Arabian Journal for Science and Engineering (2020) 45:7445–7454
https://doi.org/10.1007/s13369-020-04598-0

RESEARCH ARTICLE-CHEMICAL ENGINEERING

In Silico Prediction of Critical Micelle Concentration (CMC) of Classic and Extended Anionic Surfactants from Their Molecular Structural Descriptors

S. Rahal[1] · N. Hadidi[1] · M. Hamadache[1]

Received: 13 February 2020 / Accepted: 26 April 2020 / Published online: 8 May 2020

Abstract

CMC is an important parameter for the characterization of surfactants. Compared to other properties, the CMC can be correlated with surfactants performance characteristics on an industrial scale. In this investigation, QSPR models were established to identify the relation between the molecular structures and the critical micelle concentration (CMC) of 50 anionic surfactants employing four molecular structural descriptors. Three regression methods were chosen in this work to develop robust predictive models, namely multilayer perceptron–artificial neural network (MLP/ANN), multiple linear regressions, and partial least square approach. To establish the reliability and the robustness of the developed QSPR models, all available validation strategies were adopted. The best results $\left(\overline{r_m^2} = 0.87; Q^2_{\mathrm{LOO}} = 0.93; Q^2_{F1} = 0.95; \Delta r_m^2 = 0.15\right)$ were obtained for MLP/ANN with a 4-3-1 artificial neural network model trained with the Broyden–Fletcher–Goldfarb–Shanno algorithm. In this study, it is observed that electronic properties, structure and size of the molecule, as well as the number of atoms in the longest aliphatic chain play major roles in the development of the CMC model of anionic surfactants.

Keywords QSPR · Anionic surfactants · CMC · Predicting · MLP–ANN · MLR

1 Introduction

A large number of main properties of the surfactant solution such as conductivity, emulsification, surface tension, detergency, foam stability and conductivity are important for many industrial and biological systems [1–4]. In addition, it has been established that the values of these physicochemical properties change radically as soon as the molecules begin to aggregate to form micelles [5]. At this moment, the concentration of the surfactant is defined as critical micelle concentration (CMC). Therefore, the CMC is an important and extremely useful parameter for characterizing surfactants. Given the impact of CMC on the characteristics of surfactant, it seems obvious to pay special attention to it [4].

Anionic surfactants constitute the largest class of surfactants. Of all types of surfactants, they are the most widely used and account for about 70% of the surfactant consumption [6]. Various industrial processes are carried out with the addition of anionic surfactants. They are used as wetting agents, emulsifiers, dispersants and foaming agents. They play a major role in modern cleaning products (laundry detergents, electric dishwashing detergents, and in some shampoos) due to their superior detergency performance [7]. In addition, these compounds are often added to a variety of products, such as pharmaceuticals, antimicrobials and corrosion inhibitors [4]. From the moment that micellization process is important from both technological and environmental point of view, many researchers use quantitative structure–property relationships (QSPR) models [8] to provide early CMC estimations on the one hand and to design new surfactants with a special property on the other hand. The QSPR principle consists in finding a correlation between a property (or activity) of a substance (such as CMC) with its structural characteristics (molecular descriptors) reflecting topological, geometric and electrostatic aspects of a molecule. QSPRs are very practical methods involved in the early assessment of

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13369-020-04598-0) contains supplementary material, which is available to authorized users.

✉ S. Rahal
rahalsoufiane@yahoo.fr; rahal.soufiane@univ-medea.dz

[1] Department of Process Engineering and Environment, Faculty of Technology, Yahia Fares University of Medea, Medea, Algeria

physico-chemical and biological parameters of substances that have not been studied experimentally.

Indeed, founded over 50 years ago by Corwin Hansch, quantitative models of "structure–activity" (QSAR) relationships are widely used at university research structures, industry and several government agencies (OECD, REACH, etc.). The abundance of experimental databases for millions of products added to pressure from several organizations to minimize the use of animals has encouraged the use of QSAR/QSPR as a promising alternative in drug design, toxicology and ecotoxicological risk assessment [9, 10]. Reference works dealing with fundamental concepts of QSAR modeling and their basic concepts for applications in risk assessment are currently available in the literature [11, 12].

Recently, large amounts of QSPR approaches have been developed to estimate the physic-chemical parameters of a large group of important compounds for industrial applications [13–16]. In the past two decades, few studies have been published on the establishing of QSPR models to predict CMCs for anionic surfactants. A QSPR approach was designed to attempt to correlate the molecular structure of 119 anionic surfactants to their CMC [17]. With three descriptors containing information on the size, structure and hydrophobic properties of the surfactants studied ($n = 68$), a regression model was developed. A very good value of the coefficient of determination ($R^2 = 0.988$) was obtained. Recently, a set of 31 anionic surfactants was used to develop a QSPR model linking molecular structural parameters and log CMC [4]. The statistics of these models ($R^2 = 0.964$ and $R^2 = 0.982$) with each other with cross-validation performance confirm the aptitude of both models to predict the CMC of anionic surfactants. In conclusion, the authors suggest that the branching and the polarity of the compounds studied contribute significantly to the micellization process. However, they claim that the polarity contributes less to this process.

A QSPR study was developed for the quantification of CMC from the logP (octanol/water partition coefficient) for a various range of anionic surfactants [18]. Acceptable modeling was obtained by using two parameters, πh and L. In the same year, 37 anionic surfactants of sodium alkyl sulfates and 3 descriptors were used to develop a QSPR model for the prediction of CMC [19]. Given the results obtained, the authors suggest the use of this model in the context of the prediction of the CMC of anionic surfactants

In a report published in 2004 [20], 98 anionic surfactants and three descriptors were used to develop a QSPR model for the predicting of the CMC using multiple linear regression techniques. The anionic surfactants used include a wide variety of hydrophobic structures.

The results obtained ($R^2 = 0.980$ and $R^2_{\text{cv}} = 0.978$) indicate the robustness and friability of the QSPR approach. A good correlation between observed and predicted CMC was noticed. The authors suggested that the contribution of three parameters (total atom number, dipole moment, and max net atomic charge on C atom) is very important. Interesting QSPR models have been generated to predict the CMC of 37 anionic surfactants using two categories of descriptors [21]. The internal performance and the predictivity of models are satisfactory. The descriptors used highlighted the impact of ramification, hydrophobicity, and electronic properties of the surfactants on the micellization process.

