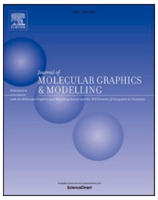
Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/jmgm



QSPR for the prediction of critical micelle concentration of different classes of surfactants using machine learning algorithms

Nada Boukelkal [*], Soufiane Rahal, Redha Rebhi, Mabrouk Hamadache

Biomaterials and Transport Phenomena Laboratory (LBMPT), University of Yahia Fares, Faculty of Technology, Department of Process Engineering and Environment, Medea, 26000, Algeria

ARTICLE INFO

Keywords:
CMC
QSPR
RFR
Surfactants
SVR

ABSTRACT

The determination of the critical micelle concentration (CMC) is a crucial factor when evaluating surfactants, making it an essential tool in studying the properties of surfactants in various industrial fields. In this present research, we assembled a comprehensive set of 593 different classes of surfactants including, anionic, cationic, nonionic, zwitterionic, and Gemini surfactants to establish a link between their molecular structure and the negative logarithmic value of critical micelle concentration (pCMC) utilizing quantitative structure-property relationship (QSPR) methodologies. Statistical analysis revealed that a set of 14 significant Mordred descriptors (SlogP, GATS6d, nAcid, GATS8dv, GATS4dv, PEOE_VSA11, GATS8d, ATS0p, GATS1d, MATS5p, GATS3d, NdssC, GATS6dv and EState_VSA4), along with temperature, served as appropriate inputs. Different machine learning methods, such as multiple linear regression (MLR), random forest regression (RFR), artificial neural network (ANN), and support vector regression (SVM), were employed in this study to build QSPR models. According to the statistical coefficients of QSPR models, SVR with Dragonfly hyperparameter optimization (SVR-DA) was the most accurate in predicting pCMC values, achieving ($R^2 = 0.9740$, $Q^2 = 0.9739$, $\bar{r}_m^2 = 0.9627$, and $\Delta r_m^2 = 0.0244$) for the entire dataset.

1. Introduction

Surface active agents, also known as surfactants, are vital chemicals that play a crucial role in many aspects of our daily lives. There are four main subclasses into which surfactants can be categorized, depending on their net charge: cationic, anionic, nonionic, and zwitterionic. Lately, researchers have shown significant interest in a specific type of surfactant called Gemini surfactants. These surfactants possess two hydrophobic tails and two head groups that are connected by a short spacer [1]. One of the fundamental characteristics of surfactants is their ability to collect at interfaces, which is related to the fact that their structure contains both hydrophilic and hydrophobic regions [2]. This characteristic of surfactants can reduce surface tension between two substances. In principle, the effectiveness of the surfactant increases as its tendency becomes stronger. At a boundary, both the structure of the surfactant and the types of the two phases that meet at the interface influence the concentration of the surfactant. Therefore, the choice of a surfactant should be based on its specific use. Micellization, or micelle formation, is the dynamic process of the creation of micelles, which is another essential characteristic of surfactants, that is, the surfactant concentration exceeds a critical value called the CMC [1]. This last refers to the concentration at which surfactant molecules start to form micelles. At this point, various characteristics of the surfactant solution, such as foaming, interfacial tension, emulsification, conductivity, and others, undergo substantial changes [3]. Due to these characteristics, surfactants are commonly used in various industrial applications such as pharmaceuticals, detergents, personal care, food, and agriculture. They play a crucial role in facilitating wetting, foaming, emulsification, and lubrication processes [4–7]. Various external factors, such as temperature, pressure, pH, ionic strength, volume of the solution, and the structural characteristics of the surfactant, including hydrophobic tail length and head group area, can significantly impact the CMC [7]. A variety of experimental techniques such as tensionmetry, conductance, nuclear magnetic resonance spectroscopy, cyclic voltammetry, and fluorescence emission spectroscopy can be employed to determine the CMC value [8].

Within the wide range of methodologies found in the literature for predicting substance properties by understanding the chemical

[*] Corresponding author.
E-mail addresses: nadaboukelkal@gmail.com, boukelkal.nada@univ-medea.dz (N. Boukelkal).

https://doi.org/10.1016/j.jmgm.2024.108757
Received 27 October 2023; Received in revised form 7 March 2024; Accepted 10 March 2024
Available online 11 March 2024

structure, QSPR modeling is a significant field of research in computational chemistry [9] because it can provide a faster, more accurate, and less expensive method for understanding and measuring structural characteristics that affect physical property. In the QSPR framework, a wide range of molecular descriptors effectively capture the intricate details of the molecular structure, such as topological, constitutional, geometrical, electronic, and more [8]. Therefore, a chosen set of descriptors is statistically associated with the studied experimental property, producing a mathematical model that can be used for discovering valuable correlations between structure and property.

Multiple QSPR models have been constructed to predict the CMC of surfactants [2,3,10–14] but few published QSPR models can accurately predict CMC values for all surfactant classes. Qin et al. [15] utilized a graph representation of molecules, where atoms are nodes and chemical bonds are edges, as the input for a graph convolutional neural network (GCN) to predict the CMC of a set of 202 surfactants, which included anionic, cationic, nonionic, and zwitterionic compounds. The test set with a ($R^2 = 0.92$, $RMSE = 0.3$), demonstrates the robustness of the GCN model. A report published in 2002 [16] presents a QSPR model to predict the CMC of 49 surfactants and seven molecular descriptors. By applying partial least squares regression, which takes into account nonionic, anionic, cationic, and zwitterionic surfactants, the study achieved an observed coefficient of determination of 0.90 for the training set.

This study aims to establish QSPR models for the predicting and correlating the negative logarithm of CMC with the molecular structure of 593 different classes of surfactants (including anionic, cationic, zwitterionic, nonionic, and Gemini surfactants) using various descriptors. These descriptors are calculated from the two-dimensional depiction of the molecules. Both internal and external validation were used to select the best model. Finally, the applicability domain based on the Williams plot was analyzed for the best model.

2. Methodology

The process of building QSPR models contains several steps, which are explained below and illustrated in Fig. 1.

