URTeC: 5072

Eagle Ford and Bakken Productivity Prediction Using Soil Microbial Fingerprinting and Machine Learning

M.H.A.A. Zijp*[1], T. Mallinson[2], J. Zwaan[1], A.G. Chitu[1] , P. David[3]
1. Biodentify BV, 2. Aramco Services Company, 3. Wintershall Dea.

This paper was prepared for presentation at the Unconventional Resources Technology Conference held in Houston, Texas, USA, 26-28 July 2021.

Abstract

This paper presents the results of a pilot project to predict the production potential of locations in the Eagle Ford Shale and Bakken Shale formations. The approach uses DNA analysis of shallow soil samples to determine the present mix of species which are affected by vertical microseepage from hydrocarbon accumulations to the surface. This is assumed to be related to well productivity. 340 samples in the Eagle Ford and 200 samples in the Bakken Shale were taken above the highest and lowest producing wells. 200 samples were blinded, and the remaining 340 samples were used to train the machine learning model based on the corresponding generated DNA fingerprints. The trained model achieved a predictive accuracy of 85% on the 200 blinded samples. Investigating whether the trained model could be exported to other locations, showed that the Eagle Ford samples could be predicted with 72% accuracy using only Bakken data. Exporting to South America, the results were that two conventional fields in Argentina could be predicted with 83% accuracy using only US data, despite the vast distance between the two study areas. An additional goal of the study was to investigate whether knowledge on microbial characteristics found in the scientific community (disseminated in scientific publications) match conclusions obtained from the DNA analysis. For mining the scientific papers, we used Natural Language Processing (NLP) tools. Many of the biospeciemarkers selected by the trained models are linked to known species/genera that are metabolizing hydrocarbons for growth or finding them toxic.

Introduction

This article describes a technology to predict the productivity potential of hydrocarbons in the subsurface, using DNA fingerprints of the microbial ecosystem of shallow soil or seabed samples. To prove its validity, two well-known areas were selected; 540 samples were taken around high and low producing wells in the Bakken oil Shale (ND) and high and low producing wells in both the oil window and the gas window in the Eagle Ford Shale (TX). To calculate an objective measure for validation, 200 samples were blinded. The first goal was to use the DNA fingerprints of the remaining samples to build a predictive model using machine learning techniques, and with that to classify the blinded samples as belonging to highly productive or low productive wells. The sample location design and project execution were done in

cooperation with Aramco Americas (Aramco) and Wintershall Dea GmbH (Wintershall). The first publication featuring the technology was at URTeC 2017 (Stroet et al., 2017), this paper describes a more mature version of the technology and a set of applications funded by these two industry partners.

The second goal of the project was to test whether the SUA trained model can be exported to other locations. This was done by first investigating the Bakken Shale trained model to predict the Eagle Ford Shale samples, and second to use the total USA trained model (Eagle Ford and Bakken combined) to predict samples in central Argentina.

Theory

The technology described in this article is based on the response elicited in the shallow soil (<30cm) microbial population by hydrocarbons that have migrated from depth to the surface through vertical microseepage (Figure 1). The trace hydrocarbons released are otherwise generally undetectable, but the microbial population acts as an extremely sensitive indicator. A small number of specific microbes in the soil sample metabolize hydrocarbons whilst others find hydrocarbons toxic. The influence of microseepage delivered hydrocarbons is therefore found by comparing DNA fingerprints of soil samples from locations with known high production and alternately known low production or known hydrocarbon absence. Each sample consists of less than a cubic centimeter (less than half a teaspoon) of soil.