In this context, the main objective of this work is to establish new robust QSPR models for predicting the CMC of a wide variety of surfactants (classical and extended anionic surfactants) from their molecular structure. The models developed using multiple linear regression and artificial neural network, which satisfies the guidelines required by the Organization for Economic Cooperation and Development (OECD), are based on different types of descriptors to access physically meaningful models. In addition, the developed QSPR models can be useful in the design of new anionic surfactants.

2 Methodology

2.1 Data Collection and Dataset Division

To establish high-performance QSPR models, experimental data must be of high quality [22]. In the present work, the experimental critical micelle concentration (CMC) data of 50 anionic surfactants (36 conventional anionic surfactants and 14 extended anionic surfactants) were extracted from the literature (Table 1). A wide variety of surfactant structures were included. The CMC values were measured at 25 °C in purified water without any added ingredient. The data were carefully analyzed to avoid any errors. The values of the CMC were converted to negative logarithmic scale [pCMC = $-\log_{10}$CMC (µmol/L)] to guarantee the linear distribution. The normality distribution was checked using different statistical tests and distribution plots are presented in Fig. 1a, b. The complete dataset (50 anionic surfactants) has been split into two sets, one for training and one for testing [23, 24] employing "Kennard Stone" division method using 'Dataset division GUI 1.2' tool (DTC Lab Software Tools). In this case, the best ratio is 75:25 ($n_{\text{training}} = 38$ anionic surfactants and $n_{\text{test}} = 12$ anionic surfactants).

2.2 Molecular Descriptor Calculation

There are more than 11145 usable molecular descriptors [25]. All descriptors considered in this study were computed using PaDEL-Descriptor (ver. 2.21) program. Anionic surfactant structures were saved as SMILES (Simplified Molecular Input Line-Entry System) notation, which is the recommended input format for PaDEL-Descriptor software [26].

Table 1 List of 50 anionic surfactants and their experimental pCMC values along with predicted pCMC values

Nos.	Anionic surfactants structures	pCMC (μmol/L)	
		Experimental[a]	Predicted[b]
Training set			
1	$C_{15}H_{31}CONHCH[CH(CH_3)_2]$ COO^- Na^+	3.28	3.31
2	$C_{10}H_{21}SO_3^-$ Na^+	4.63	4.52
4	$C_{12}H_{25}SO_3^-$ Li^+	4.04	3.82
5	$C_{12}H_{25}SO_3^-NH_4^+$	3.95	3.82
6	$C_{12}H_{25}SO_3^-K^+$	3.97	3.82
7	$C_{12}H_{25}SO_4^-Na^+$	3.91	3.76
8	$C_{10}H_{21}$ $CHCH_3C_6H_4SO_3^-Na^+$	2.52	2.60
9	$C_{12}H_{25}SO_4^-Li^+$	3.95	3.76
12	$C_{14}H_{29}SO_4^-Na^+$	3.32	3.42
14	$C_{10}H_{21}$ $OC_2H_4SO_4^-Na^+$	3.59	3.89
15	$C_{10}H_{21}$ $(O\ C_2H_4)_2SO_4^-Na^+$	3.46	3.29
17	$C_{10}H_{21}$ $CH(C_6H_{13})CH_2(O$ $C_2H_{4)5}SO_4^-Na^+$	1.93	2.02
19	C_8H_{17} $OOCCH_2SO_3^-Na^+$	4.82	4.80
20	$C_{10}H_{21}$ $OOCCH_2SO_3^-Na^+$	4.34	4.32
21	C_4H_9 $OOCCH_2$ $CH(SO_3^-Na^+)COOC_4H_9$	5.30	5.25
22	C_5H_{11} $OOCCH_2$ $CH(SO_3^-Na^+)COOC_5H_{11}$	4.72	4.81
23	C_6H_{13} $OOCCH_2$ $CH(SO_3^-Na^+)COOC_6H_{13}$	4.15	4.25
24	$C_4H_9CH(C_2H_5)CH_2$ $OOCCH_2CH(SO_3^-Na^+)$ $COOCH_2CH(C_2H_5)C_4H_9$	3.40	3.48
25	C_8H_{17} $OOCCH_2$ $CH(SO_3^-Na^+)COOC_8H_{17}$	2.96	2.89
26	$C_{12}H_{25}$ $CH(SO_3^-Na^+)COOC_2H_5$	3.35	3.35
27	$C_{12}H_{25}$ $CH(SO_3^-Na^+)COOC_4H_9$	3.13	3.29
28	C_7F_{15} COO^-K^+	4.46	4.31
29	C_7F_{15} COO^-Na^+	4.48	4.31
31	$(CF_3)CF(CF_2)_4COO^-Na^+$	4.48	4.31
32	C_8F_{17} COO^-Li^+	3.69	4.01
33	C_8F_{17} $SO_3^-Li^+$	3.80	4.13
34	$C_{11}H_{23}$ COO^-Na^+	3.45	3.96
35	$C_{12}H_{25}$ $(O\ C_2H_{4)5}$ $OCH_2COO^-Na^+$	3.54	3.44
36	$C_{12}H_{25}$ $(O\ C_2H_{4)9}$ $OCH_2COO^-Na^+$	3.77	3.80
37	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_5$ $CH_2CH(CH_3)SO_4^-Na^+$	3.83	3.93
38	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_9$ $CH_2CH(CH_3)SO_4^-Na^+$	4.28	4.20
41	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_6$ $(C_2H_4O)_2$ $SO_4^-Na^+$	3.40	3.53
42	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_{10}$ $(C_2H_4O)_2$ $SO_4^-Na^+$	4.05	3.97
45	$C_{12}H_{25}O[CH_2CH(CH_3)O]_{10}$ $CH_2COO^-Na^+$	3.99	4.05
46	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_{14}$ $CH_2COO^-Na^+$	4.44	4.38
47	C_9H_{20} $(C_6H_4)[O$ $CH_2CH(CH_3)]_3$ $SO_4^-Na^+$	4.23	3.97
48	C_9H_{20} $(C_6H_4)[O$ $CH_2CH(CH_3)]_6$ $SO_4^-Na^+$	4.52	4.71
49	C_9H_{20} $(C_6H_4)[O$ $CH_2CH(CH_3)]_9$ $SO_4^-Na^+$	4.80	4.91
Test set			
3	$C_{12}H_{25}SO_3^-$ Na^+	4.08	3.82
10	$C_{12}H_{25}SO_4^-$ $N(CH_{3)4}^{+}$	3.74	3.76
11	$C_{12}H_{25}SO_4^-$ $N(C_3H_{7)4}^{+}$	3.34	3.76
13	$C_{10}H_{21}$ $OC_2SO_3^-Na^+$	4.18	4.38
16	$C_{10}H_{21}$ $(O\ C_2H_{4)5}SO_4^-Na^+$	1.40	1.65
18	C_6H_{13} $OOCCH_2SO_3^-Na^+$	5.23	5.05
30	C_7F_{15} COO^-Li^+	4.52	4.31
39	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_{13}$ $CH_2CH(CH_3)SO_4^-Na^+$	4.89	4.53
40	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_6$ $(C_2H_4O)_2CH_2COO^-Na^+$	3.77	3.71
43	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_6$ $CH_2COO^-Na^+$	3.54	3.59
44	$C_{12}H_{25}$ $O[CH_2CH(CH_3)O]_{14}$ $(C_2H_4O)_2$ $SO_4^-Na^+$	4.29	4.31
50	C_9H_{20} $(C_6H_4)[O$ $CH_2CH(CH_3)]_{12}$ $SO_4^-Na^+$	4.92	5.03

