

MAAMAR LAIDI
ABDALLAH ABDALLAH
EL HADJ
CHERIF SI-MOUSSA
OTHMANE BENKORTEBI
MOHAMED HENTABLI
SALAH HANINI

Laboratory of Biomaterials and
Transport Phenomena (LBMPT),
University of Medea, Medea,
Algeria

SCIENTIFIC PAPER

UDC 62:5

CMC OF DIVERSE GEMINI SURFACTANTS MODELING USING A HYBRID APPROACH COMBINING SVR-DA

Article Highlights

- QSPR descriptors were used for modeling of pCMC values of 211 diverse Gemini surfactants
- A novel hybrid approach combining DFO optimization algorithm and SVR was proposed
- A comparison was made between SVR-DA and four other techniques: ANN, PLS, OLS and KNN
- The statistical quality of the SVR-DA model is better than the other models
- SVR-DA can be used for prediction of pCMC values of Gemini surfactants

Abstract

Quantitative structure-property relationship (QSPR) technique provides a suitable tool to predict the critical micelle concentration (CMC) of Gemini surfactants from their structure descriptors. In this study, a comparative work was conducted to model the CMC property of 211 diverse Gemini surfactants based on their structural characteristics using linear and non-linear quantitative structure-property relationship models. Least squares model (OLS) and partial least squares (PLS) against k-nearest neighbours regression model (KNN), artificial neural network (ANN) and support vector regression (SVR) have been developed to model the CMC. Molecular descriptors were calculated and screened to remove unsuitable descriptors and improve the learning. Results indicate that the improved performance of support vector regression when the hyper-parameters are optimized using Dragonfly algorithm (SVR-DA) was highly capable of predicting the pCMC (-log CMC) values with an average absolute relative deviation (AARD) of 0.666 and coefficient of determination (R^2) of 0.9971 for the global dataset.

Keywords: quantitative structure-property relationship, surfactants, critical micelle concentration, modelling, machine learning.

Since surfactants are considered as important interfacial active compounds, they are usually employed in different industrial fields, namely petroleum, detergent, chemical and environmental sectors [1]. Different types of surfactants have been proposed in literature by trying various chemical structures to enhance their properties; among these surfactants, Gemini surfactants (dimeric surfactants) have gained significant attention due to their attractive properties

versus the conventional surfactants. Gemini surfactants contain two hydrophobic tails and two hydrophilic heads connected by a spacer at or near the two heads (Figure 1) [2]. These chemical structure modifications grant them excellent properties, namely low CMC, high aqueous solubility and viscosity, unique micelle structure and aggregation behaviour and high yield in decreasing the surface tension of water solutions [3-12]. As the CMC is an important parameter to characterize surfactants, their measurement is usually difficult and costly. For that reason, many studies have been conducted to find an easy method to determine the CMC values of Gemini surfactants based on chemometrics techniques [13-21].

Recently, quantitative structure-property relationship (QSPR) method was used in many fields to

Correspondence: M. Laidi, Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Medea, Medea, Algeria.
E-mail: maamarw@yahoo.fr
Paper received: 7 September, 2020
Paper revised: 23 October, 2020
Paper accepted: 3 December, 2020

<https://doi.org/10.2298/CICEQ200907048L>

extract and link chemicals properties to their molecular structures [22-30]. In QSPR modeling, different computational techniques have been used, such as multiple linear regression (OLS or MLR), PLS, ANN, SVR and ANFIS [31-43].

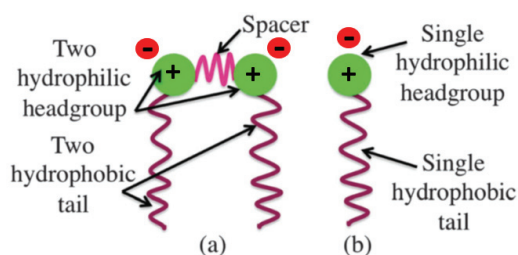


Figure 1. Surfactant molecular structure: a) Gemini; b) monomeric.

To the best to our knowledge, a few research papers have investigated the modeling techniques of Gemini surfactants based on relevant descriptors [3,13,14-20,28,44]. The novelty of this work is the application of SVR-DA algorithm to model the CMC of diverse Gemini surfactants which has not been

evaluated for CMC estimation yet and compare its performance with ANN, OLS, PLS and KNN model.

QSPR CALCULATION AND FILTERING METHODOLOGY

The methodology adopted in this work consisted of four main steps that are summarized below.

1. Dataset gathered from previously experimental published papers in literature [17,44,52-59,44-51] is presented in Table 1. The experimental CMC values of 211 diverse dibromides/di-chlorides Gemini surfactants were obtained by conductometer at as close as possible to room temperature (*i.e.*, between 20 and 25 °C). This dataset of 211 different Gemini surfactant molecules were drawn using HyperChem package (www.hyper.com, Figure 2), and then pre-optimized employing the molecular mechanics force field (MM+) presented in the HyperChem software. The molecular geometries energy minima were then obtained using the semiempirical AM1 using the Polack-Rabiere algorithm until the root mean square gradient limit as stopping criterion at 0.01 kcal/Å.

Table 1. Structures and CMC values of 211 Gemini surfactants according to Figure 1 (1-64)

No.	Tail 1	Tail 2	Ion-1	Ion-2	Spacer	Exp. pCMC	No.	Tail 1	Tail 2	Ion-1 or Ion-2	Spacer	Exp. pCMC
1	C ₈ H ₁₇	C ₈ H ₁₇	PO ₄ ⁻	(CH ₃) ₂ N ⁺	C ₂ H ₄	3.0000	33	C ₁₆ H ₃₂	C ₁₆ H ₃₂	(CH ₃) ₂ N ⁺	(CH ₂) ₆	3.6990
2	C ₈ H ₁₇	C ₁₀ H ₂₁			C ₂ H ₄	3.9586	34	C ₁₆ H ₃₂	C ₁₆ H ₃₂		(CH ₂) ₃	3.9208
3	C ₁₀ H ₂₁	C ₈ H ₁₇			C ₂ H ₄	4.0000	35	C ₁₆ H ₃₂	C ₁₆ H ₃₂		(CH ₂) ₅	1.8601
4	C ₈ H ₁₇	C ₁₂ H ₂₅			C ₂ H ₄	4.8539	36	C ₁₆ H ₃₂	C ₁₆ H ₃₂		(CH ₂) ₈	1.5817
5	C ₁₂ H ₂₅	C ₈ H ₁₇			C ₂ H ₄	4.7447	37	C ₁₆ H ₃₂	C ₁₆ H ₃₂		(CH ₂) ₁₀	1.5800
6	C ₁₀ H ₂₁	C ₁₀ H ₂₁			C ₂ H ₄	4.8861	38	C ₁₆ H ₃₂	C ₁₆ H ₃₂		(CH ₂) ₁₂	3.0809
7	C ₈ H ₁₇	C ₁₄ H ₂₉			C ₂ H ₄	5.2147	39 ^f	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₄	3.2007
8	C ₁₄ H ₂₉	C ₈ H ₁₇			C ₂ H ₄	5.0969	40 ^f	C ₁₄ H ₂₉	C ₁₄ H ₂₉		(CH ₂) ₄	3.4318
9 ^f	C ₁₀ H ₂₁	C ₁₂ H ₂₅			C ₂ H ₄	4.8861	41	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₂	3.6990
10 ^f	C ₁₀ H ₂₁	C ₁₄ H ₂₉			C ₂ H ₄	5.2757	42	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₃	3.9208
11	C ₁₄ H ₂₉	C ₁₀ H ₂₁			C ₂ H ₄	5.2596	43	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₄	1.8601
12	C ₁₂ H ₂₅	C ₁₂ H ₂₅			C ₂ H ₄	5.0458	44	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₆	2.0362
13	C ₈ H ₁₇	C ₁₈ H ₃₆			C ₂ H ₄	5.3188	45	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₈	2.1249
14	C ₁₈ H ₃₆	C ₈ H ₁₇			C ₂ H ₄	5.2291	46	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₁₀	2.3768
15	C ₁₀ H ₂₁	C ₁₆ H ₃₂			C ₂ H ₄	5.3010	47	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₁₂	2.6576
16	C ₁₄ H ₂₉	C ₁₂ H ₂₅			C ₂ H ₄	5.1549	48	C ₆ H ₁₃	C ₆ H ₁₃		(CH ₂) ₅	0.7959
17	C ₁₂ H ₂₅	C ₁₄ H ₂₉			C ₂ H ₄	5.3665	49 ^f	C ₉ H ₁₉	C ₉ H ₁₉		(CH ₂) ₅	1.6021
18	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₂	3.0969	50 ^f	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₅	2.0506
19 ^f	C ₁₂ H ₂₅	C ₁₂ H ₂₅			(CH ₂) ₃	3.0269	51	C ₁₁ H ₂₃	C ₁₁ H ₂₃		(CH ₂) ₅	2.5229
20 ^f	C ₁₂ H ₂₅	C ₁₂ H ₂₅			(CH ₂) ₄	2.9318	52	C ₁₂ H ₂₅	C ₁₂ H ₂₅		(CH ₂) ₅	2.9586
21	C ₁₂ H ₂₅	C ₁₂ H ₂₅			(CH ₂) ₆	2.9872	53	C ₁₃ H ₂₇	C ₁₃ H ₂₇		(CH ₂) ₅	3.3872
22	C ₁₂ H ₂₅	C ₁₂ H ₂₅			(CH ₂) ₈	3.0809	54	C ₁₄ H ₂₉	C ₁₄ H ₂₉		(CH ₂) ₅	3.7447
23	C ₁₂ H ₂₅	C ₁₂ H ₂₅			(CH ₂) ₁₀	3.2007	55	C ₁₃ H ₂₇	C ₁₁ H ₂₃		(CH ₂) ₆	3.0555
24	C ₁₂ H ₂₅	C ₁₂ H ₂₅			(CH ₂) ₁₂	3.4318	56	C ₁₄ H ₂₉	C ₁₀ H ₂₁		(CH ₂) ₆	3.0757
25	C ₁₂ H ₂₅	C ₁₂ H ₂₅			(CH ₂) ₁₄	3.6990	57	C ₁₆ H ₃₂	C ₈ H ₁₇		(CH ₂) ₆	3.1249
26	C ₁₂ H ₂₅	C ₁₂ H ₂₅			(CH ₂) ₁₆	3.9208	58	C ₁₈ H ₃₇	C ₆ H ₁₃		(CH ₂) ₆	3.2291

