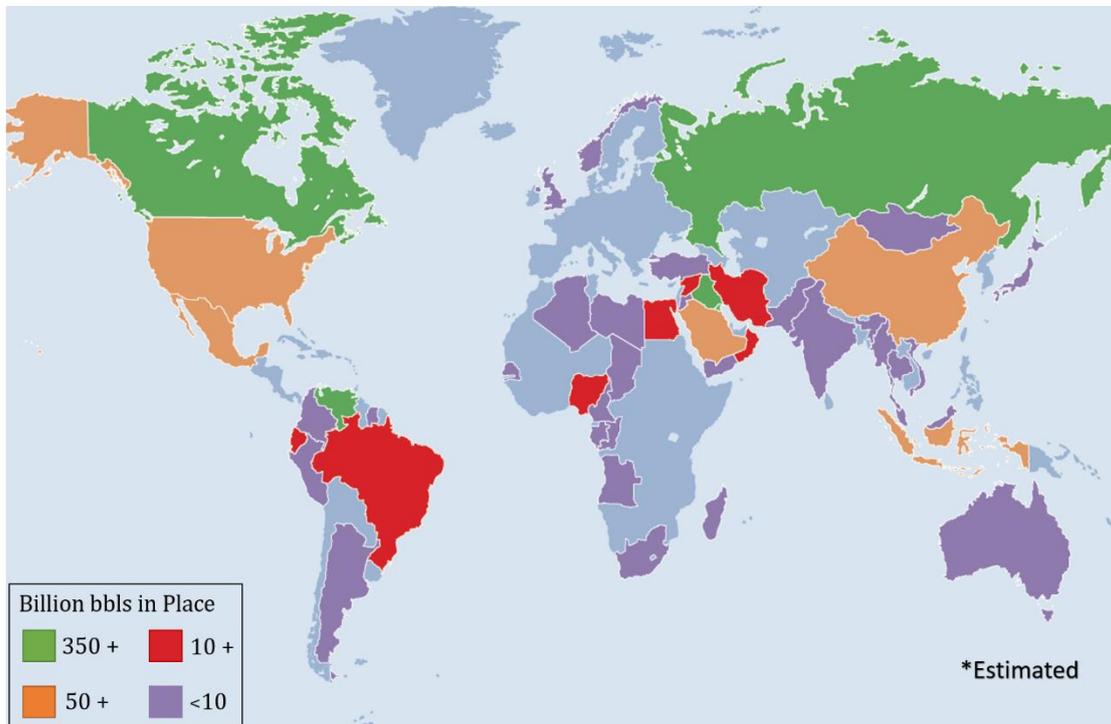
2 This work studied different types of unconventional hydrocarbons, namely oil
3 sands, heavy oils, and bitumens. These hydrocarbons represent the most
4 abundant unconventional hydrocarbon resources worldwide. Due to their high
5 viscosity, thermally enhanced oil recovery (EOR) methods are often required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

12
13 In light of the foregoing, I focused my research on advancing conventional and
14 establishing new analytical and data-driven methods for *in-situ* characterization
15 of heavy oil and bitumen resources, primarily using LF-NMR measurements.
16 Special attention is paid to developing new workflows and models for predicting
17 oil viscosity and water saturation, two critical factors for the recovery of
18 hydrocarbons.

19
20
21
22
23
24
25
26
27

1 **1.1. Heavy oil and bitumen resources**

2 Bitumens and heavy oils usually form deposits in shallower geological settings,
3 where due to the cold temperatures and a lack of the caprocks, they are being
4 subjected to bacterial biodegradation. The depth of these deposits usually does
5 not exceeds 4 km, while the temperatures are most often lower than 80 °C¹. In
6 such conditions, the biodegradation process spans over geological periods, where
7 hydrocarbon-degrading bacteria use lighter oil fractions for the metabolic
8 processes, which gradually reduces the viscosity, initial mass, and gas-to-oil ratio
9 (GOR) of the oil.



10
11 Figure 1: Estimated worldwide heavy oil reserves by Country (copyright Schlumberger).

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

1 technologies and time-demanding processing. However, their large in-place
2 volumes and high oil market prices allow their profitable extraction and
3 refinement, so major oil companies acquire licenses for their development as an
4 alternative to conventional oil resources.
5 Studies estimate that heavy oil and bitumen reserves today amount to about 55%
6 of the total world reserves (nearly 9 trillion barrels), while the most extensive
7 deposits were found in Canada, Venezuela, and Russia¹.

8

9 **1.2. Enhanced Oil Recovery (EOR) methods for heavy oil fields**

10 Heavy oils and bitumens have distinctively high viscosities, so their recovery from
11 the reservoir rocks is often performed using various thermal Enhanced Oil
12 Recovery (EOR) techniques such as hot steam injection or in-situ combustion,
13 where the heat exchange process reduces oil viscosity in the formation of
14 interest². In heavy oil reservoirs, the viscosity may vary up to a hundred times in
15 vertical and horizontal directions. Therefore information about the spatial
16 distribution of viscosity can affect not only the well placement and injection or
17 production rates but mathematical reservoir simulations as well³.

18

19 In literature, the EOR methods are generally divided into two groups: cold and
20 thermal production. The cold production methods are economically and
21 technically convenient recovery methods; however, they are limited to shallow
22 deposits since they involve open-pit mining. Waterflooding was also occasionally
23 used with some success, but in fields where oil viscosity was not more than 100
24 cP, the sweeping efficiency of the waterfront reduces with an increase of viscosity,
25 most notably due to the viscous fingering¹. Vapor-assisted extraction (VAPEX) is
26 another cold production method that proved efficient for reservoirs with oil
27 viscosities ranging from 100 cP up to 130,000 cP^{4,5}. It is based on miscible solvent
28 vapor injection into parallel stacked horizontal wells. The miscible solvent vapor
29 is injected into the upper well, where the vapor chamber is produced around the
30 well. This causes the dilution of solvent vapor into surrounding heavy oil and

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

5
6 In thermal methods, on the other hand, increasing the temperature in the well or
7 reservoir causes viscosity to decrease, thus improving the heavy oil and bitumen
8 mobility. For instance, in steam-assisted gravity drainage (SAGD), the placement
9 of the injection and producer wells is almost identical to the one in the VAPEX
10 method; however, instead of the solvent vapor, hot steam is injected. After the
11 hydrodynamic linking between wells is obtained by the initial steam treatment of
12 both wells, steam injection is continued in the injection well, where hydrocarbons
13 are affected by the steam expansion. The temperatures of this treatment can be
14 over 200 °C, thus initiating a highly mobile gravitational drainage of bitumen and
15 heavy oil towards the underlying producing well. This method and its' variations

16 were successfully utilized in numerous projects, most notably in Canadian heavy
17 oil reservoirs where viscosities vary from a couple of hundreds up to $3 \cdot 10^6$ cP ⁶. In
18 addition to the SAGD, cyclic steam stimulation (CSS) is used, where the same well
19 is used for injection and production. In this approach, hot steam is injected into
20 the targeted formation and left to soak up to heat the oil, followed by the
21 production cycle. Reports in the literature show that CSS is suitable for oil
22 viscosities from 50 cP to 350,000 cP ⁷. As a follow-up to CSS, steam flooding is
23 usually performed, where steam is continuously injected into the targeted
24 formation without a soaking period. Since this process is heavily influenced by
25 horizontal and vertical sweep efficiency, much attention is given to monitoring
26 steam fingering and channeling, which is why steam flooding has shown to be
27 most successful for reservoirs with oil viscosity ranging from 20 – 20,000 cP ⁸.

28 Another well-known thermal method is in-situ combustion (ISC) or fire flooding,
29 where an oxidizing gas (i.e., air) is injected into the formation, which causes the
30 ignition of heavy oil in-situ. In this set-up, the portion of the hydrocarbon's heavier
31 components is utilized as a fuel for further propagation of the combustion front,

1 where exothermal reaction heats the surrounding rocks and thus lowers the oil
2 viscosity. An added benefit of ISC is that the heavier components are being
3 consumed due to the thermal cracking, which results in upgraded oil. Although
4 there were many pilot studies for ISC between the 1960s - 1980s in heavy oil
5 reservoirs with oil viscosities from $40 \cdot 10^3$ to $1 \cdot 10^6$ cP, the commercially successful
6 ISC projects were performed in reservoirs with oil viscosity between 20 – 8000 cP
7 ⁹.

8

9 The success of the EOR project in heavy oil and bitumen deposits is dictated
10 mainly by the initial viscosity of the hydrocarbons and their spatial distribution in
11 the field. In steam EOR, for instance, monitoring of the steam front is often
12 performed using periodic or permanent four-dimensional (4D) seismic surveying
13 by collecting pressure and fluid saturation data within the field¹⁰. Monitoring
14 based on logging is also commonly implemented by fiber-optic distributed
15 temperature sensing (DTS), which provides constant temperature monitoring
16 along the wellbore. NMR measurements have been considered for monitoring the
17 fluid saturation changes and wettability alteration assessing^{10,11}. If this technology
18 is used for reservoir monitoring, NMR logging could also be used as a continuous,
19 non-invasive technique for monitoring viscosity variations and changes caused by
20 temperature variation in the reservoir. The NMR equipment can be used in-situ in
21 observation wells, online, or inline for the heavy oil viscosity monitoring during a
22 thermal EOR project with a 200 °C upper-temperature limit, which corresponds
23 to steam injection temperatures, and other EOR methods such as solvent or
24 miscible gas injection ¹².

25

26 **1.3. NMR theory review**

27 **1.3.1. Spin-lattice (T₁) and spin-spin (T₂) relaxation**

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

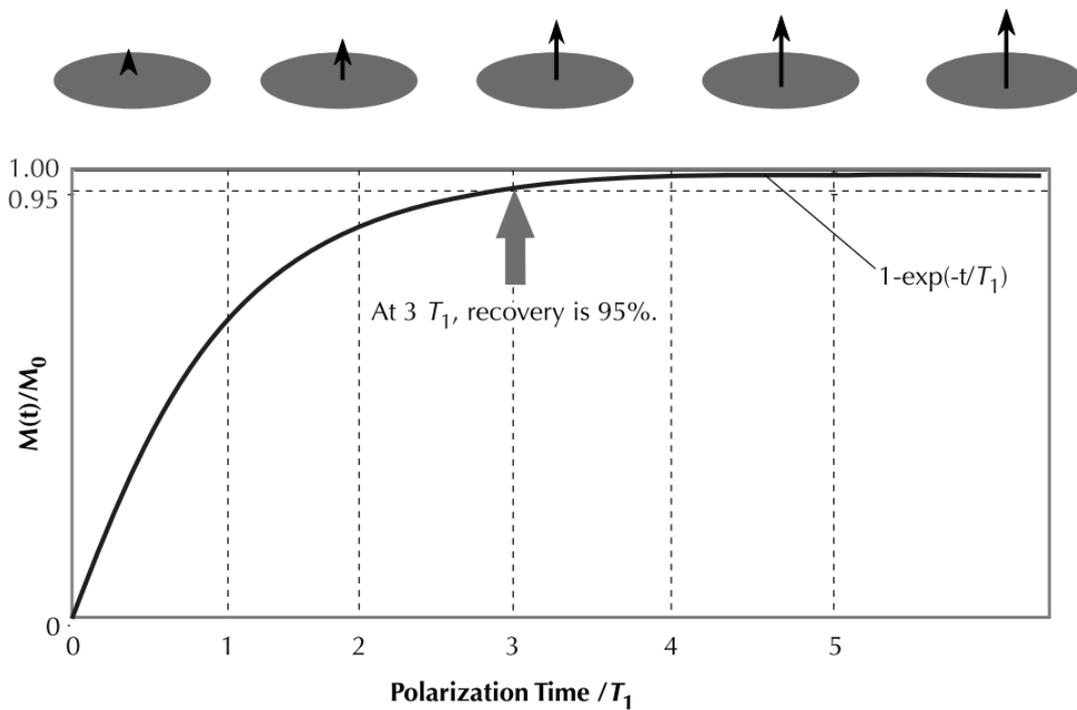
1 magnetic field can control the orientation of their spins. For instance, if a magnetic
2 field (B_0) is introduced to the H^+ proton-containing system, the proton spins will
3 tend to align along the direction of the B_0 . The quantum theory posits that the
4 protons will be distributed to a low or high energy state corresponding to spins -
5 $\frac{1}{2}$ and $\frac{1}{2}$. The difference between the number of protons in high and low energy
6 states will generate the total or bulk magnetization M_0 that NMR tools can
7 measure.

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

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

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

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



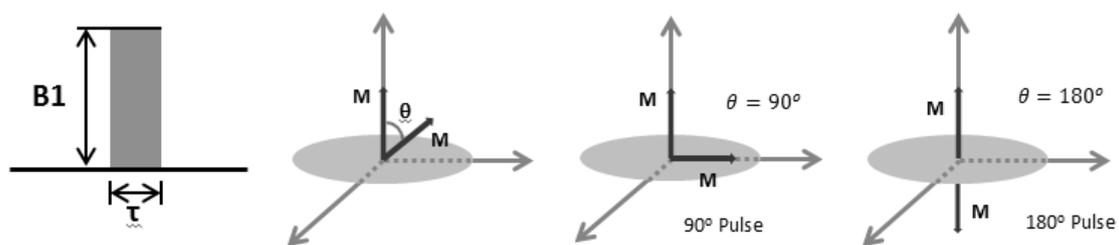
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Figure 2: The proton polarization in a B_0 static magnetic field¹³. 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:

$$\theta = \gamma B_1 \tau \tag{3}$$

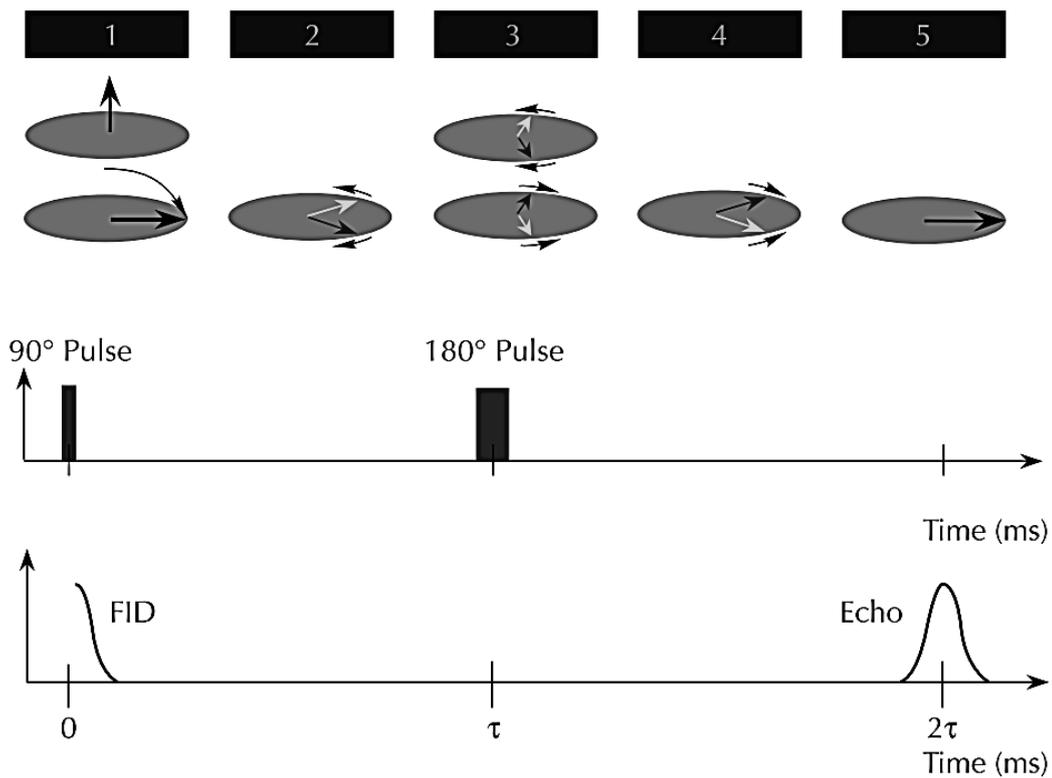
where τ is the period when the B_1 is applied to the system. The tipping angle can be controlled by both B_1 and τ .



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

3
 4 In practice, the first B_1 pulse tips the proton system into the perpendicular plane
 5 (x, y) relative to the B_0 (z), thus $\theta = 90^\circ$. At this point, the protons are precessing
 6 about B_0 and are in phase, and the NMR device can detect their signal. However,
 7 immediately as the B_1 stops, the protons start to dephase due to the B_0 magnetic-
 8 field inhomogeneity and molecular tumbling, usually at an exponential rate.
 9 Therefore, their signal decays at the particular time constant (T_2^*) and is called
 10 free induction decay (FID).

11
 12 Fortunately, the dephasing due to the B_0 inhomogeneity can be reversed by adding
 13 the B_1 pulse ($\theta = 180^\circ$) after τ time elapsed from the first pulse ($\theta = 90^\circ$). The
 14 second pulse flips the protons, which reverses the dephasing process, meaning
 15 that the protons will return to the same phase after τ time elapsed from the second
 16 pulse. This subsequent pulse is referred to as a 'refocusing pulse.' When spins
 17 return to the same phase, the 'spin echo' signal is produced. Generation of one spin
 18 echo is illustrated on Figure 4. The NMR device can be configured to produce a
 19 series of refocusing pulses, thus generating a series of spin echoes. This series or
 20 sequence of refocusing 180° pulses is known as Carr-Purcell-Meiboom-Gill
 21 (CPMG) pulse sequence, while the recorded sequence of echoes is called the 'spin
 22 echo train' ¹³.



1

2 **Figure 4: Spin echo generation steps. (1) A 90° B₁ pulse, (2) spin dephasing, (3) 180°**
 3 **pulse after τ time flips the spins, (4) rephasing, (5) spins in phase generate an echo at**
 4 **2τ¹³.**

5 Although the CPMG pulse sequence can reverse the effect of B₀ field
 6 inhomogeneity, the dephasing of spins due to molecular tumbling and diffusion is
 7 irreversible. Since the dephasing is mainly due to the interaction of spins, the
 8 decay of magnetization in the horizontal plane (x, y) is called spin-spin relaxation,
 9 or transversal relaxation, as the magnetization is in the transversal plane relative
 10 to B₀. The time constant associated with the decay rate is T₂ relaxation time.

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

11 where $M_{xy}(0)$ is magnetization magnitude at $t=0$. According to Equation 4, for one
 12 T₂ time constant, the magnitude of M_{xy} will drop to ~37% of its initial value, and
 13 after three T₂ constants to ~95%.

14

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:

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

Here, T_2 is the spin-spin relaxation time of fluids in pores and results from a sum of $T_{2\text{bulk}}$ and $T_{2\text{surface}}$ relaxation and relaxation due to $T_{2\text{diffusion}}$. The $T_{2\text{bulk}}$ 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 $T_{2\text{bulk}}$ in practice are pressure and temperature oscillations. It has been shown that $T_{2\text{bulk}}$ relaxation of water and dead oil are generally^{13,15}:

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

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

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

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

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

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 $T_{2\text{surface}}$ relaxation is that it

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

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

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

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

11 **1.3.3. CPMG pulse sequence configuration and NMR data processing**

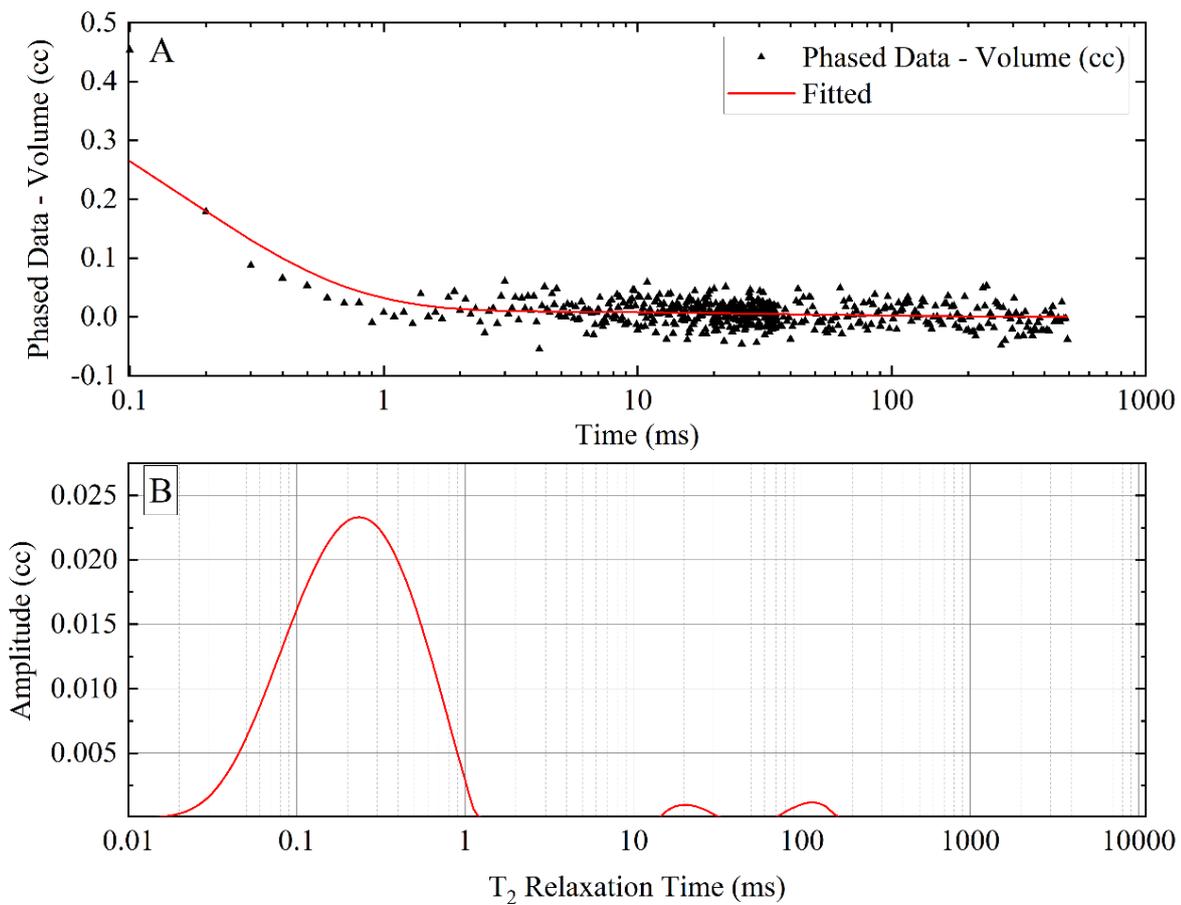
12 The CPMG pulse sequence consists of a 90° initial pulse that tips the polarized
13 protons into an x, y-plane, followed by several refocusing 180° pulses. The quality
14 of the obtained CPMG decay data, its inversion, and the quality of interpretation
15 are strongly dependent on the configuration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

19
20 Setting an NMR tool to a short echo spacing (TE) will influence the signal-to-noise
21 ratio (SNR) twofold; first, the density of spin-echoes within a train will increase;
22 second, the echoes will be recorded earlier. Consequently, this leads to increased
23 SNR. However, the experiment time will increase proportionally when the number
24 of echoes (NE) increases. The same is true for the number of trains (NT) and
25 waiting time (TW). It should be noted that TW should be configured according to
26 the sample or reservoir interval and the purpose. If the goal is to characterize
27 heavy oil and clay-bound water, the TW can be decreased since protons will return
28 to equilibrium much faster than pure water or light oil.

1 The representation of the decaying echo train in the time domain is usually
2 obtained by the Laplace Inverse Transform (ILT)¹⁶, whereas the output of a T₂
3 distribution is obtained (Figure 3). However, the inversion of the NMR signal
4 represents the ill-posed problem since minor perturbations (i.e., noise) in the
5 measurements can substantially impact the T₂ distribution form, that is, the
6 stability of the solution. A stable and unique solution can be obtained if the
7 inversion of the signal is performed numerically. In such a case, the representation
8 of the T₂ decay is achieved from echo-fitting. To simplify the echo-fitting, the
9 discretization of the T₂ decay signal can be performed, where the predefined
10 number of discrete T_{2i} relaxation times correspond to individual exponential
11 decay. Then, the set of echo trains can be expressed as a system of linear equations
12 where each equation corresponds to an individual echo train. Since the fitting is
13 performed to the sum of multi-exponentials, the stable solution can be obtained
14 by solving non-negative least squares. The standard approaches are Lawson-
15 Hanson and Tikhonov regularization¹⁷.

16

17



1

2

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

3

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 information¹⁷. 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

13

1 be carefully tuned to achieve a balanced tradeoff between signal-to-noise ratio,
2 the shape of T₂ distribution, and measurement time.

3 1.4. Oil viscosity

4 1.4.1. Conventional measurements of oil viscosity

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

$$\tau_{ss} = \eta \cdot \gamma_{sr} \quad (10)$$

16 where τ_{ss} presents the amount of shear stress applied to the fluid, γ_{sr} stands for the
17 shear rate, and η is the fluid viscosity. Based on the Eyring's theory of viscosity,
18 molecules in fluids are structured in lattices, while intermolecular space remains
19 vacant but not spacious enough for other molecules to shift through readily²⁰.
20 However, if force is applied, the molecules will reconfigure their positions until
21 the vacant space becomes large enough for another molecule to enter. Eyring
22 proposed an analytical model for such behavior (Equation 11)

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

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

1 molecules, cyclic paraffin, and branches from heavy components. To mobilize
2 these molecules, more activation energy is required (G_0) since the attractive forces
3 of surrounding molecules hinder the movement of molecules attempting to
4 occupy the vacant space. In rheological measurements of heavy oils by a cone and
5 plate viscometers, this resistance to flow translates to high shear stresses. From
6 Equation 11, it can be observed that the temperature is in an exponential
7 relationship with viscosity. If we apply heat to the heavy oil, the heavy components
8 will gain more energy while the intermolecular distance will increase, enabling the
9 less restricted motion of molecules. This is also consistent with rheological
10 viscosity measurements, where for fixed shear stress and with an increase in
11 temperature, we observe increased shear rate¹⁸. Equation 6 can be re-written by
12 substituting constants (Equation 12).

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

13 where A is a constant and E_a is the viscous free energy of activation. In this
14 Arrhenius-type equation, the values of E_a and A vary for different oil samples.
15 These variations are associated with molecular weight variation of different oil
16 components and their chemical composition and structure²⁰. These variations
17 may be substantial since the composition of heavy oils and bitumen, in particular,
18 can be significantly different, indicating that derivation of the general viscosity
19 model is challenging.

20 **1.4.2. Oil viscosity by LF-NMR measurements**

21 The oil is a blend of a diverse range of liquid hydrocarbons with inconsistent
22 molecular structure²¹. When it has a higher proportion of complex high molecular
23 weight compounds such as asphaltenes and resins, oil viscosity will be higher,
24 signifying that viscosity reflects oil's chemical complexity²². This natural
25 inconsistency of oil compositions elicits a constant demand for the development
26 of new techniques for their efficient characterization. In recent years, the wave of
27 innovation has led to the application of low-field nuclear magnetic resonance (LF-
28 NMR) tools to characterize hydrogen-bearing liquids due to their ability to rapidly

1 convey a series of contactless, non-invasive experiments. In NMR petrophysical
2 logging, the pulsed NMR tools typically generate magnetic fields between 120 to
3 550 gauss (0.012 to 0.055 tesla) ²³. In this work, low-field NMR is defined by fields
4 below 550 gauss.

5
6 To relate a fluid viscosity to NMR relaxation, it is necessary to comprehend and
7 model the molecular interactions. Equations 11 and 12 demonstrate that viscosity
8 can be expressed without macroscopic flow or shearing (Equation 10). These
9 findings are consistent with a theoretical Bloembergen-Purcell-Pound (BPP) NMR
10 relaxation model, associated Debye-Einstein-Stokes spherical molecules model,
11 which anticipates different rotational correlation times (τ_c) for various molecule
12 sizes ^{19,24,25}. The correlation time represents the mean time required for a
13 molecule to rotate one radian and is a crucial parameter for determining
14 microviscosity. It is also a fundamental component of the BPP relaxation model.
15 Based on this relationship, one can study the association of T_2 relaxation with
16 viscosity²⁵. The physics of this relationship is discussed in section 2.2.

17
18 As previously mentioned, the T_2 relaxation distribution after mathematical
19 inversion can be represented in a time domain. Distributions of T_2 relaxation
20 between different fluids can be compared using a mean T_2 distribution time, such
21 as T_2 logarithmic mean (T_{2lm}). When fluids are measured in a bulk state, the
22 primary relaxation mechanism will be bulk relaxation (T_{2B}), which is due to the
23 energy exchange of the H spins and diffusion. Straley et al.²³ and Coates et al.¹³
24 have experimentally shown that the T_{2B} is proportional to the ratio of temperature
25 (T) and viscosity (η):

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

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

1 light and medium viscous oils ($\sim 1\text{-}800\text{ cP}$)²⁸ composed of lighter hydrocarbons
2 with a less complex chemical structure. In the case of heavier oils, the T_2 relaxation
3 deviates from the classical BPP model, and the relationship described in Equation
4 13 alters significantly.