[a] Experimental pCMC values are collected from following literatures: [24, 47, 48]

[b] pCMC values predicted by MLP/ANN model

In this work, 1543 molecular descriptors were calculated for each surfactant.

2.3 Molecular Descriptor Selection

One of the important steps in QSPR modeling is the reduction of the number of descriptors. This reduction has a twofold purpose: to avoid the phenomenon of overfitting and to reduce the risk of obtaining a model by chance [27]. To keep only the most important descriptors, the selection procedure described previously in one of our articles was used [22]. The number of descriptors obtained after the selection was 868. Then, the GA approach was employed (http://teqip.jdvu.ac.in/QSAR_Tools/).

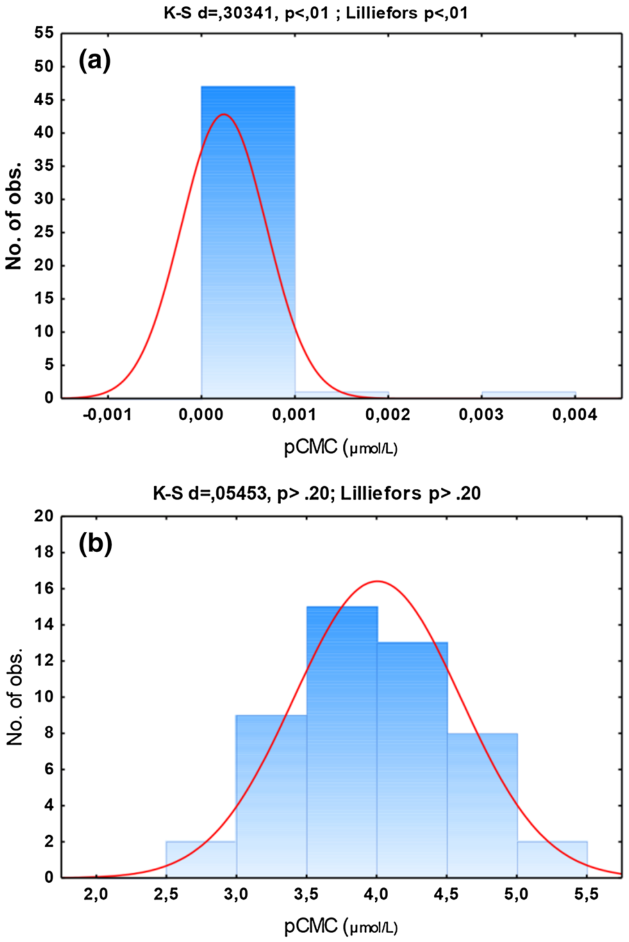

Fig. 1 Normality distribution plot of CMC data: **a** before transformation, **b** after log transformation (pCMC)

2.4 QSPR Model's Development and Validation

In this present study, models are developed employing three statistical methods: (1) multiple linear regressions (MLR) approach; (2) partial least square (PLS) approach; (3) multilayer perceptron–artificial neural networks (MLP/ANN) with BFGS (Broyden–Fletcher–Goldfarb–Shanno) as learning algorithms. For MLP/ANN approach, we have employed STATISTICA software (STATISTICA 8.0, Tulsa; StatSoft, Inc.). For second and third methods, we have employed MLR Plus Validation GUI 1.3 tools and Partial Least Squares v1.0, respectively (DTC Lab Software Tools). Theories and applications about the MLP–ANN have been reported in the literature [28, 29].

To assess model's quality and predictability, validation (internal and external) is a necessary step. For internal validation, traditional validation metrics recommended by leading research groups [30, 31] were checked: the determination coefficient (R^2), and the cross-validated correlation coefficient (Q^2_{LOO}) along with novel validation parameters (r^2_m; Δr^2_m). The external validation was performed on the test set by calculating the following parameters: Q^2_{F1}, Q^2_{F2}, average r^2_m, and Δr^2_m. The equations of these validation statistical parameters are provided in the literature [32–35]. In a recent work, Roy and his collaborators [36] suggested adding a parameter for external validation. This parameter concerns to establish of a threshold for the mean absolute error (MAE). Thus, to assess the predictive performance of QSAR models with a higher degree of confidence, we have calculated and verified the criteria based on the MAE as suggested by Roy et al. [36]. A tool (XternalValidationPlus) for computing the suggested MAE based criteria for external validation is accessible online [37].

Further, Xternal Validation Plus has been used to check the presence of high systematic errors (bias) in the ANN model. If the bias is present in the model, then it should be discarded and performing any external validation test is useless on such a biased model [38].

2.5 QSPR Model's Applicability Domain

In addition to internal and external validation, the determination of applicability domain (AD) is therefore of great importance [39]. The AD is another validation measure to check the friability of QSPR models. The QSPR model cannot be applied outside of the chemical space covered by the training set. Therefore, extrapolation is not allowed and can lead to erroneous model predictions [40]. To investigate the AD of the anionic surfactants used in this study, the Williams plot was established [22, 41].

