# Equivalent alkane carbon number of crude oils: A predictive model based on machine learning

Benoit Creton, Isabelle Lévêque, Fanny Oukhemanou

Oil & Gas Science and Technology - Rev. IFP Energies nouvelles **74**, 30 (2019)

https://doi.org/10.2516/ogst/2019002

Available online at:
ogst.ifpenergiesnouvelles.fr

REGULAR ARTICLE OPEN ACCESS

# Equivalent alkane carbon number of crude oils: A predictive model based on machine learning

Benoit Creton[1,3,*], Isabelle Lévêque[1,3], and Fanny Oukhemanou[2,3]

[1] IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France
[2] Solvay-Laboratory of the Future, 178 avenue du Dr. Schweitzer, 33600 Pessac, France
[3] The EOR Alliance, www.eor-alliance.com

Received: 1 October 2018 / Accepted: 8 January 2019

**Abstract.** In this work, we present the development of models for the prediction of the Equivalent Alkane Carbon Number of a dead oil (EACNdo) usable in the context of Enhanced Oil Recovery (EOR) processes. Models were constructed by means of data mining tools. To that end, we collected 29 crude oil samples originating from around the world. Each of these crude oils have been experimentally analysed, and we measured property such as EACNdo, American Petroleum Institute (API) gravity and $C_{20^-}$, saturate, aromatic, resin, and asphaltene fractions. All this information was put in form of a database. Evolutionary Algorithms (EA) have been applied to the database to derive models able to predict Equivalent Alkane Carbon Number (EACN) of a crude oil. Developed correlations returned EACNdo values in agreement with reference experimental data. Models have been used to feed a thermodynamics based models able to estimate the EACN of a live oil. The application of such strategy to study cases have demonstrated that combining these two models appears as a relevant tool for fast and accurate estimates of live crude oil EACNs.

### Symbols and acronyms

| | |
|---|---|
| API | American Petroleum Institute |
| CCC | Concordance Correlation Coefficient |
| cEOR | Chemical Enhanced Oil Recovery |
| EA | Evolutionary Algorithms |
| EACN | Equivalent Alkane Carbon Number |
| EACNdo | Dead Oil EACN |
| EACNg | Gas EACN |
| EACNlo | Live Oil EACN |
| EOR | Enhanced Oil Recovery |
| GA | Genetic Algorithm |
| GC | Gas Chromatography |
| GP | Genetic Programming |
| HLD | Hydrophilic–Lipophilic Deviation |
| MAE | Mean Absolute Error |
| MGGP | Multi-Gene Genetic Programming |
| MPLC | Medium Pressure Liquid Chromatography |
| *n*-CV | *n*-fold Cross-Validation |
| PCA | Principal Component Analysis |
| QPPR | Quantitative Property Property Relationship |
| QSPR | Quantitative Structure Property Relationship |
| RMSE | Root Mean Square Error |
| $S^*$ | Optimal Salinity |
| SARA | Saturates, Aromatics, Resins, Asphaltenes |
| SG | Specific Gravity |
| SRK | Soave-Redlich-Kwong |
| SVM | Support Vector Machine |

## 1 Introduction

Microemulsions are commonly encountered in many of products or formulations dedicated to various domains such as pharmaceuticals, cosmetics, or petroleum applications. Among these latter, crude oil extraction after applying primary and secondary recovery methods can be roughly estimated to half of the initial oil reservoir content, according to the considered field [1]. The development of tertiary recovery methods – Enhanced Oil Recovery (EOR) – has gained interest especially with the increase of crude oil prices [2]. The Chemical Enhanced Oil Recovery (cEOR) technique involving combinations of alkali, surfactants and/or polymers aims at decreasing water/oil Interfacial

* Corresponding author: benoit.creton@ifpen.fr

Tension (IFT) in order to mobilize the oil trapped by capillary forces. Optimizing these combinations or formulations to form an efficient microemulsion is a challenging and time-consuming task considering that each potentially eligible reservoir exhibits different conditions such as the oil composition, brine salinity and hardness, pressure, temperature.

The Hydrophilic–Lipophilic Deviation (HLD) concept as proposed by Salager has been applied in numerous studies to mimic phase behavior of {brine/surfactant/oil} systems [3, 4]. When HLD value is zero, the Salager relation linearly correlates the optimal salinity ($S^*$) – the logarithm of the optimum salinity in g/L – with alcohol amount and type ($f(A)$), the temperature and the Equivalent Alkane Carbon Number (EACN), see equation (1) [5].

$$S^* = K(\mathrm{EACN}) + f(A) + \alpha(T - T_{\mathrm{ref}}) - Cc, \qquad (1)$$

where $T_{\mathrm{ref}}$ is set to 298.15 K, $\alpha$ is a temperature coefficient, $Cc$ – the characteristic curvature – and $K$ are parameters related to the surfactant chemistry [6]. The concept of EACN is commonly considered during surfactant formulation design. It assumes that the EACN of an oil is equal to the number of carbon atoms of the $n$-alkane exhibiting a similar phase behavior. EACN of the dead oil (EACNdo) – oil degassed at standard conditions – are experimentally determined by means of test tubes and salinity scans. It consists in identifying the $n$-alkane matching {brine/surfactant/dead crude oil} and {brine/surfactant/$n$-alkane} phase behaviors [7, 8]. However, depending on the crude oil composition several weeks may be necessary to reach the thermodynamics equilibrium.

Bouton *et al.* proposed a Quantitative Structure Property Relationship (QSPR) for the prediction of the EACN of hydrocarbons by means of two theoretical descriptors, *i.e.*, the average negative softness and the Kier A3 [9], with EACN values in between −4 and +35. More recently, Lukowicz *et al.* proposed a QSPR based on COSMO-RS $\sigma$-moments to predict EACN of polar hydrocarbon oils [10] and then extended their model to the case of aprotic polar oils [11]. To determine EACN of hydrocarbon mixtures, Cayias *et al.* [12] and Cash *et al.* [13] proposed the use of a mixing rule in which individual hydrocarbon EACNs are weighted according to corresponding mole fractions. In the case of live oils – oils containing dissolved gases at specific temperature and pressure conditions – we recently proposed an approach to predict the EACN of live oil (EACNlo) on the basis of volumetric fractions of oil and gas [14]. Indeed, Marliere *et al.* have experimentally shown that EACN linearly varies with the gas volumetric fraction [7]. Our EACNlo model necessitates the *a priori* knowledge of the gas composition, gas to oil ratio, temperature and pressure conditions, and the EACN of the dead oil. The volumetric fractions of light hydrocarbons are estimated using the Soave-Redlich-Kwong (SRK) equation of state [15] with the volume correction as proposed by Péneloux *et al.* [16]. The use of our model for the prediction of live crude oil EACN [14] would gain in relevancy developing methods to predict crude oil EACNdo. A crude oil contains thousands of diverse chemicals and the exact composition is never known, as a consequence the combined use of above mentioned models and mixing rules to predict EACNdo is unrealistic.

During the past decade, we considered the use of data mining based approaches to extract information from databases and predict properties of complex fluids [17]. These approaches known by the acronym QSPR aim at identifying non-obvious correlations between property values of the matter and some features rendering information about the matter [18, 19]. In this work, we propose (i) the creation of a database containing experimental EACNdo values as well as results of experimental analysis for a series of crude oils, (ii) the application of machine learning methods to derive models for the prediction of crude oil EACN values, and (iii) the use of developed models for the prediction of live oil EACN for a set of crude oils. The article is organized as follows: the next section deals with materials and methods and gives all details regarding the database creation and methods used to generate models, a subsequent section presents the predictive performance of models and an application of generated models to predict live oil EACN, and the paper ends with concluding remarks.