No.	Tail 1	Tail 2	lon -1	lon-2	Spacer	Exp. pCMC	No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC
27	C ₈ H ₁₇	C ₈ H ₁₇			(CH ₂) ₃	1.8601	59 [†]	C ₈ H ₁₇	C ₈ H ₁₇		(C ₂ H ₄) ₂ N(CH ₃)	1.5086
28	C ₈ H ₁₇	C ₈ H ₁₇			(CH ₂) ₄	1.5817	60 [†]	C ₉ H ₁₉	C ₉ H ₁₉		(C ₂ H ₄) ₂ N(CH ₃)	1.7447
29 [†]	C ₈ H ₁₇	C ₈ H ₁₇			(CH ₂) ₅	1.5800	61	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(C ₂ H ₄) ₂ N(CH ₃)	2.0757
30 [†]	C ₈ H ₁₇	C ₈ H ₁₇			(CH ₂) ₆	3.0809	62	C ₁₁ H ₂₃	C ₁₁ H ₂₃		(C ₂ H ₄) ₂ N(CH ₃)	2.5229
31	C ₁₆ H ₃₂	C ₁₆ H ₃₂			(CH ₂) ₂	3.2007	63	C ₁₂ H ₂₅	C ₁₂ H ₂₅		(C ₂ H ₄) ₂ N(CH ₃)	2.9586
32	C ₁₆ H ₃₂	C ₁₆ H ₃₂			(CH ₂) ₄	3.4318	64	C ₁₃ H ₂₇	C ₁₃ H ₂₇		(C ₂ H ₄) ₂ N(CH ₃)	3.4815

Table 1. Structures and CMC values of 211 Gemini surfactants according to Figure 1 (65-100)





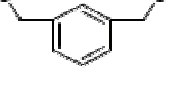

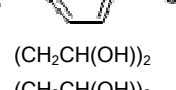














No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC	No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC
65	C ₁₄ H ₂₉	C ₁₄ H ₂₉	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ N(CH ₃)	3.9208	88	C ₈ H ₁₇	C ₈ H ₁₇		CH ₂ C≡CCH ₂	1.3206
66	C ₁₅ H ₃₁	C ₁₅ H ₃₁	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ N(CH ₃)	4.3665	89 [†]	C ₁₀ H ₂₁	C ₁₀ H ₂₁		CH ₂ C≡CCH ₂	2.2757
67	C ₁₆ H ₃₃	C ₁₆ H ₃₃	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ N(CH ₃)	4.9208	90 [†]	C ₁₀ H ₂₁	C ₁₀ H ₂₁		CH ₂ C≡CCH ₂	2.2248
68	C ₈ H ₁₇	C ₈ H ₁₇	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ O	1.3565	91	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ C≡CCH ₂	3.8861
69 [†]	C ₉ H ₁₉	C ₉ H ₁₉	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ O	1.6576	92	C ₈ H ₁₇	C ₈ H ₁₇	(CH ₃) ₂ N ⁺		1.6459
70 [†]	C ₁₀ H ₂₁	C ₁₀ H ₂₁	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ O	2.1135	93	C ₈ H ₁₇	C ₈ H ₁₇	(CH ₃) ₂ N ⁺		1.6289
71	C ₁₁ H ₂₃	C ₁₁ H ₂₃	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ O	2.5850	94	C ₈ H ₁₇	C ₈ H ₁₇	(CH ₃) ₂ N ⁺		1.6108
72	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ O	3.0000	95	C ₁₀ H ₂₁	C ₁₀ H ₂₁	(CH ₃) ₂ N ⁺	(CH ₂ CH(OH)) ₂	2.4318
73	C ₁₃ H ₂₇	C ₁₃ H ₂₇	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ O	3.4202	96	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(CH ₃) ₂ N ⁺	(CH ₂ CH(OH)) ₂	3.1549
74	C ₁₄ H ₂₉	C ₁₄ H ₂₉	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ O	3.7212	97	C ₁₄ H ₂₉	C ₁₄ H ₂₉	(CH ₃) ₂ N ⁺	(CH ₂ CH(OH)) ₂	4.0706
75	C ₁₆ H ₃₃	C ₁₆ H ₃₃	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ O	3.1938	98	C ₁₆ H ₃₃	C ₁₆ H ₃₃	(CH ₃) ₂ N ⁺	(CH ₂ CH(OH)) ₂	4.3010
76	C ₆ H ₁₃	C ₆ H ₁₃	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ S	0.8861	99 [†]	C ₁₀ H ₂₁	C ₁₀ H ₂₁		(CH ₂) ₄	2.5702
77	C ₈ H ₁₇	C ₈ H ₁₇	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ S	1.4318	100 [†]	C ₁₂ H ₂₅	C ₁₂ H ₂₅		(CH ₂) ₄	3.2518
78	C ₁₀ H ₂₁	C ₁₀ H ₂₁	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ S	2.2076						
79 [†]	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ S	3.0177						
80 [†]	C ₁₄ H ₂₉	C ₁₄ H ₂₉	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ S	4.0000						
81	C ₁₆ H ₃₃	C ₁₆ H ₃₃	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ S	5.0132						
82	C ₁₈ H ₃₇	C ₁₈ H ₃₇	(CH ₃) ₂ N ⁺	(C ₂ H ₄) ₂ S	6.0000						
83	C ₁₀ H ₂₁	C ₁₀ H ₂₁	(CH ₃) ₂ N ⁺	CH ₂ C≡CCH ₂	2.8570						
84	C ₈ H ₁₇	C ₈ H ₁₇	(CH ₃) ₂ N ⁺	CH ₂ C≡CCH ₂	1.5935						
85	C ₁₂ H ₂₅	C ₁₂ H ₂₅		CH ₂ C≡CCH ₂	3.0506						
86	C ₁₂ H ₂₅	C ₁₂ H ₂₅		CH ₂ C≡CCH ₂	3.0410						
87	C ₈ H ₁₇	C ₈ H ₁₇		CH ₂ C≡CCH ₂	1.4634						

Table 1. Structures and CMC values of 211 Gemini surfactants according to Figure 1 (101-136)

No.	Tail 1	Tail 2	lon-1	lon-2	Spacer	Exp. pCMC	No.	Tail 1	Tail 2	lon-1	lon-2	Spacer	Exp. pCMC
101	C ₁₀ H ₂₁	C ₁₀ H ₂₁			(CH ₂) ₆	2.6990	127	C ₁₃ H ₂₇	C ₁₃ H ₂₇	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₂	3.3979
102	C ₁₂ H ₂₅	C ₁₂ H ₂₅				3.2676	128	C ₁₂ H ₂₅	C ₁₂ H ₂₅			CH ₂ COO(CH ₂) ₂ - -OCOCH ₂	3.2218
103	B-C ₈ H ₁₇	B-C ₈ H ₁₇	PO ₄ ⁻	(CH ₃) ₂ N ⁺	C ₂ H ₄	3.7212	129	C ₁₂ H ₂₅	C ₁₂ H ₂₅				3.1367
104	B-C ₈ H ₁₇	B-C ₈ H ₁₇	PO ₄ ⁻	(CH ₃) ₂ N ⁺		3.1549	130	C ₁₂ H ₂₅	C ₁₂ H ₂₅				3.0132
105	B-C ₈ H ₁₇	C ₉ H ₁₉	PO ₄ ⁻	(CH ₃) ₂ N ⁺		3.6990							
106	B-C ₈ H ₁₇	C ₁₁ H ₂₃	PO ₄ ⁻	(CH ₃) ₂ N ⁺		3.8239							



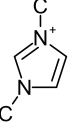
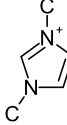
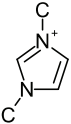

