People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

UNIVERSITY YAHIA FARES MEDEA

Faculty of Technology

Department of Process and Environmental Engineering

Biomaterial and Transport Phenomena Laboratory (LBMPT)

DOCTORAL THESIS

Presented by:

Miss. Nada BOUKELKAL

Master in Pharmaceutical Process Engineering at YAHIA FARES University of Medea.

To obtain the title of Doctorate in Pharmaceutical Engineering

Theme

Prediction and modeling of physicochemical properties of surfactants of pharmaceutical interest.

Defense publicly on 28/10/2024 against the jury composed of:

Pr. BENKORTBI Othmane	Professor at the University of MEDEA	President
Pr. LAIDI Maamar	Professor at the University of MEDEA	Examiner
Dr. ALI AGHA Hamza	Senior lecturer A at the University of MEDEA	Examiner
Dr. DANOUNI Samir	Senior lecturer A at Tipaza University Center	Examiner
Dr. RAHAL Soufiane	Senior lecturer A at the University of MEDEA	Supervisor
Dr. REBHI Redha	Senior lecturer A at the University of MEDEA	Co- Supervisor
Pr. HANINI Salah	Professor at the University of MEDEA	Guest

2024/2025

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ YAHIA FARÈS DE MÉDÉA

Faculté de Technologie

Département de Génie des Procédés et Environnement

Laboratoire de Biomatériaux et Phénomène de Transport (LBMPT)

THÈSE DE DOCTORAT

Présentée par :

M^lle^ Nada BOUKELKAL

Master en génie des procédés pharmaceutique : Université Yahia Fares de Médéa.

En vue de l'obtention du titre de Doctorat en génie pharmaceutique

Thème

Prédiction et modélisation des propriétés physico-chimiques des tensioactifs d'intérêt pharmaceutique.

Soutenue le 28/10/2024 devant le jury composé de :

Pr. BENKORTBI Othmane	Professeur, Université Yahia Fares de Médéa	Président
Pr. LAIDI Maamar	Professeur, Université Yahia Fares de Médéa	Examinateur
Dr. ALI AGHA Hamza	Maitre de Conférences A, Université Yahia Fares de Médéa	Examinateur
Dr. DANOUNI Samir	Maitre de Conférences A, centre universitaire de Tipaza	Examinateur
Dr. RAHAL Soufiane	Maitre de Conférences A, Université Yahia Fares de Médéa	Directeur de thèse
Dr. REBHI Redha	Maitre de Conférences A, Université Yahia Fares de Médéa	Co Directeur de thèse
Pr. HANINI Salah	Professeur, Université Yahia Fares de Médéa	Invité

2024/2025

Acknowledgments

First and foremost, I thank Allah for giving me the strength and the courage to accomplish this modest work

and for letting me live to see this thesis.

I'm deeply grateful to my supervisor, ***Dr. Soufiane RAHAL*** *& my co-supervisor* ***Dr. Redha REBHI****:*

for their coaching, availability, and sense of listening and exchange. and for providing guidance and feedback throughout this project.

A special thanks should be address to ***Pr. Mabrouk HAMADACHE*** *for the help and the influence that he could give.*

I would also like to extend my thanks to: the president of the jury, Professor Othmane BENKORTBI, and to the jury members, Professor Salah HANINI, Pr. Maamar LAIDI, Dr. Hamza ALI AGHA at the University of Medea, and Dr. Samir DANOUNI at Tipaza University Center, for having done us the honor and the pleasure of accepting and judging our work.

Thank you to everyone who helped me directly or indirectly during these years of research and study. I cannot conclude without a word of thanks to my teachers and colleagues.

May the Almighty Allah richly bless all of you.

Dedication

I have the immense pleasure of dedicating this modest work to:

Those I love most in the world, ***my dear parents*** *who always encourage me and push me towards success, for their support, and their countless sacrifices, may ALLAH keep and protect them.*

May this modest work be for you a small compensation and recognition for what you have done incredible for me

To my dear ***grandfathers*** *and my dear* ***grandmothers***

To all my ***BOUKELKAL & BOUKHALFA*** *family*

To all my friends

Everyone who knows Nada from near or far.

Nada BOUKELKAL

ملخص

التنبؤ ونمذجة الخواص الفيزيائية والكيميائية للمواد الخافضة للتوتر السطحي ذات الأهمية الصيدلانية.

تُعرف الأمفيفيلات، والتي تسمى أيضًا المواد الخافضة للتوتر السطحي، على نطاق واسع بقدرتها على تقليل التوتر السطحي وتستخدم بشكل شائع في المجالات الصناعية المختلفة بسبب خصائصها الفيزيائية والكيميائية المثيرة للاهتمام. كان الهدف من هذا العمل هو تطوير وتقييم نماذج QPSR التي تساهم في نمذجة خصائص الفاعل بالسطح باستخدام خوارزميات التعلم الآلي.

ركزت المساهمة الأولى على التنبؤ بلوغاريتم CMC لـ 119 من المواد الخافضة للتوتر السطحي القائمة على السكر من خلال تطبيق أربع تقنيات تعلم إحصائية، بما في ذلك الانحدار الخطي المتعدد، والمربعات الصغرى الجزئية، والشبكات العصبية الاصطناعية (ANN) ونظام الاستدلال العصبي الغامض التكيفي. تم استخدام مقاييس إحصائية مختلفة لتقييم موثوقية ومتانة النماذج. تم الحصول على أفضل النتائج($\Delta r_m^2 = 0.006, Q_{F1}^2 = 0.982, \bar{r}_m^2 = 0.803$) لـ ANN مع بنية {6-6-1}. بالإضافة إلى ذلك، تم اختيار 6 مواد خافضة للتوتر السطحي أخرى ذات أساس سكري كمجموعة خارجية لتقييم نموذج ANN المطور وأعطت النتيجة تنبؤًا جيدًا (R = 0.96).

ركزت المساهمة الثانية على التنبؤ ب pCMC لفئات مختلفة من المواد الخافضة للتوتر السطحي. كشف التحليل الإحصائي أن مجموعة من 14 واصفًا مهمًا لـ Mordred (SlogP، GATS6d، nAcid، GATS8dv، GATS4dv، PEOE_VSA11، GATS8d، ATS0p، GATS1d، MATS5p، GATS3d، NdssC، GATS6dv و EState_VSA4)، بالإضافة إلى درجة الحرارة، كانت بمثابة مدخلات مناسبة. تم فحص القدرة التنبؤية/الارتباطية والجدوى لأساليب التعلم الآلي المختلفة، مثل MLR ، والانحدار العشوائي للغابات (RFR) ، والشبكات العصبية الاصطناعية، والانحدار المتجهي الداعم المُحسَّن بواسطة نموذج خوارزمية اليعسوب (DA-SVR) . وفقًا للمعاملات الإحصائية، كان DA-SVR هو الأكثر دقة في التنبؤ بقيم pCMC، حيث وصل إلى ($R^2 = 0.9740$، $Q^2 = 0.9739$، $\bar{r}_m^2 = 0.9627$، و$\Delta r_m^2 = 0.0244$) لجميع البيانات.

الكلمات المفتاحية: المواد الخافضة للتوتر السطحي، CMC، QSPR، التعلم الآلي، ANN، ANFIS، DA-SVR، MLR، PLS

Résumé
Prédiction et modélisation des propriétés physico-chimiques des tensioactifs d'intérêt pharmaceutique.

Les amphiphiles, également appelés tensioactifs, sont largement connus pour leur aptitude à diminuer la tension superficielle et sont couramment employés dans différents domaines industriels en raison de leurs caractéristiques physico-chimiques intéressantes. L'objectif de ce travail était de développer et d'évaluer des modèles QPSR contribuant à la modélisation des propriétés des tensioactifs à l'aide des algorithmes d'apprentissage automatique.

La première contribution a porté sur la prédiction de la concentration micellaire critique (CMC) de 119 tensioactifs à base de sucre en appliquant quatre techniques d'apprentissage statistique, notamment la régression linéaire multiple, les moindres carrés partiels, les réseaux de neurones artificiels (ANN) et le système d'inférence neuro-flou adaptatif. Différentes mesures statistiques ont été utilisées pour évaluer la fiabilité et la robustesse des modèles. Les meilleurs résultats ($\bar{r}_m^2$= 0.803, Q_{F1}^2=0.982, and Δr_m^2= 0.006) ont été obtenus pour ANN avec une architecture {6-6-1}. De plus, 6 autres tensioactifs à base de sucre ont été sélectionnés comme ensemble externe pour évaluer le modèle ANN développé et le résultat a donné une bonne prédiction avec (R = 0,96).

La deuxième contribution a porté sur la prédiction de la concentration micellaire critique (pCMC) de différentes classes de tensioactifs. L'analyse statistique a révélé qu'un ensemble de 14 descripteurs Mordred significatifs (SlogP, GATS6d, nAcid, GATS8dv, GATS4dv, PEOE_VSA11, GATS8d, ATS0p, GATS1d, MATS5p, GATS3d, NdssC, GATS6dv et EState_VSA4), ainsi que la température, ont servi d'entrées appropriées. Différentes méthodes d'apprentissage automatique, telles que MLR, la régression forestière aléatoire (RFR), ANN et la regression à vecteurs de support optimisée par l'algorithme libellule (DA-SVR), modélisent leur capacité prédictive/corrélative et leur faisabilité ont été examinées. Selon les coefficients statistiques, DA-SVR était le plus précis pour prédire les valeurs de pCMC, atteignant (R^2=0.9740, Q^2=0.9739, $\bar{r}_m^2$=0.9627, and Δr_m^2=0.0244) pour l'ensemble des données.

Mots clés : tensioactif, CMC, RQSP, apprentissage automatique, RNA, ANFIS, DA-SVR, MLR, PLS

Abstract

Prediction and modeling of the physicochemical properties of surfactants of pharmaceutical interest.

Amphiphiles, also called surfactants, are widely known for their ability to reduce surface tension and are commonly used in different industrial fields due to their interesting physicochemical characteristics. This work aimed to develop and evaluate QSPR models that contribute to the modeling of surfactant properties using machine learning algorithms.

The first contribution focused on predicting the critical micelle concentration (CMC) of 119 sugar-based surfactants by applying four statistical learning techniques including Multiple linear regression, Partial least square, Artificial neural networks (ANN), and Adaptive neuro-fuzzy inference system. Different statistical metrics were employed to evaluate the reliability and robustness of the models. The best result ($\bar{r}_m^2$= 0.803, Q_{F1}^2=0.982, and Δr_m^2= 0.006) were obtained for ANN with {6-6-1} architecture. In addition, 6 other sugar-based surfactants were selected as external sets to assess the developed ANN model and the result gave good prediction with (R = 0.96).

The second contribution focused on predicting the critical micelle concentration (pCMC) of different classes of surfactants, including anionic, cationic, nonionic, zwitterionic, and Gemini surfactants. Statistical analysis revealed that a set of 14 significant Mordred descriptors (SlogP, GATS6d, nAcid, GATS8dv, GATS4dv, PEOE_VSA11, GATS8d, ATS0p, GATS1d, MATS5p, GATS3d, NdssC, GATS6dv and EState_VSA4), along with temperature, served as appropriate inputs. Different machine learning methods, such as MLR, random forest regression (RFR), ANN, and optimized Support Vector Regression by the dragonfly algorithm (DA-SVR) model their predictive/correlative capacity and feasibility have been examined. According to the statistical coefficients, DA-SVR was the most accurate in predicting pCMC values, achieving (R^2=0.9740, Q^2=0.9739, $\bar{r}_m^2$=0.9627, and Δr_m^2=0.0244) for the entire dataset.

Keywords: surfactants, CMC, QSPR, machine learning, ANN, ANFIS, DA-SVR, MLR, PLS

Table of Content

Table of Content i
List of Figures v
List of Table vii
Abbreviations viii
Nomenclature and Symbol x
General Introduction 1
Bibliographic References 4

Chapter I: Bibliography Element

I.1 Basics of surfactants 6
I.1.1 Introduction 6
I.1.2 Definition 6
I.1.3 Classification of surfactants 6
I.1.4 Characterization of surfactants 10
I.1.4.1 Adsorption at interfaces and lowering of surface energy 10
I.1.4.2 Micellization 13
I.1.4.2.1 Morphology of aggregate 14
I.1.4.3 Hydrophile-lipophile balance 15
I.1.4.4 Solubility 17
I.1.4.4.1 Krafft point 17
I.1.4.4.2 Cloud point 17
I.1.5 Surfactants applications 18
I.1.5.1 Detergent and cleaning agents 19
I.1.5.2 Coating products 19
I.1.5.3 Personal care products 20
I.1.5.4 Food industry 20
I.1.5.5 Pharmaceutical research 20
I.2 Chemoinformatic and programming tools 20
I.2.1 Introduction 20
I.2.2 Fundamental Concepts 21
I.2.2.1 Databases 21
I.2.2.2 SMILES 21
I.2.2.3 Molecular descriptors 22
I.2.3 Quantitative structure-activity/propriety Relationship 22

I.2.3.1 Property prediction (QSAR/QSPR) 22
I.2.3.2 Construction of QSAR/QSPR model 23
I.2.4 Multiple linear regression 25
I.2.5 Partial least square 26
I.2.6 Artificial neural network 26
I.2.6.1 Biological neuron Vs. Artificial neuron 27
I.2.6.2 Activation function 30
I.2.6.3 The concept of a feed-forward neuron network 31
I.2.6.4 Selection of an artificial neural network 32
I.2.6.4.1 Neuron type 33
I.2.6.4.2 Selection of network structure 33
I.2.6.4.3 Selection of activation functions 34
I.2.6.4.4 Training algorithm 34
I.2.7 Adaptive Adaptive Neuro-Fuzzy Inference System (ANFIS) 34
I.2.7.1 Fuzzy system 36
I.2.7.2 ANFIS architecture 37
I.2.8 Support vector machine (SVM) 40
I.2.8.1 Fundamental Concepts of Statistical Learning Theory 40
I.2.8.2 Support vector regression (SVR) 41
I.2.8.3 Kernal function 44
I.2.9 Dragonfly Optimization Technique 45
I.2.10 Random forest regression 48
I.3 Critical review on prediction of surfactant properties 50
I.3.1 Introduction 50
I.3.2 Critical review on CMC of surfactants 51
I.3.3 Critical review on the surface tension of surfactants 63
I.3.4 Critical review on HLB of surfactants 64
I.3.5 Critical review on cloud point of nonionic surfactants and Krafft point of anionic surfactants 67
I.3.6 Critical review on multi-property surfactants 69
I.4 Conclusion 70
Bibliographic References 72

Chapter II: Computational QSPR model to predict the CMC of classic and extended sugar-based surfactants

II.1 Introduction 89

II.2 Methodology 90
II.2.1 Data collection and dataset division 90
II.2.2 Molecular descriptor calculation 90
II.2.3 Molecular descriptor selection 91
II.2.4 QSPR model's development and validation 91
II.2.5 QSPR model's applicability domain (AD) 93
II.2.5.1 Applicability domain using leverage approach 93
II.2.5.2 Standardization approach 93
II.3 Results and discussion 94
II.3.1 Molecular descriptor selection 94
II.3.2 MLR model 94
II.3.3 PLS model 95
II.3.4 ANN model 96
II.3.5 ANFIS model 98
II.3.6 Comparison of four statistical models 99
II.3.7 Applicability domain 99
II.3.8 Comparison with previously reported models 100
II.4 Conclusion 103
Bibliographic References 104

Chapter III : QSPR for the prediction of CMC of different classes of surfactants using machine learning algorithms

III.1 Introduction 110
III.2 Methodology 111
III.2.1 Data collection 112
III.2.2 Molecular descriptors calculation 113
III.2.3 Selection of relevant descriptors 113
III.2.4 Model development 114
III.2.5 Model validation 115
III.3 Results and discussion 115
III.3.1 MLR predictive model 115
III.3.2 ANN predictive model 116
III.3.2.1 Mathematical equation application for the QSPR-ANN model 116
III.3.3 RFR predictive model 117
III.3.4 DA-SVR predictive model 117
III.3.5 Comparison of four statistical models 118

III.3.6 Applicability domain 119
III.3.7 Contribution of the selected descriptors 119
III.3.8 Comparison with previously reported models 120
III.4 Conclusion 122
Bibliographic References 123
General Conclusion and Outlook 128
Annex 130

List of Figures

Figure	Figures Name	Page
	Chapter I	
Figure I.1	Simplified Schematic of a surfactant	6
Figure I.2	Chemical structure of polyoxyethylene glycol dodecyl ether	7
Figure I.3	Chemical structure of sodium lauryl sulfate	7
Figure I.4	Chemical structure of cetylpyridinium chloride	7
Figure I.5	Chemical structure of lauryl sultaine	8
Figure I.6	Simplified Schematic of Gemini surfactant	9
Figure I.7	Examples of natural surfactants	9
Figure I.8	Representation of the forces experienced by a water molecule within the liquid	11
Figure I.9	Evolution of surface tension as a function of concentration, and the behavior of surfactants in aqueous solution.	12
Figure I.10	Evolution of a solution's physicochemical parameters in function of surfactant concentration	13
Figure I.11	Definition of the terms used in the packing parameter equation	14
Figure I.12	Aggregate characteristics as a function of the packing parameter	15
Figure I.13	Phase diagram of the water-ionic surfactant system	17
Figure I.14	Phase diagram of the water-nonionic surfactant system	18
Figure I.15	Some common areas of surfactant applications	19
Figure I.16	The basic approach to QSAR/QSPR	23
Figure I.17	Schematic representation of QSAR/QSPR model development	24
Figure I.18	Diagram of a biological neuron	27
Figure I.19	The basic operation of an artificial neuron	28
Figure I.20	A simple example of an artificial neural network	28
Figure I.21	Designing biological and artificial neurons	29
Figure I.22	A multilayer feed-forward neural network structure	31
Figure I.23	Fuzzy system structure	36
Figure I.24	ANFIS model architecture	38
Figure I.25	Schematic structure of the TSK fuzzy model	38
Figure I.26	Regression with the ε-insensitive tube	43
Figure I.27	Effect of kernel transformation	44
Figure I.28	Dynamic dragonfly swarms	45
Figure I.29	Static dragonfly swarms	45
Figure I.30	Primitive corrective patterns between individuals in a swarm	46
Figure I.31	Illustration of random forest regression	49
Figure I.32	Number of publications per year with the keywords “surfactant” and “formulation” during the period 1960-2021	50
Figure I.33	Concept map of publications related to the keyword search “surfactant” and “formulation”	51
	Chapter II	
Figure II.1	Scatter plot of the predicted values of pCMC versus the experimental values by ANN model for the training, validation, and test set	96
Figure II.2	Plot of the fraction contribution of the descriptors to the pCMC of sugar-based surfactants	97

Figure II.3	Projection of the training, validation, and test set of sugar-based surfactants in the Williams plot	100
	Chapter III	
Figure III.1	General steps for the generation of QSPR models in the current study	112
Figure III.2	Normality distribution plot of CMC data: (a) before transformation, (b) after log transformation (pCMC)	113
Figure III.3	Scatter plots Predicted vs. Observed values of pCMC for the DA-SVR for training and test sets	117
Figure III.4	William's plot for the DA-SVR model.	119
Figure III.5	The results of the relevance factor analysis carried out on the DA-SVR model	120