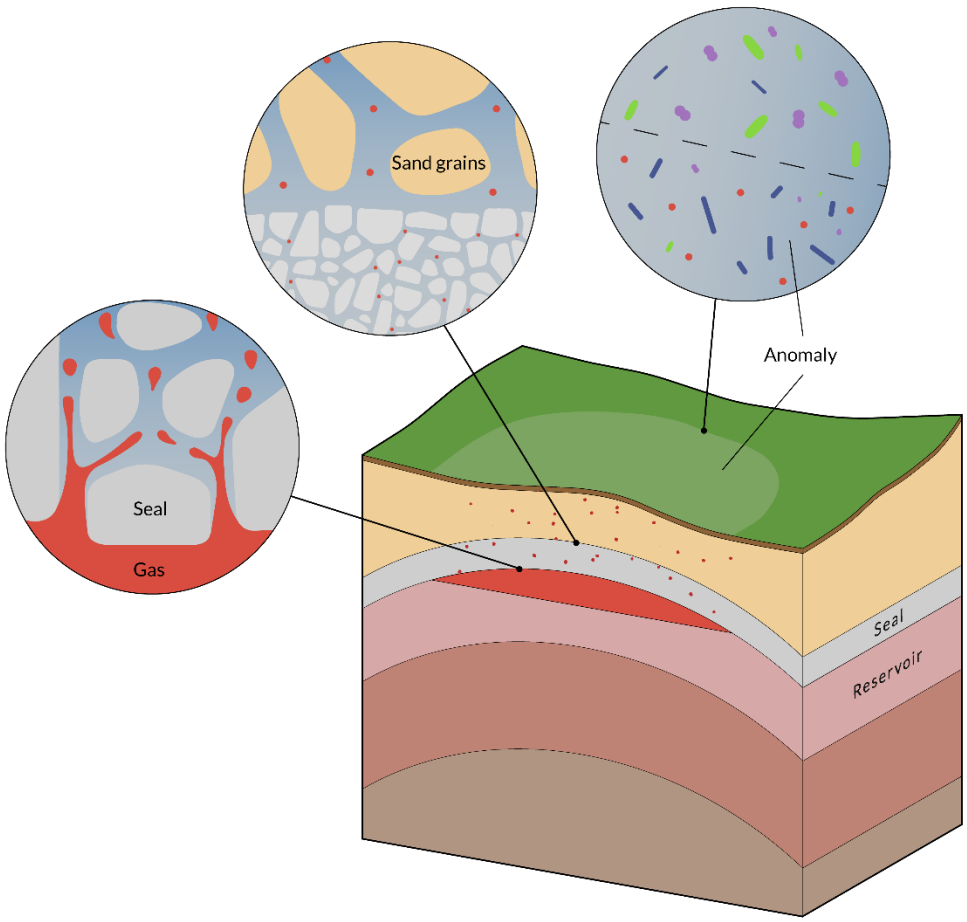

Figure 1: Microseepage from initiation at the boundary of a pressurized reservoir and its seal to the surface anomaly. Gas bubbles break off from viscous 'fingers' and are pushed vertically upward by buoyancy force (see also England, 1987). The local microbiome is affected by the increase in hydrocarbon.

Microseepage in this context refers to the vertical migration of small gas bubbles suspended in the interstitial water. These bubbles are pushed to the surface by the buoyancy force. A requirement for the forming of microbubbles is that the in-situ pressure (proportional to the hydrocarbon column) is higher than the combined forces that oppose the migration of the gas through the reservoir seal. As the buoyancy driven migration is mainly vertical (the slight horizontal local deviations are averaged out in the long run), there are sharp differences in hydrocarbon concentration expected to be seen at the surface, located above the field boundary. This contrasts with the more known phenomenon called macroseepage, which involves orders of magnitude higher volumes. As a result of the larger volumes involved, the horizontal component of the migration is in this case more pronounced. Macro and microseepage are related phenomena that take place simultaneously only at different spatial and time scales. Also, while macroseepage gives rise to easily visible differences, microseepage outcome is more subtle. In short, microseepage involves very small volumes, is fast, and its surface expression is expected to approximately overlay the underlying reservoir. The idea of the surface expression of underlying hydrocarbons accumulations is not new, with seepage, surface flora characterization, and other techniques being described at least as far back as the 1930s. This surface expression is essentially the basis of geochemical exploration. Due to the very low quantities provided through microseepage, direct measurement is not always possible. The microbiome analysis using machine learning provides a very sensitive indirect measurement means that can dramatically improve the quality of the analysis.

Methods

This work builds on the methodology shown in Figure 2, was presented in URTeC 2671117 and relies on two recent innovations:

- Next Generation DNA sequencing: microbial quantification of all species in the collected soil samples.
- Big Data: the species counted in the soil samples are correlated with presence or absence of hydrocarbons using machine learning.

Figure 2: Steps to produce a map with an estimation of productivity: 1) sampling in the Eagle Ford Shale (TX), 2) DNA fingerprinting, 3) training on labeled locations, and 4) estimating/mapping of productivity potential.

To determine the presence or absence of micro-seepage, microbial genetic material is extracted from shallow soil or seabed samples, producing tagged 16S rRNA gene sequences to be interpreted into prokaryotic phylogenies. The result of this analysis is the microbial diversity, which consists of hundreds of thousands of nucleotide-strings per sample. This is referred to as the 'DNA fingerprint' of the soil sample. The database with these DNA fingerprints is subsequently used as modeling input to build a tailored and localized machine learning model which predicts the productivity of an area based on classification as a high or low producer.

