

# Computational QSPR Model to Predict the Critical Micelle Concentration (CMC) of Classic and Extended Sugar-Based Surfactants.

Boukelkal Nada<sup>\*1</sup>, Rahal Soufiane<sup>1</sup>, Rebhi Redha<sup>1,2</sup>, Hamadache Mabrouk<sup>1</sup>, and Ibrir Abdellah<sup>1</sup>

<sup>1</sup>Biomaterials and Transport Phenomena Laboratory (LBMPT), University of Yahia Fares, Faculty of Technology, Department of Process and Environmental Engineering, Medea 26000, Algeria

<sup>2</sup>Department of Mechanical Engineering, Faculty of Technology, University of Medea, Medea 26000, Algeria

Received: 24-10-2023

Accepted: 04-01-2024

Published: 26-02-2024

**Abstract:** As known, critical micelle concentration is a crucial characteristic in product formulation. In the current work, six molecular structure descriptors were used to create QSPR models that predicted the critical micelle concentration (CMC) of 119 sugar-based surfactants. The analysis of the qualities of descriptors shows that the micellization process is specifically affected by electronic properties (electronegativity and charges), electro-topology, and symmetry of a molecule. Four statistical learning techniques including Multiple linear regression, Partial least square, Artificial neural networks (ANN), and Adaptive neuro-fuzzy inference system to develop the QSPR models. different statistical metrics were employed to evaluate the reliability and robustness of the models. The best result ( $\bar{r}_m^2 = 0.803$ ,  $Q_{loo}^2 = 0.856$ ,  $Q_{F1}^2 = 0.982$ , and  $\Delta r_m^2 = 0.006$ ) were obtained for ANN with {6-6-1} architecture. In addition, estimating the CMC of 6 other sugar surfactants based on simulate of the network gave very good results ( $R = 0.96$ ). Therefore, these findings suggest that the developed model is appropriate for predicting and correlating CMC value for sugar-based surfactants.

**Keywords:** ANN, ANFIS, CMC, QSPR, Sugar-based surfactant.

**Tob Regul Sci.™ 2024 ;10(1): 988-1009**

**DOI : [doi.org/10.18001/TRS.10.1.64](https://doi.org/10.18001/TRS.10.1.64)**

## 1. Introduction

Surfactants are amphiphilic molecules, that is to say, composed of two parts: a polar head that is hydrophilic and a hydrophobic alkyl chain, thus allowing them to adsorb at the immiscible bulk phases interface and self-assemble, in the form of aggregates commonly called micelles[1–3]. These properties make it possible to envisage a wide range of applications, such as detergent, cosmetics, food processing, or even the pharmaceutical field[2,4]. Depending on the nature of the hydrophilic head, four major surfactant groups can be distinguished: anionic (with a negative charge), cationic (with a positive charge), nonionic (uncharged), and amphoteric (zwitterions carrying both positive and negative charges)[1].

---

\* Author to whom correspondence should be addressed: [nadaboukelkal@gmail.com](mailto:nadaboukelkal@gmail.com) ;  
[boukelkal.nada@univ-medea.dz](mailto:boukelkal.nada@univ-medea.dz)

Nonionic surfactants, such as sugar-based surfactants, have polar heads consisting of carbohydrates[1] like glucose[5], maltose[6], or sucrose[7] and their derivatives. Sugar-based surfactants have a green chemistry character as lower toxicity, are biocompatible, and easily biodegradable, and may thus be made from renewable resources[3,8,9]. Therefore, the pharmaceutical and cosmetics sectors have more favorable profiles for using sugar-based surfactants[3,10].

A fundamental characteristic of surfactants is their CMC, which is the concentration of surfactants at which micelles begin to form in solution[3,11]. The polar heads always remain in contact with the aqueous phase, whereas the alkyl chains are housed in the micellar core during the formation of micelles to reduce the contact between the hydrophobic part and water molecules[12,13]. Experimentalists observed, that the CMC of non-ionic surfactants is affected by polar head size, alkyl chain size, and unsaturated or branched chains[2,13–15]. Currently, CMC can be determined using a variety of methods, the most common of which is tensiometry, which involves plotting the surface tension of surfactants versus log concentration value[1,16]. In silico prediction approaches like QSPR have an early estimation of CMC based on the knowledge of molecular structure to decrease the time and expense of experimental screening[1,17–19]. The QSPR model establishes a mathematical correlation between a molecular structure and a specific property using molecular descriptors[1].

Few QSPR models consider taking into account sugar-based surfactants [1,3,13], even though several QSPR models have been developed to estimate CMC [11,20–25]. Gaudin et al.[1] developed a QSPR model using the best multiple linear regression of 83 sugar-based surfactants with an  $RMSE$ ,  $R^2_{training}$ ,  $Q^2_{Loo}$  and  $R^2_{test}$  of 0.32, 0.93, 0.9, and 0.91 respectively. Additionally, Wang et al.[3] created an MLR model of 83 sugar surfactants. Indeed, they obtained  $MAE$  after removing 5% high residual compounds reaching 0.19. Baghban et al.[13] also investigated the same data set as Gaudin et al. and Wang et al. with least squares support vector machine (LLSVM) and reported the  $RMSE$  value of 0.02.

This review aims to create a new QSPR model that could be used to predict CMC from the molecular structure of a range of surfactants, including classical and extended sugar-based surfactants. Artificial neural networks (ANN), Multiple linear regression, Partial least square, and Adaptive neuro-fuzzy inference system were the statistical methods employed to develop the QSPR models, which were built following the Organization for Economic Cooperation and Development guidelines.

## 2. Methodology

### 2.1 Data collection and dataset division

It is generally recognized that the creation of high-quality QSPR models requires high-quality experimental data [26]. The dataset for the current investigation included 125 sugar-based surfactants (109 conventional sugar surfactants and 16 extended sugar surfactants). The CMC of these surfactants was determined at temperatures close to room temperature (20–25°C). The CMC values were adjusted to a negative logarithmic scale [ $pCMC = -\log_{10} CMC \text{ (mol/L)}$ ] to guarantee the linear distribution [27]. The whole dataset has been split into two classes. The first class, consisting of 119 sugar surfactants (Table 1) collected from previously mentioned literature [1,4,28–30], was used to create the QSPR model (75% of the data comprise the training set, whereas 25% include the validation and test set). The second class, which included six sugar-based surfactants that were not employed in creating the QSPR model, was used to predict  $pCMC$  using the most efficient model.

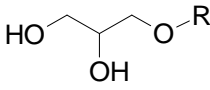
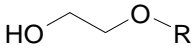
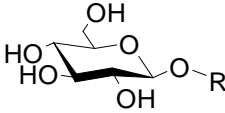
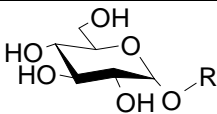
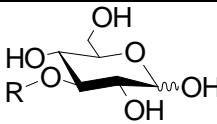
## 2.2 Molecular descriptor calculation

Molecular descriptors are numerical that represent the specificities of a compound's structure. For each surfactant, 1544 molecular descriptors were determined using PaDEL-descriptor software, which is available online. The molecular structure of compounds used in model development were drawn manually using Chem sketch software and the structures were saved as SMILES (Simplified Molecular Input Line-Entry System) notation, which is the approved input format for PaDEL-descriptor software[31].

## 2.3 Molecular descriptor selection a

An important step in the QSPR model is decreasing the number of descriptors. This decrease serves two objectives: it prevents overfitting and it reduces the possibility of finding a model by chance[27]. The stepwise-multiple linear regression (S-MLR) tool ([http://teqip.jdvu.ac.in/QSAR\\_Tools/](http://teqip.jdvu.ac.in/QSAR_Tools/)) was used to find the least number of meaningful descriptors that could predict the output property. The dataset of descriptors acquired following S-MLR selection using the F-value consisted of 6 descriptors.