## 2 Materials and methods

### 2.1 Experimental data and database creation

Wan *et al.* proposed a review and comparisons of applicable approaches to experimentally determine the EACN of a dead crude oil [20]. As detailed in previous works [7, 8], all EACNdo values reported hereafter were obtained following the method referred to as the direct method by Wan *et al.* This method is mainly based on the use of equation (1) and consists in performing several phase diagrams for a {brine/surfactant/crude oil} system varying the salt concentration, *i.e.* a salinity scan. The $S^*$ for the {brine/surfactant/crude oil} system is reached when the phase diagram exhibits an equal repartition of the microemulsion between the oil and the aqueous phase. The so-obtained salinity scan and the $S^*$ value are then compared to those previously obtained for similar conditions (surfactant formulation, temperature, brine composition…) in the case of linear alkanes such as $n$-decane ($C_{10}$), $n$-dodecane ($C_{12}$) and $n$-tetradecane ($C_{14}$). The EACNdo value for the crude oil is finally determined by solving equation (1) using $K$ and $Cc$ parameters obtained for $C_{10}$, $C_{12}$ and $C_{14}$. This analytical methodology has been applied on 29 crude oils from around the world, and obtained EACNdo values are reported in Table 1. EACNdo values for crude oils of interest lie in between 1.2 and 18.0. The experimental determination of EACNdo for crude oils requires days to weeks to reach thermodynamic equilibrium depending on their nature, whether they are light or heavy. The interest of a fast and accurate theoretical method to predict EACNdo for a crude oil thus becomes evident but its parameters should be easily determined.

**Table 1.** Experimental EACNdo, API gravity, $C_{20^-}$ fraction (%wt) and fractions (%wt) of saturates (Sat.), aromatics (Aro.), resins (Res.), and asphaltenes (Asp.) measured for the 29 crude oils considered in this study.

| Crude oil | EACNdo | API gravity | $C_{20^-}$ | Sat. | Aro. | Res. | Asp. |
|---|---|---|---|---|---|---|---|
| #01 | 18.0 | 22.3 | 31.36 | 21.01 | 15.62 | 27.49 | 4.52 |
| #02 | 1.2 | 49.8 | 94.18 | 2.68 | 1.52 | 1.50 | 0.12 |
| #03 | 12.0 | 27.6 | 39.71 | 11.81 | 13.42 | 19.63 | 15.42 |
| #04 | 9.0 | 39.0 | 64.08 | 15.94 | 10.33 | 9.03 | 0.62 |
| #05 | 14.7 | 37.9 | 12.60 | 13.45 | 23.90 | 38.58 | 11.48 |
| #06 | 11.0 | 30.5 | 46.74 | 18.62 | 14.79 | 18.20 | 1.65 |
| #07 | 17.5 | 29.5 | 24.39 | 38.59 | 7.62 | 22.69 | 6.70 |
| #08 | 15.5 | 37.9 | 44.83 | 36.34 | 7.46 | 10.28 | 1.08 |
| #09 | 13.5 | 37.5 | 50.77 | 36.81 | 4.63 | 6.86 | 0.93 |
| #10 | 12.0 | 26.3 | 50.63 | 30.97 | 7.19 | 10.33 | 0.88 |
| #11 | 6.7 | 28.9 | 50.49 | 22.12 | 10.27 | 16.45 | 0.66 |
| #12 | 13.0 | 34.3 | 30.91 | 35.30 | 15.72 | 16.35 | 1.72 |
| #13 | 14.0 | 22.6 | 39.90 | 10.24 | 12.32 | 19.58 | 17.97 |
| #14 | 13.6 | 31.9 | 40.21 | 24.57 | 13.78 | 18.15 | 3.29 |
| #15 | 15.5 | 37.8 | 29.26 | 44.41 | 9.39 | 15.35 | 1.59 |
| #16 | 16.4 | 31.1 | 25.22 | 32.49 | 7.94 | 30.29 | 4.04 |
| #17 | 12.4 | 24.4 | 39.29 | 19.52 | 12.31 | 22.26 | 6.62 |
| #18 | 10.2 | 32.3 | 44.61 | 15.46 | 15.89 | 21.74 | 2.29 |
| #19 | 13.6 | 11.0 | 13.87 | 13.21 | 16.64 | 40.26 | 16.03 |
| #20 | 16.5 | 37.6 | 34.54 | 42.24 | 8.02 | 13.88 | 1.31 |
| #21 | 7.2 | 30.9 | 45.07 | 26.51 | 12.33 | 15.09 | 1.00 |
| #22 | 12.0 | 29.5 | 41.90 | 13.29 | 16.98 | 25.07 | 2.76 |
| #23 | 13.8 | 26.1 | 25.41 | 12.08 | 20.02 | 24.28 | 18.21 |
| #24 | 15.5 | 34.8 | 34.61 | 36.14 | 8.36 | 13.74 | 7.15 |
| #25 | 16.5 | 35.9 | 32.67 | 40.51 | 6.73 | 15.24 | 4.85 |
| #26 | 14.2 | 27.1 | 23.06 | 23.38 | 19.56 | 32.78 | 1.23 |
| #27 | 11.9 | 23.7 | 40.38 | 25.88 | 14.22 | 18.68 | 0.84 |
| #28 | 13.0 | 31.1 | 47.86 | 14.16 | 14.83 | 19.25 | 3.89 |
| #29 | 12.5 | 27.5 | 46.10 | 27.19 | 11.24 | 11.64 | 3.83 |

The American Petroleum Institute (API) gravity measures whether a crude oil is lighter or heavier than water. It is defined using the following expression:

$$\text{API} = \frac{141.5}{\text{SG}} - 131.5, \tag{2}$$

with the Specific Gravity (SG) $= \rho_{\text{crude oil}}/\rho_{\text{water}}$, where $\rho_{\text{crude oil}}$ and $\rho_{\text{water}}$ denote the density of the crude oil and the density of water, respectively at 15.5 °C (60 °F). Densities for the 29 crude oils were measured using an Anton Paar density meter (model DMA$^{\text{TM}}$ 4500 M) including an oscillating U-tube sensor, and the uncertainty associated to density measurements is 0.1%. API gravity values were then calculated using equation (2). Table 1 presents API gravity values for crude oils #01 to #29, it reveals that the set of crude oils covers a broad range of API gravity values from 11 to 50 denoting heavy and light crude oils, respectively. It is interesting to note that in the set of crude oils, crude oil #19 appears as an outlier in terms of API gravity.

