CMST 22(3) 133-141 (2016) DOI:10.12921/cmst.2016.0000022

Quantitative Modeling of Physical Properties of Crude Oil Hydrocarbons Using Volsurf^{+} Molecular Descriptors

Saadi Saaidpour*, **Faraidon Ghaderi**

Department of Chemistry, Faculty of Science
Sanandaj Branch, Islamic Azad University, Sanandaj, Iran
**E-mail: sasaaidpour@iausdj.ac.ir*

Received: 29 April 2016; revised: 18 July 2016; accepted: 26 July 2016; published online: 24 August 2016

Abstract: The quantitative structure-property relationship (QSPR) method is used to develop the correlation between structures of crude oil hydrocarbons and their physical properties. In this study, we used VolSurf+ descriptors for QSPR modeling of the boiling point, Henry law constant and water solubility of eighty crude oil hydrocarbons. A subset of the calculated descriptors selected using stepwise regression (SR) was used in the QSPR model development. Multivariate linear regressions (MLR) are utilized to construct the linear models. The prediction results agree well with the experimental values of these properties. The comparison results indicate the superiority of the presented models and reveal that it can be effectively used to predict the boiling point, Henry law constant and water solubility values of crude oil hydrocarbons from the molecular structures alone. The stability and predictivity of the proposed models were validated using internal validation (leave one out and leave many out) and external validation. Application of the developed models to test a set of 16 compounds demonstrates that the new models are reliable with good predictive accuracy and simple formulation.

Key words: boiling point, water solubility, Henry's law constant, crude oil hydrocarbons, volsurf+ descriptors

I. INTRODUCTION

The aim of this work is to obtain Quantitative Structure-Property Relationship (QSPR) models of three physicochemical properties: boiling point, Water solubility and Henry's law constant, for a set of 80 Crude oil hydrocarbons, a special class of chemicals that has been of concern to the scientific community due to their pollutant potential.

The boiling point (BP) is one of the main physicochemical properties used to characterize and identify compounds. The BP is the temperature at which a liquid boils at 1 atmosphere of pressure and an indication of attractive forces between the molecules. These intermolecular forces are directly related to the structure of the compound, and hence the BP may be correlated to the structure. The BP of a compound is an important property for the simulation of processes in chemical and petroleum industries. With the increased need of reliable data for optimization of industrial processes, it is important to develop Quantitative Structure-Property Relationship (QSPR) models for the estimation of normal BP for compounds that are not yet synthesized or whose BP is unknown.

Numerous QSPR models for calculating the BPs of organic compounds have been introduced using various numerical descriptors of a chemical structure [1-11].

Henry's law is one of the gas laws formulated by William Henry in 1803 and is defined as the amount of a given gas that dissolves in a given type and volume of liquid is directly proportional to the partial pressure of that gas in equilibrium with that liquid. In other words, the Henry's law constant (H) is as a ratio partial pressure in the vapor on the concentration in the liquid. Several papers are published about the prediction and modeling of H [12-16]. As the air-water partition coefficient, H represents a key physical property of a compound with respect to its distribution and fate in the environment as well as to the applicability of potential treatment methods such as air-stripping for treatment of contaminated ground water. The estimation methods for H for environmental purposes can be categorized as (1) property-

property relationships (PPR) methods; (2) bond and group contribution methods; (3) continuum-solvation methods; (4) UNIFAC (universal quasi-chemical functional group activity coefficient) and structural, quantum chemical or physico-chemical descriptor- based QSPR methods. The most well-known PPR is the VP/AS (vapor pressure/ aqueous solubility) method [17].

The aqueous solubility (Sw) of organic compounds is an important molecular property, playing a vital role in the behavior of compounds in many areas of interest. The importance of solubility of water in crude oil will increase in view of processing, safety, hazard, and environmental considerations focusing on product quality and equipment sustainability. Any processing that lowers temperatures to near the freezing point of water may result in formation of solids (freezing of water or hydrate formation). Such formation will affect both fluid flows in piping and operational characteristics of equipment. For catalytic reactions, any water in the hydrocarbon may poison the catalyst that promotes the hydrocarbon reaction. For reactions in general, any water in the reaction species may result in formation of undesirable by-products issuing from the hydrocarbon reaction. The solubility of a substance is the amount of substance that will dissolve in a given amount of solvent. Solubility is a quantitative term that depends on the physical and chemical properties of the solute and solvent as well as on temperature, pressure. The production of gas and oil is often accompanied by water; this water at the top of the pipe becomes saturated with acid gases and corrodes the pipe. Corrosion control in oil and gas production is carried out using corrosion inhibitors. The first step in formulating corrosion inhibitors is determining the solubility and other factors [18-21]. The importance of the water solubility in crude oil will increase in view of processing, safety, hazard, and environmental considerations focusing on product quality and equipment sustainability. Numerous QSPR models for prediction the Sw of organic compounds have been introduced using various molecular descriptors of chemical structure [22-26].

In our previous papers we reported on the application of QSPR techniques in developing a new, simplified approach to prediction of organic compounds properties using different models [27-36].