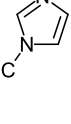

No.	Tail 1	Tail 2	lon-1	lon-2	Spacer	Exp. pCMC	No.	Tail 1	Tail 2	lon-1	lon-2	Spacer	Exp. pCMC
107	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(C ₂ H ₄ O) ₂ C ₂ H ₄	3.0362	131	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(C ₃ H ₇)	(C ₂ H ₅) N ⁺	CH ₂ COO(CH ₂) ₂ -OCOCH ₂	3.0362
108	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(C ₂ H ₄ O) ₃ C ₂ H ₄	3.0655	132	C ₁₂ H ₂₅	C ₁₂ H ₂₅			C ₃ H ₇	3.0809
109 ^f	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(C ₂ H ₄ O) ₄ C ₂ H ₄	3.0458	133	C ₁₂ H ₂₅	C ₁₂ H ₂₅			CH ₂ S(CH ₂) ₂ SCH ₂	3.3979
110 ^f	C ₁₂ H ₂₅	C ₁₂ H ₂₅	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(C ₂ H ₄ O) ₅ C ₂ H ₄	2.9101							
111	E-C ₁₂ O	E-C ₁₂ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₃	3.2596							
112	E-C ₁₂ O	E-C ₁₂ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₄	3.2218							
113	E-C ₁₂ O	E-C ₁₂ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₆	3.2403							
114	E-C ₁₄ O	E-C ₁₄ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₃	3.5800							
115	E-C ₁₄ O	E-C ₁₄ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₄	3.3799							
116	E-C ₁₄ O	E-C ₁₄ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₆	3.4802	134	C ₁₂ H ₂₅	C ₁₂ H ₂₅			CH ₂ S(CH ₂) ₃ SCH ₂	3.4559
117	E-C ₁₆ O	E-C ₁₆ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₃	4.5406	135	C ₁₂ H ₂₅	C ₁₂ H ₂₅			CH ₂ S(CH ₂) ₄ SCH ₂	3.5086
118	E-C ₁₆ O	E-C ₁₆ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₄	4.3507							
119 ^f	E-C ₁₆ O	E-C ₁₆ O	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₆	4.4101							
120 ^f	C ₁₄ H ₂₉	C ₁₄ H ₂₉	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺	(CH ₂) ₆	3.8013							
121	C ₈ H ₁₇	C ₈ H ₁₇	CH ₃) ₂ N ⁺	CH ₃) ₂ N ⁺	CH ₂ COO(CH ₂) ₂ -OCOCH ₂	1.9574							
122	C ₁₀ H ₂₁	C ₁₀ H ₂₁	CH ₃) ₂ N ⁺	CH ₃) ₂ N ⁺		2.6326	136	C ₁₄ H ₂₉	C ₁₄ H ₂₉			CH ₂ S(CH ₂) ₂ SCH ₂	4.0000
123	C ₁₂ H ₂₅	C ₁₂ H ₂₅	CH ₃) ₂ N ⁺	CH ₃) ₂ N ⁺		2.9747							
124	C ₁₄ H ₂₉	C ₁₄ H ₂₉	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺		3.8239							
125	C ₁₆ H ₃₃	C ₁₇ H ₃₅	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺		4.1549							
126	C ₁₈ H ₃₇	C ₁₈ H ₃₇	(CH ₃) ₂ N ⁺	(CH ₃) ₂ N ⁺		4.6990							

Table 1. Structures and CMC values of 211 Gemini surfactants according to Figure 1 (137-193)

No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC	No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC
137	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ S(CH ₂) ₃ -SCH ₂	4.0706	165	C ₈ H ₁₇	C ₈ H ₁₇		3.3760	2.7900
138	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ S(CH ₂) ₄ -SCH ₂	4.1487	166	C ₁₀ H ₁₁ CH-	C ₁₀ H ₁₁ CH-		C ₈ H ₁₆	2.7990
139 ^f	C ₁₆ H ₃₃	C ₁₆ H ₃₃		CH ₂ S(CH ₂) ₂ -SCH ₂	4.6576	167	C ₁₀ H ₁₁ CH-	C ₁₀ H ₁₁ CH-		C ₆ H ₁₂	2.8040
140 ^f	C ₁₆ H ₃₃	C ₁₆ H ₃₃		CH ₂ S(CH ₂) ₃ -SCH ₂	4.6778	168	C ₉ H ₁₉	C ₉ H ₁₉		CH ₂ O(CH ₂) ₂ -OCH ₂	2.8100
141	C ₁₆ H ₃₃	C ₁₆ H ₃₃		CH ₂ S(CH ₂) ₄ -SCH ₂	4.6990	169 ^f	C ₈ H ₁₇	C ₈ H ₁₇			2.8300
142	CH ₃	CH ₃		CH ₂ S(CH ₂) ₃ -SCH ₂	2.1800	170 ^f	C ₁₀ H ₁₁ CH-	C ₁₀ H ₁₁ CH-		C ₅ H ₁₀	2.8330
143	C ₂ H ₅	C ₂ H ₅		CH ₂ S(CH ₂) ₂ -SCH ₂	2.2300	171	C ₁₀ H ₁₁ CH-	C ₁₀ H ₁₁ CH-		C ₄ H ₈	2.8540
144	C ₂ H ₅	C ₂ H ₅		CH ₂ S(CH ₂) ₃ -SCH ₂	2.2600	172	C ₁₀ H ₁₁ CH-	C ₁₀ H ₁₁ CH-		C ₃ H ₆	2.8630
145	C ₂ H ₅	C ₂ H ₅		CH ₂ S(CH ₂) ₄ -SCH ₂	2.3000	173	C ₈ H ₁₇	C ₈ H ₁₇		CH ₂ O(CH ₂) ₆ -OCH ₂	2.8700
146	C ₂ H ₅	C ₂ H ₅		CH ₂ S(CH ₂) ₅ -SCH ₂	2.3300	174	C ₉ H ₁₉	C ₉ H ₁₉		CH ₂ O(CH ₂) ₄ -OCH ₂ H ₂	2.8800
147	C ₈ H ₁₇	C ₈ H ₁₇		CH ₂ CH- -(OH)CH ₂	2.3600	175	C ₉ H ₁₉	C ₉ H ₁₉		CH ₂ O(CH ₂) ₅ -OCH ₂	2.9100
148	C ₂ H ₅	C ₂ H ₅		CH ₂ S(CH ₂) ₆ -SCH ₂	2.3700	176	C ₁₀ H ₂₁	C ₁₀ H ₂₁		CH ₂ O(CH ₂) ₂ -OCH ₂	2.9200
149 ^f	C ₄ H ₉	C ₄ H ₉		CH ₂ S(CH ₂) ₂ -SCH ₂	2.4100	177	C ₁₀ H ₂₁	C ₁₀ H ₂₁		CH ₂ O(CH ₂) ₃ -OCH ₂	2.9500

No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC	No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC
150 [†]	C ₃ H ₇	C ₃ H ₇		CH ₂ S(CH ₂) ₅ - -SCH ₂	2.4100	178	C ₉ H ₁₉	C ₉ H ₁₉		CH ₂ O(CH ₂) ₆ - -OCH ₂	2.9500
151	C ₄ H ₉	C ₄ H ₉		CH ₂ S(CH ₂) ₅ - -SCH ₂	2.4900	179 [†]	C ₁₁ H ₂₃	C ₁₁ H ₂₃		CH ₂ O(CH ₂) ₃ - -OCH ₂	3.0700
152	C ₅ H ₁₁	C ₅ H ₁₁		CH ₂ S(CH ₂) ₄ - -SCH ₂	2.5300	180 [†]	C ₁₀ H ₂₁	C ₁₀ H ₂₁		CH ₂ O(CH ₂) ₅ - -OCH ₂	3.0700
153	C ₄ H ₉	C ₄ H ₉		CH ₂ S(CH ₂) ₆ - -SCH ₂	2.5400	181	C ₁₁ H ₂₃	C ₁₁ H ₂₃		CH ₂ O(CH ₂) ₄ - -OCH ₂	3.1300
154	C ₅ H ₁₁	C ₅ H ₁₁		CH ₂ S(CH ₂) ₅ - -SCH ₂	2.5800	182	C ₁₂ H ₂₅	C ₁₂ H ₂₅		C ₆ H ₁₂	3.1310
155	C ₆ H ₁₃	C ₆ H ₁₃		CH ₂ S(CH ₂) ₂ - -SCH ₂	2.5800	183	C ₁₂ H ₂₅	C ₁₂ H ₂₅		C ₄ H ₈	3.1370
156	C ₁₀ H ₂₁	C ₁₀ H ₂₁		CH ₂ CH- -(OH)CH ₂	2.5970	184	C ₁₂ H ₂₅	C ₁₂ H ₂₅		CH ₂ O(CH ₂) ₂ - -OCH ₂	3.1500
157	C ₅ H ₁₁	C ₅ H ₁₁		CH ₂ S(CH ₂) ₆ - -SCH ₂	2.6200	185	C ₁₁ H ₂₃	C ₁₁ H ₂₃		CH ₂ O(CH ₂) ₅ - -OCH ₂	3.1600
158	C ₆ H ₁₃	C ₆ H ₁₃		CH ₂ S(CH ₂) ₄ - -SCH ₂	2.6400	186	C ₁₂ H ₂₅	C ₁₂ H ₂₅		CH ₂ O(CH ₂) ₃ - -OCH ₂	3.1800
159 [†]	C ₇ H ₁₅	C ₇ H ₁₅		CH ₂ S(CH ₂) ₂ - -SCH ₂	2.6500	187	C ₁₂ H ₂₅	C ₁₂ H ₂₅		(CH ₂) ₂	3.1940
160 [†]	C ₆ H ₁₃	C ₆ H ₁₃		CH ₂ S(CH ₂) ₅ - -SCH ₂	2.6700	188	C ₁₁ H ₂₃	C ₁₁ H ₂₃		CH ₂ O(CH ₂) ₆ - -OCH ₂	3.2000
161	C ₇ H ₁₅	C ₇ H ₁₅		CH ₂ S(CH ₂) ₄ - -SCH ₂	2.7100	189 [†]	C ₁₂ H ₂₅	C ₁₂ H ₂₅		CH ₂ O(CH ₂) ₄ - -OCH ₂	3.2200
162	C ₆ H ₁₃	C ₆ H ₁₃		CH ₂ S(CH ₂) ₆ - -SCH ₂	2.7100	190 [†]	C ₁₂ H ₂₅	C ₁₂ H ₂₅		CH ₂ O(CH ₂) ₅ - -OCH ₂	3.2500
163	C ₈ H ₁₇	C ₈ H ₁₇		CH ₂ S(CH ₂) ₂ - -SCH ₂	2.7200	191	C ₁₂ H ₂₅	C ₁₂ H ₂₅		CH ₂ O(CH ₂) ₆ - -OCH ₂	3.2900
164	C ₈ H ₁₇	C ₈ H ₁₇		CH ₂ S(CH ₂) ₃ - -SCH ₂	2.7500	192	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ O(CH ₂) ₂ - -OCH ₂	3.3400
						193	C ₁₆ H ₃₃ OCH ₂ - -CH(OH)CH ₂	C ₁₆ H ₃₃ OCH ₂ - -CH(OH)CH ₂		(CH ₂) ₂	3.3760