This machine learning approach must first manage the enormous amount of data produced from a set of samples, determine the relatively few microbial genera that provide a positive or negative signal in response to the presence of hydrocarbons, and then provide a robust link between the productivity data in the field and these positive and negative DNA signals.
The case study described in this article was performed as a proof of technology project for the sponsors. Therefore, all samples are taken at locations with known productivity data. For this, the 2-year cumulative production data from all wells in both the Eagle Ford and Bakken Shale was analyzed and filtered. In this way lab results can be linked to production values and used to train the machine learning models and validate estimation results. A set of blinded samples were kept separate during the model building and validation work for later confirmation of model accuracy by the project sponsors.

The "exportability" of our models was tested in two ways: first by using a model trained only on Bakken data to predict the productivity of the Eagle Ford samples, and second, by using the combined Bakken/Eagle Ford model to predict productivity we performed in the Neuquén Basin of central Argentina.

To broaden the scope of the investigation and further increase our confidence that microseepage is occurring and the microbiome is reacting on it (i. e., we are indeed finding causality and not only correlation,) we also analyzed 60,000+ scientific papers on microbes and their relation to hydrocarbons using Natural Language Processing (NLP) tools. This information was used to determine whether the specific microbes identified by the machine learning models were known in literature and described either as "hydrocarbon-philic" or "hydrocarbon-phobic".

Workflow description

1. Sampling design and execution
Three different sets of samples were taken from two plays (see Figure 3): 200 samples above oil-producing wells in the North Dakota Bakken and in the Eagle Ford of Texas 270 samples above gas wells and 70 samples above oil wells; one sample above each well. In each set, half the sample locations corresponded to wells classified as highest producers and the other half to lowest producers. The classification is based on a top 5% or bottom 5% 2-year cumulative production quantile within the play using data from shaleprofile.com, after normalizing for the year of first production and lateral length as a way of reducing the impact of non-geological variation. The wells were of various vintages with at least 2 years of production history. Further work was performed to ensure that a particular well location was not anomalous compared to its neighbors and that a physical sampling location was publicly accessible above the lateral.

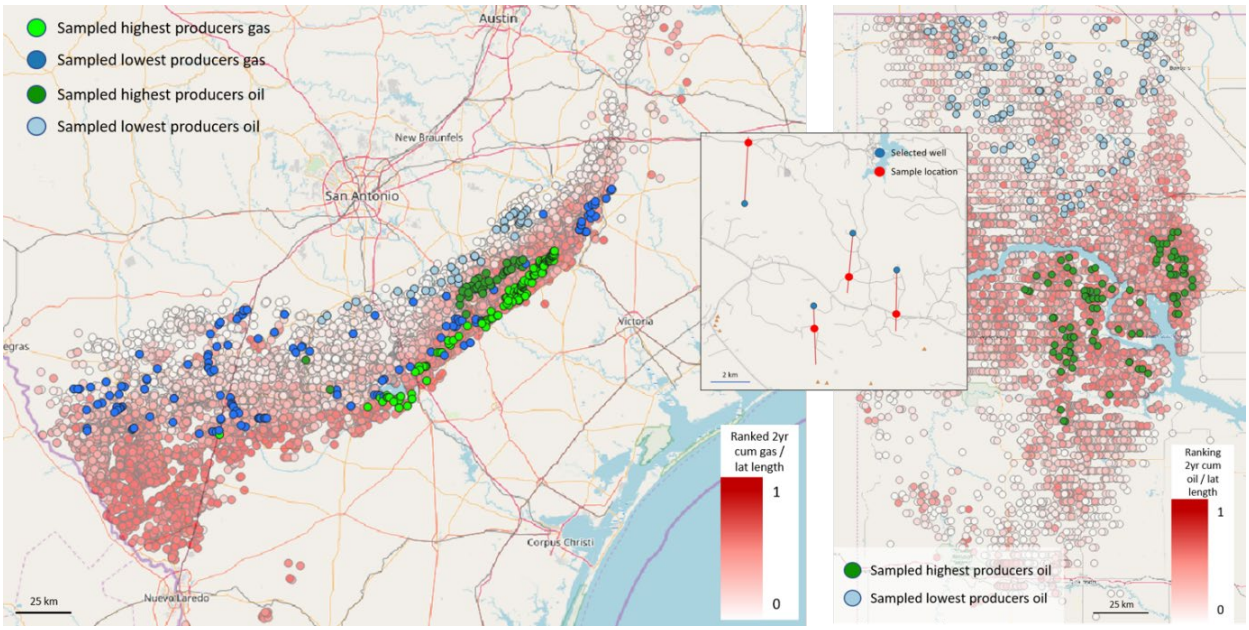

Figure 3: Highest (green) and lowest (blue) producing wells sampled in the Eagle Ford (left) and the Bakken (right) on top of all Bakken and Eagle Ford wells in white and red (higher production more intense red color); example of sample locations above selected laterals (inset)