The purpose of this study is to develop QSPR models for the estimation of boiling points (BP), Henry law constant (H) and water solubility (Sw) of crude oil hydrocarbons using the VolSurf+ program. In this study we present new QSPR models for prediction of the BP, logH and logSw of various crude oil hydrocarbons. A stepwise regression (SR) and multiple linear regression (MLR) procedure were used to select relevant descriptors and mathematical modeling. Also, in this work we applied back propagation neural network (BPNN) and support vector machine regression (SVMR) on this data set, but no significant difference between results with the MLR method, so we preferred to report on results of the MLR method. The predictive power of the resulting model is demonstrated by testing them on unseen data that were not used during model generation. A physicochemical interpretation of the selected descriptors is also given.

II. DATA AND METHODS

The QSPR models for the estimation of the boiling point, Henry law constant and water solubility of various crude oil hydrocarbons are established in the following six steps: the molecular structure input and generation of the files containing the chemical structures is stored in a computer-readable format; quantum mechanics geometry is optimized with a semi-empirical (AM1) method; molecular descriptors are computed; molecular descriptors are selected; and the molecular descriptors-BP, H and logSw models are generated by the multi-linear regression analysis (MLR), and statistical approval techniques and prediction analysis.

II. 1. Experimental Data

The total data set of the boiling point (Kelvin), Henry's law constant (atm mol^{-1} frac^{-1}) and water solubility (ppm (wt)) in crude oil collected from the Handbook of physical properties for Hydrocarbons and chemicals [37]. For evaluating the predictive capability of the proposed model, before model generation both datasets were split into a training set ($\sim$80% of compounds), used for model development, and a prediction set ($\sim$20% of compounds), used for external validation. The training set was used to adjust the parameters of the SR-MLR and the test set was used to evaluate its prediction ability. Hydrocarbons are partitioned randomly

Tab. 1. Experimental data of boiling point, Henry law constant and water solubility of crude oil hydrocarbons

No	Objects	Formula	Name	Case no.	BP(K)	logH (atm/mol frac)	logSw (ppm-wt)
1	Object 1	C5H12	pentane	109-66-0	309.22	1.889302	2.004192
2	Object 2	C5H12	isopentane	78-78-4	300.99	1.911317	1.982181
3	Object 3	C5H12	neopentane	463-82-1	282.65	1.893651	1.999826
4	Object 4	C6H14	hexane	110-54-3	341.88	1.862251	1.954098
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
80	Object 80	C18H38	octadecane	593-45-3	589.86	1.644242	1.701827

into a training set (64 hydrocarbons) and a test set (16 hydrocarbons). A complete list of the compound names and corresponding experimental properties are given as Tab. 1.

II. 2. Molecular Modeling and Descriptor Generation

All numerical calculations have been performed by a computer with Intel Core^{TM} i7 processor and 6 Gb RAM characteristics. The ChemDraw Ultra version 15.0 (ChemOffice 2015, CambridgeSoft Corporation; Cambridge, MA) software was used for drawing the molecular structures[38]. The optimizations of molecular structures were done by the HyperChem 8.0 (Hypercube, Inc., Gainesville, 2011) using AM1 method[39], and descriptors were calculated by VolSurf+ (Molecular Discovery Ltd., 2008) Version 1.0.4 software[40]. The models have been developed by multiple linear regression (MLR) using the ordinary least squares (OLS) method and the stepwise regression have been applied for variable selection using the in-house software for QSPR modeling, Molegro Data Modeller (MDM 2011.2.6.0) [41].

VolSurf+ is an advanced computational procedure aimed to produce and explore the physicochemical property space of a molecule (or library of molecules) starting from 3D maps of interaction energies between the molecule and chemical probes (GRID based Molecular Interaction Fields, or MIFs).

Interaction fields with a water probe (OH2), a hydrophobic probe (DRY) plus an H-bond donor (NH) and an H-bond acceptor (=O) probes are calculated all around the target molecules as in the program GRID. The basic concept of VolSurf+ is to compress the information present in 3D maps into a few quantitative numerical that is very simple to understand and to interpret. The molecular descriptors obtained refer to molecular size and shape or the originality of VolSurf+ resides in the fact that surface, volume and other related descriptors can be directly obtained from three dimensional molecular fields with simple computation algorithms [42-44]. In this study, VolSurf+ software was used to generate many descriptors (128 descriptors) by H_2O, DRY and other probe characterize structural properties.

II. 3. Descriptor Selection

The selection descriptor is important to construct a predictive model. In the work, the stepwise multiple linear regression was used as the feature selection method to select the best calculated descriptors. Stepwise regression is the most known subset descriptor selection methods. Stepwise combines the forward selection and backward elimination. Forward selection begins with one variable and continues to add variable at a time until no further improvement is possible. Backward elimination begins all variable available and repeatedly removes variable until no move important is possible. Stepwise regression methods are basically a forward selection procedure that a descriptor entered the model in the earlier stages of selection may be elimination at the later stages [45-47], but at each stage the possibility of eliminating a variable, as in backward elimination. In this work, with the stepwise regression method for each property two descriptors were selected, that has high correlation to the dependent variable and used to build the models. We selected for each property (dependent variable) two descriptors and models were constructed by using them.