2.1. Data collection

A total of 593 distinct classes of surfactants, including anionic, cationic, nonionic, zwitterionic, and Gemini surfactants, were collected from a previous publication [2,10,17–29]. These surfactant CMCs were measured at temperatures ranging from 10 °C to 60 °C. To ensure a

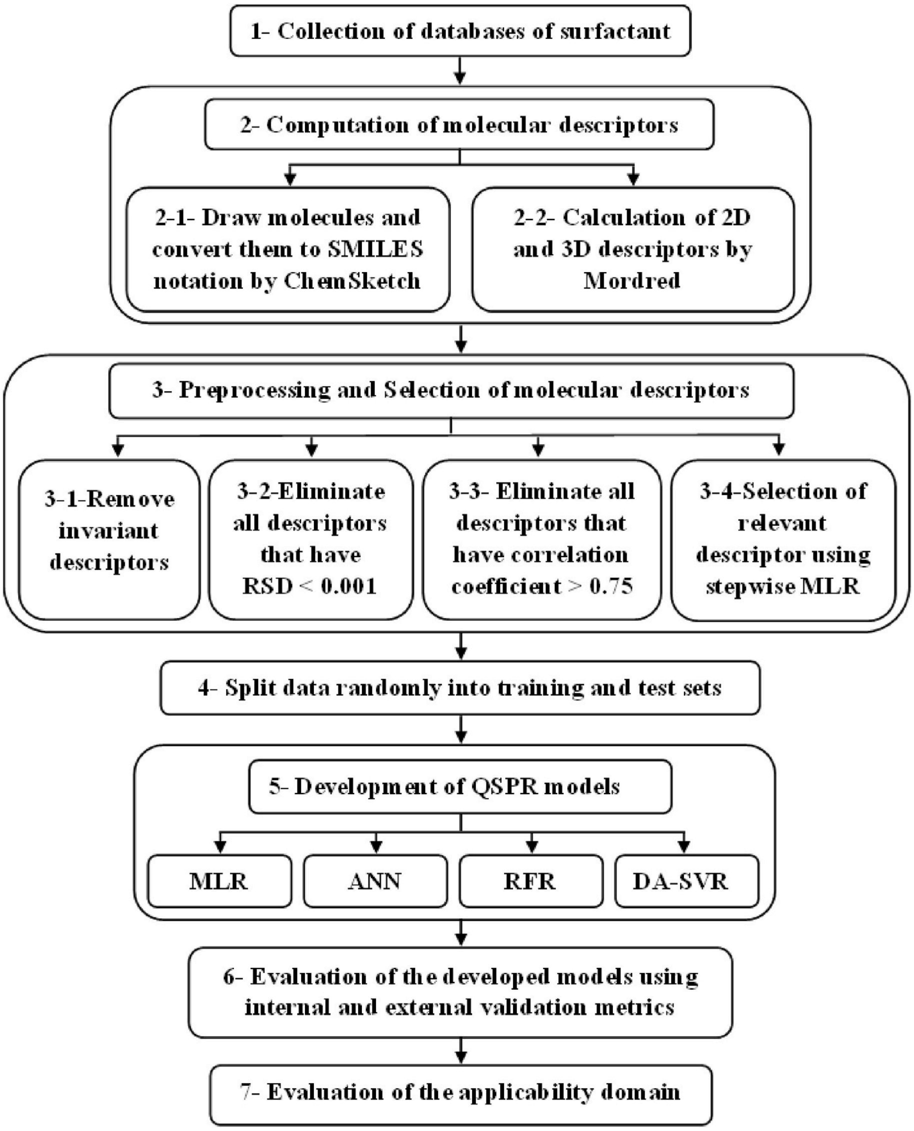

1- Collection of databases of surfactant
2- Computation of molecular descriptors
2-1- Draw molecules and convert them to SMILES notation by ChemSketch
2-2- Calculation of 2D and 3D descriptors by Mordred
3- Preprocessing and Selection of molecular descriptors
3-1-Remove invariant descriptors
3-2-Eliminate all descriptors that have RSD < 0.001
3-3- Eliminate all descriptors that have correlation coefficient > 0.75
3-4-Selection of relevant descriptor using stepwise MLR
4- Split data randomly into training and test sets
5- Development of QSPR models
MLR
ANN
RFR
DA-SVR
6- Evaluation of the developed models using internal and external validation metrics
7- Evaluation of the applicability domain

Fig. 1. General steps for the generation of QSPR models in the current study.

normal distribution, CMC values were transformed into the negative logarithmic form of pCMC (in mol/L), as shown in Fig. 2. The surfactant compounds and their associated experimental pCMC values are listed in the supplemental file as Table S1.

2.2. *Molecular descriptors calculation*

A molecular descriptor is a numerical number that is calculated based on a compound's molecular structure [30]. Molecular descriptors act as features in building a machine learning model. Mordred was utilized to compute 1826 descriptors for each surfactant within a Python environment. These descriptors consist of 1613 2D descriptors and 213 3D descriptors [31].

Before calculating these descriptors, you need to first draw the surfactant molecules and then convert them to SMILES (Simplified Molecular Input Line-Entry System) notation using "Chem Sketch" software and save them as a CSV file. Then, convert the file into Mol objects using RDKit to compute the Mordred descriptor.

2.3. *Selection of relevant descriptors*

To ensure the relevance of molecular descriptors, it is necessary to implement a feature selection method. First, we excluded invariant descriptors that had missing values. Any descriptors with zero values are also removed. Then, descriptors with a relative standard deviation less than 0.001 were excluded. In addition, descriptors with an absolute value of the correlation coefficient greater than 0.75 were finally eliminated. We have acquired a total of 77 descriptors through the process of selection. Then stepwise MLR with the F value method was used to select the most pertinent descriptors (https://sites.google.com/site/dtclabsmlr/). A set of 15 features containing temperature data obtained from feature pre-screening was used to build our predictive models. Variance inflation factors (VIF) were calculated to describe the selected model using SPSS 19.0 software to detect multi-collinearity. A model is considered appropriate if the VIF value is between 1 and 5 [32]. As shown in Table 1, all the descriptors' VIF values were less than 3, indicating that the generated model is statistically significant, and the descriptors were discovered to be sufficiently orthogonal. For more details, Table 1 gives an overview of the characterization of the selected descriptors.