The Saturates, Aromatics, Resins, Asphaltenes (SARA) analysis is a method based on fractionation to characterize the crude oil content in terms of saturates, aromatics, resins, and asphaltenes [21]. The basic idea is to divide the crude oil into smaller fractions playing with oil component solubilities in solvents such as linear alkanes. Different SARA methodologies have been described for instance varying the used *n*-alkane, *i.e.* *n*-pentane or *n*-heptane changing the amount of precipitated asphaltenes [22]. All considered crude oils were characterized using a SARA analysis similar to that used by Behar *et al.* [23], and the analytical procedure can be briefly described as follows: Each crude oil is dissolved in *n*-pentane at 43 °C, and the resulting solution is filtered (Durapore® membrane in polyvinylidene fluoride with 0.45 µm pore size) to separate by-products from other crude oil components. By-products are treated by adding dichloromethane at the same temperature to recover the precipitated asphaltenes, and fractions are weighted during the entire procedure. An aliquote of the *n*-pentane rich solution is analyzed by Gas Chromatography (GC) to quantify the

$C_{20-}$ fraction – the crude oil fraction containing compounds with a number of carbon atoms lower than 20. An another aliquote is evaporated and separated using Medium Pressure Liquid Chromatography (MPLC) to characterize weight amounts of saturates, aromatics, and resins. All fractions are then standardized according to measured masses, and the crude oil is assumed as the blend of $C_{20-}$, saturates, aromatics, resins, and asphaltenes. Table 1 presents results of the SARA analysis for the 29 considered crude oils. Fractions of saturates, aromatics, resins, and asphaltenes are determined with associated uncertainties of 2%wt, in agreement with conclusions drawn by Aske *et al.* [24]. Experiments necessary to determined $C_{20-}$, saturate, aromatic, resin, and asphaltene fractions demand approximatively 4 days whatever the nature of the oil. Note that fractions of saturates obtained with different SARA methodologies should be similar, and that the Aro., Res. or Asp. fraction determined with one SARA analysis should correlate with its corresponding fraction issued from a different SARA methodology. These assumptions result from comparisons of data reported in Table 1 with fractions determined using another SARA methodology (not shown here). From data reported in Table 1, it is interesting to note that in this set of crude oils, crude oil #02 appears as an outlier in terms of $C_{20-}$ and SARA fractions.

Hereafter, each crude oil is characterized by descriptors presented in Table 1, *i.e.* API gravity, $C_{20-}$ fraction and fractions of saturates (Sat.), aromatics (Aro.), resins (Res.), and asphaltenes (Asp.). The development of models will then consist in relating EACNdo to descriptors leading to relations between properties, *i.e.* Quantitative Property Property Relationships (QPPR). Several works in the literature report correlations of crude oil properties with SARA analysis outputs [25–30].

## 2.2 Modeling methods

### *2.2.1 Data sets*

The accuracy of predictive QPPR (similarly to QSPR) is in part related to the quality of data, hence the quality of the database is a keystone for the success of such modeling works. Possible correlations between descriptors have been investigated by generating a correlation matrix by means of the Materials Studio software [31]. No evidence of highly correlated descriptors has been found, and the highest values in the correlation matrix were obtained for couples: Res. with Aro. and Res. with $C_{20-}$. Fan and Buckley proposed on the basis of six medium-gravity dead crude oils (with API gravity values from 22.6 to 37.2) a relation between API gravity and SARA fractions [25]. No evidence of such a relation has been found in our data set. The best found correlation between API gravity and SARA fractions has a low coefficient of determination ($R^2 \simeq 0.4$). Data presented in the Table 1 have been used to constitute our database.

As a preprocessing of the data, we performed a Principal Component Analysis (PCA) applied on API gravity, $C_{20-}$ and SARA fractions measured for the 29 crude oil samples. Figure 1 presents projections of the 29 crude oils in the space formed by the three main principal components resulting from the PCA. The diagram thus provides a representation of crude oil distributions in the chemical space of our database. This distribution reveals that one crude oil (crude oil #02) is isolated from all other samples that thus confirms crude oil #02 as an outlier. In Figure 1, each symbol is filled as a function of their EACNdo value, and there is no obvious relation between the location of a crude oil on the diagram and the value of its EACNdo.

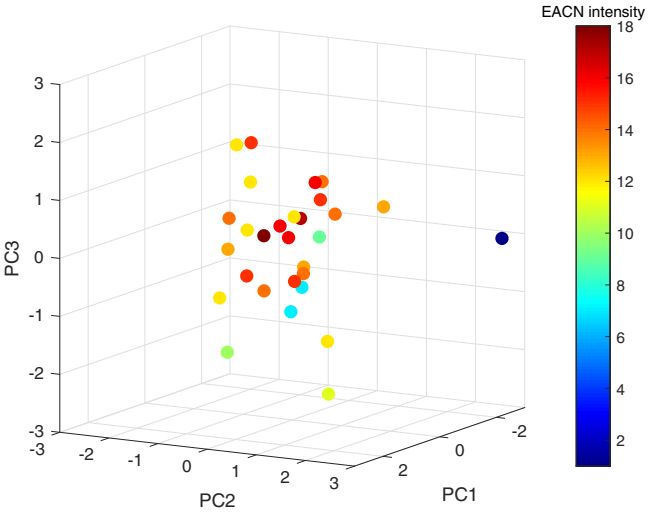


**Fig. 1.** Projections of crude oils on PC1, PC2 and PC3, the three first principal components resulting from the PCA. Symbols are filled using a colorbar and intensities are as follows: low and high EACNdo values are indicated in blue and red, respectively.

Application of external validation has been shown as necessary to validate model's robustness when predicting new compound property values, meaning candidates not used during the model development [32]. One of the popular methods is the $n$-fold Cross-Validation ($n$-CV) in which the data set is randomly divided in approximately equal $n$ portions, the leave-one-out being its extreme version with $n$ equals the number of samples in the database. An aggregate of $(n-1)$ portions forms the Training set used to optimize predictive models, the remaining portion constituting the Test set. We emphasize that no data point belonging to external sets is used to derived models. This procedure is repeated $n$ times choosing at each new fold another portion of data as a Test set. From conclusions drawn during the preprocessing of the database, we choose to impose crude oils #01 and #02 in all Training sets in order to keep EACNdo ranges constant. The 27 remaining crude oils were randomly distributed into nine portions, therefore the Training sets and Test sets represent 90% and 10% of the database, respectively.

Perfomances of models are evaluated on both Training and Test sets calculating values for some statistical indicators such as the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), $R^2$, and the Concordance Correlation Coefficient (CCC) [33]. Chirico and Gramatica have shown that the use of this latter coefficient is advocated considering various scenarios such as location shifts, scale shifts, and location plus scale shifts [34, 35].

### 2.2.2 Machine learning methods

The application of machine learning methods for thermophysical property predictions has been the subject of our numerous past and ongoing research works [17]. From comparisons drawn in these previous studies, we have observed that Support Vector Machine (SVM) based models outperform other evaluated learning algorithms such as neural networks, partial least squares, Genetic Algorithm (GA). However, the number of data points is in this work quite small and we hereafter search for explicit multilinear equations that can be easely implemented in a spreadsheet. One possible manner to generate multilinear models is to use Evolutionary Algorithms (EA). EA are based on the Darwinian evolution theory of biological species in nature, and they represent promizing methods for optimization problems. When applied to regression problems, the resolution consists in the iterative evolution of a population of equations initially randomly generated and respecting the general following form:

$$\text{Property} = \lambda_0 + \sum_{i=1}^{N} \lambda_i G_i, \tag{3}$$

where $\lambda_0$ is the inercept, $\lambda_i$ denotes a weight associated to the gene $i$ ($G_i$), and $N$ is the total number of genes in the model. To derive models, two approaches were tested varying from the level of information in genes.

In the first approach, a gene (see the proposed schematic representation in Fig. 2) consists of a tree built by combining descriptors (API gravity, fractions of $C_{20^-}$, Sat., Aro., Res., and Asp.) and mathematical functions (see Tab. 2) allowing to catch non-linearity in property variation. Multi-Gene Genetic Programming (MGGP) based models were generated using the Genetic Programming Toolbox for the Identification of Physical Systems (GPTIPS) coded in the MATLAB environment [36–38]. We applied the tournament method to select individuals in the population of equations on the basis of their fitness and complexity. We fixed the tournament size to 25 corresponding to 10% of the population size. Generations are constructed by survival of fitter individuals, and reproduction of individuals consists in applying crossover as well as mutation operations to produce child equations. The iterative procedure ends when one of the fixed criteria such as maximum number of generations, best fitness values is reached. Note that during the iterative procedure, the structure of trees evolves through out crossover and mutation operations applied to sub-tree elements. Clearly, the maximum numbers of genes and nodes per tree must be limited to prevent overfitting problems. Additionally, the maximum numbers of generations and runs – repetition of the calculation – should be optimized to ensure convergence of calculations for reasonable computational ressources. Table 2 reports details about values and/or ranges of investigated GPTIPS settings in this work.