2. DNA analysis to get DNA fingerprints

The DNA analysis used for this paper is derived from the workflow to determine a DNA fingerprint in life sciences (extraction, multiplication with PCR, and sequencing of DNA) and modified for this specific application. The methodology uses the 16S ribosomal RNA gene (hereafter 16S sequence), which is commonly used as a genetic fingerprint for bacteria due to its ubiquity and its slow evolution. This gene has 1500 base pairs consisting of interlaced variable and invariable regions used for classification and detection, respectively. A schematic representation of these DNA analysis steps using the Illumina MiSeq automatic sequencing platform is shown in Figure 4.

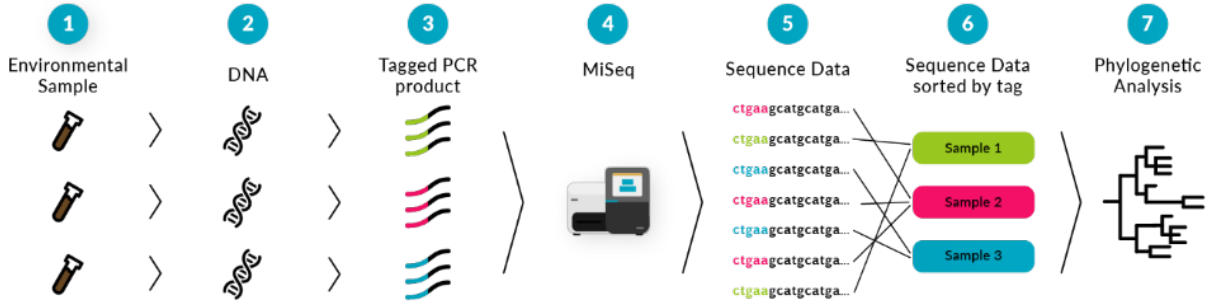

Figure 4: Workflow for getting DNA fingerprints for soil samples: 1) soil sample preparation for DNA analysis, 2) extraction of microbial DNA from soil sample, 3) amplification of the 16S rRNA by polymerase chain reaction, 4) 16S sequence analysis using Illumina MiSeq, 5) processing raw data from MiSeq to verify 16S sequences, 6) processing verified 16S sequences back to individual soil samples and 7) interpretation of 16S sequence data to categorize microbial genera (families).

3. Building a model to correlate DNA fingerprints with productivity

Predicting the productivity of locations where only the soil sample DNA fingerprints are known is performed by a machine learning model which attempts to correlate the results of the sequencing process (the microbial diversity in the DNA fingerprints of all soil samples) with production data for the selected training set, which is based on the unblinded data (see Figure 5).

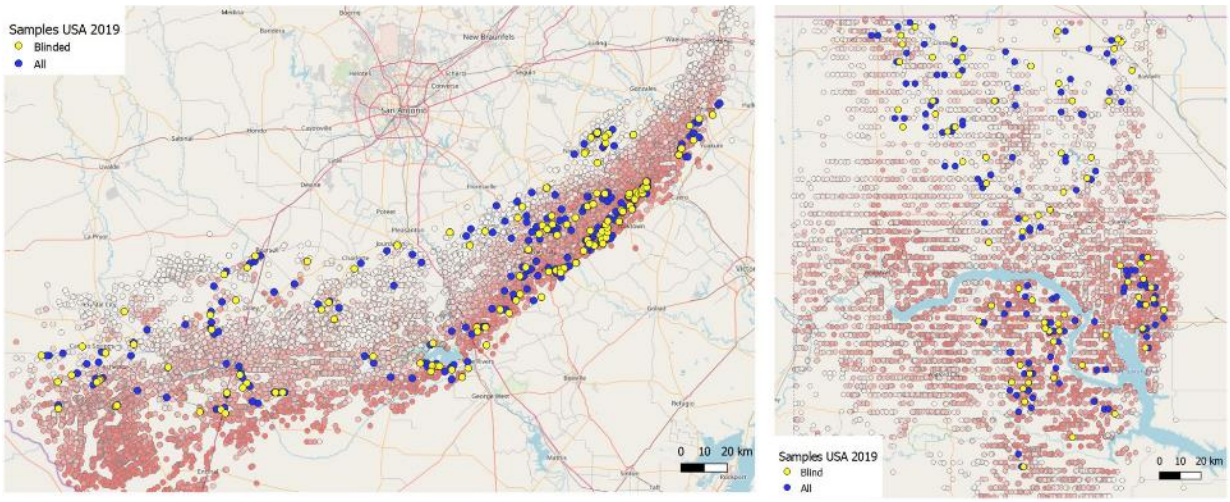

Figure 5: Unblinded (blue) locations in the Eagle Ford (left) and Bakken (right) that are used for training. The yellow locations (100 oil and 100 gas samples) are to be predicted later on and are blinded by the operators.