II. 4. MLR Modeling

The general purpose of multiple linear regression (MLR) is to model the relationship between two or more independent variables and a dependent variable by fitting a linear equation to observed data. Every value of the independent variable X is associated with a value of the dependent variable Y. Formally, the model for multiple linear regression, given n observations, is

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n, \qquad (1)$$

where in the presented study, the dependent variable Y is the BP, logH and logSw property, $X_1 - X_n$ represents the specific descriptor, while $a_1 - a_n$ represents the coefficients of those descriptors, and a_0 is the intercept of the equation. A detailed description of theories of MLR can be found in the literature [48, 49].

II. 4. 1. Model Validation

In order to estimate the predictive power of a QSAR/QSPR model, it can be conveniently estimated by statistical parameters. Reliability of the proposed method was explored using the cross-validation methods. In this study we applied three most well-known validation tools: external and internal validation, and a randomization test.

II.4.1.1. Internal and external validation

In the constructed model internal validation is usually done by leave-one-out (LOO) or leave-many-out (LMO) procedures [50]. The Q^2 is quality of prediction, if in the model squared correlation coefficient of the training set (R^2) increased artificially by adding more descriptors whereas Q^2decreases in such over-fitting conditions. During the leave-one-out (LOO) procedure by elimination each time one data from the training set and a new model is constructed without this data. The building model and leaving out is continued until predicted all data. The new QSPR models are expected to have low $\mathrm{R}^2{}_{cal}$ and LOO-cross-validation ($Q^2{}_{loo}$) values. The leave-many-out (LMO-CV) in comparison with LOO-CV is stronger and LMO-CV is more reliable [51, 52]. In the LMO-CV by removing each time more one data from the training set (leave 10 out) and constructed model. It is suggested for big datasets. The leave-one-out cross-validation ($Q^2{}_{\mathrm{LOO}}$) (or $Q^2{}_{\mathrm{LMO}}$) was calculated by the following equation:

$$Q^2_{\mathrm{LOO}} \text{ or } Q^2_{\mathrm{LMO}} = 1 - \frac{\sum_{i=1}^{\mathrm{training}} \left(Y_i - \overline{Y}_i\right)^2}{\sum_{i=1}^{\mathrm{training}} \left(Y_i - \overline{Y}\right)^2} \qquad (2)$$

Tab. 2. Experimental, descriptors, predicted and residual data for train(64 compounds) and test (16 compounds) sets of BP

No	Objects	BP(exp)	logVP	MW	BP(pred)	Residual
1	Object 1	309.22	2.71012	72.1488	316.09	-6.87
2	Object 3	282.65	3.11394	72.1488	300.561	-17.911
3	Object 4	341.88	2.17609	86.1754	343.107	-1.227
4	Object 5	322.88	2.49136	86.1754	330.983	-8.103
⋮	⋮	⋮	⋮	⋮	⋮	⋮
64	Object 79	575.3	-2.48945	240.468	593.806	-18.506
1	Object 2	300.99	2.83759	72.1488	311.188	-10.198
⋮	⋮	⋮	⋮	⋮	⋮	⋮
16	Object 80	589.86	-2.83565	254.494	613.6	-23.74

where Y_i, $\widehat{Yi}$ and $\bar{Y}$are the experimental, predicted, and averaged (over the entire training dataset) values of the samples in the training set.

$$Q_{ext}^2 = 1 - \frac{\sum_{i=1}^{test}\left(Y_i - \overline{Y}_i\right)^2}{\sum_{i=1}^{test}\left(Y_i - \overline{Y}\right)^2}, \tag{3}$$

where Y_i and $\hat{Y}i$ are experimental and predicted values of the test set, respectively. The other useful parameters named squared correlation coefficient (R^2) and root mean-squared error (RMSE) were also employed to evaluate the performance of developing models, which are important indicators for linear correlation between predicted and experimental data. They characterize an ability of the model to reproduce quantitatively the experimental data. R^2 is an indicator that measures the linear correlation degree between one variable and another. RMSE indicates the dispersion degree of the random error, which summarizes the overall error of the model.

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n}\left(Y_{i,\mathrm{pred}} - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_{i,\mathrm{exp}} - \overline{Y}\right)^2}. \tag{4}$$

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_{i,\mathrm{exp}} - Y_{i,\mathrm{pred}}\right)^2\right]^{0.5}, \tag{5}$$

where $Y_{i,\mathrm{exp}}$ is the experimental property in the sample *i*, $Y_{i,\mathrm{pred}}$ represented the predicted property in the sample *i*, $\bar{Y}$is the mean of experimental property in the prediction set and n is the total number of samples in the prediction set.

II.4.1.2. Randomization Test

Randomization test (y-randomization or y-scrambling) is a technique to protect them against the risk of chance correlation [53]. This technique ensures stableness of the QSAR/QSPR model. The randomization test suggests that whenever a model has been trained on a dataset, the same procedure should be applied to a data set where the order of the dependent variable has been randomized. To exclude the possibility of chance correlation between modeling descriptors and the response, the Y-Scrambling method has been applied, which verifies the fitting of the model developed on randomly re-ordered responses (2000 scrambling iterations); where a low value of the averaged R^2 scrambled ($\mathrm{R}^2{}_{ys}$) is indicative of a well- founded original model.