5 In the past 30 years, many analytical NMR viscosity models have been proposed
6 for characterizing crude oil. However, the reports in the literature show
7 inconsistency in the prediction accuracy of these models due to three main
8 reasons: use of the light oil NMR models for the prediction of heavy oil
9 viscosity^{15,23,27,29}, use of models (including heavy oil models) without prior tuning
10 for a given reservoir or a suit of oils, and due to use of ambiguous mathematical
11 procedures for model tuning^{3,30,31}. Moreover, there were several attempts to
12 develop a “universal model” for *in-situ* heavy oil viscosity prediction aiming to
13 estimate viscosity in the formations with weak prior knowledge about the oil
14 properties³⁰⁻³³. These models have default parameters derived for heavy oils from
15 a particular oil field. However, when applied to different heavy oils, they generate
16 significant prediction errors, in some cases over a factor of three³¹. To develop a
17 universal analytical NMR viscosity model for systems with oils of various
18 compositions would be contradictory to Debye-Stokes-Einstein’s findings stating
19 that different correlation times are expected for different molecule sizes ²⁴.
20 However, for the order-of-magnitude *in-situ* viscosity evaluation, existing
21 analytical models can be improved to a degree where more reliable estimates can
22 be utilized to optimize the decision-making process for viscosity variation
23 monitoring during EOR projects.

24

25 Although LF-NMR technology has been proved to be a viable tool for observing
26 differences in variable viscosity oils, numerous constraints arise as a consequence
27 of not only the embedded chemical complexity of oils and limitations of LF-NMR
28 devices but also from analytical tools and models used for the interpretation of
29 experimental results ^{31,33-37}. Since the former two are technologically challenging
30 to change, one can attempt to improve the analytical tools and frameworks using
31 new mathematical approaches. In such circumstances, the supervised learning

1 (SL) methods have been proven helpful in developing more reliable mathematical
2 models in many relevant fields such as fuel processing, petrophysical studies of
3 porous mediums, and oil viscosity monitoring equipment in mechanical systems
4 ³⁸⁻⁴².

5 **1.4.3. LF-NMR oil viscosity measurements in other industrial fields**

6 The application of NMR viscosity models is not relevant only for petrophysical
7 well-logging. One potential application is in fuel processing, where there has been
8 a surge for the last few years in the development of fast methods for the
9 characterization of petroleum fractions by LF-NMR⁴³. Among many studied
10 physicochemical properties, the oil viscosity showed to be of the principal
11 importance in determining the rate of interaction with fuel during combustion
12 processes in internal combustion engines ^{44,45}. In these studies, the LF-NMR
13 predictive models were typically derived using multivariate calibration with
14 partial least squares (PLS) regression or artificial neural networks (ANN), which
15 proved efficient. However, in nonlinear datasets, the reports in the literature show
16 that PLS did not provide satisfactory accuracy, whereas ANN tended to overfit the
17 data, thus leading to poor model generalization^{46,47}. In the LF-NMR examination of
18 petroleum fractions, this nonlinearity can occur due to their chemical intricacy,
19 leading to the degradation of model forecasting performance²⁴.

20

21 Moreover, in mechanical systems (tribosystems), viscosity reflects the oil's
22 capacity to render the sufficient thickness of the lubricating film between the
23 surfaces exposed to friction. In order to efficiently buffer the rate of machinery
24 wear, the oil selection is made under the speed-load and temperature conditions
25 of the system⁴⁸. The prevention of malfunctions in tribosystems is usually
26 performed by monitoring oil viscosity, where its relative increase may indicate
27 excessive oxidation or contamination of the oil by other fluids. In contrast, its
28 decrease may indicate the beginning of a thermal cracking process, occurring at
29 high temperatures⁴⁹. In earlier studies, LF-NMR measurements were proposed as
30 an alternative to conventional monitoring approaches that involve direct-contact

1 instruments based on vibration, acoustic, and micro-displacement methods ⁴². In
2 circumstances where these instruments would be difficult to utilize, the LF-NMR
3 tools could be used instead for non-invasive, real-time viscosity monitoring^{42,50,51}.
4 As the operating conditions of these systems may lead to significant oil viscosity
5 fluctuations, the robust data-driven or analytical NMR model could be used to
6 measure the fluctuations accurately and, in that manner, help in the early
7 detection of equipment failure.

8
9

10 **1.5. Water saturation**

11 **1.5.1. Water saturation by resistivity measurements**

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

$$S_o = 1 - S_w \quad (14)$$

16 where S_o is oil saturation and S_w is water saturation. Conventionally, in well-
17 logging practice, the water saturation was determined using resistivity logs, and
18 depending on the reservoir type, resistivity data would be used in combination
19 with other standard logs such as density (RHOB) and neutron (NPHI)⁵². The
20 widely used empirical model for water saturation estimation in conventional
21 hydrocarbon reservoirs was developed by Archie⁵³ (Equation 17).

$$F = \frac{R_o}{R_w} = \frac{a}{\phi^m} \quad (15)$$

$$I_r = \frac{R_t}{R_o} = \frac{1}{S_w^n} \quad (16)$$

$$S_w = \sqrt[n]{\frac{aR_w}{R_t\phi^m}} \quad (17)$$

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

11

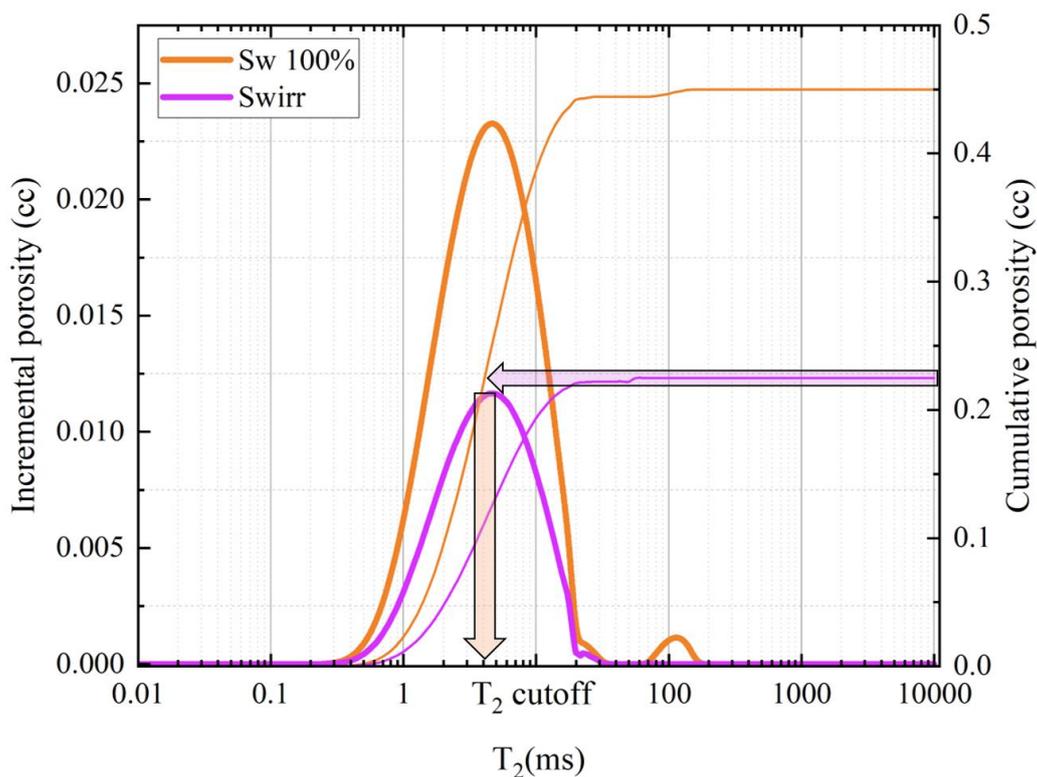
12 Archie's equation was successfully used in systems with simple, uniform pore
13 space saturated by water⁵⁴. However, issues arise in reservoirs with large
14 amounts of clay-bound and capillary-bound water, strong variations of salinity
15 with depth, and formations containing clays or conductive minerals such as
16 pyrite^{55,56}. This is also true for Canadian oil-sands⁵⁷. In these terms, the presence
17 of bound water and clays will cause the underprediction of OOIP, while the
18 variable salinity can cause either overprediction or underprediction of OOIP.

19

20 **1.5.2. Water saturation by LF-NMR measurements - T₂ cutoff** 21 **approach**

22 Since the LF-NMR tools measure the response H⁺ protons of the fluids, many of the
23 issues relevant for resistivity logging can be avoided. Another advantage of NMR
24 measurements is differentiating between irreducible water saturation (capillary
25 and clay bound water) and producible fluids. The primary assumption is that
26 larger pores are saturated by producible fluids, where the flow can occur in the
27 presence of pressure gradient, while smaller pores contain fluids trapped by
28 capillary forces or are bound within the lattice of clay minerals. If this condition is
29 true, a T₂ cutoff value (location in T₂ distribution) can be defined, separating the
30 T₂ distribution to signals corresponding to clay-bound, capillary-bound and

1 producible fluids. Integration of the separated regions can be performed and
 2 related to producible and bound fluid volumes. Straley et al.,²³ were the first to
 3 empirically identify the universal T_2 cutoff for clay-bound water at 3 ms in
 4 conventional sandstones. The cutoff value was determined by comparing the clay-
 5 bound water calculated from the ratio of cation exchange capacity and pore
 6 volume (Q_{ve}) with cumulative T_2 distribution porosity, using sandstone core plugs
 7 from 45 American and European oilfields. To evaluate the producible porosity or
 8 free-fluid index (FFI), it was necessary to perform NMR T_2 measurements on core
 9 samples in two states – cleaned and fully saturated state (S_w 100%) and after
 10 centrifuging to the irreducible water saturation (S_{wirr}). These experiments were
 11 performed for the suit of 86 sandstone samples, and the universal cutoff was found
 12 to be at 33 ms²³.



13
 14 Figure 6: An example of the determination of T_2 cutoff value by LF-NMR and centrifuge.
 15 The orange curve presents the T_2 distribution of the 100% water-saturated sample (S_w
 16 100%). The purple curve presents the T_2 distribution of the sample centrifuged to
 17 irreducible water saturation (S_{wirr}).
 18

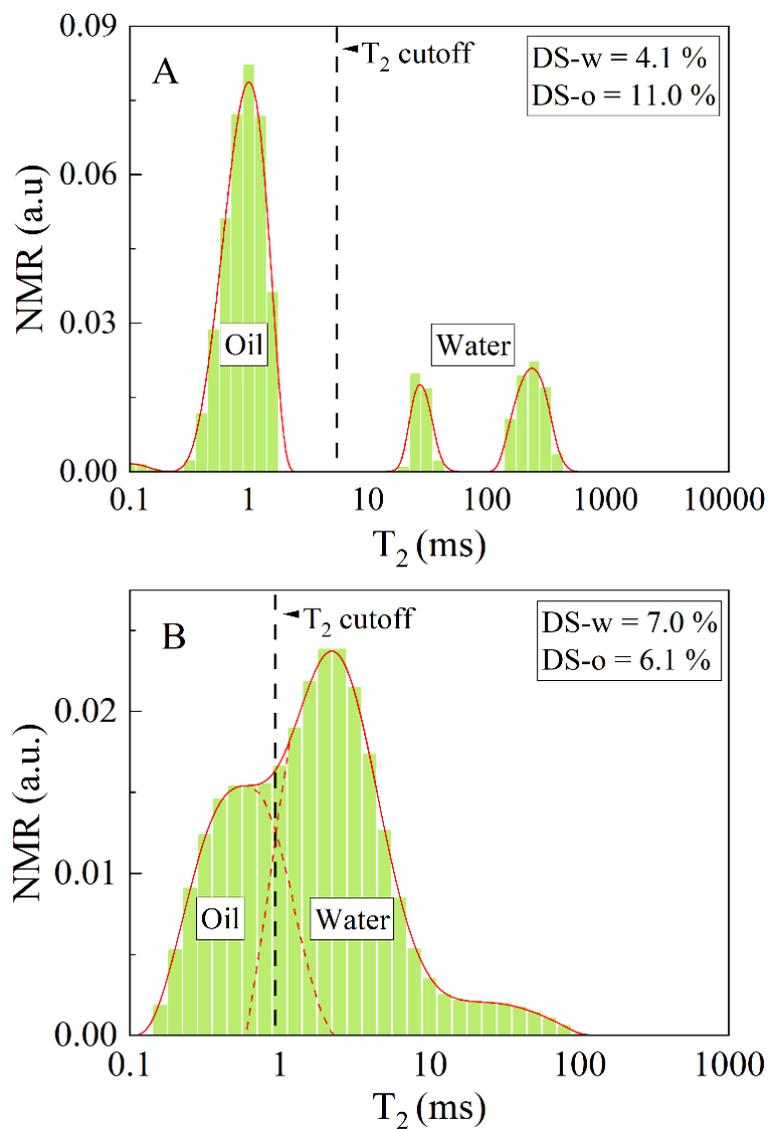
1 Although these universal cutoff times work for conventional sandstone reservoirs
2 where porosity and permeability are generally uniform, many recent studies have
3 shown that T_2 cutoff values vary dramatically for reservoirs of other lithologies,
4 such as shales, carbonates, oil-sands, coals, and tight sandstones⁵⁸. In addition,
5 even if the T_2 cutoffs are determined experimentally, it is not recommended to use
6 a single T_2 cutoff for the same well or oilfield since the T_2 distribution can vary
7 drastically both vertically along the well and laterally, which would potentially
8 cause erroneous estimation of OOIP⁵⁸.
9

10 **1.5.3. Water saturation by LF-NMR measurements in oil-sands**

11 The T_2 cutoff determination for oil-sands is even more problematic for two
12 principal reasons. First, the centrifuging to S_{wirr} cannot be adequately performed
13 since the sand exposed to severe centrifugal forces will lose its structure and
14 native filtration properties, rendering the subsequent NMR experiments
15 inadequate. Second, the heavy oil and bitumen in oil-sands have a fast T_2
16 relaxation time and produce a signal in the same region as capillary and clay-
17 bound water, causing a significant overlap. One of the well-known approaches that
18 had considerable success in addressing this effect is based on NMR spin-spin
19 relaxation (T_2) distribution peak deconvolution⁵⁹. The principal assumption
20 behind this approach is that bitumen relaxes faster in an NMR distribution than
21 surface-bound water, so early T_2 signals are attributed to bitumen, and later T_2
22 signals correspond to the capillary bound and free water saturation in the rock.
23 For reference, Canadian bitumens produce a T_2 signal at approximately 0.1 – 4 ms
24 range ^{60,61}, the same as clay bound water in sandstones, while capillary bound
25 water produces a signal roughly from 3 – 33 ms in sandstones ²³. Consequently,
26 deconvolution approach works best for oil sands with low clay and capillary
27 bound water content, where the overlapping of bitumen and water signals is not
28 extensive.

29

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



11

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

7 However, in fines and clays, where pores are tiny, the water protons relax faster
8 due to the surface relaxation at the water-rock interface, thus generating the signal
9 in the fast-relaxing part of distribution where it can overlap with the signal
10 originating from heavy oil and bitumen (Figure 7B). In addition to that, the
11 diffusion coupling effect may further decrease the interpretability of the oil and
12 water signals. This effect occurs in saturated and connected micro- and
13 macropores when water is in diffusional exchange, causing the change in the
14 relationship between T_2 relaxation and pore size distribution⁶². In strong
15 diffusive-coupling conditions, macro- and micropore water signals will merge into
16 a single peak, rendering the single T_2 cutoff and deconvolution approach
17 inaccurate⁶³. Another limitation of this approach is that it requires the separate
18 determination of water and oil NMR amplitudes and the independent
19 measurement of their volume or mass.

20

21 Alternative methods involve 2D LF-NMR measurements, where instead of using
22 one NMR parameter (i.e., T_2 relaxation), additional parameters are employed (i.e.,
23 T_1 relaxation or diffusion) to obtain so-called 2D NMR maps⁶⁴⁻⁶⁶, which can
24 theoretically help to separate these overlapping bitumen and water signals.
25 Application of 2D maps showed considerable success in fluid saturation
26 evaluation, compared to 1D T_2 relaxation distribution analysis, since T_1 relaxation
27 or diffusion of reservoir fluids can be sufficiently different, thus enabling relatively
28 simple separation of their signals. Unfortunately, 2D NMR is slower and more
29 expensive to run, and there can still be instances where these signals are not
30 distinct, in which case estimation of fluid types and fluid volumes can be
31 challenging and require advanced analysis involving blind-source signal

1 separation (BSS), clustering algorithms, and a certain degree of knowledge in 2D
2 NMR maps interpretation⁶⁷.

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

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

3

4 **2.1. MOTIVATION**

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

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

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

24

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

1 relaxation) and relative hydrogen index (RHI_v) were measured for all samples and
2 were used as model inputs, while model tuning was achieved by NLS regression.
3 To quantify the effect of tuning on the reduction of prediction errors, we evaluated
4 the performance of models both in their default form (reported fitting
5 parameters) and after tuning by NLS regression (fitting parameters used in this
6 work).
7 Model performance was evaluated using root mean square error (RMSE),
8 maximum absolute error (MaAE), and adjusted coefficient of determination (adj.
9 R^2).

10 **2.2. THEORY AND EXPERIMENTS**

11 **2.2.1. SPIN-SPIN RELAXATION (T_2)**

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

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

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

22

23

24

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 interaction¹⁹:

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

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

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

where τ_c is molecule correlation time, and ω_0 is a Larmor frequency. The parameter C is defined for the $\frac{1}{2}$ 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

$$\tau_c = \frac{1}{6D_r} \quad (22)$$

$$D_r = \frac{kT}{8\pi\eta a^3} \quad (23)$$

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

1 any universal NMR viscosity model, discrepancies in predictive ability will grow
2 with the complexity of the chemical composition ^{24,32,70}.

3

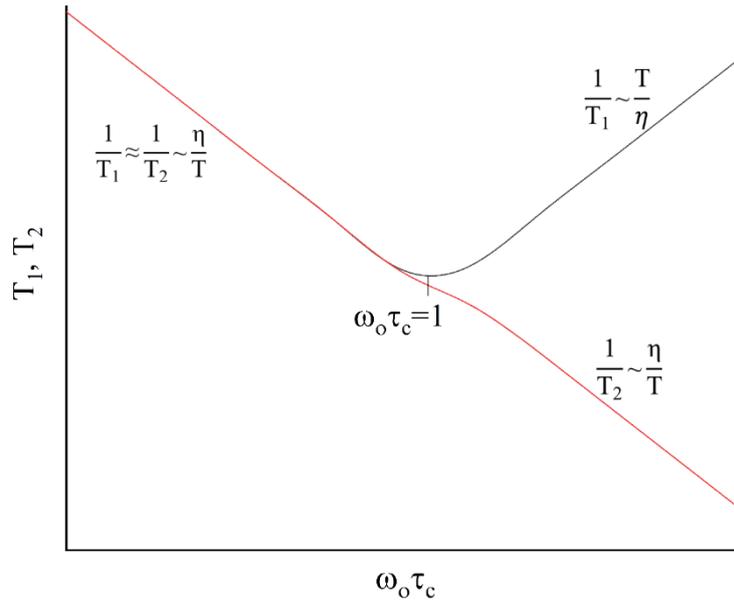
4 **2.2.3. T₂-RELAXATION MECHANISMS IN HEAVY OILS**

5 Recall that in the porous medium the total T₂ relaxation represents the sum of
6 three relaxation components (Equation 5). The total spin-spin relaxation is
7 governed by a sum of bulk relaxation T_{2bulk}, relaxation influenced by the pore
8 surface T_{2surface}, and relaxation caused by the gradients of magnetic field T_{2diffusion}.

9 In the case of water-wet porous media, surface relaxation is dominant for water
10 and bulk relaxation for oil. **Since the bulk relaxation remains a primary mechanism**
11 **in such scenario, it can be considered that for heavy oil T₂ ≅ T_{2bulk}.** According to
12 the nuclear spin relaxation theory⁷¹, there are two extreme cases:

- 13 1. Fast motion or extreme narrowing case (**ω₀τ_c << 1**), characteristic for
14 small molecules, low viscosities, or high temperatures. In such cases,
15 T₁ ≈ T₂ (Figure 8).
- 16 2. Slow-motion case (**ω₀τ_c >> 1**) characteristic for relaxation of large
17 molecules in high viscous substances or at low temperatures.

18 In both cases, the 1/T₂ relaxation rate is proportional to correlation time τ_c and
19 η/T ratio, respectively.



1
2 Figure 8: T_1 (black) and T_2 (red) dependence on correlation time (τ_c), according to the
3 BPP relaxation model.

4 Equations 19-23 describe the molecular relaxation processes governed by a single
5 exponential function which explains why the viscosity models for short
6 correlation times (Figure 8) provide sufficiently accurate predictions for light
7 oils^{15,23,27,29}. Solid-like components in heavy oils induce various relaxation rates as
8 opposed to light oils, and the cumulative spin-echo decay can exhibit non-
9 exponential behavior. This behavior can be approximated by the stretched-
10 exponential function, also known as the Kohlrausch-Williams-Watts function³³:

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

11 Function $F(t)$ is a time-dependent function of molecule location and orientation.
12 $G(\tau)$ describes a relationship between function $F(t)$ in different time steps, and $(\tau/\tau_c)^\gamma$
13 is a stretch parameter. Equation 24 does not have an analytical Fourier
14 transform, but a modified Cole-Davidson function approach^{31,72} can be used
15 instead.

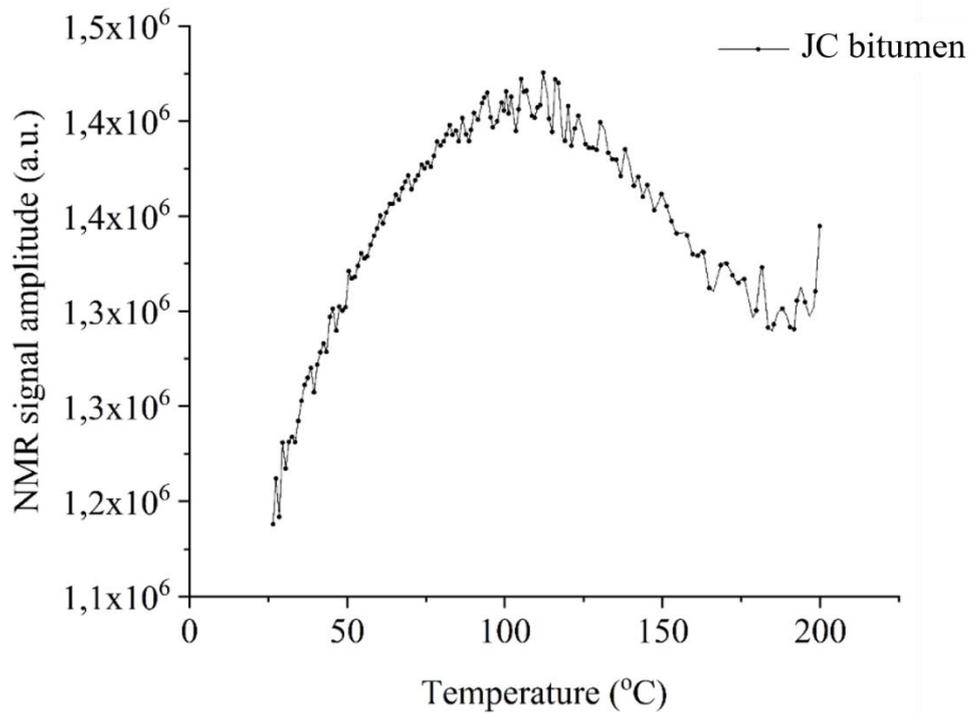
$$T_2 \sim \left(\frac{\eta}{T} \right)^{-\beta} \quad (25)$$

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

1 authors^{3,25,31,73}. It should be noted that T_1 dependence on correlation time also
2 does not follow the classical BPP model for heavy oils and bitumen in the slow-
3 motion case, which was experimentally proved by several authors^{25,60}, most
4 recently by Singer et al.³⁵. In Figure 8, instead of the anticipated increase, T_1
5 plateaus to 3 ms value on a frequency-normalized scale for various viscosity
6 samples. The authors explained the observed plateau effect by combining the
7 dipole-dipole interaction model for intramolecular interactions and the modified
8 Lipari-Szabo model for internal motions of the non-rigid structure.
9

10 **2.2.4. ECHO SPACING (TE) AND RELATIVE HYDROGEN INDEX (RHI_v)**

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



1

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

5 From Figure 9, it can be observed that at temperatures above 100 °C, the NMR
 6 amplitude decreases. This is known as a Curie effect^{60,76}, where magnetization loss
 7 (NMR signal loss) occurs due to the high temperature of the sample. One of the
 8 means to account for this loss is the implementation of the relative hydrogen index
 9 (RHI). The RHI represents the relative amount of measured (detectable) hydrogen
 10 protons by the NMR device in the oil sample. It is expressed as the ratio of oil and
 11 water NMR signal amplitudes per unit mass⁷⁷. In the case of using NMR tools at
 12 elevated temperatures, it is compulsory to implement temperature correction for
 13 RHI to account for the magnetization loss:

$$\text{RHI} = \left(\frac{A_o m_w}{A_w m_o} \right) \left(\frac{T(^{\circ}\text{K})}{T_{\text{amb}}(^{\circ}\text{K})} \right) \quad (26)$$

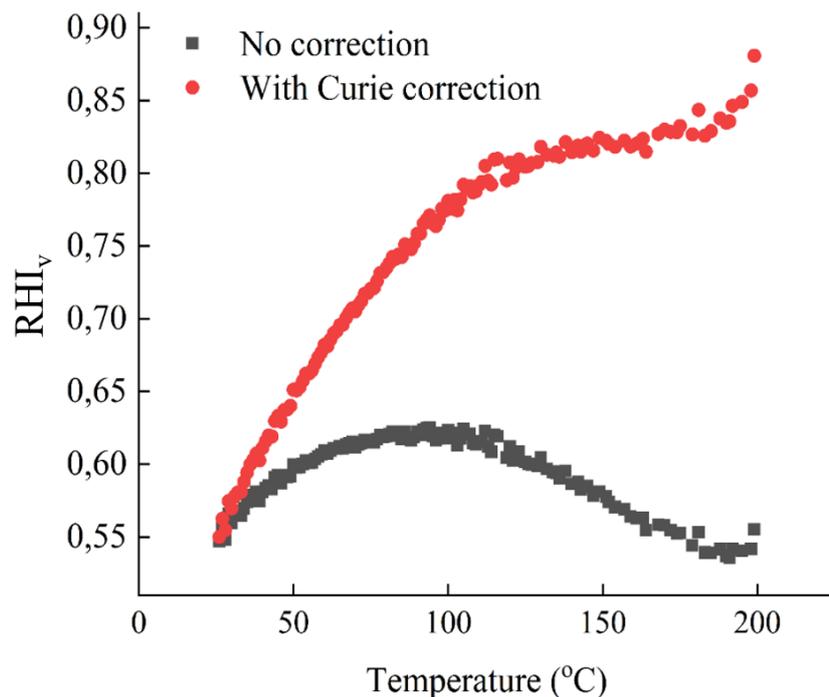
14 where A_o and A_w are the amplitudes of the oil and water signal respectively, m_o and
 15 m_w are masses of oil and water respectively, and T/T_{amb} is a temperature
 16 correction term in kelvins. Also, note that T is a sample temperature while T_{amb} is

1 the temperature at which water sample was measured (ambient). If the RHI is
 2 normalized to the sample volume, a relative hydrogen index for a defined sample
 3 volume can be obtained (RHI_v). This normalization is consistent with Curie's
 4 expression for magnetic susceptibility, where magnetization is expressed per unit
 5 volume⁷⁵. Also, RHI_v is more suitable for application to well logs because the NMR
 6 tool detects a defined volume of the formation⁶⁸. Burcaw et al. proposed a simple
 7 approach for conversion of RHI to HI⁶⁸:

$$RHI_v = \frac{\rho_o}{\rho_w} \cdot RHI \quad (27)$$

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

13



14

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

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

8

9 **2.2.5. ENHANCED NMR VISCOSITY MODEL**

10 The RHI_v works as a correction factor by compensating the magnetization loss
 11 (Curie effect) at high temperatures. The RHI_v also accounts to a degree for the long
 12 TE, meaning that the correction coefficient for the TE term is redundant. For
 13 longer correlation times characteristic for very viscous oils ($\omega_0\tau_c \gg 1$), the right
 14 addend in Equation 28 contains the stretching parameter d (or β in Equation 25),
 15 which accounts for non-exponential relaxation, i.e., the power-law relationship
 16 with T_{2lm} . Lastly, the enhanced model contains a T_{2lm} term, inversely proportional
 17 to viscosity, which properly works for light oils. This model is expected to mitigate
 18 the following known pitfalls in NMR viscosity prediction:

- 19 • Magnetization loss due to the high temperature (Curie effect).
- 20 • Long echo times (TE) which hinder the detection of solid-like components
 21 in heavy oils.
- 22 • Non-exponential relaxation (i.e., glass transition) in viscous oils where
 23 $\omega_0\tau_c \gg 1$.