Table 1. Structures and CMC values of 211 Gemini surfactants according to Figure 1 (194-211)

No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC	No.	Tail 1	Tail 2	lon-1 or lon-2	Spacer	Exp. pCMC
194	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ O(CH ₂) ₄ OCH ₂	3.4000	203 [†]	C ₆ H ₁₃	C ₆ H ₁₃		CH ₂ O(CH ₂) ₃ OCH ₂	2.6100
195	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ O(CH ₂) ₆ OCH ₂	3.4700	204 [†]	C ₇ H ₁₅	C ₇ H ₁₅		CH ₂ O(CH ₂) ₃ OCH ₂	2.6800
196	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ O(CH ₂) ₂ OCH ₂	3.5200	205 [†]	C ₇ H ₁₅	C ₇ H ₁₅		CH ₂ O(CH ₂) ₅ OCH ₂	2.7500
197	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ O(CH ₂) ₃ OCH ₂	3.5800	206 [†]	C ₇ H ₁₅	C ₇ H ₁₅		CH ₂ O(CH ₂) ₆ OCH ₂	2.7900
198	C ₁₄ H ₂₉	C ₁₄ H ₂₉		CH ₂ O(CH ₂) ₆ OCH ₂	3.6500	207 [†]	C ₉ H ₁₉	C ₉ H ₁₉		CH ₂ O(CH ₂) ₃ OCH ₂	2.8400
199 [†]	CH ₃	CH ₃		CH ₂ O(CH ₂) ₂ OCH ₂	2.1500	208 [†]	C ₁₀ H ₂₁	C ₁₀ H ₂₁		CH ₂ O(CH ₂) ₄ OCH ₂	3.0000
200 [†]	CH ₃	CH ₃		CH ₂ O(CH ₂) ₅ OCH ₂	2.2500	209 [†]	C ₁₁ H ₂₃	C ₁₁ H ₂₃		CH ₂ O(CH ₂) ₂ OCH ₂	3.0400
201 [†]	C ₄ H ₉	C ₄ H ₉		CH ₂ O(CH ₂) ₄ OCH ₂	2.4500	210 [†]	C ₁₀ H ₂₁	C ₁₀ H ₂₁		CH ₂ O(CH ₂) ₆ OCH ₂	3.0700
202 [†]	C ₃ H ₇	C ₃ H ₇		CH ₂ O(CH ₂) ₆ OCH ₂	2.4500	211 [†]	C ₁₄ H ₂₉	C ₁₄ H ₂₉		(CH ₂) ₄	3.9210

Notes: 1-141, 166-167, 170-172, 182-183, 187, 195, 211 di-bromide Gemini surfactants, the rest are di-chloride; C_mH_n (No. 103-106 Gemini surfactants) means branched alkyl group; E-C_mO (No. 111-119 Gemini surfactants) represents esterquat Gemini surfactants with the formula: [H_{2m-1}C_{m-1}COOCH₂CH₂(CH₃)₂N⁺(CH₂)_nN⁺(CH₃)₂CH₂CH₂COOCC_{m-1}H_{2m-1}]₂Br⁻; [†]the chemicals were in the test set, and others were in the training set

2. The optimized 2D-structures of the studied surfactants were after that uploaded to alvaDesc software to compute 5305 descriptors for each surfactant.

3. To avoid overfitting, the number of variables (descriptors) must be reduced based on the following steps and using alvaDesc software.

a. Constant and near constant descriptor values were removed;

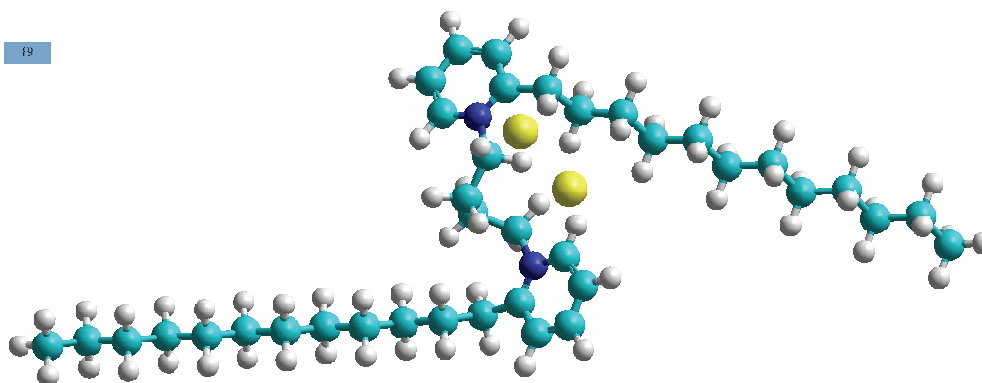


Figure 2. Example of the drawn structures of the investigated Gemini surfactants by HyperChem software (compound number 100).

b. All missing values or at least one missing value were removed;

c. Every descriptor with a relative standard deviation <0.001 was removed;

d. Descriptors with pair correlation ≥ 0.95 were removed.

4. Relevant descriptors were screened after performing stepwise MLR with F value method. 20 descriptors were selected among 5305 with stepwise regression. For each model, we have used the same type and number of descriptors, as well as the same composition of the training and test sets. A set of the selected descriptors is summarized in Table 2 with their signification and types. The datasets with linear

chemical formula of each compound are available from the author upon request).

Modeling techniques

alvaDesc is a tool for the calculation of molecular descriptors. It calculates almost 5305 1, 2, 3-dimensional descriptors divided into 30 logical blocks (3885-2D and 1420-3D) as represented in Figure 3.

alvaModel is a software tool to create quantitative structure-property relationship (QSPR) models using descriptors previously calculated in alvaDesc. This study deals with five different modelling techniques cited below.

- OLS: an ordinary least squares model;
- KNN: a K -nearest neighbours regression model;

Table 2. Descriptions of the retained descriptors obtained from alvaDesc software

No.	Descriptor code	Descriptor description	Descriptor type
1	nP	Number of Phosphorous atoms	Constitutional indices
2	Mi	Mean first ionization potential (scaled on Carbon atom)	
3	IC5	Information Content index (neighbourhood symmetry of 5-order)	Information indices
4	CATS2D_03_AP	CATS2D Acceptor-Positive at lag 03	Pharmacophore descriptors
5	F02[C-N]	Frequency of C-N at topological distance 2	
6	F04[C-S]	Frequency of C-S at topological distance 4	2D Atom Pairs
7	F03[O-O]	Frequency of O-O at topological distance 3	
8	F08[C-C]	Frequency of C-C at topological distance 8	
9	nCp	Number of terminal primary C(sp3)	Functional group counts
10	Wap	All-path Wiener index	Topological indices
11	VE1_B(i)	Coefficient sum of the last eigenvector (absolute values) from Burden matrix weighted by ionization potential	2D matrix-based descriptors
12	SM6_B(p)	Spectral moment of order 6 from Burden matrix weighted by polarizability	2D matrix-based descriptors
13	Chi_Dz(p)	Randic-like index from Barysz matrix weighted by polarizability	
14	H_Dz(Z)	Harary-like index from Barysz matrix weighted by atomic number	
15	Eta_sh_y	Eta y shape index	ETA indices
16	Eig05_AEA(dm)	Eigenvalue n. 5 from augmented edge adjacency mat. weighted by dipole moment	Edge adjacency indices
17	MATS2e	Moran autocorrelation of lag 2 weighted by Sanderson electronegativity	2D autocorrelations
18	MATS3v	Moran autocorrelation of lag 3 weighted by van der Waals volume	
19	SM05_EA(b0)	Spectral moment of order 5 from edge adjacency mat. weighted by bond order	Edge adjacency indices
20	D/Dtr05	Distance/detour ring index of order 5	Ring descriptors

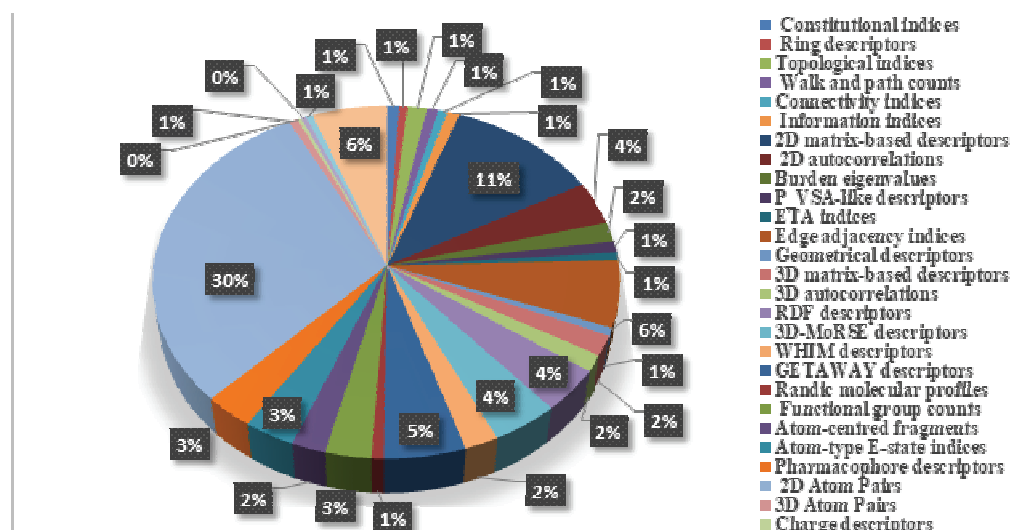


Figure 3. Block name of descriptors.