Tab. 3. Experimental, descriptors, predicted and residual data for train (64 compounds) and test (16 compounds) sets of logH

No	Objects	logH(exp)	logVP	CW2	logH(pred)	Residual
1	Object 1	1.8893	2.71012	0.139704	1.88985	-0.00055
2	Object 2	1.91132	2.83759	0.122463	1.90076	0.01056
3	Object 4	1.86225	2.17609	0.194731	1.85248	0.00977
4	Object 6	1.86225	2.35603	0.179248	1.86359	-0.00134
⋮	⋮	⋮	⋮	⋮	⋮	⋮
64	Object 80	1.64424	-2.83565	0.440419	1.63303	0.01121
1	Object 3	1.89365	3.11394	0.116121	1.90658	-0.01293
⋮	⋮	⋮	⋮	⋮	⋮	⋮
16	Object 73	1.6695	-1.36051	0.428491	1.66965	-0.00015

III. RESULTS AND DISSCUSIONS

III. 1. Model Analysis

The MLR analysis has been carried out to derive the best QSPR model. The MLR technique was performed on the molecules of the training set. After regression analysis, a few suitable models were obtained among which the best model was selected and presented in equations 6, 7, and 8. MLR analysis provided a useful equation that can be used to predict the BP, logH and logSw of crude oil hydrocarbons based VolSurf+ descriptors.

$$\begin{gathered} BP = -38.45\,(\pm\,8.57)\,\mathrm{logVP} \\ +0.46\,(\pm\,0.25)\,\mathrm{MW} + 386.97\ (\pm\,41.55) \\ n = 64, \mathrm{R}^2 = 0.9938, s = 6.04, \\ F = 4896.52, Q^2 = 0.9927 \end{gathered} \tag{6}$$

$$\begin{gathered} \mathrm{logH} = +0.02\,(\pm 0.01)\,\mathrm{logVP} \\ -0.48\,(\pm\,0.12)\,\mathrm{CW2} + 1.90\,(\pm\,0.04) \\ n = 64, \mathrm{R}^2 = 0.9731,\ s = 0.011, \\ F = 1105.27,\ \mathrm{Q}^2 = 0.9701 \end{gathered} \tag{7}$$

$$\begin{gathered} \mathrm{logSw} = -0.29\,(\pm 0.1019)\,\mathrm{R} \\ -0.038\,(\pm 0.0021)\,\mathrm{logP} + 2.46\,(\pm\,0.13) \\ n = 64, \mathrm{R}^2 = 0.9804,\ s = 0.01, \\ F = 1524.18,\ \mathrm{Q}^2 = 0.9762. \end{gathered} \tag{8}$$

MW is the molecular mass, logVP is the logarithm of vapor pressure, CW2 (capacity factor) is the hydrophilic volume per surface unit, R is the ratio of volume/surface, and logP is the n-octanol/water partition coefficient. The statistical terms are the number of molecules used to calculate the regression (n), squared correlation coefficient (R^2), standard error (s), F statistic (F), and the Q^2 is the squared correlation coefficient of leave one out cross validation. Positive values of the regression coefficients indicate that the indicated descriptor contributes positively to the value of variable property, whereas negative values indicate that the greater the value of the descriptor the lower the value of variable property.

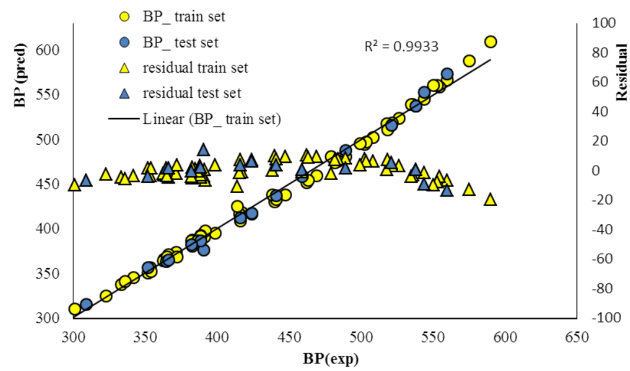

Fig. 1. Experimental, predicted and residual of boiling points for train and test sets

The plot of predicted BP, logH and logSw versus experimental BP, logH and logSw and the residuals (experimental-predicted) versus experimental values, obtained by the SR-MLR modeling, and the random distribution of residuals about zero means are shown in Fig. 2 and 3, respectively.

In Tables 2, 3 and 4 the results of the BP, logH and logSw experimental, predicted and related descriptors of training and test sets are shown, respectively. In Tab. 5. results of the statistical data are shown. The statistical parameters of the model are satisfying and prove that the MLR model is stable, robust and predictive. In addition, the low value of $\mathrm{R}^2{}_{Y\_\mathrm{scrambling}}$ and the high value of $\mathrm{RMS}D_{Y\_\mathrm{scrambling}}$ of randomization test indicating that the obtained models have no chance correlations.