Table 1. List of 119 sugar-based surfactants and their experimental pCMC values along with predicted pCMC values.

N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
1		Octyl	2.237	2.394
2		Octyl	2.310	2.319
3		Octyl	1.669	2.125
4		1-butylhexyl	1.824	1.776
5		1-propylheptyl	1.921	1.806
6		1-ethyloctyl	2.071	1.877
7		1-methylnonyl	2.347	2.276
8		3,7-dimethyloctyl	2.398	1.995
9		Decyl	2.699	2.928
10		Nonyl	2.161	2.535
11		Heptyl	1.141	1.601
12		Hexyl	0.645	1.055
13		Dodecyl	3.721	4.019
14		Dodecyl	4.377	4.019
15		Octyl	5	2.125
16		Octyl	2.854	2.542
17		Octanoyl	3.215	2.402
18		Octyl	1.638	1.713

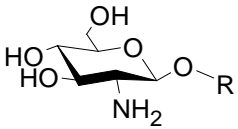
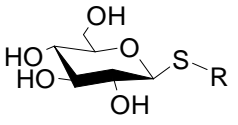
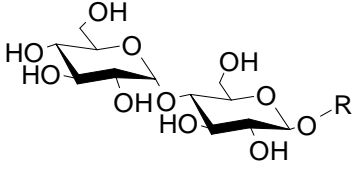
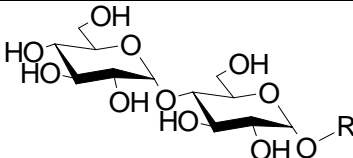
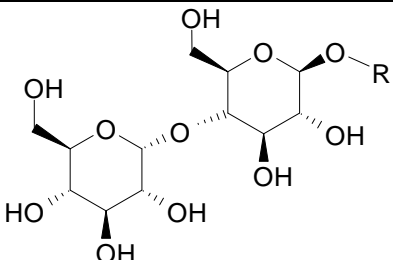
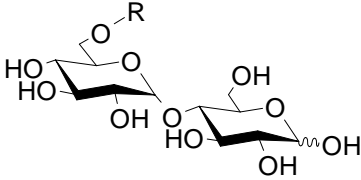
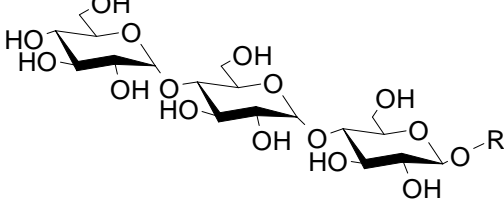
19		Nonyl	2.155	2.258
20		Hexyl	1.086	0.896
21		Heptyl	1.565	1.135
22		Octyl	1.983	1.975
23		Nonyl	2.553	2.643
24		Decyl	3.046	3.253
25		3,7-dimethyloctyl	2.276	2.070
26		Decyl	2.699	2.554
27		Dodecyl	3.769	3.770
28		Octyl	1.635	1.557
29		Nonyl	2.187	2.044
30		Tetradecyl	4.777	5.162
31		Hexadecyl	5.839	6.218
32		3,7,11 trimethyldodecyl	4.526	4.467
33		Dodecyl	3.921	3.770

Table 1. (Continued)

N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
34		Hexadecyl	6.222	6.219
35		Dodecanoyl	3.481	3.327
36		3,7-dimethyloctyl	2.301	2.327
37		Oleyl	4.377	4.503
38		Octyl	2.086	2.307

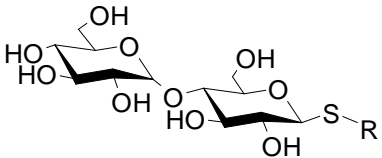
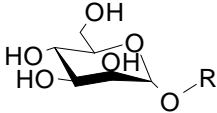
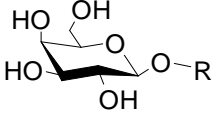
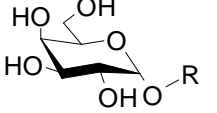
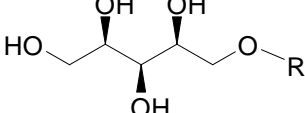
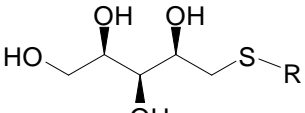
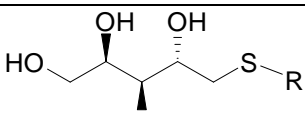
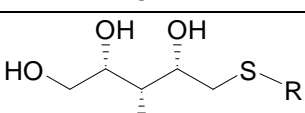
39		Decyl	3.222	3.294
40		Octyl	2.222	2.125
41		Decyl	3.602	2.928
42		Dodecyl	4.301	4.019
43		Heptyl	1.514	1.601
44		Nonyl	2.398	2.535
45		Octyl	1.796	2.125
46		Decyl	3.155	2.928
47		Dodecyl	3.699	4.019
48		Heptyl	1.757	1.601
49		Heptyl	1.488	1.601
50		Octyl	1.724	2.155
51		Nonyl	2.284	2.535
52		Butyl	1.237	1.175
53		Pentyl	1.420	1.310
54		Hexyl	1.023	1.639
55		Heptyl	2.036	2.146
56		Octyl	2.174	2.532
57		Nonyl	2.678	2.773
58		Pentanoyl	0.921	1.141
59		Hexanoyl	1.237	1.272
60		Heptanoyl	2	1.415
61		Nonanoyl	2.357	2.372
62		decanoyl	2.745	2.957

Table 1. (Continued)

N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
63		Butyl	0.745	1.249
64		Pentyl	1.337	1.442
65		Hexyl	1.745	1.908
66		Octyl	2.745	2.980
67		Hexyl	1.921	1.908
68		Octyl	2.921	2.980
69		Decyl	3.398	3.732
70		Octyl	3.319	3.569

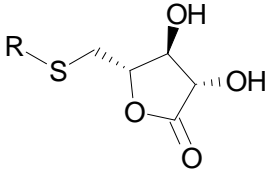
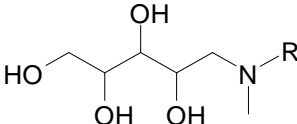
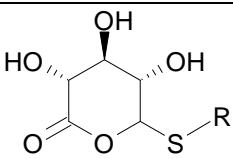
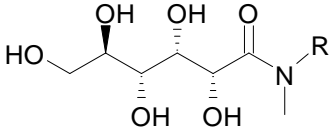
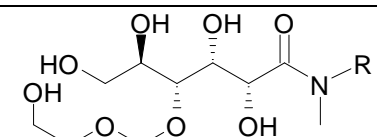
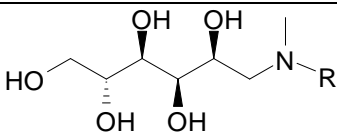
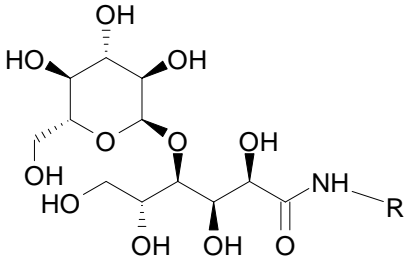
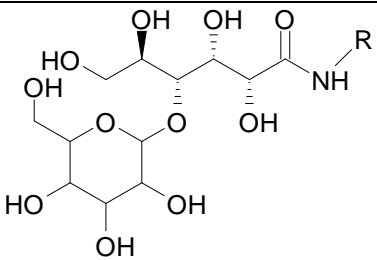
71		Decyl	4.481	4.505
72		Dodecanoyl	3.444	3.104
73		Hexyl	2.301	2.629
74		Octyl	3.276	3.167
75		Decyl	4.638	4.387
76		Decyl	2.886	3.039
77		Dodecyl	3.854	3.913
78		Tetradecyl	4.619	5.358
79		Olelyl	4.495	5.419
80		Decyl	2.638	2.836
81		Tetradecyl	4.444	4.452
82		Hexadecyl	5.032	5.133
83		Octadecyl	5.481	5.571
84		Olelyl	4.268	4.241
85		Octanoyl	1.161	1.359
86		Nonanoyl	1.678	1.801
87		Decanoyl	2.174	2.336
88		Hexyl	1.081	1.179
89		Dodecyl	3.509	3.812

