# Clustering NMR: Machine learning assistive rapid two-dimensional relaxometry mapping

Weng Kung Peng[1]

[1]Affiliation not available

May 14, 2020

## Abstract

Low-field nuclear magnetic resonance (NMR) relaxometry is an attractive approach for point-of-care testing medical diagnosis, industrial food science, and in situ oil-gas exploration. However, one of the problems is the inherently long relaxation time of the (liquid) sample (and hence low signal-to-noise ratio) which causes unnecessarily long repetition time. In this work, a new methodology is presented for a rapid and accurate object classification using NMR relaxometry with the aid of machine learning techniques. It is demonstrated that the sensitivity and specificity of the classification are substantially improved with a higher order of (pseudo)-dimensionality (e.g., 2D or multidimensional). This new methodology (the so-called Clustering NMR) may be extremely useful for rapid and accurate object classification (in less than a minute) using the low-field NMR.

## Introduction

High resolution nuclear magnetic resonance (NMR) spectroscopy is a powerful and attractive approach in biochemistry (e.g., protein analysis[1], metabolomics[2–4]) and inorganic chemistry[5]. In the recent years however, with the rapid advances in NMR engineering (e.g., IC-based spectrometer[6–12], microfluidic-based chip[13–17], artificial intelligence[18,19]) utilizing small foot-print permanent magnet, the time-domain NMR instrumentations have seen a myriad of interesting applications from point-of-care testing (PoCT) medical diagnosis[7,20–23], industrial food science [24,25], and *in-situ* oil-gas exploration[26,27].

Biochemical information is typically detected and encoded in the frequency domain (´chemical shift´) in the high-field NMR. In contra, the low-field NMR, information is encoded in the time domain, with the dephasing of the spin-spin relaxation ($T_2$ relaxation) of the water-proton of the observed sample used as diagnostic criterion[20,21]. Time domain NMR however, suffers from inherently long relaxation time of the (liquid) sample, (and hence low signal-to-noise ratio (SNR)) causes unnecessarily long repetition time[28,29]. Furthermore, the $T_2$-relaxation measurement (in one-dimensional) which is frequently reported in NMR relaxometry experiments has limited number of dimensionality (e.g., healthy/non-healthy)[20,21].

In this work, a new class of methodology is presented for rapid and accurate object classification using PoCT NMR relaxometry with the aid of machine learning (Fig. 1). It is demonstrated (using various edible oils as proof-of-concept) that the sensitivity (´true positive rate´) and specificity (´true negative rate´) of the classification is substantially improved using higher order of (pseudo)-dimensionality (e.g., 2D or multidimensional). Further, by leveraging on the advances in machine learning techniques (e.g., pre-trained dataset) the detection time was sped up (in minutes) as compared to conventional 2D or multidimensional NMR (>hours), without resorting to using Ultrafast NMR[30]. This methodology (termed as Clustering NMR) is extremely useful for rapid and accurate classification of objects (in less than a minute) using the low-field NMR at point-of-need.

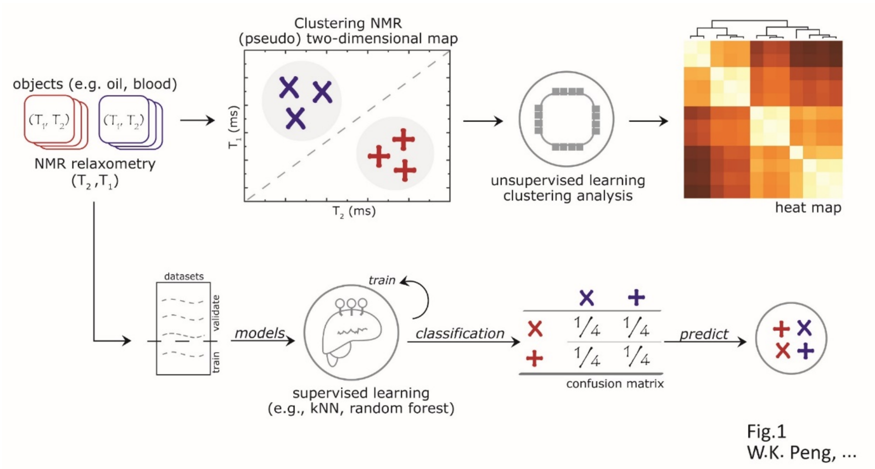


Figure 1: Conceptualization of the (pseudo) two-dimensional mapping using the Clustering NMR method proposed in this work. A pair of ($T_1$, $T_2$) relaxation time for each objects (e.g., edible oils, blood) were measured using micro NMR relaxometry system. A (pseudo) two-dimensional map is constructed with ($T_2$, $T_1$) relaxation time with a (X, Y) scatter plot (Fig. 2c), where the object clustering became obvious in comparison to its´ one-dimensional counterparts (i.e, $T_1$relaxation or $T_2$ relaxation). The efficacies of Clustering NMR method were validated using both the supervised and unsupervised learning methods. The relationship between each objects is established using (unsupervised) clustering analysis methods (e.g., tree classfication, hierarchical clustering) and its´ quantitative linkage (e.g., inter/intra cluster similarity) of each objects which is depicted on a dendogram with a heat map (details in Supp. Figs. 2-3). Supervised learning techniques (e.g., kNN, random forest, logistic regression) were used to train the classification of objects and the best trained model is subsequently chosen to predict the object classification. (e.g., oils content, infection/non-infection).

## Methods

NMR measurement and detection. The relaxometry measurements ($T_1$ relaxation, $T_2$ relaxation) were carried out on four group edible oils (i.e., peanut, olive, sunflower, corn) labelled as (A, B, C, D), respectively (Figs. 2a-b). In order to avoid bias, more than one different manufacturers were used for the same oil (with the exception of corn oil) and the detail on fat compositions were presented in Supp. Fig. 1. (A, A´, B, B´, C, C´, C´´, D) were the variants of the same oil from various manufacturers. The manufacturer labelling indicated 100% of oil contents (no mixture of oils). The edible oils were cooking oils bought locally in Braga, Portugal. No further alteration was made before the NMR measurements.