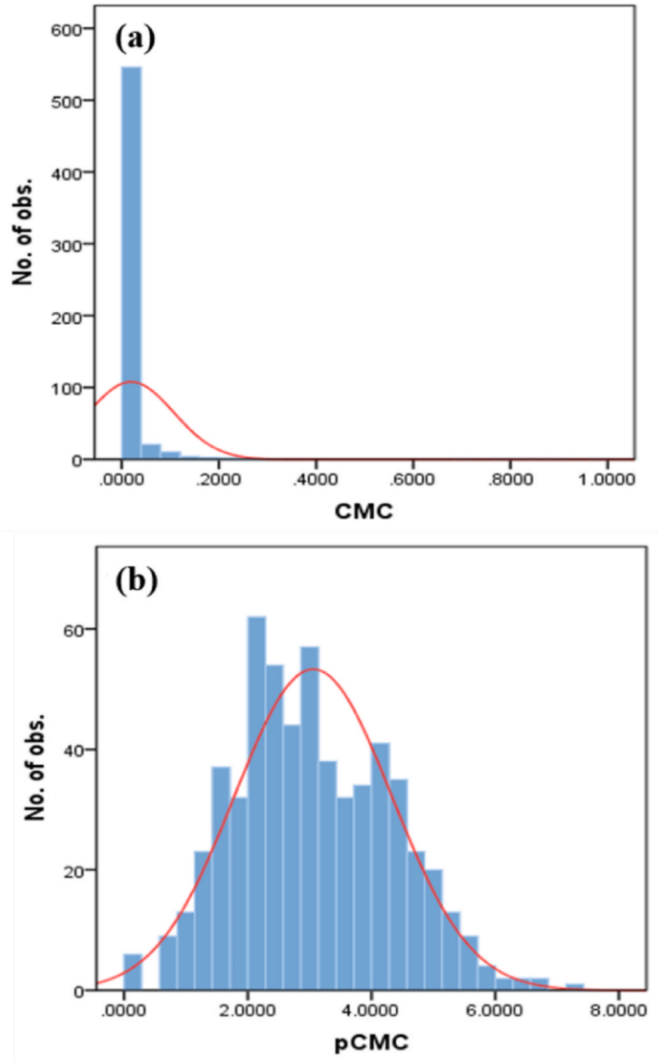


Fig. 2. Normality distribution plot of CMC data: (a) before transformation, (b) after log transformation (pCMC).

2.4. *Model development*

The dataset was randomly divided into two sets using the MATLAB® divider function. 80% of the data is utilized as a training set for model development, while the test set comprises the remaining 20% for evaluating the performance of the model.

2.4.1. *MLR model*

The use of MLR led to the creation of a linear model. Various research groups have demonstrated the MLR approach for developing predictive

Table 1
Characteristics of the selected descriptors by S-MLR and their VIF values.

Descriptors names	type	class	Definition	VIF
SlogP	2D	Physicochemical properties	Logarithm of the octanol/water partition coefficient	1.457
GATS6d	2D	Autocorrelation	Geary autocorrelation descriptor of lag6	1.445
nAcid	2D	Count descriptor	Number of acidic groups	1.773
GATS8dv	2D	Autocorrelation	Modified Geary autocorrelation descriptor of lag8	2.609
GATS4dv	2D	Autocorrelation	Modified Geary autocorrelation descriptor of lag4	3.000
PEOE_VSA1	2D	MOE-type	MOE Charge VSA Descriptor 1	2.223
GATS8d	2D	Autocorrelation	Geary autocorrelation descriptor of lag8	2.395
ATS0p	2D	Autocorrelation	Broto-Moreau autocorrelation - lag 0/ weighted by polarizabilities	1.228
GATS5d	2D	Autocorrelation	Geary autocorrelation descriptor of lag5	2.350
MATS5p	2D	Autocorrelation	Moran autocorrelation - lag 5/weighted by polarizabilities	1.660
GATS3d	2D	Autocorrelation	Geary autocorrelation descriptor of lag3	2.245
ndssC	2D	Electrotopological State Atom Type	Count of atom-type E-State:=C<	1.198
GATS6dv	2D	Autocorrelation	Modified Geary autocorrelation descriptor of lag6	1.725
EState_VSA4	2D	MOE-type descriptor	MOE-type descriptors using EState indices and surface area contributions	1.472
T	–	–	Temperature	1.152

QSPR models [33–37]. MLR requires establishing a quantitative relationship between independent variables X (descriptors) and response Y (property), as demonstrated by Eq. (1)

$$Y = b_0 + \sum_{i=1}^{N} X_i b_i \tag{1}$$

Where Y is the dependent variable; X_i represents the explanatory variable (descriptors); b_i represents the coefficient of those descriptors and b_0 is the intercept of the equation. The MLRplusValidation 1.3 tool (https://sites.google.com/site/mlrplusvalidation) was used to conduct the MLR calculations.

2.4.2. ANN model

McCulloch and Pitts [38] created artificial neural networks, inspired by biological nervous systems. ANN consists of three layers: input, hidden, and output. Each layer consists of interconnected processing components called neurons, which send their output to the next layer's input. The neurons in the hidden and output layers are computed with active functions such as the linear or sigmoid function. The number of neurons in the input and output layers depends on the input and output variables. The number of neurons in the hidden layer is not fixed, but it depends on the difficulty of the problem and is determined through trial and error [39]. There are several neural networks with different training algorithms. In this work, we used a multilayer perceptron (MLP) trained with the trainbr function, that updates the weight and bias value according to Bayesian regularization back propagation optimization. It reduces a linear combination of mean squared errors (MSE) and weights [40]. The ANN algorithm was implemented using the MATLAB® R 2018a. environment.

2.4.3. RFR model

Tin Kam Ho [41,42] introduced Radom Forest in 1995 as a method for creating decision trees for classification or regression applications. Breiman and Culter [42,43] extended the bagging approach to include a random selection aspect during training.

A random forest is formed of a large number of individual decision trees that work as an ensemble. Each tree makes a category prediction, and the category with the most votes becomes our model's forecast [44]. For this study, we used Python as a programming language and utilized Jupyter, with the assistance of the SKlearn library.

2.4.4. DA-SVR model

The concept of SVR was initially presented in 1997 by Drucker et al. [45]. SVR is a supervised model within the SVM [46] designed for regression tasks. SVR has hyperparameters [47], like the kernel and C. Kernel parameters control the decision boundary, while C parameters control error penalties. Dragonfly [48] was used to determine the best hyperparameter values for the SVR model in our database, resulting in remarkable accuracy and minimal error. The DA-SVR algorithm was implemented in the MATLAB® R 2018a environment.