Table 1. (Continued)

N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
90		Decyl	2.886	2.869

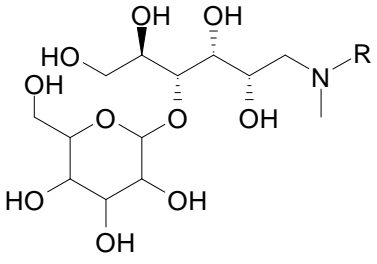
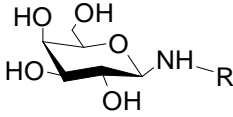
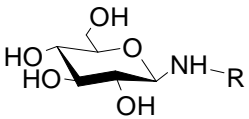
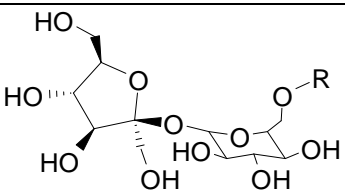
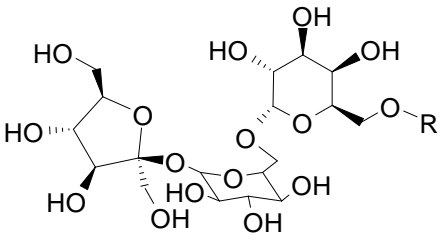
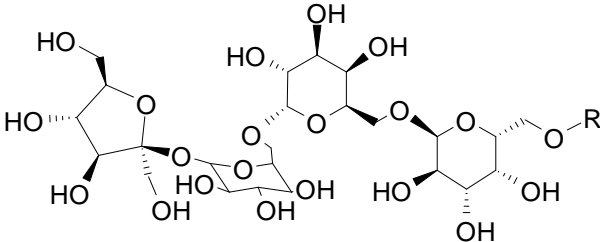
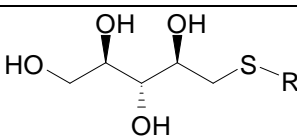
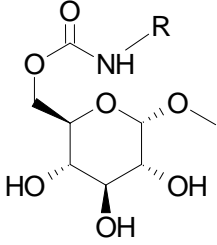
91		Decanoyl	2.481	2.622
92		Dodecanoyl	3.347	3.219
93		Tetradecanoyl	4.167	3.969
94		Octanoyl	1.347	1.250
95		Octanoyl	1.155	1.250
96		Dodecanoyl	3.337	3.389
97		Dodecanoyl	3.022	2.997
98		Dodecanoyl	2.638	2.647
99		Hexyl	1.991	1.908
100		Octyl	3.420	2.980

Table 1. (Continued)

N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
101		heptyl	1.710	1.851

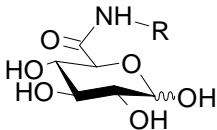
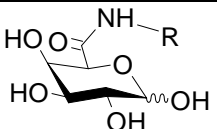
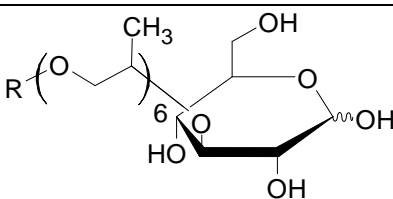
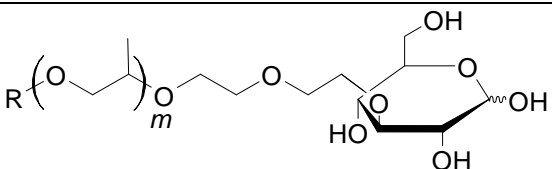
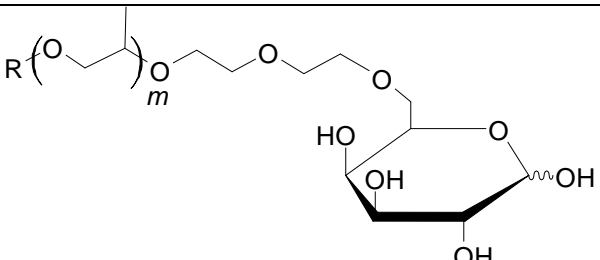
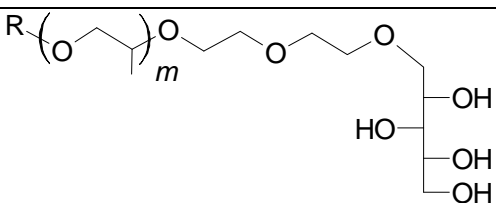
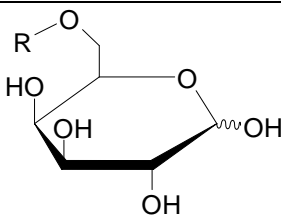
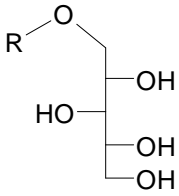
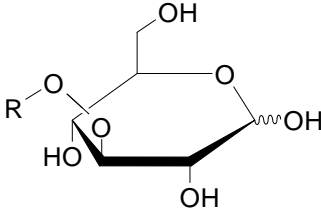
102		Octyl	2.481	2.685
103		Octyl	2.398	2.685
104		Dodecyl	4.398	4.230
		Dodecyl		
105		m=6	3.959	4.032
106		m=10	4.398	4.610
107		m=14	4.699	4.862
		Dodecyl		
108		m=6	4.301	4.303
109		m=10	4.523	4.733
		Dodecyl		
110		m=6	4.699	4.573
111		m=10	4.699	4.917
112		Octyl	2.469	2.583
113		Decyl	3.432	3.244

Table 1. (Continued)

N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
114		Octyl	2.174	2.532



115		Decyl	3.092	3.031
116		Octyl	2.921	2.901
117		Decyl	3.108	3.036
118		Dodecyl	3.638	3.424
119		Tetradecyl	3.745	3.866

\*Experimental pCMC values are collected from the following literature: [1, 4, 28–30]\*\* pCMC values predicted by ANN

#### 2.4 QSPR model's development and validation

Multiple linear regression (MLR), Partial least square (PLS), Artificial neural networks (ANN), and Adaptive neuro-fuzzy inference system (ANFIS) were used to create the models. For the ANFIS and ANN approach, we have employed MATLAB® R 2018a. while MLR PlusValidation 1.3 tools and Partial Least Squares Y1.0 were used for MLR and PLS methods, respectively ([http://teqip.jdvu.ac.in/QSAR\\_Tools/](http://teqip.jdvu.ac.in/QSAR_Tools/)). Validation is an important part of QSPR modeling, and the predictive models generated with varied parameters were subjected to both internal and external validation criteria. The statistical parameters [32–36] described in Equations (1) – (17) were employed, and the terms in these equations are specified appropriately:

$$Q_{Loo}^2 = 1 - \frac{\sum_{i=1}^{n_{INT}} (y_{i(training)}^{obs} - y_{i(training)}^{pred})^2}{\sum_{i=1}^n (y_{i(training)}^{obs} - \bar{y}_{(training)}^{obs})^2} \quad (1)$$