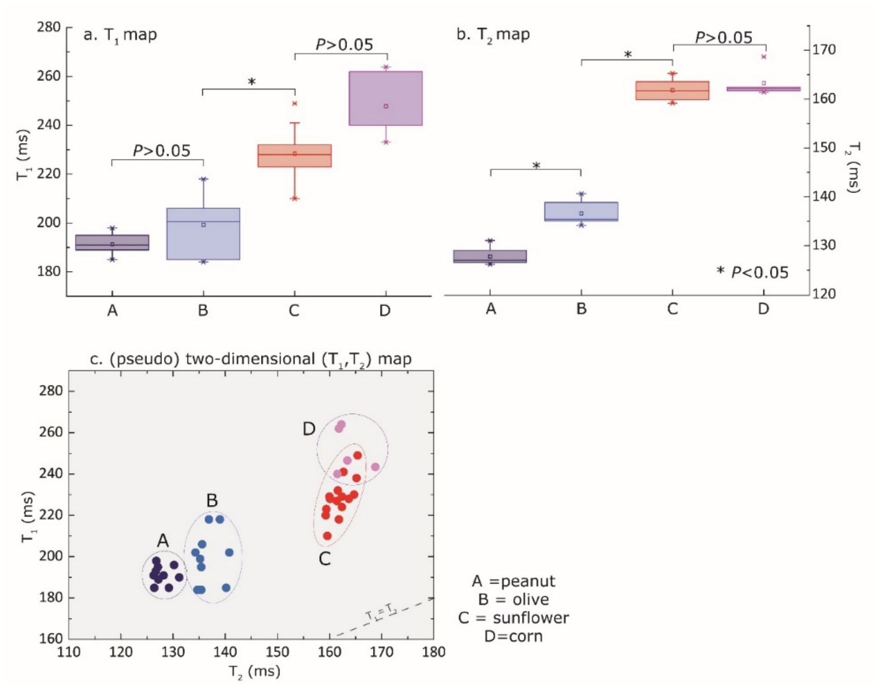


Fig. 2
W.K. Peng, ...

Figure 2: NMR measurements and (pseudo) two-dimensional mapping with Clustering NMR approach. $T_1$ and $T_2$relaxation times were carried on various edible oils (i.e., peanut, olive, sunflower, corn), with the label of (A, B, C, D) respectively using micro NMR relaxometry. One-dimensional mapping with (a) $T_1$ relaxation time, (b) $T_2$ relaxation time, and (c) (pseudo) two-dimensional mapping using a pair of ($T_1$, $T_2$) relaxation times. NMR measurements were carried out on each edible oils in quintuplicate manner with a total of 40 points (datasets). The clustering circles were drawn for eye-balling purposes. Details of the oils (e.g., manufacturers, fat compositions) were presented in Supp. Fig. 1. The box plots represent 25% and 75% quantile of the entire measurements. The diaganol line ($T_1$=$T_2$) represents the border limit where it is physically non-measurable. Two tailed Student´s T-test was used to calculate the $P$ -value.

NMR measurements were carried out (in single blinded manner) on each oils in quintuplicate manner (i.e., five repeated times) with a total of 40 points for all the samples. Details on NMR parameter are reported in Supplementary Methods. Clustering NMR method uses a pair of ($T_2$, $T_1$) relaxation time for each objects (e.g., edible oils, blood) to construct a (pseudo) two-dimensional map (Fig. 2c). The pseudo two-dimensional map can be used a referencing map (control).

Machine learning learning algorithm and workflows. Using a statistical programming languages (e.g., *R* or Orange 3.1.2), the raw datasets can be processed using supervised and unsupervised learning techniques. The machine learning algorithms were written and runs on a personal laptop (Intel Core Pentium i7 CPU @ 2.70GHz, 8.00 GB RAM). Once the model in machine learning is built, all the tasks run simultaneously and completes typically in less than 1 minute.

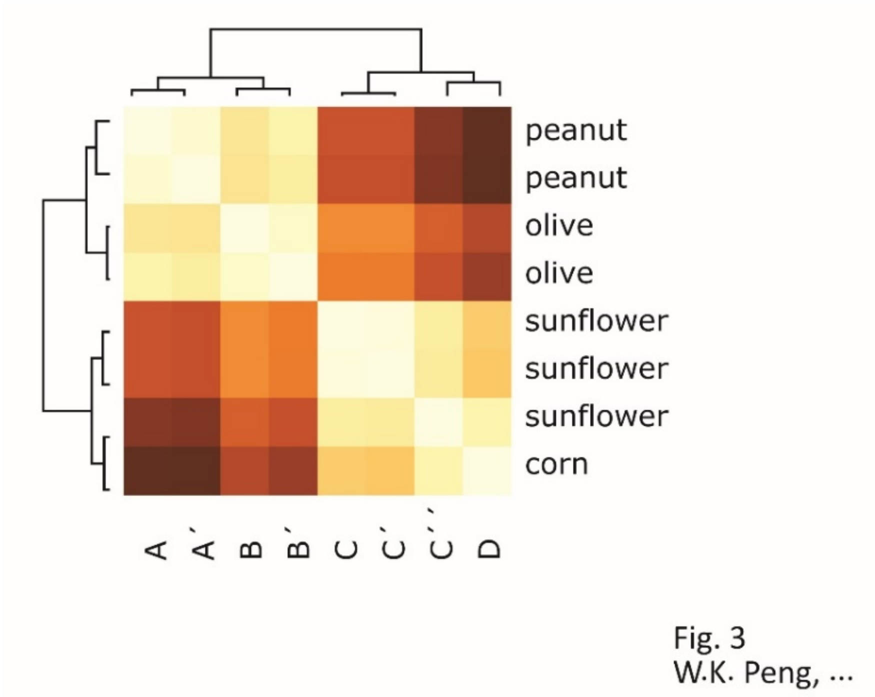


Figure 3: Unsupervised learning techniques (e.g., hierarchical clustering, tree classification) can be used for clustering analysis. This hierarchical clustering was constructed based on Euclidean distance (between T$_1$ relaxation and T$_2$relaxation) and its´ quantitative linkages (e.g., inter/intra cluster similarity) shown in a heat map. (A, A´, B, B´, C, C´, C´´, D) were the variants of the same oil content taken from different manufacturers. Tree classification method is shown for comparison (Supp. Fig. 2).

Using unsupervised learning, the relationship between each objects were rapidly constructed using clustering analysis (e.g., tree classification, hierarchical clustering) and its´ quantitative linkages (e.g., inter/intra cluster similarity) were shown on a dendogram and a heat map (Fig. 3). Supervised learning models (i.e., neural network, kNN, logistic regression, naïve Bayes, and random forest) can be used to train the datasets and the best model with the highest accuracy can be chosen to predict the object classification (e.g., oil classification, infection/non-infection) using pre-trained datasets (Fig. 4 and Table 1).