2.5. Model validation

Various significant statistical parameters and applicability domains were applied to evaluate the efficacy of the model. The metrics assessed were the coefficient of determination (R^2), different regression metrics (Q^2, Q^2_{F1}, Q^2_{F2}), average and delta of r^2_m ($\overline{r}^2_m$, Δr^2_m), concordance correlation coefficient (CCC) to evaluate reproducibility, root mean squared error ($RMSE$), mean absolute error (MAE), average absolute relative deviation ($AARD$) and Akaike's information criterion (AIC) [49], which is an additional statistical parameter that is utilized to compare different models and determine which model best represents the experimental results. A low AIC value suggests a better-fitting model. The statistical parameters' formulas are available in the literature [50–53].

3. Results and discussion

3.1. MLR predictive model

The MLR model's performance depends on the size and quality of the dataset. For a reliable model, the ratio of datasets to variables must exceed 5 [54]. The linear equation representing the generated MLR model is given below.

$$\begin{aligned}&\text{pCMC} = 2.12274(\pm 1.0075) + 0.33915(\pm 0.02052)\ \text{SLogP} + 3.45516\\&(\pm 0.38652)\ \text{GATS6d} - 0.17197(\pm 0.06904)\ \text{nAcid - } 1.86614(\pm 0.40426)\\&\text{GATS8dv - } 0.25316(\pm 0.37152)\ \text{GATS4dv} + 0.02768(\pm 0.01231)\\&\text{PEOE\_VSA11} + 1.69493(\pm 0.35779)\ \text{GATS8d} + 0.00014(\pm 0.00014)\ \text{ATS0p}\\&+ 0.07065(\pm 0.42478)\ \text{GATS1d} + 1.03934(\pm 1.5928)\ \text{MATS5p - } 1.34056\\&(\pm 0.43476)\ \text{GATS3d - } 0.10528(\pm 0.08413)\ \text{NdssC - } 1.35554(\pm 0.33952)\\&\text{GATS6dv} + 0.02183(\pm 0.00494)\ \text{EState\_VSA4 - } 0.01049(\pm 0.00714)\ \text{T}\end{aligned} \tag{2}$$

Where.

$n_{train} = 475$, $R^2 = 0.5194$, $Q^2 = 0.5194$, $\overline{r}^2_m = 0.3597$, $\Delta r^2_m = 0.3193$, $MAE = 0.6901$, $RMSE = 0.8841$ for training set.

The standard error of the regression coefficients is indicated in parentheses. Table 2 provides an overview of the external criteria of the MLR model.

3.2. ANN predictive model

The ANN model was constructed using two activation functions; the hyperbolic tangent (tansig) function as a transfer function for hidden layers, and the linear transfer (purelin) function as a transfer function for output layers. The network was trained using a Bayesian regularization back propagation training function (trainbr). The architecture of the final model is {15,10,1}. In the training set, this model achieved statistical parameters of $R^2 = 0.9299$, $Q^2 = 0.9299$, $\overline{r}^2_m = 0.9852$, $\Delta r^2_m = 0.0626$, $MAE = 0.2026$, $RMSE = 0.3345$. Table 2 provides an overview of the external criteria of the ANN model.

3.2.1. Mathematical equation application for the QSPR-ANN model

The network architecture developed in this study is a multilayer perceptron with a structure of {15,10,1}. So, the network includes fifteen inputs (X_i, $i = 1$ to 15), one output (Z), and a hidden layer of ten neurons. We used hyperbolic tangent and linear functions as activation functions, which are mathematically defined in Eqs. (3) and (4), respectively.

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3}$$

$$f(x) = x \tag{4}$$

The output signal of each hidden neuron (Y_j) is calculated by Eq. (5), and the output of the network is given by Eq. (6).

Table 2
Statistical criteria of all developed models for the test set.

Validation metrics	MLR	ANN	RFR	SVR
R^2	0.4343	0.8877	0.8121	0.9864
Q^2_{F1}	0.4485	0.8905	0.8142	0.9868
Q^2_{F2}	0.4252	0.8866	0.8063	0.9862
r^2_0	0.4256	0.8870	0.8063	0.9862
r^2_m	0.3938	0.8642	0.7503	0.9725
$\overline{r}^2_m$	0.2958	0.8531	0.6915	0.9751
Δr^2_m	0.2590	0.0693	0.2295	0.0223
CCC	0.6337	0.9409	0.8854	0.9930
MAE	0.7354	0.3126	0.3907	0.0566
RMSE	0.9322	0.4350	0.5411	0.1443

$$Y_j = f\left[\sum_{i=1}^{15} w_{i,j}X_i + b_j\right] = \frac{exp\left(\sum_{i=1}^{15} w_{i,j}\,X_i + b_j\right) - exp\left(-\sum_{i=1}^{15} w_{i,j}\,X_i + b_j\right)}{exp\left(\sum_{i=1}^{15} w_{i,j}\,X_i + b_j\right) + exp\left(-\sum_{i+1}^{15} w_{i,j}\,X_i + b_j\right)} \tag{5}$$

$$Z = f\left[\sum_{j=1}^{10} w_{1,j}Y_j + b_1\right] = \sum_{j=1}^{10} w_{1,j}\,Y_j + b_1 \tag{6}$$

Where $w_{i,j}$ are the weights of the connections between the input and hidden neurons, X_i are the input variables and b_j is the bias on hidden neuron j. $w_{1,j}$ represent the weights of the connections between the hidden and output neuron and b_1 is the bias on the output neuron.

The final mathematical formula for predicting the critical micelle concentration of different classes of surfactants derived from the ANN approach is given in Eq. (6). The values of the weights and bias for each layer are given in the supplemental file as Table S2 and Table S3, respectively. Table 5 shows the predicted values for 11 surfactants that were not included in the construction of the QSPR models. These predictions were based on Eq. (6).

3.3. RFR predictive model

We create the RFR model using Scikit-learn's RandomForestRegressor class. Our hyperparameters include setting 100 estimators and a minimum sample split of 2. In the training set, this model achieved statistical parameters of $R^2 = 0.9717$, $Q^2 = 0.9665$, $\bar{r}_m^2 = 0.8861$, $\Delta r_m^2 = 0.0327$, $MAE = 0.1553$, $RMSE = 0.2334$. Table 2 provides an overview of the external criteria of the RFR model.