In the second approach, a gene (see Eq. (3)) simply stands for one of the descriptors. A GA based variable selection method was followed, and the Genetic Function Approximation (GFA) as implemented in the Materials Studio software was used to build multilinear models [31]. The GFA procedure consists in iterations of selection, crossover, and mutation operations, coupled with objective criteria such as $R^2$ in order to extract the best fitting models. In this work, the adjusted $R^2$ was used as the objective criteria. The initial population (*i.e.*, initial number of equations) was set to 6, and the maximum generation number to 50 000. This procedure was performed on each of the nine Training sets, noting that the same decompositions (folds, Training, and Test sets) are used during GFA and GPTIPS based procedures.

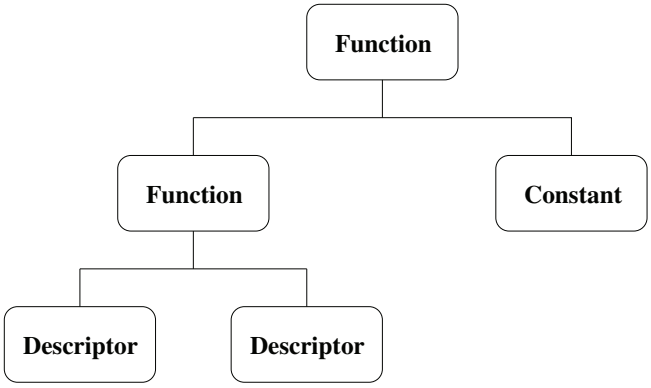


**Fig. 2.** Schematic representation of a tree. Function stands for a mathematical operator belonging to the function set (see Tab. 2), for instance +, −, × or ÷. Descriptor denotes either API gravity, fractions of $C_{20^-}$, Sat., Aro., Res., or Asp.

**Table 2.** Investigated parameter settings for the MGGP based method.

| Parameter | Corresponding values |
|---|---|
| Function set | +, −, ×, ÷, √, exp, ln |
| Population size | 250 |
| Number of runs | 1, 5, 10, 15, 20, 25, 30, 40 |
| Tournament size | 25 |
| Maximum tree depth | 4 |
| Number of generations | 100, 500, 1000, 2000 |
| Maximum number of genes | 1, 2, 3, 4, 5, 6 |
| Maximum number of nodes per tree | 2, 4, 6, 8, 10, 12 |
| Mutation events | 0.10 |
| Crossover events | 0.85 |
| Reproduction events | 0.05 |

## 3 Results and discussion

### 3.1 Development of QPPR models

In this section, we report various QPPR models to predict EACN of dead crude oil knowing a series of experimental data such as API gravity, $C_{20^-}$ fraction, and fractions of saturates, aromatics, resins, and asphaltenes, see Table 1. Two machine learning methods based on EA were used: GFA and GPTIPS.

**Table 3.** Parameter values for the GFA model presented in equation (4), determined using each fold as external data (Test set). <> stands for the average of parameter values taken over the nine folds.

| | Fold-01 | Fold-02 | Fold-03 | Fold-04 | Fold-05 | Fold-06 | Fold-07 | Fold-08 | Fold-09 | <> |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_0$ | 13.761 | 14.508 | 13.312 | 13.270 | 13.765 | 14.007 | 13.033 | 12.127 | 13.371 | 13.462 |
| $\lambda_1$ | −0.120 | −0.133 | −0.124 | −0.120 | −0.129 | −0.132 | −0.121 | −0.116 | −0.122 | −0.124 |
| $\lambda_2$ | 0.141 | 0.140 | 0.148 | 0.149 | 0.146 | 0.141 | 0.160 | 0.175 | 0.147 | 0.150 |
| $\lambda_3$ | 0.112 | 0.056 | 0.162 | 0.140 | 0.129 | 0.154 | 0.146 | 0.180 | 0.146 | 0.136 |

**Table 4.** Performance characteristics for the GFA based model. Values taken by some statistical indicators when applied to predictions of equation (4) for crude oils in Test sets (Pred.). <> stands for the average of parameter values taken over the nine folds.

| | Pred. | <> |
|---|---|---|
| MAE | 1.32 | 1.25 |
| RMSE | 1.90 | 1.80 |
| $R^2$ | 0.702 | 0.732 |
| CCC | 0.827 | 0.846 |

During the development of GFA based models, the maximum number of variables allowed to form equations was set to four. This value meets the statistical criteria $n \geq 4k$, where $k$ and $n$ are the number of variables in the model and the number of data points in the Training set, respectively [39]. GFA based models were optimized following a 9-fold cross-validation procedure. Samples #01 and #02 are fixed in Training sets, and therefore each of the nine folds contains three randomly selected crude oils. Equation (4) presents the so-obtained model. This model is composed of three descriptors weighted by $\lambda_i$ coefficients, and a constant $\lambda_0$ as follows:

$$\text{EACNdo} = \lambda_0 + \lambda_1 \text{C}_{20^-} + \lambda_2 \text{Sat.} + \lambda_3 \text{Asp.}, \qquad (4)$$

$\lambda_i$ values obtained considering successively each fold as external data (not used during the optimization procedure) are reported in Table 3. Equation (4) is in line with some physical intuitions. For instance, the negative value of $\lambda_1$ agrees with the fact that EACNdo should decrease with increasing $C_{20^-}$ fractions (the light fraction of the crude oil). The opposite effect is observed through out the positive values for $\lambda_2$ and $\lambda_3$, *i.e.* EACNdo values should increase with Sat. and Asp. fractions (heavy fractions of the crude oil). Note that the value obtained for $\lambda_0$ roughly corresponds to the average of EACNdo values in the database (12.9 points of EACN). Table 4 presents performance characteristics for equation (4), calculated as follows: (i) For all crude oils belonging to Test sets and corresponding $\lambda_i$ coefficients (see Tab. 3). For instance, for crude oils belonging to Fold-01 we used $\lambda_0 = 13.761$, $\lambda_1 = -0.120$, $\lambda_2 = 0.141$, and $\lambda_3 = 0.112$. Performance characteristics reported in Table 4 so reflect the predictive capabilities of GFA based models. (ii) For all crude oils and average parameters (see Tab. 3). In this case, no conclusion can be drawn regarding the predictive capabilities of the GFA based model as all crude oils have been indirectly involved in the learning procedure.

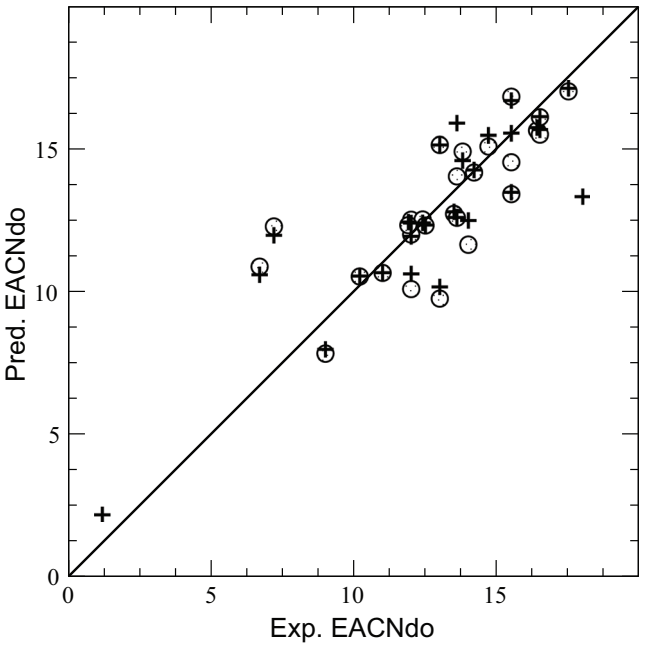


**Fig. 3.** Scatterplots of experimental EACNdo values *versus* predicted EACNdo values using equation (4). Circles stand for pure predictions using GFA based models while plus symbols denote the use of average parameters.