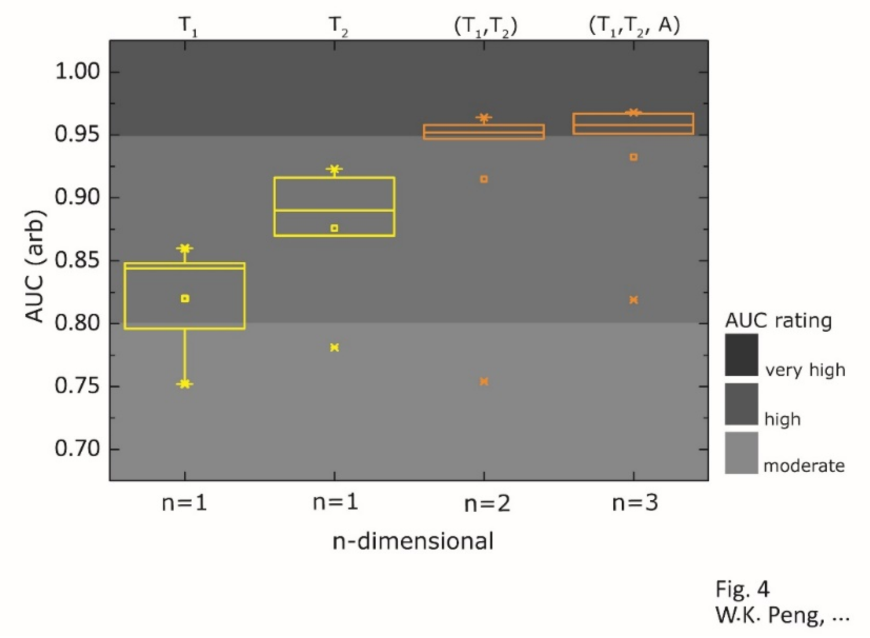


Figure 4: The Area Under Curve (AUC) plot as evaluated by Receiver Operating Characteristic (ROC) of various supervised models (i.e., kNN, random forest, neural network, naïve bayes, logistic regression) evaluated using target variables of (a) one-dimensional ($T_1$-relaxation), (b) one-dimensional ($T_2$-relaxation), (c) two-dimensional ($T_1$-relaxation, $T_2$-relaxation), and (d) three-dimensional ($T_1$-relaxation, $T_2$-relaxation, A-ratio) using leave-one-out training method. Other details can be found in Table 1.

## Hosted file

`image5.emf` available at https://authorea.com/users/322019/articles/451126-clustering-nmr-machine-learning-assistive-rapid-two-dimensional-relaxometry-mapping

Table 1: The sensitivity and specificity of the various supervised models evaluated using target variable/s of (a) one-dimensional ($T_1$-relaxation), (b) one-dimensional ($T_2$-relaxation), (c) two-dimensional ($T_1$-relaxation, $T_2$-relaxation), and (d) three-dimensional ($T_1$-relaxation, $T_2$-relaxation, A-ratio) from leave-one-out training method. The synonyms used were: $T_1$-relaxation/$T_2$-relaxation (A-ratio), area under the curve (AUC), classification accuracy (CA), F1 score – the balance between precision and recall, Precision – how many selected items were relevant, Recall - how many relevant items are selected. The training method using cross validation of k=5 was also evaluated for comparison (Supp. Fig. 4).

Results

Each edible oils (i.e., peanut, olive, sunflower, corn) were assigned to its´ respective label (A, B, C, D) following the blinded NMR measurements. As depicted in the one-dimensional map, each of oil contents has a specific $T_1$ relaxation and $T_2$ relaxation characteristic reading (Figs. 2a-b). The means for $T_1$ relaxation time were (191.3, 199.3, 228.4, 247.8) ms and means for $T_2$ relaxation time were (127.9, 136.8, 162, 163) ms for (A, B, C, D), respectively.

The spread of the readings were, however, substantially large making objects (A and B) and objects (C and D) inseparable in the $T_1$ relaxation dimension ($P$ >0.05) (Fig. 2a). Further in the $T_2$ relaxation dimension, the objects (C and D) were also inseparable (Fig. 2b). The undesirable spread causes (similarly to spectral) cluster overlapping and hence making classification difficult (if not impossible). One straightforward solution is to increase the SNR (e.g., increasing the number scans) or/and increase the number of samplings, which unfortunately, came at the expenses of acquisition time. In addition, the relaxation time of liquid sample is inherently long. On the other hand, using the Clustering NMR method (as proposed in this work), one can leverages on the combined characteristic of ($T_1$, $T_2$) relaxation times of the oil contents. It forms (visibly) unique and specific cluster based on the oil contents (´molecular fingerprint´) in (pseudo) two-dimensional map (Fig. 2c). With the minor exception of corn oil (which partially overlapped with sunflower oils), which could be due to possible adulteration or factory processes. Upon further investigation, we found that this artifact can be removed with higher SNR.

Interestingly, unsupervised techniques based clustering analysis (e.g., hierarchical clustering (HC), tree-based classification, and k-means) can be performed in conveniently using (open-source code) user friendly third party software (e.g., *R* , or Orange 3.1.2). A front-end statistical programming language allows the clustering analysis (once compiled), can be executed in the next occasion. The HC analysis successfully separated the (peanut and olive) cluster from the (sunflower and corn) cluster, and subsequently split between themselves (Fig. 3). The HC was constructed based on Euclidean distance (between $T_1$ relaxation and $T_2$ relaxation) and its´ quantitative linkages (e.g., inter/intra cluster similarity) were shown in a heat map. The HC methods also confirmed the oil variants (A, A´, B, B´, C, C´, C´´, D) based on its´ respective manufacturer. Similarly, the Chemometric approach[31] based on fat compositions (Supp. Fig. 2) and tree-based classification technique based on the $T_1$-relaxation cutoff and $T_2$-relaxation cutoff criterion (Supp. Fig. 3) appear to be in good agreement (qualitatively) with the HC classification using Euclidian distance of $T_1$ relaxation and $T_2$ relaxation obtained with NMR experimentally. It is worth noting, however, that the figures (i.e., fat compositions) given by the manufacturers are for references (and not for scientific) purposes. The clustering analysis models despite using various differential clustering criterions (e.g., Euclidean distance, fat compositions, relaxation cutoff) were in agreement with our observation (Clustering NMR, Fig. 2c). This demonstrated the robustness of Clustering NMR method, which can be validated using unsupervised techniques.