III. 2. Interpretation of Descriptors

The QSPR model of equation 6 developed indicated that vapor pressure of compound at 25°C (logVP) and molecular mass (MW) significantly influence hydrocarbons normal boiling points. Vapor pressure (logVP) is the pressure of a vapor

Tab. 4. Experimental, descriptors, predicted and residual data for train(64 compounds) and test (16 compounds) sets of logSw

No	Objects	logSw(exp)	Rugosity (R)	logP o/w	logSw(pred)	Residual
1	Object 1	2.00419	1.27768	2.765	1.97649	0.0277
2	Object 2	1.98218	1.31676	2.614	1.97068	0.0115
3	Object 4	1.9541	1.33188	3.256	1.9418	0.0123
4	Object 6	1.9541	1.37371	2.954	1.94092	0.01318
⋮	⋮	⋮	⋮	⋮	⋮	⋮
64	Object 80	1.70183	1.42749	9	1.69509	0.00674
1	Object 3	1.99983	1.31262	2.479	1.97704	0.02279
⋮	⋮	⋮	⋮	⋮	⋮	⋮
16	Object 73	1.75504	1.42764	7.524	1.75118	0.00386

Tab. 5. Statistical parameters obtained by using Molegro Data Modeller software for MLR models

logSw	logH	BP	Statistical parameters
$R^2 = 0.9804$	$R^2 = 0.9760$	$R^2 = 0.9938$	Squared correlation coefficient (training set)
RMSD = 0.0101	RMSD = 0.0111	RMSD = 5.8971	Root Mean Squared Deviation (training set)
$Q^2{}_{\mathrm{Loo}} = 0.9762$	$Q^2{}_{\mathrm{Loo}} = 0.9701$	$Q^2{}_{\mathrm{Loo}} = 0.9927$	Squared Correlation coefficient LOO-CV
RMSD = 0.0111	RMSD = 0.0117	RMSD = 6.4058	Root Mean Squared Deviation (LOO-CV)
$Q^2{}_{\mathrm{LMO}} = 0.9757$	$Q^2{}_{\mathrm{LMO}} = 0.9697$	$Q^2{}_{\mathrm{LMO}} = 0.9924$	Squared Correlation coefficient LMO-CV
RMSD = 0.0112	RMSD = 0.0118	RMSD = 6.5077	Root Mean Squared Deviation (LMO-CV)
RMSD = 0137	RMSD = 0.0131	RMSD = 10.8742	Root Mean Squared Deviation (test set)
$Q^2{}_{\mathrm{Ext}} = 0.9735$	$Q^2{}_{\mathrm{Ext}} = 0.9720$	$Q^2{}_{\mathrm{Ext}} = 0.9833$	Squared correlation coefficient (test set)
$R^2{}_{y-scr} = 0.0108$	$R^2{}_{y-scr} = 0.0194$	$R^2{}_{y-scr} = 0.0418$	Squared correlation coefficient (Y-scrambling)
RMSD = 0.0716	RMSD = 0.0670	RMSD = 73.36	Root Mean Squared Deviation (Y-scrambling)

in thermodynamic equilibrium with its condensed phases in a closed container. The boiling point of a substance is the temperature at which the vapor pressure of the liquid equals the pressure surrounding the liquid. The vapor pressure is the pressure exerted by vapor molecules on a solid/liquid surface with which it is in a state of equilibrium, which means that as long as there is equilibrium, vapor molecules enter the liquid phase and liquid molecules enter the vapor phase.

temperature. The second descriptor is molar mass (MW). Among the size descriptors, molar mass is the simplest and most commonly used molecular 0D-descriptor, calculated as the sum of the atomic masses of all the atoms in a molecule. It is related to molecular size and is atom-type sensitive. It is defined as $\mathrm{MW} = \sum_{i=1}^{A} m_i$ where m is the atomic mass and i runs over the A atoms of the molecule. By increasing molecular mass of compounds the BP increases.