- PLS: partial least squares method;
- ANN: artificial neural networks;
- SVR: support vector regression.

The performance and predictive power of the QSAR developed models in this study was performed between the experimental and predicted critical micelle concentration vectors using different statistical criterion, including the root mean squared error (*RMSE*) [60], average absolute relative deviation (*AARD*) [34,61], correlation and determination coefficient and other regression coefficients (R , R^2 , R_0^2 , $R_0'^2$, R_m^2) [61], accuracy and bias factors (A_f , B_f) [61], four different regression metrics Q^2 (Q_{F1}^2 , Q_{F2}^2 , Q_{F3}^2 , Q_{CCC}^2) [61], slopes k and k' of the regression lines through the origin [61]. A QSPR model can be predictive, if the listed errors satisfy the following conditions: *AARD* and *RMSE* are close to zero, $Q^2 > 0.5$, $R^2 > 0.6$, $0.85 \leq k \leq 1.15$, $0.85 \leq k' \leq 1.15$, A_f and $B_f \geq 1$ [61-63]. The values of these assessing parameters are calculated and presented in greater detail in Table 3.

RESULTS AND DISCUSSION

Linear modeling of CMC

Ordinary least squares model modeling results

The ordinary least squares is a way for modeling the relationship between a target value and one or more features. Since the performance of any model depends mainly and strongly on the size and the quality of the dataset, the number of datasets over the number of variables would have to be higher than 5 to realize a reliable and safe model [64]. In this work, this ratio is equal to $211/5 \approx 11$, which shows that the number of datasets is satisfactory for modeling the investigated problem. Figure 4 shows the workshop of the study and Figure 5a shows dispersion between *pCMC* values estimated by the OLS model and measured. *pCMC* estimated by OLS are in linear accordance with the measured values with a very acceptable correlation and determination coefficient. Table 3

Table 3. Statistical quality of all developed models for the global dataset

Model	Q_{F1}^2	Q_{F2}^2	Q_{F3}^2	Q_{CCC}^2	R^2	R	k	k'	R_0^2	$R_0'^2$	R_m^2	A_f	B_f	<i>AARD</i>	<i>RMSE</i>
KNN	0.1946	0.1941	0.2049	0.4876	0.1941	0.5081	0.9672	0.9631	0.9781	0.9842	0.0222	1.2300	1.0390	22.0808	0.8741
PLS	0.8295	0.8295	0.8297	0.9062	0.8295	0.9108	1.0028	0.9826	0.9999	0.9965	0.4871	1.1196	1.0093	11.8391	0.4021
OLS	0.9682	0.9682	0.9686	0.9838	0.9682	0.9840	1.0004	0.9969	1.0000	0.9999	0.7954	1.0447	1.0050	4.3317	0.1737
ANN-MLP	0.9265	0.9264	0.9265	0.9661	0.9264	0.9713	1.0101	0.9839	0.9990	0.9970	0.6769	1.0548	1.0019	4.1815	0.2641
SVR-DA	0.9971	0.9971	0.9972	0.9985	0.9971	0.9985	1.0002	0.9996	1	1	0.9433	1.0067	1.0001	0.666	0.0525

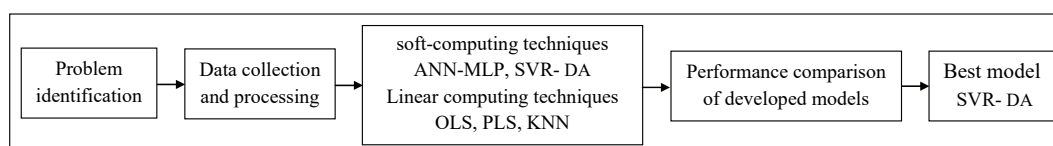


Figure 4. Workshop of the study.

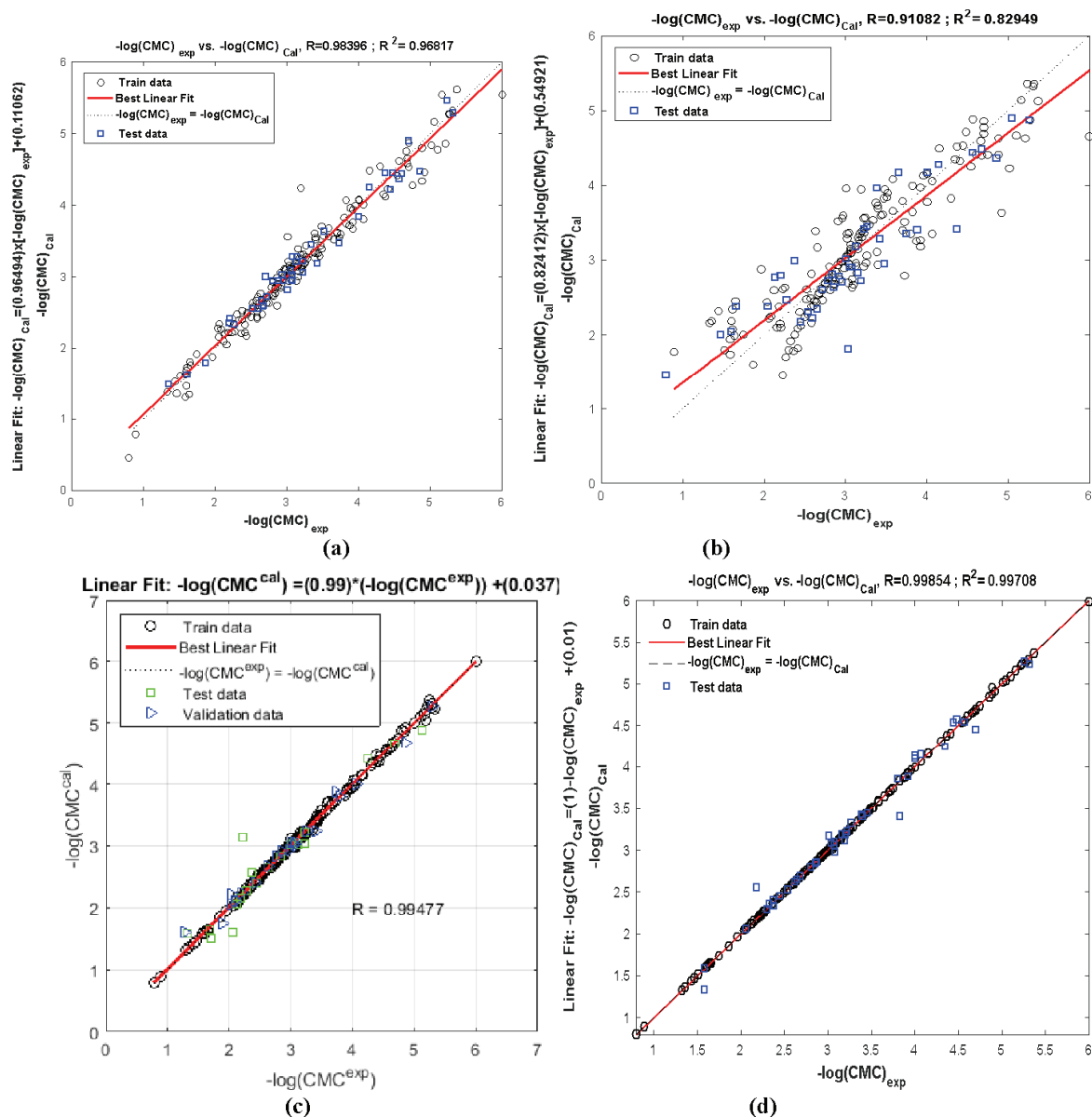


Figure 5. Scatter plot of the global data adjusted by the optimal model between predicted and experimental $p\text{CMC}$ values for the global dataset: a) OLS; b) PLS; c) ANN; d) SVR.

represents different statistical indicators for the best OLS-QSPR model; more details are available from the authors upon request.