List of Table

Table	Table Name	Page
	Chapter I	
Table I.1	Main properties characterizing the adsorption, aggregation of surfactants as well as the solubility of their aggregates	10
Table I.2	Surfactants properties/application based on the HLB parameter	15
Table I.3	The group numbers obtained by the Davies	16
Table I.4	Common transfer functions	30
	Chapter II	
Table II.1	Acceptance criterion (A.C) of a model	93
Table II.2	Validation parameters for MLR model's train, validation, test, and global set	95
Table II.3	Validation parameters for the PLS model's train, validation, test, and global set	95
Table II.4	Validation parameters for ANN model's train, validation, test, and global set	97
Table II.5	Validation parameters for ANFIS model's train, test, and global set	98
Table II.6	Results of our best model's (ANN) internal and external validation compared with those from previous research	101
Table II.7	Observed value of pCMC and those predicted by ANN-simulate for 6 sugar surfactants.	102
	Chapter III	
Table III.1	Characteristics of the selected descriptors by S-MLR and their VIF values	114
Table III.2	Statistical criteria of all developed models for the test set	118
Table III.3	Statistical criteria of all developed models for the global set	118
Table III.4	Performance of our DA-SVR model with previously published models for internal and external validations	121
Table III.5	Observed value of pCMC and those calculated by Eq. (III.5) and DA-SVR for 11 surfactants.	122

Abbreviations

Abbreviation	Signification
AAE	Average absolute error
AARD	Average absolute relative deviation
AD	Applicability domain
AI	Artificial intelligence
AIC	Akaike Information Criterion
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural network
ARD	Average relative difference
BMLR	Best multicollinearity regression
CAGR	Compound annual growth rate
CCC	Concordance correlation coefficient
CMC	Critical micelle concentration
CODESSA	Comprehensive descriptors for structural and statistical analysis
COSMO-RS	Conductor screenlike model for real solvents
DA-SVM	Dragonfly-support vector machine
DBMS	Database management system
EO	Polyoxyethylene
ETA	Extended topochemical atom
FIS	Fuzzy inference system
GA	Genetic algorithm
GA-MLR	Genetic algorithm- Multiple linear regression
GC	Group contribution
GCNs	Graph convolutional neural networks
GFA	Genetic function approximation
GP	Gaussian process
HLB	Hydrophile-lipophile balance
HQSAR	Hologram structure-activity relationship
LSSVM	Least squares support vector machine
MAE	Mean absolute error
MFs	Membership function
MLP	Multi layer perceptrons.
MLP-ANN	multilayer perceptron–artificial neural network
MLR	Multiple linear regression
OECD	Organization for Economic Co-operation and Development
PCA	Principal component analysis
pH	potential of hydrogen
PLS	Partial least square
PO	Polyoxypropylene
QSAR	Quantitative structure activity relationship
QSPR	Quantitative structure property relationship

RBF	Radial Basis Function
RBFNN	Radial basis function neural network
RF	Random forest
RFE	Recursive feature elimination
RFR	Random forest regression
RMSE	Root mean square error
RSD	Relative standard deviation
SMILES	Simplified Molecular Input Line Entry Specification
S-MLR	Stepwise-multiple linear regression
SVM	Support vector machine
SVR	Support vector regression
T_C	Cloud point
T_K	Krafft point
VIF	Variance inflation factors
WNN	wavelet neural network

Nomenclature and Symbol

Symbol	Designation	Unity
ΔH_f	Heat of formation	$J.mol^{-1}$
2D	2 Dimensions	--
3D	3 Dimensions	--
b	bias	--
C	Concentration	$mol.L^{-1}$
E_{HOMO}	The energy of the highest occupied molecular orbital	eV
E_{LUMO}	The energy of the lowest unoccupied molecular orbital	eV
f	fittest function	--
K	Kernal function	--
n	Number of compounds in the data set	--
p	Packing parameter	--
R	Universal ideal gas constant	
R	Coefficient of corrélation	--
R^2	Coefficient of détermination	--
T	Temperature	°C
w_i	weight	--
Γ	Maximum surface excess	$mol.m^{-2}$
γ	Surface tension	$J.m^{-2}$

General Introduction

General Introduction

Throughout time, humans have always felt the need to understand and explore the phenomena of our world, which led to the development of chemistry and physical sciences. Surface chemistry studies phenomena that occur on the surfaces or interfaces of materials, including adsorption, heterogeneous catalysis, colloidal formation, corrosion, crystallization, and solubility [1]. Surface chemistry has been used by humans for a long time, through soap and cleaning. Then it manifested itself in various scientific fields, which led to the creation and industrial production of surfactants, not to mention its applications in industry as well as in daily life.

The worldwide surfactants market is expected to grow at a compound annual growth rate (CAGR) of 4.9% between 2021 and 2028, from $41.22 billion in 2021 to 57.81 billion in 2028. Surfactants are produced at a rate of around 17 million metric tons per year [2].

Surfactants encompass a wide range of chemicals that play critical roles in many fields, and their effect is directly related to their chemical structure. All surfactants have an asymmetric structure with hydrophobic and hydrophilic parts. Therefore, they are soluble in organic solvents and water.

Understanding the fundamental properties of surfactants, such as micellization, adsorption behavior, and interfacial properties, is crucial for designing processes and products and creating correlation and prediction models [3].

However, critical micelle concentration (CMC) is a key parameter for both surface chemistry and chemists and systematic data collection and daily progress in this field are essential. The CMC is defined as the concentration of surfactants upon which micelles form and all surfactants added to the system are directed toward the micelles. Hence CMC is an important surface property. Measuring them using traditional experimental methods is sometimes time-consuming and expensive. In this paper, we have tried to solve these problems by developing mathematical methods to estimate this property.

Molecular modeling offers the possibility of representing structures in two or three dimensions, but it also allows any theoretical chemist to carry out in-depth and precise studies of molecular systems and to express them using mathematical models [4]. Nowadays, computational methods of artificial intelligence are widely used in various sectors for modeling purposes. These methods allow the creation of reliable models that can accurately predict

multiple properties and activities related to physical, chemical, and biological phenomena [5,6]. Quantitative structure-activity/property relationships (QSAR/QSPR) are particularly relevant predictive methods. For a long time, QSAR/QSPR models were mainly designed for biological activities [7,8], such as surfactant toxicity [9]. However, they are increasingly used in the physicochemical characteristics of molecules [10,11], such as the amphiphilic properties of surfactants [12]. QSAR/QSPR models can predict specific properties based only on their molecular structure. Relationships between molecular structures, and their properties are usually established using statistical learning modeling techniques.

The main objective of this thesis is to develop mathematical models from structural information incorporated in surfactant molecules to predict CMC based on QSPR concepts with computational intelligence techniques.

Several steps were used during the development of the QSPR models, the main ones being: (1) the collection and processing of the database, (2) the calculation and selection of molecular descriptors, and (3) the development and selection of the best developed models. The efficiency and precision of these models were tested using several statistical parameters such as the coefficient of determination (R^2), and the Root Mean Square Error (RMSE). The models were also validated using the Applicability Domain method.

This thesis is mainly structured around three chapters:

The first chapter is subdivided into three parts. The first part includes a bibliographic overview of surfactants such as the definition, classification, and characteristics of surfactants as well as their areas of use. The second part is devoted to the study of modeling methods, in particular the concept of (QSPR/QSAR), and artificial intelligence techniques based on machine learning for linear and non-linear modeling. The methods described are multiple linear regression (MLR), partial least square (PLS), artificial neural network (ANN), adaptive neuron fuzzy inference system (ANFIS), random forest regression (RFR), support vector machine (SVM), and also the dragonfly optimization algorithm. The third part is devoted to the critical review of the modeling of the physicochemical parameters of surfactants.

The second chapter presents the results of contributions made by the four computational intelligence techniques, MLR, PLS, ANN, and ANFIS, for the estimation of the CMC of sugar-based surfactants followed by a comparative study between the proposed models and other models from the literature. In addition, a sensitivity study of the input parameters and application domain for the best proposed model is developed.

The third chapter presents the results of contributions made for the modeling of the CMC of the different classes of surfactants (anionic, cationic, nonionic, zwitterionic, and Gemini) at temperatures ranging from 10°C to 60°C using the predictive MLR, ANN, RF, and SVR optimized by the dragonfly algorithm (DA-SVR) are presented, as well as a comparison between the proposed models and other models in the literature, and a study of the sensitivity of the input parameters and the applicability domain for best model.

The study ends with a general conclusion summarizing the main results obtained throughout the work of this thesis as well as some perspectives which can be considered as future work which will be carried out in this line of research.

Bibliographic References

[1] C. Sur, C. Surface, Unit 5 Objectives, 16 (2015).

[2] V.S. Nagtode, C. Cardoza, H.K.A. Yasin, S.N. Mali, S.M. Tambe, P. Roy, K. Singh, A. Goel, P.D. Amin, B.R. Thorat, J.N. Cruz, A.P. Pratap, Green Surfactants (Biosurfactants): A Petroleum-Free Substitute for Sustainability─Comparison, Applications, Market, and Future Prospects, ACS Omega. 8 (2023) 11674–11699. https://doi.org/10.1021/acsomega.3c00591.

[3] L.L. Schramm, S.M. Kutay, Emulsions and Foams in the Petroleum Industry, Surfactants. (2010) 79–118. https://doi.org/10.1017/cbo9780511524844.004.

[4] J. Gasteiger, T.E. Eds, Related Titles from WILEY-VCH Molecular Modeling Neural Networks in Chemistry and Drug Design 150 and More Basic, 2003.

[5] W. Duch, K. Swaminathan, J. Meller, Artificial Intelligence Approaches for Rational Drug Design and Discovery, Curr. Pharm. Des. 13 (2007) 1497–1508. https://doi.org/10.2174/138161207780765954.

[6] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature. 521 (2015) 452–459. https://doi.org/10.1038/nature14541.

[7] H. Gao, J.A. Katzenellenbogen, R. Garg, C. Hansch, Comparative QSAR Analysis of Estrogen Receptor Ligands, Chem. Rev. 99 (1999) 723–744. https://doi.org/10.1021/cr980018g.

[8] E. Lo Piparo, A. Worth, Review of QSAR Models and Software Tools for predicting Developmental and Reproductive Toxicity, 2010. https://doi.org/10.2788/9628.

[9] D.W. Roberts, QSAR issues in aquatic toxicity of surfactants, Sci. Total Environ. 109–110 (1991) 557–568. https://doi.org/10.1016/0048-9697(91)90209-W.

[10] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, ChemInform Abstract: Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction, ChemInform. 42 (2011) 5714–5789. https://doi.org/10.1002/chin.201120272.

[11] C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.C. De Hemptinne, P. Ungerer, B. Rousseau, C. Adamo, A General Guidebook for the Theoretical Prediction

of Physicochemical Properties of Chemicals for Regulatory Purposes, Chem. Rev. 115 (2015) 13093–13164. https://doi.org/10.1021/ACS.CHEMREV.5B00215.

[12] J. Hu, X. Zhang, Z. Wang, A review on progress in QSPR studies for surfactants, Int. J. Mol. Sci. 11 (2010) 1020–1047. https://doi.org/10.3390/ijms11031020.

Chapter I
Bibliography Element

Chapter I Bibliography Element

I.1 Basics of surfactants

I.1.1 Introduction

Advances in chemistry and industrial production processes in the 1920s and 1930s led to a significant increase in surfactants and detergents, increasing the number of valuable materials.

Their numerous applications in pharmaceuticals, para pharmaceutics, and crude oil recovery highlighted the importance of understanding surfactants and their properties.

I.1.2 Definition

Surfactants, abbreviated as surface-active agents, are active organic chemicals at the surface. When introduced between two immiscible phases, they modify their interfacial properties.

The term "amphiphiles" is frequently used synonymously with "surfactants." The term "amphi," which means "both," comes from Greek. These compounds have two parts of different polarity. The hydrophobic groups, represented by a hydrocarbon chain, can be soluble in organic solvents. On the other hand, the hydrophilic groups are soluble in water [1] (see Figure I.1).

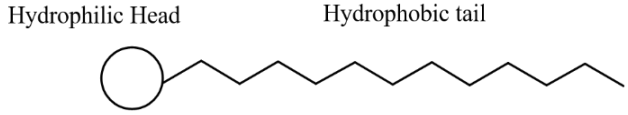


Figure I.1. Simplified Schematic of a surfactant.

I.1.3 Classification of surfactants

Surfactant compounds are classified into several classes:

- The first classification is based on their solubility. Hydrophilic surfactants possess the ability to dissolve in water, whereas lipophilic surfactants dissolve in lipids. In general, ionic surfactants are hydrophilic, but nonionic surfactants can have either a hydrophilic or lipophilic characteristic depending on the composition of their groups. Consequently, the solubility of nonionic surfactants is determined by the hydrophilic-lipophilic ratio [2].
- Surfactants are often commercially classed based on their intended functions. However, this is not very advantageous because many surfactants have several uses, which might confuse [3].

- The most frequently accepted and scientifically adapted classification of surfactants is based on their hydrophilic head charge. Charge variation at the hydrophilic head provides surfactant molecules with a very diverse range of properties [3–5]. We identify the following families:

Nonionic surfactants

Nonionic surfactants are one of the most frequently utilized surfactants in industry. Non-ionic surfactants are distinguished by a lack of charge at their hydrophilic head. They are typically low toxicity and stable in hard water [6]. This group comprises ethoxylated alcohols, alkylphenols, fatty acid esters, and more. Due to their mildness, nonionic surfactants are widely employed in the home and personal care, as well as the agrochemical industry.

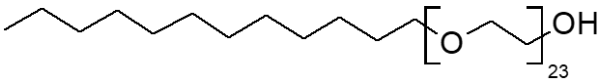


Figure I.2. Chemical structure of polyoxyethylene glycol dodecyl ether.

Anionic surfactants

Anionic surfactants are distinguished by a negatively charged ionic group, such as carboxylate, sulfate, or sulfonate. They are more used due to their low manufacturing cost. They are therefore used in the formulation of detergents. However, they have been shown to irritate the skin [7].

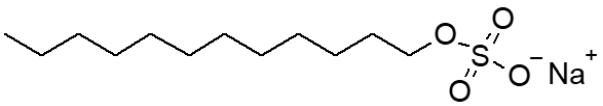


Figure I.3. Chemical structure of sodium lauryl sulfate.

Cationic surfactants

These molecules have a positively charged hydrophilic head and a halide counter ion. It mainly includes ammonium salts. They have a range of purposes, including antistatic properties, antimicrobial effects, and fabric softening [6]. However, due to their synthesis difficulty, cationic surfactants are generally more expensive than anionic surfactants.

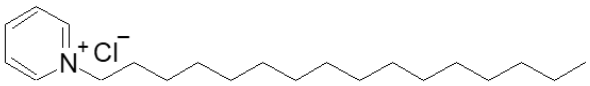


Figure I.4. Chemical structure of cetylpyridinium chloride.

Amphoteric surfactants

Amphoteric surfactants are known as zwitterionic molecules that have both negative and positive charges in their hydrophilic group. These surfactants are soluble in water and combine well with other surfactants to create micelles. Amphoteric surfactants' charge changes with pH alter their wetting, detergent, and foaming properties [8]. Amphoteric surfactants have qualities similar to those of non-ionic. Zwitterionic surfactants possess excellent dermatological qualities, as well as demonstrating little eye irritation, and are commonly used in personal care products such as shampoos and are currently being employed in agrochemical formulation. This category includes, among others, betaines or amino acids. However, these surfactants are costly to produce.

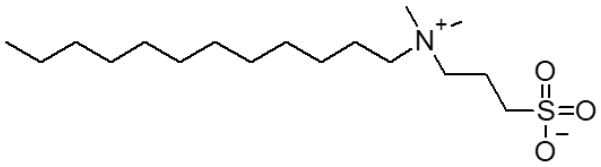


Figure I.5. Chemical structure of lauryl sultaine.

The hydrophobic part comes from petrochemicals, plants, and animals [9]. It can have a variety of structures, including linear, branching, unsaturated, and saturated hydrocarbon chains. The number of carbon atoms, unsaturation, and branching all affect this part's hydrophobicity and solubility [10]. Indeed, more carbon atoms increase hydrophobicity, whereas more branches decrease it. A length of 8 to 16 carbon atoms is ideal for properties like cleaning, wetting, adsorption, and emulsification. Linear hydrophobic chains are also favored for improving these compounds' biodegradability [9].

There are other types of surfactants besides those which have just been named, among which:

Gemini surfactants

Surfactants possess valuable properties and are extensively utilized. A new generation of surfactants is emerging as a result of chemists improving these properties by modifying their molecular structures.

In 1991, Menger and Littau identified a novel class of surfactants called Gemini surfactants. As illustrated in Figure I.6, it consists of more than one hydrophilic group connected by a spacer [11–15]. This spacer group increases the Gemini surfactants' hydrophobicity compared to the monomeric surfactants [12,16]. The hydrophilic head can

display a range of charges from cationic, anionic, neutral, and zwitterionic, while the hydrophobic tail can vary in length. Moreover, the spacer group might be polar or nonpolar, and rigid or flexible [12,17].

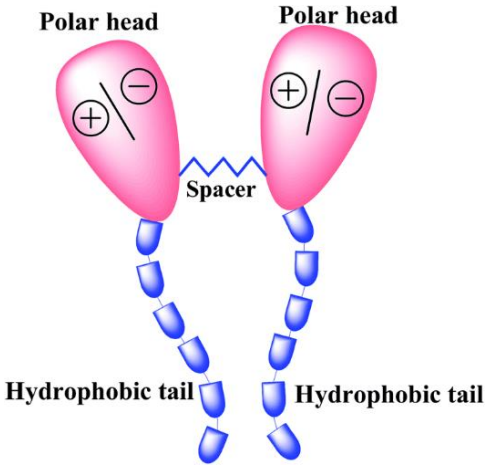


Figure I.6. Simplified Schematic of Gemini surfactant.

Gemini surfactants have been more heavily focused on for many years for a variety of purposes, including genetics, cosmetics, paint, textile industries, and especially oil recovery [18], due to their low CMC [19], high solubilization capacity [20,21], unique aggregation behavior, and micelle formation [22], intriguing rheological properties [23,24], high surface activity [25], and lower Krafft temperature [13] when compared to conventional surfactants.

Natural surfactants

Natural surfactants can have either their hydrophobic or hydrophilic parts derived from natural sources [26]. These surfactants can be generated directly from biomass, known as biosurfactants, produced by microorganisms [27], or synthesized from biomass-derived compounds. Figure I.7 presents some examples of natural surfactants.

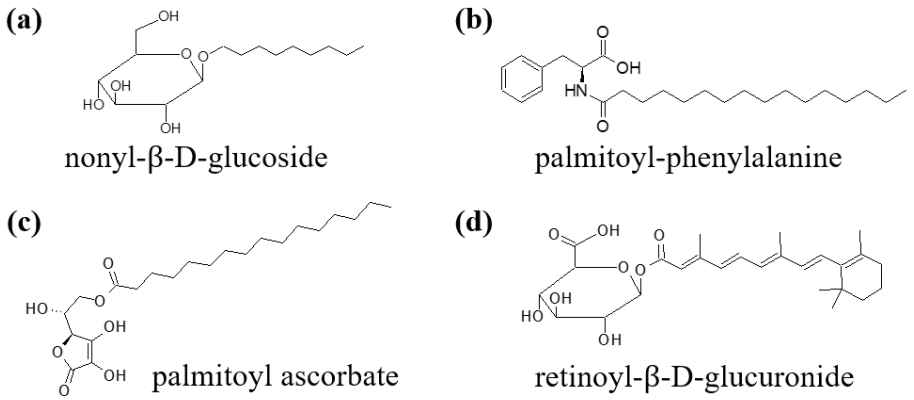


Figure I.7. Examples of natural surfactants.