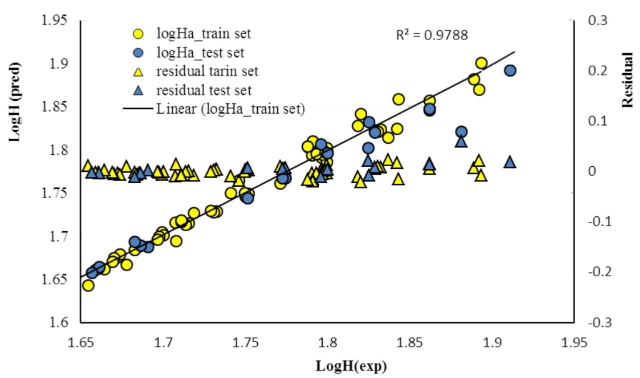

Fig. 2. Experimental, predicted and residual of logH for train and test sets

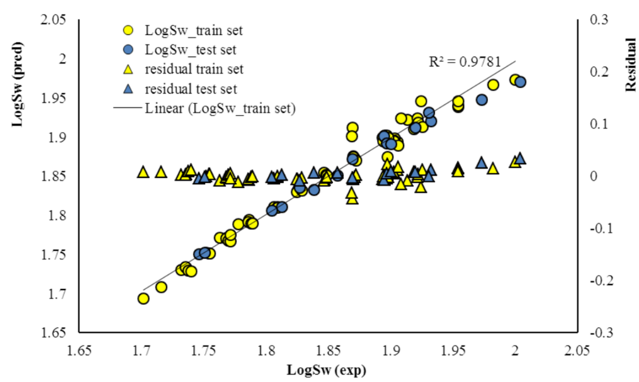

Fig. 3. Experimental, predicted and residual of logSw for train and test sets

If the intermolecular forces between the liquid molecules are strong it will not easily leave the liquid phase and hence reduce the vapor pressure and consequently its boiling point will be higher. If the intermolecular forces between the liquid molecules are weak, they will easily leave the liquid phase and enter the vapor phase and hence have low boiling points. The boiling point of hydrocarbons with high molecular weight can be increased. The boiling points of straight chain alkanes are related to the number of carbon atoms in their molecules. Increased intermolecular attractions are related to the greater molecule-molecule contact possible for larger alkanes. The boiling point downwards due to branched hydrocarbons is due to the spherical surface molecule that reduced the surface size and intermolecular forces become weaker and boil at a low

The larger the molar mass, the greater the polarizability of the molecule and hence also the van der Waals attractive forces between near neighbors. Increasing molecular mass leads to increasing the boiling point of hydrocarbons. However, it should be noted that substances of high molecular mass evaporate more slowly than similar substances of low molecular mass. The compounds with the highest vapor pressures have the lowest normal boiling points. The developed QSPR of equation 7 showed that the vapor pressure (logVP) and capacity factor (CW2) descriptors significantly influence the Henry law constant of hydrocarbons. The relationship between the vapor pressure and Henry's law constant is directly that by increasing the logarithm vapor pressure increases logarithm Henry's law constant. The capacity factor

(CW2) is the hydrophilic volume per surface unit that by decreasing capacity factor, logH increases. So, with increase of vapor pressure and decrease of hydrophilic property (decrease water solubility) of compounds, Henry law constant increases. In the equation 8, two parameters of Rugosity (R) and log$P_{o/w}$are effective at prediction of logSw. The rugosity is a measure of molecular wrinkled surface; it represents the ratio of volume/surface. The smaller the ratio, the larger the rugosity. With increased rugosity (decreases volume/surface ratio), water solubility decreases. The *n*-octanol-water partition coefficient, respectively, its logarithmic value is called log$P_{o/w}$. The log$P_{o/w}$is defined as the ratio of the concentration of a chemical in n-octanol and water at equilibrium at a specified temperature. The typical quantitative descriptor of lipophilicity is the logPo/w of a given compound between two immiscible solvents. The logPo/w is frequently used as a measure of the lipophilic character of the molecules and molecular hydrophobicity. With increased octanol/water partition coefficients, water solubility decreases. This brief discussion indicates that solubility of water in hydrocarbons contained in crude oil is important in engineering applications involving processing, safety, hazard, and environmental considerations.

IV. CONCLUSION

Prediction of the boiling point, Henry's law constant and water solubility are important properties of oil and gas industry. In this study, we use calculation molecular descriptors from 3D molecular fields of interaction energies with physicochemical properties. VolSurf+ descriptors are easy to interpret. The MLR method was used for QSPR modeling of physical properties of 80 hydrocarbons in crude oil. MLR analysis provided useful equations that can be used to predict the BP, logH and logSw of hydrocarbons based upon logVP, MW, CW2, Rugosity and logPo/w parameters. The results indicated that a strong correlation exists between the experimental and predicted properties of compounds. The obtained molecular descriptors are effective and meaningful. The results are usable in engineering applications involving processing, safety, hazard, and environmental considerations.

References

[1] L.M. Egolf, P.C. Jurs, *Prediction of boiling points of organic heterocyclic compounds using regression and neural network techniques*, J. Chem. Inf. Comp. Sci., **33**, 616-625(1993).

[2] L.H. Hall, L.B. Kier, *Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information*, J. Chem. Inf. Comput. Sci., **35**, 1039-1045 (1995).

[3] O. Ivanciuc, T. Ivanciuc, A.T. Balaban, *Quantitative structure-property relationship study of normal boiling points for halogen-/ oxygen-/ sulfur-containing organic compounds using the CODESSA program*, Tetrahedron, **54**, 9129-9142(1998).

[4] D. Plavsic, N. Trinajstic, D. Amic, et al., *Comparison between the structure-boiling point relationships with different descriptors for condensed benzenoids*, New J. Chem., **22**, 1075-1078 (1998).

[5] D. Sola, A. Fer, M. Banchero, L. Manna, S.Sicardi, *QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method*, Fluid Phase Equilib., **263**(1), 33-42 (2008).

[6] D. Yi-min , Z.Z. ping , Z.Z. Cao, Y.F. zhang, J.I. Zeng, X. Li, *Prediction of boiling points of organic compounds by QSPR tools*, J. Mol. Graph. Model., **44**,113-119 (2013).

[7] D. Abooali , M.A. Sobati , *Novel method for prediction of normal boiling point and enthalpy of vaporization at normal boiling point of pure refrigerants: A QSPR approach*, Int. J. Refrig., **40**, 282-293(2014).