3 Results and Discussion

3.1 Molecular Descriptor Selection

To select the optimal number of suitable descriptors, the effect of the number of descriptors on the statistic parameters (R^2, Q^2, $\bar{r}^2_m$, Δr^2_m) was investigated for 2–7 descriptors. The results are shown as plots of R^2, Q^2, $\bar{r}^2_m$, and Δr^2_m for the training set as a function of the number of descriptors for the 2–7 descriptors in the model (Fig. 2). Analyzing Fig. 2 and Table 2 (prediction quality), we can confirm that the best numbers of descriptors are 4 and 5. However, it is plausible to note that there is no significant improvement in their statistical parameters. For this reason, we have chosen the following 4 descriptors: ATSC7v; ATSC5e; nAtomLAC; ETA_Epsilon_3 (Table S1 in supplementary files).

The correlation matrix of four relevant descriptors has been obtained and is presented in the additional files of Table S2. From this table, the 4 relevant variables (descriptors) are independent of the fact that each pair of descriptors has a correlation coefficient value of less than 0.57.

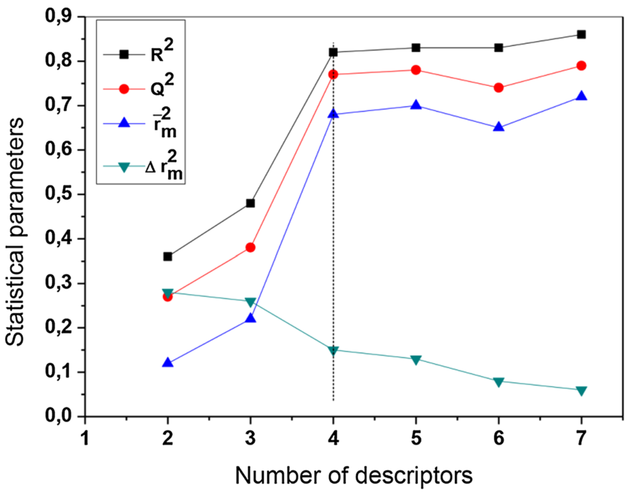

Fig. 2 Influences of the number of descriptors on statistical parameters

3.2 MLR Model

As mentioned in the methodology section, the MLR model was developed based on 4 relevant molecular descriptors. The final equation of the MLR model is as follows (Eq. 1):

$$
\begin{aligned}
pCMC = {} & 76.89501\,(\pm 15, 94778) \\
& + 0.00041\,(\pm 0, 00007)\ ATSC7v \\
& - 0.13615\,(\pm 0.02508)\ ATSC5e \\
& - 165.43679\,(\pm 37.07955)\ ETA\_Epsilon\_3 \\
& - 0.18446\,(\pm 0.01934)\ nAtomLAC
\end{aligned}
\tag{1}
$$

$n_{\text{train}} = 38\ R^2 = 0.80\ Q^2_{\text{LOO}} = 0.72\ \overline{r^2_{m(\text{scaled})}} = 0.62$
$\Delta r^2_{m(\text{scaled})} = 0.16$
$n_{\text{test}} = 12\ Q^2_{F1} = 0.87\ Q^2_{F2} = 0.87\ \overline{r^2_{m(\text{scaled})}} = 0.64$
$\Delta r^2_{m(\text{scaled})} = 0.15$
$Q^2 = 0.72$, (Threshold value $Q^2 > 0.5$), *Passed*
$r^2 = 0.92$, (Threshold value $r^2 > 0.6$), *Passed*
$|r_0^2 - r_0'^2| = 0.11$ (Threshold value $|r_0^2 - r_0'^2| < 0.3$), *Passed*
$[(r^2 - r_0^2)/r^2] = 0.05 < 0.1$ or, $[(r^2 - r_0^{2'})/r^2] = 0.16 < 0.1$, *Passed*
$0.85 \le k = 1.00 \le 1.15$ or $0.85 \le k' = 0.99 \le 1.15$, *Passed*

The statistical parameters values indicate the robustness and friability of the MLR model. The predicted pCMC values of the surfactants studied as well as the values of the descriptors of the model are presented in Table S3 (supplementary files).

According to the recommendation of Tropsha and Golbraikh [42, 43], if the difference between R^2 and Q^2_{LOO} is less than 0.3, the model is without overfitting. In the actual study, R^2-$Q^2_{\text{LOO}} = 0.08$, indicating no overfitting in the MLR model. Moreover, the concrete prediction error of the model is estimated by the PRESS parameter value [44]. To have a credible QSPR model, the PRESS/SSY ratio should be smaller than 0.4. As part of this study, the PRESS/SSY ratio was equal to 0.25 (PRESS = 3.32 and SSY = 13.19), so this proves that the developed model predict is better than chance. In addition, to confirm the absence of a chance factor during the development of the MLR model, a Y-randomization analysis was performed by generating 50 random models. The average values of R^2 and Q^2 obtained (0.12 and − 0.17) are below than the acceptable limit of 0.5 for both parameters.

The impact of a descriptor in a model is characterized by its sign-in model mathematical equation. According to the regression coefficients of Eq. (1), the ETA_Epsilon_3 descriptor was the main contributor to the CMC of anionic surfactants. This descriptor has highest and negative contribution and therefore has a negative impact on CMC. Thus, for specified anionic surfactants, low values of ETA_Epsilon_3 descriptor would help in improving its CMC. In addition, the regression coefficients of the descriptors ATSC5e and nAtomLAC had negative signs, thus giving a negative impact on CMC. On the contrary, ATSC7v has positive contribution

Table 2 Calculated parameters for selection of optimal descriptors

Number of descriptors	Descriptors	R^2	Q^2	$\overline{r}_m^2$	Δr_m^2	Prediction quality
2	GATS5s; nAtomLAC	0.36	0.27	0.12	0.28	Bad
3	GATS5s; nAtomLAC; GATS3p	0.48	0.38	0.22	0.26	Bad
4	ATSC7v; ATSC5e; nAtomLAC; ETA_Epsilon_3	**0.82**	**0.77**	**0.68**	**0.15**	**Good**
5	ATSC5m; nAtomLAC; VE2_DzZ; ATSC6m; ATSC7e;	0.83	0.78	0.70	0.13	Good
6	ATSC6s; GATS5s; nAtomLAC; GATS3p; ATSC6m; ATSC2i; piPC10.	0.83	0.74	0.65	0.08	Moderate
7	ATSC2i; GATS3p; nAtomLAC; ATSC6m; GATS5s; ATSC6s; piPC10	0.86	0.79	0.72	0.06	Moderate

The bold signifies the best model chosen for the rest of the study towards CMC and the highest values were conducive to the improvement of the CMC of anionic surfactants.