Surfactants (a)-(c) are synthesized from various molecules derived from biomass, while the surfactant (d) is a biosurfactant, derived from microorganisms.

More precisely, the surfactant octyl α-D glucoside (Figure I.7., a) is generated from sugar. The most common bio-based surfactant family is sugar surfactants. Sugar-based surfactants have sugar as their polar head and can be obtained from plant starch [28]. Many types of sugar exist, such as glucose [29], maltose [30], and sucrose [31]. The nonpolar parts are derived from fats or fatty alcohols.

A natural surfactant can be termed "green" or "eco-friendly" due to their low toxicity, high biodegradability, and environmental compatibility, with relative stability at extreme pH or temperature [28,32,33]. These surfactants have significant potential in the pharmaceutical and cosmetic industries [34].

I.1.4 Characterization of surfactants

As previously stated, surfactants often comprise hydrophilic and lipophilic parts that contribute to their particular properties. These molecules can exhibit a variety of intriguing behaviors due to their affinity variations. Some properties are directly linked to adsorption at interfaces, some to self-assembly, and others to the solubility of the aggregates formed by self-assembly. Table I.1 summarizes the amphiphilic characteristics that characterize the diverse behaviors.

Tableau I.1 Main properties characterizing the adsorption, aggregation of surfactants as well as the solubility of their aggregates.

symbol	property	associated behavior
CMC	critical micelle concentration	self-assembly
$\boldsymbol{\gamma_{CMC}}$	surface tension at CMC	adsorption/self-assembly
$\boldsymbol{pC_{20}}$	efficiency	Adsorption
$\boldsymbol{\Gamma_{20}}$	maximum surface excess	Adsorption
$\boldsymbol{A_{20}}$	minimum surface area per surfactant at the interface	Adsorption
$\boldsymbol{T_K}$	Krafft point	solubility of aggregates
$\boldsymbol{T_C}$	Cloud point	solubility of aggregates

I.1.4.1 Adsorption at interfaces and lowering of surface energy

The main characteristic of surfactant molecules in solution is their ability to adsorb to interfaces which reduces surface energy [35].

In the liquid phase, each molecule is influenced by attractive intermolecular forces that play a crucial role in enhancing the overall cohesion of the liquid. The force acting on the molecules in the core of liquid is isotropic and equal in all directions since they are surrounded by identical surrounding molecules. Thus, the molecules in the liquid are in equilibrium and the forces are zero. However, molecules at surfaces and interfaces are not equally connected by molecules of the same kind, the forces acting on them are unbalanced, and the resulting force is non-zero. Molecules on surfaces and interfaces are consequently drawn to the liquid's core. It causes a spontaneous contraction of the liquid's surface or interface, reducing the liquid molecules' contact area with air or another fluid. This behavior is illustrated in Figure I.8 [4].

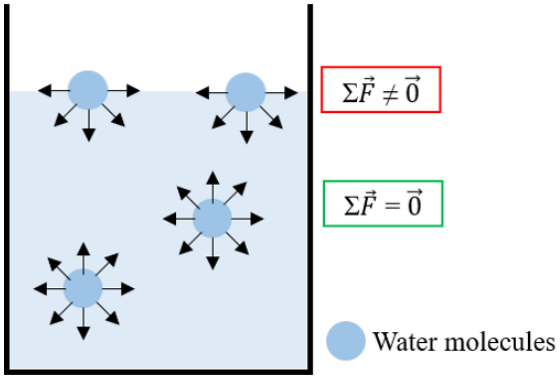


Figure I.8. Representation of the forces experienced by a water molecule within the liquid [4].

From these considerations, there is free energy noted γ, expressed in N/m and defined as the minimal work (W_{min}) required to increase the surface area or interface (ΔA) of a unit (Eq I.1). Surface free energy can also be described by the force applied perpendicularly to the surface or interface per unit length, known as surface tension for an air/water surface or interfacial tension for an interface between two immiscible liquids. The surface energy is important when the phases in contact are very diverse [5,35].

$$W_{min} = \gamma \Delta A \qquad \text{I.1}$$

Where,

W_{min}: work [J].

γ : surface free energy [J.m^{-2}].

ΔA : increase in surface area [m^2].

When surfactant molecules dissolve and in a low-concentration aqueous medium, amphiphilic molecules adsorb in monomeric form and replace water molecules (Figure I.9).

The hydrophilic part will move toward the aqueous solution, while the lipophilic part will move toward the second phase to minimize contact. Due to their structure, the surfactant will have a significant affinity with one of the two phases and will contribute to reducing the surface free energy through favorable interactions. According to Gibbs' law (Eq I.2), the decrease in surface tension depends on the nature and concentration of the surfactant molecules and the temperature. The following law applies to non-ionic surfactants in diluted media [4,36].

$$d\gamma = -RT\Gamma dlnC \qquad \text{I.2}$$

Where,

γ : Surface tension [N.m-1].

R : Universal ideal gas constant [8.314 J.mol^{-1}K^{-1}].

T : Temperature [K].

Γ : Excess surface area [mol.m^{-2}].

C : Concentration of surfactants [mol.L^{-1}].

When the concentration of surfactant molecules in an aqueous medium rises, this leads to an accumulation of molecules on the surface, thereby reducing surface tension until the water/air surface becomes saturated, leading to constant surface tension.

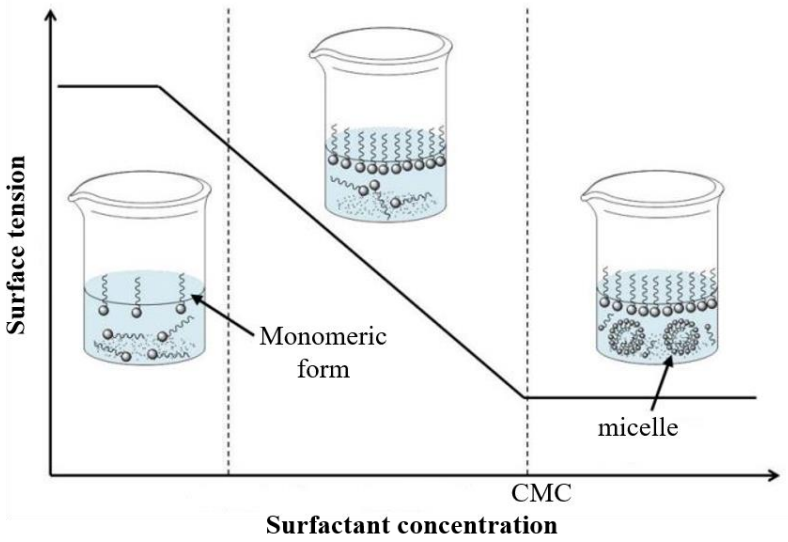


Figure I.9. Evolution of surface tension as a function of concentration, and the behavior of surfactants in aqueous solution.

I.1.4.2 Micellization

Once the surface is fully saturated, the surfactant begins to form aggregates in the solution, known as micelles, at a concentration referred to as the "critical micelle concentration" (CMC) [37]. We note that aggregates develop when the hydrophobic component of the monomers, which cannot establish hydrogen bonds in the aqueous phase, causes a significant rise in the system's free energy. One method for reducing this energy is to isolate the nonpolar fraction by adsorbing it on an organic matrix or building aggregates with hydrophobic regions in the middle of the solution [38].

Several physicochemical parameters of the solution, such as conductivity, turbidity, surface tension, and more, change suddenly at the CMC [39], as seen in Figure I.10.

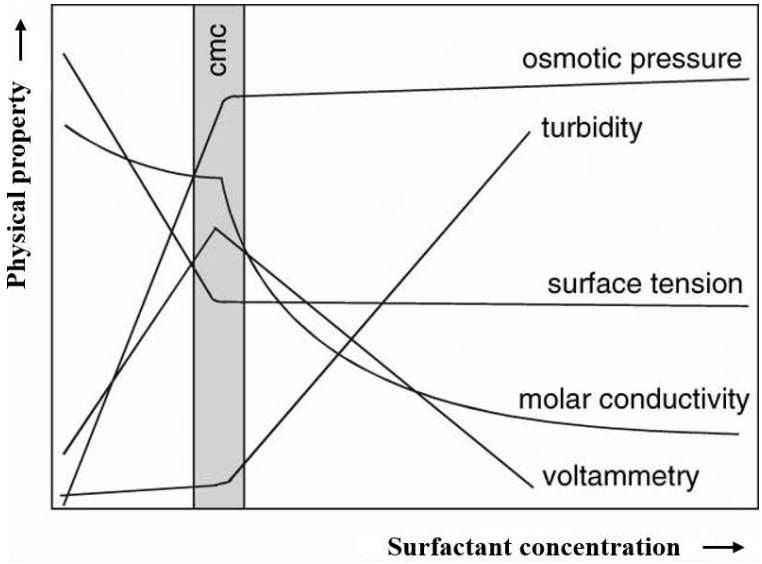


Figure I.10. Evolution of a solution's physicochemical parameters in function of surfactant concentration [39].

According to Figure I.10, relevant measures may be utilized to calculate the CMC of a surfactant. The CMC value at the curve's break point is obtained by graphing the physicochemical characteristic versus surfactant concentration.

Furthermore, the CMC of surfactant is influenced by several factors, including the molecular structure of the surfactant [40], solution salinity [41], pressure conditions [42], temperature, pH, ionic composition, and others [43].

I.1.4.2.1 Morphology of aggregate

The molecular packing parameter (Eq I.3) is a concept that represents the link between the surfactant geometry and its aggregate structure in aqueous solutions, as follows [44]:

$$p = \frac{v_0}{I_0 a} \qquad \text{I.3}$$

Where,

p : packing parameter

v_0 : the volume of the surfactant tail [Å³]

I_0 : the length of the surfactant tail [Å²]

a : the surface area of the hydrophilic headgroup at the aggregate surface [Å].

Figure I.11 Displays the terms involved in the packing parameter equation (I.3).

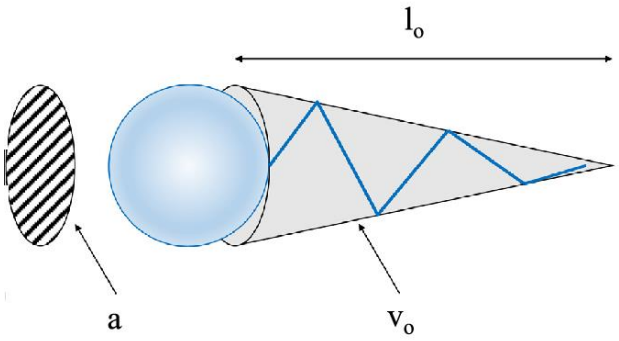


Figure I.11. Definition of the terms used in the packing parameter equation [45,46].

The surfactant molecules self-assemble into various aggregate shapes dependent on the packing parameters, as illustrated in Figure I.12 These shapes influence the solution's properties, including its viscoelastic behavior [47] and solubilization capacity [44].

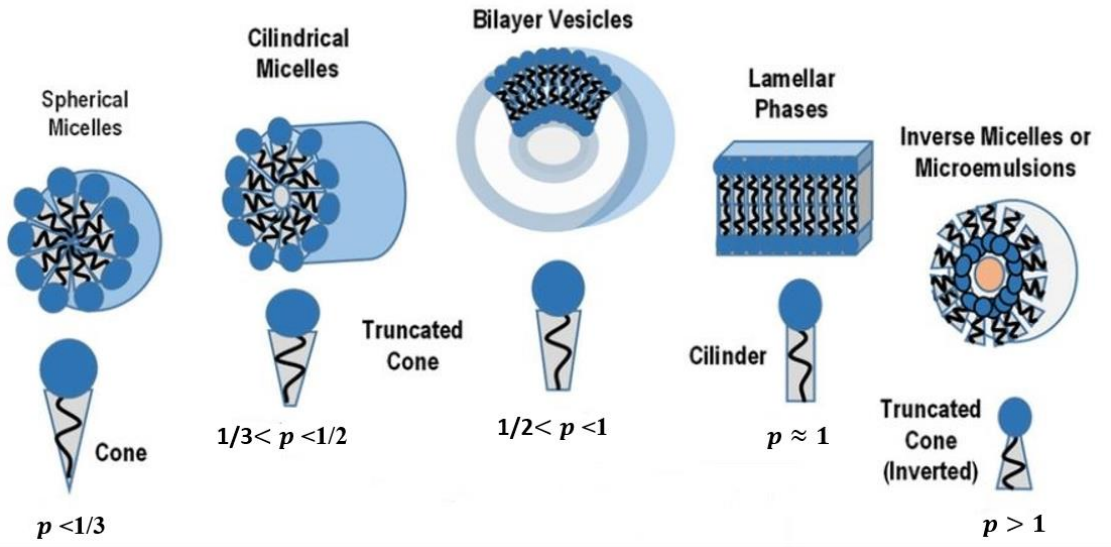


Figure I.12. Aggregate characteristics as a function of the packing parameter [48].

I.1.4.3 Hydrophile-lipophile balance

The hydrophilic-lipophilic balance (HLB) is essential for determining a surfactant's ability to dissolve in water or oil. It is represented by a number ranging from 0 to 20. A value of 0 represents a completely hydrophobic molecule, whereas a value of 20 represents a molecule made entirely of hydrophilic components [46,49,50]. Table I.2 highlights the HLB values associated with the surfactants' application properties. It can be calculated using empirical equations from surfactant molecules.

Tableau I.2 Surfactants properties/application based on the HLB parameter [46].

HLB value	Property/application
0-3	Antifoaming agent
4-6	W/O emulsifier
7-9	Wetting agent
8-18	O/W emulsifier
13-15	Detergent
10-18	Hydrotrope or solubilize

Griffin suggested in 1949 that the HLB parameter of nonionic surfactants be computed using the following formula [51]:

$$HLB = 20 * \frac{M_h}{M} \qquad \text{I.4}$$

Where,

M_h: molar mass of the hydrophilic part [g.mol^{-1}].

M: molar mass of the surfactant [g.mol^{-1}].

Another equation established by Davies in 1957 [52] for determining the HLB values of nonionic and charged surfactants based on the functional group, as seen in (Eq I.5), and a set of group numbers is presented in Table I.3 provided by Davies (1957).

$$HLB = 7 + \sum (hydrophilic\ group\ numbers) - \sum (hydrophobic\ group\ numbers) \quad \text{I.5}$$

Tableau I.3 The group numbers obtained by the Davies [52].

Hydrophilic groups	**value**
–$\mathbf{SO_4^-Na^+}$	38.7
–$\mathbf{COO^-K^+}$	21.1
–$\mathbf{COO^-Na^+}$	19.1
N (tertiary amine)	9.4
Ester (sorbitan ring)	6.8
Ester (free)	2.4
–**COOH**	2.1
Hydroxyl (free)	1.9
–**O**–	1.3
Hydroxyl (sorbitan ring)	0.5
Lipophilic groups	
–CH–	-0.475
–$\mathbf{CH_2}$–	-0.475
$\mathbf{CH_3}$–	-0.475
=CH–	-0.475
Derived groups	
–($\mathbf{CH_2CH_2O}$)–	0.33
–($\mathbf{CH_2CH_2CH_2O}$)–	-0.15

The enhancement made to the equations for calculating HLB has partially made it possible to consider certain structural considerations, but now the QSPR approach appears to be more reliable due to the diversity and complexity of new surfactant molecules [51].

I.1.4.4 Solubility

knowing surfactant solubility helps us select the best items for certain applications and establish the conditions under which they should be used.

At a critical temperature, two phenomena were observed: cloud point and Krafft point. These are specific to each surfactant and can be observed on the surfactant/water phase diagram.

I.1.4.4.1 Krafft point

Ionic surfactants are distinguished by their Krafft point, also known as the critical micelle temperature, at which solubility equals CMC [53]. At temperatures lower than Krafft surfactant cannot form micelles [54] and it tends to be insoluble [55]. However, at the Krafft temperature, the surfactant's solubility increases significantly, and a further temperature rise allows the formation of micelles once the CMC is reached [56]. Furthermore, the Krafft point corresponds to a triple point in the phase diagram of the binary water-ionic surfactant, as shown in Figure I.13. In many applications, knowledge of this parameter enables selecting a surfactant having a Krafft point lower than its temperature of use. Generally, the Krafft point for the same polar head increases as the length of the alkyl chain increases [57].

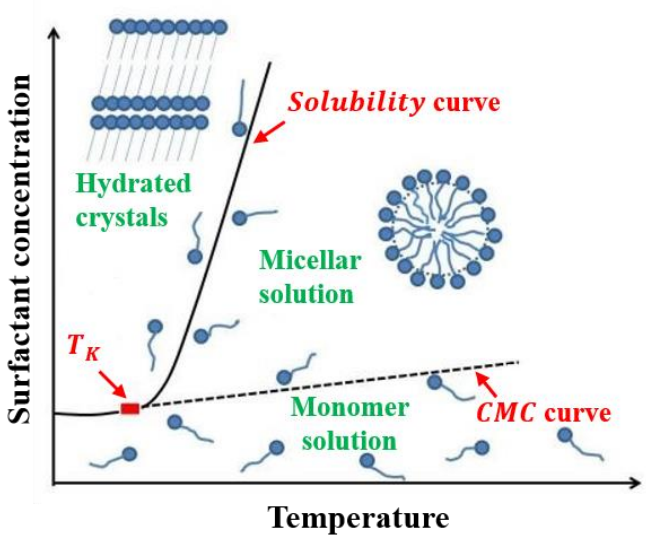


Figure I.13. Phase diagram of the water-ionic surfactant system [4].

I.1.4.4.2 Cloud point

The cloud point is the temperature at which aqueous solution of nonionic surfactants, especially polyethoxylated type, becomes cloudy, due to dehydration of the hydrophilic part of the surfactant molecules. Figure I.14 represents a simplified phase diagram of an aqueous

solution containing ethoxylated nonionic surfactants. A cloudy biphasic solution with a surfactant-rich phase and a surfactant-poor phase is present above T_C. The presence of a clear micellar solution is observed below T_C. it should be noted that at higher concentrations, a liquid crystal phase may appear in the form of a cubic, hexagonal, or lamellar phase [4].

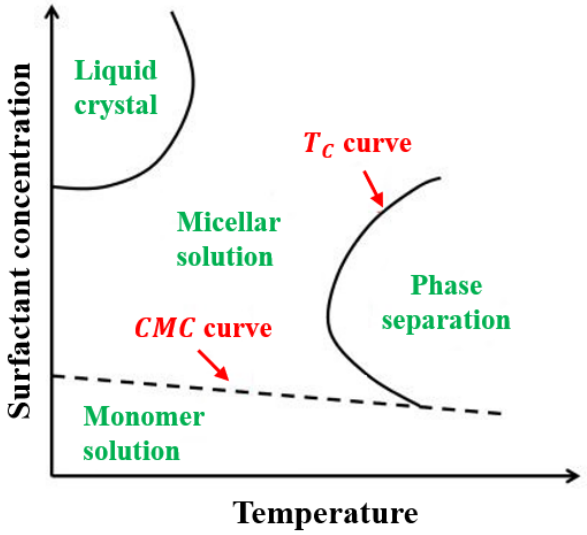


Figure I.14. Phase diagram of the water-nonionic surfactant system [4].