3.4. DA-SVR predictive model

The optimal value of the SVR hyperparameter was found using the dragonfly algorithm. Our model used the radial basis function (rbf) as a kernel function. The highest performance of the DA-SVR model was obtained with the following optimal hyperparameter values: regularization constant ($C = 200$), insensitivity function ($\varepsilon = 1.9939$), and gaussian function parameter ($\gamma = 0.0080$). The predicted pCMC values for all 593 surfactants are presented in the supplemental file as Table S1. The scatter plot shown in Fig. 3, shows a strong correlation between the experimental values and the predicted results obtained by the DA-SVR. In the training set, this model achieved statistical parameters of $R^2 = 0.9712$, $Q^2 = 0.9710$, $\bar{r}_m^2 = 0.9581$, $\Delta r_m^2 = 0.0222$, $MAE = 0.0871$, $RMSE = 0.2171$. Table 2 provides an overview of the external criteria of the DA-SVR model.

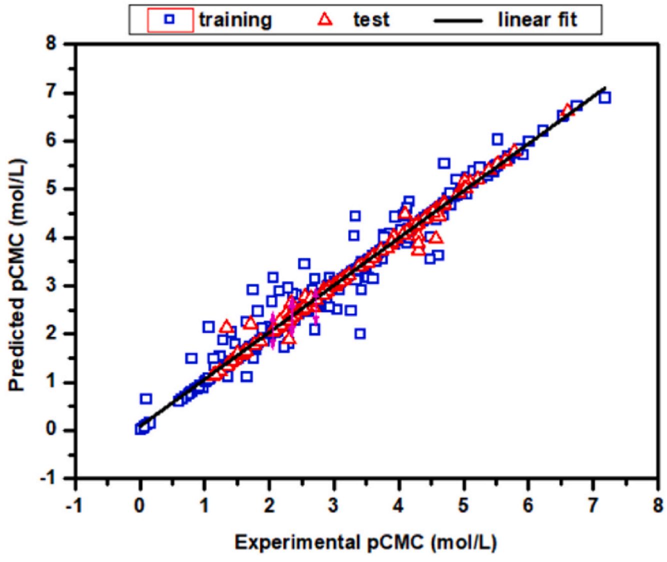


Fig. 3. Scatter plots Predicted vs. Observed values of pCMC for the DA-SVR for training and test sets.

3.5. Comparison of four statistical models

Various validation metrics were utilized to assess the level of quality of the model. The validation metrics, both internal and external, demonstrate that the SVR, ANN, and RFR models are of acceptable quality; however, when it comes to the MLR model, the regression and error values fall short of ensuring a satisfactory level of predictive ability. The statistical metrics for the test set and whole set, as can be seen in Tables 2 and 3, respectively, show that the DA-SVR model surpasses all other models. It attains the highest R^2 (0.9740), Q^2 (0.9739) and lower $RMSE$ (0.2047), $AARD\%$ (4.4447). Also, the DA-SVR exhibits a significantly lower AIC value. Overall, statistical parameters indicated that the DA-SVR prediction had significantly higher accuracy compared to the RFR, ANN, and MLR models.

3.6. Applicability domain

Once the model is validated, it is crucial to determine its applicability domain as defined by the third OECD principle. This step ensures the reliability of the model and defines the range of samples whose predictions can be considered accurate. The Williams plot [55,56] was used in this study to establish the applicability domain of the SVR model. As shown in Fig. 4, The threshold leverage computed, denoted by h^*, equals 0.101. A total of 569 compounds were found in both the training and test sets within the domain, and only 24 compounds were identified as outliers, where 8 compounds are outside the applicability domain (h > h*) and 16 compounds are outside the range of the ± standard deviation unit. Thus, an accuracy rate of 95.95% was achieved through predictions within the applicability domain. Fortunately, the DA-SVR model produced accurate predictions for these compounds.

3.7. Contribution of the selected descriptors

The set of 15 features obtained from feature pre-screening was used to build our predictive models. According to Table 1, the selected descriptors have been categorized into five distinct groups according to their properties and characteristics. The relevance factor [57] method was utilized to assess the impact of each of the 15 descriptors of the DA-SVR model in predicting surfactant CMC. Fig. 5 demonstrates that the CMC is directly influenced by descriptors known as SLogP, GATS6d, GATS8d, MATS5p, GATS3d, GATS6dv, and EState_VSA4. Conversely, the temperature and other descriptors exhibit an inverse relationship with CMC. Moreover, the greatest relevance input variables were SLogP and GATS4dv, with relevance factors of +0.57 and −0.18, respectively. GATS3d and NdssC showed the least effect, with a relevance factor of +0.01 and −0.02, respectively.

Table 3
Statistical criteria of all developed models for the global set.

Validation metrics	MLR	ANN	RFR	SVR
R^2	0.5024	0.9247	0.9418	0.9740
Q^2	0.5021	0.9247	0.9365	0.9739
r_0^2	0.5021	0.9247	0.9366	0.9740
r_m^2	0.4937	0.9247	0.8739	0.9740
$\bar{r}_m^2$	0.3478	0.8898	0.8819	0.9627
Δr_m^2	0.3085	0.0700	0.1095	0.0244
CCC	0.6743	0.9607	0.9649	0.9868
MAE	0.6991	0.2229	0.2022	0.0810
RMSE	0.8939	0.3476	0.3192	0.2047
AARD (%)	81.1891	10.7150	24.2263	4.4447
AIC	−102.2222	−1222.3	−1323.3	−1850.05

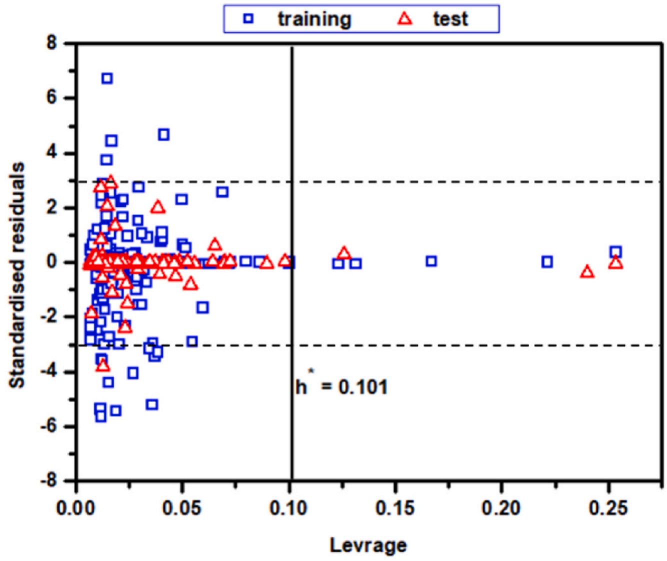


Fig. 4. William's plot for the DA-SVR model.