In order to evaluate the classification accuracy on the quantitative basis, various supervised learning models (i.e., kNN, random forest, neural network, naïve Bayes, and logistic regression) were used to train, validate and predict the datasets. The Area Under Curve (AUC) as evaluated with Receiver Operating Characteristic (ROC) were on average (0.820, 0.876, 0.915, 0.933) with (one-dimensional ($T_1$-relaxation), one-dimensional ($T_2$-relaxation), two-dimensional ($T_1$-relaxation, $T_2$-relaxation), and three-dimensional ($T_1$-relaxation, $T_2$-relaxation, A-ratio)), respectively, using the leave-one-out training method (Fig. 4). A-ratio is the ratio between $T_1$-relaxation and $T_2$ -relaxation. Similar conclusions were observed using cross validation method (e.g., k=5) (details in Supp. Table 1). This confirmed that the sensitivity and specificity of the proposed Clustering NMR method has substantially improved at the higher order of (pseudo)-dimensionality (e.g., 2D or multidimensional) over low dimensionality (e.g., n=1). With the (minor) exception of logistic regression, all the supervised models performed reasonably well (AUC>0.80) (Table 1). Furthermore, all the machine learning tasks run simultaneously and computational time taken were typically in less than 1 minute (in this work).

## Discussion

The proposed Clustering NMR method works on the rational that accumulative characteristic of each dimensionality would forms a specific and unique signature (´molecular fingerprint´). This is the same concept which borrowed from the data mining[32]. Fortunately, the characteristic of ($T_1$, $T_2$) relaxation times in the relaxometry is rather specific and prominent, and as the results suggested, an optimal n=2 to 3 of dimensionality are essential to attain a high AUC (Fig. 4)[33]. With the recent advances in machine learning, however, its´ becoming computationally cheaper (e.g., shorter analysis time) to calculate a big dataset. The computational time reported in this analysis (less than one minute) much shorter than a conventional two- or multidimensional NMR (>hours), without resorting to the use of Ultrafast NMR.

Two- or multidimensional relaxometry experiments (e.g., $T_1$-$T_2$ correlation spectroscopy), however, may provides much more information (e.g., cross peaks) but are far more time consuming than that of Clustering NMR method. One way to speed up acquisition time is to employ the use of gradient fields (e.g., Ultrafast NMR[30], continuous spatial encoding[34]) which require modification to the radio-frequency probe. Machine learning in the form of dimension reductionist (e.g., principal component analysis (PCA), partial least squares (PLS)) have also been used to reduce the dimensionality in multidimensional spectroscopy (e.g., NMR metabolomics[19,35,36]). A recent deep learning assistive NMR spectroscopy[18], which signals reconstructing were demonstrated. We summarized and compared Clustering NMR method with the state-of-the-art methodologies in a SWOT-like analysis (Table 2).

In conclusion, this proposed methodology, termed as Clustering NMR is extremely powerful for rapid and accurate classification of objects using the low-field NMR. This methodology is highly distruptive to the low-field NMR applications, in particularly, the recent reported NMR-based PoCT medical diagnostic. These include the immuno-magnetic labelled detection (e.g., tumour cells[14,20], tuberculosis[37] and magneto-DNA detection of bacteria[38]) and the label-free detection of various pathological states (e.g., blood oxygenation[15]/oxidation level[10] and malaria screening[21,22,39]). Interestingly, with the recent advances on machine learning technique, it has become remarkably efficient that a large data run in almost in ´real-time mode´, which open-up opportunity to combine real-time NMR (or MRI) with machine learning simultaneously.

(1675 words)

Table 2: State-of-the-art (with/without) machine learning assistive NMR works in comparison to the current work (Clustering NMR).

| Authors/Year | NMR | Applications | n-Dimensional | Machine Learning | Advantageous | Weakness |
|---|---|---|---|---|---|---|
| Wishart (2008)[19], Karaman (2015)[36], Rocha (2018)[35] | spectroscopy | metabolomics | 2D | PCA/PLA | informative | slow |
| Frydmann (2014)[30] | spectroscopy | ultrafast NMR | 2D | no | rapid | gradient field |
| Qu (2019)[18] | spectroscopy | generic | n-dimensional | deep learning | speed up | information lose? |
| Haun (2010)[40], Haun[20] (2011), Liong (2013)[37], Peng (2014)[21], Neely (2016)[41], Robinson (2017)[42] | relaxometry | medical diagnosis | 1D | no | rapid, PoCT | low specificity and sensitivity, missing out cross peaks?! |
| Robinson (2014)[43], Ok (2016)[44] | relaxometry | food science | 1D | no | rapid, PoCT | |
| Santos (2016)[45], Zhu (2016)[38] | relaxometry | food science | 1D | PCA/PLA | rapid, PoCT | |

| Authors/Year | NMR | Applications | n-Dimensional | Machine Learning | Advantageous | Weakness |
|---|---|---|---|---|---|---|
| Xu (2014)[46], Rudszuck (2019)[27] | relaxometry | food science | 2D | no | PoCT, high specificity and sensitivity | slow |
| Hurlimann (2002)[47] | relaxometry | oil-gas exploration | 1D, 2D | no | *in situ*, high specificity and sensitivity | slow |
| Lewis (2013)[48] | relaxometry | oil-gas exploration | 2D | no |  | slow |
| Birdwell (2015)[49] | relaxometry | oil-gas exploration | 2D | PCA/PLA |  | slow |
| Clustering NMR (2020) | relaxometry-clustering | generic | (pseudo) n-dimensional | clustering analysis, supervised model | rapid, PoCT, high specificity and sensitivity | missing out cross peaks? |

Supplementary Methods

NMR setup and parameters. The $^1$H magnetic resonance measurements of edible oils were carried out at the resonance frequency of 21.67 MHz using a portable permanent magnet (Metrolab Instruments, Switzerland), $B_o$=0.5T using a benchtop-type console (Kea Magritek, New Zealand). A temperature controller was set to maintain the measurement chamber at 30°C. The $T_1$ relaxation and $T_2$ relaxation pulse sequences were set at standard inversion recovery, followed by Carr-Purcell-Meiboom-Gill (CPMG) train pulses, respectively. The experimental parameters used were echo time=200 µs, number of echoes=2000 and signal averaging=4. A recycle delay of 4s was set between each experiment to provide sufficiently long time to allow all the molecular spins to return to thermal equilibrium.

Statistical methods. Two tailed Student´s T-test was used to calculate the $P$ -value.

Data availability statement. The machine learning algorithms and raw NMR datasets are available upon reasonably request at weng.kung@inl.int.