Literature shows that sugar-based surfactants such as alkyl polyglucoside and sugar amides have both a cloud point and a Kraft point [4].

I.1.5 Surfactants applications

Surfactants have amphiphilic characteristics and can interact at different interfaces to stabilize mixtures with different polarities. They are often utilized in scientific research to industrial manufacturing. They are referred to by many depending on their application, such as detergents, wetting agents, dispersion agents, and so on. They have become an essential part of our daily lives in every aspect. Figure I.15 illustrates some of the most common applications of surfactants.

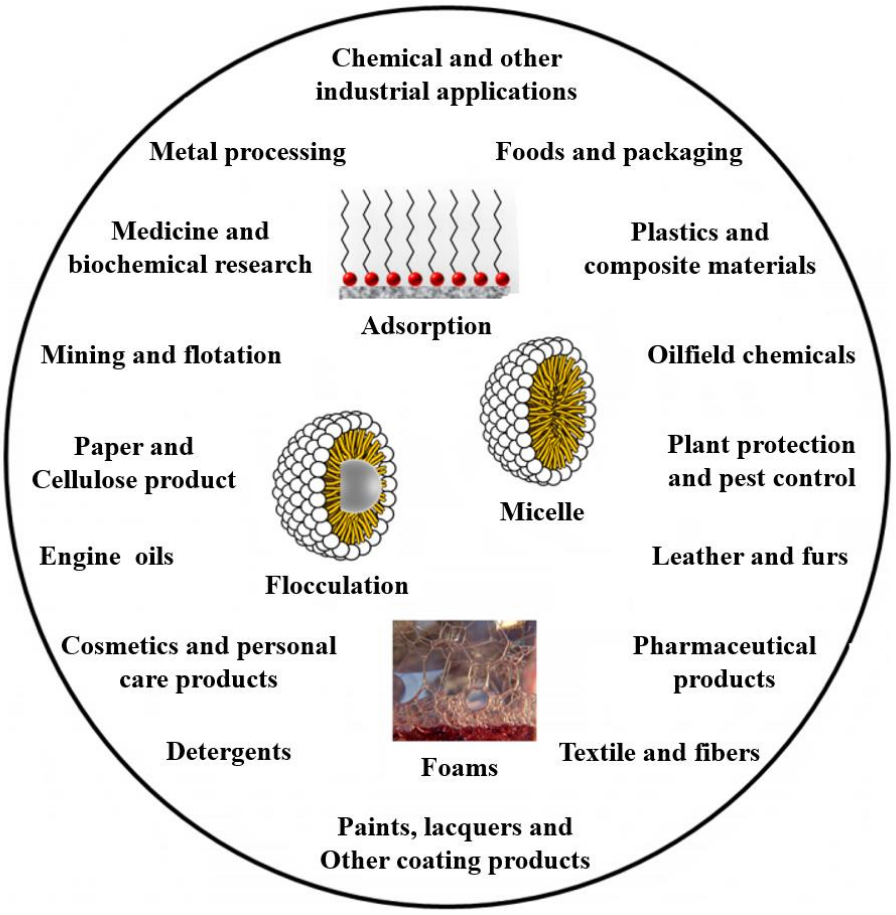


Figure I.15. Some common areas of surfactant applications.

I.1.5.1 Detergent and cleaning agents

One common use for surfactants is as detergents or cleaning agents. The ability of the surfactant to wet hydrophobic substances (undesirable dirt) and enclose them in surfactant micelles is what ensures their cleaning effect. Anionic surfactants, like alkyl benzene sulfonates, alkyl sulfate, alkyl ether sulfate, and soaps, are the most commonly used types in this field [58].

I.1.5.2 Coating products

Surfactants play a crucial role in paint production processes. They help to grind the pigments in paint by reducing the energy of the particles. The barriers that prevent the aggregation of pigments are also created by the adsorption of surfactant molecules on the surface of the particles. Surfactants also play a crucial role in stabilizing pigment particles in the final paint product, allowing the paint to be stored for a longer time [59].

I.1.5.3 Personal care products

Surfactant is a major element in many personal care products, including shampoo, hand cream, and facial cleaning. One reason is that such products are always a combination of diverse chemicals with varying polarities, and surfactants improve their mixability and stability. Surfactants can also assist active substances in penetrating the skin layer [60].

I.1.5.4 Food industry

Surfactants are used in many food manufacturing processes, including cholesterol extraction, oil solubilization, liquor emulsification, component separation prevention, and nutrient solubility [60].

I.1.5.5 Pharmaceutical research

Surfactants have a high potential as carriers of drugs or targeting systems due to their distinctive phase structure in different conditions. They can form micelles in aqueous solutions, which helps to encapsulate drugs that are not easily soluble in water. This allows to creation of stable solutions for intravenous or oral administration [61]. Researchers are interested in utilizing surfactants for drug targeting. Some studies suggest that using surfactant micelle as micro-containers can improve neuroleptic targeting from blood flow into the brain [62]. Nonionic surfactant-based vesicles can also be used to target anti-cancer drugs in tumor tissues, taking advantage of the unique blood vessel structure in these tissues [63,64].

I.2 Chemoinformatic and programming tools

I.2.1 Introduction

Since the invention of the term Informatik by Professor Karl Steinbush in a publishing paper entitled Informatik: Automatische Informationsverarbeitung ("Informatics: Automatic Information Processing"), researchers have noted the emergence of computer science in various fields including artificial intelligence, data mining, data science, and more.

Chemoinformatics is a scientific discipline that applies computer science tools to solve chemical problems. The concept of Chemoinformatics had its literary debut by F.K. Brown in 1998 [65].

“Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization”

Machine learning is an area of artificial intelligence that focuses on creating techniques and procedures that allow computers to learn and execute tasks. It entails the creation of algorithms that will enable machines to do jobs and activities, and it has been linked to statistics in many ways over time [66].

This section focuses only on the general exposure of the informatics bases, models, and algorithms associated with the theme.

I.2.2 Fundamental Concepts

I.2.2.1 Databases

A database, also known as an electronic database or dataset, is a collection of information that is organized for easy access, search, management, and updates. Typically, a database management system (DBMS) controls it [67].

The choice of the initial experimental database is a critical step for developing QSAR/QSPR models. To ensure quality, a database should consist of reliable experimental data. This is because the uncertainties in these data will propagate into the final modal.

Most statistical methods assume that the observed values follow a normal distribution. So, it's important to check the normality of the distribution of measured values. There are various statistical tests available for this purpose, a more straightforward approach is to visually evaluate the distribution through a histogram representation of the data, allowing this characteristic to be evaluated [68].

I.2.2.2 SMILES

The initial step in formalizing chemistry involves naming a chemical compound. we need an unambiguous and reproducible notation, whether for the simplest atom or a complex structure.

SMILES stands for Simplified Molecular Input Line Entry Specification, which is a chemical notation designed for modern chemical information processing [68]. It’s Based on principles of molecular graph theory [69,70].

For more details about the coding rules of the SMILES notation, you can consult the work of Weininger and his collaborators [69,71,72].

I.2.2.3 Molecular descriptors

A molecular descriptor is the outcome of a systematic experiment or a logical and mathematical process that converts chemical information encoded in a symbolic representation of a molecule into a usable number [73]. More precisely, they are quantitative representations of a molecule's physical, chemical, or topological properties that summarize our knowledge and understanding of molecular structure and activity from several perspectives [74].

Molecular descriptors play an important role in studies on the relationship between quantitative structure-activity/property. We use them as independent variables to predict a dependent variable (activity/property).

The use of molecular descriptors in QSAR/QSPR models is a difficult task, as a large number of molecular descriptors, complexities, and diverse conceptions presented in recent years. Despite this, no strict rules have been established for selecting suitable descriptors, often based on researchers' chemistry intuition [68].

I.2.3 Quantitative structure-activity/propriety Relationship

QSAR/ QSPR methods are based on the assumption that the structure of a chemical compound is related to its activity or property. A QSAR/ QSPR is, by definition, a mathematical model that combines one or more quantitative parameters from a chemical structure with a quantitative measure of a biological activity or chemical property.

The mathematical expression obtained can serve as a prediction of the activity or property studied for new molecules or molecules for which experimental data are not available. This can be represented as depicted in equation I.6 [68].

$$activity\ /\ property = f(molecular\ descriptors\ \dots) \qquad \text{I.6}$$

I.2.3.1 Property prediction (QSAR/QSPR)

The relationships between various chemicals and, especially, the biological data of the compounds and their structure are too complex to directly predict using the first principles, necessitating an indirect, two-step process (Figure I.16) [75].

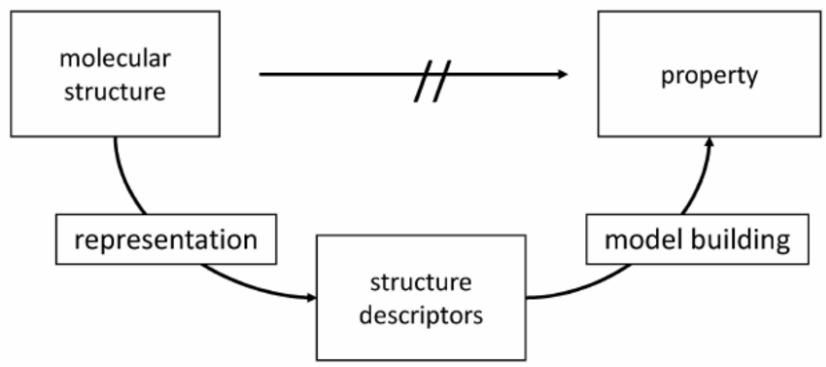


Figure I.16. The basic approach to QSAR/QSPR.

The first step is to represent a molecule structure by structural descriptors. The second stage involves analyzing and creating models based on structural descriptors and characteristics [75].

I.2.3.2 Construction of QSAR/QSPR model

The quality of the QSAR/QSPR model can be directly associated to [76]:

- The data used for its construction. Hence, before building the model, insights into the data must be gained.
- A comprehensive understanding of the problem and influencing factors helps identify relevant relationships.
- Relevant background data concerning the system studied, whether biological or chemical, must be collected through a literature search.
- Data sets for the model built should be selected carefully, as poor and inconsistent data could result in the creation of a corrupt model.
- Other elements such as data division into training and test datasets, molecular descriptors, and statistical methods for model development, impact model quality.

The construction of the QSAR/QSPR model is illustrated schematically in Figure 1.17. The following chapter briefly outlines the different steps that are involved in developing the QSAR/QSPR model.

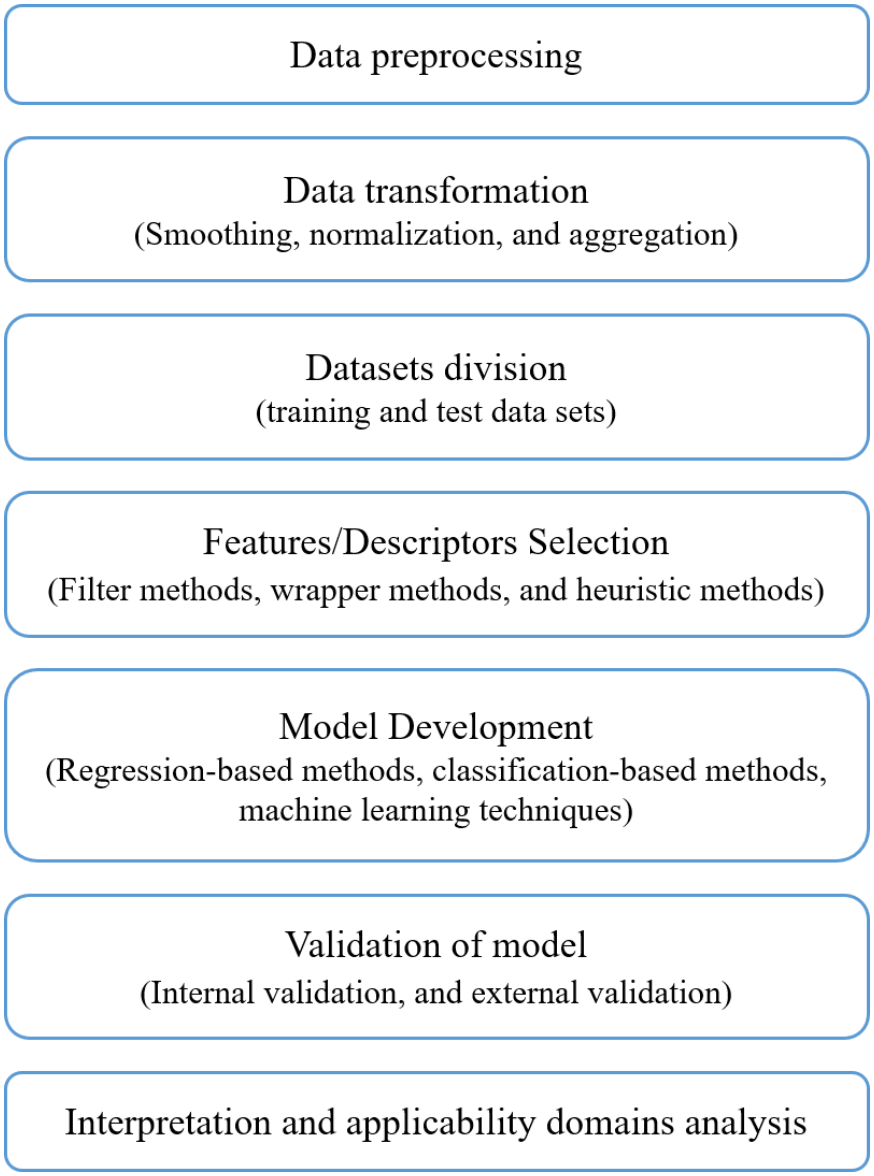


Figure I.17. Schematic representation of QSAR/QSPR model development [76].

The major challenge of QSARs is to choose the group of descriptors that describe the most essential structural and physicochemical characteristics related to the activity. The QSAR modeling process mainly relies on the quality of biological data and the effective selection of descriptors or variables. Implementing QSAR models relies on their statistical significance and ability to predict [77].

In order to validate and accept the QSAR/QSPR models, the Organization for Economic Co-operation and Development (OECD) has adopted five guidelines, which encompass the following criteria [78,79]:

i. **a defined endpoint:** The endpoint (activity/property) must be reliable and follow a stated experimental process.
ii. **an unambiguous algorithm:** The model must be based on an algorithm transparency that predicts an endpoint from information on chemical structure and/or physicochemical properties.
iii. **a defined domain of applicability:** Confident predictions from QSAR/QSPR models can only be made within a certain set of conditions, called the applicability domain, which defines the limits and scope of the model.
iv. **appropriate measures of goodness-of-fit, robustness, and predictivity:** This concept expresses the importance of statistical validation for model accuracy, encompassing both internal validation for the training set (fit and robustness) and external validation for the test set (predictability).
v. **a mechanistic interpretation, if possible:** to ensure the potential mechanistic association between model descriptors and predicted endpoints, and to ensure that this relationship is recorded.

These guidelines are now called the OECD principles for the validation of QSARs. Thus, several groups of researchers have expressed strong interest in developing more appropriate validation measures for developing accurate and predictive QSAR models [79].

I.2.4 Multiple linear regression

Due to its simplicity of use and ease of interpretation, multiple linear regression MLR is one of the most commonly used modeling methods in the development of predictive models.

The MLR method is based on the hypothesis that the property (Y) is linearly influenced by the different variables (the descriptors). The mathematical definition of this approach is given by equation (I.7).

$$Y = b_0 + \sum_{i=1}^{N} X_i b_i \qquad \text{I.7}$$

Where,

Y : is the dependent variable

X_i : represents the explanatory variable (descriptors)

b_i : represents the coefficient of those descriptors

b_0 : is the intercept of the equation.

The degree of impact of the corresponding molecular descriptors on the target activity/property is indicated by the size of these coefficients. A positive coefficient suggests a positive contribution of the molecular descriptor corresponding to the target, while a negative coefficient indicates a negative contribution.

I.2.5 Partial least square

Partial least squares (PLS) regression is an extension of multiple linear regression. It can be used when the number of descriptors is high and they are highly correlated [80,81]. This process combines the principles of Principal Component Analysis (PCA) and multilinear regression. The method makes it possible to determine, through a linear transformation, the axes that best reflect the data in space. To put it another way, this approach will make it possible to determine the axes that best explain the dispersion of the cloud of points. In the case where the data is represented according to n descriptors, PLS will therefore make it possible to find a maximum of n axes classified according to the variance they express. This involves substituting a matrix of predictive data X, which includes n rows and m columns, with a new matrix derived from k much weaker than m. The columns of the T matrix must also be linear combinations of the original variables. In matrix terms, the relationship is represented by Eq I.8 [68].

$$T = XW \qquad \text{I.8}$$

Where,

W $(m * k)$ is the matrix of coefficients defining the linear combinations;

T is the new matrix whose columns form “artificial variables”, obtained by a linear combination of the original variables;

After this transformation, multiple linear regression is applied on the table T instead of X.

I.2.6 Artificial neural network

Artificial neural networks (ANN) are the most advanced technology in the last two decades, which has been widely employed in many applications in different fields [82,83]. ANNs are computer programs based on parallel information processing methods that attempt to simulate some aspects of the functioning of the human brain, notably learning, storage of empirical knowledge, and making it available for use.

I.2.6.1 Biological neuron Vs. Artificial neuron

The notion of artificial neural networks was developed based on the way biological neurons process information. This concept is used to create massively parallel process simulation software that involves processing elements linked in a network architecture. Three major functional components of a biological neuron are its dendrites, cell body, and axon (Figure I.18).

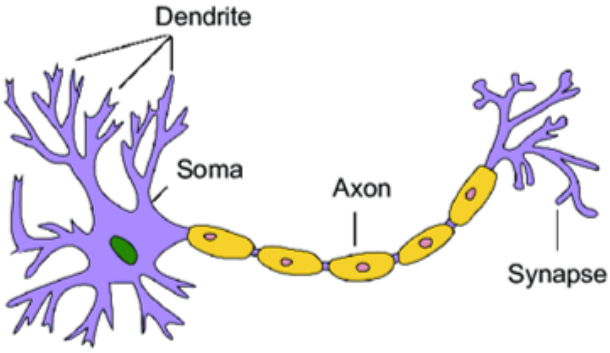


Figure I.18. Diagram of a biological neuron [84].

The cell body or soma has a nucleus which is the information processing center. the branching of the cell body forms the dendrites. Through the dendrites, information is transmitted from the outside to the soma, the body of the neuron. The information processed by the neuron travels along the (single) axon to be transmitted to other neurons. The synapse is the connection between two neurons [85]. The flow of information between two biological neurons is influenced by varying synaptic resistance. Dendrites collect, via synapses, electrochemical signals or action potentials coming from other cells or the outside, then transfer them to the cell body or soma. In addition to its role in cellular metabolism, the soma groups and concentrates information, making it a so-called “spatiotemporal” summation [86].

In Figure I.19, the artificial neuron receives inputs identical to the electrochemical pulses that the dendrites of biological neurons receive from other neurons. It is therefore a mathematical model that tries to translate the structure and function of biological brain networks.