The OLS model was done by alvaModel 1.0.4 software and can be written in linear way *versus* descriptors and non-standardized coefficients:

$$\begin{aligned}
 -\log\text{CMC} = & 5.863 + 1.887 \times nP - 0.561 \times IC5 - 0.413 \times \text{CATS2D}_{03_Ap} + 0.116 \times F04[C-S] + \\
 & + 12.831 \times Mi - 0.154 \times nCp + 0 \times Wap - 1.307 \times VEI_B(i) - 2.817 \times \text{Eig05_AEA}(dm) + 0.777 \times \text{SM6_B}(p) + \\
 & + 2.719 \times \text{MATS3v} + 16.828 \times \text{Eta_sh_y} - 41.677 \times \text{Chi_Dz}(p) - 1.537 \times \text{SM05_EA}(bo) + 0.025 \times H_Dz(Z) - \\
 & - 0.232 \times F03[O-O] - 1.822 \times \text{MATS2e} - 0.001 \times \frac{D}{Dtr05} - 0.018 \times F08[C-C] + 0.066 \times F02[C-N]
 \end{aligned} \quad (1)$$

PLS modeling results

The PLS-QSPR model of the CMC of 211 Gemini surfactants was successfully developed based on various molecular descriptors. The plot of predicted

and experimental critical micellar concentration is shown in Figure 5b with determination coefficient of 0.8295 and average absolute relative deviation of 11.8391 for all dataset indicates medium correlation

between the experimental and predicted values and does not confirm the good predictive ability of PLS-QSPR model.

$$\begin{aligned}
 -\log\text{CMC} = & -3.98 + 0.66 \times nP - 0.49 \times IC5 - 0.06 \times \text{CATS2D}_{03_Ap} + 0.05 \times F04[C-S] + \\
 & + 1.46 \times Mi - 0.38 \times nCp + 0 \times Wap - 0.26 \times VEI_B(i) + 0.43 \times \text{Eig05_AEA(dm)} + 1.13 \times M6_B(p) + \\
 & + 0.43 \times \text{MATS3v} - 1.39 \times \text{Eta_sh_y} - 14.46 \times \text{Chi_Dz(p)} - 0.06 \times \text{SM05_EA(bo)} + 0.01 \times H_Dz(Z) - \\
 & - 0.14 \times F03[O-O] + 0.82 \times \text{MATS2e} + 0 \times \frac{D}{Dtr05} + 0.01 \times F08[C-C] - 0.12 \times F02[C-N]
 \end{aligned} \quad (2)$$

More details of the performance of this model for the global data set are illustrated in Table 3.

Non-linear modeling of CMC

KNN modelling results

The *K*-nearest neighbors algorithm is a non-parametric method often used for classification, but it can also be used for regression. In this paper, we have applied the KNN-QSPR approach to a data set of 211 compounds. The obtained statistical criterion is shown in Table 3 where the regression and error values do not guarantee the acceptable predictive ability of a KNN-QSPR model. The number of neighbors was selected to be $k = 5$, and the distance that must be used to identify the closest ones was determined by Euclidean distance method.

Artificial neural network modeling results

In this study, the non-linear relationship between CMC of Gemini surfactants and their descriptors was investigated using the artificial neural network (ANN) model. The training, validation and testing of the feedforward artificial neural network (FFANN) with Levenberg-Marquardt (LM) training algorithm was carried out using the Visual Gene Developer free software under optimization of several parameters and using trial and error method such as the type of transfer functions, number of hidden neurons, learning rate and momentum coefficient. In order to measure the performance of the developed ANN for CMC

The PLS-QSPR model can be written in linear way *versus* descriptors and non-standardized coefficients:

modeling, different types of statistical parameters were employed to calculate the generalization error and regression. In this study, fourteen parameters were selected to investigate the network response with respect to the experimental CMC. The datasets with 211 values of CMC were divided randomly to avoid the overfitting into 80, 10 and 10% for training, validation and test, respectively. Twenty descriptors were selected to build the ANN and inputs/outputs were scaled in the range of -1 to 1 to avoid numerical round-off effects. Weights and threshold were randomly assigned during the training process. Characteristics of the network are summarized in Tables 3 and 4.

A graphical visualization of weights and thresholds is depicted in Figure 6, where lines represent weight constants and nodes represent threshold values. In this plot, the red color represents to the high positive number, and the violet color means high negative number. The line width is proportional to the absolute number of weight factor or threshold value.

Results of the linear regression analysis for the best developed model are presented in Figure 5c. The values of R^2 and mean *AARD* for all databases were closer to 1 and 0, respectively, and the values of slope and *Y*-intercept for the global dataset were closer to 1 and 0, respectively. Based on the high values of coefficients of regression and low errors, it can be said that the developed ANN model gave excellent prediction values for the design outputs.

Table 4. Optimized parameters of the ANN model with its statistical quality

Training setting		Topology setting	
Learning rate	0.01	Number of input variables	20
Momentum coefficient	0.1	Number of output variables	01
Hidden transfer function	Hyperbolic tangent	Number of hidden layers	01
Output transfer function	Linear	Node in the hidden layer	09
Maximum number of training cycles	10^7	<i>AARD</i> _training	0.792
Target error	10^{-5}	<i>AARD</i> _validation	3.179
Initialization method of threshold	Random	<i>AARD</i> _global	1.597
Initialization of weight factor	Random	R^2 _training	0.998
Analysis update interval	500 cycles	R^2 _validation	0.987
Processing time	16 h-15 min-19 s	R^2 _global	0.990

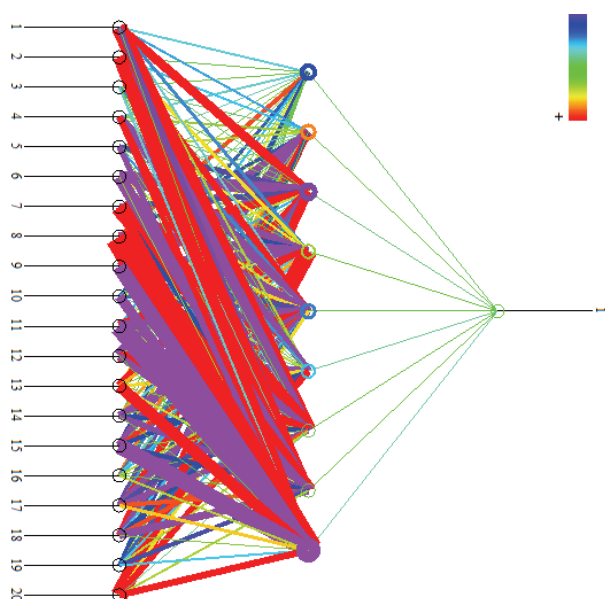


Figure 6. ANN map analysis.

The best ANN model can be written in non-linear form as in Eq. (3), where inputs must be scaled before introducing them in the model and weights and biases obtained during the training stage:

$$-\log(\text{CMC}) = \sum_{s=1}^9 \left\{ W_{o(1,s)} \left[\frac{2}{1 + e^{-2 \left(b_{1(s)} + \sum_{k=1}^{20} W_{i(s,k)} \cdot \text{Input}(k) \right)}} - 1 \right] \right\} + b_{2(1)} \quad (3)$$

Dragonfly-support vector machine for regression modeling results

The SVR algorithm was first proposed by Cortes and Vapnik as one of the most accurate and robust algorithms in data mining [65]. To design the SVR model, three kernel functions have been tested, namely linear, Gaussian and polynomial kernel functions. The SVR model coupled with Dragonfly algorithm (DA) was developed by writing a MATLAB script. Dragonfly algorithm was used to find the optimized parameters of the model. The best kernel function prediction against and the optimal values of parameters obtained are summarized in Table 5.

Figure 5d shows the correlation between experimental and calculated CMC by SVR-DA model for all databases.

Statistical indices, including regression coefficient R , R^2 and AARD of 0.9985, 0.9971 and 0.666,

respectively, revealed the high accuracy of the SVR-DA prediction against ANN, OLS, PLS and KNN models, respectively. Also, another comparison in Figure 7 shows the clear superiority of SVR-DA in modeling CMC of diverse Gemini surfactants.

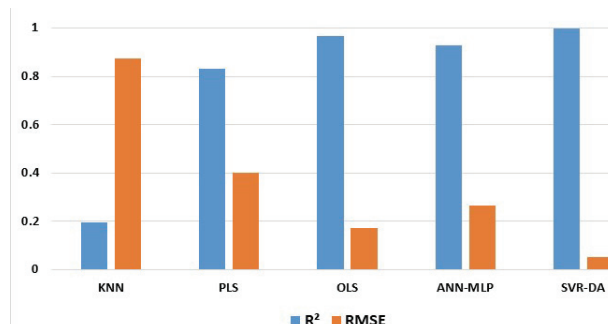


Figure 7. Performance comparison of the developed models.