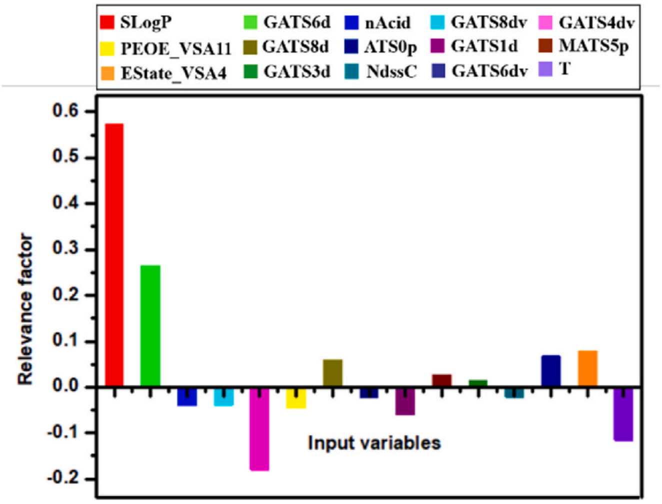


Fig. 5. The results of the relevance factor analysis carried out on the DA-SVR model.

3.8. Comparison with previously reported models

As part of model validation, our DA-SVR model for predicting the pCMC of 593 surfactants has been compared to a previously published model in the literature. Comparing the current study with previous ones is difficult due to differences in data sets and modeling methods involving molecular descriptors and algorithms. It is important to note that the majority of these QSPR models have exclusively used a single class of surfactant. For instance, Wang et al. [58] considered nonionic surfactants, Rahal et al. [37] focused on anionic surfactants, Guo et al. [21] considered Gemini surfactants, and Roy et al. [59] explored cationic surfactants. However, Qin et al.'s [15] and Anoune et al.'s [16] studies were exceptional as they encompassed anionic, cationic, nonionic, and zwitterionic surfactants. Table 4 displays the statistical parameters for the results obtained in this study and previous studies found in the literature. All of these models have the potential to provide excellent prediction abilities. However, our model surpasses all previously published models in all statistical metrics available for comparison, with a higher coefficient correlation and a lower root mean squared error. Furthermore, we used our ANN-based equation and DA-SVR models to predict the pCMC of 11 surfactants. These surfactants were not included in the construction of the QSPR models. The results are shown in Table 5. The findings indicate that the DA-SVR model is more consistent with the actual data compared to the ANN model.

4. Conclusion

This study aims to compare linear and non-linear methods for predicting the CMC of various surfactant classes based on their molecular descriptors. After applying multiple reduction techniques, a total of 14 molecular descriptors were identified as highly relevant and retained. A total of 593 experimental data points were used to train and validate the proposed models. These data points represent the critical micelle concentration of surfactants at various temperatures. The proposed statistical parameters were examined to assess the predictive performance of the improved model. The results demonstrate that by utilizing the dragonfly approach for hyperparameters optimization, the SVR model surpasses the ANN, RFR, and MLR models in accurately forecasting the non-linear characteristics of CMC of surfactants. In summary, the validation results demonstrate the strength and reliability of the SVR-QSPR model that has been built. Its performance is highly satisfactory, making it a valuable tool for predicting the critical micelle concentration of these surfactant compounds.

Table 5
Observed value of pCMC and those calculated by Eq. (6) and DA-SVR for 11 surfactants.

Structure	T (°C)	pCMC (mol/L)		
		observed	Predicted by ANN Eq. (6)	Predicted by DA-SVR
$C_9H_{19}SO_4^-Na^+$	25	1.22	1.15	1.27
$C_{10}H_{21}(OC_2H_4)_2SO_4^-Na^+$	25	1.91	1.80	1.68
$C_{12}H_{25}(OC_2H_4)_4SO_4^-Na^+$	25	2.77	2.86	2.47
$C_{14}H_{29}COO^-Na^+$	40	2.14	1.95	2.26
$C_{16}H_{33}COO^-Na^+$	40	2.74	2.30	2.32
$C_7H_{15}CH(C_3H_7)C_6H_4SO_3^-Na^+$	40	2.40	0.83	2.27
$C_6H_{13}CH(C_4H_9)C_6H_4SO_3^-Na^+$	40	2.30	2.40	2.23
$C_5H_{11}CH(C_5H_{11})C_6H_4SO_3^-Na^+$	40	2.25	2.89	2.37
$C_8H_{17}C_6H_4SO_3^-Na^+$	40	1.91	2	1.91
$C_8H_{17}CH(CH_3)C_6H_4SO_3^-Na^+$	40	2.30	1.58	2.43
$C_7H_{15}CH(C_2H_5)C_6H_4SO_3^-Na^+$	40	2.20	0.58	2.37

Table 4
Performance of our DA-SVR model with previously published models for internal and external validations.

Models	QSPR method	Data points	Internal validation				External validation			
			R^2	Q^2	$\bar{r}_m^2$	Δr_m^2	R^2	Q_{ext}^2	*MAE*	*RMSE*
This study	**DA-SVR**	**593**	**0.971**	**0.971**	**0.958**	***0.087***	**0.986**	**0.986**	**0.056**	**0.144**
Qin et al. [15]	GCN	202	–	–	–	–	0.920	–	–	0.300
Anoune et al [16]	PLS	49	0.900	-	-	-	-	-	-	-
wang et al. [58]	MLR	83	0.959	0.947	–	–	0.946	–	–	–
Rahal et al. [37]	MLP-ANN	50	0.940	0.930	0.890	0.060	–	0.950	–	–
Guo et al. [21]	GA-LSSVM	120	0.964	0.917	–	–	0.855	–	–	0.500
Roy et al. [59]	GFA	35	0.949	0.893	0.851	0.020	0.904	–	–	–

CRediT authorship contribution statement

Nada Boukelkal: Writing – original draft. **Soufiane Rahal:** Writing – original draft. **Redha Rebhi:** Writing – original draft. **Mabrouk Hamadache:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmgm.2024.108757.