Acknowledgement. This research was supported by the International Iberian Nanotechnology Laboratory (INL) Start Up and INL Seed Grant. W.K. Peng would like to personally thank internship student Daniele Paesani and Dr Murali Kumarasamy, who carried out the NMR measurements in blinded manner and for assisting other laboratory works.

Author Contribution**.** W.K.P conceived the original idea, wrote the paper, designed the protocols, and built the entire hardware setup.

Competing interests. The authors declare no competing interests.

Additional information. Supplementary information is available for this paper.

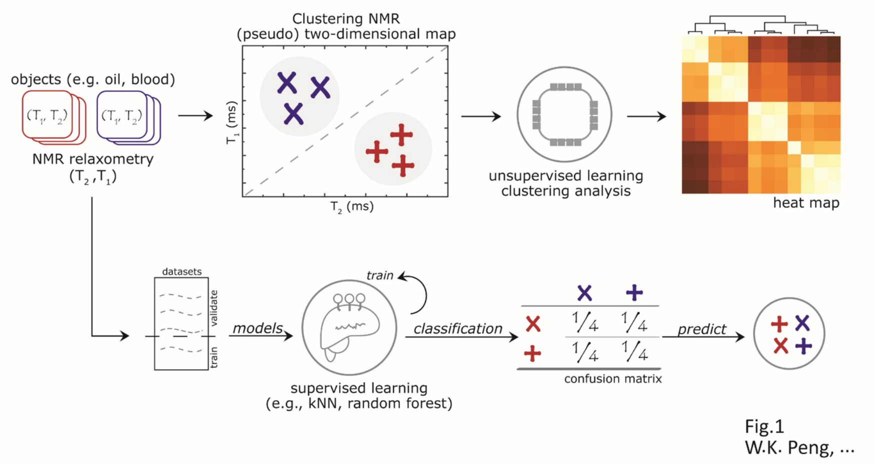


Figure 1: Conceptualization of the (pseudo) two-dimensional mapping using the Clustering NMR method proposed in this work. A pair of ($T_1$, $T_2$) relaxation time for each objects (e.g., edible oils, blood) were measured using micro NMR relaxometry system. A (pseudo) two-dimensional map is constructed with ($T_2$, $T_1$) relaxation time with a (X, Y) scatter plot (Fig. 2c), where the object clustering became obvious in comparison to its´ one-dimensional counterparts (i.e, $T_1$relaxation or $T_2$ relaxation). The efficacies of Clustering NMR method were validated using both the supervised and unsupervised learning methods. The relationship between each objects is established using (unsupervised) clustering analysis methods (e.g., tree classfication, hierarchical clustering) and its´ quantitative linkage (e.g., inter/intra cluster similarity) of each objects which is depicted on a dendogram with a heat map (details in Supp. Figs. 2-3). Supervised learning techniques (e.g., kNN, random forest, logistic regression) were used to train the classification of objects and the best trained model is subsequently chosen to predict the object classification. (e.g., oils content, infection/non-infection).

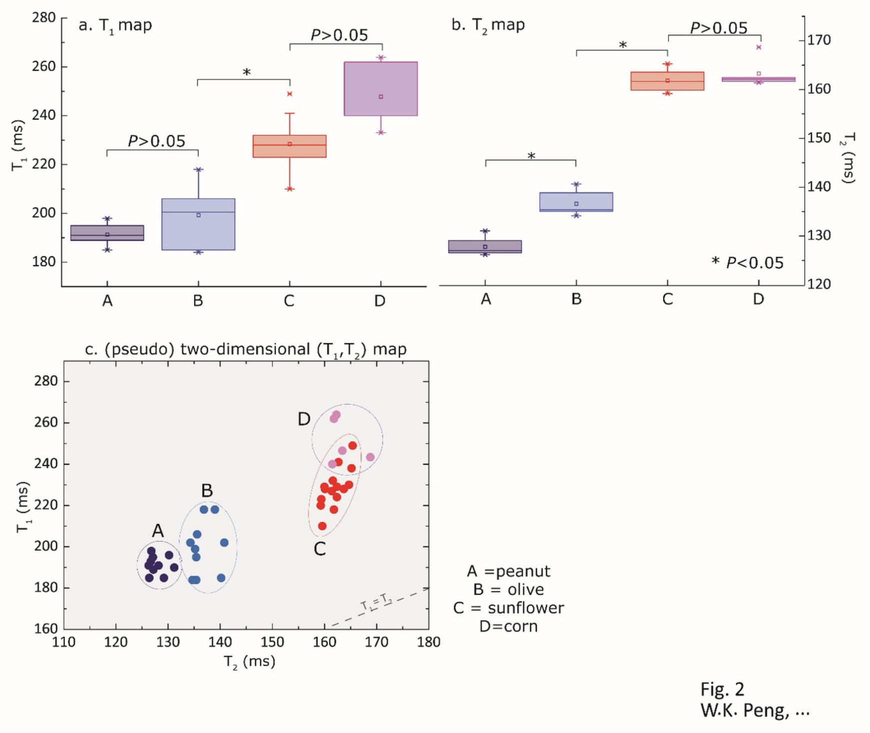


Figure 2: NMR measurements and (pseudo) two-dimensional mapping with Clustering NMR approach. $T_1$ and $T_2$relaxation times were carried on various edible oils (i.e., peanut, olive, sunflower, corn), with the label of (A, B, C, D) respectively using micro NMR relaxometry. One-dimensional mapping with (a) $T_1$ relaxation time, (b) $T_2$ relaxation time, and (c) (pseudo) two-dimensional mapping using a pair of ($T_1$, $T_2$) relaxation times. NMR measurements were carried out on each edible oils in quintuplicate manner with a total of 40 points (datasets). The clustering circles were drawn for eye-balling purposes. Details of the oils (e.g., manufacturers, fat compositions) were presented in Supp. Fig. 1. The box plots represent 25% and 75% quantile of the entire measurements. The diaganol line ($T_1$=$T_2$) represents the border limit where it is physically non-measurable. Two tailed Student´s T-test was used to calculate the $P$ -value.