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

$$\eta = \frac{a}{RHI_v^b T_{2lm}} + c T_{2lm}^{-d} \tag{28}$$

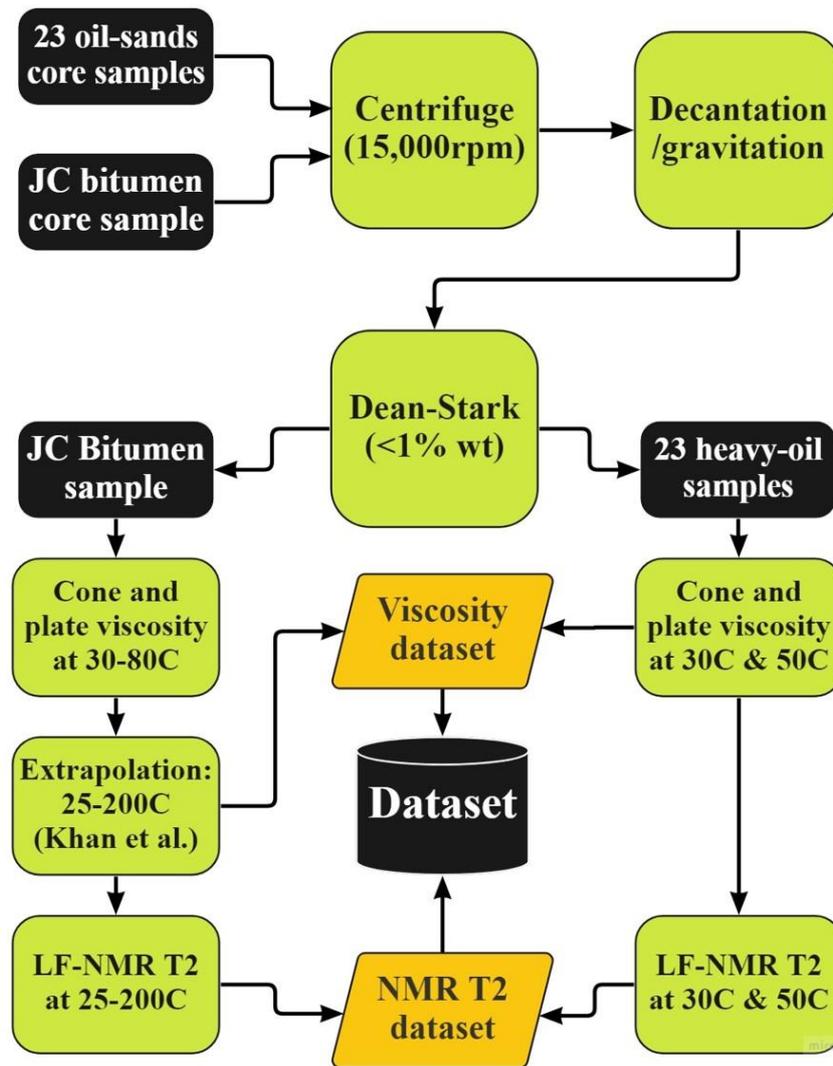
26 where a , b , c , and d are obtained from the nonlinear least squares (NLS) regression.
 27 The left-hand addend of Equation 28 is adapted from the Bryan et al. model¹⁸,

1 which correlates measured hydrogen content and T₂ logarithmic mean with oil
2 viscosity.

3 **2.2.6. PREPARATION OF OIL SAMPLES**

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

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



1

2

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

3

4

5

2.2.7. RHEOLOGICAL MEASUREMENTS – 23 HEAVY OIL SAMPLES

6

Rheological measurements were executed on 23 heavy oils to obtain a reference

7

dataset compared with predictions from NMR viscosity models. Previous studies

8

have shown that cone and plate rotational viscometers provide higher accuracy in

9

measuring viscous fluids than glass-capillary and oscillating-piston viscometers⁷⁹.

10

The Brookfield DV-II-Pro cone and plate viscometer was used, which meets ASTM

11

D4287 industry standard for oil viscosity measurement⁸⁰. As previously described

12

by Bryan et al.³², the cone diameter was 12 mm with an angle of 1.5°. Shear rate

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

8

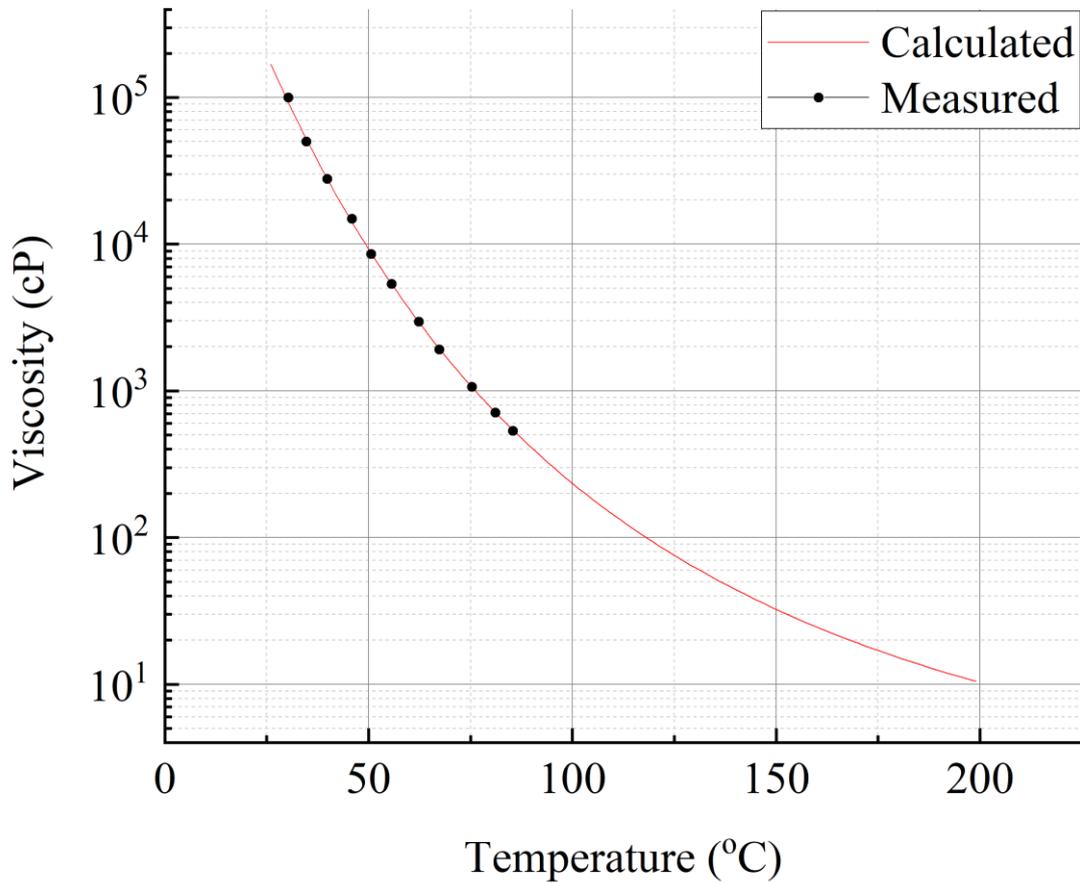
9 **2.2.8. RHEOLOGICAL MEASUREMENTS AT HIGH TEMPERATURE– JC BITUMEN**

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

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

17 where A and B are empirical constants calculated as a slope and intercept of
18 absolute temperature and measured dynamic viscosity, respectively. The
19 relationship between JC bitumen viscosity and temperature is depicted in Figure
20 12.

21



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

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 T₂ relaxation distribution in the time domain using NNLS inversion software ExpFit⁷⁷. 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 °C (~0.7 °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 $T_2 \sim 10\text{-}20$ ms, representing <1% of the total signal produced by the bitumen sample, which is negligible. Similar reports can be found in the literature⁸¹.

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

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

3

4 The NLS regression was performed using Levenberg – Marquardt (M-L) iteration
5 algorithm⁸².

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

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

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

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

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

15 Like in Tikhonov regularization, a damping factor λ is added to regulate the
16 reduction rate of S and for more efficient discovery of a gradient direction. For the
17 initialization of L-M, the λ was set to 0.001, and after each successive iteration was
18 automatically increased or decreased by a factor of 10 relative to the gradient
19 direction, depending on whether the squared residuals were reduced or
20 increased. The L-M algorithm converges when the sum of squared residuals
21 remains unchanged relative to the set tolerance or is equal to zero. The tolerance
22 parameter was reduced chi-square (χ^2), which can be calculated by dividing a sum
23 of squared residuals (RSS) by degrees of freedom. The tolerance was set to χ^2
24 $< 1 \cdot 10^{-9}$.

25

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

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

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

6
 7 In the existing literature, the NMR model accuracy is usually assessed visually on
 8 cross-plots relative to the prediction bands and by comparing coefficients of
 9 determination - R^2 3,30,33,73,74,77. In this study, an adjusted coefficient of
 10 determination and standard coefficient of determination is used for the evaluation
 11 of prediction variation captured by the model

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

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \quad (35)$$

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

22
 23
 24
 25

1 **2.3. RESULTS AND DISCUSSION**

2

3 **2.3.1. NMR VISCOSITY PREDICTION – 23 VARIOUS HEAVY OIL SAMPLES**

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

6

7 Table 1: Tested literature NMR viscosity correlations

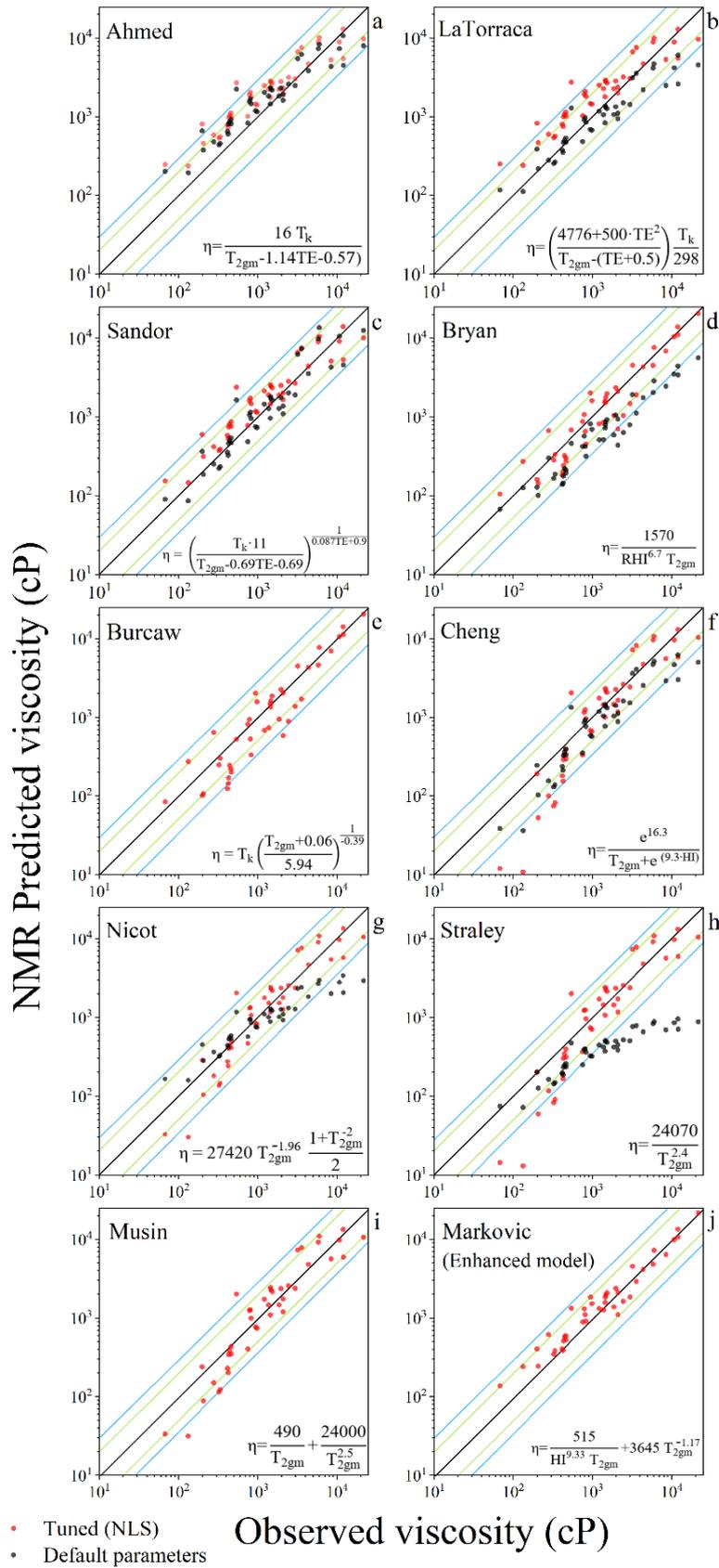
Model	Input data	Fitting parameters
Straley, 1997	T_{2lm}	2
LaTorraca, 1999	T_{2lm} , TE, T	2
Bryan, 2003	T_{2lm} , RHI	2
Nicot, 2007	T_{2lm} , proton radius (a), inter proton distance (b)	1
Burcaw, 2008	T_{2lm} , HI	3
Cheng, 2009	T_{2lm}	3
Ahmed, 2014	T_{2lm} , TE, T	2
Musin, 2016	T_{2lm}	3
Sandor,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

8 Figure 13 shows NMR viscosity predictions and observed viscosity for a suite of
9 23 heavy oil samples at 30 °C and 50 °C with viscosities ranging from 70 – 21,600
10 cP. The NMR viscosity predictions generated by nine well-known literature
11 models by Ahmed et al.³⁰, LaTorraca et al.⁷⁴, Sandor et al.³¹, Bryan et al.¹⁸, Burcaw
12 et al.⁶⁸, Cheng et al.³³, Nicot et al.²⁵, Straley et al.²³, and Musin et al.³. Predictions
13 made by the enhanced model are shown in Figure 13j. To emphasize the effect of
14 NLS regression, predictions were produced with NLS tuned parameters (red) and
15 with default model form, where used parameter values were reported by the
16 authors (black). The analytical form of models with fitting parameters obtained by
17 nonlinear least squares (NLS) regression is depicted in the lower right corners.

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

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

18
19 Heavy oil models by LaTorraca et al., Ahmed et al., and Sandor et al. show a similar
20 performance since they are based on TE correction and temperature. Sandor et al.,
21 and Ahmed et al., proposed multiple models in their work, but for this work were
22 selected the ones with the highest reported score. After NLS regression, the model
23 by Sandor et al. (Figure 13c) shows that most predictions fall within a factor of one
24 and two at the viscosity range between 30 and 3,000 cP. However, at viscosities
25 $>3,000$ cP, predictions scatter into a factor of two and three, causing the inflation
26 of RMSE and MaAE scores. (Figure 14). The models by LaTorraca et al. and Ahmed
27 et al. have less variance in $>3,000$ cP domain after NLS regression but tend to
28 overpredict viscosity in the $<6,000$ cP domain, with most predictions falling within
29 a factor of two and three. This is due to the cost function minimization, where the
30 iteration algorithm minimizes larger squared errors in the higher viscosity
31 domain at the expense of accuracy in the $<6,000$ cP domain.

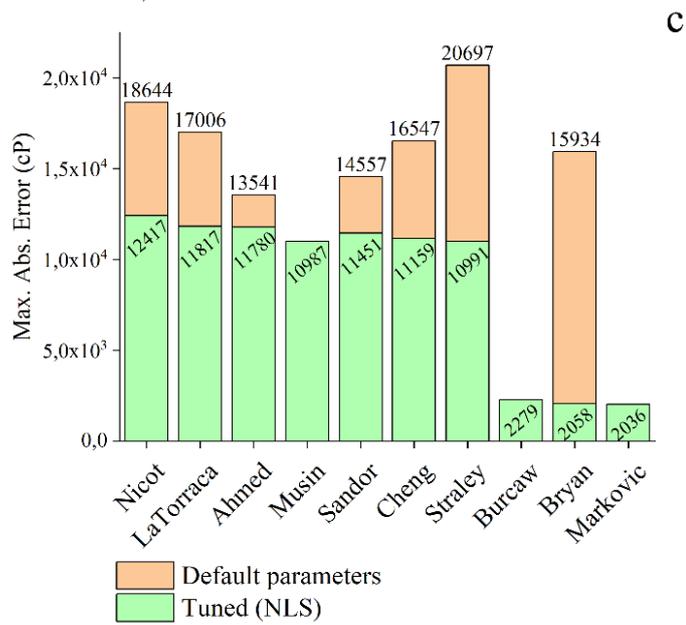
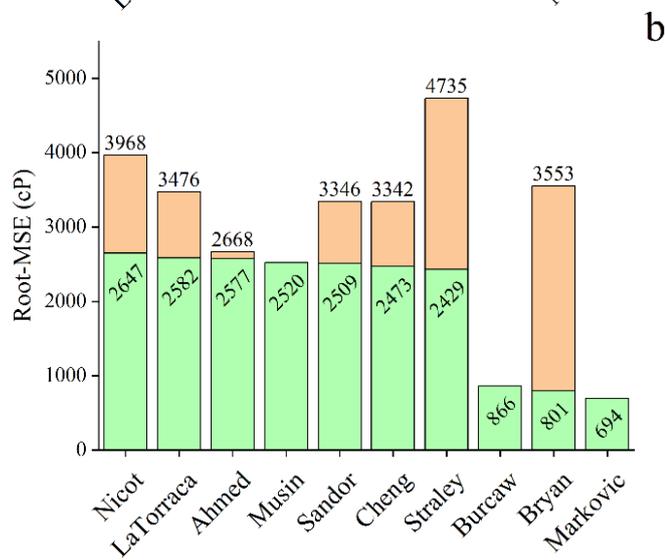
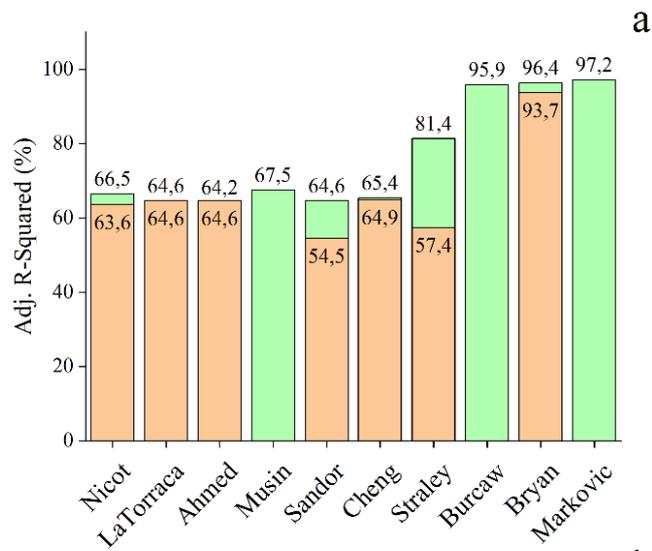


1 Figure 13: Rheological viscosities compared to NMR viscosities of 23 heavy oils. The
2 Markovic et al. (j) model demonstrates the highest accuracy. A solid black line ($x=y$)
3 presents a perfect prediction.

4 The models by Nicot et al., and Cheng et al. are heavy oil models, while Straley et
5 al. is a light oil model. All three are based on the power-law T_{2lm} , which accounts
6 for long TE (echo spacing) and have similar performance. Figure 13 and Figure 14
7 show that the models are affected by NLS to a different extent but generally
8 achieve considerably improved scores after NLS. Most of the predictions in the
9 >600 cP domain fall within a factor of one and two. Recall that Nicot et al., Cheng
10 et al., and Straley et al. models produce large prediction residuals in the domain
11 <600 cP after NLS regression. Again, this is caused by minimizing squared errors
12 in the higher viscosity domain at the expense of accuracy in the lower viscosity
13 domain. This issue can be solved either by constraining fitting parameters to a
14 specific range, at the expense of the accuracy at higher viscosities, or by using
15 these models for oilfields or wells where the viscosity **varies** no more than three
16 orders of magnitude.

17
18 Figure 14 shows the segregation of the three models (Markovic et al., Bryan et al.,
19 and Burcaw et al.), which perform substantially better than the remaining seven
20 based on all three statistics. Predictions by Markovic et al. fall within a factor of
21 one and two along with the whole range, with only four exceptions in the <600 cP
22 domain. The most accurate predictions are distributed in the $>1,000$ cP domain.
23 Also, Bryan et al. and Burcaw et al. perform almost on par with the enhanced
24 model. The new model has a marginally better score than the latter two. This
25 marginal improvement is due to the power-law parameter, which was not
26 considered in the models by Bryan et al., and Burcaw et al.

27
28
29



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

4 These results demonstrate that the integration of RHI_v , into the correlation,
5 substantially improves the prediction accuracy of heavy oil models in the 70–
6 21,600 cP viscosity range at 30 °C and 50 °C, especially in the >3,000 cP domain,
7 as demonstrated by correlations from Bryan et al., Burcaw et al., and the newly
8 proposed model from Markovic et al. (Figure 13d, Figure 13e, Figure 13j). This
9 finding complements the adj. R^2 scores in Figure 14a show that these three models
10 have the highest explained variability (>96 %). In addition to the RHI_v , the
11 enhanced model contains the power-law term in T_{2lm} , which marginally improves
12 prediction capacity by rectifying the non-exponential relaxation effect of heavy
13 components, for which T_{2lm} and measured RHI_v cannot account.

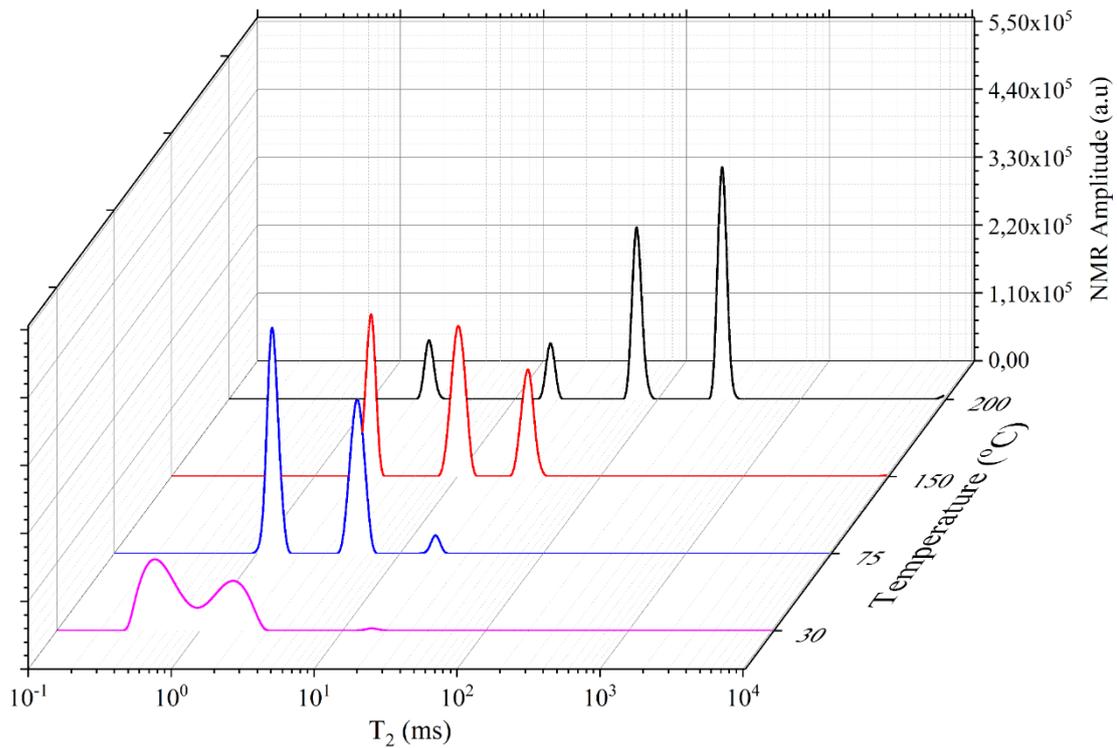
14 In conclusion, the new model (Markovic et al.) achieved the most favorable
15 statistical scores, while the models by Sandor et al., Bryan et al., Burcaw et al., and
16 Musin et al. have satisfactory performance only after the NLS regression. The
17 prediction capacity of the models in high temperatures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

19

20 **2.3.2. NMR VISCOSITY PREDICTION AT HIGH TEMPERATURES – JC BITUMEN**

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

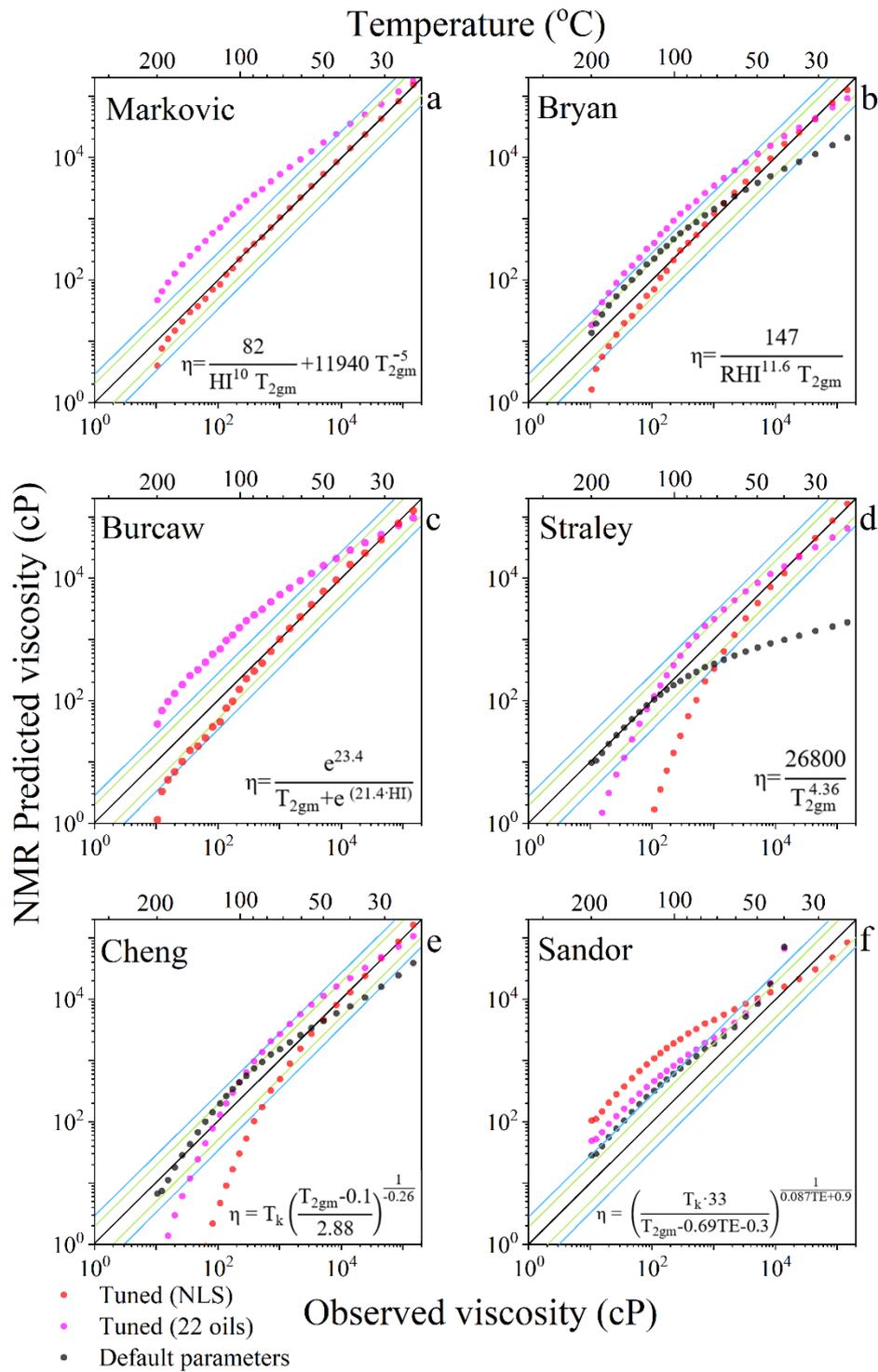
29



1

2 Figure 15: T₂ distribution curves of a single bitumen sample (JC bitumen) in the function
 3 of temperature.

4 Figure 15 presents the T₂ spectrum of JC bitumen, in the time domain, obtained
 5 using Tikhonov regularization¹⁸. The NMR relaxation spectra shift to the right-
 6 hand side (slow-relaxing part) with increasing temperature, and the NMR signal
 7 amplitude varies with temperature. The distribution curve at 200 °C shows four
 8 distinct peaks, which might indicate oil separation to several relaxometry
 9 components, possibly heavy and light fractions of the bitumen⁸³.



1

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

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

8

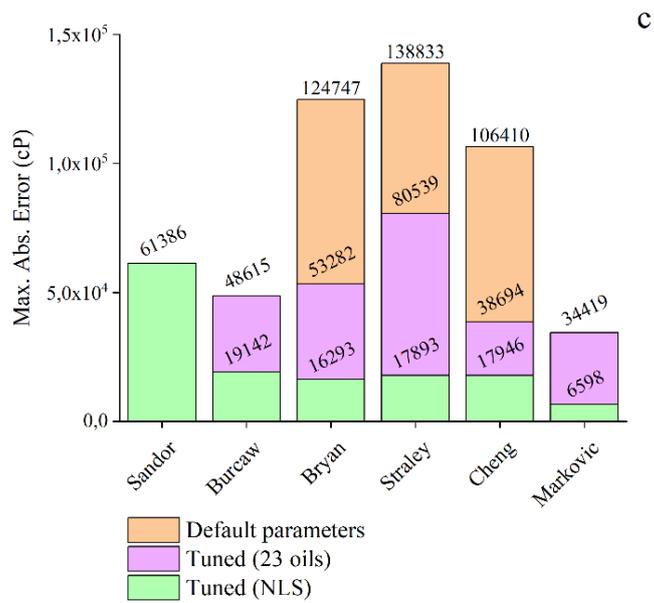
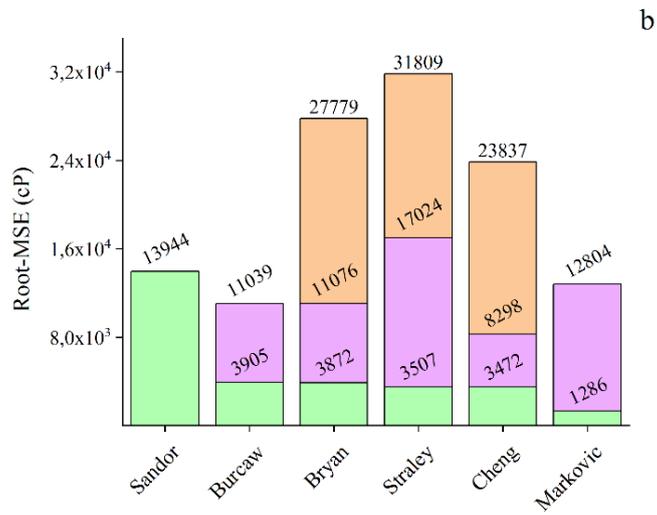
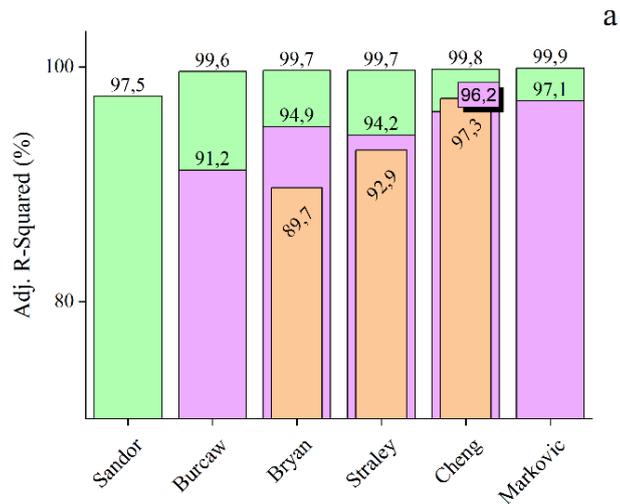
9 Figure 17 shows three statistical scores used to compare the model forecasting
10 performance. Note that the y-axis in Figure 17a is truncated for convenience. The
11 first observation comes from Figure 17a, where high adj. R^2 scores indicate a low
12 variability of response data (i.e., viscosity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

17

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

27

28



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

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

14

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

27

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

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

12

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

18

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

22 • Evaluation of the model performance on the set of different heavy oil and
23 bitumen samples from other heavy oil reservoirs would serve as an
24 additional validation, further reducing the forecasts' uncertainty and
25 enabling us to verify whether the proposed model is overfitted.