Applicability domain investigation

The applicability domain (AD) is an alternative validation method to assess the ability of QSPR models by assessing structural similarity to a set of chemicals. Therefore, the applicability domain of the best developed QSPR model was assessed by the Euclidean distance-based method suitable when similarity is assessed in descriptor space and Williams plot. Euclidean distance-based method was included into the AMBIT Discovery software (version 0.04). The Williams plot is illustrated between standardized residuals (δ) versus leverage values (h), where δ is calculated by:

$$\delta = \frac{y_i - \hat{y}_i}{\sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - A - 1)}} \quad (4)$$

where y_i and \hat{y}_i are the experimental and the calculated value for the i -th compound, respectively; A is the number of descriptors and n is the number of compounds. The leverage values (h) can be defined as:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (5)$$

where x_i , x_i^T , X and X^T are the descriptor vector of the i -th compound, transpose of x_i , descriptor matrix and the transpose of X . The warning leverage value (h^*) was computed as:

$$h^* = 3(k + 1) / n \quad (6)$$

Table 5. Results of SVR-DA modeling

Penalty parameter ($C > 0$)	Gaussian kernel constant (γ)	Size of the insensitive zone (ϵ)	Kernel function	Quantity of support vectors (n)	Loss regression error
223.3974	1.9923	0.0135	Gaussian	151	0.0028

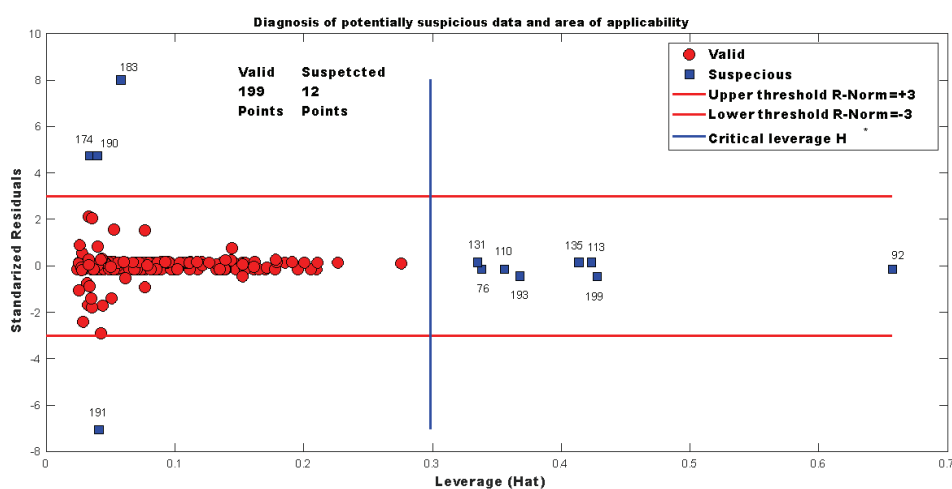
where k is the number of forecaster variables included in the model.

Figure 8a shows the applicability domain represented by Williams plot of the developed SVR-DA QSPR model where 199 compounds in both the training and the test set were in the domain, and 12 compounds were considered as outliers. For the Euclidean distance, all the compounds in both the training set and the test set were located in the acceptable region.

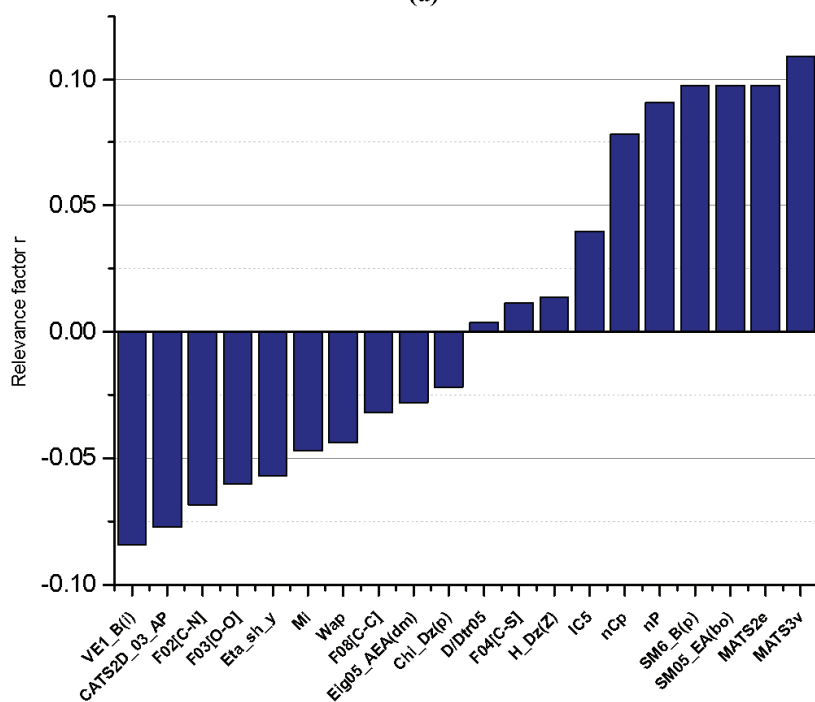
The obtained deviation against the experimental values can be explained by the fact that the selected descriptors for SVR-DA QSPR analysis could not capture some relevant structural features that may be present in these compounds. The obtained model can be used to estimate CMC of new Gemini surfactants if these are located later within the AD of the model.

Sensitivity analysis

In this section, a sensitivity analysis of the selected descriptors on the CMC has been undertaken



(a)



(b)

Figure 8. Applicability domains for the developed SVR-DA QSPR model based on Williams plot (a) and relevance factor plot of the descriptors to the pCMC of Gemini surfactants (b).

using the relevance factor (r) method to determine the existing dependency between inputs and the output. The relevance factor can be calculated using Eq. (7) and given in the range of [-1, 1]:

$$r_k = \frac{\sum_{i=1}^n (X_{k,i} - \bar{X}_k)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_{k,i} - \bar{X}_k)^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (7)$$

where $X_{k,i}$ is the i -th value of the k -th input vector with its average \bar{X}_k , Y_i is i -th output value and its average \bar{Y} and n is the number of compounds. As can be seen in Figure 8b, the CMC shows a straight dependency on 10 descriptors, and an opposite dependency on the rest of descriptors. In addition, MATS3v, VE1_B(i) are the most relevant input variables with a relevance factor of +0.1092 and -0.0842, respectively. In contrast, the Chi_Dz(p) and D/Dtr05 are the least relevant with a relevance factor of -0.0219 and 0.0037, respectively.

Comparison with literature studies

The performance of the optimized SVR-DA model is compared to the previously developed ones in literature for prediction of pCMC of 211 Gemini surfactants and only those containing sufficient statistical parameters for comparison. Details of the comparison are given in Table 6 in terms of R^2 and root mean squared error (RMSE). Results show the superiority of the proposed model to fit the non-linear behaviour of pCMC against the selected features.

Table 6. Performance of the proposed SVR-DA model against literature models developed for modeling pCMC of Gemini surfactants for the global dataset

Authors	Model name	Data points	R^2	RMSE
This study	SVR-DA QSPR	211	0.9971	0.0525
[10]	GA-LSSVM	120	0.964	0.222
[17]	PLS QSPR	94	0.985	0.153
[60]	ASNN QSPR	70	0.96	0.08

The obtained deviation against the experimental values can be explained by the fact that the selected descriptors for QSPR analysis could not capture some relevant structural features that may be present in these compounds.

CONCLUSION

In this work, a comparison between linear and non-linear fitting techniques to model the CMC of diverse Gemini surfactants has been conducted. Results show that the improved performance of SVR model for regression when the hyper-parameters are opti-

mized using dragonfly algorithm (SVR-DA) and using the calculated descriptors was capable to accurately predict the non-linear behaviour of critical micelle concentration of Gemini surfactants, more than ANN, OLS, PLS and KNN models for the global dataset. In addition, the applicability domain of the five models and for all Gemini surfactants has been conducted using Euclidean distance-based method and Williams plot. The SVR-DA can be used to estimate CMC of new Gemini surfactants if these are later located within the applicability domain of the model. Moreover, the sensitivity analysis of each input on the output has been carried on using the relevance factor method.

Acknowledgment

The authors thank the editor of this journal and the anonymous reviewers for their constructive comments, which helped us to improve the quality and the presentation of this paper.

REFERENCES

- [1] A.R. Katritzky, L. Pacureanu, D. Dobchev, M. Karelson, J. Chem. Inf. Model. (2007) 782-793
- [2] C.Y. Hu, S.J. Hua, Y.L. Lin, Y.G. Deng, Y.Z. Hou, Y.F. Du, C. Di Dong, C.W. Chen, C.H. Wu, Separ. Purific. Technol. 243 (2020) 116797
- [3] S.K. Yadav, K. Parikh, S. Kumar, Colloids Surfaces, A 514 (2017) 47-55
- [4] H.L. Zhu, Z.Y. Hu, J.L. Wang, D.L. Cao, J. Mol. Liquids 195 (2014) 54-58
- [5] S.M. Shakil Hussain, M.S. Kamal, M. Murtaza, J. Mol. Liquids 296 (2019) 111837
- [6] M. Yi, Z. Huang, J. Hao, Langmuir 35 (2019) 9538-9545
- [7] M.S. Kamal, J. Surfact. Deterg. 19 (2016) 223-236.
- [8] T. Ma, H. Feng, H. Wu, Z. Li, J. Jiang, D. Xu, Z. Meng, W. Kang, Colloids Surfaces, A 581 (2019) 123800
- [9] A.M. Cardoso, C.M. Morais, A.R. Cruz, A.L. Cardoso, S.G. Silva, M.L. Do Vale, E.F. Marques, M.C. Pedroso De Lima, A.S. Jurado, Mol. Pharm.. 12 (2015) 716-730
- [10] C. Guo, P. Zhou, J. Shao, X. Yang, Z. Shang, Chemosphere 84 (2011) 1608-1616
- [11] L. Jiao, Y. Wang, L. Qu, Z. Xue, Y. Ge, H. Liu, B. Lei, Q. Gao, M. Li, Colloids Surfaces, A 586 (2020) 124226
- [12] Z. Kardanpour, B. Hemmateenejad, T. Khayamian, Anal. Chim. Acta 531 (2005) 285-291
- [13] Z. Wang, G. Li, X. Zhang, R. Wang, A. Lou, Colloids Surfaces, A 197 (2002) 37-45
- [14] K. Roy, H. Kabir, Chem.Eng. Sci. 81 (2012) 169-178
- [15] D.W. Roberts, Langmuir 18 (2002) 345-352
- [16] A.R. Katritzky, L.M. Pacureanu, S.H. Slavov, D.A. Dobchev, D.O. Shah, M. Karelson, Comput. Chem. Eng. 33 (2009) 321-332
- [17] G. Absalan, B. Hemmateenejad, M. Soleimani, M. Akhond, R. Miri, QSAR Comb.Sci. 23 (2004) 416-425