The model is meant to find specific bacteria whose presence or absence correlates to hydrocarbon production, the so-called biospeciemarkers. These may number only hundreds out of the hundreds of thousands of DNA strings present in the dataset. A biospeciemarker is the DNA fingerprint of a microbe that oxidizes the hydrocarbons transported by microseepage and is, therefore, more abundant above high producing well locations, or a microbe that finds hydrocarbons toxic and is therefore less abundant above high producing well locations. The difference in abundance of biospeciemarkers cannot however be used as an absolute indicator. The need to use more than one biospeciemarkers is described in te Stroet (2017) and Figure 9 of that article with at around 70+ markers the high productivity area becoming visible.

The optimization of the machine learning model is then to find which biospeciemarkers best predict a reserved part of the non-blinded dataset. This is accomplished using Biodentify's "Triple Loop" approach (Figure 6), which is comprised of: 1) an inner loop of non-blinded samples used to correlate DNA fingerprints with known well productivity (referred to as the training set), 2) a highly iterative middle loop of non-blinded samples not used for training that are predicted (referred to as the cross-validation set), and 3) an outer loop which predicts the productivity of blinded samples by using the aggregated loop 2 models.

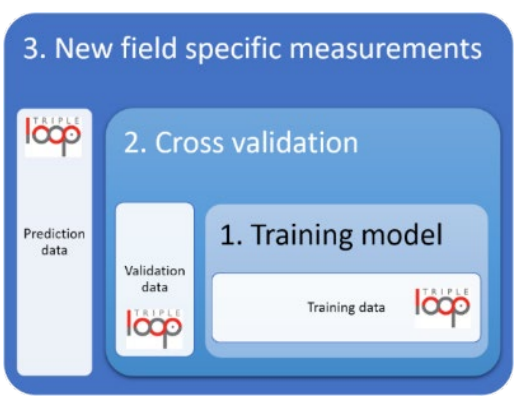

Figure 6: Biodentify's Triple Loop model showing 1) inner loop of non-blinded samples, 2) highly iterative middle lop of non-blinded samples not used for training that are predicted and 3) prediction of the productivity on new or blinded samples.

The procedure is schematized in Figure 7, where the leftmost green column represents all data from the 540 samples. A test set is reserved by masking the location data for 200 samples (orange). For this work, 130 samples in the test (blinded) set were from the Eagle Ford, and 70 from the Bakken. Of these 200

samples, half are oil (100), and half are gas (100) samples. A random subset of 30% of the remaining 340 samples is set aside for checking the model, colored in red (cross validation data). The remaining 70% of the data is then subdivided into two parts: half for developing a model (green), and half as predicted with that model (yellow). The predictive error or misfit is calculated and can be thought of as loop 1.
For loop 2, the same procedure is repeated by swapping these sets, and a model which minimizes the average misfit of the predictions is used to predict the cross-validation set (red). The random segregation to training and cross validation sets is re-run 1000 times to highlight the most predictive biospeciemarkers over all simulations. On average, each point in the non-blinded data is a member of the cross validation set 300 times (30% left out x 1000 random repetitions) and an accuracy measure estimate of the model can be derived from the success of these cross validations.