26 • In this work, we used non-linear least squares regression which strongly
27 penalizes squared residuals in the high viscosity range, giving less weight
28 to lower viscosity samples. Depending on a task, other cost functions which
29 give less weight to large residuals (mean absolute error) could be
30 minimized. Another option for model optimization could be orthogonal

1 distance regression (ODR) which proved to work well in instances with
2 asymmetric distribution of observations by providing a less biased fit⁸⁴.
3 Note that ODR was also tested in Chapter 3.
4

5 **2.4. SUMMARY**

6 This study demonstrates that LF-NMR relaxometry can be applied for viscosity
7 prediction in a broad viscosity range and at a broad range of temperatures (26-
8 200 °C). The results show that published NMR viscosity models cannot accurately
9 predict heavy oil viscosity at this range of temperatures. The enhanced NMR
10 model was associated with an NLS regression (parameter tuning) and used to
11 predict the viscosity of two distinct datasets: 23 heavy oils at 30°C and 50°C from
12 several wells and reservoirs in Alberta and a single bitumen sample (JC bitumen
13 dataset) at 26–200 °C. The prediction quality was evidenced by the root mean
14 square error (RMSE), maximum absolute error (MaAE), and adjusted coefficient
15 of determination (adj. R²). The new model scored an RMSE of 1,286 cP for the JC
16 bitumen sample compared to the RMSE of 23,837 cP generated by the first runner-
17 up model in default calibration from the literature. For the suite of 23 heavy oils,
18 the enhanced model scored an RMSE of 2,036 cP compared to the RMSE of 15,934
19 cP generated by the first runner-up literature model. The results also indicate that
20 the new heavy oil NMR viscosity model can be configured for monitoring purposes
21 in high-temperature conditions for order-of-magnitude viscosity monitoring.
22
23
24
25
26
27

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 water³². 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

1 (SVR)⁹⁰, several other machine learning algorithms were tested as well, including
2 decision trees (DT)⁹¹, random forests (RF)⁹², and k-nearest neighbors (KNN)⁹³.
3 Multiple linear regression (MLR)⁹⁴ was also used. A feature engineering (FE)
4 approach was integrated to maximize the forecasting capacity of the models by
5 deriving new features using the empirical findings from the NMR oil
6 characterization domain ⁹⁵. The study results indicate that this strategy can be
7 successfully applied even to small datasets. As most of the underlying
8 mathematical principles of tested algorithms are substantially different, we could
9 observe the study task from different perspectives. The database used for
10 calibration of models in the study was formed from the previously published LF-
11 NMR crude oil data, containing over 130 light and heavy oil samples recovered
12 from various reservoirs in Canada and the USA^{18,60,86}. The database consisted of
13 282 observations in total. The study was segmented into two stages. In the first
14 stage, the feature engineering was performed, and the dataset was randomly
15 shuffled and split into training and testing sets in the 0.75:0.25 proportion. A split
16 seed was selected randomly (random_state=42) and fixed for the reproducibility
17 of the results. Therefore, the training set consisted of 211 observations and a test
18 set of 71. The generalization ability of the models was assessed by the K-fold cross-
19 validation, while model performance was recorded using root mean squared error
20 (RMSE), mean absolute error (MAE), mean square log error (MSLE) mean absolute
21 percentage error (MAPE) and coefficient of determination (R²). In the second
22 stage, the performance of models was compared against another four well-known
23 empirical NMR viscosity models that were trained using the same framework. The
24 code and the data have been uploaded to the GitHub repository and are available
25 for use.
26

1 3.2. METHODOLOGY

2 3.2.1. GRADIENT BOOSTED REGRESSION TREES

3 In supervised learning, gradient boosting represents an ensemble (additive)
4 model that can be used for solving supervised regression and classification
5 problems. The main idea is to derive a model from a set of weak learners, typically
6 decision trees (DTs) or their simplified versions known as decision tree stumps.
7 The construction of the viscosity model $\hat{\eta} = F(x)$, evolves in sequences or
8 boosting iterations (m). For each iteration, a new decision tree (h) is added to the
9 existing model to minimize the loss function further. This way, an updated and
10 improved version of the model is obtained $F_{m+1}(x)$. This process is repeated until
11 the specified number of boosting iterations is reached ^{91,96,97}.

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

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

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

$$h_m(x_i) = \eta_i - F_{m-1}(x_i) \quad (37)$$

17 From Equation 37, it can be observed that each added h is fitted to prediction
18 residuals. In gradient boosting regression, the residuals are integrated into the
19 concept of negative gradients, enabling the use of other loss functions such as
20 absolute loss and Huber loss⁹¹. When dealing with datasets with a large number
21 of outliers, the commonly used squared error loss function $L = \Sigma(y_i - F(x_i))^2$ will
22 emphasize the larger residuals. Absolute loss function is not squaring the errors
23 $L = \Sigma|y_i - F(x_i)|$, making it therefore more resistant to outliers. The negative
24 gradient with an absolute loss function can be denoted as:

$$-\frac{\partial L(\eta_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} = \text{sign}(\eta_i - F_{m-1}(x_i)) \quad (38)$$

25 since the loss function is minimized by adding a new DT and fitting it to F_{m-1} . The
26 number of DTs can become excessively large, which can result in overfitting the

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

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

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

9 **3.2.2. SUPPORT VECTOR MACHINES FOR REGRESSION (SVR)**

10 The SVR is a sophisticated and straightforward supervised learning (SL) algorithm
11 used in classification and regression tasks. The SVR is based on the structural risk
12 minimization (SRM) principle, which was confirmed to have better performance
13 compared to empirical risk minimization (ERM) used, for instance, in neural
14 networks. In simpler terms, SRM prevents the overfitting of the model by
15 balancing two inversely related hyperparameters and consequently making a gap
16 between the training set errors and test set errors smaller while reducing model
17 complexity. In contrast, in ERM, a single objective minimizes the training error.
18 What made support vector machines so famous was the introduction of kernels –
19 the arbitrary functions whose purpose is to map the dot product of input features
20 into the higher-dimension feature space. This functionality enables the utilization
21 of hyperplanes, which are particularly useful in non-linear classification problems.
22 Fortunately, the same concept was generalized for regression tasks ¹⁰⁰. In
23 addition, SVR has been proven to be an effective method even in application to
24 small datasets, which is a necessary implication for the task at hand.

25 In terms of viscosity prediction by NMR parameters, SVR has to be associated with
26 our input features (T_{2lm} , TE, and T(°K)) and output vector η (Tables 2 and 3).
27 Suppose we arrange all the preprocessed input features in a matrix form as $x =$
28 $[x_1, x_2, x_3 \dots x_n]$, where x_n are column vectors of inputs. The measured viscosity
29 instances can be rewritten into a response vector $\eta = [\eta_1, \eta_2, \eta_3, \dots, \eta_n]$. Thus the

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

$$\eta: f(x) = W \cdot \phi(x) + b \quad (40)$$

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

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

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

7 where, the $\|W\|$ term is a magnitude of a vector of feature weights, which reduces
 8 the function's sensitivity to the perturbations in input x (i.e., flatness), thus
 9 gauging the robustness of a model. The right-hand side term quantifies the
 10 prediction error, measured by the L_ε loss function (Equation 42). The magnitude
 11 of residuals $|\eta_i - f(x_i)|$ is compared with the predefined value of ε , so that the
 12 residuals smaller than ε are ignored, but residuals larger than ε are
 13 penalized. Since any ε can be defined, the C parameter is introduced to regulate
 14 the tradeoff between the flatness of the $f(x_i)$ and penalty size for residuals larger
 15 than ε ¹⁰⁰. The optimization of Equations 41 and 42 can be simplified by
 16 introducing slack variables (ξ_i, ξ_i^*) instead of prediction residuals ⁹⁰:

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

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

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

$$\eta: f(x) = \sum_{i=1}^n (\alpha - \alpha_i^*) \cdot K(x_i, x_j) + b \quad (45)$$

1 where α and α_i^* are Lagrange multipliers and $K(x_i, x_j)$ is the kernel function,
2 which maps the input features into the higher-dimension space. Further details
3 about support vector machine regression can be found elsewhere ^{90,100}.

4

5 **3.2.3. DATABASE OF RHEOLOGICAL AND NMR MEASUREMENTS**

6 The oil data were collected from our previous research and other published works
7 ^{18,60,86}. In all studies, the experimental procedure was similar: the dynamic
8 viscosity of oils was determined using conventional laboratory instruments (i.e.,
9 cone and plate rheometers), whereas the T₂-relaxation spectra of the samples
10 were obtained after raw materials data mathematical inversion from the
11 measurements made by LF-NMR relaxometers. For this study, 282 data points
12 were used for model development.

13

14 **3.2.4. PREPROCESSING AND ANALYSIS OF THE DATASET**

15 The preprocessing and analysis of all rheological and NMR data were performed
16 using Python environment version 3.7.2 with the scikit-learn package and
17 OriginPro 2019b ⁹⁸. The feature dataset consists of T_{2lm}, TE, and T(°K), while
18 viscosity observations were stored as output vectors. The data was divided into
19 the training set and a test set in the 3:1 proportion, respectively. This way, we
20 obtained a training set of 211 data points and a test set (unseen data) of 71 data
21 points, which was used to estimate model accuracy only.

22

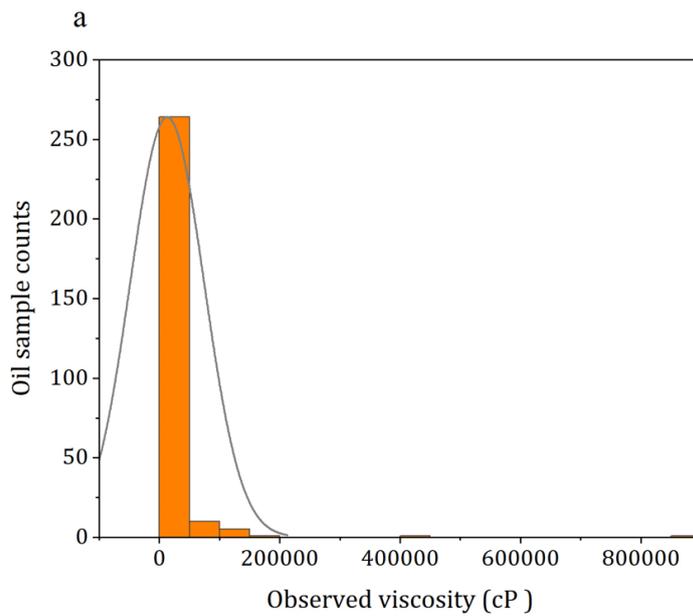
23 **3.2.5. FEATURE ENGINEERING AND TRANSFORMATION**

24 Feature engineering (FE) is a process in which domain knowledge is applied to
25 perform appropriate transformations of the inputs and to extract new information
26 from their known empirical relationships. This strategy proved to be effective in

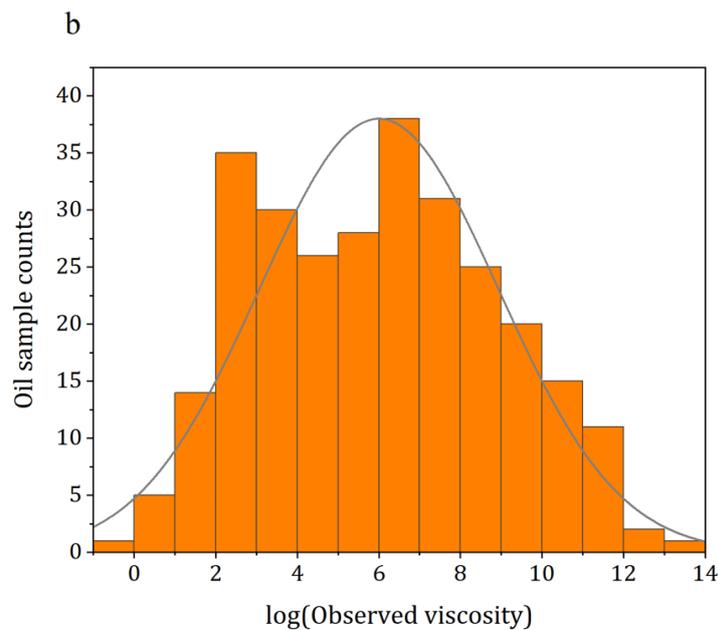
1 reducing the complexity of SL models, which in turn led to an increase in
2 prediction performance⁹⁵. In our case, this entailed: (1) the transformation of
3 inputs T_{2lm} , TE, $T(^{\circ}K)$, and target output η , and (2) deriving new inputs from
4 empirical relationships of T_{2lm} , TE, and $T(^{\circ}K)$ with target output η .

5
6 Table 2 shows that the ranges of inputs and outputs are out of scale, which implies
7 that a particular transformation should be applied to normalize the data. Also, the
8 observed viscosity data has a long-tailed distribution as it is skewed to the right-
9 hand side of Figure 18a, with over 95 % of samples distributed between 0.8 –
10 100,000 cP range. In the field of statistics, the observations outside three standard
11 deviations (outliers) typically degrade the forecasting performance of the models
12 and can be, therefore, omitted¹⁰¹. In our case, however, the outliers correspond to
13 extra-heavy oils and bitumens (e.g., > 180,000 cP). In practice, the natural
14 reservoirs in which these oils reside are often thermally treated in order to
15 facilitate their recovery¹⁰². Therefore, if these samples were omitted from the
16 training data, the valuable information about their T_2 -relaxation behavior at high
17 temperatures would be lost. This information was preserved by applying a simple
18 logarithmic transformation to all features, which normalized the distribution of
19 the data. The effect of log-transformation is illustrated in Figure 18b, on the
20 example of target output η . Also, the log-transformation reduced nonlinearity of
21 the dataset, which, in theory, should improve the performance of the SL regression
22 models, which are efficient in solving linear problems (i.e., multiple linear
23 regression and support vector regression).

24
25
26
27
28
29



1



2

3 Figure 18: Distribution of oil viscosity η before (a), and after the log transformation (b).

4 **Note that η is a target variable.**

5

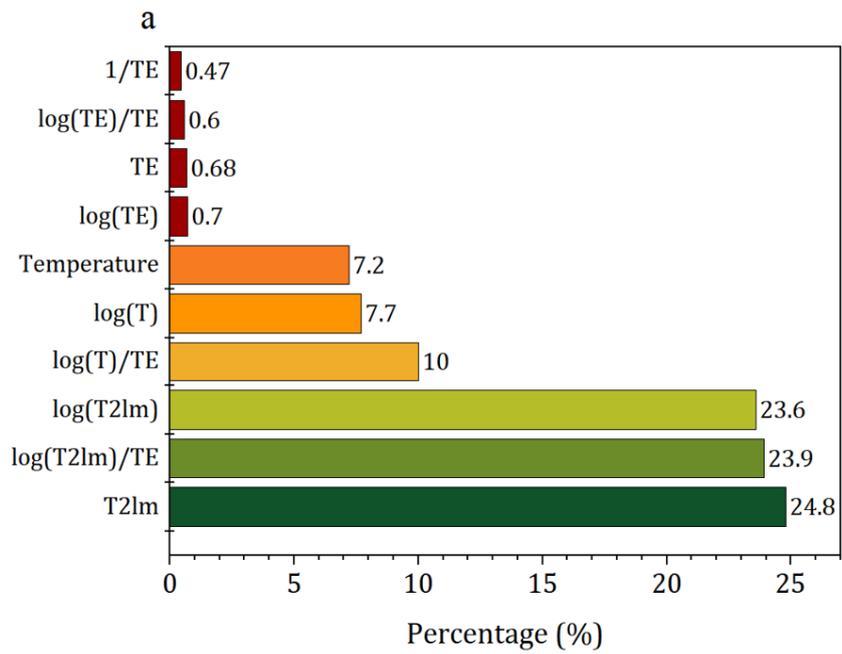
6 In the second stage, we derived the new features by employing the findings from
 7 previous studies^{31,33,103}. To evaluate the importance of newly derived features, we
 8 employed the GBRT algorithm. One of the benefits of ensemble models such as
 9 GBRT is their capability of feature ranking by their relative contribution to the

1 prediction accuracy, thus making the interpretation and selection of new features
2 more convenient. It should be noted that there are some downsides to feature
3 ranking. For instance, two or more features may have a comparable correlation
4 with the output. During feature ranking, one feature will be assigned a higher rank,
5 which will lead others to get a lower rank, thus potentially leaving out a strong
6 predictor ¹⁰⁴. Figure 19a shows the ranking of the seven new features, alongside
7 with T_{2lm} , TE, and $T(^{\circ}K)$, with the bottom ones being the most relevant. The
8 ranking is achieved by assessing the reduction of the training error generated
9 from splitting the nodes of the DTs. Therefore, the features which reduce the
10 training error more frequently during splitting will be ranked higher. Note in
11 Figure 19a that T_{2lm} -related features ($\log(T_{2lm})/TE$, $\log(T_{2lm})$, and T_{2lm}) capture
12 most of the variability ($\sim 72\%$), while $T(^{\circ}K)$ -derived features ($\log(T)/TE$, $\log(T)$,
13 and T) capture about 25% of the variability. This variability distribution was
14 expected, considering that the T_{2lm} strongly correlates with η , whereas T_2 -related
15 features become more important at high temperatures when severe NMR signal
16 loss occurs.

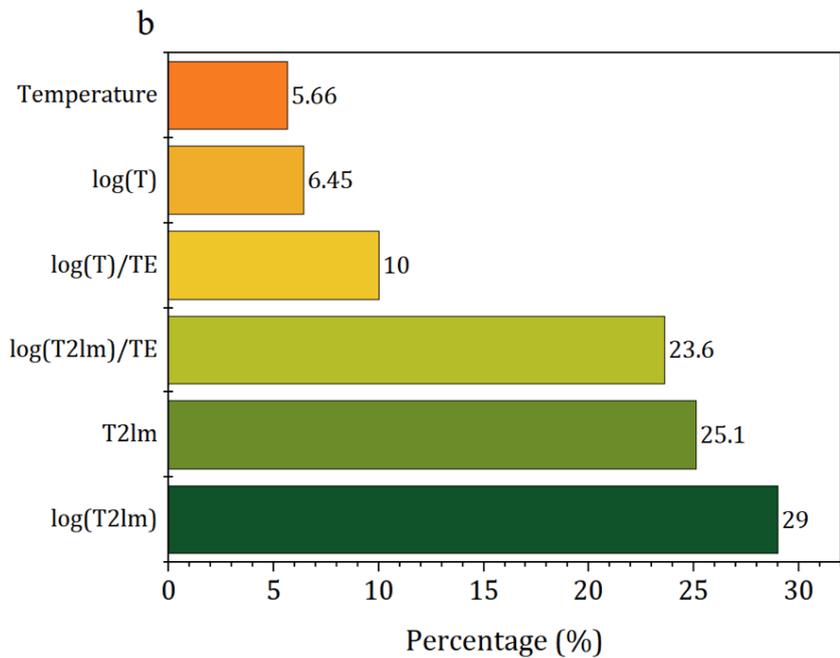
17

18 In contrast, the TE-related features ($1/TE$, $\log(TE)/TE$, TE , and $\log(TE)$) affect
19 prediction accuracy negligibly, with each being less than 1%. This is because the
20 NMR measurements used as inputs for this study were all acquired to optimize the
21 signal of fast relaxing fluids, i.e., through the use of small TE values. If this dataset
22 were to be expanded to systems with larger TE values (0.6 – 1.2 ms), TE's impact
23 would be higher. Within this dataset, the impact of TE was removed from further
24 consideration. However, even with the perceived insignificance of TE, it should be
25 noted that features that include the TE in the denominator demonstrate higher
26 relevance (e.g., $\log(T)/TE$ and $\log(T_{2lm})/TE$). Figure 19b illustrates the relative
27 importance of the remaining six features used for the training of the SL models.
28 Finally, Table 3 summarizes the statistical description of log-transformed
29 viscosity (i.e., target output) and engineered inputs used for SL viscosity
30 forecasting alongside original features (T_{2lm} , TE, and $T(^{\circ}K)$).

31



1



2

3 Figure 19: Training set relative feature importance (ranking) by GBRT model of all input

4 features, (a) before, and (b) after removal of redundant TE-derived features with less

5 than <1% relative contribution.

6

7

8

1 **Table 2: Descriptive statistics of six input features (input variables) used for training of**
 2 **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 ($^{\circ}$ K)	299.15 - 468.15	337.43	45.15	282

3 *target variable

4 **3.2.6. EVALUATION METRICS**

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

11
 12 The RMSE is regularly employed in scientific studies to evaluate model
 13 performance ^{44,105}. In this study, the RMSE is the square root of the average of
 14 squared differences between predicted viscosity and observed viscosity and is
 15 expressed in the centipoises:

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

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

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

1 In contrast to RMSE, MAE does not square the differences between predicted and
 2 observed viscosity, making MAE less sensitive to outliers¹⁰⁷. In this manner, the
 3 MAE score gives less weight to the large prediction residuals and, therefore, can
 4 be used as a control measure in RMSE interpretation. The shared disadvantage of
 5 RMSE and MAE is that both metrics do not provide any information about
 6 percentual differences between predictions and observations. MSLE accounts for
 7 this by associating squared differences between *log-scaled* predictions and
 8 observations:

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

9 In this manner, the MSLE avoids the heavy penalization of prediction errors in the
 10 high viscosity domain, as is the case with RMSE and MAE. Instead, it considers the
 11 relative percentual difference between observation and prediction rather than the
 12 size of their residual¹⁰⁸. In addition to MSLE, MAPE illustrated the relative
 13 percentual difference between sums of errors. MAPE represents the mean of the
 14 sums of absolute percentage errors of viscosity predictions. This metric enabled a
 15 more intuitive interpretation of the model forecasts since the errors are expressed
 16 in percentages¹⁰⁹:

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

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

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

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

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

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

3 where p is a number of features. Note that $\bar{R}^2 < R^2$.
 4

5 3.2.7. GBRT OPTIMIZATION

6 As mentioned in section 3.2.1, a choice of an arbitrary differentiable loss function
 7 ('loss' parameter), can be made according to the statistical properties of the
 8 dataset. The NMR viscosity dataset contains a substantial number of outliers,
 9 which implied using an outlier-resistant loss function, such as the least absolute
 10 deviation (LAD)¹⁰¹. To test this premise, 5-fold cross-validation was executed for
 11 four commonly used loss functions: Huber loss, least squares, least absolute
 12 deviation, and quantile loss. The rest of the parameters and hyperparameters
 13 were fixed to default values. Based on the lowest mean validation error, it was
 14 found that the GBRT configuration with LAD loss function generated the most
 15 stable predictions in terms of all error metrics (Table 4).
 16

17 Table 3: **The five-fold cross validation training set scores for different GBRT loss**
 18 **functions.** The LAD loss exhibits the best performance based on MAE_{cv} , $RMSE_{cv}$, $MSLE_{cv}$,

19 \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

20 The 'criterion' parameter can be determined in the same manner. This parameter
 21 allows a user to select the function that will estimate the DT node split quality.
 22 Usually, in regression tasks with DTs, the difference between the observed and

1 predicted value is quantified by mean squared error (MSE). Subsequently, the
 2 node splitting for a particular DT will be achieved so that the lowest MSE value is
 3 obtained. Since MSE heavily penalizes outliers, MAE was expected to perform the
 4 splitting task more efficiently (Table 5).

5

6 Table 4: The five-fold cross validation training set scores for different GBRT criterion
 7 hyperparameters. Node splits by MAE criterion achieve best scores in terms of MAE_{cv},
 8 RMSE_{cv}, MSLE_{cv}, \bar{R}_{cv}^2 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}_{cv}^2	0.86	0.66	0.64

9 The next step was to find the optimal hyperparameter values for the GBRT model.
 10 According to ^{91,99}, five hyperparameters have a considerable impact on GBRT
 11 model performance:

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

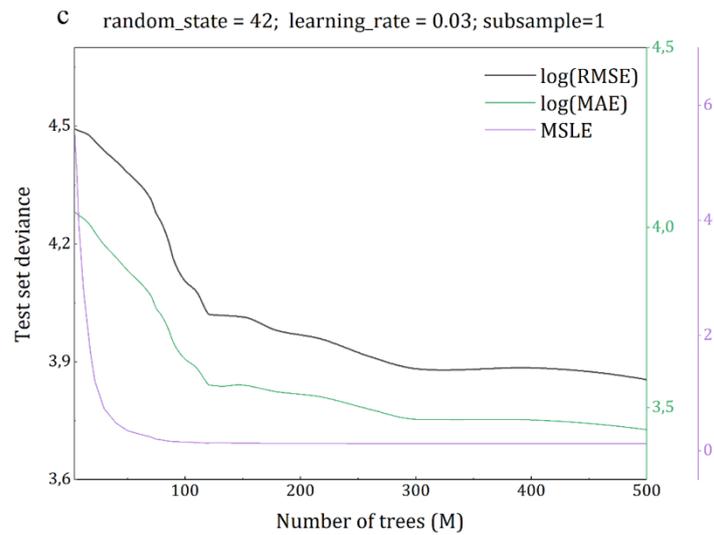
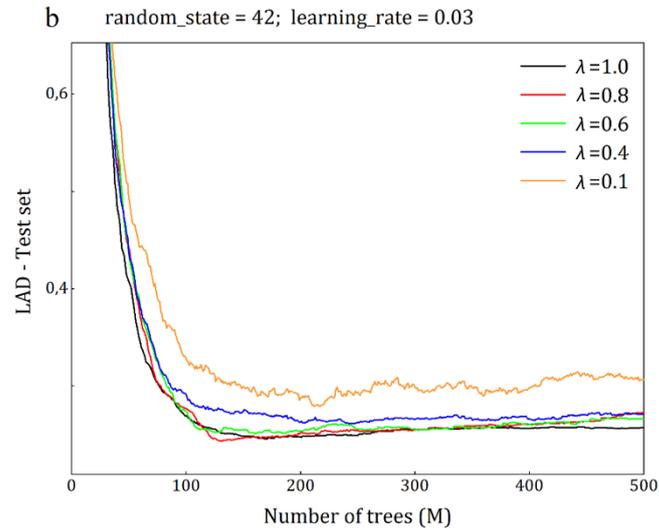
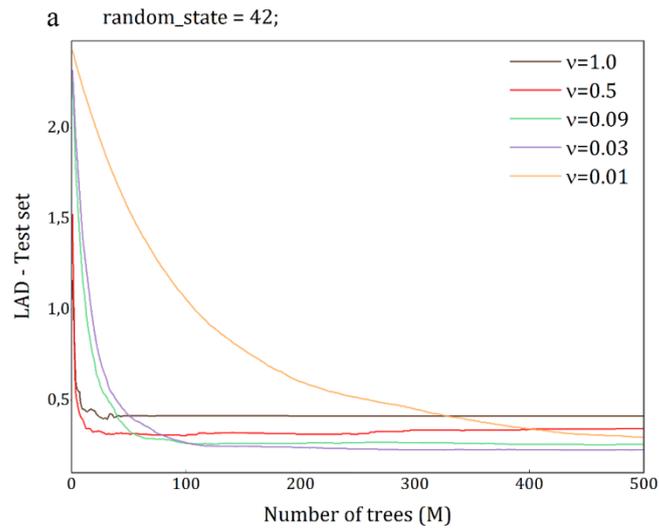
21 Mathematically speaking, these hyperparameters are mutually dependent (also
 22 observable from Equation 39), which is why it was required to use a more
 23 sophisticated optimization technique other than pure trial-and-error. Therefore,

1 these were evaluated simultaneously with the help of GridSearchCV (GS-CV), an
 2 exhaustive search cross-validation algorithm available in a scikit-learn package ⁹⁸.
 3 From the computer science standpoint, this metaheuristic approach iteratively
 4 optimizes an algorithm by searching for an appropriate combination of
 5 hyperparameters in multidimensional real-valued parameter space (grid),
 6 relative to some measure of accuracy (e.g., R^2). This approach captures the
 7 interaction between the hyperparameters, therefore significantly reducing the
 8 optimization time. However, due to discrete data (i.e., TE) and other distributions
 9 in the input data, the grid-search optimization may fail to discover the best
 10 hyperparameter configuration, even with the appropriate transformations
 11 applied to the dataset. Also, with a growing number of hyperparameters, its'
 12 utilization becomes computationally intensive. Hence, optimization was assessed
 13 further using error curves. Table 6 shows the hyperparameters and their value
 14 range, which were optimized using the GS-CV approach. Recall that 'loss' and
 15 'criterion' parameters were fixed according to results in Tables 4 and 5.