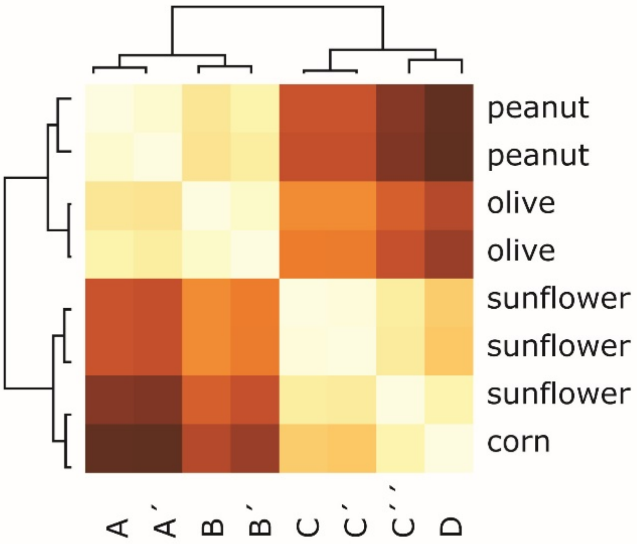


Fig. 3
W.K. Peng, ...

Figure 3: Unsupervised learning techniques (e.g., hierarchical clustering, tree classification) can be used for clustering analysis. This hierarchical clustering was constructed based on Euclidean distance (between $T_1$ relaxation and $T_2$relaxation) and its´ quantitative linkages (e.g., inter/intra cluster similarity) shown in a heat map. (A, A´, B, B´, C, C´, C´´, D) were the variants of the same oil content taken from different manufacturers. Tree classification method is shown for comparison (Supp. Fig. 2).

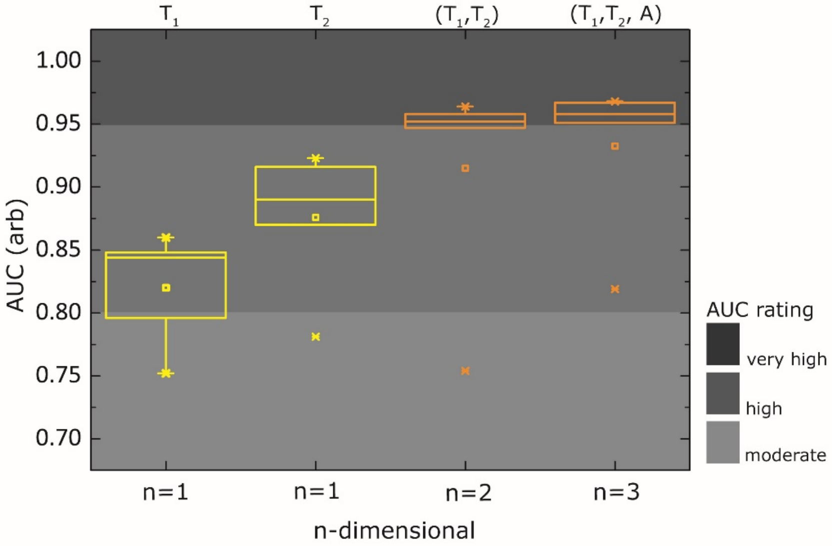
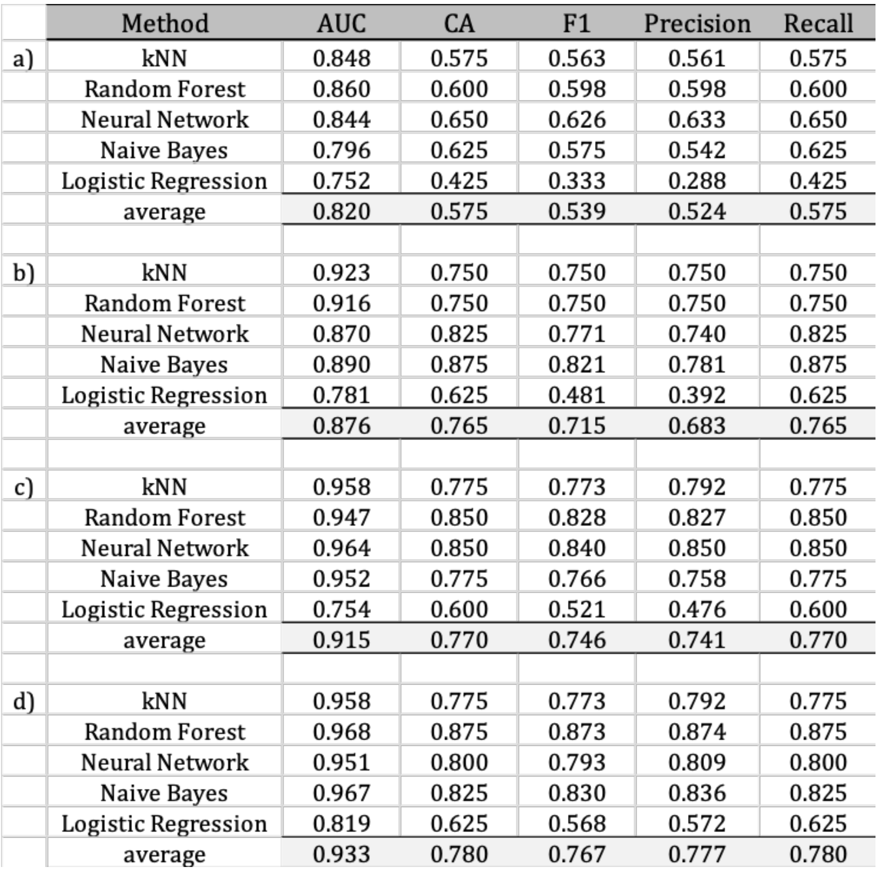


Fig. 4
W.K. Peng, ...

Figure 4: The Area Under Curve (AUC) plot as evaluated by Receiver Operating Characteristic (ROC)

of various supervised models (i.e., kNN, random forest, neural network, naïve bayes, logistic regression) evaluated using target variables of (a) one-dimensional (T$_1$-relaxation), (b) one-dimensional (T$_2$-relaxation), (c) two-dimensional (T$_1$-relaxation, T$_2$-relaxation), and (d) three-dimensional (T$_1$-relaxation, T$_2$-relaxation, A-ratio) using leave-one-out training method. Other details can be found in Table 1.