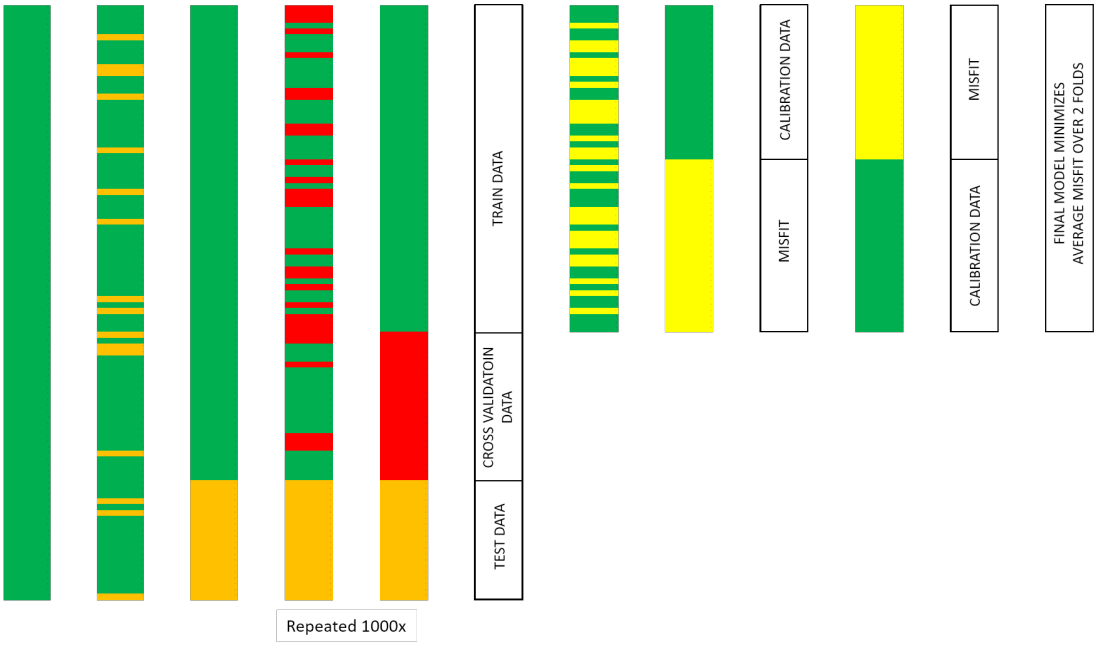

Figure 7: The use of data in machine learning: prediction data (orange), cross validation data (red) and the remaining train data (yellow/green).

4. Predicting and mapping the productivity of the target play

With a successful assemblage of models now built, the productivity can now be predicted for the remaining blinded samples (the orange 'test' samples in Figure 7). In practice, this estimate corresponds to a dimensionless value between -1 and 1 calculated as the average outcome from the 1000 models based only on the DNA fingerprint of the blinded sample. This result can be compared to the blinded data productivity values, giving an objective predictive accuracy. This is what we call loop 3. The blinded data set locations can then be unmasked and mapped along with the other data to check on any spatial patterns.

Results

The predictions of the blinded samples of loop 3 are mapped together with the productivity values for the non-blinded locations for both fields in Figure 8, comprising the full 340 samples in the Eagle Ford and the 200 in the Bakken. The high productive wells (yellow/red) and the low productive wells (green/blue) are in the correct zones (as can be seen by comparing with Figure 3).

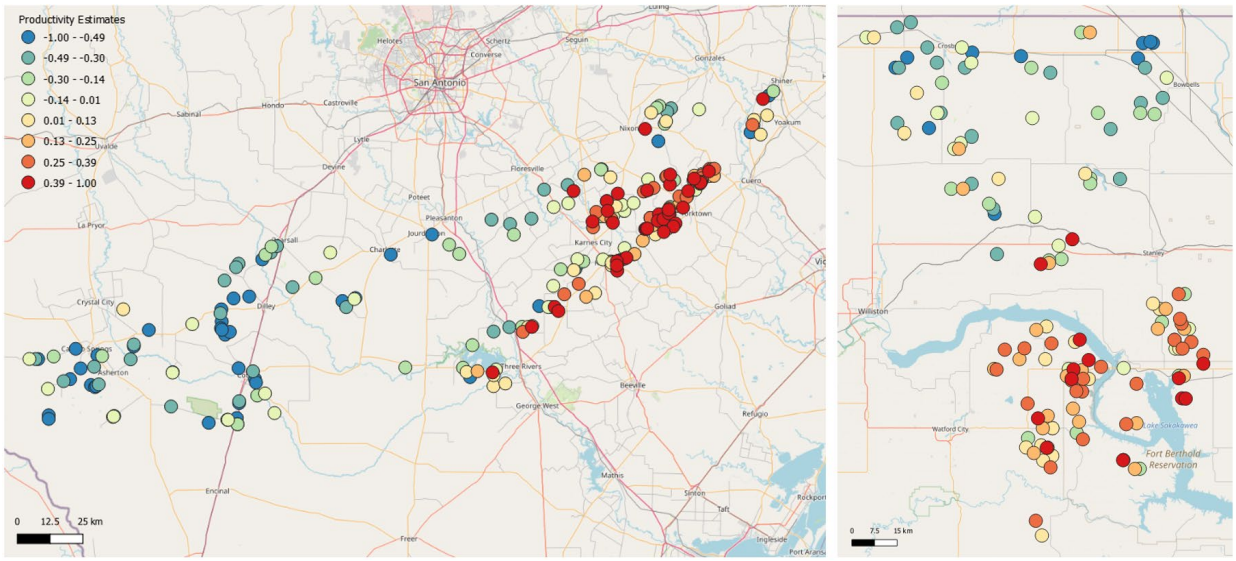

Figure 8: Productivity estimation results from the Eagle Ford Shale (left) and the Bakken Shale (right). The estimates are normalized numbers indicating production potential (i. e., their relative value is important, namely a negative value of -0.8 identifies more strongly a low production well than a value of -0.4.)