References

[1] K. Holmberg, B. Jönsson, B. Kronberg, B. Lindman, Surfactants and Polymers in Aqueous Solution, 2002.

[2] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices: exploring effects of hydrophobicity, branching and electronic parameters on logCMC values of anionic surfactants, Chem. Eng. Sci. 87 (2013) 141–151.

[3] L. Jiao, Y. Wang, L. Qu, Z. Xue, Y. Ge, H. Liu, B. Lei, Q. Gao, M. Li, Hologram QSAR study on the critical micelle concentration of Gemini surfactants, Colloids Surfaces A Physicochem. Eng. Asp. 586 (2020).

[4] D. Myers, SURFACTANT SCIENCE AND TECHNOLOGY, third ed., John Wiley & sons, 2006.

[5] M.J.L. Castro, C. Ojeda, A.F. Cirelli, Advances in surfactants for agrochemicals, Environ. Chem. Lett. 12 (2014) 85–95.

[6] K. Hill, O. Rhode, Sugar-based surfactants for consumer products and technical applications, Lipid - Fett 101 (1999) 25–33.

[7] A.R. Katritzky, L.M. Pacureanu, S.H. Slavov, D.A. Dobchev, M. Karelson, QSPR study of critical micelle concentrations of nonionic surfactants, Ind. Eng. Chem. Res. 47 (2008) 9687–9695.

[8] X. Li, G. Zhang, J. Dong, X. Zhou, X. Yan, M. Luo, Estimation of critical micelle concentration of anionic surfactants with QSPR approach, J. Mol. Struct. THEOCHEM. 710 (2004) 119–126.

[9] A.R. Katritzky, U. Maran, V.S. Lobanov, M. Karelson, Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties, J. Chem. Inf. Comput. Sci. 40 (2000) 1–18.

[10] T. Gaudin, P. Rotureau, I. Pezron, G. Fayet, New QSPR models to predict the critical micelle concentration of sugar-based surfactants, Ind. Eng. Chem. Res. 55 (2016) 11716–11726.

[11] A. Baghban, J. Sasanipour, M. Sarafbidabad, A. Piri, R. Razavi, On the prediction of critical micelle concentration for sugar-based non-ionic surfactants, Chem. Phys. Lipids 214 (2018) 46–57.

[12] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices: modeling of critical micelle concentration of non-ionic surfactants, Chem. Eng. Sci. 73 (2012) 86–98.

[13] P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, S. Dinesh, O. Shah, M. Karelson, Prediction of critical micelle concentration using a quantitative structure-property relationship approach. 1. Nonionic surfactants, Langmuir 12 (1996) 1462–1470.

[14] A.R. Katritzky, L. Pacureanu, D. Dobchev, M. Karelson, QSPR study of critical micelle concentration of anionic surfactants using computational molecular descriptors, J. Chem. Inf. Model. 47 (2007) 782–793.

[15] S. Qin, T. Jin, R.C. Van Lehn, V.M. Zavala, Predicting critical micelle concentrations for surfactants using graph convolutional neural networks, J. Phys. Chem. B 125 (2021) 10610–10620.

[16] N. Anoune, M. Nouiri, Y. Berrah, J.Y. Gauvrit, P. Lanteri, Critical micelle concentrations of different classes of surfactants: a quantitative structure property relationship study, J. Surfactants Deterg. 5 (2002) 45–53.

[17] Z. qiang He, M. jun Zhang, Y. Fang, G. yong Jin, J. Chen, Extended surfactants: a well-designed spacer to improve interfacial performance through a gradual polarity transition, Colloids Surfaces A Physicochem. Eng. Asp. 450 (2014) 83–92.

[18] A. Fernández, C. Scorzza, A. Usubillaga, J.L. Salager, Synthesis of new extended surfactants containing a carboxylate or sulfate polar group, J. Surfactants Deterg. 8 (2005) 187–191.

[19] I.J. Lin, L. Marszall, CMC, HLB, and effective chain length of surface-active anionic and cationic substances containing oxyethylene groups, J. Colloid Interface Sci. 57 (1976) 85–93.

[20] L.H. Lin, Y.C. Lai, K.M. Chen, H.M. Chang, Oxyethylene chain length affects the physicochemical properties of sugar-based anionic surfactants with phosphates groups, Colloids Surfaces A Physicochem. Eng. Asp. 485 (2015) 118–124.

[21] C. Guo, P. Zhou, J. Shao, X. Yang, Z. Shang, Integrating statistical and experimental protocols to model and design novel Gemini surfactants with promising critical micelle concentration and low environmental risk, Chemosphere 84 (2011) 1608–1616.

[22] M.J. Rosen, Surfactants and Interfacial Phenomena, third ed., John Wiley & sons, 2004.

[23] M. Mattei, G.M. Kontogeorgis, R. Gani, Modeling of the critical micelle concentration (CMC) of nonionic surfactants with an extended group-contribution method, Ind. Eng. Chem. Res. 52 (2013) 12236–12246.

[24] S.K. Hait, S.P. Moulik, Determination of critical micelle concentration (CMC) of nonionic surfactants by donor-acceptor interaction with iodine and correlation of CMC with hydrophile-lipophile balance and other parameters of the surfactants, J. Surfactants Deterg. 4 (2001) 303–309.

[25] M.T. Lima, V.J. Spiering, S.N. Kurt-Zerdeli, D.C. Brüggemann, M. Gradzielski, R. Schomäcker, The hydrophilic-lipophilic balance of carboxylate and carbonate modified nonionic surfactants, Colloids Surfaces A Physicochem. Eng. Asp. 569 (2019) 156–163.

[26] T. Gaudin, H. Lu, G. Fayet, A. Berthauld-Drelich, P. Rotureau, G. Pourceau, A. Wadouachi, E. Van Hecke, A. Nesterenko, I. Pezron, Impact of the chemical structure on amphiphilic properties of sugar-based surfactants: a literature overview, Adv. Colloid Interface Sci. 270 (2019) 87–100.