16

17 Table 5: Training set GBRT hyperparameter optimization by grid-search based on 5-fold
 18 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



1 **Figure 20:** The test set GBRT model performance in terms of least absolute deviations
2 (LAD) for various learning rates (a), and subsample sizes (b) relative to the number of
3 trees M . Bottom plot (c) illustrates the model accuracy evaluation as a function of M , in
4 terms of MAE, RMSE, and MSLE.

5 Figure 20a shows how the GBRT model deviance evolves with different learning
6 rates (ν) as a function of a number of trees M . The GBRT loss is expressed in the
7 least absolute deviations. Larger values of ν (e.g., $\nu=1$) lead to faster convergence,
8 that is, smaller values of M are needed for the deviance to converge. However,
9 when a value is decreased ($\nu=0.01$), the contribution of every additional estimator
10 is reduced further, leading M to increase to ensure smooth convergence, which
11 also means an increase in computational cost. Since the apparent tradeoff exists
12 between these two parameters, the parameter grid-search cross-validation was
13 used as a strategy for obtaining their appropriate values.

14

15 Additionally, subsampling λ is a parameter that enforces the variance reduction of
16 the sample population. In GBRT applications for large datasets, this technique
17 proved helpful for improving computing performance and accuracy⁹¹. Figure 20b,
18 however, shows that no subsampling ($\lambda =1$) leads to the smoothest and lowest
19 deviance for the given input, possibly due to the small number of data points. Also,
20 alternating the λ parameter has only a minor effect on deviance magnitude. In fact,
21 the variations are so minor that one must zoom in on the y-axis to observe this
22 behavior (note y-axis scales of Figure 20b).

23

24 Finally, the maximum depth of all trees was restricted to the same size ($J=4$), as
25 determined by the GS-CV, which agrees with recommendations in literature⁹¹.
26 The optimal number of features for the best split of the tree node was found to be
27 $\psi=2$ (i.e., $max_features = "log2"$). It should be noted that the value of the latter has
28 the least impact on the prediction performance of the GBRT model, and therefore,
29 using the default value (i.e., "auto") is also acceptable.

1 3.2.8. SVR OPTIMIZATION

2 In Equation 45, the term $K(x_i, x_j)$ represents the kernel function. The standard
3 kernel functions used in SVR are linear, polynomial, sigmoid, and Gaussian. The
4 NMR input parameters are in the nonlinear relationship with the oil viscosity;
5 therefore, the kernel must capture this relationship once the input features are
6 mapped into a higher dimension space. In these circumstances, the Gaussian, or
7 radial basis function (RBF) kernel has proven to be effective ¹¹¹. From Equation
8 45, the kernel function can be expressed as:

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

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

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

19 Literature findings show that the behavior of these hyperparameters is
20 interrelated, which should be considered during their optimization¹⁰⁰. Thus, the
21 GS-CV approach was used to simultaneously approximate C , ϵ , and RBF kernel
22 parameter γ (Table 7). Also, their in-depth assessment was performed from the
23 analysis of the error curves (Figure 21).

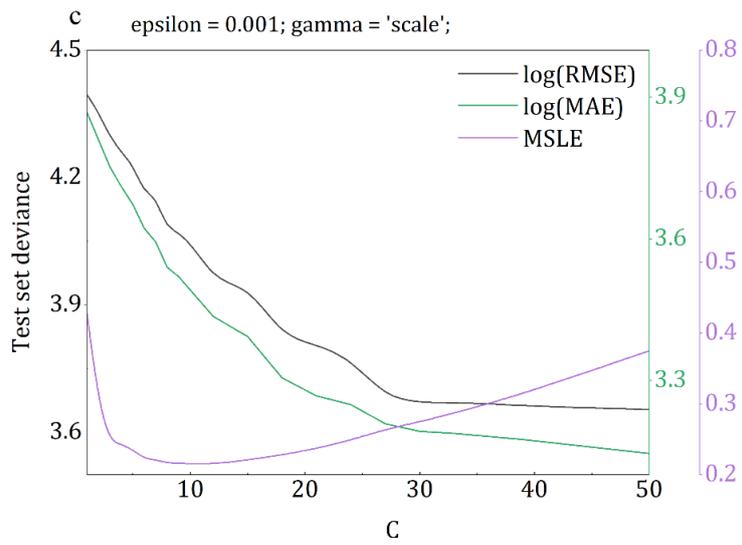
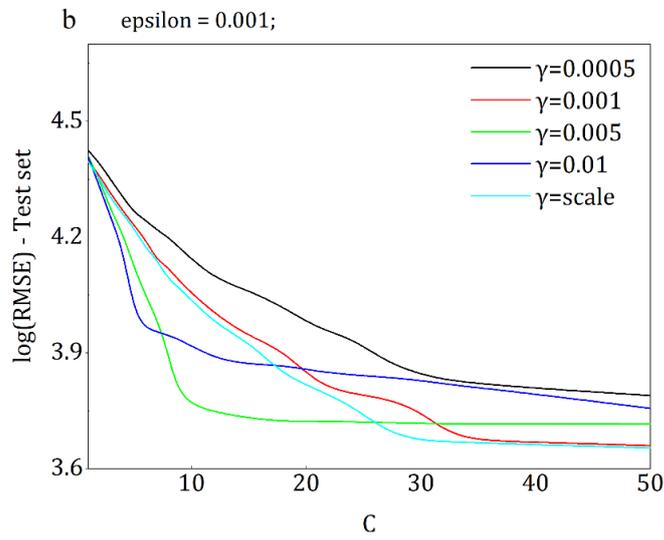
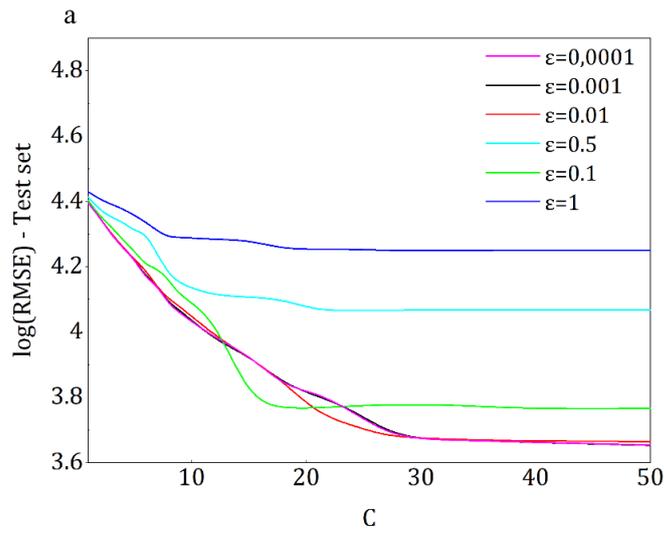
24

25

26

1 Table 6: Training set SVR hyperparameter optimization by grid-search based on 5-fold
 2 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



1 Figure 21: Test set SVR model performance in terms of log-normalized RMSE for various
2 values of ϵ (a), and γ (b) with respect to regularization C . Bottom plot (c) illustrates the
3 accuracy of optimized SVR model as a function of C , in terms of three error metrics;
4 $\log(\text{RMSE})$, $\log(\text{MAE})$ and MSLE.

5 Figure 21a presents how SVR model prediction accuracy behaves for various
6 values of radius- ϵ in the function of C . For the $\epsilon=10^{-4}$ obtained by GS-CV, it was
7 found that the SVR model utilized over 70 % of the data samples as support
8 vectors, which indicates overfitting¹¹². For values of $\epsilon=10^{-3}$ and $\epsilon=10^{-2}$, the
9 deviance converged smoothly at $C=30$. The number of SVs was reduced by
10 increasing ϵ to 10^{-3} (Figure 21a, black) while preserving nearly identical accuracy.
11

12 Figure 21b shows deviance for the fixed ϵ and various radii of individual SV impact
13 γ . The smoothest convergence and lowest deviation are achieved when $\gamma='scale'$,
14 which is the value when an inverse of the number of features is scaled by their
15 standard deviation. Interestingly, the GS-CV obtained $\gamma=5 \cdot 10^{-4}$, but according to
16 its' plot (black curve), the deviance converges when $C>50$, at which point the SVR
17 model attempts to perfectly predict each entry from the training set (hard-margin
18 SVM behavior). Since this might lead to overfitting and increased model
19 complexity, the γ was set to '0.001.'

20
21 As a final step, the model with fixed ϵ and γ hyperparameters was evaluated in
22 Figure 21c, where three metrics were utilized to evaluate the tuned SVR model.
23 While both RMSE and MAE follow the same decreasing trend, the MSLE error
24 decreases until the $C=12$ inflection point, after which it starts increasing. This
25 behavior is due to the inflation of residuals in the low viscosity domain. To restrict
26 the further growth of residuals and to preserve the overall model performance,
27 the regularization was set to $C=25$, in line with the grid-search results (see Table
28 7).

1 **3.3. RESULTS AND DISCUSSION**

2 This section is divided into two parts. In the first part, we compare SVR and GBRT
3 model performance against four other popular regression models, whereas in the
4 second part, a performance of four well-known empirical NMR viscosity models
5 was considered. The models are compared using the five error metrics introduced
6 in chapter 3.2.6. Also, the cross-plots with predicted and observed viscosities are
7 provided for the in-depth analysis.

8

9 **3.3.1. SUPERVISED LEARNING MODELS**

10 The performance of the GBRT and SVR was evaluated against an additional four
11 SL algorithms, including multiple linear regression (MLR) ⁹⁴, K-nearest neighbors
12 (K-NN) ⁹³, decision trees (DTs) ⁹¹, and random forests (RF) ⁹². Their optimization
13 was performed with GS-CV, similarly as for GBRT and SVR. Recall that the dataset
14 contains 282 observations which were shuffled and split into train and test set in
15 0.75:0.25 proportion, making a training and test set of 211 and 71 observations,
16 respectively.

17

18

19

20

21

22

23

24

25

26

27

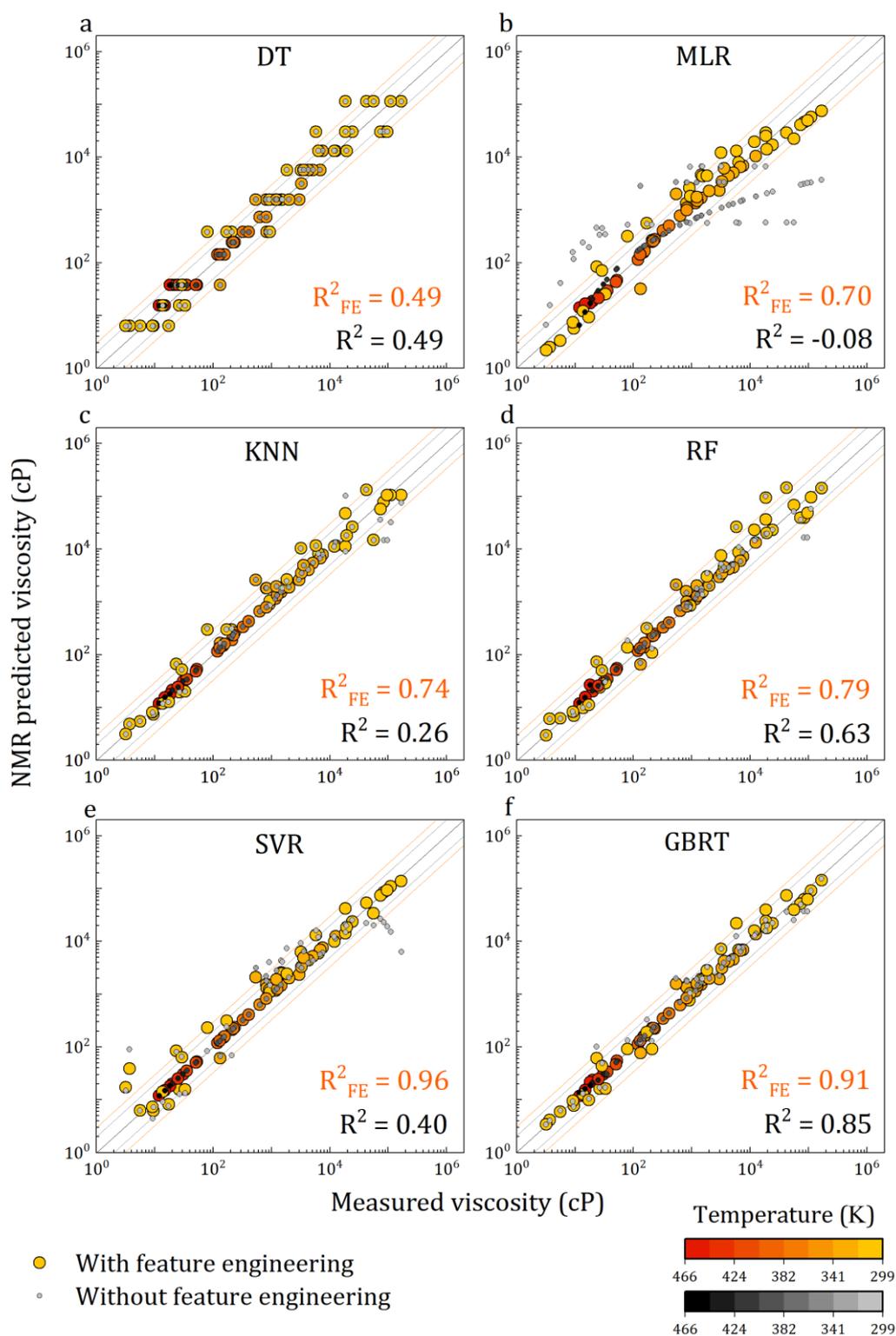
28

29

30

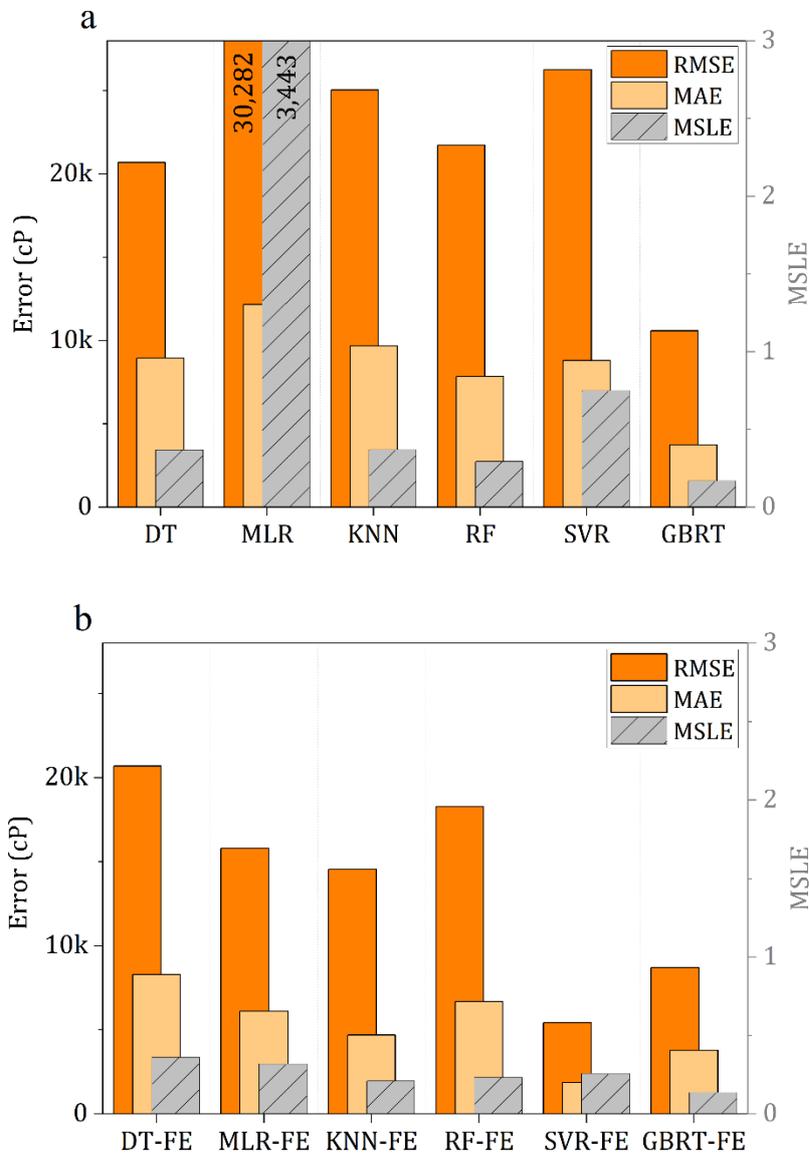
1 Table 7: Training set GS-CV hyperparameter optimization results for all supervised
 2 learning algorithms which were tested in this work.

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]



1
 2 Figure 22: Comparison of NMR SL viscosity model **test set predictions** and observations.
 3 Note that the grayscale points are predictions of models generated without FE, while
 4 warmer color points are predictions with FE. Lighter colors indicate lower temperatures

1 (from 25 °C/299 K), and more intense, darker colors indicate higher temperatures (up
 2 to 200 °C/466 K). GBRT and SVR models with integrated FE demonstrate the best
 3 performance.



4

5

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

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

1 model's prediction variance has reduced dramatically after employing FE (Figure
2 22b). Interestingly, for GBRT and RF models, the variance-reducing effect from FE
3 is much smaller than in the latter's case (Figures 22e and 22c). The same
4 observation applies to the DT model, the base estimator of GBRT, and RF models .
5

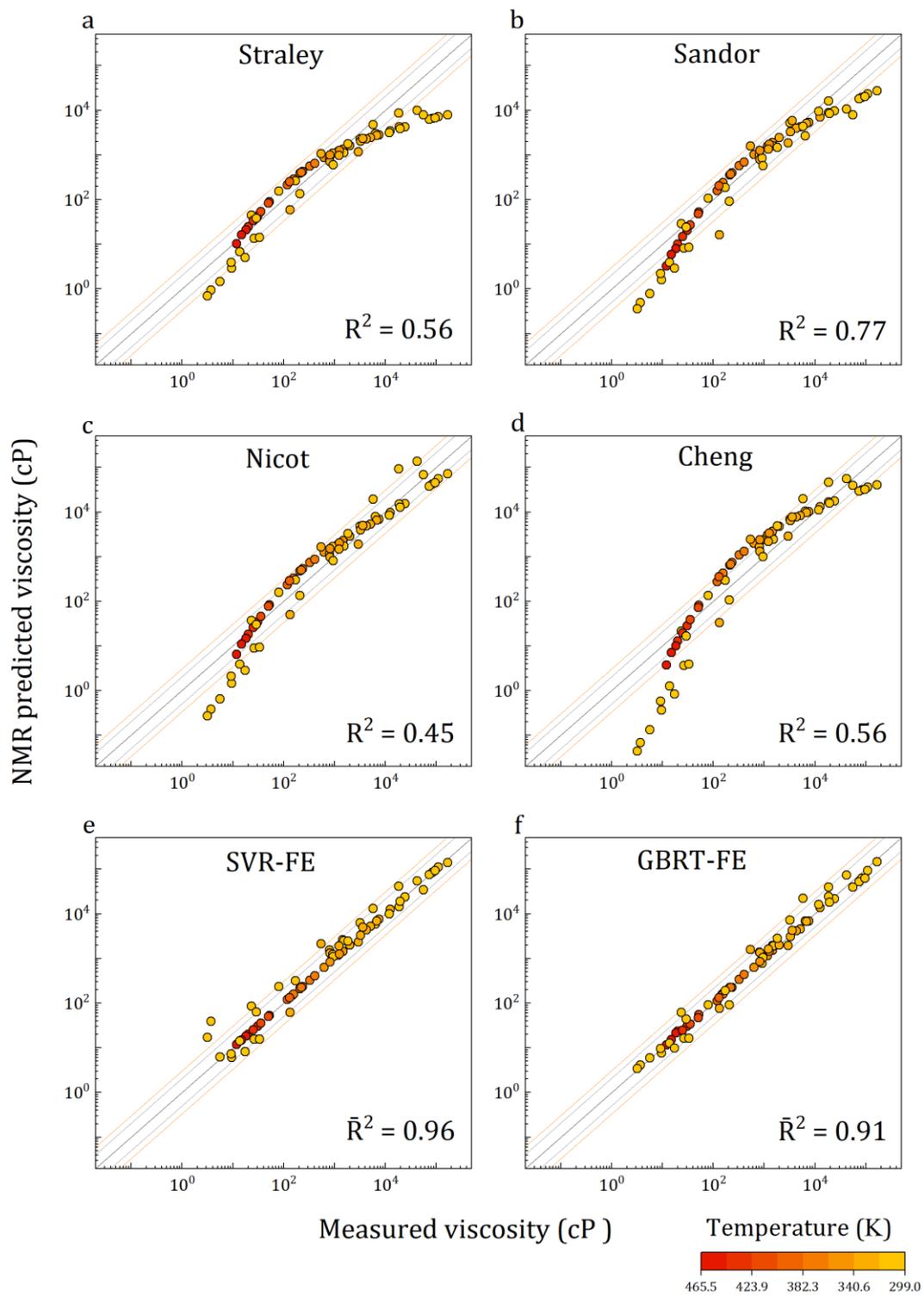
6 This difference in performance comes from the difference in the underlying
7 mathematical principles of these models. As the MLR model is *linear*, the
8 nonlinearity reduction from the log transformation naturally improved the
9 model's generalization and stability. Additionally, the integration of new features
10 reduced the variance of the predictions, which resulted in a further shrinking of
11 residuals. A similar is valid for the SVR model, though to a smaller extent (Figure
12 22f) ¹¹³. However, in the case of RF, GBRT, and DT models, the log transformation
13 did not impact the performance because the background tree-branching process
14 does not rely on numerical values of the features but instead uses the rank of the
15 features, which remained the same after transformations ⁹¹. Thus, the variance
16 reduction came solely from capturing more variability from the new features
17 derived in the second step of FE. Furthermore, when we compare the DT scores in
18 Figure 23, with those by GBRT, we observe a massive gap in performance, which
19 perfectly illustrates the advantage of ensembles of DTs over a single DT. One of the
20 reasons for the poor performance of single DT models is their 'habit' of overfitting
21 the training data, making them unstable with unseen data. Therefore, ensembles
22 of DTs generate a variance that minimizes the overfitting ⁹¹.
23

24 Lastly, the KNN is a simple algorithm where the output values are forecasted based
25 on the similarity between the input features. This similarity is calculated as a
26 distance (e.g., Euclidian, Manhattan, etc.) from k-instances, defined by the user ⁹³.
27 Feature engineering improves KNN scores almost proportionally to RF and GBRT
28 models, however not enough to minimize the large residuals in the high viscosity
29 domain, which causes the RMSE and MAE scores to rise (Figures 20a and 20b).
30

1 Another remark is that the significant temperature variations seem to have a
2 negligible effect on the performance of all supervised learning algorithms. The
3 predictions in the highest temperature domain overlap with the $x=y$ line in all six
4 cases, demonstrating that each algorithm has appropriately captured the
5 relationship between observed oil viscosity and NMR signal loss that occurs at
6 high temperatures. In comparison, empirical models tested in this work^{23,31,33,103},
7 exhibit poor performance in these conditions⁸⁶, as seen in the following chapter.
8

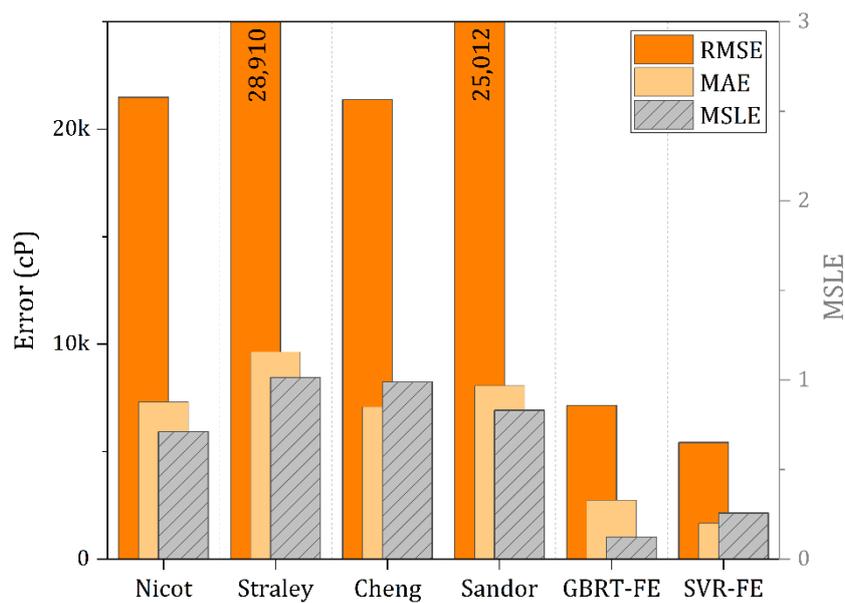
9 **3.3.2. EMPIRICAL NMR MODELS**

10 The performances of GBRT-FE and SVR-FE models are compared with four well-
11 known empirical NMR viscosity models based on T_{2lm} , TE, and T. These models
12 were developed by Straley et al.²³, Nicot et al.¹⁰³, Cheng et al.³³, and Sandor et al.
13³¹. Previous research showed that tuning by non-linear least squares (NLS)
14 improves the performance of empirical models⁸⁶. However, the viscosity dataset
15 in the present study has long-tailed distribution with many outliers at higher
16 viscosities, which dominate the sum of squares minimization, thus ultimately
17 leading to erroneous model fit and misleading statistical scores^{114,115}. Hence, the
18 model fitting was performed using the orthogonal distance regression (ODR),
19 proved to be a successful technique for dealing with outliers¹¹⁶. Figure 24 and
20 Figure 25 demonstrate the superior performance of both SVR-FE and GBRT-FE
21 models in terms of all statistical metrics. The curvature of the viscosity forecasts
22 by empirical models in Figure 24 reflects the combined influence of NMR signal
23 loss due to the Curie effect, which occurs at high temperatures, and fast relaxation
24 by solid-like components in heavy oil, which led to poor model generalization.



1
2
3
4

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.



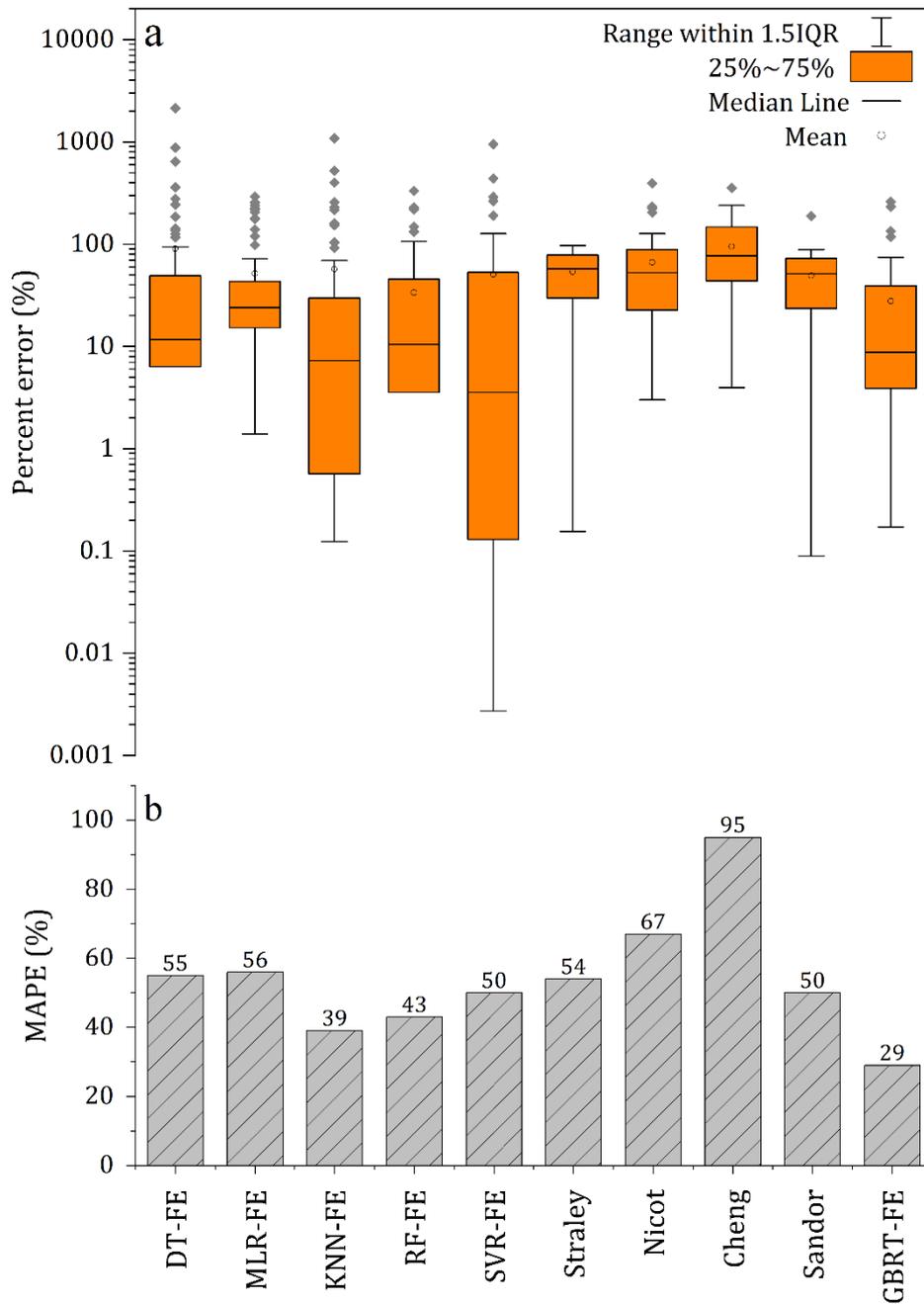
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

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.