This model comprises three simple sets of rules: multiplication, summation, and activation. The inputs to the artificial neuron are weighted, which implies that each input value is multiplied by its specific weight. The sum function in the middle of an artificial neural network

sums all weighted inputs and biases. At the exit of the artificial neuron, the total of previously weighted inputs and bias passes via the activation function, also known as the transfer function.

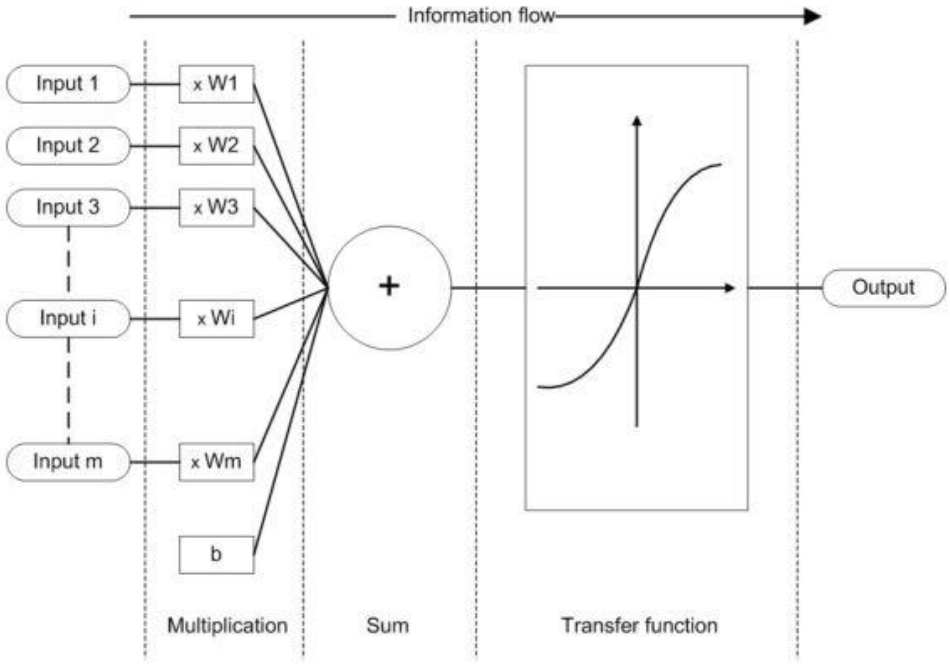


Figure I.19. The basic operation of an artificial neuron [83].

Despite the ordinary appearance of the basic operation and set of rules of artificial neurons, the models' full potential and computing power become apparent when they are interconnected into ANNs (Figure I.20). These ANNs simply make use of the notion that complexity may evolve from a few simple, fundamental rules [83].

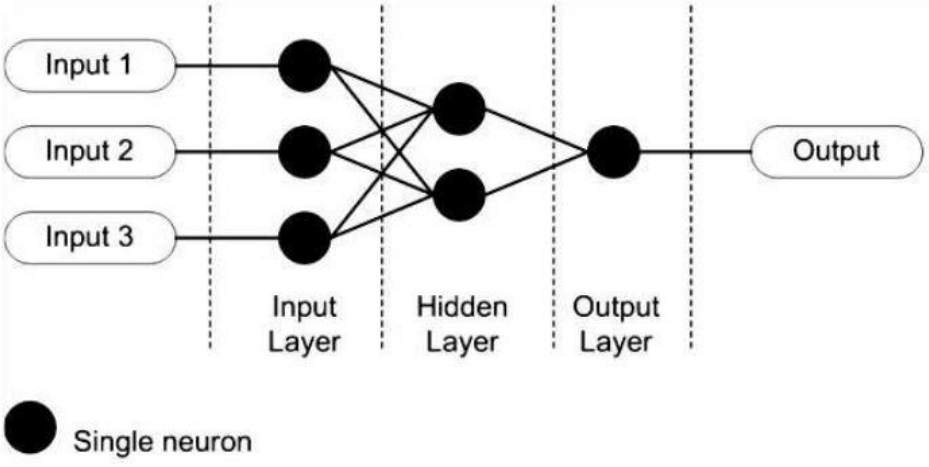


Figure I.20. A simple example of an artificial neural network [83].

When observing Figure I.20, it is essential to note that for the simpler mathematical manipulation and description of ANN, neurons are individually grouped into couches. Each

network contains an input layer, an output layer, and a variable number of hidden layers. A neuron is associated with all the neurons of the next layer.

The relationship between the artificial neuron and the biological neuron is summarized in Figure I.21. We can see in this schematic description that each element of the artificial neuron finds its inspiration from its biological nature.

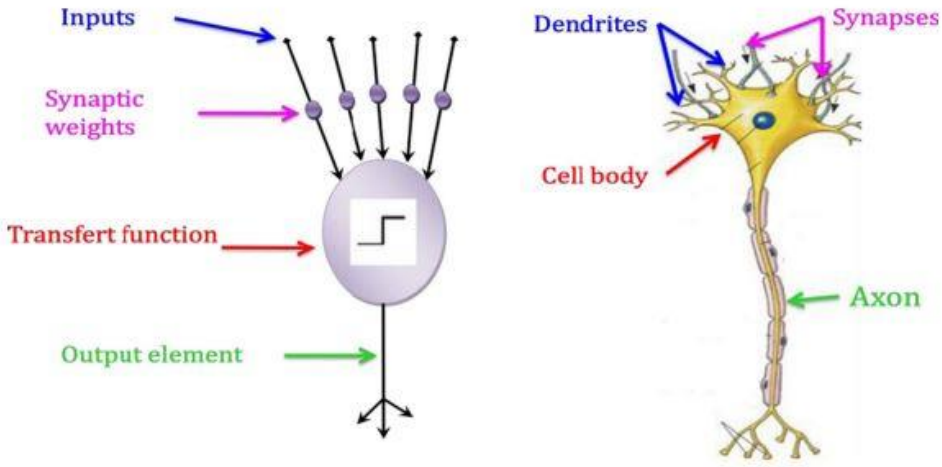


Figure I.21. Designing biological and artificial neurons [87].

The mathematical explanation below demonstrates the advantage of the simplicity of the artificial neuron model [83].

$$y(k) = F\left(\sum_{i=0}^{m} w_i(k) \cdot x_i(k) + b\right) \qquad \text{I.9}$$

Where,

$x_i(k)$: is input value in discrete time k where i goes from 0 to m,

$w_i(k)$: is weight value in discrete time k where i goes from 0 to m,

b : is bias,

F is a transfer function,

$y(k)$: is output value in discrete time k .

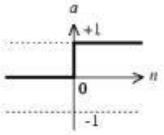
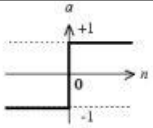
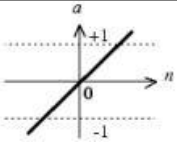
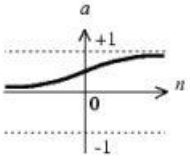
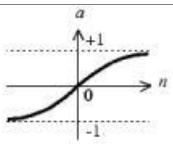
As seen by an artificial neuron model and its equation (Eq I.9), The transfer function is the main unknown in our model.

I.2.6.2 Activation function

The transfer function converts input signals into output signals. It is crucial for artificial neuron networks and any mathematical model as it determines their essential properties. The transfer function is chosen as the basis for the artificial neuron problem, which must be solved in most cases [83]. Unlike biological neurons whose state is binary, most transfer functions are continuous, offering an infinity of possible values within the range [0, +1] or [-1, +1]. There are many possible forms for the transfer function. Table I.4 summarizes some transfer functions used in the MATLAB® technical computing program.

Furthermore, since current optimization algorithms require the continuity and differentiability of activation functions at all points, the common activation functions that meet these requirements are the linear function, the hyperbolic tangent function, and the logarithmic sigmoid function. Linear functions are more suitable for output neurons, while sigmoid functions are more suited for hidden layer neurons [86]

Table I.4 Common transfer functions [86].

Function name	Input/output relationship	Graph of the function	Corresponding MATLAB function
Step (threshold)	$a = 0$ if $n < 0$ $a = 1$ if $n \geq 0$		***hardlim***
Symmetrical threshold	$a = -1$ if $n < 0$ $a = 1$ if $n \geq 0$		***hardlims***
Linear	$a = n$		***purelin***
logarithmic sigmoid	$a = \frac{1}{1 + e^{-n}}$		***logsig***
hyperbolic tangent	$a = \frac{e^{+n} - e^{-n}}{e^{+n} + e^{-n}}$		***tansig***

I.2.6.3 The concept of a feed-forward neuron network

Feed-forward Artificial Neural Networks are non-recurrent networks in which signals must only flow from input to output in one direction [83]. The artificial neural are arranged in layers (Figure I.22), where the input layer receives input signals (u_i) from the real world and each successive layer receives the weighted outputs ($w_{ij}\ u_i$) of the previous layer as its input thus resulting in a feedforward type ANN (case of multilayer perceptrons), in which each output of a layer is fed forward to the succeeding layer where it is processed. The outputs of the last layer constitute the outputs of the real world [86].

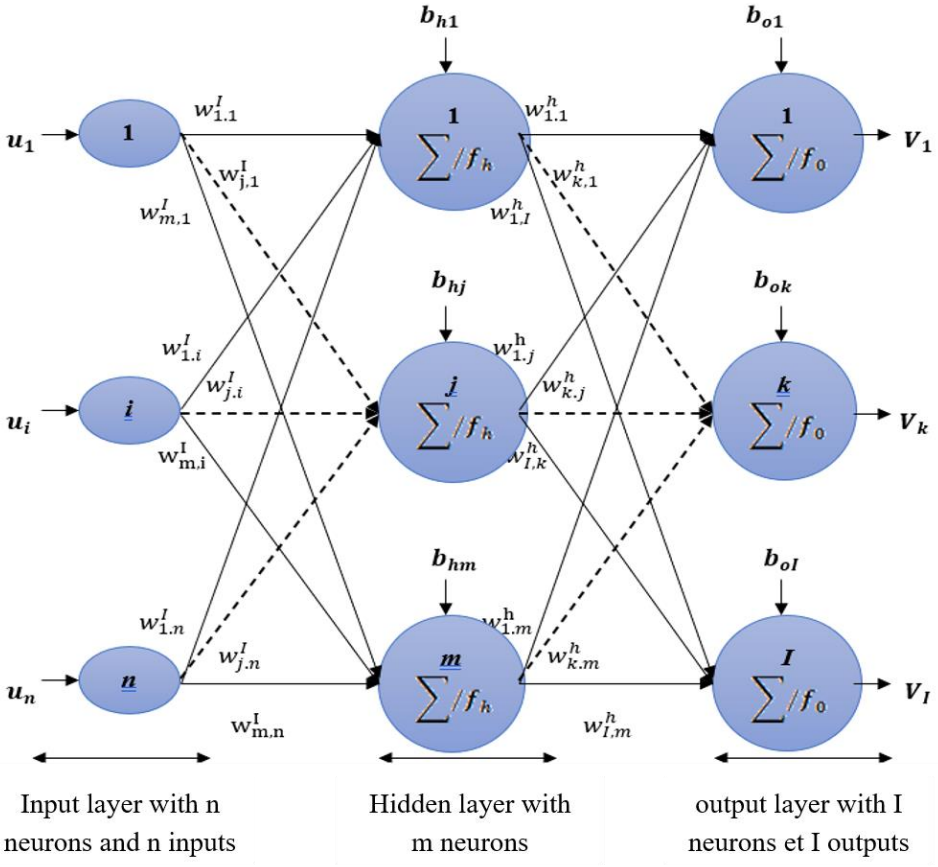


Figure I.22. A multilayer feed-forward neural network structure [88].

In such a feedforward type ANN, a neuron, whether in a hidden layer or an output layer, has two tasks:

- adds the bias to the weighted inputs of several connections, and then an activation function is applied to this sum, as given in equation I.10 (for neuron j in the hidden layer of Figure I.22):

$$z_j = f_h\left(\sum_{i=1}^{n} w_{ji}^{I}\, u_i + b_{hj}\right); \quad j = 1, 2, \ldots, m \qquad \text{I.10}$$

- The resulting value is propagated by the outgoing connections to the neurons of the next layer, where it is subjected to the same process in equation I.11 (for example the outputs of the hidden layer fed to neuron k of the output layer give the output):

$$V_k = f_0\left(\sum_{i=1}^{m} w_{kj}^{h}\, z_j + b_{ok}\right); \quad k = 1, 2, \ldots, I \qquad \text{I.11}$$

The relationship between the output and the inputs of the ANN is obtained by combining equations I.10 and I.11 :

$$V_k = f_0\left(\sum_{i=1}^{m} w_{kj}^{h}\, f_h\left(\sum_{i=1}^{n} w_{ji}^{I}\, u_i + b_{hj}\right) + b_{ok}\right); \quad k = 1, 2, \ldots, I \qquad \text{I.12}$$

The number of input and output variables determines the number of neurons in the input and output layers. Depending on the complexity of the problem, the modeler defines the number of neurons in the hidden layer. The model is developed using a learning process where an experimental dataset of independent variables is presented to the input layer of the network. The outputs of the output layer include a prediction of the dependent variables. The network learns the relationships between independent and dependent variables by iteratively comparing the calculated outputs and experimental outputs and adjusting the weight matrix and bias vector of each layer by a back-propagation training algorithm. Thus, the network develops a neural model capable of predicting with acceptable precision the output variables included in the domain of the training data. Therefore, the goal of neural modeling is to minimize the prediction errors of the validation data set presented to the network at the end of the training step.

I.2.6.4 Selection of an artificial neural network

The choice of a neural model for engineering issue modeling is a topic of debate due to its art rather than science. Four aspects determine the effectiveness of ANN applications:

- Network type
- Network topology (number of hidden layers, and number of neurons per layer).
- Activation functions.
- Training algorithm.

I.2.6.4.1 Neuron type

The forward-type neural networks having the universal function approximator characteristic are MLPs and RBFs. Despite this common characteristic of these two types of networks, they present a certain number of differences which are mainly:

- An RBF has only one hidden layer, while an MLP can have one or more hidden layers.
- For an MLP, the neurons of the hidden layers or the output layer have the same model, while the neurons of the hidden layer of the RBF do not have the same model and in addition, in terms of their functions of activation of neurons in the output layer.
- The hidden layer of an RBF is nonlinear and the output layer is linear while the hidden layers and the output layer of an MLP are all nonlinear. However, in the case of nonlinear regression, the output layer is preferably linear.

Given that it is a black box static modeling to approach a bounded and continuous function, the best possible neural method is that of MLPs with a single hidden layer.

I.2.6.4.2 Selection of network structure

The variability in the number of neurons in the hidden layer has a significant impact on the prediction ability of the network. The appropriate number of hidden neurons depends in a complex way on:

- The number of neurons in the input and output layer.
- The size or dimension of the training sample.
- The amount of noise in the targets (network outputs).
- The complexity of the function or classification to be learned by the network.
- Network architecture.
- The type of activation function of the hidden neurons.
- The learning algorithm.
- Regularization.

In general, there is no rule for determining the appropriate number of hidden neurons without having to train multiple networks and then evaluate the generalization error of each network. Very few hidden neurons give too high learning errors and generalizations due to underfitting. On the other hand, too many hidden neurons give a very low learning error but too high generalization error due to overfitting.

The simplest, most reasonable, and probably the most popular way to limit the training error is the root mean square error minimization method.

I.2.6.4.3 Selection of activation functions

The activation function, also known as the neuronal transfer function, is the mathematical function that is used at the output of each neuron. Again, there is no method to predict in advance which function will give the best results.

I.2.6.4.4 Training algorithm

Learning is the procedure by which the values of the network's synaptic weights are determined and establish the network structure. The performance of learning algorithms depends on the modeled process, the training sample, and the learning mode. This explains that the choice of algorithms, although of primordial importance, depends closely on the application concerned and remains mainly experimental. Back-propagation algorithms offer the possibility of using different training algorithms (function). They belong to one of the following five classes:

- Descent of the gradient
- Descent of the conjugate gradients
- Quasi-Newton
- Levenberg-Marquardt
- Regularization.

I.2.7 Adaptive Neuro-Fuzzy Inference System (ANFIS)

Adaptive Neuro-Fuzzy Inference System (ANFIS) combines fuzzy systems and neural networks. In ANFIS, we use a neural network to select the parameters of the fuzzy system. ANFIS mainly removes the necessity for manual optimization of the fuzzy system parameters.

System parameters are automatically adjusted using the neural network. For example, Adjust the limits of membership functions, resulting in improved performance without operator intervention.

The neuro-fuzzy system, which has the learning ability of neural networks and the benefits of a rule-based fuzzy system, can enhance significant performance and provide a mechanism for incorporating past observations into the process of classification or regression.

In neural networks, training builds the system. However, with a neuro-fuzzy graph, the system is constructed from fuzzy logic concepts and is further enhanced using neural network training algorithms.

Fuzzy logic and neural networks are universal estimators. They can estimate any function with approved accuracy if sufficient hidden neurons and fuzzy rules are available. For fuzzy neural networks, backpropagation and gradient descent algorithms are used to adjust the parameters of the membership functions (fuzzy sets) and the defuzzification weights (neural networks). To update the parameters, ANFIS uses these two methods.

ANFIS is a FIS (Fuzzy Inference System) implemented within the framework of an adaptive fuzzy neural network. In ANFIS, the knowledge representation of a FIS is associated with the learning power of ANN.

ANFIS mainly aims to integrate the characteristics of fuzzy systems and neural networks. Fuzzy has the advantage of prior knowledge in a set of constraints to limit the optimization search space. To automate the parametric adjustment of fuzzy control, we use the neural network to adapt back-propagation to a structured network.

The gradient descent algorithm is used by ANFIS to fine-tune the premise parameters that determine the membership functions. ANFIS uses the least squares method to identify the consequent parameters that describe the coefficients of each equation.

This approach combines the gradient descent algorithm and the least squares method, known as the hybrid learning method. To ensure effective generalization to unseen data, the size of the training data set must be at least equivalent to the number of adjustable parameters in ANFIS [89].

The neurons in ANFIS have different structures:

- The Membership function is described by parameterized soft trapezoids (Generalized Bell Functions).
- Functions are linear regressions and multiplication with *w*, that is, normalized weights ω, and Output (Algebraic Sum).
- The Normalization is by Sum and arithmetic division.
- The rules are differentiable T-norm usually product.

I.2.7.1 Fuzzy system

Generally, "system" refers to a set of components that interact with a well-defined and organized structure as a complex whole that can be distinguished from the "external" environment. A system communicates with the environment through what we call inputs and outputs. Fuzzy systems are structures that use fuzzy techniques to process information, where it is impossible or difficult to use classical set theory and binary logic. In the literature, expressions like fuzzy system, fuzzy model, fuzzy rule-based system, fuzzy controller or fuzzy associative memory are used, depending on the type of application. They are mainly distinguished by the symbolic representation of knowledge in the form of fuzzy conditional rules (if-then).

The fuzzy system has four main blocks (Figure I.23): the fuzzifier, the fuzzy inference engine, the knowledge base, and the defuzzifier. An entry for a flow system can be linguistic values (defined by fuzzy sets) and crisp (numerical). When dealing with crisp data, the inference process is preceded by fuzzification, which assigns the appropriate fuzzy set to the nonfuzzy input. The appropriate method of approximate reasoning (inference engine) uses expert knowledge to transform the values of the input variables into linguistic values of the output variable. These values are represented by a set of fuzzy conditional rules (knowledge base). In addition to linguistic values, numerical data may be required as the output of the fuzzy system. In these situations, defuzzification techniques are employed, which assign the representative crisp data to the resulting output fuzzy set.