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_{i(test)}^{obs} - y_{i(test)}^{pred})^2}{\sum_{i=1}^n (y_{i(test)}^{obs} - \bar{y}_{(training)}^{obs})^2} \quad (2)$$

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_{i(test)}^{obs} - y_{i(test)}^{pred})^2}{\sum_{i=1}^n (y_{i(test)}^{obs} - \bar{y}_{(test)}^{obs})^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})^2} \quad (4)$$

$$k = \frac{\sum_{i=1}^n (y_i^{obs} \cdot y_i^{pred})}{\sum_{i=1}^n (y_i^{pred})^2} \quad (5)$$

$$k' = \frac{\sum_{i=1}^n (y_i^{obs} \cdot y_i^{pred})}{\sum_{i=1}^n (y_i^{obs})^2} \quad (6)$$

$$r_0^2 = 1 - \frac{\sum_{i=1}^n (y_i^{obs} - ky_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})^2} \quad (7)$$

$$r_0'^2 = 1 - \frac{\sum_{i=1}^n (y_i^{pred} - k' y_i^{obs})^2}{\sum_{i=1}^n (y_i^{pred} - \bar{y}^{pred})^2} \quad (8)$$

$$r_m^2 = r^2 (1 - \sqrt{r^2 - r_0'^2}) \quad (9)$$

$$r_m'^2 = r^2 (1 - \sqrt{r^2 - r_0'^2}) \quad (10)$$

$$\bar{r}_m^2 = \frac{(r_m^2 + r_m'^2)}{2} \quad (11)$$

$$\Delta r_m^2 = |r_m^2 - r_m'^2| \quad (12)$$

$$\bar{y}^{obs} = \frac{\sum_{i=1}^n y_i^{obs}}{n} \quad (13)$$

$$\bar{y}^{pred} = \frac{\sum_{i=1}^n y_i^{pred}}{n} \quad (14)$$

$$CCC = \frac{2 \sum (y_{i(test)}^{obs} - \bar{y}_{(test)}^{obs}) (y_{i(test)}^{pred} - \bar{y}_{(test)}^{pred})}{\sum (y_{i(test)}^{obs} - \bar{y}_{(test)}^{obs})^2 + \sum (y_{i(test)}^{pred} - \bar{y}_{(test)}^{pred})^2 + n (\bar{y}_{(test)}^{obs} - \bar{y}_{(test)}^{pred})^2} \quad (15)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{n}} \quad (16)$$

$$MAE = \frac{\sum_{i=1}^n |y_i^{obs} - y_i^{pred}|}{n} \quad (17)$$

Where,  $Q^2$ : cross-validation correlation coefficient,  $R^2$ : coefficient of determination,  $k$  &  $k'$ : slopes of the corresponding regression lines,  $r_0'^2$ : squared correlation coefficient between the observed and predicted value of compounds without intercept,  $r_0'^2$ : bears the same meaning as  $r_0'^2$ , but uses the reversed axes,  $\bar{r}_m^2$  &  $\Delta r_m^2$ : average and delta of  $r_m^2$ ,  $y_i^{obs}$  is the experimental value of Y,  $y_i^{pred}$  is the predicted Y-value of training set, test set or validation set,  $n$ : number of compounds in the data set,  $\bar{y}^{obs}$  &  $\bar{y}^{pred}$ : average of  $Y^{obs}$  and  $Y^{pred}$  respectively,  $CCC$ : concordance correlation coefficient,  $RMSE$ : Root Mean Squared Error,  $MAE$ : mean absolute error.

The different thresholds for these indicators are listed in the Table 2[37].

Table 2. Acceptance criterion (A.C) of a model [37].

Parameter	$Q_{ext}^2$	$R_{ext}^2$	k	k'	$\bar{r}_m^2$	$\Delta r_m^2$	CCC
A.C value	>0.7	>0.7	0.85 < k & k' < 1.15		>0.65	<0.2	>0.85

Roy and his collaborators[38], proposed two metrics for measuring the external predictability of QSPR models using, mean absolute error (MAE and MAE +3 $\sigma$ ) after omitting 5% high residual compounds. An ExternalValidationPlus ([http://teqip.jdvu.ac.in/QSAR\\_Tools/](http://teqip.jdvu.ac.in/QSAR_Tools/)) is an online tool that may be used to compute these MAE-based external validation requirements[39].

## 2.5 QSPR model's applicability domain (AD)

Application domain is described as the area in which a compound can be predicted with confidence[40], so it's a crucial factor to consider [41]. The leverage strategy (Williams plot) and the standardization approach were employed in this research.

### 2.5.1 Applicability domain using leverage approach

The Williams plot is a graphical representation of standardized residuals displayed against the leverage value of each compound[42].The applicability domain was specified as a square rectangle spanning  $\pm 3$  standard deviations and a warning leverage value ( $h^*$ ). The leverage values ( $h_{ii}$ ) correspond to the diagonal elements of the hat matrix (H) as defined[40] by:

$$H = X(X^t X)^{-1} X^t \quad (18)$$

Where  $X$  is the matrix built on the values of the model descriptor and compounds of the learning set and the warning leverage ( $h^*$ ) was determined[40,43] as:

$$h^* = \frac{3(p + 1)}{n} \quad (19)$$

Where  $p$  is the number of model descriptors and  $n$  is the total number of samples in the training set.

### 2.5.2 Standardization approach

This approach is an easy way to identify outliers (in the case of the training set) and the compounds that reside outside the AD (in the case of the test set). An open access standalone application has been created for estimating the AD for QSPR models. The software is available at ([http://teqip.jdvu.ac.in/QSAR\\_Tools/](http://teqip.jdvu.ac.in/QSAR_Tools/)). The basic theory, algorithm, and methodology, as well as the benefits of the suggested approach, are all available in the literature[44].

## 3. Results and discussion

### 3.1 Molecular descriptor selection

The six (6) descriptors obtained after selection by S-MLR techniques were: SssCH2, ATSC5e, AATSC4m, BIC3, MATS1c, and SpMin1\_Bhe. To conduct the research, a correlation matrix comprising six descriptors was created. The correlation coefficient for each pair of descriptors is less than 0.63, indicating that the descriptors chosen were independent.

### 3.2 MLR model

The MLR model, generated by using 6 variables (descriptors) is illustrated by a linear equation (Eq.20) and the statistical parameters represented in Table 3).

$$\begin{aligned} \text{pCMC} = & 12,09731 (\pm 5,88554) + 0,16059 (\pm 0,0293) \text{ SssCH2} + 0,17617 (\pm 0,06241) \text{ ATSC5e} + 0,25775 \\ & (\pm 0,06659) \text{ AATSC4m} - 7,58599 (\pm 2,16343) \text{ BIC3} - 6,20729 (\pm 2,10092) \text{ MATS1c} - 4,55738 \\ & (\pm 2,92592) \text{ SpMin1\_Bhe} \quad (20) \end{aligned}$$

The standard error of regression coefficients is denoted in parenthesis, and the statistical parameter values represent the MLR model's robustness and friability. In addition to indicating whether the model is not

overfitted, Tropsha and Golbraikh [45,46] propose that  $R^2 - Q_{loo}^2$  should be less than 0.3. As part of this study, the  $R^2 - Q_{loo}^2 = 0$ , demonstrating that the MLR model is not overfitting. Moreover, a PRESS/SSY ratio of less than 0.4 suggests a reasonable QSPR model [47]. The ratio is 0.13 ( $PRESS = 22.75$ ,  $SSY = 170.3$ ), indicating that the established model prediction is better than chance. The influence of a descriptor in a model is characterized by its mathematical equation for the sign-in model.