Figure 3 presents scatterplots of experimental EACNdo *versus* predicted EANCdo values using equation (4) with either parameters associated to each fold in Test set, or average parameters. The diagram exhibits no huge difference between the two sets of parameters in terms of predicted EACNdo values. Values taken by some statistical indicators as reported in Table 4 quantitatively confirm this observation. Noting that EACNdo values for samples fixed in Training sets (*i.e.*, 18.0 and 1.2 for crude oils #01 and #02, respectively) can only be predicted using average parameters, and GFA based model fails in predicting the EACNdo value for crude oil #01.

GP is a generalization of GA. The main difference between GP and GA is that models obtained by the latter are strings of weighted descriptors, while the former returns tree expressions in which a node is either a descriptor, a mathematical function or a coefficient. MGGP can be seen as a combination of GA and GP as MGGP returns strings of genes, each gene having a tree structure [40]. We have

used MGGP to develop correlations able to catch possible non-linearity effects in EACNdo modelling. In MGGP as implemented in the GPTIPS, important settings should be parameterized, for instance numbers of runs, generations, genes, nodes. To our knowledge there is no methodology to parameterize GPTIPS and it often results from trials and errors. Mohamadi-Baghmolaei *et al.* indicated that some of MGGP parameters are commonly randomly set [41], and Garg *et al.* proposed to adjust MGGP parameters according to the nature of the regression problem [40, 42]. For our problem, the maximum number of genes has been chosen according to the statistical criteria $n \geq 4k$, where $k$ and $n$ are the number of variables (*i.e.* genes) in the model and the number of data points in the Training set, respectively. Considering both the database content and the Training/Test splitting, a reasonable value for the maximum number of genes is 6. Regarding settings dedicated to tree's structure, the maximum number of nodes per tree and the maximum depth of tree have been set to 12 and 4, respectively. For settings related to convergence of calculations, the maximum number of generations and the maximum number of runs have been set to 2000 and 40, respectively. In order to allow a great possibility in nonlinear models, we have chosen a large number of mathematical operators such as addition, subtraction, multiplication, division, square root, exponential, logarithm. We propose an approach to optimize the numbers of runs, generations, genes, and nodes that will further be used to develop a model. The convergence matter was first addressed, and the trade-off between convergence, computational time, and accuracy was treated as follows:

(i) The numbers of genes and nodes were set to their respective maximum possible value in order to generate models with the highest complexity. We then compared model's performances exploring the space formed by the number of generations discretized as follows: 100, 500, 1000, and 2000, and by the number of runs discretized as follows: 1, 5, 10, 15, 20, 25, 30, and 40, as indicated in Table 2. Performances were evaluated using the RMSE statistical indicator calculated over all samples in the dataset. No Training and Test sets division was considered in this step as the idea is to roughly select appropriate numbers of runs and generations, leading to reasonable accuracy and computational time. Figure 4 presents the power law evolution of model's performance as a function of numbers of runs and generations. Clearly, models poorly perform when optimized with only 100 generations. Considering 500, 1000 or 2000 generations leads roughly to similar performances after 20 runs. However, the deviation between raw data and the corresponding power law is in the case of 2000 generations twice that obtained for 500 or 1000 generations (not shown here). Contour plots on Figure 4 indicate that for a similar computational cost, equivalent performances are obtained with 500 or 1000 generations. The consideration of 30 runs seemed relevant to ensure the convergence of calculations, and in order to limit the computational time we selected 500 generations.

(ii) Numbers of generations and runs were respectively set to 500 and 30, and models were developed all along the space formed by the number of genes discretized as follows: 1, 2, 3, 4, 5, and 6, and by the number of nodes discretized as follows: 2, 4, 6, 8, 10, and 12, as indicated in Table 2. Contrary to step (i), Training and Test sets were considered here using the nine folds previously generated for the GFA based model development. For each point of the space, sum of squared errors were calculated for Test sets, and the optimum numbers of genes and nodes were determined minimizing the total sum of squared errors on Test sets. The optimization procedure led to a number of nodes of 8 and a number of genes of 4.

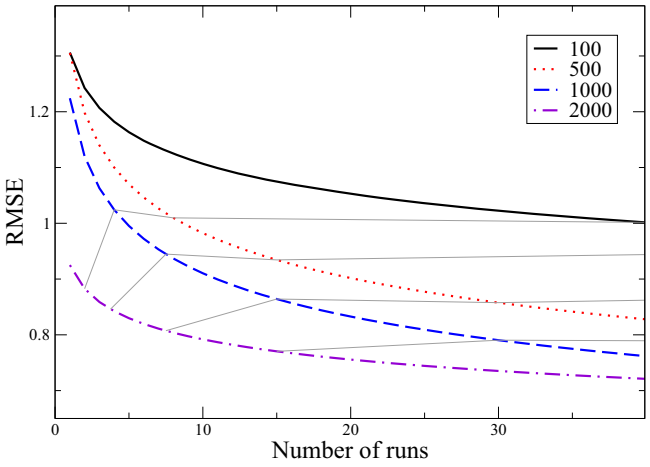


**Fig. 4.** Power law evolution of model's performance (on the basis of RMSE values) as a function of the number of runs, for number of generations in between 100 and 2000. Contour plots (thin grey lines) show lines that share the same computational cost.

The proposed optimization procedure applied to our regression problem led to numbers of runs, generations, genes, and nodes of 30, 500, 4, and 8, respectively. Nine GPTIPS based models were optimized following a 9-fold cross-validation procedure, noting that the nine folds generated for the GFA based model development have been reused. Performances of models are presented in Table 5. All models outperform GFA based models, and exhibit RMSE values in between 0.97 and 1.40. The model that best generalizes the database has been developed using Fold-08 as Test set. Details about this latter model such as the four weighted genes and the intercept are presented in equation (5).

$$\begin{aligned} \lambda_0 &= \text{Intercept} = -44.28, \\ \lambda_1 G_1 &= -0.39 \frac{\text{Aro.}}{\text{Asp.}}, \\ \lambda_2 G_2 &= 0.30 \exp\left(\frac{\text{Sat.}}{\text{API}-\text{Aro.}}\right), \\ \lambda_3 G_3 &= 4.67 \times 10^{-5}\left(\text{API}^3 + \text{Sat.}^3 + \text{Res.}^3\right), \\ \lambda_4 G_4 &= 55.12 \exp\left(-\exp\left(\text{Aro.}\right)\right). \end{aligned} \tag{5}$$

**Table 5.** Performance characteristics (statistical indicators) GPTIPS based models applied to crude oils in the database. Fold-0$i$ means that the fold is external to the learning procedure.