[27] S. Iglauer, Y. Wu, P. Shuler, Y. Tang, W.A. Goddard, Analysis of the influence of alkyl polyglycoside surfactant and cosolvent structure on interfacial tension in aqueous formulations versus n-octane, Tenside, Surfactants, Deterg 47 (2010) 87–97.

[28] G. Lemahieu, J. Aguilhon, H. Strub, V. Molinier, J.F. Ontiveros, J.M. Aubry, Hexahydrofarnesyl as an original bio-sourced alkyl chain for the preparation of glycosides surfactants with enhanced physicochemical properties, RSC Adv. 10 (2020) 16377–16389.

[29] C. Scorzza, P. Godé, G. Goethals, P. Martin, M. Miñana-Pérez, J.L. Salager, A. Usubillaga, P. Villa, Another new family of "extended" glucidoamphiphiles. Synthesis and surfactant properties for different sugar head groups and spacer arm lengths, J. Surfactants Deterg. 5 (2002) 337–343.

[30] F. Grisoni, V. Consonni, R. Todeschini, Impact of Molecular Descriptors on Computational Models, 2018.

[31] H. Moriwaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, J. Cheminf. 10 (2018) 1–14.

[32] D. Wang, Y. Yuan, S. Duan, R. Liu, S. Gu, S. Zhao, L. Liu, J. Xu, QSPR study on melting point of carbocyclic nitroaromatic compounds by multiple linear regression and artificial neural network, Chemometr. Intell. Lab. Syst. 143 (2015) 7–15.

[33] M. Hamadache, L. Khaouane, O. Benkortbi, C. Si Moussa, S. Hanini, A. Amrane, Prediction of acute herbicide toxicity in rats from quantitative structure-activity relationship modeling, Environ. Eng. Sci. 31 (2014) 243–252.

[34] M. Sarkhosh, J.B. Ghasemi, M. Ayati, A quantitative structure- property relationship of gas chromatographic/mass spectrometric retention data of 85 volatile organic compounds as air pollutant materials by multivariate methods, Chem. Cent. J. 6 (2012) 2–9.

[35] S. Bitam, M. Hamadache, S. Hanini, QSAR model for prediction of the therapeutic potency of N-benzylpiperidine derivatives as AChE inhibitors, SAR QSAR, Environ. Res. 28 (2017) 471–489.

[36] S. Bitam, M. Hamadache, S. Hanini, 2D-QSAR, docking, molecular dynamics, studies of PF-07321332 analogues to identify alternative inhibitors against 3CLpro enzyme in SARS-CoV disease, J. Biomol. Struct. Dyn. 41 (2023) 6991–7000.

[37] S. Rahal, N. Hadidi, M. Hamadache, Silico prediction of critical micelle concentration (CMC) of classic and extended anionic surfactants from their molecular structural descriptors, Arabian J. Sci. Eng. 45 (2020) 7445–7454.

[38] W.S. Mccuulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 52 (1943) 115–133.

[39] H. Harandizadeh, D. Jahed Armaghani, M. Khari, A new development of ANFIS–GMDH optimized by PSO to predict pile bearing capacity based on experimental datasets, Eng. Comput. 37 (2021) 685–700.

[40] D. Nielsen, T. Lott, S. Dutta, J. Lee, Artificial neural network (ANN)-based predictive tool for estimating lightning damage in composites, 36th, Tech. Conf. Am. Soc. Compos. 2021 Compos. Ingen. Tak. Challenges Environ. ASC 2 (2021) 1019–1034, 2021.

[41] T. Kam Ho, Random decision forests, in: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 1955.

[42] I.C. Suherman, Implementation of Random Forest Regression for COCOMO II Effort Estimation, International Sem inar on Application for Technology of Inform ation and Communication iSemantic, 2020, pp. 476–481. L.

[43] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.

[44] A. Pandey, V. Rastogi, S. Singh, Car's selling price prediction using random forest machine learning algorithm, SSRN Electron. J. (2020).

[45] H. Drucker, C.J.C. Surges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Adv. Neural Inf. Process. Syst. 1 (1997) 155–161.

[46] C. Corte, V. Vapnic, support-vector network, Mach. Learn. 20 (1995) 273–297.

[47] A. Ben-hur, J. Weston, Chapter 13: a user's guide to support vector machines, data min, Tech. Life Sci. (2010) 223–239.

[48] S. Mirjalili, Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems, Neural Comput, Apple 27 (2016) 1053–1073.

[49] O. Falyouna, O. Eljamal, I. Maamoun, A. Tahara, Y. Sugihara, Magnetic zeolite synthesis for efficient removal of cesium in a lab-scale continuous treatment system, J. Colloid Interface Sci. 571 (2020) 66–79.

[50] K. Roy, R.N. Das, P. Ambure, R.B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, Chemometr. Intell. Lab. Syst. 152 (2016) 18–33.

[51] K. Roy, I. Mitra, On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design, Comb. Chem. High Throughput Screen. 14 (2011) 450–474.

[52] 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 (2011) 2320–2335.

[53] G. Schüürmann, R.U. Ebert, J. Chen, B. Wang, R. Kühne, External validation and prediction employing the predictive squared correlation coefficient - test set activity mean vs training set activity mean, J. Chem. Inf. Model. 48 (2008) 2140–2145.

[54] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.

[55] H.F. Chen, Silico log p prediction for a large data set with support vector machines, radial basis neural networks and multiple linear regression, Chem. Biol. Drug Des. 74 (2009) 142–147.

[56] Z. Ghomisheh, A.E. Gorji, M.A. Sobati, Journal of molecular graphics and modelling prediction of critical properties of sulfur-containing compounds : new QSPR models, J. Mol. Graph. Model. 101 (2020) 107700.

[57] A. Baghban, A. Jalali, M. Shafiee, M.H. Ahmadi, K. wing Chau, Developing an ANFIS-based swarm concept model for estimating the relative viscosity of nanofluids, Eng. Appl. Comput. Fluid Mech. 13 (2019) 26–39.

[58] Y. Wang, F. Yan, Q. Jia, Q. Wang, Quantitative structure-property relationship for critical micelles concentration of sugar-based surfactants using norm indexes, J. Mol. Liq. 253 (2018) 205–210.

[59] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices, 3: modeling of critical micelle concentration of cationic surfactants, Chem. Eng. Sci. 81 (2012) 169–178.