The regression coefficients of Eq (20) show that the BIC3, MATS1c, and SpMin1\_Bhe descriptors have negative signs, implying that they have a negative influence on the CMC of sugar-based surfactants. On the contrary, the regression coefficients of the descriptors SssCH2, ATSC5e, and AATSC4m have positive contributions and the greatest values contributing to the improvement of the CMC of sugar surfactants.

Table 3. Validation parameters for MLR model's train, validation, test, and global set.

Statistical values	Train	Validation	Test	Global
n	83	18	18	119
$R^2$	0,834	0,902	0,945	0,868
$Q_{Loo}^2$	0,834	--	--	--
$Q_{F1}^2$	--	0,887	0,934	--
$Q_{F2}^2$	--	0,834	0,930	--
K	1	0,903	1,042	0,997
k'	0,979	1,089	0,949	0,983
$r_0^2$	0,834	0,902	0,940	0,866
$r_0'^2$	0,806	0,888	0,928	0,839
$r_m^2$	0,834	0,882	0,879	0,8389
$r_m'^2$	0,693	0,795	0,821	0,719
$\bar{r}_m^2$	0,764	0,838	0,850	0,779
$\Delta r_m^2$	0,141	0,087	0,057	0,119
CCC	--	0,922	0,960	0,926
MAE	0,288	0,304	0,269	0,287
RMSE	0,453	0,409	0,386	0,437
MAE (95%data)	--	--	0.217	--
MAE +3 $\sigma$ (95%data)	--	--	0.786	--

### 3.3 PLS model

The PLS regression model generated with six descriptors displayed a significant correlation between the predicted and experimental values of pCMC ( $R^2 = 0.783$ ,  $RMSE = 0.519$ ). Table 4 shows the statistical parameters for the prediction set.

Table 4. Validation parameters for the PLS model's train, validation, test, and global set.

Statistical values	Train	Validation	Test	Global
n	83	18	18	119
$R^2$	0,783	0,832	0,973	0,832
$Q_{Loo}^2$	0,782	--	--	--
$Q_{F1}^2$	--	0,836	0,956	--
$Q_{F2}^2$	--	0,758	0,953	--

Statistical values	Train	Validation	Test	Global
K	0,999	0,900	1,054	0,998
k'	0,973	1,078	0,944	0,977
$r_0^2$	0,783	0,828	0,968	0,832
$r_0'^2$	0,737	0,821	0,963	0,797
$r_m^2$	0,773	0,783	0,909	0,820
$r_m'^2$	0,614	0,746	0,879	0,676
$\bar{r}_m^2$	0,694	0,764	0,894	0,749
$\Delta r_m^2$	0,159	0,037	0,030	0,144
CCC	--	0,889	0,974	0,907
MAE	0,362	0,379	0,237	0,345
RMSE	0,519	0,494	0,317	0,490
MAE (95%data)	--	--	0.203	--
MAE +3 $\sigma$ (95%data)	--	--	0.704	--

### 3.4 ANN model

In this investigation, the training function trainlm “Levenberg-Marquardt” was utilized for training the network. For the hidden layer and output layer, the tansig “hyperbolic tangent” and purelin “linear transfer” functions were used, respectively. Several computations were performed using different numbers of hidden nodes (1-9) to optimize the number of hidden neurons. The observed pCMC was represented by one output neuron. The 119 sugar-based surfactants were split into three groups: training set (70%), validation set (15 %) and test set (15 %). The model with a minimum *RMSE* was chosen [48]. The ANN model with {6-6-1} architecture was constructed as the final model. Table 1 provides the prediction pCMC from the ANN model for 119 sugar surfactants. The scatter plot of the observed vs predicted results for the training, validation, and test set is shown in Fig1. A close correlation was found between the predicted and observed values of pCMC, except for some outlier points, considered as accepted out of range.

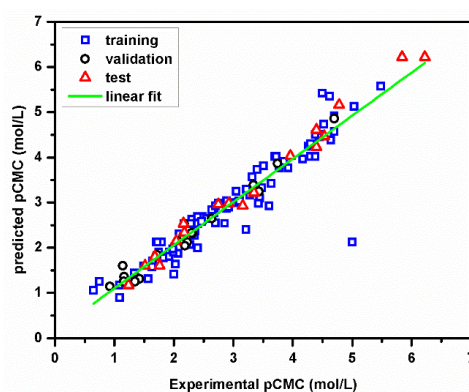


Figure 1. Scatter plot of the predicted values of pCMC versus the experimental values by ANN model for the training, validation, and test set.

The statistical parameters for validation results, given in Table 5, meet the required standards, indicating that the ANN model is robust and provides excellent predictive performance.

Table 5. Validation parameters for ANN model's train, validation, test, and global set.

Statistical value	Train	Validation	Test	Global
n	83	18	18	119
R <sup>2</sup>	0,861	0,972	0,985	0,908
Q <sub>Loo</sub> <sup>2</sup>	0,856	--	--	--
Q <sub>F1</sub> <sup>2</sup>	--	0,978	0,982	--
Q <sub>F2</sub> <sup>2</sup>	--	0,967	0,981	--
K	0,991	0,974	0,975	0,986
K'	0,990	1,022	1,023	0,999
r <sub>0</sub> <sup>2</sup>	0,857	0,971	0,985	0,906
r <sub>0</sub> ' <sup>2</sup>	0,856	0,969	0,985	0,906
r <sub>m</sub> <sup>2</sup>	0,804	0,939	0,976	0,873
r <sub>m</sub> ' <sup>2</sup>	0,802	0,913	0,982	0,864
r <sub>m</sub> <sup>2</sup>	0,803	0,926	0,979	0,868
Δr <sub>m</sub> <sup>2</sup>	0,001	0,026	0,006	0,009
CCC	--	0,983	0,991	0,953
MAE	0,246	0,144	0,165	0,218
RMSE	0,422	0,182	0,199	0,368
MAE (95%data)	--	--	0.152	--
MAE +3σ (95%data)	--	--	0.466	--

To see the relationship between the predicted property and the descriptor in the ANN model, a connection weights approach was used in this method, proposed by Olden[49]. From Fig 2 the order of relative contribution level of the descriptors was: ATSC5e > BIC3 > SpMin1\_Bhe > MATS1c > SssCH2 > AATSC4m. In this situation, the variables with the biggest relative contribution were ATSC5e, BIC3, and SpMin1\_Bhe (40.97%, 36.44%, and 27.34% respectively).

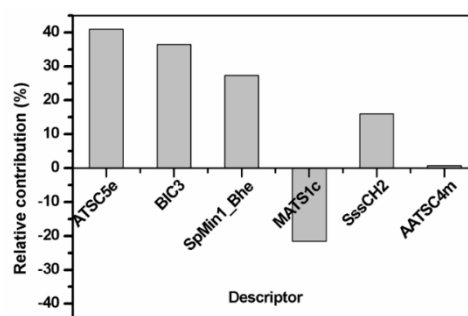


Figure 2. Plot of the fraction contribution of the descriptors to the pCMC of sugar-based surfactants.

ATSC5e is centred broto-moreau autocorrelation of lag 5 weighted by Sanderson electronegativity. BIC3 is a bond information content index defined as neighborhood symmetry of 3-order. SpMin1\_Bhe is burden eigenvalues n.1 of burden matrix weighted by Sanderson electronegativity. MATS1c is a 2D autocorrelation, which is known as the moran coefficient of lag 1 weighted by gasteiger charge. SssCH2

describes the electropological state. SssCH2 represents the sum of  $\text{-CH}_2$  group connected with two double single bonds. On the other hand, AATSC4m is defined as averaged and centered moreau-broto autocorrelation of lag 4 weighted by mass.