1 **3.3.3. SVR-FE vs. GBRT-FE**

2 When cross-plots from Figure 22 and Figure 24 are analyzed along with scores in
3 Figure 23, Figure 25, one can conclude that SVR-FE and GBRT-FE models have
4 superior test performance than any other SL model. For instance, in the case of
5 MLR-FE, K-NN-FE, and RF-FE, the SVR-FE model, on average, scores 2.5 times
6 lower RMSE and MAE, while GBRT-FE achieves nearly two times lower scores.
7 When their performance is compared to empirical models, the difference in
8 performance is even more substantial; the SVR-FE model has about 4.5 times
9 lower RMSE and MAE scores, whereas GBRT-FE achieves nearly 3.5 times lower
10 scores.

11
12 The principal difference in the performance of these two models is related to their
13 precision (i.e., variance), which is evidenced by their different MSLE and MAPE
14 scores. For instance, the SVR-FE model has a better MSLE score than empirical
15 models (~4.5 times lower) but compared to SL models, KNN-FE and RF-FE
16 marginally outperform SVR-FE. The same is true for MAPE scores and percentage
17 error box plots when further examined in Figure 26. The GBRT-FE model, on the
18 other hand, scored the best MSLE and MAPE scores in this work. These results
19 imply that SVR-FE has the highest accuracy but somewhat lower precision (i.e.,
20 variance), relative to GBRT-FE, KNN-FE, and RF-FE models. For more convenience,
21 all evaluation scores are summarized in Table 9.



1
 2
 3
 4
 5
 6

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.

1 Table 8: Compared view of test set statistical scores for all SL and empirical models.
 2 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

4 3.3.4. PHYSICAL IMPLICATIONS OF SVR-FE AND GBRT-FE PERFORMANCE

5 From the physical point of view, meaningful insight can be extracted from the
 6 models' fundamental understanding. Although SVR-FE is moderately stable and
 7 achieves good accuracy, it struggles with capturing additional variability from the
 8 oil samples' diverse chemical compositions. This behavior can be observed in
 9 Figure 24f, where SVR-FE predictions favor the high-temperature heavy oil
 10 sample over the other samples. This occurs due to the structural risk minimization
 11 principle, which balances model complexity to avoid overfitting the training data
 12 and ensures the best possible generalization of new data. In other words, the
 13 model can be adapted to perform with more precision, but that would likely

1 deteriorate its generalization ability. However, two possible strategies could
2 rectify this; firstly, SVR heavily relies on feature engineering, which implies that
3 the SVR training on a set of lighter or more chemically alike oils would improve
4 the forecasts' precision by preserving good generalization. The second strategy
5 would be to expand the database by adding more NMR data from new samples.

6
7 On the other hand, GBRT-FE effectively handles the discrepancies from chemically
8 diverse set oil samples and a wide span of temperature and viscosity. This is due
9 to its stage-wise estimator addition principle, where overfitting is controlled by
10 tuning the learning rate and restriction of tree sizes. In this way, the GBRT
11 hyperparameters limit individual trees' contribution, but by adding many
12 estimators, the model manages to "learn" nuanced relationships that stem from
13 mixed oil chemistry, thus outperforming the SVR approach. As a result, GBRT-FE
14 achieves the best tradeoff between variance and bias for the task at hand at a
15 negligible increase in computational costs.

16
17 On another note, models presented in this study have certain limitations
18 originating from (a) NMR hardware configuration and (b) data availability.

19 a) One of the limitations of the presented SL models is that they were trained
20 on NMR oil data acquired with echo-spacing (TE) ranging from 0.1-0.3 ms.
21 Thus, the NMR data acquired using older NMR tools where echo-spacing
22 (TE) values are hardware-limited to longer TE (0.3-1.2 ms) might have less
23 reliable predictions. Reliability might be particularly problematic for heavy
24 oils and bitumens, where due to the fast relaxation of solid-like
25 constituents, the NMR device fails to measure the whole T₂-relaxation
26 spectra, which would cause the models to underpredict the real viscosity
27 ^{31,74,103}. However, by adding new NMR data to the dataset acquired using
28 longer TE, preferably from heavy oil samples, the SL algorithm could
29 capture the relationship between long TE and viscosity, compensating for
30 the undetected part of the T₂ spectra.

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

7

8 **3.4. SUMMARY**

9 In this study, we used SVR and GBRT algorithms to develop NMR models for oil
10 viscosity prediction using NMR T₂-relaxation time, echo-spacing and temperature
11 as an input, and dynamic oil viscosity as the target output. Also, a strategy to
12 reduce the variance of the forecasts was introduced, where domain knowledge
13 was used to implement feature engineering. Model performance was assessed
14 against four other popular SL algorithms and another four analytical models from
15 the literature. The SVR-FE and GBRT-FE have achieved statistically most favorable
16 scores in the study in terms of five error metrics: RMSE, MAE, MSLE, MAPE, and
17 \bar{R}^2 .

18

19 In summary, GBRT-FE demonstrated the best overall generalization ability, thus
20 producing predictions with a well-balanced variance-bias tradeoff. Consequently,
21 the use of GBRT-FE might prove as a viable solution in circumstances where a wide
22 span of oil types (light oils, heavy oils, and bitumens) is being tested at various
23 temperatures. Environments like these correspond to heavy oil reservoirs
24 undergoing or being screened for thermal treatment and other EOR approaches
25 such as solvent injection or a miscible gas injection¹². On the other hand, the SVR-
26 FE model exhibited a high accuracy but could not account for the variability
27 originating from the diverse chemical composition of the oils at the level that the
28 GBRT-FE model did. These findings indicate that SVR-FE would be a better choice
29 when sets of chemically more similar oils are being studied (e.g., only light or only
30 heavy oils) at various temperatures. In such conditions, the variance of SVR-FE

1 predictions would reduce to the degree where high accuracy and precision come
2 into play, such as in laboratory NMR characterization of petroleum fractions or
3 contactless non-invasive oil viscosity monitoring in mechanical systems.

4

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

9

1 Chapter 4 APPLICATION OF XGBOOST MODEL FOR *IN-SITU* WATER 2 SATURATION DETERMINATION IN CANADIAN OIL-SANDS BY LF-NMR AND 3 BULK DENSITY MEASUREMENTS

4 5 4.1. MOTIVATION

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

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

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

17
18 In this work, we postulated that the combination of LF-NMR T_2 data and bulk
19 density data could be used to effectively separate the contributions of oil and
20 water signals to a degree at which an accurate determination of relative water
21 fraction is possible. For model derivation, two machine learning approaches based
22 on Extreme Gradient Boosting (XGB) were employed. The first modeling approach
23 is based on a feature engineering process that reduces the number of inputs while
24 maximizing model generalization capacity. This was achieved by deriving new
25 features using empirical knowledge from the T_2 distribution analysis domain and
26 a feature extraction technique based on information theory. In contrast, the
27 second approach considers as input the whole NMR T_2 distribution of the sample,
28 aiming to preserve all available information originating from fluids residing in the
29 sample pore space. The dataset comprised 82 oil-sands core samples recovered

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

9 4.2. THEORY

10 4.2.1. LF-NMR MEASUREMENTS FOR WATER SATURATION DETERMINATION

11 Three main processes comprise the total T_2 relaxation; bulk relaxation, surface
 12 relaxation, and diffusion relaxation due to the gradient in a magnetic field. In this
 13 work, the benchtop LF-NMR relaxometer was used in which the gradient is absent,
 14 thus the diffusion term can be neglected. In such case, in the Equation 5 the
 15 $T_{2diffusion}$ term can be omitted:

$$\frac{1}{T_2} = \frac{1}{T_{2Bulk}} + \frac{1}{T_{2Surface}} \quad (53)$$

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

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

23 4.2.2. XGBOOST PRINCIPLES

24 XGBoost stands for Extreme Gradient Boosting (XGB), and it presents an
 25 implementation of the gradient boosting decision trees¹¹⁷. The main principle of
 26 gradient boosting is to utilize the individual weak learner, such as a decision tree,

1 and in a stage-wise manner, add iteratively new trees, to minimize further the
 2 objective function. This process continues for the specified number of boosting
 3 iterations, after which the prediction model is obtained in a final form. The
 4 algorithm uses gradient descent to minimize the loss function by finding the
 5 direction of the “steepest” descent. The loss function is minimized by finding a new
 6 best model (tree), and fitting it to the prediction residual. As this is performed in
 7 the step-wise manner, the size of the step is controlled by learning rate. If the step
 8 is too small the gradient descent will be slow, or may just converge to the local
 9 minimum. If it is too large, it will diverge. To find a minimum, the gradient descent
 10 uses a first order approximation which assumes that the loss surface is planar in
 11 the direction of the descent (i.e., in 3D, a plane tangent to the error surface). If the
 12 loss surface is not planar but convex, the gradient descent will ignore information
 13 about convexity, which may lead to a slow convergence or convergence to a local
 14 minimum. XGBoost, on the other hand, behaves much like Newton’s method¹¹⁸,
 15 which uses a second order approximation (i.e., in 3D, a convex quadratic surface
 16 with greater overlap with error surface) which assumes that the loss function is
 17 twice differentiable. The second order approximation allows considering the
 18 curvature of the loss function, and therefore ensures a faster convergence to the
 19 loss function minimum compared to the gradient descent. Although, Newton’s
 20 method requires calculating the Hessian matrix for storing the quadratic term
 21 coefficients of local second order function, the XGBoost instead computes the
 22 second partial derivative of the loss function (element-wise), which is less
 23 computationally demanding.

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

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

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

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

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

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

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

8 Here, γ is L1 and λ is L2 regularization parameters, T is the number of leaf nodes
 9 in a tree, and w_j^2 are the squared scores of leaves. Equation 57 shows that
 10 γ penalizes the growing complexity of the model (large T), while λ serves as a
 11 smoothing parameter for the learnt scores w_i to prevent overfitting. Like in
 12 gradient boosting regression trees, the model is trained by adding new trees in a
 13 stage-wise manner. For the prediction $\hat{y}_i^{(t)}$ at i-th instance and t-th stage:
 14

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

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

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

...

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

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

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

18

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

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

3 where g_i is the first order gradient and h_i is a second order gradient. The Equation
4 63 can be further improved by removing constant terms and expanding the
5 penalty term. These details can be found in the work of Chen et al. ¹¹⁷.

6

7 4.3. METHODOLOGY

8 4.3.1. EXPERIMENTAL PROCEDURE AND DATA PREPROCESSING

9 Oil-sand samples were collected in northern Alberta in Canada from a single
10 delineation well. Two sets of 82 whole core samples were recovered. The first set
11 was used for laboratory LF-NMR measurements, and a second set represented
12 sister samples used in Dean-Stark extraction for determining the relative fraction
13 of water, oil, and solids. Samples for NMR experiments were stored in glass vials
14 and measured using a Corespec 1000™ benchtop LF-NMR relaxometer at
15 reservoir temperature (6 °C) and ambient temperature (25 °C). The Carr-Purcell-
16 Meiboom-Gill (CPMG) pulse sequence was used for obtaining T₂-relaxation
17 distribution. The CPMG parameters were predetermined after a series of test NMR
18 experiments on different oil-sand samples. There were two aspects which had to
19 be taken into account. The first was to tune the CPMG parameters to detect the fast
20 relaxing heavy oil and clay-bound water signals. This was achieved by setting the
21 shortest echo time TE that the equipment allowed (0.2 ms). The second aspect was
22 achieving a lower signal-to-noise ratio (SNR) to simulate the well-logging *in-situ*
23 NMR tool output by reducing the number of trains, which in turn resulted in a
24 noisier signal. After trial rounds of measurements, the following parameters
25 produced optimal T₂ distribution and SNR (Table 10).

26

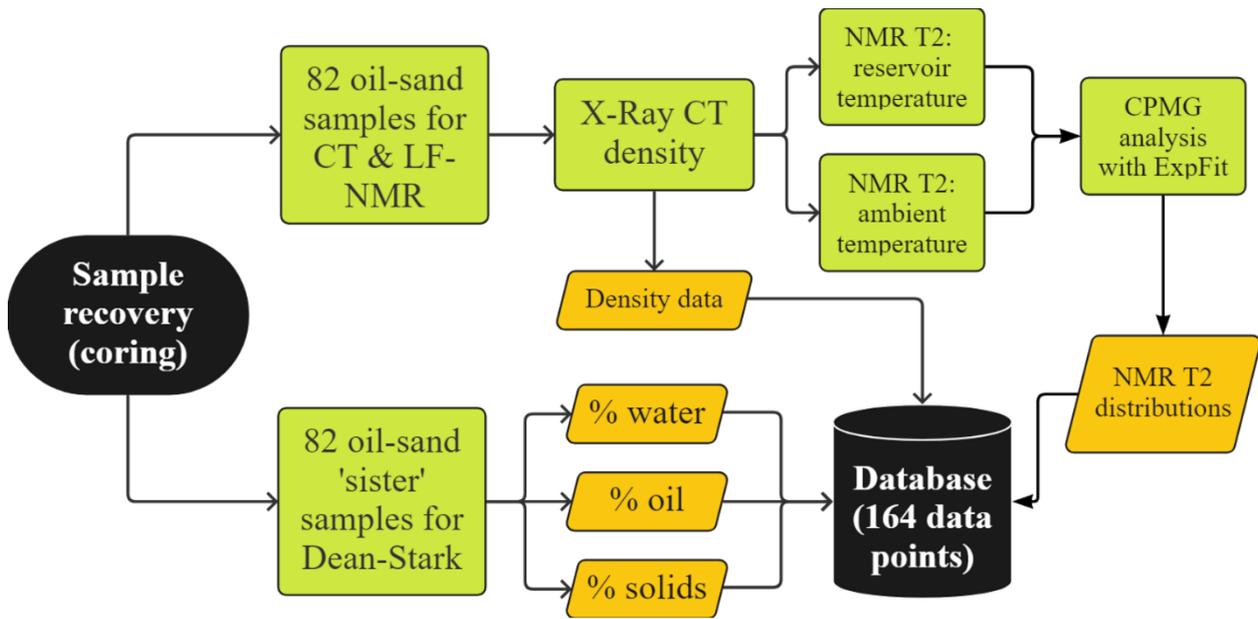
27

1 Table 9: Optimal CPMG pulse sequence parameters for detection of fast relaxing clay-
2 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

3
4 For the dataset, the range of SNR varied from 5 to 56, with an average of 23. The
5 ExpFit in-house software for multi-exponential analysis of the NMR signal was
6 used. The representation of the signal after Inverse Laplace Transform (ILT) was
7 obtained using Tikhonov regularization¹¹⁹. The practice has shown that the
8 regularization parameter helps avoid oscillations in solution associated with noise
9 and provides smooth T₂ distributions¹⁷. The regularization parameter can be
10 determined by direct and indirect methods such as Butler-Reed-Dawson, L-curve,
11 or generalized cross-validation¹⁷. In the case of oil-sands, after initial analysis, the
12 regularization parameter was determined directly and $\alpha=0.05$ was found to
13 provide the most stable solution for most samples. The density values of these
14 samples were measured beforehand by X-ray Computed Tomography (X-ray CT)
15 using GE 9800 CT scanner as a substitute for the density logging data.

16
17 The experimental program for Dean-Stark extraction, LF-NMR measurements,
18 and X-ray CT density measurements is illustrated in the flowchart (Figure 27).



1

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

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

11

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

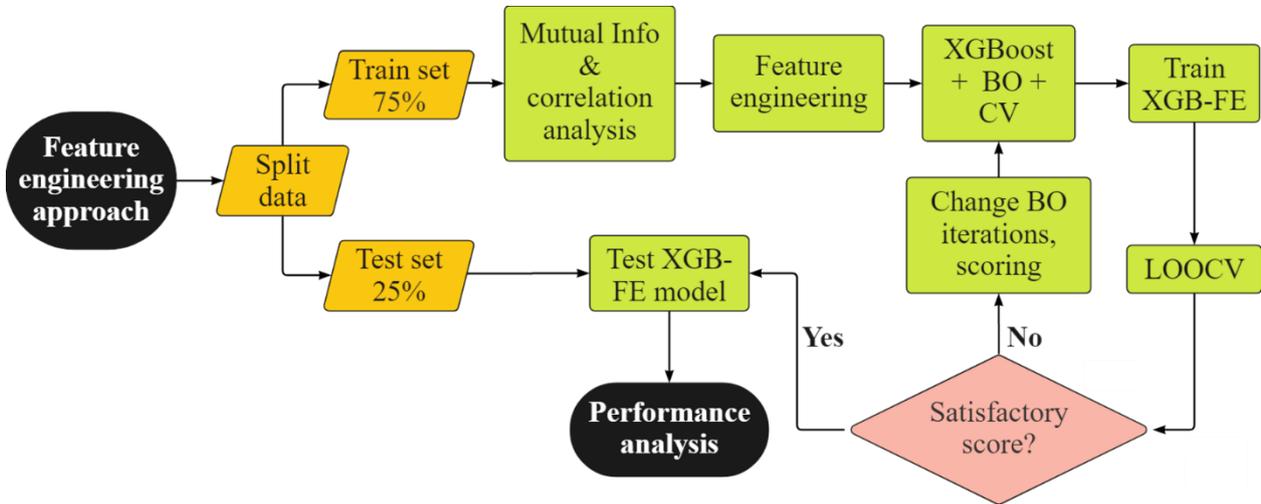
1 and residual distribution analysis. These steps will be discussed in detail in the
2 following sections.

3

4 **4.3.2. XGBOOST MODEL BASED ON FEATURE ENGINEERING (XGB-FE)**

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

10



11

12

Figure 28: Flowchart for XGB-FE model development.

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

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

7
8 As the T_2 distribution of relaxation times is represented on the semi-logarithmic
9 scale, the standard parameter for representing the average T_2 relaxation is T_2
10 logarithmic mean (T_{2lm}):

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

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

17
18 To account for the variation in T_2 distribution (i.e. narrow vs. wide peaks), the T_2
19 standard deviation was defined as:

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

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

21
22 The T_{2p} was defined as a location of a maximum value (peak) of the T_2 amplitude
23 on T_{2i} axis. This parameter is used in the petrophysical practice for the separation
24 of bound and producible fluids and fluid typing since T_{2p} gives an indication of
25 whether the largest amplitude portion of the signal corresponds to low or high T_2
26 values.

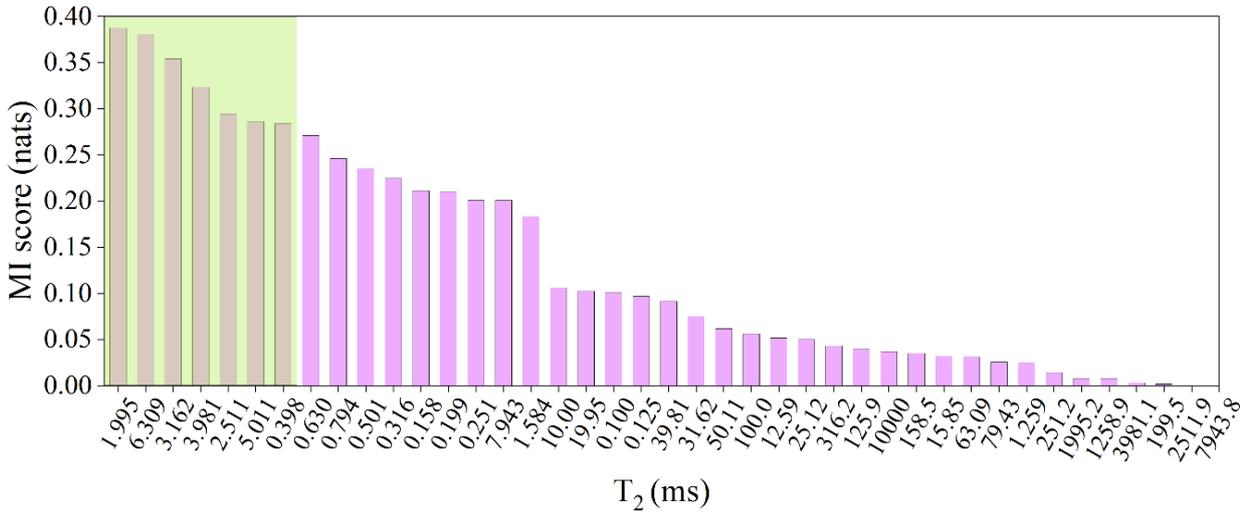
$$T_{2p} = \max(f(T_{2_1}), \dots, f(T_{2_n})) \quad (66)$$

1 Intelligent algorithms like XGBoost have gained popularity due to their ability to
 2 generalize complex data dependencies in large datasets and achieve state-of-the-
 3 art forecasting results. However, a small dataset is used in this study, where the
 4 overlapping of water and oil T₂ signals are likely to remain hidden or poorly
 5 represented. So, instead of allowing the algorithm to search through the whole
 6 NMR T₂ distribution, we can ‘show’ it where to look for the patterns and changes
 7 in the amplitude. One of the essential parts of the T₂ distribution in sandstones is
 8 the empirical clay-bound water T₂ cutoff located at 3 ms, which presents the
 9 boundary between capillary-bound and clay-bound fluids in a water-saturated
 10 core^{13,120}. In order to capture the possible T₂ response of clay-bound water and
 11 monitor its signal variation with different training samples, we defined a T₂ bound
 12 fluid (T_{2bf}) interval as:

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

13 However, this parameter cannot be used on its own to describe the changes in
 14 water content since the oil signal may also be located in the relevant interval. The
 15 true T₂ cutoff value in petrophysical practice is usually determined by performing
 16 lab tests on the saturated core samples (i.e., centrifuging), and even then, the use
 17 of a fixed or averaged T₂ cutoff value leads to the erroneous prediction of
 18 producible fluids. Instead, we attempt to obtain insights into the true T₂ cutoffs
 19 using a feature extraction technique called Mutual Information (MI) regression,
 20 based on the information entropy between variables. In classical regression
 21 analysis, statistical tests like F-test are carried out to study the degree of the linear
 22 association or continuous analysis of covariance (CANOVA) for the non-linear
 23 association between variables. However, mutual information is not ‘concerned’
 24 whether the variables have apparent linear correlation or covariance of zero, and
 25 they may still be stochastically dependent. This is the case in studying the changes
 26 in the conditional probability of one variable when another is modified¹²¹. In other
 27 words, by using MI regression, one can measure the level of association of the

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



6
 7 Figure 29: The mutual information regression results applied to the training set T_2
 8 distributions of the oil-sand samples relative to the Dean-Stark water content (DS-w).
 9 The shaded area presents the continuous cluster of T_2 responses with a strong mutual
 10 association with DS-w, which were used to calculate the T_2 cutoff range parameter – T_{2cr} .
 11 Figure 29 shows the relative mutual information scores of T_2 responses, where
 12 higher values indicate a stronger association with water content by Dean-Stark.
 13 For this dataset, the responses from 1.99-6.30 ms have the highest association
 14 with the water signal and form a continuous cluster between 10^0 and 10^1 decades
 15 along the T_2 semi-log scale, suggesting that most theoretical T_2 cutoff values lie in
 16 this interval. Therefore, the T_2 cutoff range (T_{2cr}), was defined as:

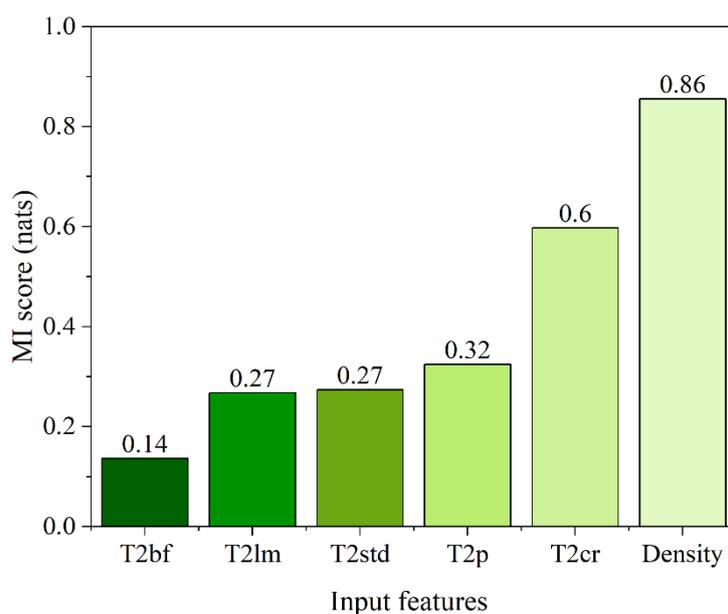
$$T_{2cr} = \sum_{1.99(ms)}^{6.30(ms)} A_i \quad (68)$$

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

1 recovery for the subsequent lab experiments is required. However, a common
 2 practice in well-logging is to combine NMR and bulk density logs to improve
 3 interpretation. Therefore, the bulk density was used as an additional parameter
 4 which we postulate is associated with T₂ surface relaxation and diffusive coupling.



5
 6 Figure 30: The diagonal correlation matrix shows the linear dependence between six
 7 input features with Dean-Stark water content (DS-w) in the training set. Scores
 8 represent Pearson's correlation coefficient and are color-coded (heatmap).



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

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, T₂ cutoff range and T₂ peak, and T₂ logarithmic mean features exhibit the strongest positive correlation with the water content by Dean-Stark (DS-w). The T₂ standard deviation shows a moderate degree of positive correlation, while T₂ 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).

1 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/m³)
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

2

3 Table 11: Descriptive statistics of the target variable Dean-Stark water saturation (DS-
4 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

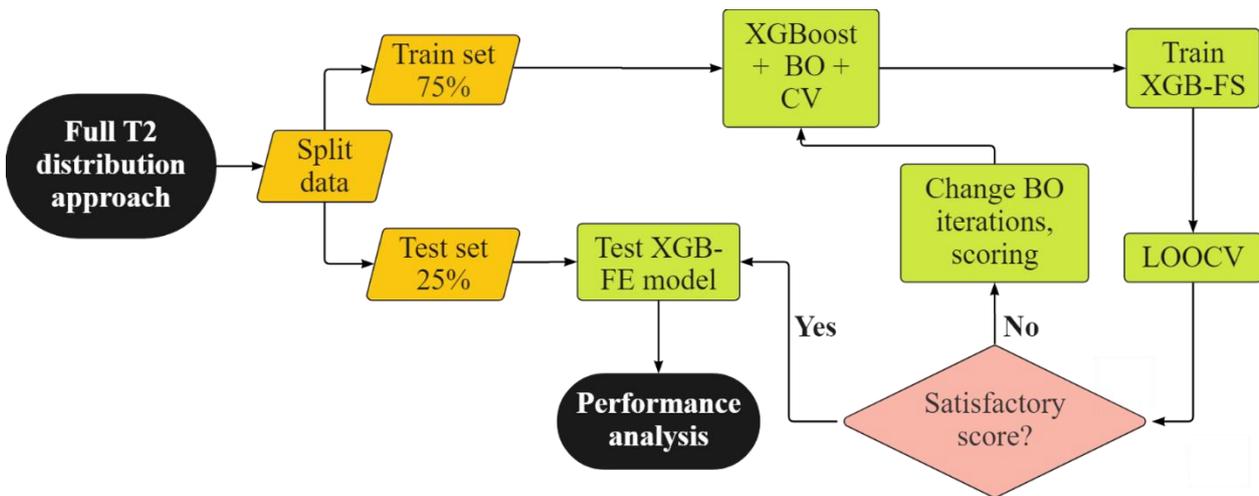
5

6 4.3.3. XGBOOST MODEL BASED ON THE FULL T₂ RELAXATION DISTRIBUTION 7 (XGB-FS)

8 The second modeling method facilitates the complete sample T₂ distribution.
9 There are two main incentives for this approach. First, the T₂ relaxation
10 distribution contains a large amount of information about the fluids residing in
11 the pore space, indicating that using a single or even a few features to characterize

1 the whole distribution may lead to significant information loss and, therefore to
 2 poor model forecasting performance⁷³. By using the entire T_2 distribution,
 3 variations such as changes in slope or local minima can implicitly be used to help
 4 separate oil and water signals. Secondly, predictions generated by the full- T_2
 5 distribution model provide a good baseline for comparison with the feature
 6 engineering and conventional deconvolution approaches. Therefore, the input
 7 features were arranged as $X = [A_1, A_2, A_3, \dots, A_{52}, \rho_i]$, where A_i is the i -th column
 8 vector of the amplitudes at the corresponding T_{2i} bin, and ρ_i is a column vector of
 9 density measurements. The water content by Dean-Stark (DS-w) was arranged
 10 as $Y = [DS-w_1, DS-w_2, \dots, DS-w_n]$, thus defining the dataset as $\{(X_i, Y_i)\}_{i=1}^n$ where n is
 11 the number of oil-sands samples. The complete XGB-FS model derivation the
 12 procedure is illustrated in Figure 32. **Statistical description of input features is**
 13 **summarized in Table 12, and the target variable description is in Table 11.**

14



15

16

Figure 32: Flowchart for XGB-FS model development.