Since we were dealing with a binary classification problem, namely predicting which location is part of the low production wells group as opposed to the high production wells group, we can calculate the accuracy of the model estimates. The prediction of the actual well group for a sample is calculated from the model output based on a thresholding constant that is estimated based on the training set. The accuracy is calculated as the ratio of the number of correctly classified samples (i.e., either as highly productive or as low productive) to the total number of samples analyzed. This accuracy measure is usually calculated on the validation sets generated from the training set, providing a measure of confidence in the predictions. However, in this case, the accuracy of the estimates can be quantified objectively, since Aramco/Wintershall team applied a randomization mask to the locations of the blinded samples prior to analysis. This means that the true group of the blinded samples was not known, thus not used, during modeling. After Biodentify used the models developed to make predictions for the blinded samples, Aramco/Wintershall evaluated the results by removing the randomization mask. The resulting model achieved a well productivity predictive accuracy of 85% on the blinded samples, see Figure 9.

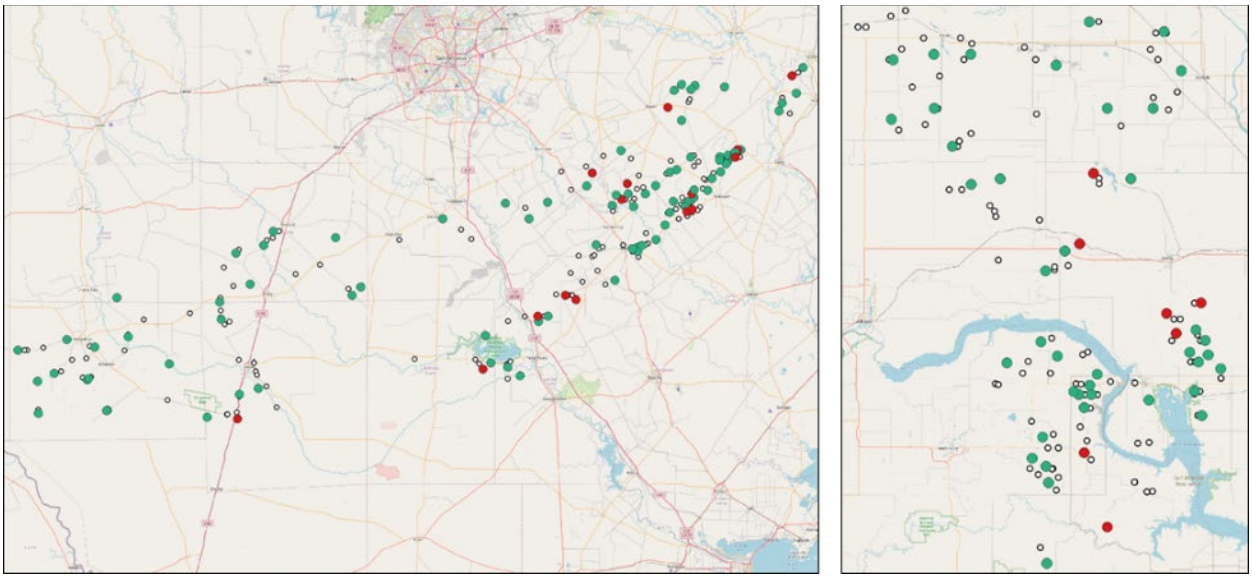

Figure 9: Productivity classification accuracy results from the Eagle Ford Shale (left) and the Bakken (right). The green dots indicate correct predictions, red dots are incorrect prediction, and white dots are the training set.

The 'exportability' of this technology model was tested in two ways: 1) using a model trained on Bakken samples only and using the same Triple Loop approach to predict the productivity of the Eagle Ford locations, and 2) by using the full USA data trained model (all 540 samples) to predict the productivity of a study area in Argentina's Neuquén Basin (a separate Wintershall Dea study). The Eagle Ford locations were predicted with 72% accuracy using only the Bakken data trained model. The complete USA dataset trained model predicted samples above two conventional fields in Argentina with an accuracy of 83%, despite the study areas being separated by several thousand kilometers.

Conclusions and Future Work

Given the good results of the predictive models generated through the work described, we surmise that microseepage has a measurable and indicative influence on the microbial ecosystem. The exportability of the model from North Dakota to Texas and even further to Argentina indicates that the biospeciemarkers have a wide distribution despite the differences in climate, geography, geology, and other potential characteristics of the various locations. Of course, rainfall, temperature, and other factors influence the microbial ecosystem and therefore the DNA extracted from the soil samples. However, based on the work presented, it appears that these effects, whether on the biospeciemarkers, the other microbes present, or the entire population, are filtered through the machine learning process. The similarity of these areas is only that they have been sampled above high or low producers.