In summary, it can be concluded that electronic properties (electronegativity and charges), electrotopology, and symmetry of a molecule are of major importance in the micellization of sugar-based surfactants.

### 3.5 ANFIS model

Both artificial neural networks and neural-fuzzy systems are used in ANFIS architecture[50]. The  $R^2(0.916)$  value and  $\bar{r}_m^2(0.845)$  for the test set obtained within an acceptable range. Table 6 reports a value for various parameters used in the ANFIS model.

Table 6. Validation parameters for ANFIS model's train, test, and global set.

Statistical value	Train	Test	Global
n	101	18	119
$R^2$	0,935	0,916	0,932
$Q_{Loo}^2$	0,935	--	--
$Q_{F1}^2$	--	0,917	--
$Q_{F2}^2$	--	0,894	--
K	1,000	0,972	0,994
K'	0,990	1,018	0,996
$r_0^2$	0,935	0,902	0,931
$r_0'^2$	0,931	0,915	0,930
$r_m^2$	0,935	0,806	0,911
$r_m'^2$	0,877	0,884	0,893
$\bar{r}_m^2$	0,906	0,845	0,902
$\Delta r_m^2$	0,058	0,079	0,019
CCC	--	0,952	0,965

Table 6. (Continued)

Statistical value	Train	Test	Global
MAE	0,089	0,2701	0,116
RMSE	0,299	0,384	0,314
MAE (95%data)	--	0,227	--
MAE +3 $\sigma$ (95%data)	--	0,882	--

### 3.6 Comparison of four statistical models

The statistical quality of a model for an external (test) set is the most important factor for evaluating its predictive power [51]. We used the same kind and number of descriptors for each model. As evidenced by the validation metrics previously reported in (Tables 3, 4, 5, and 6), all models (MLR, PLS, ANN, ANFIS) are of acceptable quality. ANN model outperforms the other three regression models in terms of the external validation metrics namely  $R^2$ ,  $Q_{F1}^2$ ,  $Q_{F2}^2$  and CCC. A further analysis using MAE (MAE and MAE +3 $\sigma$ ) after removing 5% of the data compounds with large prediction residuals gave the quality of the ANN model as

good, which is also consistent with the judgment offered by classical external validation metrics. we used cross-validation with the LOO method to evaluate the robustness of the models. In the ANN model,  $Q^2$  was 0.86 and in MLR, PLS and ANFIS were 0.83, 0.78, and 0.93 respectively.

### 3.7 Applicability domain

The Williams plot and standardization technique were used to examine the applicability domain of the ANN model. In Figure 3 we can see that no compound in the entire dataset had leverage greater than the warning  $h^*$  value of 0.25. the standardization approach recognizes the training compound 15 within the AD but the leverage approach identifies it outside the two horizontal lines. However, the training compounds 109,110, and 111 are identified as outliers, validation compounds 58 and 107 and test compound 106 are identified outside by standardization technique but recognized inside the AD by leverage approach. Importantly, more than 99.15%, and 95% by Williams plot and standardization approach respectively of the domain covered, indicating that the ANN model compiles with the third principle of the OECD. As a result, the ANN model provided a good prediction for these compounds. It can be used to estimate the CMC of sugar-based surfactants, especially for untested substances and novel compounds.

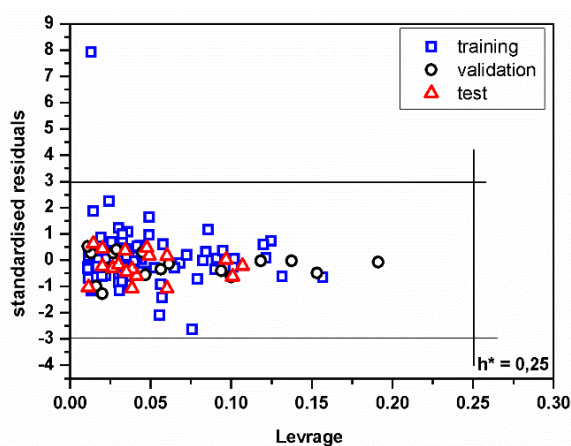


Figure 3. Projection of the training, validation, and test set of sugar-based surfactants in the Williams plot.

### 3.8 Comparison with previously reported models

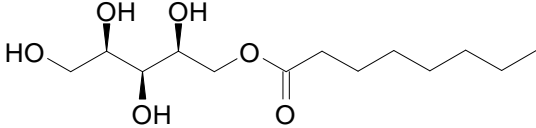
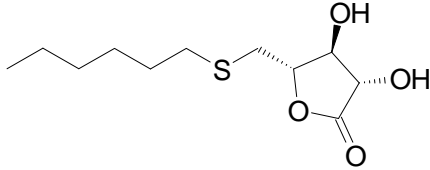
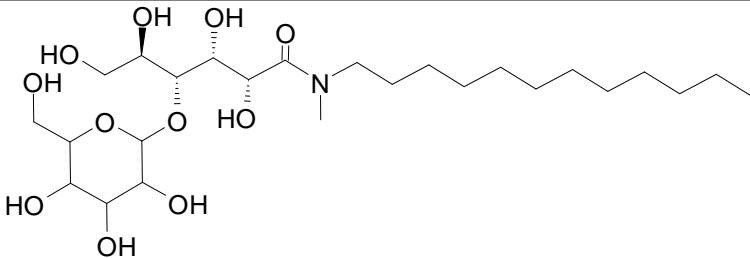
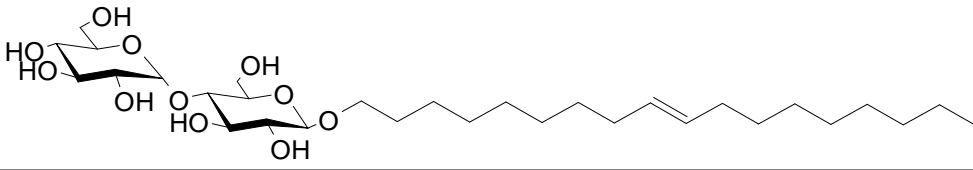
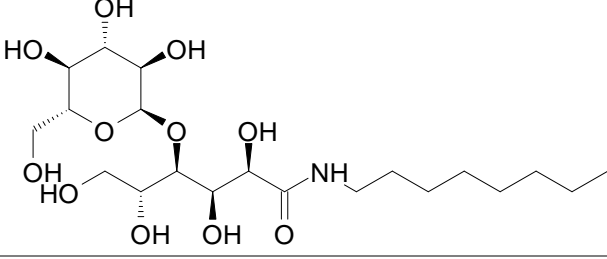
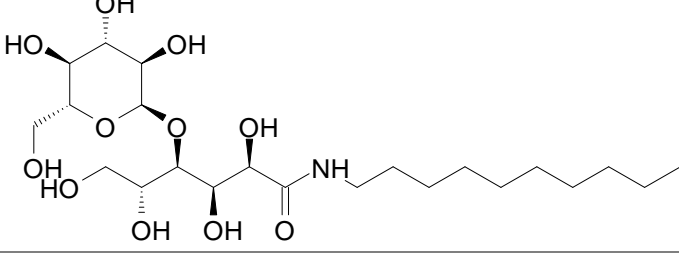
To predict the CMC of sugar-based surfactants, we compared the statistical results of our ANN model with a limited number of QSPR models available in the literature (Table 7). The comparison with the work of Gaudin et al.[1] shows that our model gives higher statistical quality and predictive performance in terms of external validation. Additionally, a limited number of statistical parameters are used from these QSPR models in the internal validation, unlike our model which has several statistical indices. In addition, we used our model (ANN-simulate network) to forecast the CMC of 6 sugar-based surfactants that had not been used for the construction of the QSPR models. The promising results are summarized in Table 8. These results demonstrate a high degree of consistency between the computed estimates and observed data, and the developed ANN model shows great potential for determining the CMC of new surfactants while reducing time and money.