17

18

19

20

21

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

	1	2	...	51	52
Statistic	$T_2(0.10\text{ms})$	$T_2(0.12\text{ms})$...	$T_2(7943.3)$	$T_2(10,000)$
	(a.u.)	(a.u.)		(a.u.)	(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

3

4 **4.3.4. MODEL OPTIMIZATION**

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

18

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

4

5 Table 12: Results of Bayesian Optimization on training set with 5-fold cross-validation,
 6 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']

7 **4.3.5. PERFORMANCE METRICS AND MODEL VALIDATION**

8 The forecasting performance of the models was evaluated using three
 9 performance metrics, including coefficient of determination (R^2), root mean
 10 squared error (RMSE), and mean absolute error (MAE). The R^2 is the positively
 11 oriented metric used in regression for representing the amount of model variance,
 12 and how well the model predictions generalize the observations. However, R^2
 13 alone does not provide information on prediction errors. The RMSE is another
 14 regularly employed error metric, used alongside R^2 , but under the assumption that
 15 residuals follow the normal distribution¹²². As a result of the heavy penalization
 16 of larger residuals, the RMSE is a convenient metric for revealing the differences
 17 in performance between multiple models with normally distributed residuals. At
 18 the same time, large residuals can cause the inflation of the RMSE score, which is
 19 why MAE can be used for additional evaluation. The MAE measures the mean
 20 magnitude of model prediction errors, but in contrast to RMSE, the errors are not

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

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

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

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

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

7

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

20

21 4.4. RESULTS

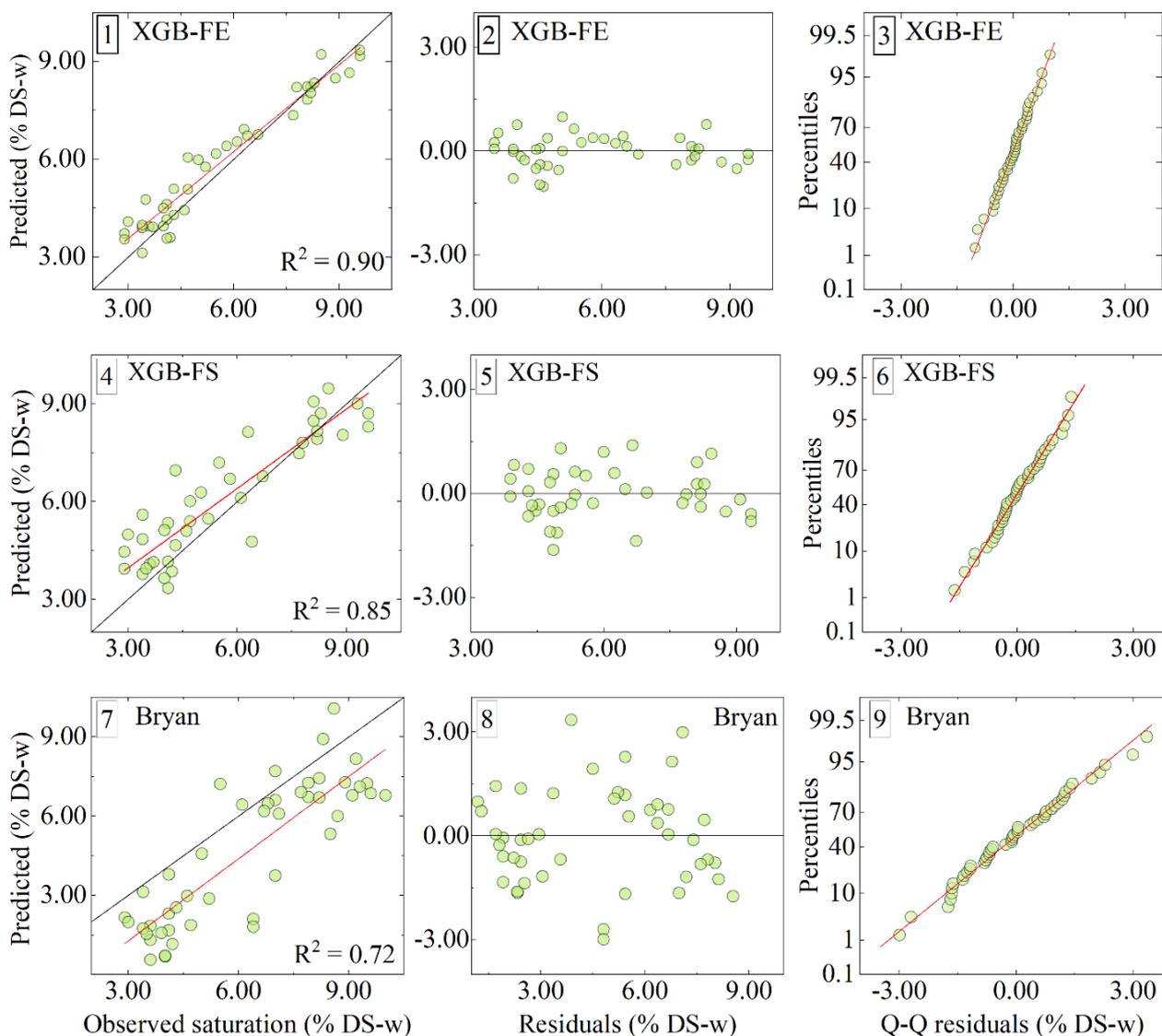
22 In this section, the performance of three models is presented, including the XGB-
 23 FE model based on the XGBoost algorithm with feature engineering, the XGB-FS

1 model based on the XGBoost algorithm using the whole sample T₂ distribution,
2 and a peak deconvolution approach (Bryan et al⁷). To assess the model
3 performance in more detail, residual plots (Figures 33-2, 33-5, and 33-8) and
4 quantile-quantile plots (Figures 33-3, 33-6, and 33-9) are used for the analysis of
5 the residual normality, and model variance and bias. All results are summarized in
6 Figures 33-34.

7

8 Analysis and comparison of error statistics, cross-plots, and distribution of
9 residuals indicate that the XGB-FE model achieves the highest accuracy and
10 generalization ability in the prediction of water content in oil-sand samples. Figure
11 33-1 shows that apart from slight overprediction in the 3-5% DS-w range, all XGB-
12 FE predictions spread along the x=y line with low variance, achieving the highest
13 R² score in the study (R²=0.90). Figure 33-2 shows the constant low variance of
14 the residuals, indicating that the model inputs capture variation in the data
15 correctly. Finally, the Q-Q plot (Figure 33-3) confirms the residual normality and
16 thereby the underlying assumption that XGB-FE model residuals follow the normal
17 distribution (low bias, low variance). Finally, the XGB-FE model achieves 1.5-3
18 times lower RMSE and MAE scores compared to the XGB-FS and Bryan et al.⁷
19 models indicating the best generalization ability of the three.

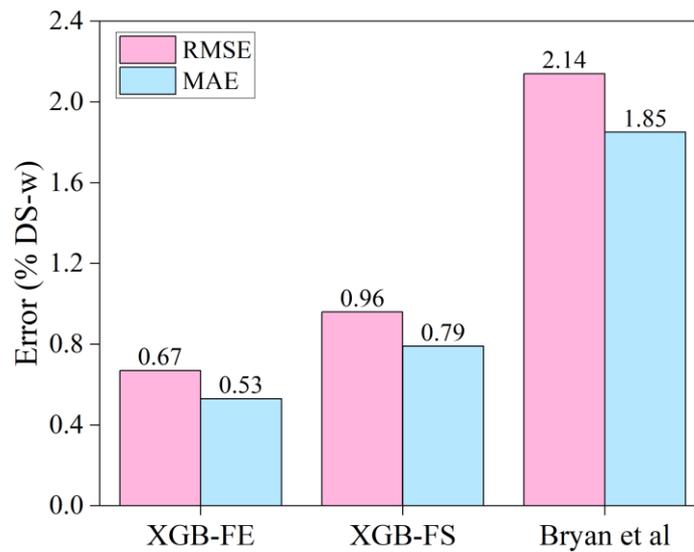
20



1

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

7



1

2

Figure 34: Comparison of RMSE and MAE **test prediction** scores for the three models

3

(*random_state=2*).

4

As for the XGB-FS model predictions, Figures 33-4, 33-5, and 33-6 show a similar

5

residual distribution to XGB-FE (normality and bias). However, the residual

6

variance is increased but constant, therefore achieving a somewhat lower R^2 score

7

($R^2=0.85$) and 1.5 times higher RMSE and MAE than XGB-FE. From Figure 33-7, it

8

can be observed that the Bryan et al.⁷ model generally tends to underpredict the

9

water content in samples. In addition, Figures 33-8 and 33-9 show inflated but the

10

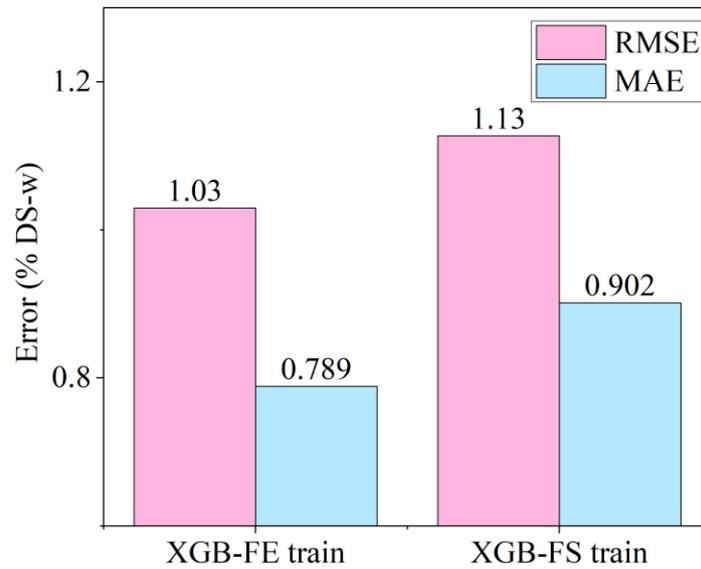
constant variance in the distribution of residuals, while residual normality still

11

holds with some local perturbing. As a result, Bryan et al.⁷ model RMSE and MAE

12

scores are the highest (Figure 34).



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

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 T₂ 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*).

1 Table 13: Comparison of XGB-FS model **test set prediction scores** with and without *bulk*
 2 *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

3
 4 Table 14: Comparison of XGB-FE model **test set prediction scores** with and without *bulk*
 5 *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

6
 7 Bulk density measurements are used together with LF-NMR measurements in
 8 petrophysical practice to improve the interpretation of well logs^{28,123}. LF-NMR
 9 measures the response of the fluids in the rock pore space, therefore carrying
 10 information about the fluids and pore size distribution of the rock. On the other
 11 hand, density logging equipment measures the response of the solids (rock
 12 matrix) together with fluids. The two are related in terms of T₂ surface relaxation,
 13 which depends on the rock pore to surface ratio with the fluid volume and the
 14 diffusive-coupling effect. This dependence can also be observed from the
 15 prediction test scores of the XGB-FS and XGB-FE models with and without bulk
 16 density as one of the model inputs. Prediction scores from Tables 13 and 14
 17 indicate that models achieve better scores with the integration of bulk density,

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

3
4 To affirm the second hypothesis: when XGB-FS and XGB-FE are compared (Figure
5 33), it can be observed that XGB-FE achieves better performance, especially in
6 terms of prediction variance. The XGB-FE variance reduction supports two
7 premises. First, the engineered features properly capture all the relevant
8 information from the T_2 distribution, indicating negligible information loss.
9 Secondly, in the feature engineering case, the XGBoost algorithm generalized the
10 variability in the data with the output more effectively, suggesting that for smaller
11 datasets, the appropriate feature engineering enables the XGBoost algorithm to
12 discover dependencies within the data more effectively than for a large number of
13 raw information (53 features in case of XGB-FS), **due to the high dimensionality**. In
14 other words, the new features contain all the relevant parts of the T_2 distribution
15 compressed into a few values, which reduces the XGB-FE model complexity and
16 enables better generalization of the relationship between inputs and a target
17 variable (DS-Sw).

18
19 According to Figure 31, along with a bulk density (MI=0.86 nats), the T_2 cutoff
20 range feature ranks second by MI score (0.60 nats), indicating the variability of the
21 sum of T_2 responses between 1.99 – 6.30 ms has a strong relationship with water
22 signal. The location of the T_2 peak (T_{2p}), T_2 standard deviation of the spectrum
23 (T_{2std}), and T_2 logarithmic mean (T_{2lm}) achieve similar MI scores (0.27 and 0.30
24 nats, respectively), signifying that these features alone do not capture enough
25 information about the water content. Finally, the sum of T_2 responses
26 representative of the empirical clay-bound water part of the T_2 distribution (0.1-
27 3.0 ms) shows the least association with the target (DS-w). Although these features
28 alone cannot explain variance in data effectively, their mutual interaction can
29 improve it. Since MI does not consider this mutual interaction between features
30 relative to the target output, the correlation matrix can be used. For instance,
31 Figure 30 shows that the T_{2cr} vs. T_{2p} and T_{2cr} vs. T_{2lm} have a strong positive

1 correlation (0.76 and 0.63, respectively), while T_{2bf} vs. T_{2lm} have a moderate
2 negative correlation. These interactions are likely to be generalized in the XGB-FE
3 model training process, thus explaining improved performance. Furthermore, the
4 permutation test score of XGB-FE using 150 permutations generated a P-value of
5 0.001, compared to the XGB-FS P-value of 0.007. In both cases, the P-value is well
6 below 0.05, showing a very low likelihood of obtaining such model performance
7 purely by chance.

8

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

19

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

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

27 (b) The SNR achieved by the benchtop LF-NMR relaxometers can be up
28 to 30 times higher than the SNR values obtained using well-logging
29 tools. In this study, the NMR signal-to-noise ratio was, on average 20,
30 which can still be considered high relative to the logging tools where the
31 SNR of 3-5 is considered satisfactory¹²⁴. Although the recent research

1 demonstrated that the XGBoost algorithm is sufficiently robust even
2 with noisy data¹²⁵, an additional validation using the data obtained by
3 the LF-NMR logging tools would be desirable. It is worth noting that, in
4 lower SNR samples, the deconvolution approach will be even more
5 challenging, and the value of using just the general properties of the T₂
6 distribution and XGBoost may be even further enhanced.

7
8 The XGB-FS method which uses all T₂ discrete responses as an input, is not optimal
9 approach as it has been shown, due to the high dimensionality of inputs. It is worth
10 noting that dimensionality reduction techniques could be used such as principal
11 component analysis, non-negative matrix factorization, linear discriminant
12 analysis and other methods, which would enable construction of less complex
13 models. This is planned to be done as a follow-up study.

In addition, the procedures
14 for NMR measurements with a controlled saturation and desaturation of samples,
15 similar to those reported in recent literature⁶⁶, would enable deeper sensitivity
16 analysis of the features derived in this work and further improvement XGB-FE
17 model. In such a setup, the Dean-Stark measurements could be replaced by the
18 more cost and time-effective mass-volume measurements, ultimately allowing the
19 collection of a larger database, at which point the application of artificial neural
20 networks (ANNs) would be possible.

21
22 It is also worth noting that logging equipment configuration can be substantially
23 different from desktop NMR relaxometers, which may cause inconsistencies
24 between NMR T₂ distributions obtained in the lab and the field. This can cause the
25 variable performance of proposed NMR data-driven model, which is why the
26 parameters of the NMR logging device, such as TW, TE and number of trains,
27 should be relatively consistent to the values reported in this study.

28

1 **4.6. SUMMARY**

2 This study presents the approach which integrates extreme gradient boosting with
3 LF-NMR measurements and bulk density data for the water saturation
4 determination in oil-sands. Two models were developed using full NMR T₂
5 distribution (XGB-FS), and feature engineering (XGB-FE). It is concluded that;

- 6 • Feature extraction methods such as mutual information regression can
7 effectively select the most relevant information from NMR T₂ distribution.
- 8 • Integrating bulk density data as a model input notably improves the XGB-
9 FS and XGB-FE forecasting performance.
- 10 • XGB-FE achieved RMSE = 0.67%, MAE = 0.53% and R² = 0.90 in predicting
11 relative water content by Dean-Stark, a substantial improvement compared
12 to deconvolution method.

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

15
16

1 Chapter 5 CONCLUSIONS

2

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

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

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

8
9 This work also lays the foundations for further research. The following
10 recommendations can be made:

- 11
12 1. Integrating bulk density measurements into the machine learning models
13 for viscosity determination to perform additional validation for *in-situ*
14 applications.
- 15 2. Additional LF-NMR and bulk density measurements (dataset expansion)
16 would further improve the understanding of T₂ surface relaxation and
17 diffusive coupling in connected micro- and macropores. This would also
18 enable us to use artificial neural networks with more success.
- 19 3. Performing a deeper study on the effect of bulk density data on the water
20 saturation prediction by LF-NMR measurements and studying the effects
21 of noise. Another interesting study would be the application of
22 dimensionality reduction methods on NMR T₂ distribution.
- 23 4. The combination of other conventional logging data with NMR data for
24 machine learning modeling shows excellent potential for other *in-situ*
25 applications such as rock wettability characterization or quantification of
26 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 T₂ 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 1.1 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

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

3

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

Statistic	T_{2std}	T_{2p}	T_{2lm}	T_{2cr}	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

5

6

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

Statistic	T_{2std}	T_{2p}	T_{2lm}	T_{2cr}	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

8

9

10

11

12

13

14

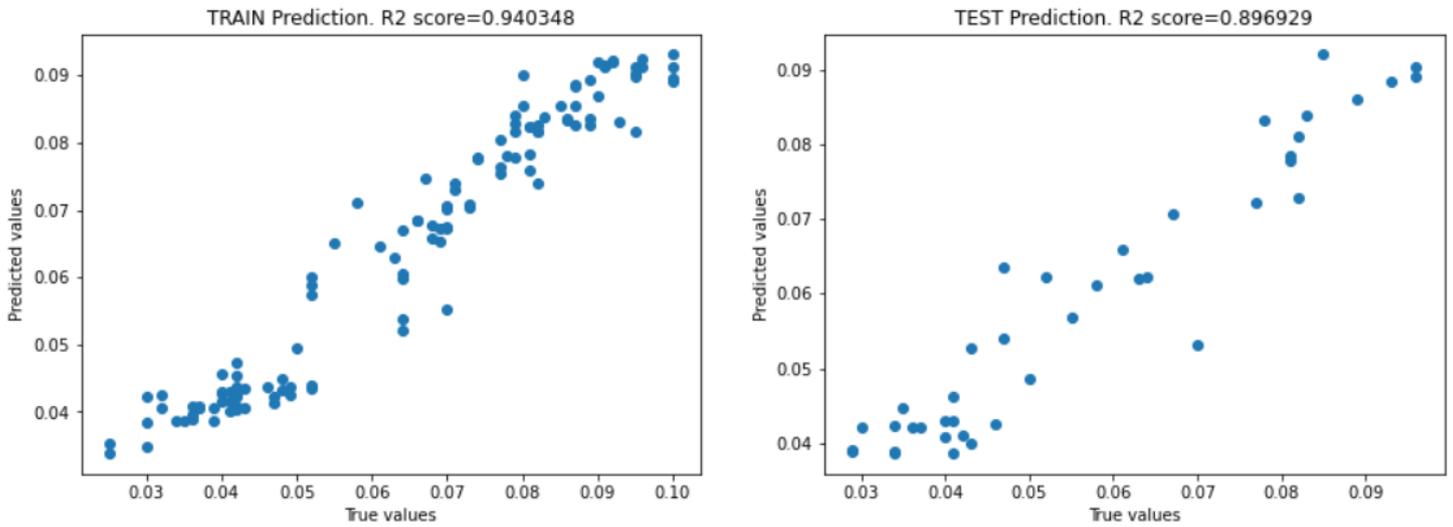
1 Table 17: Train and test set descriptive statistics of the target feature (Dean-Stark water
 2 saturation) after scaling.

Target-Train set		Target-Test set	
Statistic	DS-Sw	Statistic	DS-Sw
Count	123,00	Count	41,00
Mean	-0,08	Mean	-0,25
Std	0,54	Std	0,53
Min	-1,05	Min	-0,95
25%	-0,63	25%	-0,68
50%	0,00	50%	-0,43
75%	0,38	75%	0,28
Max	0,83	Max	0,73

3
 4
 5
 6
 7
 8
 9
 10
 11
 12
 13
 14
 15
 16
 17
 18
 19
 20
 21

1 **A 1.2 Random forests**

2



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

4

5 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]

6

7

8

9

10

11

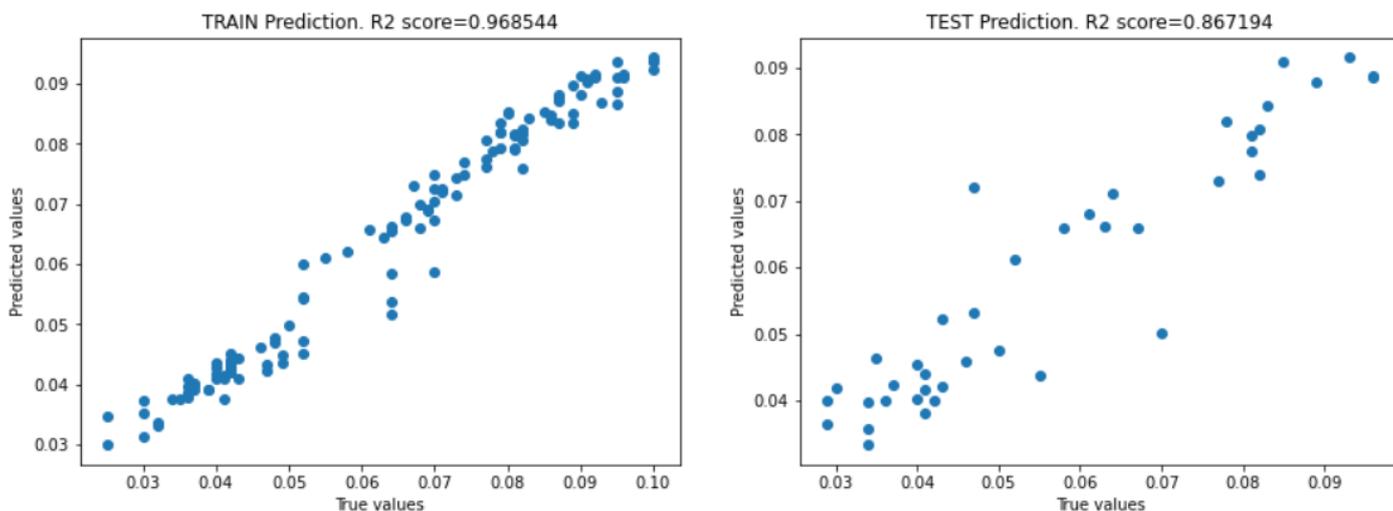
12

13

14

1 A 1.3 Gradient Boosting Regression Trees

2



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

4

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

6

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]

7

8

9

10

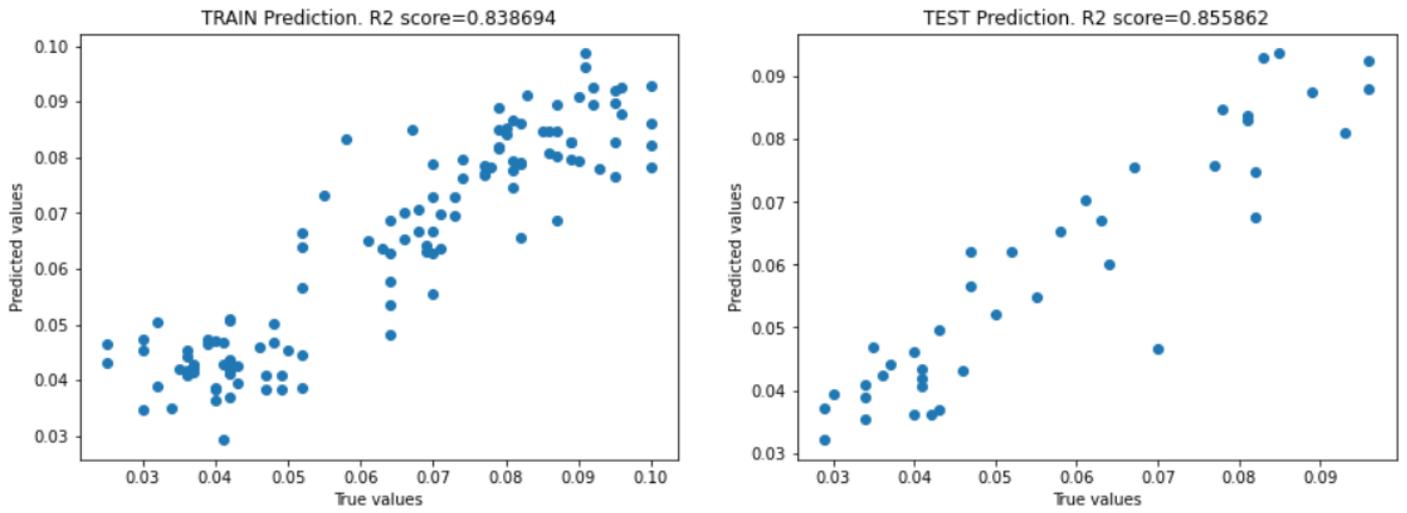
11

12

13

1 **A 1.4 Gaussian Process Regression**

2



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

4

5 **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]

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

7

8

9

10

11

12

13

14

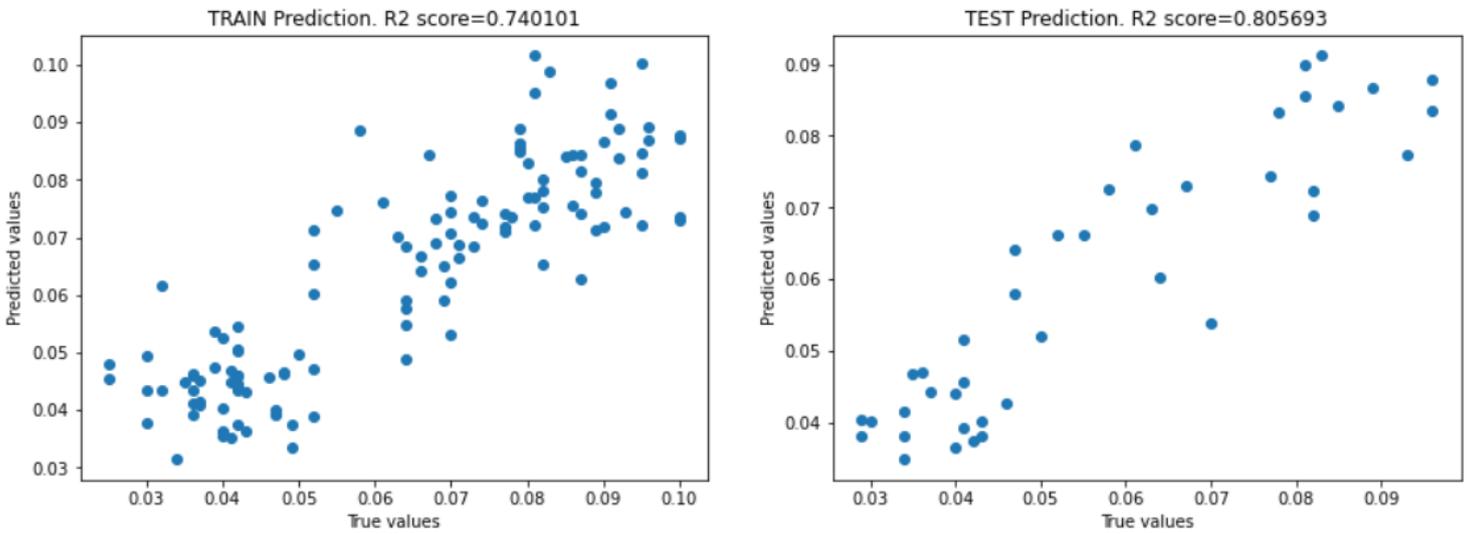
15

16

17

1 **A 1.5 Elastic Net**

2



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

4

5 **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]