[8] I. Oprisiu, G. Marcou , D. Horvath, D.B. Brunel, F. Rivollet, A.Varnek, *Publicly available models to predict normal boiling point of organic compounds*, Thermochim. Acta, **553**, 60-67 (2013).

[9] K. Panneerselvam, C.V.S. Brahmmananda Rao, M.P. Antony, *Correlation of normal boiling points of dialkylalkyl phosphonates with topological indices on the gas chromatographic retention data*, Thermochim. Acta, **600**, 77-81(2015).

[10] J. Ghasemi, S. Saaidpour, *Artificial Neural Network Based Quantitative Structural Property Relationship for Predicting Boiling Points of Refrigerants*, QSAR Comb. Sci., **28**, 1245-1254 (2009).

[11] S. Saaidpour, A. Bahmani, A. Rostami, *Prediction the Normal Boiling Points of Primary, Secondary and Tertiary Liquid Amines from their Molecular Structure Descriptors*, CMST, **21**(4), 201-210 (2015).

[12] M. Goodarzi , E.V. Ortiz, L.D.S. Coelho, P.R. Duchowicz, *Linear and non-linear relationships mapping the Henry's law parameters of organic pesticides*, Atmos. Environ., **44**(26), 3179-3186 (2010).

[13] P.R. Duchowicz , J.C.M. Garro, E.A. Castro, *QSPR study of the Henry's Law constant for hydrocarbons*, Chemom. Intell. Lab. Sys., **91**(2), 133-140 (2008).

[14] H. Modarresi , H. Modarress, J.C. Dearden, *QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach*, Chemosphere, **66**(11), 2067-2076 (2007).

[15] D.R. O'Loughlin, N.J. English, *Prediction of Henry's Law Constants via group-specific quantitative structure property relationships*, Chemosphere, **127**, 1-9 (2015).

[16] S. Sahoo, S. Patel, B.K. Mishra, *Quantitative structure property relationship for Henry's law constant of some alkane isomers*, Thermochim. Acta, **512** (1-2), 273-277 (2011).

[17] D. Mackay, W.S. Shiu, K.C. Ma, *Henry's law constant*. In: R.S. Boethling, D. Mackay, (Eds.), *Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences*. Lewis, Boca Raton, FL, USA, pp. 69-87, 2000.

[18] A. Chapoy, A.H. Mohammadi, D. Richon, B. Tohidi, *Gas solubility measurement and modeling for methane-water and methane-ethane-n-butane-water systems at low temperature conditions*, Fluid Phase Equilib., **220**,113-121(2004).

[19] J.H. Gary, G.E. Handwerck, *Petroleum Refining Technology and Economics*, 2001.

[20] S. Mokhatab, W.A. Poe, J.G. Speight, *Handbook of Natural Gas Transmission and Processing*, 2006.

[21] A. Chapoy, S. Mokraoui, A. Valts, D. Richon, A.H. Mohammadi, B. Tohidi, *Solubility measurement and modeling for the system propane-water from 277.62 to 368.16 K*, Fluid Phase Equilib., **226**, 213-220 (2004).

[22] J. Ghasemi, S. Saaidpour, *QSPR prediction of aqueous solubility of drug-like organic compounds*, Chem. Pharm. Bull., **55**(4), 669-674 (2007).

[23] A.R. Katritzky, L. Mu, *A QSPR Study of the Solubility of Gases and Vapors in Water*, J. Chem. Inf. Comput. Sci., **36** (6), 1162-1168 (1996).

[24] P.R. Duchowicz, A. Talevi, C. Bellera, L.E.B. Blanch, E.A. Castro, *Application of descriptors based on Lipinski's rules in the QSPR study of aqueous solubilities*, Bioorgan. Med. Chem., **15**, 3711-3719 (2007).

[25] P.D.T. Huibers, A.R. Katritzky, *Correlation of the Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure*, J. Chem. Inf. Comput. Sci., **38**, 283-292 (1998).

[26] P.V. Khadikar, D. Mandloi, A.V. Bajaj, Sh. Joshi, *QSAR Study on Solubility of Alkanes in Water and Their Partition Coefficients in Different Solvent Systems Using PI Index*, Bioorg. Med. Chem. Lett., **13**, 419-422 (2003).

[27] J. Ghasemi, S. Saaidpour, *Quantitative structure-property relationship study of n-octanol-water partition coefficients of some of diverse drugs using multiple linear regression*, Anal. Chim. Acta, **604**, 99-106 (2007).

[28] J. Ghasemi, S. Saaidpour, S.D. Brown, *QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis*, J. Mol. Struct. (Theochem), **805**, 27-32 (2007).

[29] J. Ghasemi, S. Saaidpour, *QSPR modeling of stability constants of diverse 15-crown-5 ethers complexes using best multiple linear regression*, J. Incl. Phenom. Macrocycl. Chem., **60**(3), 339-351 (2008).

[30] J. Ghasemi, S. Saaidpour, *QSRR prediction of the chromatographic retention behavior of painkiller drugs*, J. Chromatogr. Sci., **47**(2), 156-163 (2009).