Table7. Results of our best model's (ANN) internal and external validation compared with those from previous research.

MOD ELS	N <sub>total</sub>	Training				Validation				Test				
		$R^2$	$Q_{Loo}^2$	$\overline{r}_m^2$	$\Delta r_m^2$	$R^2$	$Q_{F1}^2$	$\overline{r}_m^2$	$\Delta r_m^2$	$R^2$	$Q_{F1}^2$	$\Delta r_m^2$	MAE (95% data)	MAE +3 $\sigma$ (95% data)
Present work (ANN)	119	0.861	0.856	0.803	0.001	0.972	0.978	0.926	0.026	0.985	0.982	0.006	0.152	0.466
Gaudin et al.	83	0.93	0.91	--	--	--	--	--	--	0.91	0.90	0.1	--	--
Wang et al.	83	0.959	0.947	--	--	--	--	--	--	0.946	--	--	0.197	0.674
Baghdan et al.	83	0.999	--	--	--	--	--	--	--	0.997	--	--	--	--

Table 8. Observed value of pCMC and those predicted by ANN-simulate for 6 sugar surfactants.

Sugar surfactants	Observed pCMC (mol/L)	Predicted pCMC (mol/L)
	1.745	1.749
	2.174	2.879
	3.602	3.591
	4.699	4.746
	2.244	2.077
	2.886	2.868

#### 4. Conclusion

In this work, we propose four regression models to estimate the CMC value of 119 sugar surfactants by utilizing six relevant molecular descriptors. The metrics of the four developed models fell within the acceptable range. The ANN model was trained using the training function trainlm “Levenberg-Marquard” gave a better performance in CMC prediction with  $Q_{ext}^2$  and  $\bar{r}_m^2$  values (0.98 and 0.97) were higher, and

$\Delta r_m^2$  value (0.006) was acceptable for a testing dataset in comparison to models previously reported. Based on MAE metrics, the ANN model makes accurate predictions, as indicated by the removal of 5% of the test set objects with high residual values. In addition, estimating the CMC of 6 other sugar surfactants based on simulate of the network gave very good results ( $R = 0.96$ ). The In-silico models used in this study demonstrated the significance of electronic properties (electronegativity and charges), electro-topology, and symmetry of a molecule in contributing to the micellization process. In conclusion, all validations showed that the built QSPR model was reliable, acceptable, consistent with the OECD principle, and able to accurately predict the CMC for the surfactant that not contained in the data set.

## Reference

1. Gaudin. T., Rotureau. P., Pezron. I and Fayet. G. (2016). New QSPR models to predict the critical micelle concentration of sugar-based surfactants. *Ind. Eng. Chem. Res*, vol. 55 pp. 11716–11726.
2. Rosen. M.J. (2004). *Surfactants and interfacial phenomena*. 3rd edition.
3. Wang. Y., Yan. F., Jia. Q and Wang. Q. (2018). Quantitative structure-property relationship for critical micelles concentration of sugar-based surfactants using norm indexes. *J. Mol. Liq*, vol. 253, pp. 205–210.
4. Iglaier. S., Wu. Y., Shuler. P., Tang. Y and Goddard. W.A. (2010). Analysis of the influence of alkyl polyglycoside surfactant and cosolvent structure on interfacial tension in aqueous formulations versus n-octan. *Tenside, Surfactants, Deterg*, vol.47, pp. 87–97.
5. Shinoda. K., Yamanaka. T and Kinoshita. K. (1959). Surface chemical properties in aqueous solutions of non-ionic surfactants: Octyl glycol ether,  $\alpha$ -octyl glyceryl ether and octyl glucoside. *J. Phys. Chem*, vol. 63, pp. 648–650.
6. Liljekvist. P., Kjellin. M and Christer Eriksson. J. (2001). Surface pressure effect of pentaoxyethylene and maltoside surfactant head groups. *Adv. Colloid Interface Sci*, vol. 89–90, pp. 293–302.
7. Bazin. H.G., Polat. T and Linhardt. R.J. (1998). Synthesis of sucrose-based surfactants through regioselective sulfonation of acylsucrose and the nucleophilic opening of a sucrose cyclic sulfate. *Carbohydr. Res*, vol. 309, pp. 189–205.
8. Kjellin. M. and Johansson. I. (2010). *Surfactants from Renewable Resources, Surfactants from Renew. Resour.*
9. Matsumura. S., Imai. K., Yoshikawa. S., Kawada. K and Uchibor. T. (1990). Surface activities, biodegradability and antimicrobial properties of n-alkyl glucosides, mannosides and galactosides. *J. Am. Oil Chem. Soc*, vol. 67, pp. 996–1001.
10. Rojas. O.J., Stubenrauch. C., Lucia. L.A and Habibi. Y. (2009). *Interfacial Properties of Sugar-based Surfactants*, pp. 457–480.
11. Huibers. P.D.T., Lobanov. V.S., Katritzky. A.R., Dinesh. S., Shah. O and Karelson. M. (1996). Prediction of critical micelle concentration using a quantitative structure-property relationship approach. 1. Nonionic Surfactants, vol. 12, pp. 1462–1470.
12. Israelachvili. J.N., Mitchell. D.J and Ninham. B.W. (1976). Theory of self-assembly of hydrocarbon amphiphiles into micelles and bilayers. *J. Chem. Soc. Faraday Trans. 2 Mol. Chem. Phys*, vol. 72, pp. 1525–1568.
13. Baghban. A., Sasanipour. J., Sarafbidabad. M., Piri. A. and Razavi. R. (2018). On the prediction of