| | Method | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| a) | kNN | 0.848 | 0.575 | 0.563 | 0.561 | 0.575 |
| | Random Forest | 0.860 | 0.600 | 0.598 | 0.598 | 0.600 |
| | Neural Network | 0.844 | 0.650 | 0.626 | 0.633 | 0.650 |
| | Naive Bayes | 0.796 | 0.625 | 0.575 | 0.542 | 0.625 |
| | Logistic Regression | 0.752 | 0.425 | 0.333 | 0.288 | 0.425 |
| | average | 0.820 | 0.575 | 0.539 | 0.524 | 0.575 |
| | | | | | | |
| b) | kNN | 0.923 | 0.750 | 0.750 | 0.750 | 0.750 |
| | Random Forest | 0.916 | 0.750 | 0.750 | 0.750 | 0.750 |
| | Neural Network | 0.870 | 0.825 | 0.771 | 0.740 | 0.825 |
| | Naive Bayes | 0.890 | 0.875 | 0.821 | 0.781 | 0.875 |
| | Logistic Regression | 0.781 | 0.625 | 0.481 | 0.392 | 0.625 |
| | average | 0.876 | 0.765 | 0.715 | 0.683 | 0.765 |
| | | | | | | |
| c) | kNN | 0.958 | 0.775 | 0.773 | 0.792 | 0.775 |
| | Random Forest | 0.947 | 0.850 | 0.828 | 0.827 | 0.850 |
| | Neural Network | 0.964 | 0.850 | 0.840 | 0.850 | 0.850 |
| | Naive Bayes | 0.952 | 0.775 | 0.766 | 0.758 | 0.775 |
| | Logistic Regression | 0.754 | 0.600 | 0.521 | 0.476 | 0.600 |
| | average | 0.915 | 0.770 | 0.746 | 0.741 | 0.770 |
| | | | | | | |
| d) | kNN | 0.958 | 0.775 | 0.773 | 0.792 | 0.775 |
| | Random Forest | 0.968 | 0.875 | 0.873 | 0.874 | 0.875 |
| | Neural Network | 0.951 | 0.800 | 0.793 | 0.809 | 0.800 |
| | Naive Bayes | 0.967 | 0.825 | 0.830 | 0.836 | 0.825 |
| | Logistic Regression | 0.819 | 0.625 | 0.568 | 0.572 | 0.625 |
| | average | 0.933 | 0.780 | 0.767 | 0.777 | 0.780 |

Table 1: The sensitivity and specificity of the various supervised models evaluated using target variable/s of (a) one-dimensional (T$_1$-relaxation), (b) one-dimensional (T$_2$-relaxation), (c) two-dimensional (T$_1$-relaxation, T$_2$-relaxation), and (d) three-dimensional (T$_1$-relaxation, T$_2$-relaxation, A-ratio) from leave-one-out training method. The synonyms used were: T$_1$-relaxation/T$_2$-relaxation (A-ratio), area under the curve (AUC), classification accuracy (CA), F1 score – the balance between precision and recall, Precision – how many selected items were relevant, Recall - how many relevant items are selected. The training method using cross validation of k=5 was also evaluated for comparison (Supp. Fig. 4).

Table 2: State-of-the-art (with/without) machine learning assistive NMR works in comparison to the current work (Clustering NMR).

| Authors/Year | NMR | Applications | n-Dimensional | Machine Learning | Advantageous | Weakness |
|---|---|---|---|---|---|---|
| Wishart (2008)[19], Karaman (2015)[36], Rocha (2018)[35] | spectroscopy | metabolomics | 2D | PCA/PLA | informative | slow |
| Frydmann (2014)[30] | spectroscopy | ultrafast NMR | 2D | no | rapid | gradient field |
| Qu (2019)[18] | spectroscopy | generic | n-dimensional | deep learning | speed up | information lose? |
| Haun (2010)[40], Haun[20] (2011), Liong (2013)[37], Peng (2014)[21], Neely (2016)[41], Robinson (2017)[42] | relaxometry | medical diagnosis | 1D | no | rapid, PoCT | low specificity and sensitivity, missing out cross peaks?! |
| Robinson (2014)[43], Ok (2016)[44] | relaxometry | food science | 1D | no | rapid, PoCT | |
| Santos (2016)[45], Zhu (2016)[38] | relaxometry | food science | 1D | PCA/PLA | rapid, PoCT | |
| Xu (2014)[46], Rudszuck (2019)[27] | relaxometry | food science | 2D | no | PoCT, high specificity and sensitivity | slow |
| Hurlimann (2002)[47] | relaxometry | oil-gas exploration | 1D, 2D | no | *in situ*, high specificity and sensitivity | slow |
| Lewis (2013)[48] | relaxometry | oil-gas exploration | 2D | no | | slow |
| Birdwell (2015)[49] | relaxometry | oil-gas exploration | 2D | PCA/PLA | | slow |
| Clustering NMR (2020) | relaxometry-clustering | generic | (pseudo) n-dimensional | clustering analysis, supervised model | rapid, PoCT, high specificity and sensitivity | missing out cross peaks? |

## References

[1] M. Salzmann, K. Pervushin, G. Wider, H. Senn, K. Wuthrich,*Proceedings of the National Academy of Sciences* **1998** ,*95* , 13585–13590.

[2] M. E. Bollard, E. G. Stanley, J. C. Lindon, J. K. Nicholson, E. Holmes, *NMR in Biomedicine* **2005** , *18* , 143–162.

[3] I. F. Duarte, S. O. Diaz, A. M. Gil, *Journal of Pharmaceutical and Biomedical Analysis* **2014** , *93* , 17–26.

[4] M. R. Viant, B. G. Lyeth, M. G. Miller, R. F. Berman, *NMR in Biomedicine* **2005** , *18* , 507–516.

[5] L. Ronconi, P. J. Sadler, *Coordination Chemistry Reviews***2008** , *252* , 2239–2277.

[6] J. Anders, P. SanGiorgio, G. Boero, in *2009 IEEE Custom Integrated Circuits Conference* , IEEE, San Jose, CA, USA, **2009** , pp. 471–474.

[7] H. Lee, E. Sun, D. Ham, R. Weissleder, *Nature Medicine***2008** , *14* , 869–874.

[8] K. Takeda, *Journal of Magnetic Resonance* **2008** ,*192* , 218–229.

[9] N. Sun, T.-J. Yoon, H. Lee, W. Andress, R. Weissleder, D. Ham,*IEEE Journal of Solid-State Circuits* **2011** , *46* , 342–352.

[10] W. K. Peng, L. Chen, J. Han, *Review of Scientific Instruments* **2012** , 095115.

[11] A. Dupré, K.-M. Lei, P.-I. Mak, R. P. Martins, W. K. Peng,*Microelectronic Engineering* **2019** , *209* , 66–74.

[12] W. K. Peng, D. Paesani, *JPM* **2019** , *9* , 39.

[13] T. F. Kong, W. Ye, W. K. Peng, H. W. Hou, Marcos, P. R. Preiser, N. T. Nguyen, J. Han, *Scientific Reports* **2015** , 11425.