[31] S. Saaidpour, *Prediction of drug lipophilicity using back propagation artificial neural network modeling*, Orient. J. Chem., **30**(2), 793-802 (2014).

[32] S. Saaidpour, *Prediction of the Adsorption Capability onto Activated Carbon of Liquid Aliphatic Alcohols using Molecular Fragments Method*, Iranian J. Math. Chem., **5**(2), 127-142 (2014).

[33] S. Saaidpour, S.A. Zarei, F. Nasri, *QSPR study of molar diamagnetic susceptibility of diverse organic compounds using multiple linear regression analysis*, Pak. J. Chem., **2**(1), 6-17 (2012).

[34] S. Saaidpour, *Quantitative Modeling for Prediction of Critical Temperature of Refrigerant Compounds*, Phys. Chem. Res., **4**(1), 61-71(2016).

[35] S. Saaidpour, S. Khaledian, *Quantitative Structure-property Relationship Modelling of Distribution Coefficients (logD7.4) of Diverse Drug by Sub-structural Molecular Fragments Method*, Orient. J. Chem., **31**(4), 1969-1976 (2015).

[36] S. Saaidpour, *Computational Model For Chromatographic Relative Retention Time of Polychlorinated Biphenyls Using Sub-structural Molecular Fragments*, CMST, **22**(1) 41-53 (2016).

[37] C.L. Yaws, *Handbook of Physical Properties for Hydrocarbons and Chemicals*, Houston: Gulf Publishing Co., 2005.

[38] ChemOffice 15.0, PerkinElmer, Inc., Waltham, MA, USA, 2015, http://www.cambridgesoft.com

[39] HyperChem (TM) Professional 8.0, Hypercube, Inc., 2011, Gainesville, Florida, USA, http://www.hyper.com.

[40] VolSurf+, Version 1.0.4, Molecular Discovery Ltd., 2008, http://www.moldiscovery.com.

[41] Molegro Data Modeller (MDM 2011.2.6.0), Molegro ApS., 2011, C.F. Mřllers Alle, Building 1110, DK-8000 Aarhus C, Denmark, http://www.molegro.com/mdm-product.php.

[42] G. Cruciani, P. Crivori, P.A. Carrupt, B. Testa, *molecular fields in quantitative structure permeation relationships: the VolSurf+ approach*, J. Mol. Struct. (Theochem), **503**, 17-30 (2000).

[43] G. Cruciani, M. Pastor, W. Guba, *VolSurf+, a new tool for the pharmacokinetic optimization of lead compounds*, Europ. J. Pharm. Sci., **11**, S29-S39 (2000).

[44] S. Clementi, G. Cruciani, P. Fifi, D. Riganelli, R. Valigi, G. Musumarra, *A New Set of Principal Properties for Heteroaromatics Obtained by GRID*, Quant. Struct. Act. Relat., **15**, 108-120 (1995).

[45] D.C. Young, *Computational Chemistry*, John Wiley & Sons Inc., 2001.

[46] M.J. Sharma and Y.S. Jin, *Stepwise regression data envelopment analysis for variable Reduction*, Appl. Math. Comput., **253**, 126-134 (2015).

[47] K. Baumann, *Cross-validation as the objective functions for variable-selection techniques*, TrAC - Trends Anal. Chem., **22**, 395-406 (2003).

[48] S. Weisberg, *Applied Linear Regression*, 3rd edn. Wiley, New York, 2005.

[49] D.C. Montgomery, E.A. Peck, G.G. Vining, *Introduction to Linear Regression*, 4th edn. Wiley, New York, 2006.

[50] K. Baumann, N. Stiefl, *Validation tools for variable subset regression*, J. Comput. Aided. Mol. Des., **18**, 549-562 (2004).

[51] N. Chirico, P. Gramatica, *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**, 2320-2335 (2011).

[52] N. Chirico, P. Gramatica, *Real External Predictivity of QSAR Models. Part2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection*, J. Chem. Inf. Model., **52**, 2044-2058 (2012).

[53] C. Rücker, G. Rücker, M. Meringer, *y-Randomization and Its Variants in QSPR/QSAR*, J. Chem. Inf. Model., **47**, 2345-2357 (2007).

Saadi Saaidpour is an Assistant Professor at the Department of Chemistry, IAU Sanandaj branch, Sanandaj, Iran. In 2004, he received his MSc degree in Analytical Chemistry and in 2008 he received his PhD degree in Analytical Chemistry and Chemometric at Razi University, Kermanshah, Iran. His research interests concern chemometrics methods, computational chemistry and QSAR & QSPR studies.

Faraidon Ghaderi Faraidon Ghaderi-was born in Kermanshah, Iran (1980). He received his BSc in Chemistry at the Razi University of Kermanshah and his MSc at the Department of Chemistry, IAU Sanandaj Branch, Sanandaj, Iran, in 2016, in Analytical Chemistry. His MSc dissertation (Modeling of Henry's law constant, Boiling Point and Water Solubility of crude oil Hydrocarbons using chemomertrics methods) is being prepared under Assistant Professor Saadi Saaidpour supervision.