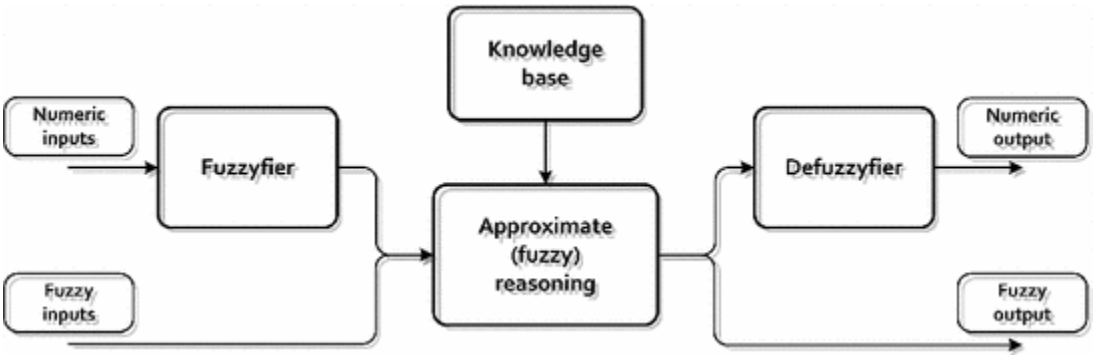


Figure I.23. Fuzzy system structure [90].

Fuzzy systems can be used in problems where the full mathematical description is not accessible or where the use of the precise (non-fuzzy) model is uneconomical or very impractical. Indeed, due to its ability to process inaccurate information, a fuzzy system is a

very useful tool, particularly for control processes, system identification, decision-making support, and the processing of signals and images [90].

I.2.7.2 ANFIS architecture

ANFIS is a neural network that focuses on the Takagi-Sugeno fuzzy inference system. It is an artificial intelligence (AI) technique currently employed in hydrological phenomena. Jang introduced ANFIS in 1993, based on the first-order Sugeno fuzzy model. ANFIS frequently uses back-propagation or a combination of back-propagation and least squares estimation to estimate the membership function (MFs) parameter. The integration of neural networks with fuzzy systems mainly aims to exercise the learning ability of the neural network, while the learning ability represents an advantage in terms of fuzzy systems. On the other hand, a combined system has more advantages from a neural network perspective.

In ANFIS, a Takagi-Sugeno fuzzy inference system is employed, with each rule's output being either a constant term or a linear mixture of input variables in addition to a constant. The final result is the weighted average of each rule's outputs [91].

An adaptive network is a multi-layer feed-forward network. There are five layers in this network. This network's input and output layers are the descriptors and response, respectively. The hidden layer contains the node functions as the MFs and the rules, with each node performing a specific function. In this system, the "square nodes" (adaptive nodes) have parameters, while the "circle nodes" (fixed nodes) have no parameters (see Figure I.24). Consider a Sugeno fuzzy model with two inputs, x and y, and one output. A common rule set with two fuzzy “if-then” rules is as follows [92]:

Rule 1: if $x = A_1$ and $y = B_1$, then

$$f_1 = p_1 x + q_1 y + r_1 \qquad \text{I.13}$$

Rule 2: if $x = A_2$ and $y = B_2$, then

$$f_2 = p_2 x + q_2 y + r_2 \qquad \text{I.14}$$

Where,

f : an output parameter (response)

$p, q,$ and r : are linear parameters

$A,$ and B : are nonlinear parameters.

The equivalent ANFIS structure of the first order Takagi-Sugeno inference system is shown in Figure I.25 Consider a first order Takagi-Sugeno fuzzy model comprising a system of two inputs, and one output (Figure I.24).

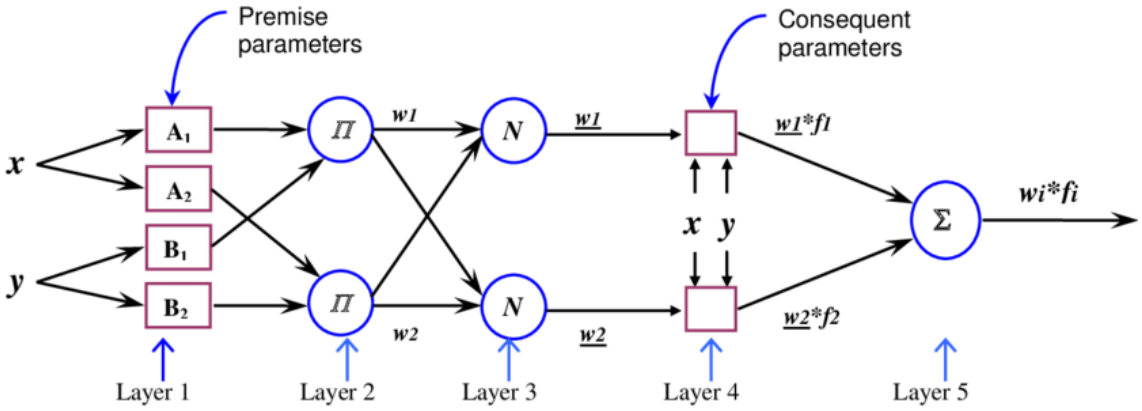


Figure I.24. ANFIS model architecture [93].

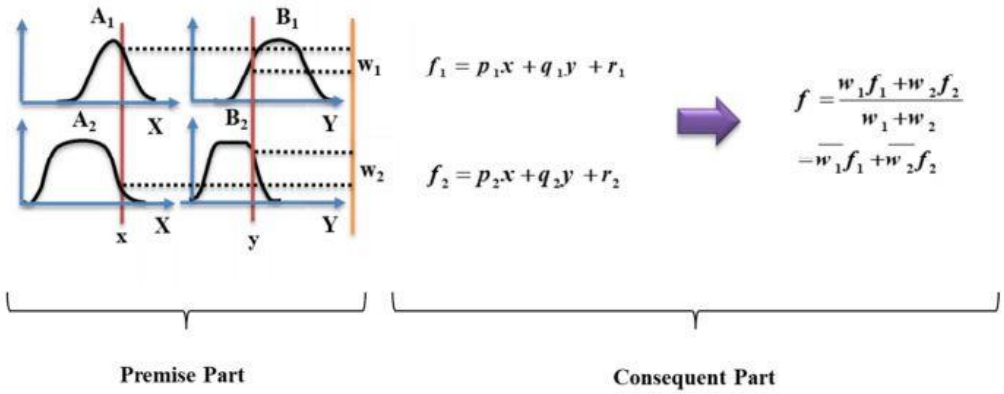


Figure I.25. Schematic structure of the TSK fuzzy model [94].

The functioning of the five layers of the ANFIS system is summarized below.

Layer 1. Input nodes. Every node of this layer produces membership grades to which each of the appropriate fuzzy sets belongs using membership functions. It's "a square node" with a node function.

$$\begin{cases} O_i^1 = \mu_{Ai}(x) \\ O_i^1 = \mu_{Bi}(y) \end{cases} \quad \text{I.15}$$

Where,

$i = 1, 2$

x, y The inputs to node i

A_i and B_i is a fuzzy set associated with this node, characterized by the shape of the MFs in this node, and can be any appropriate functions that are continuous and piecewise differentiable such as Gaussian, generalized bell-shaped, trapezoidal shaped and triangular shaped function.

Layer 2. rule nodes. The AND operator is used in the second layer to obtain an output that matches the antecedent result for that rule, i.e. the firing force. firing strength refers to the degrees where the antecedent part of a fuzzy rule is fulfilled and which shapes the output function of the rule. The outputs O_i^2 of this layer are the products of the corresponding degrees from the Layer.

Every node in this layer is identified as “a circular node” with a label Π (see Figure I.24), which allows the incoming signals to be multiplied and sends the product out. For example,

$$O_i^2 = w_i = \mu A_i(x) \cdot \mu B_i(y) \quad \text{I.16}$$

Layer 3. average nodes. in the third layer, each node is identified as "a circle node" with a label N. The main objective is to determine the ratio of each rule's firing strength to the sum of all rules' firing strength. Thus, the normalized firing force is taken as w_i.

$$O_i^3 = \overline{w_i} = \frac{w_i}{w_1 + w_2} \quad \text{I.17}$$

Layer 4. consequent nodes. The node function in this layer computes the contribution of the i^{th} rule towards the model output, and the functions given as:

$$O_i^4 = \overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i) \quad \text{I.18}$$

Where,

$\overline{w_i}$ The output of the ***layer 3***

$p_i, q_i, and\ r_i$ The parameter set.

The parameter of this is referred to as consequent parameters. Each node is identified as "a circle node" with a node function.

Layer 5. Output nodes. The single node in this layer is "a circle node" with a label Σ that computes the overall output O_i^5as the summation of all incoming signals.

$$O_i^5 = overall\ output = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad \text{I.19}$$

I.2.8 Support vector machine (SVM)

In the field of statistical learning, SVMs are currently a trendy topic. In the late 1990s, the difficulty of generalizing and producing models using traditional neural network approaches led to the development of SVMs. It was created by Vladimir Vapnik in 1995 and quickly gained popularity due to its many attractive features [95].

SVMs are a supervised learning approach that comes with learning algorithms that analyze a set of data. Initially, it was implemented as an approach to solving classification problems. However, due to its multiple attractive properties, it has recently been extended to regression analysis, called support vector regression (SVR). The regression version of SVM is the most common type of application [95,96].

SVMs are a category of general linear classifiers. Differently, SVM is a classification and regression prediction tool that uses machine learning theory to maximize prediction accuracy while automatically avoiding overfitting data [66].

SVMs are systems that use a hypothesis space of linear functions in a high-dimensional feature space. These systems are trained using a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. It is also employed in a wide range of applications, including handwriting and facial analysis, as well as pattern classification and regression-based applications [66].

The advantage of this method is more convenient and efficient than other methods during training. This algorithm determines the optimal separation line for classifying data with a higher safety margin. Vectors are used as a selection criterion in this method, and they are used to determine the optimal data boundaries and categories. These vectors are known as Support Vectors [97].

I.2.8.1 Fundamental Concepts of Statistical Learning Theory

Statistical learning theory is a framework for machine learning that combines statistics and functional analysis. The objective of statistical learning is to determine the fittest function

f that represents the systematic information of the response from its predictors, which can be utilized for prediction and inference.

To estimate this unknown function, the method employed in statistical learning is called training data, which comes from the use that trains our algorithm how to predict f. Every point in the training data is an input-output pair, with the input mapping the output. The statistical learning issue entails inferring the function that maps between the input and the output, so the learned function may be used to predict output from future input.

Statistical learning theory is commonly utilized for regression modeling and data categorization. In regression, responses are quantitative values with a continuous range of values, but in classification, responses are qualitative values with discrete labels. The technique for finding the function f in statistical learning differs in many areas, such as the choice of loss function, However, there are some similarities.

I.2.8.2 Support vector regression (SVR)

SVR, an extended version of SVM, has become an essential computational tool due to its effective use in regression and prediction issues. SVR is designed to predict regression functions and relies on non-linear implicit data conversion into a larger feature space.

The essential principles of SVR are summarized here by considering a regression function F which is estimated based on the training data in the form $\{(x_i, y_i) \quad i = 1, 2, \dots n\}$, where x_i, and y_i are the input and output sets respectively, and n is the overall number of the dataset. SVR uses regression function to predict target values according to Eq I.20 [96] .

$$F = \omega^T \varphi(x) + b \qquad \text{I.20}$$

Where,

F : is the predicted output

ω : is the weight vector

b : is the bias

$\varphi(x)$: is the high-dimensional input vector.

Flatness in Eq I.20 involves finding a small ω. Thus, it is necessary to reduce the Euclidean norm, i.e. $\|\omega\|^2$. The coefficients ω and b were calculated by reducing the risk function, as mentioned below.

$$R(F) = \frac{1}{2}\|\omega\|^2 + C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(y_i, F_i) \quad \text{I.21}$$

Where,

$$L_\varepsilon(y_i, F_i) = \begin{cases} 0, & \text{if } |y_i - F_i| \le \varepsilon \\ |y_i - F_i| - \varepsilon, & \text{otherwise} \end{cases} \quad \text{I.22}$$

The penalty parameter C is used to identify the trade-off between function complexity and losses. As shown in Figure I.26, ε is the loss function which corresponds to the range between the true values and the regression function. This range can be interpreted as a tube around the regression function. Every point outside the tube is considered a training error. Lε(yi, Fi) is known as the insensitive loss function ε. It is possible to modify equation (I.21) into the following constrained form.

$$\text{minimize } \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad \text{I.23}$$

$$\text{Subject to} \begin{cases} \omega \cdot \varphi(x) + b - y_i \le \varepsilon + \xi_i \\ y_i - (\omega \cdot \varphi(x) + b) \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \quad i = 1,2,\dots,n \end{cases} \quad \text{I.24}$$

Where,

ξ and $\xi^*$are two positive slack variables, which assume nonzero values on the exterior of the ε-tube and zero inside (Figure I.26).

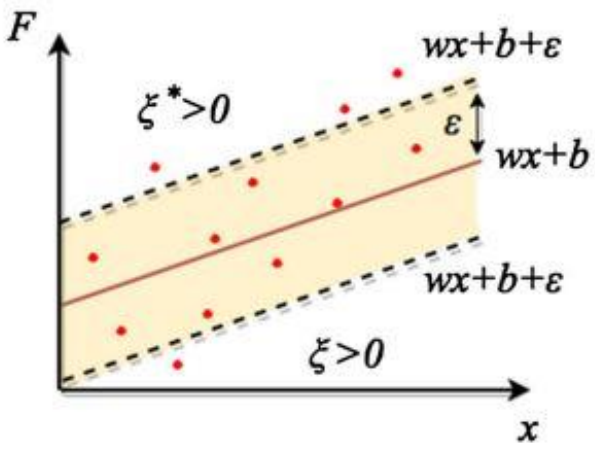


Figure I.26. Regression with the ε-insensitive tube [96].

We can more easily solve the optimization problem of (Eq I.23) in its dual formulation. Thus, a classical dualization method was employed using Lagrangian multipliers. Using Lagrangian multipliers, this problem can be expressed as follows in the dual formulation:

$$\text{minimize } \left\{-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i,\alpha_i^*)\left(\alpha_j,\alpha_j^*\right)\left(\varphi(x_i)\cdot\varphi(x_j)\right)-\varepsilon\sum_{i=1}^{n}(\alpha_i,\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i,\alpha_i^*)\right\} \quad \text{I.25}$$

$$\text{Subject to}\begin{cases}\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0\\ 0\le\alpha_i\le C \quad i=1,2,\dots,n\\ 0\le\alpha_i^*\le C \quad i=1,2,\dots,n\end{cases} \quad \text{I.26}$$

Where,

α_i, α_i^*: are nonlinear Lagrangian multipliers.

The SVR function can be obtained by solving the dual maximization problem in Eq I.25 as follows:

$$F(x,\alpha_i,\alpha_i^*)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)(\varphi(x_i)\cdot\varphi(x_j))+b \quad \text{I.27}$$

The vector inner-product $(\varphi(x_i)\cdot\varphi(x_j))$represents the mapping function from the input space to feature space, and it can be replaced by a kernel function $K(x_i,x_j)$. Hence, Equation I.27 becomes:

$$F(x,\alpha_i,\alpha_i^*)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i,x_j)+b \quad \text{I.28}$$

I.2.8.3 Kernal function

The fundamental principle behind nonlinear SVMs is to create a separating hyperplane in the input space with the greatest margin and transfer training data into a higher dimensional feature space using some mapping $\varphi(x)$. The linear decision function in the feature space correlates to a non-linear decision border in the original input space, as Figure I.27 illustrates. Without explicitly performing the mapping into feature space, the separating hyperplane may be computed by using a kernel function, $K(x_i, x_j)= (\varphi(x_i) \cdot \varphi(x_j))$. For a kernel function to match a feature space, it needs to be symmetric, meaning that $K(x_i, x_j)= K(x_j, x_i)$. Typically, kernels need to satisfy Mercer's Theorem, which states that the kernel matrix $K = (K(x_i, x_j))_{i,j=1}^{n}$ must be positive semi-definite,i.e. it has no non-negative eigenvalues.

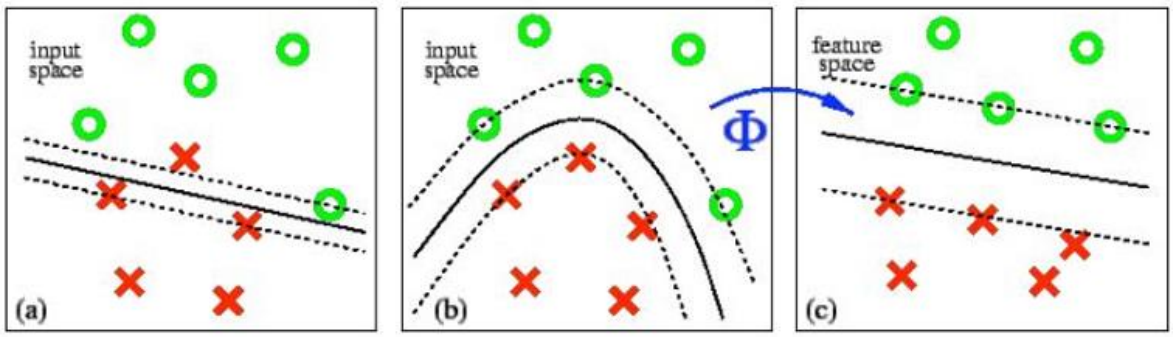


Figure I.27. Effect of kernel transformation [98].

The data is not linearly separable in the input space (a), and the data can only be separated by a non-linear surface (b). This non-linear surface in the input space corresponds to a linear surface in kernel-mapped feature space [98].

Common selections for kernels are:

- Linear Kernel: $K(x_i, x_j)= \varphi(x_i) \cdot \varphi(x_j)$
- Polynomial Kernel: $K(x_i, x_j) = (x_i x_j + 1)^d$
- Radial Basis Kernel Function (RBF): $K(x_i, x_j) = e^{(-\|x_i - x_j\|^2/2\sigma^2)}$

Since there is no explicit mapping to the feature space that each kernel belongs to, it is possible to efficiently find optimum linear separators in feature spaces with millions of dimensions [98].

I.2.9 Dragonfly Optimization Technique

Mirjalili proposed the Dragonfly Optimization (DFO) algorithm, an effective method for optimizing swarm intelligence. It is inspired by the action of dragonflies in nature. Dragonflies are carnivores that capture and eat a wide variety of small insects, from gnats and mosquitoes to wasps and butterflies. Dragonfly swarms are, in general, both active and immobile in the natural environment. Dynamic swarms, also called migratory swarms, consist of large groups (hundreds of thousands of dragonflies) and move in a single direction over long distances, as shown in Figure I.28. During static swarms, where dragonflies hunt their prey, they fly in small groups, often over a small specific area and much closer to land, as shown in Figure I.29 [96].



Figure I.28. Dynamic dragonfly swarms.



Figure I.29. Static dragonfly swarms.

The instinct of each individual in the swarm requires attracting food sources and deflecting external enemies. The MPO algorithm was inspired by these two behaviors. Figure I.30

illustrates the updating of the position of each individual in the swarm and explains it mathematically as follows [96].