The results from the NLP work show that many of the biospeciemarkers, that are selected during training the predictive models, are indeed linked to species/genera that are known to metabolize hydrocarbons or are finding them toxic. However, currently, we are still seeing that many important biospeciemarkers have not yet been cultured and studied in the scientific community, thus they have yet no phylogenetic information associated. This remains for the time being an area for future work.

This technology provides not only a development criterion complementary to other available techniques for existing fields, but new fields in noncontiguous areas can also use the approach for prospect evaluation, de-risking, and initial well selection prior to production data being generated. While the Eagle Ford, Bakken, and Neuquén Basin lack some of the complexity of highly stacked areas of the Permian and other fields, previous work has shown depleted fields are highly responsive due to the speed of microseepage, meaning that the driving in situ pressure is a factor in surface expression. It is, therefore, possible to separate different hydrocarbon settings in the vertical, e.g., separate the signal from Eagle

Ford Shale from the overlaying Austin Chalk fields or even separate between signals from deep gas and shallow oil fields in the Neuquén Basin when there is previous development in the area. Only when stacked fields are laterally overlapping and are similarly charged is the technology incapable of distinguishing the various layers from the accumulated signal. In any case, the technology is proposed not as a standalone approach, but as an inexpensive and relatively quickly characterized method to be used in combination with other existing methods. With more projects being executed, the biospeciemarker database grows. Every additional sample contributes to the breadth and quality of the database, and subsequently, more generic models with higher predictive power are possible.

Finally, we would like to recognize the contribution of Aramco and Wintershall. Through their funding and involvement in this work, including the sampling programs in North Dakota, Texas, and the Neuquén Basin, the DNA analysis in our labs, and certainly also the continued discussions where new ideas were shared and our views were critically tested along the way, we have achieved a more robust methodology and believe the work constitutes an independent and critical proof of the technology.

References

Davis, J. B. 1956. Microbial decomposition of hydrocarbons. *Industrial and Engineering Chemistry*, vol. 48, no. 9, 1444-1448.

Horvitz, L. 1939. On geochemical prospecting-I. *Geophysics*, Vol. 4, 210-228.

Klusman, R. W., Saeed, M. A. 1996. Comparison of Light Hydrocarbon Microseepage Mechanisms. *In "Hydrocarbon migration and its near-surface expression: AAPG Memoir 66"*, eds. D. Schumacher and M.A. Abrams, 157-168.

Mogilevskii, G. A. 1940. The bacterial method of prospecting for oil and natural gases. *Razvedka Nedr*, vol 12, 32-43.

Mogilewskii, G. A. 1959. Geochemical methods of prospecting and exploration for petroleum and natural gas. *Berkely Univ. of California Press*, 349.

Rasheed, M. A., Patil, D. J. and Dayal, A. M. 2013. Microbial Techniques for Hydrocarbon Exploration. *in "Hydrocarbon" Chapter 9*, edts. Vladimir Kutcherov and Anton Kolesnikov, ISBN 978-953-51-0927-3, InTech.

Saunders, D.F., Burson, K. R. and Thompson C. K. 1999. Model for Hydrocarbon Microseepage and Related Near-Surface Alterations. *AAPG Bulletin*, vol. 83, no. 1, 170-185.

Schumacher, D. 1996. Hydrocarbon-induced alteration of soils and sediments. *In "Hydrocarbon migration and its near-surface expression: AAPG Memoir 66"*, eds. D. Schumacher, M.A. Abrams, 71-89.

Schumacher, D. 2012. Pre-Drill Prediction of Hydrocarbon Charge: Microseepage-Based Prediction of Charge and Post-survey Drilling Results. *AAPG Datapages, GeoConvention 2012* (Vision).

Sealy, J. R. 1974. A geomicrobial method of prospecting for oil. *In Oil and Gas Journal*, 15, 98-102.

te Stroet, C. B. M., Zwaan, J., de Jager, G., Montijn, R. and Schuren, F. 2017. Predicting Sweet Spots in Shale Plays by DNA Fingerprinting and Machine Learning. *URTeC:2671117*. https://dx.doi.org/10.15530/URTEC-2017-2671117

Wagner, M., Wagner, M., Piske, J. and Smit, R. 2002. Case histories of microbial prospection for oil and gas. *AAPG Studies in Geology 48* and *SEG Geophysical References Series*, vol. 1