- critical micelle concentration for sugar-based non-ionic surfactants. *Chem. Phys. Lipids*, vol. 214, pp. 46–57.
14. Hu. J., Zhang. X and Wang. Z. (2010). A review on progress in QSPR studies for surfactants. *Int. J. Mol. Sci*, vol. 11, pp. 1020–1047.
15. Myers, D. (2006). *Surfactant science and technology*, 3rd edition.
16. Lecomte Du Noüy. P. An intrfacial tensiometer for universal use. (1925). *J. Gen. Physiol.* vol. 7, pp. 625.
17. Fayet. G and Rotureau. P. (2016). How to use QSPR-type approaches to predict properties in the context of green chemistry. *Biofuels, Bioprod. Biorefining*, vol. 10, pp. 738–752.
18. Creton. B., Nieto-Draghi. C and Pannacci. N. (2012). Prediction of surfactants' properties using multiscale molecular modeling tools: A review. *Oil Gas Sci. Technol. – Rev. d'IFP Energies Nouv.*, vol. 67, pp. 969–982.
19. Nieto-Draghi. C., Fayet. G., Creton. B., Rozanska. X. Rotureau. P., De Hemptinne. J.C., Ungerer. P., Rousseau. B and Adamo. C. (2015). A general guidebook for the theoretical prediction of physicochemical properties of chemicals for regulatory purposes. *Chem. Rev*, vol. 115, pp. 13093–13164.
20. Katritzky. A.R., Pacureanu. L.M., Slavov. S.H., Dobchev. D.A and Karelson. M. (2008). QSPR study of critical micelle concentrations of nonionic surfactants. *Ind. Eng. Chem. Res.*, vol. 47, pp. 9687–9695.
21. Saunders. R.A and Platts. J.A. (2004). Correlation and prediction of critical micelle concentration using polar surface area and LFER methods. *J. Phys. Org. Chem*, vol. 17, pp. 431–438.
22. Mozrzymas. A and Rozycka-Roszak. B. (2009). Prediction of critical micelle concentration of nonionic surfactants by a quantitative structure property relationship. *Comb. Chem. High Throughput Screen.*, vol. 13, pp. 39–44.
23. Roy. K and Kabir. H. (2012). QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants. *Chem. Eng. Sci.*, vol. 73, pp. 86–98.
24. Mattei. M., Kontogeorgis. G.M and Gani. R. (2013). Modeling of the critical micelle concentration (CMC) of nonionic surfactants with an extended group-contribution method. *Ind. Eng. Chem. Res.*, vol. 52, pp. 12236–12246.
25. Anoune. N., Nouiri. M., Berrah. Y., Gauthier. J.Y and Lanteri. P. (2002). Critical micelle concentrations of different classes of surfactants: A quantitative structure property relationship study. *J. Surfactants Deterg.*, vol. 5, pp. 45–53.
26. Hamadache. M. Benkortbi. O, Hanini. S and Amrane. A. (2018). 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.*, vol. 25, pp. 896–907.
27. Rahal. S., Hadidi. N and Hamadache. M. (2020). In silico prediction of critical micelle concentration (CMC) of classic and extended anionic surfactants from their molecular structural descriptors. *Arab. J. Sci. Eng.*, vol. 45, pp. 7445–7454.
28. Lemahieu. G., Aguilhon. J., Strub. H., Molinier. V., Ontiveros. J.F and Aubry. J.M. (2020). Hexahydrofarnesyl as an original bio-sourced alkyl chain for the preparation of glycosides surfactants with enhanced physicochemical properties. *RSC Adv*, vol. 10, pp. 16377–16389.
29. Gaudin. T., Lu. H., Fayet. G., Berthault-Drelich. A., Rotureau. P., Pourceau. G., Wadouachi. A.,

- Van Hecke. E., Nesterenko. A and Pezron. I.(2019). Impact of the chemical structure on amphiphilic properties of sugar-based surfactants: A literature overview. *Adv. Colloid Interface Sci*, vol, 270, pp. 87–100.
30. Scorzza. C., Godé. P., Goethals. G., Martin. P., Miñana-Pérez. M., Salager. J.L., Usabillaga. A and Villa. P.(2002). Another new family of “extended” glucidoamphiphiles. Synthesis and surfactant properties for different sugar head groups and spacer arm lengths.*J. Surfactants Deterg*, vol. 5, pp. 337–343.
31. Yap. C.W.(2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints.*J. Comput. Chem*, vol. 32, pp.1466–1474.
32. Roy. K.(2007). On some aspects of validation of predictive quantitative structure-activity relationship models. *Expert Opin. Drug Discov*, vol. 2, pp. 1567–1577.
33. Chirico. N and Gramatica. P.(2011). Real external predictivity of QSAR models: How to evaluate It? Comparison of different validation criteria and proposal of using the concordance correlation coefficient.*J. Chem. Inf. Model*, vol. 51, pp. 2320–2335.
34. Gramatica. P and Sangion. A.(2016). A historical excursus on the statistical validation parameters for QSAR models: A clarification concerning metrics and terminology.*J. Chem. Inf. Model*, vol. 56, pp. 1127–1131.
35. Ojha. P.K., Mitra. I., Das. R.N and Roy. K. (2011).Further exploring rm2 metrics for validation of QSPR models. *Chemom. Intell. Lab. Syst*, vol. 107, pp. 194–205.
36. Ahmadi. R. Sepehri. B and Ghavami. R.(2019). Development linear and non-linear QSAR models for predicting AXL kinase inhibitory activity of N-[4-(quinolin-4-yloxy)phenyl]benzenesulfonamides.*J. Recept. Signal Transduct*,vol. 39, pp. 264–275.
37. Chirico. N and Gramatica. P. (2012). Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection.*J. Chem. Inf. Model*, vol. 52, pp. 2044–2058.
38. Roy. K., Das. R.N., Ambure. P and Aher. R.B.(2016). Be aware of error measures. Further studies on validation of predictive QSAR models, *Chemom. Intell. Lab. Syst*, vol. pp. 152 18–33.
39. 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/](http://teqip.jdvu.ac.in/QSAR_Tools/).
40. N’dri. J.S., Ouattara. B., Koné. M.G.-R., Kablan. A.L.C., Dembélé. G.S., Kodjo. C.G and Ziao. N. (2021). Quantitative structure-activity atudy against plasmodium falciparum of a series of derivatives of azetidine-2-carbonitriles by the method of density functional theory. *Mediterr. J. Chem*, vol. 11, pp. 162.
41. OECD. (2009). Principles for the validation, for regulatory purposes, of quantitative structure–activity relationship models.
42. Rybinska. A., Sosnowska. A., Barycki. M and Puzyn. T. (2016).Geometry optimization method versus predictive ability in QSPR modeling for ionic liquids.*J. Comput. Aided. Mol. Des*, vol. 30, pp. 165–176.
43. Asadollahi. T., Dadfarnia. S., Shabani. A.M.H., Ghasemi. J.B and Sarkhosh. M.(2011). QSAR models for cxcr2 receptor antagonists based on the genetic algorithm for data preprocessing prior to application of the pls linear regression method and design of the new compounds using in silico virtual

screening. *Molecules*, vol.16, pp. 1928–1955.

44. Roy. K., Kar. S and Ambure. P.(2015). On a simple approach for determining applicability domain of QSAR models. *Chemom. Intell. Lab. Syst*, vol. 145, pp. 22–29.
45. Tropsha. A., Gramatica. P and Gombar. V.K.(2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci*, vol 22, pp. 69–77.
46. Golbraikh. A., Shen. M., Xiao. Z., De Xiao. Y., Lee. K.H and Tropsha. A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Mol. Des*, vol. 17, pp. 241–253.
47. Sawant. S.D., Nerkar. A.G., Pawar. N.D and Velapure. A.V. (2014). Design, synthesis, QSAR studies and biological evaluation of novel triazolopiperazine based B-amino amides as dipeptidyl peptidase-IV (DPP-IV) inhibitors: part-II. *Int. J. Pharmacy. Pharmaceutical. Sci*, vol.6, pp. 812-817.
48. Hamadache. M., Benkortbi. O., Hanini. S., Amrane. A., Khaouane. L and Si Moussa. C. (2016). A Quantitative Structure Activity Relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction. *J. Hazard. Mater*, vol. 303, pp. 28–40.
49. Olden. J. (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol. Modell*, vol. 178, pp.389-397.
50. Buyukbingol. E., Sisman. A., Akyildiz. M., Alparslan. F.N., Adejare. A.(2007). Adaptive neuro-fuzzy inference system (ANFIS): A new approach to predictive modeling in QSAR applications: A study of neuro-fuzzy modeling of PCP-based NMDA receptor antagonists, *Bioorganic Med. Chem*, vol. 15, pp. 4265–4282.
51. Lavado. G.J., Baderna. D., Carnesecchi. E., Toropova. A.P., Toropov. A.A., Dorne. J.L.C.M and Benfenati. E. (2022). QSAR models for soil ecotoxicity: Development and validation of models to predict reproductive toxicity of organic chemicals in the collembola *Folsomia candida*. *J. Hazard. Mater*, vol. 423, pp. 127236.