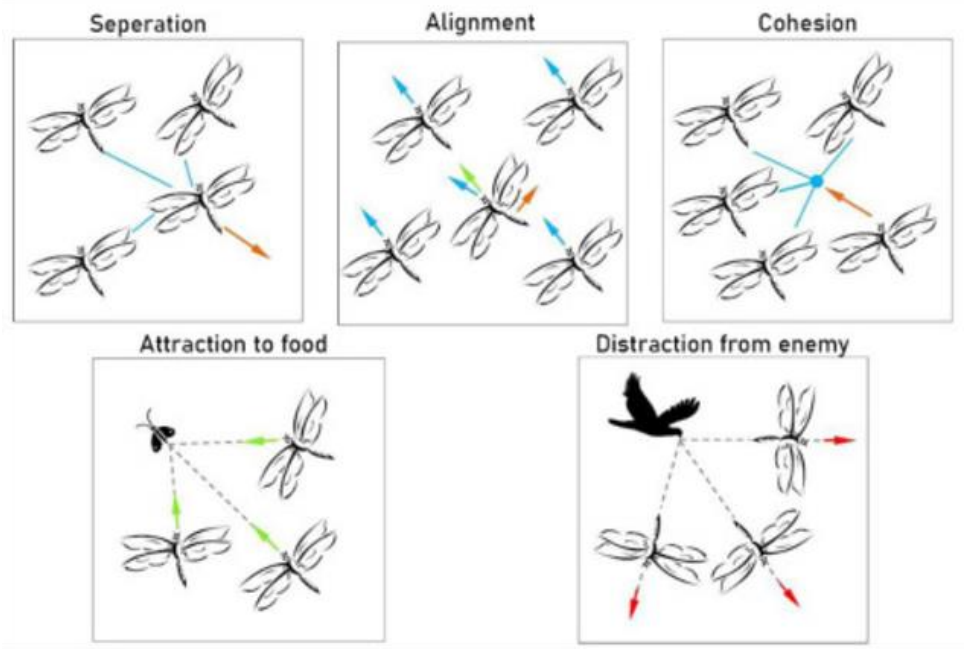


Figure I.30. Primitive corrective patterns between individuals in a swarm.

Separation (S) This step aims to avoid individuals colliding with their neighbors in the static swarm. The equation below expresses this separation.

$$S_i = \sum_{j=1}^{n} X - X_j \qquad \text{I.29}$$

Where,

X and X_j are the positions of the current individual

j^{th} neighboring individual

n is the number of neighboring individuals.

Alignment (A) The purpose of this step is to match the velocity of any individual to the other. The alignment is determined by:

$$A_i = \frac{\sum_{j=1}^{n} V_j}{n} - X \qquad \text{I.30}$$

Where,

V_j is the velocity of neighboring individual j.

Cohesion (C) is the movement of individuals to the center of the swarm's group.

$$A_i = \frac{\sum_{j=1}^{n} X_j}{n} - X \quad \text{I.31}$$

Attraction towards the food (F) all individuals tend to move towards the food.

$$F_i = X^+ - X \quad \text{I.32}$$

Where,

X^+ shows the position of the food source

Distraction outwards an enemy (E) is computed as follows:

$$E_i = X — X \quad \text{I.33}$$

Where,

X^- shows the position of the enemy.

The location of each dragonfly is updated using the step vector ΔX, obtained as follows.

$$\Delta X_i = (sS_i + aA_i + cC_i + fF_i + eE_i) + w\Delta X_t \quad \text{I.34}$$

Where,

$s, a,$ and c represent, respectively, the separation, alignment, and cohesion weights,

f and e : are the food and the enemy factors, respectively,

w : is the inertia weight,

t : is the iteration counter.

The updated position vector is computed as follows:

$$X_{t+1} = X_t + \Delta X_{t+1} \quad \text{I.35}$$

When no neighboring solutions are identified, dragonflies use a random walk, also known as Lévy flying, to improve their exploration and randomization. Dragonflies update their positions using the equation below.

$$X_{t+1} = X_t + \text{Lévy}\,(d) \cdot X_t \quad \text{I.36}$$

Where,

t : is the current iteration

d : is the dimension of the search space. Levy flight is given by:

$$\text{Lévy}(X) = 0.01 \cdot \frac{r_1 \sigma}{|r_2|^{\frac{1}{\beta}}} \quad \text{I.37}$$

Where,

r_1 and r_2 : are the random numbers in the range of [0,1]

β : is a constant,

σ : is given by

$$\sigma = \left(\frac{\Gamma(1+\beta) \cdot \sin(\frac{\pi\beta}{2})}{\Gamma(\frac{1+\beta}{2}) \cdot \beta \cdot 2(\frac{\beta-1}{2})} \right)^{\frac{1}{\beta}} \quad \text{I.38}$$

where,

$$\Gamma(X) = (X-1)! \quad \text{I.39}$$

More information may be found in [99].

I.2.10 Random forest regression

Random forest is an ensemble learning approach that generates many regression tree models in training samples. Random forest is generally used in two tasks: for classification and for regression, which Breiman and Cutler introduced [100]. For regression problems, prediction is performed using the average of all decision tree outputs, where the outcome variable is a real-valued number. A random forest fits the regression model using samples of independent variables [101,102]. Figure I.31 shows the framework for predicting with RF regression.

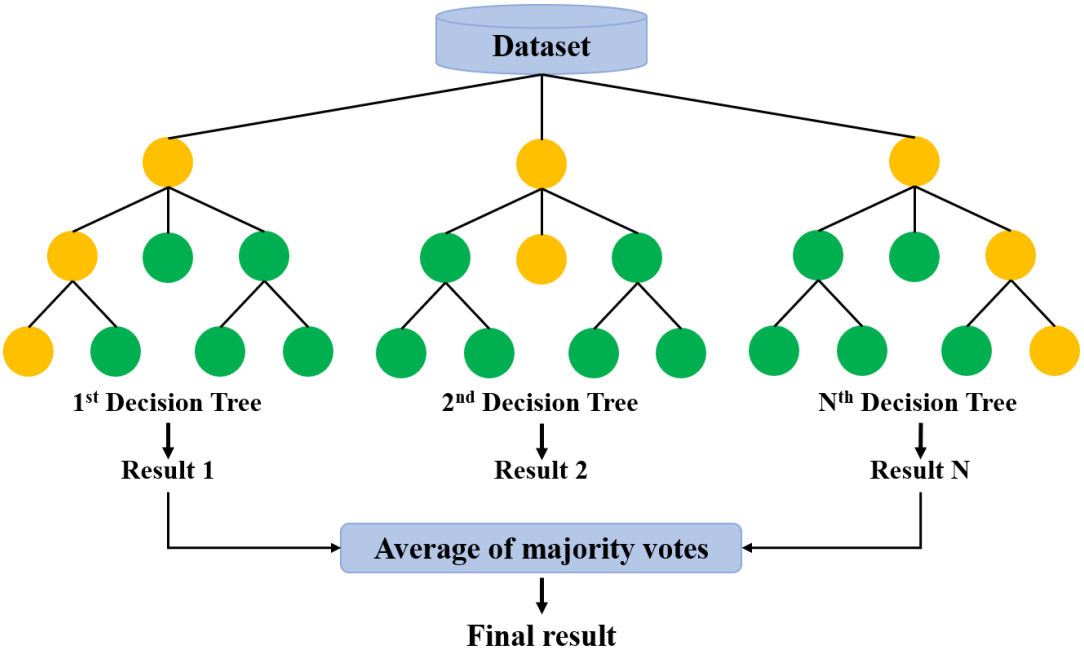


Figure I.31. Illustration of random forest regression.

The random forest regression model is an extension of the CART method. It entails creating numerous uncorrelated decision trees, each with a randomly chosen subset of predictors. The ensemble approach combines these decision trees via a technique known as 'bagging' or bootstrap aggregation. Bagging decreases the variation associated with prediction while improving performance. RF can be created by randomly sampling a feature or training data subset for each decision tree, often known as 'bootstrap'.

There are two possible training parameters for RF: the number of trees (n_{tree}) and the number of random features for each split (m_{try}) in the forest. The more trees grow the more robust and accurate the forecast will be. However, more trees might raise the computational burden. The generalization error converges as the number of trees rises, implying that estimate accuracy cannot be enhanced beyond a certain point. The default value, n_{tree}= 500, is commonly used for prediction. m_{try} is a sensitive parameter that determines model strength and defines the strength of each tree as well as the correlation between any two trees in the forest. increasing m_{try} may enhance the strength of each tree while increasing the correlation between trees might weaken it. The default value for m_{try}, one-third of the number of predictor variables is frequently a good choice [101].

The bootstrap sample used during tree growth is a random selection with the replacement of molecules in the dataset. molecules that are not used for tree growth are called “out-of-bag

samples." Each tree provides a prediction for its out-of-bag sample, and averaging these results across all trees provides an in-situ cross-validation called "out-of-bag validation."

Random Forest has a mechanism for assessing the relevance of descriptors in the model. When each descriptor is replaced in turn by random noise, the resulting degradation in model quality is determined by descriptor significance. The change in mean-squared error during out-of-bag validation can be used to determine model quality degradation [100].

More information may be found in [103,104].

I.3 Critical review on prediction of surfactant properties

I.3.1 Introduction

The development of fundamental and applied research has grown exponentially, highlighting the importance of surfactants for science, with their multiple potential applications in the fields of chemistry, biology, physics, etc. According to a search using the terms "surfactant" and "formulation" in the abstract and title of the papers in the databases Dimensions and Scopus, there are now between 12,000 and 17,000 publications on the topic. However, since the end of the 1970s, as seen in Figure I.32, there has been an exponential rise in the amount of research conducted in the sector, leading to numerous advancements in both academia and industry [105].

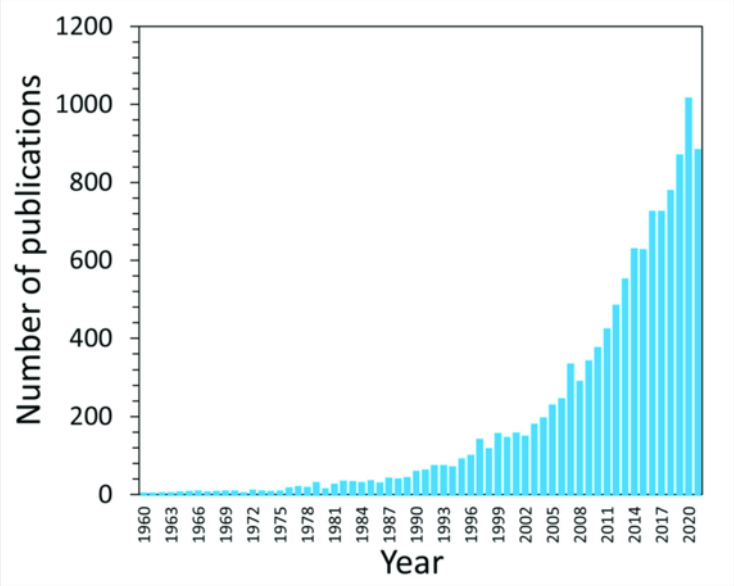


Figure I.32. Number of publications per year with the keywords "surfactant" and "formulation" during the period 1960-2021 [105].

Furthermore, using the VoSViewer text analytics module, a concept map was created for keywords that appeared at least 160 times; the result is shown in Figure 1.33. It shows the connections between three study areas that are categorized into surfactant properties (green), macro- to nano- and micro-emulsions properties (yellow), and formulation systems (blue) [105].

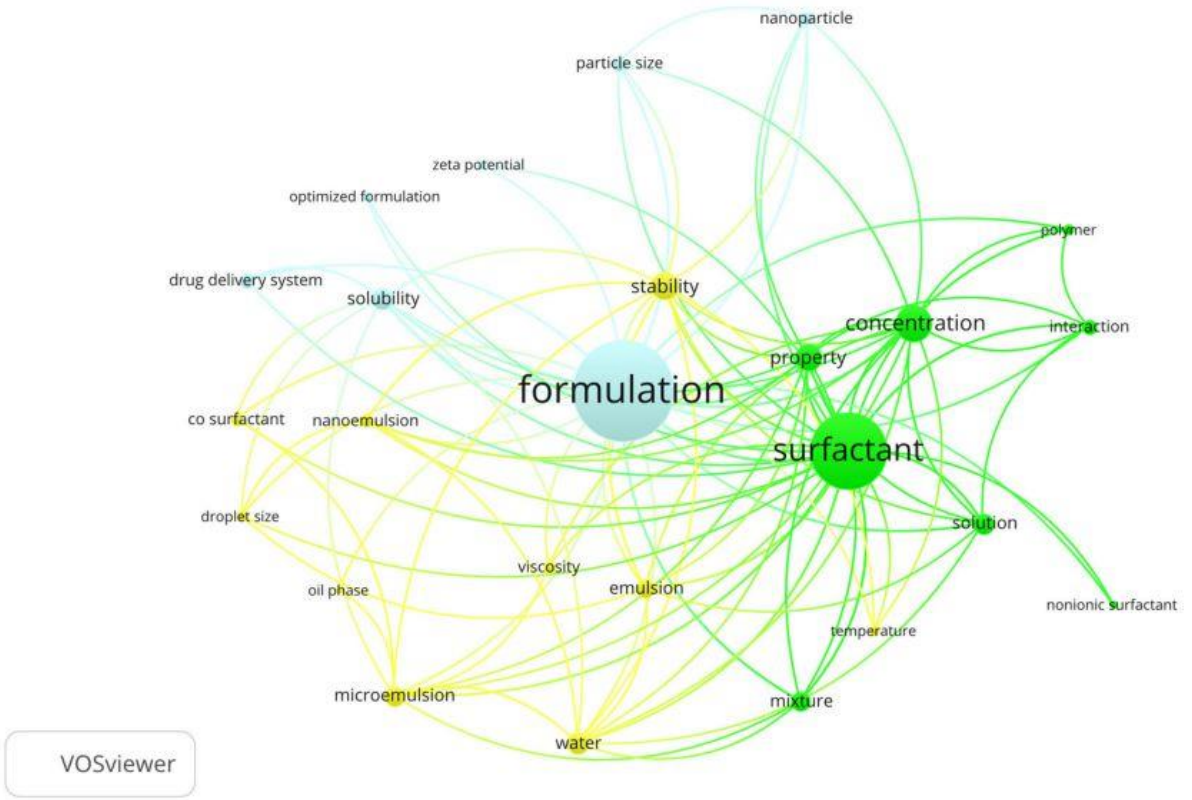


Figure I.33. Concept map of publications related to the keyword search “surfactant” and “formulation” [105].

After presenting the different characteristics of surfactants, as well as their applications, this part will deal with a critical review of the literature concerning methods for predicting certain physicochemical properties of surfactants.

I.3.2 Critical review on CMC of surfactants

For about 40 years, the challenge of theoretically determining the CMC has begun. First, an attempt was made by Rosen [106], who suggested an equation with two terms. Becher [107] published an equation with three terms in 1984. Finally, Ravey et al. [108] suggested adding a nonlinear element to enhance Becher's equation (see Eq I.40).

$$\log\text{CMC} = A + Bn + Cm + Dnm \quad \text{I.40}$$

where n denotes the number of carbon atoms, m the ethylene oxide number, A = 1.77, B = −0.52, C = 0.032, and D = 0.002.

During the past decades, there has been a significant increase in the number of published QSPR studies, mainly due to the evolution of computing power and the development of powerful statistical approaches and software.

Huibers and colleagues [109,110] presented the first QSPR models for predicting CMC using the CODESSA program, which developed equations for nonionic and anionic surfactants. Using the multiple linear regression (MLR) technique, both of these QSPR models were produced over a collection of descriptors derived from the molecular topology and constitution.

In 1996, a dataset of 77 samples was chosen for nonionic surfactants, and classified into seven groups according to the characteristics of their hydrophobic and hydrophilic moieties [109]. The best representation obtained includes three descriptors with a coefficient of R^2= 0.98, using topological descriptors calculated for the hydrophobic fragment of the surfactant molecule. The size of the hydrophobic group, the size of the hydrophilic group, and the structural complexity of the hydrophobic group are the three contributing descriptors.

Regarding anionic surfactants, a dataset of 119 samples was selected, including sodium alkyl sulfates and sodium sulfonates, which have different hydrophobic and hydrophilic structures [110]. Equation (I.41) presented the best model involving three features with a correlation coefficient of R^2 = 0.940.

$$\begin{aligned} \log_{10}\text{CMC} = {} & (1.89 \pm 0.11) - (0.314 \pm 0.010)t\text{-sum-KH0} \\ & - (0.034 \pm 0.003)\text{TDIP} - (1.45 \pm 0.18)h\text{-sum-RNC} \\ & R^2 = 0.940, F = 597, s^2 = 0.0472, n = 119 \end{aligned} \quad \text{I.41}$$

Using three descriptors for a subset of 68 structures (see equation I.42), very good correlations were obtained (R^2 = 0.988), with structural variation only in the hydrophobic domain.

$$\log_{10}\mathrm{CMC} = (2.42 \pm 0.07) - (0.537 \pm 0.009)\mathrm{KH1} - (0.019 \pm 0.002)\mathrm{KS3} + (0.096 \pm 0.005)\mathrm{HGP} \quad \text{I.42}$$

$$R^2 = 0.988, F = 1691, s^2 = 0.0068, n = 68$$

In these regressions, one can conclude that CMC is mainly influenced by the size (volume or surface area) of the hydrophobic domain, and to a lesser extent, by the structural complexity of the surfactant molecule.

Jalali-Heravi and Konouz (2000) [111] utilized MLR to predict the CMC of 31 anionic surfactants from the alkyl sulfate and alkanesulfonate categories. They selected the two best equations and found that they were capable of predicting the CMC based on the statistics of these models and cross-validation results. The models included the Wiener number, the reciprocal of the dipole moment, and the reciprocal of the Randic index. Results showed that topological characteristics, such as compactness and branching, significantly influence micelle formation. The polarity of molecules also played a role, but its effect is less than that of the topology of the surfactants.

Apart from the work of Anoune et al in 2002 [112], the QSPR models intended for CMC prediction were trained on 49 compounds from specific families which are: anionic, cationic, nonionic, and zwitterionic surfactants, among them 30 derived from α-hydroxy acids or gluconolactone synthesized and characterized in the laboratory. They used principal component analysis (PCA) to compare the behavior of the synthesized surfactants to commercial ones that are used as detergents. They found that gluconolactone derivatives exhibited the same activity as those observed for glucopyranoside derivatives. The researchers used partial least squares regression (PLS) to generate regression models. The best model has seven descriptors with a relative standard deviation (RSD) of 0.49 and $R^2 = 0.90$.

Yuan et al. (2002) [113] proposed a QSPR approach based on the MLR model to predict the CMC of anionic surfactant in aqueous solution. They used conformational, electronic, spatial, and thermodynamic properties as descriptors, including the octanol/water partition coefficient, the total area, and the dipole moments of the surfactant molecules to build the model. Four-parameter structure-property relationship between log CMC and descriptors, which was developed for a set of 37 anionic surfactants from sodium alkyl sulfate ($r^2 = 0.996$, *F-test* = 1838.773), can be utilized to estimate the CMC of anionic surfactants.