3.3 PLS Model

The results obtained with PLS model for the prediction of CMC of anionic surfactants, using 50 compounds, are summarized below:

3.3.1 Internal Validation Parameters

R^2 (Train): 0.79896
Q^2 (LOO): 0.72238
Scaled average R_m^2 (train; LOO): 0.6165
Scaled Delta R_m^2 (train; LOO): 0.16275
Mean absolute errors (MAE; train):0.2673
Standard deviation of absolute errors (SD; train):0.2247
Training set prediction quality (based on MAE-based criteria*): MODERATE

3.3.2 External Validation Parameters

Q_{F1}^2: 0.87167
Q_{F2}^2: 0.87034
Scaled average R_m^2 (test): 0.63903
Scaled Delta R_m^2 (test): 0.15178
CCC (test): 0.91628
Standard deviation of absolute errors (SD; test): 0.0952
Test set prediction quality (based on MAE-based criteria*): good

3.4 MLP/ANN Model

In this investigation, the learning algorithm used to develop an MLP/ANN nonlinear model to predict the critical micelle concentration (CMC) of anionic surfactants is called BFGS. The database has been divided into a training set (75%) and a test set (25%). the ANN network selected for this study is the multilayer perceptron (with an input layer, a hidden layer and an output layer). Several studies have shown that this category of the network is able to model any activity (or property) of a substance whatever its complexity [28]. One output neuron was used to represent the predicted pCMC. The two transfer functions used in this study are the hyperbolic tangent (tanh) and the identity function, respectively. Furthermore, the following rule has been taken into account to optimize the number of neurons in the hidden layer:

$$\begin{aligned}&\left[\,(\text{Number of input neurons} \times \text{number of hidden neurons})\right.\\&\quad + (\text{number of hidden neurons}\\&\quad \left.\times \text{ number of output neurons})\,\right] \leq (\text{size of database})\end{aligned}$$

In order to ensure the best possible model, many trials sometimes involving more than 800 iterations were carried out. The model with the lowest value of the RMSE was selected [28]. Then, the best model with MLP/ANN architecture {4-3-1} was selected.

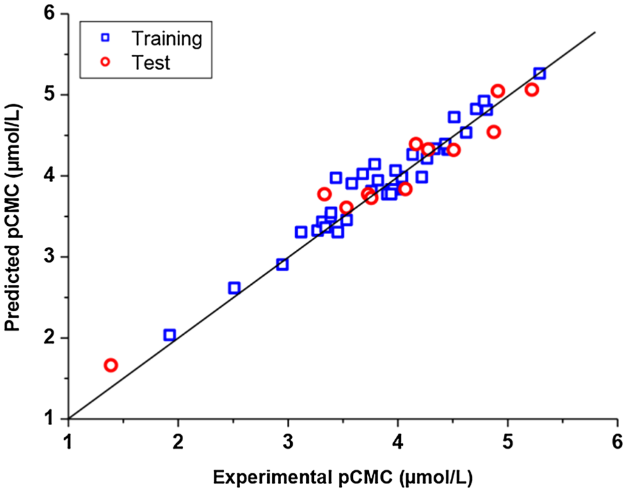

Fig. 3 Scatter plot of the predicted values of logCMC versus the experimental values by ANN model for the training, and test set

The predictive pCMC from the MLP/ANN model for 50 anionic surfactants is given in Table 1. The observed versus predicted pCMC of the training and test set is shown in Fig. 3. From this figure, a close correlation between the predicted and observed values of pCMC was obtained. The values of the validation (internal and external) statistical parameters reported in Table 3 comply with the acceptability criteria [29], suggesting that the MLP/ANN model is robust and provides excellent predictive quality.

Due to the complexity of the relationship between the predicted property and the descriptors (variables) in an ANN model [45], the effect of variables in the micellization process is relatively easy to interpret in the case of linear regression. The relative contribution [46] of the MLP/ANN model descriptors was calculated and is represented in Fig. 4. The importance of these descriptors decreases in the order: ATSC7v > nAtomLAC > ATSC5e > ETA_Epsilon_3.

ATSC7v (weighted by van der Waals volume) belongs to the 2D autocorrelation descriptors. This descriptor describes the distribution of van der Waals volume with a lag of 7 along the topological structure of the anionic surfactants. The physic-chemical significance of the descriptor ATSC7v concerns the volume of the molecule. Thus, the increase in the volume of a molecule leads to the increase in the value of ATSC7v. The second descriptor in MLP/ANN model was nAtomLAC, which involves the number of atoms in the longest aliphatic chain. ATSC5e (weighted by Sanderson electronegativity) and ETA_Epsilon_3 (Extended Topochemical Atom descriptor) are the third and fourth descriptors in the MLP/ANN model. All these quantities are well defined in the literature.

Table 3 Statistical quality of all developed QSPR models

Model	Internal validation parameters					External validation parameters					
	R^2	R^2_{adj}	Q^2_{LOO}	$\overline{r^2_{m(\text{training})}}$	$\Delta r^2_{m(\text{training})}$	Q^2_{F1}	Q^2_{F2}	$\overline{r^2_{m}}_{(\text{test})}$	Δr^2_{m} (test)	MAE (95% data)	MAE + 3σ (95% data)
MLP/ANN	0.94	0.93	0.93	0.89	0.06	0.95	0.95	0.87	0.15	0.1564	0.4935
PLS	0.79	0.76	0.72	0.61	0.15	0.87	0.87	0.63	0.16	0.2590	0.5446
MLR	0.80	0.77	0.72	0.62	0.16	0.87	0.87	0.64	0.15	0.2055	0.6336