6

7

8

9

10

11

12

13

14

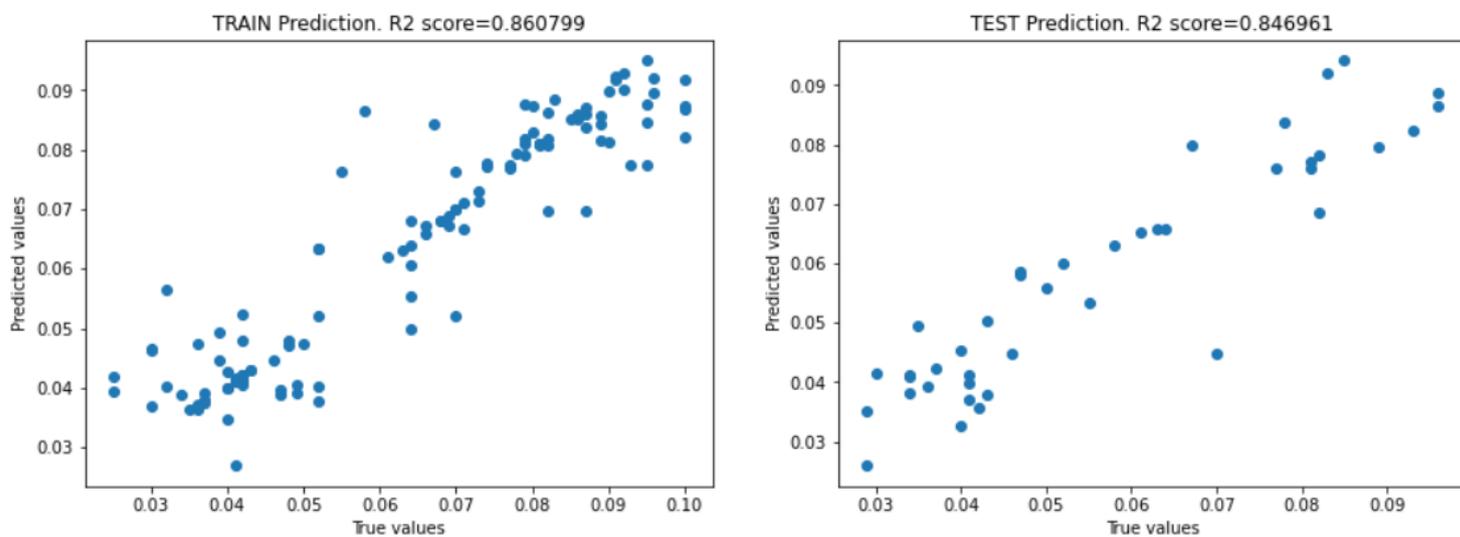
15

16

17

1 **A 1.6 Support Vector Regression**

2



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

4

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

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

7

8

9

10

11

12

13

14

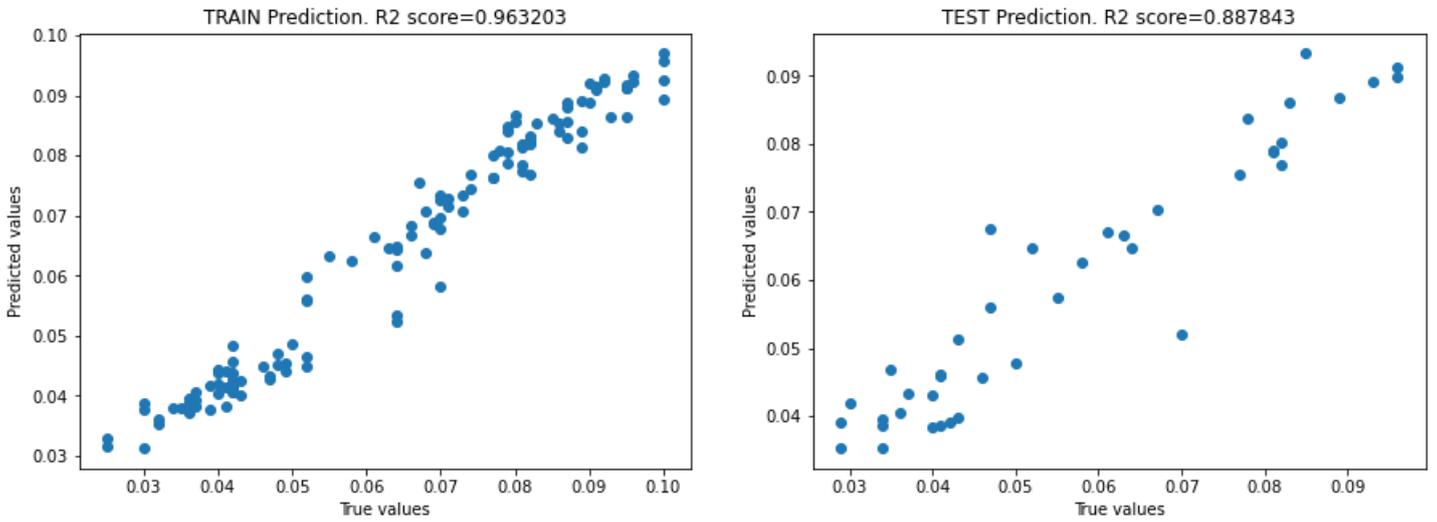
15

16

17

1 **A 1.7 XGBoost (constrained)**

2



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

4

5 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']

6

7

8

9

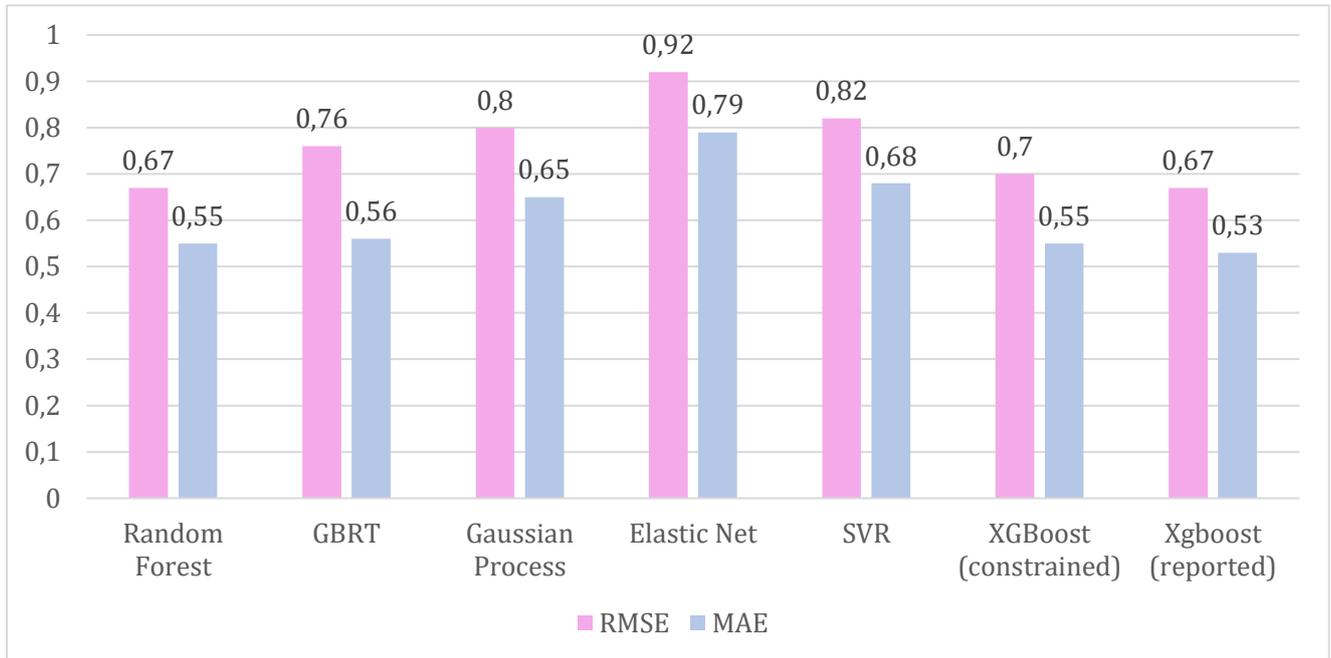
10

11

12

1 **A 1.8 Results summary**

2



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

4

5

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

6

7

1 Bibliography

- 2 1. Alboudwarej, H. *et al.* Highlighting Heavy Oil. *Oilf. Rev.* **18**, 34–53 (2006).
- 3 2. Hascakir, B. Introduction to Thermal Enhanced Oil Recovery (EOR) special
4 issue. *J. Pet. Sci. Eng.* **154**, 438–441 (2017).
- 5 3. Musin, K., Abdullin, T., Shipunov, T. & Khisamov, R. Estimation Viscosity and
6 its Heterogeneity by NMR Logging Tool in Reservoir Conditions in Oilfield
7 with Heavy Oil: Practical Results. *SPE Russ. Pet. Technol. Conf. Exhib.* 1–19
8 (2016) doi:10.2118/182066-RU.
- 9 4. James, L. A., Rezaei, N. & Chatzis, I. VAPEX, warm VAPEX and hybrid VAPEX
10 - The state of enhanced oil recovery for in situ heavy oils in Canada. *J. Can.*
11 *Pet. Technol.* **47**, 12–18 (2008).
- 12 5. Das, S. K. Vapex: An Efficient Process for the Recovery of Heavy Oil and
13 Bitumen. *SPE J.* **3**, 232–237 (1998).
- 14 6. Cui, G., Liu, T., Xie, J., Rong, G. & Yang, L. A review of SAGD technology
15 development and its possible application potential on thin-layer super-
16 heavy oil reservoirs. *Geosci. Front.* **13**, 101382 (2022).
- 17 7. Sheng, J. J. *Cyclic Steam Stimulation. Enhanced Oil Recovery Field Case Studies*
18 (Elsevier Inc., 2013). doi:10.1016/B978-0-12-386545-8.00016-6.
- 19 8. Sheng, J. J. *Steam Flooding. Enhanced Oil Recovery Field Case Studies* (Elsevier
20 Inc., 2013). doi:10.1016/B978-0-12-386545-8.00015-4.
- 21 9. Turta, A. *In Situ Combustion. Enhanced Oil Recovery Field Case Studies*
22 (Elsevier Inc., 2013). doi:10.1016/B978-0-12-386545-8.00018-X.
- 23 10. Djuraev, U. A review on conceptual and practical oil and gas reservoir
24 monitoring methods. *J. Pet. Sci. Eng.* **152**, 586–601 (2017).
- 25 11. Shikhov, I., Li, R. & Arns, C. H. Relaxation and relaxation exchange NMR to
26 characterise asphaltene adsorption and wettability dynamics in siliceous
27 systems. *Fuel* **220**, 692–705 (2018).
- 28 12. Fayazi, A., Kryuchkov, S. & Kantzas, A. Evaluating Diffusivity of Toluene in
29 Heavy Oil Using Nuclear Magnetic Resonance Imaging. *Energy and Fuels* **31**,
30 1226–1234 (2017).
- 31 13. Coates, G. R., Xiao, L. & Prammer, M. G. NMR logging. *Ebooks* 253 (1999).

- 1 14. Kantzas, A. & Bryan, J. L. Tipping process for NMR: Fundamentals of Fluid
2 Flow in porous media. [https://perminc.com/resources/fundamentals-of-](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/)
3 [fluid-flow-in-porous-media/chapter-3-molecular-diffusion/diffusion-](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/)
4 [coefficient/measurement-techniques/nmr-method/principles-nmr-](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/)
5 [processing/tipping-process/](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).
- 6 15. Zhang, Q., Lo, S. & Huang, C. Some exceptions to default NMR rock and fluid
7 properties. *SPWLA 39th Annu. ... FF* (1998).
- 8 16. Testamanti, M. N. & Rezaee, R. Determination of NMR T₂ cut-off for clay
9 bound water in shales: A case study of Carynginia Formation, Perth Basin,
10 Western Australia. *J. Pet. Sci. Eng.* **149**, 497–503 (2017).
- 11 17. Testamanti, M. N. & Rezaee, R. Considerations for the acquisition and
12 inversion of NMR T₂ data in shales. *J. Pet. Sci. Eng.* **174**, 177–188 (2019).
- 13 18. Bryan, J., Kantzas, A. & Bellehumeur, C. Oil-Viscosity Predictions From Low-
14 Field NMR Measurements. *SPE Reserv. Eval. Eng.* **8**, 44–52 (2005).
- 15 19. Bloembergen, N., Purcell, E. M. & Pound, R. V. Relaxation Effects in Nuclear
16 Magnetic Resonance Absorption. *Phys. Rev.* **73**, 679–712 (1948).
- 17 20. Khan, M. A. B., Mehrotra, A. K. & Svrcek, W. Y. Viscosity Models for Gas-Free
18 Athabasca Bitumen. *J. Can. Pet. Technol.* **23**, 47–53 (1984).
- 19 21. Meyer, R., Attanasi, E. & Freeman, P. Heavy Oil and Natural Bitumen
20 Resources in Geological Basins of the World. *Usgs* **1084**, 36 (2007).
- 21 22. Luo, P. & Gu, Y. Effects of asphaltene content on the heavy oil viscosity at
22 different temperatures. *Fuel* **86**, 1069–1078 (2007).
- 23 23. Straley, C., Rossini, D., Vinegar, H., Tutunjian, P. & Morriss, C. Core Analysis
24 by Low Field NMR. *Log Anal.* **38**, 84–94 (1997).
- 25 24. Straley, C. Reassessment of Correlations of between Viscosity and NMR
26 Measurements. *Spwla 47 AA* (2006).
- 27 25. Nicot, B., Fleury, M. & Leblond, J. Improvement of viscosity prediction using
28 NMR relaxation. *48th Annu. Logging Symp.* 1–7 (2007).
- 29 26. Morriss, C. E. *et al.* Hydrocarbon saturation and viscosity estimation from
30 NMR logging in the belridge diatomite. *SPWLA 35th Annu. Logging Symp.*
31 *1994* (1994).

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

- 1 and applications of machine learning in solid-state materials science. *npj*
2 *Comput. Mater.* **5**, 83 (2019).
- 3 40. Lei, Y. *et al.* Applications of machine learning to machine fault diagnosis: A
4 review and roadmap. *Mech. Syst. Signal Process.* **138**, 106587 (2020).
- 5 41. Gao, Z., Zou, X., Huang, Z. & Zhu, L. Predicting sooting tendencies of
6 oxygenated hydrocarbon fuels with machine learning algorithms. *Fuel* **242**,
7 438–446 (2019).
- 8 42. Myshkin, N. K. *On-line Condition Monitoring in Industrial Lubrication and*
9 *Tribology*.
- 10 43. Barbosa, L. L. *et al.* Relative hydrogen index as a fast method for the
11 simultaneous determination of physicochemical properties of petroleum
12 fractions. *Fuel* **210**, 41–48 (2017).
- 13 44. Constantino, A. F. *et al.* Determination of physicochemical properties of
14 biodiesel and blends using low-field NMR and multivariate calibration. *Fuel*
15 **237**, 745–752 (2019).
- 16 45. Alptekin, E. & Canakci, M. Determination of the density and the viscosities
17 of biodiesel–diesel fuel blends. *Renew. Energy* **33**, 2623–2630 (2008).
- 18 46. Li, G. Z., Meng, H. H., Yang, M. Q. & Yang, J. Y. Combining support vector
19 regression with feature selection for multivariate calibration. *Neural*
20 *Comput. Appl.* **18**, 813–820 (2009).
- 21 47. Li, H. *et al.* OPEN A novel multi-target regression framework for time-series
22 prediction of drug efficacy. *Nat. Publ. Gr.* 1–9 (2017)
23 doi:10.1038/srep40652.
- 24 48. Truhan, J. J., Qu, J. & Blau, P. J. The effect of lubricating oil condition on the
25 friction and wear of piston ring and cylinder liner materials in a
26 reciprocating bench test. *Wear* **259**, 1048–1055 (2005).
- 27 49. Khakimova, L. *et al.* High-pressure air injection laboratory-scale numerical
28 models of oxidation experiments for Kirsanovskoye oil field. *J. Pet. Sci. Eng.*
29 **188**, 106796 (2020).
- 30 50. US6794865 Method and apparatus for monitoring the health of a fluid
31 system , particularly a gas turbine oil system. (2004).

- 1 51. Arola, D. F., Barrall, G. A., Powell, R. L., McCarthy, K. L. & McCarthy, M. J. Use
2 of nuclear magnetic resonance imaging as a viscometer for process
3 monitoring. *Chem. Eng. Sci.* **52**, 2049–2057 (1997).
- 4 52. Liang, X., Chang-chun, Z., Zhi-qiang, M., Yu-jiang, S. & Yan, J. Estimation of
5 water saturation from nuclear magnetic resonance (NMR) and
6 conventional logs in low permeability sandstone reservoirs. *J. Pet. Sci. Eng.*
7 **108**, 40–51 (2013).
- 8 53. Donaldson, E. C. *Well logging for earth scientists. Journal of Petroleum Science*
9 *and Engineering* vol. 2 (1989).
- 10 54. Tiab, D. Formation Resistivity and Water Saturation. in *Petrophysics*
11 (Elsevier, 2012). doi:10.1016/B978-0-12-383848-3.00004-9.
- 12 55. Wang, W. & Wei, W. Chapter 3 - Water chemistry. in *Fluid Chemistry, Drilling*
13 *and Completion* (ed. Wang, Q.) 95–114 (Gulf Professional Publishing, 2022).
14 doi:https://doi.org/10.1016/B978-0-12-822721-3.00005-8.
- 15 56. Mohammadlou, M. H. & Langeland, H. Use of the NMR and Resistivity Logs
16 to Quantify Movable Hydrocarbon; Solution for the Tight and Low-
17 Resistivity Carbonate Reservoirs. *SPE* **141047**, (2011).
- 18 57. Cowie, B. R., James, B. & Mayer, B. Distribution of total dissolved solids in
19 McMurray Formation water in the Athabasca oil sands region, Alberta,
20 Canada: Implications for regional hydrogeology and resource development.
21 *Am. Assoc. Pet. Geol. Bull.* **99**, 77–90 (2015).
- 22 58. Zheng, S. *et al.* Nuclear magnetic resonance T2 cutoffs of coals: A novel
23 method by multifractal analysis theory. *Fuel* **241**, 715–724 (2019).
- 24 59. Bryan, J., Mai, A., Hum, F. M. & Kantzas, A. Oil- and water-content
25 measurements in bitumen ore and froth samples using low-field NMR. *SPE*
26 *Reserv. Eval. Eng.* **9**, 654–663 (2006).
- 27 60. Yang, Z. & Hirasaki, G. J. NMR measurement of bitumen at different
28 temperatures. *J. Magn. Reson.* **192**, 280–293 (2008).
- 29 61. Bryan, J., Moon, D. & Kantzas, A. In situ viscosity of oil sands using low field
30 NMR. *J. Can. Pet. Technol.* **44**, 23–29 (2005).
- 31 62. Anand, V. & Hirasaki, G. J. Diffusional coupling between micro and

- 1 macroporosity for NMR relaxation in sandstones and grainstones. *SPWLA*
2 *46th Annu. Logging Symp. 2005* **48**, 289–307 (2005).
- 3 63. Singer, P. M., Chen, Z., Wang, X. & Hirasaki, G. J. Diffusive coupling in heptane-
4 saturated kerogen isolates evidenced by NMR T1-T2 and T2-T2 maps. *Fuel*
5 **280**, 118626 (2020).
- 6 64. Mukhametdinova, A., Habina-Skrzyniarz, I., Kazak, A. & Krzyżak, A. NMR
7 relaxometry interpretation of source rock liquid saturation — A holistic
8 approach. *Mar. Pet. Geol.* **132**, (2021).
- 9 65. Newgord, C., Tandon, S. & Heidari, Z. Simultaneous assessment of
10 wettability and water saturation using 2D NMR measurements. *Fuel* **270**,
11 117431 (2020).
- 12 66. Krzyżak, A. T., Habina-Skrzyniarz, I., Machowski, G. & Mazur, W. Overcoming
13 the barriers to the exploration of nanoporous shales porosity. *Microporous*
14 *Mesoporous Mater.* **298**, (2020).
- 15 67. Venkataramanan, L. *et al.* An unsupervised learning algorithm to compute
16 fluid volumes from NMR T1-T2 logs in unconventional reservoirs.
17 *Petrophysics* **59**, 617–632 (2018).
- 18 68. Burcaw, L. *et al.* Improved Methods for Estimating the Viscosity of Heavy
19 Oils From Nmr Data. *Spwla* **49**, 1–14 (2008).
- 20 69. Miller, K. A. Should You Trust Your Heavy Oil Viscosity Measurement ?
- 21 70. Nicot, B., Fleury, M. & Leblond, J. A New Methodology For Better Viscosity
22 Prediction Using Nmr Relaxation. *SPWLA 47th Annu. Logging Symp.* 1–12
23 (2006).
- 24 71. Bloch, F. Nuclear Induction. *Phys. Rev.* **70**, 460–474 (1946).
- 25 72. P. Lindsey, C. & Patterson, G. *Detailed Comparison of the Williams-Watts and*
26 *Cole-Davidson Functions. The Journal of Chemical Physics* vol. 73 (1980).
- 27 73. Ahmad, K. *et al.* Radial-basis-function-based nuclear magnetic resonance
28 heavy oil viscosity prediction model for a Kuwait viscous oil field.
29 *Interpretation* **4**, SF81–SF92 (2016).
- 30 74. LaTorraca, G. A., Stonard, S. W., Webber, P. R., Carison, R. M. & Dunn, K. J.
31 Heavy oil viscosity determination using NMR logs. *SPWLA 40th Annu.*

- 1 *Logging Symp.* **3**, 11 (1999).
- 2 75. Galford, J. E. & Marschall, D. M. Combining NMR and Conventional Logs to
3 Determine Fluid Volumes and Oil Viscosity in Heavy-Oil Reservoirs. *SPE*
4 *Annu. Tech. Conf. Exhib.* **12** (2007) doi:10.2118/63257-ms.
- 5 76. Mirotchnik, K. D. & Allsopp, K. Low-Field NMR Method for Bitumen Sands
6 Characterization : A New Approach. *SPE J.* 88–96 (2001).
- 7 77. Bryan, J., Mirotchnik, K. & Kantzas, A. Viscosity determination of heavy oil
8 and bitumen using NMR relaxometry. *J. Can. Pet. Technol.* **42**, 29–34 (2003).
- 9 78. Yang, Z. & Hirasaki, G. J. NMR measurement of bitumen at different
10 temperatures. *J. Magn. Reson.* **192**, 280–293 (2008).
- 11 79. Bryan, J., Mirotchnik, K. & Kantzas, A. Viscosity Determination of Heavy Oil
12 and Bitumen Using NMR Relaxometry. **42**, (2003).
- 13 80. American Society for Testing Materials, (ASTM D4287). Standard Test
14 Method for High-Shear Viscosity Using a Cone/Plate Viscometer. *B. Stand.*
15 **06.01**,.
- 16 81. Tang, Y., Mccowan, D. & Song, Y. OPEN A miniaturized spectrometer for NMR
17 relaxometry under extreme conditions. *Sci. Rep.* 1–9 (2019)
18 doi:10.1038/s41598-019-47634-2.
- 19 82. Madsen, K. & H. B. Nielsen, O. T. Methods for non-linear least squares
20 problems. at (2004).
- 21 83. Barbosa, L. L., Kock, F. V. C., Almeida, V. M. D. L., Menezes, S. M. C. & Castro,
22 E. V. R. Low-field nuclear magnetic resonance for petroleum distillate
23 characterization. *Fuel Process. Technol.* **138**, 202–209 (2015).
- 24 84. P. T. Boggs and J. E. Rogers. “Orthogonal Distance Regression,” in “Statistical
25 analysis of measurement error models and applications: proceedings of the
26 AMS-IMS-SIAM joint summer research conference held June 10-16, 1989,”.
27 *Contemp. Math.* **112**, 186 (1990).
- 28 85. Hirasaki, G. J., Lo, S. W. & Zhang, Y. NMR properties of petroleum reservoir
29 fluids. in *Magnetic Resonance Imaging* vol. 21 269–277 (2003).
- 30 86. Markovic, S. *et al.* In-situ heavy oil viscosity prediction at high temperatures
31 using low-field NMR relaxometry and nonlinear least squares. *Fuel* **260**,

- 1 (2020).
- 2 87. Li, H. & Misra, S. Prediction of subsurface NMR T2 distribution from
3 formation-mineral composition using variational autoencoder. in *SEG*
4 *Technical Program Expanded Abstracts 2017* 3350–3354 (2017).
5 doi:10.1190/segam2017-17798488.1.
- 6 88. Li, H., Misra, S. & He, J. Neural network modeling of in situ fluid-filled pore
7 size distributions in subsurface shale reservoirs under data constraints.
8 *Neural Comput. Appl.* **32**, 3873–3885 (2020).
- 9 89. Bühlmann, P. & Hothorn, T. Boosting algorithms: Regularization, prediction
10 and model fitting. *Stat. Sci.* **22**, 477–505 (2007).
- 11 90. Cortes, C., Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297
12 (1995).
- 13 91. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*.
- 14 92. Ho, T. K. Random decision forests. in *Proceedings of 3rd International*
15 *Conference on Document Analysis and Recognition* vol. 1 278–282 vol.1
16 (1995).
- 17 93. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric
18 regression. *Am. Stat.* **46**, 175–185 (1992).
- 19 94. Cohen, P., West, S. G. & Aiken, L. S. *Applied Multiple Regression/Correlation*
20 *Analysis for the Behavioral Sciences*. (Taylor & Francis, 2014).
- 21 95. Zheng, A. & Casari, A. *Feature engineering for machine learning*. O'Reilly
22 *Media* (O'Reilly Media, Inc., 2018). doi:10.13140/RG.2.1.3564.3367.
- 23 96. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**,
24 367–378 (2002).
- 25 97. Schapire, R. E. *Boosting: Foundations and Algorithms*. *Kybernetes* vol. 42
26 (2013).
- 27 98. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn.*
28 *Res.* **12**, 2825–2830 (2011).
- 29 99. Ridgeway, G. Generalized Boosted Models : A guide to the gbm package. 1–
30 15 (2019).
- 31 100. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat.*

- 1 *Comput.* **14**, 199–222 (2004).
- 2 101. Dodge, Y. LAD Regression for Detecting Outliers in Response and
3 Explanatory Variables. *J. Multivar. Anal.* **61**, 144–158 (1997).
- 4 102. Askarova, A. *et al.* Thermal enhanced oil recovery in deep heavy oil
5 carbonates: Experimental and numerical study on a hot water injection
6 performance. *J. Pet. Sci. Eng.* **194**, 107456 (2020).
- 7 103. Nicot, B., Fleury, M. & Leblond, J. Improvement Of Viscosity Prediction Using
8 NMR Relaxation. *SPWLA 48th Annual Logging Symposium 7* at
9 <https://doi.org/> (2007).
- 10 104. Misra, S. & Wu, Y. Chapter 10 - Machine learning assisted segmentation of
11 scanning electron microscopy images of organic-rich shales with feature
12 extraction and feature ranking. in (eds. Misra, S., Li, H. & He, J. B. T.-M. L. for
13 S. C.) 289–314 (Gulf Professional Publishing, 2020).
14 doi:<https://doi.org/10.1016/B978-0-12-817736-5.00010-7>.
- 15 105. Valentín, M. B. *et al.* Estimation of permeability and effective porosity logs
16 using deep autoencoders in borehole image logs from the brazilian pre-salt
17 carbonate. *J. Pet. Sci. Eng.* **170**, 315–330 (2018).
- 18 106. Wherity, S., Sidley, T., Cowling, M., Ismayilov, A. & Olie, M. Observation and
19 Monitoring Well : in Situ Window to Assess Recovery Introduction to the
20 Halfdan Field. (2014).
- 21 107. Willmott, C. J. & Matsuura, K. Advantages of the mean absolute error (MAE)
22 over the root mean square error (RMSE) in assessing average model
23 performance. *Clim. Res.* **30**, 79–82 (2005).
- 24 108. Chollet, F. & others. Keras. at (2015).
- 25 109. de Myttenaere, A., Golden, B., Le Grand, B. & Rossi, F. Mean Absolute
26 Percentage Error for regression models. *Neurocomputing* **192**, 38–48
27 (2016).
- 28 110. Theil, H., Cramer, J. S., Moerman, H. & Russchen, A. *Economic Forecasts and*
29 *Policy.* (North-Holland Publishing Company, 1961).
- 30 111. Biancolini, M. E. *Radial Basis Functions for Engineering Applications.*
31 (Springer US, 2017).

- 1 112. Cherkassky, V. & Ma, Y. Selection of meta-parameters for support vector
2 regression. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell.*
3 *Lect. Notes Bioinformatics)* **2415 LNCS**, 687–693 (2002).
- 4 113. Sarle, W.S. *Neural Network FAQ, part 1 of 7: Introduction, periodic posting to*
5 *the Usenet newsgroup comp.ai.neural-nets.* url:
6 <ftp://ftp.sas.com/pub/neural/FAQ.html> (1997).
- 7 114. Riazoshams, H., Midi, H. & Ghilagaber, G. Outlier Detection in Nonlinear
8 Regression. *Robust Nonlinear Regression* 107–141 at
9 <https://doi.org/10.1002/9781119010463.ch6> (2018).
- 10 115. Sohn, B. Y. & Kim, G. B. Detection of outliers in weighted least squares
11 regression. *Korean J. Comput. Appl. Math.* **4**, 441–452 (1997).
- 12 116. Jones, E., Oliphant, T., Peterson, P. SciPy: Open source scientific tools for
13 Python. Retrieved from 'http://www.scipy.org/'. <http://www.scipy.org/>
14 (2001).
- 15 117. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in
16 *Proceedings of the ACM SIGKDD International Conference on Knowledge*
17 *Discovery and Data Mining* vols 13-17-Aug 785–794 (2016).
- 18 118. Nocedal, J. & Wright, S. J. Numerical optimization. *Springer Ser. Oper. Res.*
19 *Financ. Eng.* 1–664 (2006) doi:10.1201/b19115-11.
- 20 119. Tikhonov, A. N. & Arsenin, V. Y. *Solutions of ill-posed problems.* (V. H. Winston
21 & Sons, 1977).
- 22 120. Prammer, M. G., Drack, E. D., Bouton, J. C. & Gardner, J. S. Measurements of
23 clay-bound water and total porosity by magnetic resonance logging. *Log*
24 *Anal.* **37**, 61–69 (1996).
- 25 121. Ross, B. C. Mutual information between discrete and continuous data sets.
26 *PLoS One* **9**, (2014).
- 27 122. Chai, T. & Draxler, R. R. Root mean square error (RMSE) or mean absolute
28 error (MAE)? -Arguments against avoiding RMSE in the literature. *Geosci.*
29 *Model Dev.* **7**, 1247–1250 (2014).
- 30 123. Chen, J. & Bryan, J. In situ bitumen viscosity and saturation estimation from
31 core log integration for Canadian oil sands. *Soc. Pet. Eng. - SPE Heavy Oil*

- 1 *Conf. Canada 2013* **3**, 1686–1693 (2013).
- 2 124. Jin, G., Xie, R., Liu, M. & Guo, J. Petrophysical Parameter Calculation Based on
3 NMR Echo Data in Tight Sandstone. *IEEE Trans. Geosci. Remote Sens.* **57**,
4 5618–5625 (2019).
- 5 125. Gómez-Ríos, A., Luengo, J. & Herrera, F. A study on the noise label influence
6 in boosting algorithms: Adaboost, GBM and XGBoost. *Lect. Notes Comput.*
7 *Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*
8 **10334 LNCS**, 268–280 (2017).
- 9