| | Fold-01 | Fold-02 | Fold-03 | Fold-04 | Fold-05 | Fold-06 | Fold-07 | Fold-08 | Fold-09 |
|---|---|---|---|---|---|---|---|---|---|
| MAE | 0.80 | 0.83 | 0.94 | 1.04 | 0.93 | 0.89 | 0.88 | 0.81 | 0.96 |
| RMSE | 1.02 | 1.30 | 1.27 | 1.40 | 1.18 | 1.08 | 1.07 | 0.97 | 1.38 |
| $R^2$ | 0.915 | 0.861 | 0.868 | 0.838 | 0.886 | 0.903 | 0.905 | 0.922 | 0.843 |
| CCC | 0.954 | 0.925 | 0.934 | 0.919 | 0.940 | 0.951 | 0.950 | 0.960 | 0.921 |

Clearly, each gene non-linearly contributes to the predicted EACNdo value, and the model involves all descriptors excepted the $C_{20-}$ fraction. Figure 5 presents the scatterplot of experimental EACNdo *versus* predicted EANCdo values using equation (5). All data points are less scattered from both sides of the bisector (predicted EACNdo equals experimental EACNdo) as compared to observations performed on Figure 3. Contrary to the GFA based model, equation (5) well predicts crude oil #01 with a value of 18.2. Although crude oil #02 has been sensed as an outlier in terms of EACN value and composition, its EACN is well estimated using both GFA and GPTIPS based models. The largest deviation between experimental and predicted EACNdo values (2.5 EACN points) is measured for crude oil #21, noting that none of GPTIPS based models developed on each fold succeeds in predicting this value and that the GFA based model also failed in predicting this EACNdo value.

We investigated the sensitivity of equation (5) regarding uncertainties associated to input data. Indeed, due to filtration operations, evaporation losses and/or incomplete solvent removal during SARA analysis, fractions of saturates, aromatics, resins, and asphaltenes are determined with associated uncertainties of about 2%wt [24]. API is calculated from the density of the crude oil and using equation (2). The uncertainty associated to density measurements is 0.1%. Although the fraction of aromatics is involved in genes $G_1$, $G_2$, and $G_4$, a 2%wt deviation applied on Aro. fractions in Table 1 only slightly deteriorates predictions with MAE and RMSE of 0.87 and 1.00, respectively. Performances of equation (5) only falls to MAE = 0.96 and RMSE = 1.12, considering a 0.1% deviation on density values and a 2%wt deviation on each SARA fraction (Tab. 1).

### 3.2 Application to the prediction of live crude oil EACN

Live oils are oils containing dissolved gases at specific temperature and pressure conditions. Creton and Mougin proposed a model based on thermodynamics to predict the EACN of a live oil (EACNlo) knowing the EACN of the dead oil (stock tank oil), reservoir pressure ($P$) and temperature ($T$) conditions, as well as the gas to oil ratio ($R_{si}$) [14]. This model is based on a volumetric mixing law applied to EACNdo and EACN of gas (EACNg), as follows:

$$\text{EACNlo} = (1 - \phi_g)\text{EACNdo} + \phi_g\text{EACNg}, \qquad (6)$$

where, EACNg equals the sum of $n$-alkane carbon atom numbers (ACN, alkane carbon number) weighted by their respective volumetric fraction, *i.e.* when solely methane is used as representative gas: EACNg equals 1. Molar volumes were calculated using the SRK [15] Equation of State (EoS) applied with the volume correction proposed by Péneloux *et al.* [16]. Creton and Mougin validated their model for several crude oils covering broad ranges of reservoir characteristics, and studied impact of pressure, temperature, and gas composition on predicted live oil EACNs [14]. These latter crude oils are part of our database (Tab. 1), and we propose hereafter to feed the model proposed by Creton and Mougin with EACNdo values obtained using equation (5). Table 6 presents for crude oils in common between reference [14] and this work, some oil properties and their original reservoir characteristics. Figure 6 presents scatterplots of experimental EACNlo values *versus* predicted EACNlo values. Using the full predictive approach – the model proposed by Creton and Mougin together with equation (5) – a MAE of 1.5 EACN point is observed on EACNlo predictions as compared to a MAE of 1.1 EACN point when experimental EACNdo are used to feed the model by Creton and

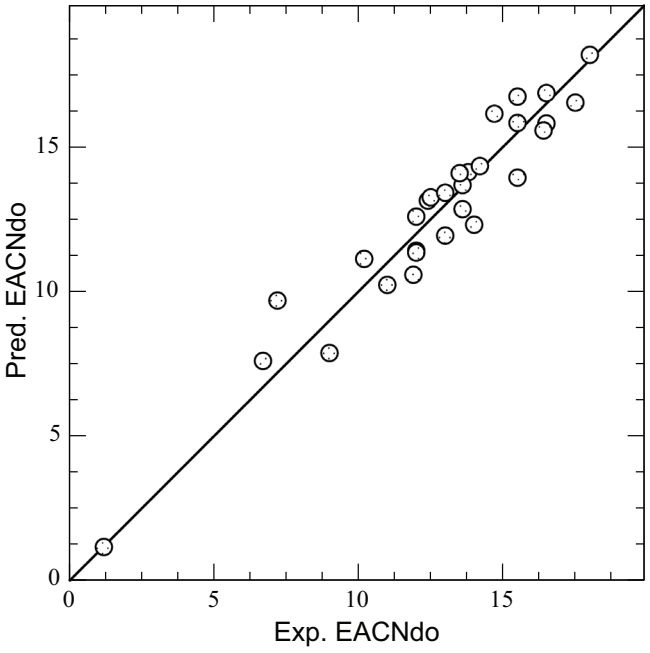


**Fig. 5.** Scatterplots of experimental EACNdo values *versus* predicted EACNdo values using equation (5).

**Table 6.** Summary of live crude oil properties. For each case study, the predicted EACNdo using equation (5), reservoir temperature and pressure, the gas to oil ratio, the solution gas composition, and experimental and predicted EACNlo values are indicated [14].

| Crude oil | #04 | #08 | #10 | #11 | #12 | #13 | #14 | #15 | #20 | #23 | #24 | #25 | #26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EACNdo pred. | 7.9 | 13.9 | 11.4 | 13.4 | 13.4 | 12.3 | 12.8 | 16.7 | 15.8 | 14.1 | 15.8 | 16.9 | 14.3 |
| $T$ (°C) | 120 | 95 | 82 | 110 | 110 | 100 | 102 | 95 | 65 | 108 | 75 | 85 | 40 |
| $P$ (bar) | 180 | 143 | 190 | 215 | 155 | 186 | 90 | 125 | 83 | 82 | 140 | 108 | 125 |
| $R_{si}$ ($Sm^3/m^3$) | 214 | 137 | 84 | 52 | 40 | 48 | 53 | 69 | 54 | 35 | 75 | 77 | 51 |
| $x_{methane}$ | 0.339 | 0.418 | 0.446 | 0.348 | 0.291 | 0.345 | 0.216 | 0.294 | 0.400 | 0.261 | 0.290 | 0.289 | 0.359 |
| $x_{ethane}$ | 0.256 | 0.099 | – | – | – | – | 0.067 | 0.082 | – | – | 0.107 | 0.092 | – |
| $x_{propane}$ | – | 0.097 | – | – | – | – | 0.054 | 0.069 | – | – | 0.069 | 0.106 | – |
| $x_{n\text{-pentane}}$ | 0.022 | – | – | – | – | – | – | – | – | 0.016 | – | – | – |
| EACNlo exp. | 6.5 | 9.0 | 12.0 | 13.0 | 13.0 | 14.0 | 10.0 | 11.5 | 13.5 | 13.0 | 11.0 | 11.5 | 12.0 |
| EACNlo pred. | 5.1 | 10.0 | 9.5 | 11.8 | 12.1 | 11.0 | 11.1 | 13.9 | 13.8 | 12.7 | 13.1 | 13.8 | 12.8 |