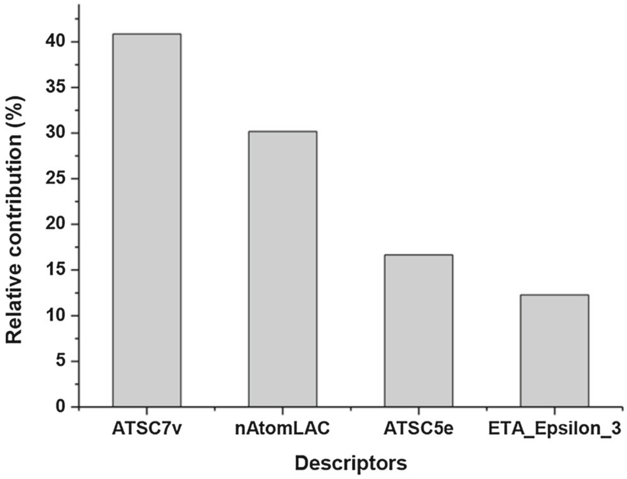

Fig. 4 Plot of the fraction contribution of the descriptors to the pCMC of anionic surfactant

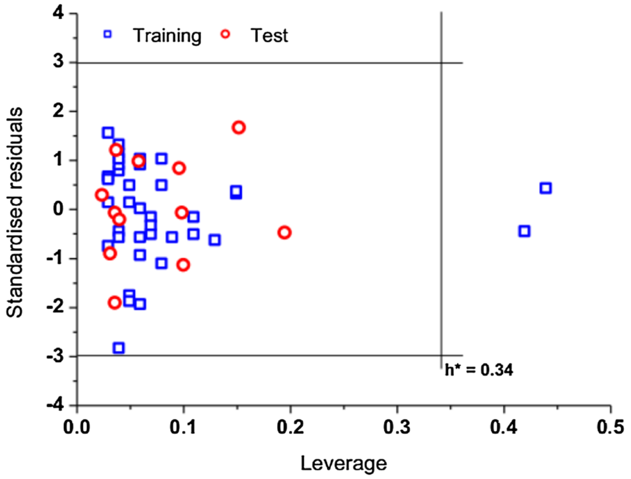

Fig. 5 Projection of the training, and test set of anionic surfactants in the Williams plot

As can be seen in Fig. 4, the CMC widely depends on the two descriptors ATSC7v and nAtomLAC. ATSC7v and nAtomLAC which accounts, respectively, 40.85% and 30.18% of the total contribution. The remaining 28.97% is from ATSC5e (16.68%), and ETA_Epsilon_3 (12.29%). Summarizing, it can be concluded that atomic electronegativity, molecular size, and the number of atoms in the longest aliphatic chain, all play an important role in micellization of anionic surfactants.

3.5 Statistical Comparison of the QSPR Models

Comparative statistics of the MLR, PLS, and MLP/ANN regression models is presented in Table 3. For each model, we used the same type and number of descriptors, as well as the same composition of the training and test sets. From Table 3, it is observed that all the reported models (MLP/ANN, MLR, and PLS) are of acceptable quality. Among all three regression models, MLP/ANN model shows the highest values for quality parameters, i.e., R^2 (0.94), R^2_{adjusted} (0.93), Q^2_{LOO} (0.93), and Q^2_{F1} (0.95). In addition, the MLP/ANN model exhibits an improvement in terms of external statistics compared to the PLS and MLR models.

The MAE-based metrics (MAE and MAE + 3 × σ) after omitting 5% data points with high prediction residuals estimated that the predictions of the ANN model are classified as 'good' (see results in supplementary file) which is also in agreement with the judgment provided by the classical metrics for external validation (Table 3). In addition, the output file of Xternal Validation tool Plus (see results in supplementary file) indicated the absence of systematic error (bias) in the ANN model.

3.6 Applicability Domain Investigation

After the validation of a model, the domain of applicability (third principle of the OECD) must be established. As part of this study, the applicability domain of the MLP/ANN model was determined based on the Williams plot (Fig. 5). The computed threshold leverage (h^*) is 0.34. As shown in Fig. 5, none of the 50 surfactants in the model are outside the range of ± 3 standard deviation units. Also, compound 24 [$C_4H_9CH(C_2H_5)CH_2OOCCH_2$ $CH(SO_3{}^-Na^+)COOCH_2CH(C_2H_5)C_4H_9$] and compound 25 [$C_8H_{17}$ $OOCCH_2$ $CH(SO_3{}^-Na^+)COOC_8H_{17}$] are outside the applicability domain (with $h > h^*$). Thus, 96% of surfactants belong to the applicability domain and therefore was covered by the MLP/ANN model. Fortunately, in this work, the data predicted by the MLP/ANN model is good for these compounds; thus, these are “good leverage” chemicals, implying that these compounds were very influential on the model, and can stabilize the QSPR model and make it more precise. Consequently, Williams’s plot provides the acceptance of built MLP/ANN model to predict the CMC. In conclusion, we can assert that the MLP/ANN model adheres to the third OECD principle.

3.7 Comparison with Previously Reported Models

The statistical results of the MLP/ANN model were compared with those of some previously developed QSPR models (Table 4). In Table 4, it can be observed that no applicability domain according to the OECD guidelines has been determined and no external quality measurement approach has been carried out in the other models with the exception of the model developed by Roy and Kabir [21]. Also, unlike other models, the MLP/ANN model and that of Roy and Kabir [21] are those that offer a better predictive power. If the statistical parameters of the internal validation are almost identical, our model slightly exceeds the model of Roy and Kabir [21] in terms of external validation, since the criteria based on the MAE have not been verified. It should be noted, however, that the results provided by Roy are those obtained by the linear regression-based techniques. We can conclude that the MLP/ANN model developed in this work is encouraging and can therefore be used for the determination of CMC of new surfactants, thus contributing to substantial amounts of money and time.

4 Conclusions

For the prediction of the CMC values for anionic surfactants, three regression methods were utilized (MLR, PLS and MLP/ANN) to develop robust predictive models. The proposed models trained and validated using a dataset comprised of 50 anionic surfactants were based on four molecular descriptors. By applying all available validation strategies, we were able to deduce that the models adopted were robust for both internal validation and external validation parameters. The multilayer perceptron–artificial neural network model (MLP/ANN) trained with the Broyden–Fletcher—Goldfarb–Shanno (BFGS) algorithm gave better performance in CMC predictions with a higher Q^2_{ext} and $\overline{r^2_m}$ values (0.95 and 0.87) and acceptable Δr^2_m value (0.15) for testing dataset compared to that of previously reported models. The MAE-based metrics estimated that the MLP/ANN model shows ‘GOOD’ predictions (after removing 5% test set objects with high residual values).