- [18] K. Roy, H. Kabir, *Chem. Eng. Sci.* 87 (2013) 141-151
- [19] B. Bhhatarai, P. Gramatica, *Environ. Sci. Technol.* 45 (2011) 8120-8128
- [20] X. Li, G. Zhang, J. Dong, X. Zhou, X. Yan, M. Luo, *J. Mol. Struc: THEOCHEM* 710 (2004) 119-126
- [21] M. Poša, *Steroids* 76 (2011) 85-93
- [22] L. Jiao, X. Wang, S. Bing, Z. Xue, H. Li, *Quím. Nova* 38 (2015) 510-517
- [23] H.A. Martens, P. Dardenne, *Chemomet. Intell. Lab. Sys.* 44 (1998) 99-121
- [24] N. Pal, N. Saxena, A. Mandal, *Colloids Surfaces, A* 533 (2017) 20-32
- [25] T. Gaudin, P. Rotureau, I. Pezron, G. Fayet, *Ind. Eng. Chem. Res.* 55 (2016) 11716-11726
- [26] L. Jiao, X. Wang, S. Bing, Z. Xue, H. Li, *RSC Adv.* 5 (2015) 6617-6624
- [27] L. Jiao, X. Zhang, Y. Qin, X. Wang, H. Li, *Chemomet. Intell. Lab. Sys.* 157 (2016) 202-207
- [28] K. Parikh, S. Singh, A. Desai, S. Kumar, *J. Mol. Liquids* 278 (2019) 290-298
- [29] L. Jiao, Z. Xue, G. Wang, X. Wang, H. Li, *Chemomet. Intell. Lab. Sys.* 137 (2014) 91-96
- [30] A.S. Kulkarni, A.J. Kasabe, M.S. Bhatia, N.M. Bhatia, V.L. Gaikwad, *AAPS PharmSciTech* 20 (2019)
- [31] H. Benimam, C.S. Moussa, M. Hentabli, S. Hanini, M. Laidi, *J. Chem. Eng. Data* 65 (2020) 3161-3172
- [32] C.B. Carvalho, E.P. Carvalho, M.A.S.S. Ravagnani, *Chem. Ind. Chem. Eng. Q.* 26 (2020) 125-134
- [33] M. Beigi, M. Toriki-Harchegani, M. Mahmoodi-Eshkaftaki, *Chem. Ind. Chem. Eng. Q.* 23 (2017) 251-258
- [34] A.A. El Hadj, C. Si-Moussa, S. Hanini, M. Laidi, *Chem. Ind. Chem. Eng. Q.* 19 (2013) 449-460
- [35] P.S. Milić, K.M. Rajković, P.M. Milićević, S.M. Milić, T.P. Brdarić, V.M. Pavelkić, *Chem. Ind. Chem. Eng. Q.* 19 (2013) 141-152
- [36] H. Golmohammadi, A. Rashidi, S.J. Safdari, *Chem. Ind. Chem. Eng. Q.* 19 (2013) 321-331
- [37] J.B. Savković-Stevanović, D.R. Francescetti, *Chem. Ind. Chem. Eng. Q.* 6 (2000) 361-371.
- [38] M. Esfahanian, M. Nikzad, G. Najafpour, A. Asghar Ghoreyshi, *Chem. Ind. Chem. Eng. Q.* 19 (2013) 241-252
- [39] A. Ghaderi, S. Abbasi, A. Motevali, S. Minaei, *Chem. Ind. Chem. Eng. Q.* 18 (2012) 283-293
- [40] M.S. Leite, M.A. Santos, E.M.F. Costa, A. Balieiro, Á.S. Lima, O.L. Sanchez, C.M.F. Soares, *Chem. Ind. Chem. Eng. Q.* 25 (2019) 369-382
- [41] A.A. Amooey, M. Ahangarian, F. Rezazadeh, *Chem. Ind. Chem. Eng. Q.* 20 (2014) 565-569
- [42] S.K. Lahiri, K.C. Ghanta, *Asia-Pacific J. Chem. Eng.* 5 (2010) 763-777
- [43] J. Ding, H. Wang, K. Dai, Y. Zi, Z. Shi, *Chem. Ind. Chem. Eng. Q.* 20 (2014) 29-38
- [44] Z. Kardanpour, B. Hemmateenejad, T. Khayamian, *Anal. Chim. Acta* 531 (2005) 285-291
- [45] J. Wu, H. Gao, D. Shi, Y. Yang, Y. Zhang, W. Zhu, *J. Mol. Liquids* 299 (2020) 112248
- [46] M.J. Rosen, L. Liu, *JAOCS, J. Am. Oil Chem. Soc.* 73 (1996) 885-890
- [47] E. Alami, G. Beinert, P. Marie, R. Zana, *Langmuir* 9 (1993) 1465-1467
- [48] M. Dreja, K. Heine, B. Tieke, G. Junkers, *J. Colloid Interface Sci.* 191 (1997) 131-140
- [49] G. Savelli, L. Brinchi, R. Germani, *Surfact. Sci. Ser.* (2001) 175-246
- [50] F.M. Menger, J.S. Keiper, *Angew. Chem. Int. Ed.* 39 (2000) 1906-1920
- [51] M. Dreja, W. Pyckhout-Hintzen, H. Mays, B. Tieke, *Langmuir* 15 (1999) 391-399
- [52] M. Dreja, S. Gramberg, B. Tieke, *Chem. Commun.* (1998) 1371-1372
- [53] A. Kumar, E. Alami, K. Holmberg, V. Serebyuk, F.M. Menger, *Colloids Surfaces, A* 228 (2003) 197-207
- [54] Kabir-ud-Din, P.A. Koya, *J. Chem. Eng. Data* 55 (2010) 1921-1929
- [55] M. Dreja, W. Pyckhout-Hintzen, B. Tieke, *Macromolecules* 31 (1998) 272-280
- [56] X. Jiang, L. Zhou, Y. Li, Z. Chen, X. Hu, *Langmuir* 23 (2007) 11404-11408
- [57] M.M. Khalaf, A.H. Tantawy, K.A. Soliman, H.M. Abd El-Lateef, *J. Mol. Struc.* 1203 (2020) 127442
- [58] M. Dreja, B. Tieke, *Langmuir* 14 (1998) 800-807
- [59] A.M. Asemu, N.G. Habtu, M.A. Delele, B. Subramanyam, S. Alavi, *J. Food Proc. Eng.* 43 (2020) 1-19
- [60] E. Setiawan, Mudasir, K. Wijaya, *IOP Conf. Ser.* 742 (2020)
- [61] R. Soleimani, A.H. Saeedi Dehaghani, N.A. Shoushtari, P. Yaghoubi, A. Bahadori, *Kor. J. Chem. Eng.* 35 (2018) 1556-1569
- [62] A. Golbraikh, A. Tropsha, *Mol. Divers.* 5 (2000) 231-243
- [63] R. Todeschini, D. Ballabio, F. Grisoni, *J. Chem. Inform. Model.* 56 (2016) 1905-1913
- [64] I.E. Frank, R. Todeschini, *The data analysis handbook*, Elsevier, Amsterdam, 1994
- [65] C. Cortes, V. Vapnik, *Machine Learning* 20 (1995) 273-297.

MAAMAR LAIDI
 ABDALLAH ABDALLAH
 EL HADJ
 CHERIF SI-MOUSSA

MODELOVANJE KRITIČNE KONCENTRACIJE
 MICELA RAZLIČITIH GEMINI SURFAKTANATA
 PRIMENOM HIBRIDNOG PRISTUPA KOJI
 KOMBINUJE REGRESIJU POTPORNIH VEKTORA

OTHMANE BENKORTEBI
MOHAMED HENTABLI
SALAH HANINI

Laboratory of Biomaterials and
Transport Phenomena (LBMPT),
University of Medea, Medea, Algeria

NAUČNI RAD

I ALGORITAM *DRAGONFLY*

Kako tehnika kvantitativne zavisnosti svojstva od strukture pruža pogodno sredstvo za predviđanje kritične koncentracije micela (CMC) gemini surfaktanata iz njihovih deskriptora strukture. U ovom radu, izveden je uporedni rad za modelovanje CMC svojstva 211 različitih gemini surfaktanata na osnovu njihovih strukturnih karakteristika koristeći linearne i nelinearne kvantitativne modele odnosa struktura-svojstvo. Za modelovanje CMC-a razvijeni su model najmanjih kvadrata i parcijalni najmanji kvadrati za regresijski model k-najbližih suseda, veštačku neuronsku mrežu i regresiju potpornih vektora (SVR). Molekularni deskriptori su izračunati i pregledani kako bi se uklonili neprikladni deskriptori i poboljšalo učenje. Rezultati ukazuju da su poboljšane performanse SVR kada se hiper-parametri optimizuju pomoću algoritma Dragonfly (SVR-DA) bile izuzetno sposobne za predviđanje vrednosti pCMC (-logCMC) sa prosečnim apsolutnim relativnim odstupanjem od 0,666 i koeficijentom determinacije (R^2) od 0,997 za globalni skup podataka.

Ključne reči: kvantitativni odnos struktura-svojstvo, tenzidi, kritična koncentracija micela, modeliranje, Mašinsko učenje.