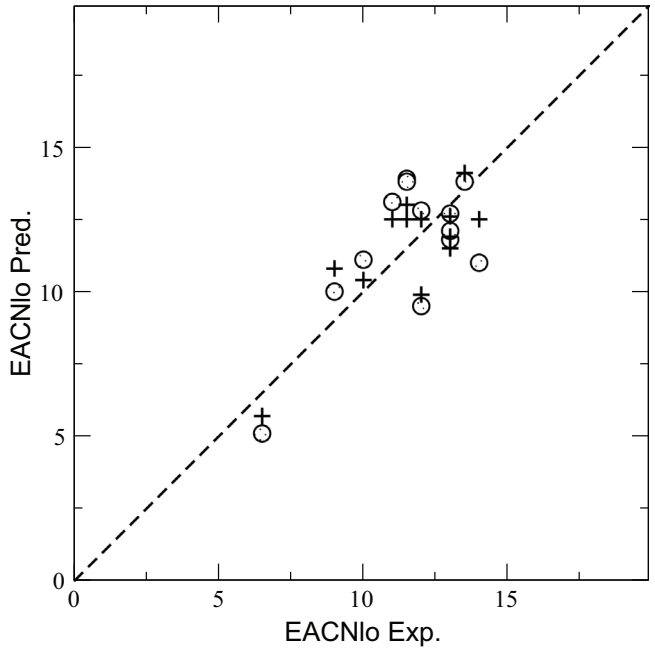


**Fig. 6.** Scatterplots of experimental EACNlo values *versus* predicted EACNlo values. Circles stand for the full predictive approach (the model proposed by Creton and Mougin together with Eq. (5)) while plus symbols denote the use of the model proposed by Creton and Mougin together with experimental EACNdo.

Mougin [14]. Therefore, the proposed combining of the two models appears as a relevant tool to estimate EACN of live crude oils.

### 3.3 Conclusions and perspectives

In the context of EOR, some chemical EOR techniques involve surfactant formulations to mobilize oil trapped by capillary forces. In order to assist and speed up experiments necessary for the formulation design, we recently proposed a model based on thermodynamics to predict EACN of live crude oil. This model consists in a linear mixing rule based on volumetric fractions of the EACN of the dead crude oil and the EACN of the representative gas. The objective of the present work was to use data mining based approaches to investigate and develop relations between the EACN and the composition of dead crude oils.

We collected 29 crude oil samples originating from around the world and performed analysis to obtained compositional information. Each sample has been described in terms of EACNdo, API gravity, and fractions of $C_{20^-}$, saturates, aromatics, resins, and asphaltenes. The database covers a broad range of API gravity with values ranging from 11 to 50 denoting heavy and light crude oils, respectively. Machine learning methods based on EA have been applied to our database in order to generate QPPR to predict EACNdo. In the case of MGGP, we proposed an approach to parameterize GPTIPS. Obtained QPPR models were compared to each other in terms of capacity to generalizing the database. Note that this work could be done using any SARA analysis but fractions of $C_{20^-}$, saturates, aromatics, resins, and asphaltenes within the database must be consistent. The best QPPR model was then used to feed a thermodynamics based model to predict EACNlo for crude oils. Comparisons carried out demonstrate that the proposed combining of the two models appears as a relevant tool for fast and accurate estimates of live crude oil EACNs.

To the best of our knowledge, this work represents the first attempt to predict EACN of crude oils using data mining. When new samples of crude oils will be available, API gravity, and fractions of $C_{20^-}$, saturates, aromatics, resins, and asphaltenes will be experimentally determined. The new samples will be used to supplement our database. The QPPR model (Eq. (5)) developed in this work will be applied to predict EACNdo values for the new crude oils. According to the accuracy of the property predictions and as the QPPR model is more statistical than a physical law, the QPPR model may necessitate an update using MGGP.

*Acknowledgments.* Authors wish to thank the EOR Alliance team for their technical assistance and all relevant discussions.

## References


1 Lu J., Liyanage P.J., Solairaj S., Adkins S., Arachchilage G.P., Kim D.H., Britton C., Weerasooriya U., Pope G.A. (2014) New surfactant developments for chemical enhanced oil recovery, *J. Pet. Sci. Eng.* **120**, 94–101. doi: 10.1016/j.petrol.2014.05.021.

2 Creton B., Nieto-Draghi C., Pannacci N. (2012) Prediction of surfactants' properties using multiscale molecular modeling tools: A review, *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles* **67**, 6, 969–982. doi: 10.2516/ogst/2012040.

3 Salager J.-L., Forgiarini A.M., Bullón J. (2013) How to attain ultralow interfacial tension and three-phase behavior with surfactant formulation for enhanced oil recovery: A review. Part 1. Optimum formulation for simple surfactant–oil–water ternary systems, *J. Surfactants Deterg.* **16**, 4, 449–472. doi: 10.1007/s11743-013-1470-4.

4 Budhathoki M., Hsu T.-P., Lohateeraparp P., Roberts B.L., Shiau B.J., Harwell J.H. (2016) Design of an optimal middle phase microemulsion for ultra high saline brine using Hydrophilic Lipophilic Deviation (HLD) method, *Colloids Surf. A: Physicochem. Eng. Aspects* **488**, 36–45. doi: 10.1016/j.colsurfa.2015.09.066.

5 Salager J.L., Morgan J.C., Schechter R.S., Wade W.H., Vasquez E. (1979) Optimum formulation of surfactant/water/oil systems for minimum interfacial tension or phase behavior, *Soc. Pet. Eng. J.* **19**, 02, 107–115. doi: 10.2118/7054-PA.

6 Acosta E.J., Yuan J.Sh., Bhakta A.Sh. (2008) The characteristic curvature of ionic surfactants, *J. Surf. Deterg.* **11**, 2, 145–158. doi: 10.1007/s11743-008-1065-7.

7 Marliere C., Creton B., Oukhemanou F., Wartenberg N., Courtaud T., Féjean C., Betoulle S., Defiolle D., Mougin P. (2016) Impact of live crude oil composition on optimal salinity of a surfactant formulation, *Paper SPE 179792-MS presented at the SPE EOR Conference at Oil and Gas West Asia*, 21–23 March, Muscat, Oman, (179792-MS). doi: 10.2118/179792-MS.

8 Oukhemanou F., Courtaud T., Morvan M., Moreau P., Mougin P., Fejean C., Pedel N., Bazin B., Tabary R. (2014) Alkaline surfactant-polymer formulation evaluation in live oil conditions: The impact of temperature, pressure and gas on oil recovery performance, *Paper SPE 169130-MS presented at the SPE Improved Oil Recovery Symposium*, 12–16 April, Tulsa, Oklahoma, USA, (169130-MS). doi: 10.2118/169130-MS.

9 Bouton F., Durand M., Nardello-Rataj V., Borosy A.P., Quellet C., Aubry J.-M. (2010) A QSPR model for the prediction of the fish-tail temperature of cie4/water/polar hydrocarbon oil systems, *Langmuir* **26**, 11, 7962–7970. doi: 10.1021/la904836m.

10 Lukowicz T., Benazzouz A., Nardello-Rataj V., Aubry J.-M. (2015) Rationalization and prediction of the equivalent alkane carbon number (EACN) of polar hydrocarbon oils with COSMO-RS $\sigma$-moments, *Langmuir* **31**, 41, 11220–11226. doi: 10.1021/acs.langmuir.5b02545.