Table 4 Comparison of the results of internal and external validation of our best model (MLP/ANN) with previously published models

Models	Type of descriptor used	N_{total}	Internal validation				External validation				
			R^2	Q^2_{LOO}	$\overline{r^2_m}$	Δr^2_m	Q^2_{ext}	$\overline{r^2_m}$	Δr^2_m	MAE (95% data)	MAE +3σ (95% data)
Present work (MLP/ANN)	Electronic and structural	**50**	**0.94**	**0.93**	**0.89**	**0.06**	**0.95**	**0.87**	**0.15**	**0.1564**	**0.4935**
Roy and Kabir [16]	ETA and hydrophobicity	37	0.96	0.94	0.87	0.05	0.92	0.82	0.07	The MAE based criteria was not used	
Li et al. [15]	Electronic and constitutional	98	0.98	0.98	–	–	The external validation was not used				
Yuan et al. [43]	Thermodynamic. Electronic. and constitutional	37	0.99	–	–	–	The external validation was not used				
Wang et al. [14]	KHO. ET. Delta H(f).D. E-LUMO and E-HOMO	40	0.98	–	–	–	The external validation was not used				
Jalili-Heravi and Konouz [7]	Topological. geometric. and electronic	31	0.99	–	–	–	The external validation was not used				
Huibers et al. [12]	Constitutional. topological. and quantum chemical.	119	0.94			–	The external validation was not used				

The bold refers to the best model among the models reported in Table 4

By studying the properties of the four descriptors used to develop QSPR models, it appears that the length of the aliphatic chain, the electronic properties (electronegativity), and the structure of the molecules play a crucial role in the micellization process. In conclusion, the QSPR model developed in this work is in line with OECD principles and is useful to provide early CMC estimations on the one hand and to design new surfactants with a special property on the other hand.

References

1. Gaudin, T.; Rotureau, P.; Pezron, I.; Fayet, G.: New QSPR models to predict the critical micelle concentration of sugar-based surfactants. Ind. Eng. Chem. Res. **55**(45), 11716–11726 (2016)
2. Wang, Z.W.; Li, G.Z.; Zhang, X.Y.; Li, L.: Prediction on critical micelle concentration of anionic surfactants in aqueous solution: quantitative structure–property relationship approach. Acta. Chim. Sin. **60**(9), 1548–1552 (2002)
3. Mozrzymas, A.; Różycka-Roszak, B.: Prediction of critical micelle concentration of cationic surfactants using connectivity indices. J. Math. Chem. **49**(1), 276–289 (2010)
4. Jalali-Heravi, M.; Konouz, E.: Prediction of critical micelle concentration of some anionic surfactants using multiple regression techniques: a quantitative structure–activity relationship study. J. Surfactants Deterg. **3**(1), 47–52 (2000)
5. Saunders, R.A.; Platts, J.A.: Correlation and prediction of critical micelle concentration using polar surface area and LFER methods. J. Phys. Org. Chem. **17**(5), 431–438 (2004)
6. Kronberg, B.; Holmberg, K.; Lindman, B.: Types of surfactants, their synthesis, and applications. In: Surface Chemistry of Surfactants and Polymers, 1st edn. Wiley, Berlin (2014)
7. Gwaltney-Brant, S.M.: Miscellaneous indoor toxicants. J. Small. Anim. Pract. **2013**, 291–308 (2013)
8. Nieto-Draghi, C.; Fayet, G.; Creton, B.; Rozanska, X.; Rotureau, P.; deHemptinne, J.C.; Adamo, C.A.: General guidebook for the theoretical prediction of physicochemical properties of chemicals for regulatory purposes. Chem. Rev. **115**(24), 13093–13164 (2015)
9. Dearden, J.C.: The history and development of quantitative structure–activity relationships (QSARs). Int. J. Quant. Struct. Prop. Relationsh. **1**, 1–44 (2016)
10. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Tropsha, A.: QSAR modeling: Where have you been? Where are you going to? J. Med. Chem. **57**(12), 4977–5010 (2014)
11. Roy, K.; Kar, S.; Das, R.N.: Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, pp. 1–46. Academic Press, Berlin (2015)
12. Roy, K.; Kar, S.; Das, R.N.: A Primer on QSAR/QSPR Modeling: Fundamentals Concepts (Springer Briefs in Molecular Science), pp. 1–35. Springer, Berlin (2015)
13. Haratipour, P.; Baghban, A.; Mohammadi, A.H.; Nazhad, S.H.; Bahadori, A.: On the estimation of viscosities and densities of CO_2-loaded MDEA, MDEA + AMP, MDEA + DIPA, MDEA + MEA, and MDEA + DEA aqueous solutions. J. Mol. Liq. **242**, 146–159 (2017)
14. Safder, U.; Nam, K.; Kim, D.; Shahlaei, M.; Yoo, C.: Quantitative structure–property relationship (QSPR) models for predicting the physicochemical properties of polychlorinated biphenyls (PCBs) using deep belief network. Ecotoxicol. Environ. Saf. **162**, 17–28 (2018)
15. Fariba, Z.; Baghban, A.: Phase behavior modelling of asphaltene precipitation utilizing MLP–ANN approach. Pet. Sci. Technol. **35**, 2009–2015 (2017)
16. Olguin, C.J.M.; Sampaio, S.C.; Do-Reis, R.R.; Remor, M.B.; Olguin, C.F.A.: QSPR modelling of the soil sorption coefficient from training sets of different sizes. SAR. QSAR Environ. Res **30**(5), 299–311 (2019)
17. Huibers, P.D.T.; Lobanov, V.S.; Katritzky, A.R.; Shah, D.O.; Karelson, M.: Prediction of critical micelle concentration using a quantitative structure–property relationship approaches. 2. Anionic surfactants. J. Colloid Interface Sci. **187**, 113–120 (1997)
18. Roberts, D.W.: Application of octanol/water partition coefficients in surfactant science: a quantitative structure–property relationship for micellization of anionic surfactants. Langmuir **18**(2), 345–352 (2002)
19. Yuan, S.; Cai, Z.; Xu, G.; Jiang, Y.: Quantitative structure–property relationship of surfactants: critical micelle concentration of anionic surfactants. J. Dispers. Sci. Technol. **23**, 465–472 (2002)
20. Li, X.; Zhang, G.; Dong, J.; Zhou, X.; Yan, X.; Luo, M.: Estimation of critical micelle concentration of anionic surfactants with QSPR approach. J. Mol. Struct. **710**(1–3), 119–126 (2004)
21. Roy, K.; Kabir, H.: QSPR with extended topochemical atom (ETA) indices: exploring effects of hydrophobicity, branching and electronic parameters on logCMC values of anionic surfactants. Chem. Eng. Sci. **8**(7), 141–151 (2013)
22. Hamadache, M.; Benkortbi, O.; Hanini, S.; Amrane, A.: QSAR modeling in ecotoxicological risk assessment: application to the prediction of acute contact toxicity of pesticides on bees (*Apis mellifera* L.). Environ. Sci. Pollut. Res. **25**(1), 896–907 (2017)
23. Martin, T.M.; Harten, P.; Young, D.M.; Muratov, E.N.; Golbraikh, A.; Zhu, H.; Tropsha, A.: Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. **52**(10), 2570–2578 (2012)
24. Roy, P.P.; Leonard, J.T.; Roy, K.: Exploring the impact of size of training sets for the development of predictive QSAR models. Chemometr. Intell. Lab. Syst. **90**(1), 31–42 (2008)
25. Roubehie Fissa, M.; Lahiouel, Y.; Khaouane, L.; Hanini, S.: QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP–ANN methods. J. Mol. Graph. Model. **87**, 109–120 (2018)
26. Yap, C.W.: PaDEL-descriptor: anopen source software to calculate moleculardescriptors and fingerprints. J. Comput. Chem. **32**(7), 1466–1474 (2010)
27. Khan, K.; Benfenati, E.; Roy, K.: Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. Ecotoxicol. Environ. Saf. **168**, 287–297 (2019)
28. Hamadache, M.; Benkortbi, O.; Hanini, S.; Amrane, A.; Khaouane, L.; Si Moussa, C.: A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction. J. Hazard. Mater. **303**, 28–40 (2016)
29. Hamadache, M.; Hanini, S.; Benkortbi, O.; Amrane, A.; Khaouane, L.; Si Moussa, C.: Artificial neural network-based equation to predict the toxicity of herbicides on rats. Chemometr. Intell. Lab. Syst. **154**, 7–15 (2016)
30. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S.: QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. J. Comput. Chem. **34**(24), 2121–2132 (2013)
31. Chirico, N.; Gramatica, P.: 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**, 2044–2058 (2012)
32. Chirico, N.; Gramatica, P.: Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation cri-