[14] C. M. Castro, A. A. Ghazani, J. Chung, H. Shao, D. Issadore, T.-J. Yoon, R. Weissleder, H. Lee, *Lab Chip* **2014** ,*14* , 14–23.

[15] T. F. Kong, W. K. Peng, T. D. Luong, N.-T. Nguyen, J. Han,*Lab Chip* **2012** , DOI 10.1039/C1LC20853E.

[16] G. Guan, P. C. Y. Chen, W. K. Peng, A. A. Bhagat, C. J. Ong, J. Han, *Journal of Micromechanics and Microengineering***2012** , *22* , 105037.

[17] T.-Y. Tu, Z. Wang, J. Bai, W. Sun, W. K. Peng, R. Y.-J. Huang, J.-P. Thiery, R. D. Kamm, *Adv. Healthcare Mater.* **2014** ,*3* , 609–616.

[18] X. Qu, Y. Huang, H. Lu, T. Qiu, D. Guo, T. Agback, V. Orekhov, Z. Chen, *Angew. Chem.* **2019** , ange.201908162.

[19] D. S. Wishart, *TrAC Trends in Analytical Chemistry***2008** , *27* , 228–237.

[20] J. B. Haun, C. M. Castro, R. Wang, V. M. Peterson, B. S. Marinelli, H. Lee, R. Weissleder, *Science Translational Medicine***2011** , *3* , 71ra16-71ra16.

[21] W. K. Peng, T. F. Kong, C. S. Ng, L. Chen, Y. Huang, A. A. S. Bhagat, N. T. Nguyen, P. R. Preiser, J. Han, *Nature Medicine***2014** , DOI 10.1038/nm.3622.

[22] J. Han, W. K. Peng, *Nature Medicine* **2015** ,*21* , 1387–1389.

[23] W. K. Peng, L. Chen, B. O. Boehm, J. Han, T. P. Loh,*bioRxiv* **2019** , 565325.

[24] B. P. Hills, in *Annual Reports on NMR Spectroscopy* , Elsevier, **2006** , pp. 177–230.

[25] E. Curti, E. Carini, M. F. Cobo, T. Bocher, E. Vittadini,*Food Chemistry* **2017** , *237* , 766–772.

[26] Y.-Q. Song, S. Ryu, P. N. Sen, *Nature* **2000** ,*406* , 178–181.

[27] T. Rudszuck, E. Förster, H. Nirschl, G. Guthausen, *Magn Reson Chem* **2019** , *57* , 777–793.

[28] W. K. Peng, K. Takeda, M. Kitagawa, *Chemical Physics Letters* **2006** , DOI 10.1016/j.cplett.2005.10.012.

[29] W. K. Peng, A. Samoson, M. Kitagawa, *Chemical Physics Letters* **2008** , DOI 10.1016/j.cplett.2008.06.027.

[30] P. Giraudeau, L. Frydman, *Annual Rev. Anal. Chem.***2014** , *7* , 129–161.

[31] A. Biancolillo, F. Marini, *Front. Chem.* **2018** ,*6* , 576.

[32] J. Han, M. Kamber, J. Pei, in *Data Mining* , Elsevier,**2012** , pp. 243–278.

[33] A. Zimek, E. Schubert, H.-P. Kriegel, *Statistical Analy Data Mining* **2012** , *5* , 363–387.

[34] S. Ahola, V. V. Zhivonitko, O. Mankinen, G. Zhang, A. M. Kantola, H.-Y. Chen, C. Hilty, I. V. Koptyug, V.-V. Telkki, *Nature Communications* **2015** , *6* , DOI 10.1038/ncomms9363.

[35] W. F. de C. Rocha, D. A. Sheen, D. W. Bearden, *Anal Bioanal Chem* **2018** , *410* , 6305–6319.

[36] İ. Karaman, N. P. Nørskov, C. C. Yde, M. S. Hedemann, K. E. Bach Knudsen, A. Kohler, *Metabolomics* **2015** , *11* , 367–379.

[37] M. Liong, A. N. Hoang, J. Chung, N. Gural, C. B. Ford, C. Min, R. R. Shah, R. Ahmad, M. Fernandez-Suarez, S. M. Fortune, et al.,*Nature Communications* **2013** , *4* , 55–65.

[38] H. J. Chung, C. M. Castro, H. Im, H. Lee, R. Weissleder,*Nature Nanotechnology* **2013** , *8* , 369–375.

[39] M. I. Veiga, W. K. Peng, *Malar J* **2020** ,*19* , 68.

[40] J. B. Haun, N. K. Devaraj, S. A. Hilderbrand, H. Lee, R. Weissleder, *Nature Nanotechnology* **2010** , *5* , 660–665.

[41] L. A. Neely, M. Audeh, N. A. Phung, M. Min, A. Suchocki, D. Plourde, M. Blanco, V. Demas, L. R. Skewis, T. Anagnostou, et al.,*Science Translational Medicine* **2013** , *5* , 182ra54-182ra54.

[42] M. D. Robinson, I. Mishra, S. Deodhar, V. Patel, K. V. Gordon, R. Vintimilla, K. Brown, L. Johnson, S. O'Bryant, D. P. Cistola,*Journal of Translational Medicine* **2017** , *15* , 258.

[43] M. D. Robinson, D. P. Cistola, *Biochemistry***2014** , *53* , 7515–7522.

[44] S. Ok, *Grasas y Aceites* **2017** , *68* , 173.

[45] P. M. Santos, E. R. Pereira-Filho, L. A. Colnago,*Microchemical Journal* **2016** , *124* , 15–19.

[46] Z. Xu, R. Morris, M. Bencsik, M. Newton, *Sensors***2014** , *14* , 2028–2035.

[47] M. D. Hürlimann, L. Burcaw, Y.-Q. Song, *Journal of Colloid and Interface Science* **2006** , *297* , 303–311.

[48] R. Lewis, P. Singer, T. Jiang, E. Rylander, S. Sinclair, R. H. Mclin, in *SPE Unconventional Resources Conference-USA* , Society Of Petroleum Engineers, The Woodlands, Texas, USA, **2013** .

[49] J. E. Birdwell, K. E. Washburn, *Energy Fuels***2015** , *29* , 2234–2243.

figures/fig1--concept/fig1--concept-eps-converted-to.pdf figures/fig2--pseudo-2D-NMR/fig2--pseudo-2D-NMR-eps-converted-to.pdf figures/fig3-unsupervised/fig3-unsupervised-eps-converted-to.pdf figures/fig4--AUC/fig4--AUC-eps-converted-to.pdf