11 Lukowicz T., Illous E., Nardello-Rataj V., Aubry J.-M. (2018) Prediction of the equivalent alkane carbon number (EACN) of aprotic polar oils with COSMO-RS $\sigma$-moments, *Colloids Surf. A: Physicochem. Eng. Aspects* **536**, 53–59. doi: 10.1016/j.colsurfa.2017.07.068.

12 Cayias J.L., Schechter R.S., Wade W.H. (1976) Modeling crude oils for low interfacial tension, *Soc. Pet. Eng. J.* **16**, 06, 351–357. doi: 10.2118/5813-PA.

13 Cash L., Cayias J.L., Fournier G., Macallister D., Schares T., Schechter R.S., Wade W.H. (1977) The application of low interfacial tension scaling rules to binary hydrocarbon mixtures, *J. Colloid Interface Sci.* **59**, 1, 39–44. doi: 10.1016/0021-9797(77)90336-8.

14 Creton B., Mougin P. (2016) Equivalent alkane carbon number of live crude oil: A predictive model based on thermodynamics, *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles* **71**, 5, 62. doi: 10.2516/ogst/2016017.

15 Soave G. (1972) Equilibrium constants from a modified Redlich-Kwong equation of state, *Chem. Eng. Sci.* **27**, 6, 1197–1203. doi: 10.1016/0009-2509(72)80096-4.

16 Péneloux A., Rauzy E., Fréze R. (1982) A consistent correction for Redlich-Kwong-Soave volumes, *Fluid Phase Equilib.* **8**, 1, 7–23. doi: 10.1016/0378-3812(82)80002-2.

17 Creton B. (2017) Chemoinformatics at IFP Energies nouvelles: Applications in the fields of energy, transport, and environment, *Mol. Informatics* **36**, 10, 1700028. doi: 10.1016/0009-2509(72)80096-4.

18 Katritzky A.R., Kuanar M., Slavov S., Hall C.D., Karelson M., Kahn I., Dobchev D.A. (2010) Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction, *Chem. Rev.* **110**, 10, 5714–5789. doi: 10.1021/cr900238d.

19 Nieto-Draghi C., Fayet G., Creton B., Rozanska X., Rotureau P., de Hemptinne J.-C., Ungerer P., Rousseau B., Adamo C. (2015) A general guidebook for the theoretical prediction of physicochemical properties of chemicals for regulatory purposes, *Chem. Rev.* **115**, 24, 13093–13164. doi: 10.1021/acs.chemrev.5b00215.

20 Wan W., Zhao J., Harwell J.H., Shiau B.-J. (2016) Characterization of crude oil equivalent alkane carbon number (EACN) for surfactant flooding design, *J. Dispers. Sci. Technol.* **37**, 2, 280–287. doi: 10.1080/01932691.2014.950739.

21 Jewell D.M., Weber J.H., Bunger J.W., Plancher H., Latham D.R. (1972) Ion-exchange, coordination, and adsorption chromatographic separation of heavy-end petroleum distillates, *Anal. Chem.* **44**, 8, 1391–1395. doi: 10.1021/ac60316a003.

22 Kharrat A.M., Zacharia J., Cherian V.J., Anyatonwu A. (2007) Issues with comparing SARA methodologies, *Energy Fuels* **21**, 6, 3618–3621. doi: 10.1021/ef700393a.

23 Behar F., Roy S., Jarvie D. (2010) Artificial maturation of a type I kerogen in closed system: Mass balance and kinetic modelling, *Org. Geochem.* **41**, 11, 1235–1247. doi: 10.1016/j.orggeochem.2010.08.005.

24 Aske N., Kallevik H., Sjöblom J. (2001) Determination of saturate, aromatic, resin, and asphaltenic (SARA) components in crude oils by means of infrared and near-infrared spectroscopy, *Energy Fuels* **15**, 5, 1304–1312. doi: 10.1021/ef010088h.

25 Fan T., Buckley J.S. (2002) Rapid and accurate SARA analysis of medium gravity crude oils, *Energy Fuels* **16**, 6, 1571–1575. doi: 10.1021/ef0201228.

26 Molina D., Uribe U.N., Murgich J. (2010) Correlations between SARA fractions and physicochemical properties with 1H NMR spectra of vacuum residues from colombian crude oils, *Fuel* **89**, 1, 185–192. doi: 10.1016/j.fuel.2009.07.021.

27 Chamkalani A. (2012) Correlations between SARA fractions, density, and RI to investigate the stability of asphaltene, *ISRN Anal. Chem.* **2012**, 219276. doi: 10.5402/2012/219276.

28 Mohan Sinnathambi C., Mohamad Nor N. (2012) Relationship between SARA fractions and crude oil fouling, *J. Appl. Sci.* **12**, 23, 2479–2483. doi: 10.3923/jas.2012.2479.2483.

29 Ashoori S., Sharifi M., Masoumi M., Mohammad Salehi M. (2017) The relationship between SARA fractions and crude oil stability, *Egypt. J. Pet.* **26**, 1, 209–213. doi: 10.1016/j.ejpe.2016.04.002.

30 Weigel S., Stephan D. (2018) Relationships between the chemistry and the physical properties of bitumen, *Road Mater. Pavement Des.* **19**, 7, 1636–1650. doi: 10.1080/14680629.2017.1338189.

31 Materials Studio. version 7.0, Accelrys Software Inc.: San Diego, USA, 2014

32 Gramatica P. (2007) Principles of QSAR models validation: Internal and external, *QSAR Comb. Sci.* **26**, 5, 694–701. doi: 10.1002/qsar.200610151.

33 Kuei Lin L.I. (1989) A concordance correlation coefficient to evaluate reproducibility, *Biometrics* **45**, 1, 255–268.

34 Chirico N., Gramatica P. (2011) Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, *J. Chem. Inf. Model.* **51**, 9, 2320–2335. doi: 10.1021/ci200211n.

35 Chirico N., Gramatica P. (2012) Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection, *J. Chem. Inf. Model.* **52**, 8, 2044–2058. doi: 10.1021/ci300084j.

36 Searson D.P., Leahy D.E., Willis M.J. (2010) GPTIPS: an open source genetic programming toolbox for multigene symbolic regression, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2010 (IMECS 2010)*, 17–19 March, Hong Kong, pp. 77–80.

37 Searson D.P. (2015) Chapter GPTIPS 2: an open-source software platform for symbolic data mining, in: *Handbook of genetic programming applications*, Gandomi A.H., Alavi A.H., Ryan C. (eds), Springer International Publishing, New York, NY, pp. 551–573.

38 Gandomi A.H., Alavi A.H., Ryan C. (2015) *Handbook of genetic programming applications*, Springer International Publishing, New York, NY.

39 Tropsha A., Gramatica P., Gombar V.K. (2003) The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR Comb. Sci.* **22**, 1, 69–77. doi: 10.1002/qsar.200390007.

40 Garg A., Garg A., Tai K. (2014) A multi-gene genetic programming model for estimating stress-dependent soil water retention curves, *Comput. Geosci.* **18**, 1, 45–56. doi: 10.1007/s10596-013-9381-z.

41 Mohamadi-Baghmolaei M., Azin R., Sakhaei Z., Mohamadi-Baghmolaei R., Osfouri S. (2016) Novel method for estimation of gas/oil relative permeabilities, *J. Mol. Liq.* **224**, 1109–1116. doi: 10.1016/j.molliq.2016.08.055.

42 Garg A., Garg A., Tai K., Sreedeep S. (2014) Estimation of factor of safety of rooted slope using an evolutionary approach, *Ecol. Eng.* **64**, 314–324. doi: 10.1016/j.ecoleng.2013.12.047.