teria and proposal of using the concordance correlation coefficient. J. Chem. Inf. Model. **51**, 2320–2335 (2011)

33. Ojha, P.K.; Mitra, I.; Das, R.N.; Roy, K.: Further exploring RM2 metrics for validation of QSPR models. Chemometr. Intell. Lab. Syst. **107**, 194–205 (2011)
34. Roy, K.: On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opin. Drug Discov. **2**, 1567–1577 (2007)
35. Tropsha, A.: Best practices for QSAR model development, validation, and exploitation. Mol. Inform. **29**, 476–488 (2010)
36. Roy, K.; Das, R.N.; Ambure, P.; Aher, R.B.: Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometr. Intell. Lab. Sys. **152**, 18–33 (2016)
37. XternalValidationPlus: An online tool for computing the suggested MAE based criteria for external validation is accessible from the link. http://dtclab.webs.com/software-tools. http://teqip.jdvu.ac.in/QSAR_Tools/
38. Roy, K.; Ambure, P.; Aher, R.B.: How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? Chemometr. Intell. Lab. Sys. **162**, 44–54 (2017)
39. OECD: Principles for the validation, for regulatory purposes, of (quantitative) structure–activity relationship models (2009)
40. Chen, J.W.; Li, X.H.; Yu, H.Y.; Wang, Y.N.; Qiao, X.L.: Progress and perspectives of quantitative structure–activity relationships used for ecological risk assessment of toxic organic compounds. Sci. China B **51**(7), 593–606 (2011)
41. Gramatica, P.; Cassani, S.; Roy, P.P.; Kovarich, S.; Yap, C.W.; Papa, E.: QSAR modeling is not “push a button and find a correlation”: a case study of toxicity of (benzo-)triazoles on algae. Mol. Inform. **31**(11–12), 817–835 (2012)
42. Tropsha, A.; Gramatica, P.; Gombar, V.K.: The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. **22**, 69–77 (2003)
43. Golbraikh, A.; Shen, M.; Xiao, Z.Y.; Xiao, Y.D.; Lee, K.H.; Tropsha, A.: Rational selection of training and test sets for the development of validated QSAR models. J. Comput. Aided Mol. Des. **17**, 241–253 (2003)
44. Clementi, M.; Clementi, S.; Fornaciari, M.; Orlandi, F.; Romano, B.: The GOLPE procedure for predicting olive crop production from climatic parameters. J. Chemom. **15**, 397–404 (2001)
45. Katritzky, A.R.; Pacureanu, L.M.; Slavov, S.H.; Dobchev, D.A.; Karelson, M.: *QSPR* study of critical micelle concentrations of nonionic surfactants. Ind. Eng. Chem. Res. **47**(23), 9687–9695 (2008)
46. Zheng, F.; Bayram, E.; Sumithran, S.P.; Ayers, J.T.; Zhan, C.G.; Schmitt, J.D.; Dwoskin, L.P.; Crooks, P.A.: QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release. Bioorg. Med. Chem. **14**, 3017–3037 (2006)
47. Fernández, A.; Scorzza, C.; Usubillaga, A.; Salager, J.L.: Synthesis of new extended surfactants containing a carboxylate or sulfate polar group. J. Surfactants Deterg. **8**(2), 187–191 (2005)
48. Zhi-qiang, H.; Mei-jun, Z.; Yun, F.; Guang-yong, J.; Ji, C.: Extended surfactants: a well-designed spacer to improve interfacial performance through a gradual polarity transition. Colloids Surf A. Physicochem. Eng. Asp. **450**, 83–92 (2014)