It is known that the dissolution of a surfactant in water and the formation of micelles are exothermic processes that result from the decrease in free energy in the system and that the energy change is mainly caused by the interactions between the surfactant molecules and the water molecules [1]. The study by Wang et al. in 2002 [114] was carried out to develop a QSPR model to predict the CMC of 39 nonionic surfactants in an aqueous solution. The regression model includes three descriptors including a topological descriptor (the Kier and Hall index of zero order of the hydrophobic segment *KH0*) and two quantum chemicals (the heat of formation ΔH_f and the dipole moment *D*). The estimated QSPR between the descriptors and the logarithm of the CMC gave multiple coefficients of determination: $r^2 = 0.994$, for both 29 linear alkyl ethoxylates and 10 alkyl phenyl polyethylene oxides, and $r^2 = 0.996$ for only the linear alkyl ethoxylates.

In the same year, wang et al. [115] established a new QSPR model for a dataset of 40 anionic surfactants. The best-regressed model includes six descriptors: the kier & hall molecular connectivity index of zero-th order *KH0*, molar heat of formation ΔH_f, the total energy of molecule E_T, the energy of the highest occupied molecular orbital E_{HOMO}, the energy of the lowest unoccupied molecular orbital E_{LUMO}, and the molecular dipole moment *D*.

$$\begin{aligned}\text{lgCMC} = 0.546 - 0.269KH0 - 0.0037\Delta H_f + 0.000224E_T + 0.382E_{HOMO} \\ + 0.493E_{LUMO} - 0.0134D\end{aligned} \quad \text{I.43}$$

The lg CMC, with the descriptors defined in equation I.43, provides a high coefficient of multiple determination of up to 0.9778.

A year later, Wang et al. [116] conducted a further study to derive a QSPR for 77 nonionic surfactants belonging to eight series. Based on their findings, the optimal model included four quantum chemical descriptors (ΔH_f, *D*, E_{HOMO} and E_{LUMO}), and two constitutional descriptors (the molecular weight of surfactant (M) and the number of oxygen and nitrogen atoms (n_{NO})), and one topological descriptor ($KH0$). The general QSPR between lgCMC and descriptors yielded a relevant coefficient of multiple determination of $R^2 = 0.986$.

In 2003, Jalali-Heravi and Konouz [117] utilized MLR to predict the CMC of cationic surfactants from the alkyltrimethylammonium (22 compounds) and alkylpyridinium salts (8 compounds) categories. They selected the two best equations and found that they were capable of predicting the CMC based on the statistics of these models and cross-validation results. The study reveals that geometric characteristics like the volume of the tail of the molecule, the

maximum distance between the atoms, and surface area significantly influence micelle formation in alkyltrimethylammonium salts. However, the simultaneous modeling of both salts' CMCs highlights the importance of Balaban and Randic indices and the electronic parameter of total energy.

Robert (2004) [118] used two methods, the linear Abraham free energy relationship, and the modified polar surface area approach, to develop models for critical micelle concentration (CMC) of anionic and nonionic surfactants. Both methods show predictive accuracy with R^2values around 0.90. In addition, a model based on the combination of these two sets was created to determine the CMC of the data set. Both methods show similar regression coefficients, with size being the most important factor affecting the CMC of the molecule. Strongly acidic H-bond surfactants form micelles at lower concentrations, and increasing H-bond basicity raises the CMC.

The work of Li et al. (2004) [119] aimed to optimize the geometries of the hydrophobic and hydrophilic sectors of 98 anionic surfactants, calculated by ab initio RHF/6-31G(d) for obtained quantum chemical data, such as the density of charge, molecular orbital energy and the pole at two moments. Anionic surfactants used include sodium alkyl sulfates, sodium alkyl sulfonates, sodium alkyl benzene sulfonates and potassium alkyl carboxylates with a wide range of hydrophobic structures. Based on one constitutional descriptor and two quantum chemical descriptors, a QSPR model for the CMC of anionic surfactant was obtained using a multiple linear regression technique. The correlation coefficient R^2 of 0.980 and cross-validation correlation coefficient R_{cv}^2 of 0.978 indicate excellent power and stability for the developed regression equation. The study found that the total number of atoms (N_T) in the hydrophobic surfactant moiety, the dipole moment (μ) of the surfactant moiety, and the maximum net atomic charge on the C atom (Q_{C-max}) in the surfactant moiety, play a major role in the model. Nine anionic surfactants were used to test the developed model, and the calculated critical micelle concentration was consistent with the experimental model.

In 2004, Absalan et al. [120] carried out a study on a database of 94 Gemini surfactants to predict their CMC. Two QSPR-type models based on Genetic-PLS and Genetic-MLR have been developed. The results demonstrated that the MLR (9 chosen descriptors) and PLS (20 chosen descriptors) techniques were capable of perfectly modeling the relationship between the CMC of Gemini surfactants and their topological descriptors. However, PLS regression provides better results. A cross-validation method and response value prediction were used to assess the stability and validity of the models. The best MLR and PLS models had an average

relative prediction error of 2.94 and 0.36, respectively. Furthermore, the correlation coefficients of these models were 0.979 and 0.985, respectively.

The study by Kardanpour and his co-authors [121] was carried out in 2005 on Gemini surfactants to predict CMC. The model is QSPR type, based on a wavelet neural network (WNN). Wavelet networks, which combine the wavelet theory and feed-forward neural networks, utilize wavelets as the basis function to construct the networks. The database used in this study consisted of 94 cationic Gemini surfactants. The number of inputs used to develop this model was equal to twelve (12) molecular descriptors (seven topological, three WHIMs, one geometrical, and one functional group descriptor). The model's performance was investigated by the test set and the average absolute error was 0.105 for the test set, and the WNN model was found to be superior to the MLR model.

Gad carried out a study in 2007 [122] on nonionic surfactants at room temperature to develop a model capable of predicting their CMC. The QSPR type model was developed based on seven (7) molecular descriptors which are molecular weight, hydrophobic-hydrophilic fragments molecular weight ratio, polarizability, partition coefficient, hydration energy, Van der Waal's surface area, and dipole moment. Principal component analysis (PCA) and MLR technique were performed to examine the relationship between the selected descriptor and the CMC of nonionic surfactants. The linear relationship between the selected features and the critical micelle concentration was modeled according to the best statistical results. The best regression model has the coefficient of determination ($R^2 = 0.988$), statistical significance (F=5 391.599), and standard errors ($s^2 = 0.235$). The study concluded that the obtained QSPR model allows estimation of the CMC for nonionic surfactants using theoretically calculated descriptors.

Katritzky and colleagues (2007) [123], exploited a data set comprising 181 different anionic surfactants. They used Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA Pro) software to relate the logarithm of CMC to the molecular structure of the surfactants. A fragment method was employed to create QSPR models, which showed superior statistical characteristics and predictive ability. The results of the regression equations provided insight into the structural properties of surfactants that impact CMC. According to the researchers, the hydrophobic fragments expressed by the topological and geometric descriptors represented the most obvious influence on the CMC. In contrast, the hydrophilic fragment is represented by constitutional, geometric, and charge-related descriptors. Molecular descriptors of significant importance in the selected QSPR models were topological, solvational, and

charge-related descriptors, as the driving force of intermolecular interactions between anionic surfactants and water.

In 2008, Katritzky et al. [124] published linear and nonlinear predictive models for 50 ammonium and quaternary pyridinium cationic surfactants. Multilinear models have been developed for the first and second CMCs. Furthermore, a general ANN model was deduced for the first CMC of the 50 cationic surfactants. The characteristics of these models are mainly associated with the size and charge of the hydrophobic tail and the size of the head. The multilinear model of the second CMC is more closely related to the hydrophobic domain of the surfactant than that of the first CMC. Both linear and nonlinear models used for the first CMC could provide valuable predictions on the CMC of structurally similar cationic surfactants.

Regarding a data collection containing 162 nonionic surfactants, Katritzky et al. (2008) [37] created linear and nonlinear prediction models using MLR and ANN algorithms. Descriptors in the derived models concern molecular shape and size as well as the presence of heteroatoms involved in donor-acceptor and dipole-dipole interactions. Steric hindrance in the hydrophobic region also plays an important role in micellization. The developed linear and nonlinear QSPR models can be useful for predicting CMCs for broad classes of nonionic surfactants.

A QSPR method based on the PLS model was presented by Stanton in 2008 [125] to estimate the logCMC of a dataset including 150 anionic surfactants. The best representation obtained includes five descriptors of type topological and yielded a very good fit to the observed logCMC values with $R^2 = 0.951$, and $Q^2_{LOO} = 0.947$.

Mozrzymas and Roszak (2010) [126] used the QSPR method to estimate the CMC of nonionic surfactants. The relation was established for 23 nonionic surfactants using only connectivity and valence connectivity indices with $r = 0.99$, and $F\text{-test} = 1042.5$.

A study by Guo and colleagues (2011) [127] combined statistical modeling and experimental analysis to analyze the CMC and environmental risks of 120 structurally different Gemini surfactants. The structural properties of the studied compounds were evaluated using hundreds of constitutional, topological, geometric, and electrostatic descriptors. Then, the variants for characterization were calibrated based on experimentally measured properties using different regression techniques, such as MLR, PLS, SVM, RF, and GP. In addition, two sophisticated variable selection methods were used, namely empirical heuristic strategy and nonnumerical genetic algorithm. Of all the models developed, the most predictable model was

developed based on the simplest combination of heuristic variable selection and MLR modeling, with a predictive coefficient of determination (r^2_{pred}) and an average prediction error (RMSP) on an independent external test set of 0.90 and 0.39 respectively. Subsequently, this approach was employed to explain the structural elements that primarily influence the self-assembly behavior of Gemini surfactant molecules in solution and to develop several new Gemini surfactants that exhibit potentially high CMC activity and low risk in the environment. Furthermore, these created substances were produced by quaternary ammonium reaction and were analyzed using elemental analysis, ^{1}H NMR, ^{13}C NMR, and mass spectrometry. One promising candidate demonstrated excellent CMC efficiency, reaching 0.83 mmol L^{-1} at a temperature of 25 °C. This experimental value is quite consistent with the predicted model of 0.89 mmol L^{-1}.

Using a new topological index, the extended distance matrix, Zhi-Chen and colleagues (2012) [128] examined the application of the QSPR model to predict CMC for a diverse set of 175 surfactants. According to the research, the new model based on this topological index is extremely effective and gives satisfactory results. The training set has a squared correlation coefficient (R^2) of 0.9295 and an average relative difference (ARD) of 8.20%, while the testing set has a squared correlation coefficient of 0.9257 and an ARD of 6.76%. The results mean that the developed model is of high quality.

In 2012, Roy and his co-authors [129] conducted a study on predicting logCMC values of 54 non-ionic surfactants using QSPR models with extended topochemical atom (ETA) indices. ETA descriptors were developed by the authors and can be calculated from 2D chemical structure representations without conformational analysis or alignment steps. Various chemometric tools were used, like stepwise multiple linear regression (MLR), genetic function approximation (GFA), and partial least squares (PLS). The final PLS models were found to be well-validated internally, externally, and overall. The best ETA model showed reliable logCMC values prediction, and the statistical quality was comparable to the non-ETA model. The use of ETA descriptors improved the statistical quality of the models. ETA indices encode important chemical information about topological attributes, electronegativity, molecular volume, branching, shape parameters, and atom and bond nature. Therefore, ETA descriptors can be used to model CMC values of non-ionic surfactants.

Roy et al (2012) [130], analyzed CMC data from a series of 35 cationic surfactants to establish predictive models using different classes of descriptors (ETA and non-ETA). The genetic function approximation technique was used to develop the models and verify their

validity using different validation statistics. A comparison was made between the most efficient model developed using the ETA descriptors and that obtained with the non-ETA variables. The external predictive potential of the ETA model was higher than that of the non-ETA model, with ($R^2 = 0.904$). According to the results, it was found that the ETA model was able to effectively determine the CMC profile of the untested molecules. The association of the two classes of descriptors made it possible to obtain a QSPR model which thus revealed an improved external predictive potential. ETA descriptors represented the importance of molecular size and their degree of unsaturation, while non-ETA descriptors represented the influence of bond order and the number of fragments carrying hydrogen atoms attached to the atoms. of carbon with variable hybridization. Thus, the addition of ETA descriptors made it possible to improve the quality of the model by precisely identifying the essential molecular fragments that contribute to their CMC profile, which allows it to be used exhaustively to predict the properties of new cationic surfactants.

Also in 2012, Roy and co-workers [131] developed QSPR models to determine the correlation between molecular structures and CMC of 37 anionic surfactants using extended topochemical atomic (ETA) indices and calculated hydrophobicity descriptors. ETA models were also compared to those developed with non-ETA topological descriptors and hydrophobicity terms. Training and test compounds (n=28 and 9, respectively) were selected using a principal component analysis (PCA) score plot. The best performing ETA model ($Q^2 = 0.938$; $R^2_{pred} = 0.923$) could outperform models developed with non-ETA models ($Q^2 = 0.910$; $R^2_{pred} = 0.885$) and combined sets (ETA and non-ETA) of descriptors ($Q^2 = 0.926$; $R^2_{pred} = 0.915$) in terms of statistical quality and predictive capacity. It is found that the presence of hydrophobicity plays a vital role in creating the CMC model of anionic surfactants, while the branched character and electronic nature of surfactants also play a crucial role as evidenced by various ETA parameters.

Based on the Marrero and Gani GC-method Mattei et al.(2013) [132] developed a group-contribution (GC) property prediction model for estimating the critical micelle concentration (CMC) of nonionic surfactants in water at 25°C. The model is based on a wide range of nonionic surfactants with varying molecular structures. The Marrero and Gani GC-model includes new third order groups based on these structures, assigning compounds with larger correlation errors to more detailed molecular descriptions. The model uses 150 experimental measurements covering various nonionic surfactants, including linear, branched, and phenyl alkyl ethoxylates, alkanediols, alkyl mono and disaccharide ethers and esters, ethoxylated alkyl

amines and amides, fluorinated linear ethoxylates and amides, polyglycerol esters, and carbohydrate derivate ethers, esters, and thiols. The model consists of linear group contributions, and the critical micelle concentration is estimated using the molecular structure of the surfactant alone. This model provides accurate correlations and facilitates faster application in computer-aided molecular design techniques, facilitating chemical process and product design.

In 2016, Gaudin et al. [133] proposed new QSPR models for the CMC of 82 sugar-based surfactants. They developed a series of QSPR models based on molecular descriptors of the whole molecule and the hydrophobic and hydrophilic fragments. Statistical evaluation results indicated that the best QSPR model including quantum chemical descriptors for the entire molecular structure with a root mean square error (RMSE) of 0.32 (log) was evaluated on the validation set of 27 molecules. Thus, a simpler and more efficient model with an RMSE of 0.36 (log) on the validation set was also found, including only constitution-based fragment descriptors.

Baghban et al. (2018) [134] in their work demonstrate the ability of a least squares support vector machine (LSSVM) based on the QSPR model to predict the CMC of sugar-based surfactants. The model was trained and validated using a comprehensive database of 83 sugar-based surfactants. The reliability and robustness of the model were evaluated using different parameters, revealing significant predictive capabilities. Their developed model outperforms previous models based on best multicollinearity regression (BMLR) and group contribution, with a lower RMSE value of 0.023.

For the same dataset of the work of Baghban and Gaudin, Wang et al (2018) [135] performed a multiple linear regression model to predict the CMC of sugar-based surfactants using standard index descriptors. The statistical results showed that the developed model could provide satisfactory predictive results with a squared correlation coefficient for the whole dataset, training set, and test set of 0.9545, 0.9547, and 0.9699, respectively.

In 2019, Mozrzymas [136] examined the impact of head group configuration on the CMC of cationic surfactants, primarily quaternary ammonium bromide salts with an alkyl tail. This study was carried out using molecular connectivity indices and atomic partial charges based on the QSRR approach. Two models were developed to analyze the impact of modifying the structure of the head groups on the value of the CMC. The model that includes molecular connectivity indices and localized charge on the central nitrogen atom has a broader descriptive

capacity than the model that only includes topological indices. Current semi-empirical studies of the charges of various quaternary ammonium salts reveal that the charges of the quaternary ammonium head groups are essentially identical. However, one of the models obtained reveals that the significant part of this charge, located on the central nitrogen atom, has an impact on the value of the CMC.

Jiao et al. (2019) [137] proposed a QSPR study on the CMC of 120 Gemini surfactants using the quantitative hologram structure-activity relationship (HQSAR) technique. Three fragment parameters were used in the HQSAR model, fragment distinction, fragment size, and fragment length. Two conventional validation techniques were employed to evaluate the accuracy of the model predictions. The results showed that the HQSAR model is accurate to study and preliminarily predict the CMC of Gemini surfactants with Q_{F3}^2, concordance correlation coefficient, root mean squared error, $\bar{r}_m^2$ and Δr_m^2 of external test set validation is 0.9799, 0.9803, 0.1757, 0.9429 and 0.0100, respectively. The root mean squared error, $\bar{r}_m^2$ and Δr_m^2 of LOO-CV is 0.2734, 0.9296 and 0.0436, respectively. The study also demonstrated that HQSAR is a promising approach to building a QSPR model related to the CMC of Gemini surfactants.

Setiawan and co-authors [138] used a QSPR approach to correlate the logarithm of CMC with the molecular structure of 231 diverse Gemini cationic surfactants. The QSPR models were created using the Online Chemical Modeling Environment (OCHEM), which offers various machine-learning methods and sets of molecular descriptors. Eight different software packages were used to calculate the molecular descriptors, including Dragon v6, OEstate and ALogPS, CDK, ISIDA Fragment, Chemaxon, Inductive Descriptor, SIRMS, and PyDescriptor. A total of 64 QSPR models were generated, and one consensus model was developed by using a simple average of 13 top-ranked individual models. Based on the statistical coefficient of QSPR models, a consensus model was the best. The model provided the $R^2 = 0.95$, $Q^2 = 0.95$, RMSE = 0.16, MAE = 0.11 for the training set and $R^2 = 0.87$, $Q^2 = 0.87$, RMSE = 0.35, MAE = 0.21 for the test set. Additionally, this model can be used to estimate the CMC of new Gemini cationic surfactant compounds during their early stages of development.

Rahal and colleagues [139] carried out a study in 2020 aimed at predicting the CMC of surfactants. To this end, three regression techniques were selected to create robust predictive models, namely: multilayer perceptron–artificial neural network (MLP-ANN), multiple linear regressions, and partial least squares approach, which are based on four molecular structural descriptors. The data set used to develop these models includes 50 classical and extended

anionic surfactants. Various statistical analysis parameters were used to evaluate the performance of these models including MLP/ANN which demonstrates their superiority compared to other models with ($\bar{r}_m^2 = 0.87, Q_{LOO}^2 = 0.93, Q_{F1}^2 = 0.95, \Delta r_m^2 = 0.15$). the architecture of this model {4.6.1}, trained with Broyden-Fletcher-Goldfarb-Shanno algorithm. Also found that electronic characteristics, structure, and size of the molecule, as well as the number of atoms in the longest aliphatic chain, play a vital role in creating a CMC model for anionic surfactants.

Turchi et al. (2021) [140] presented a newly developed method, which used first-principles-based surface tension calculations rooted in COSMO-RS theory, to predict the critical micellar concentration of a range of nonionic, cationic, anionic, and zwitterionic surfactants in water. This approach combines two prediction strategies to model two different phenomena involving the removal of the hydrophobic tail of the surfactant from contact with water. These strategies are based on micelle formation and thermodynamic phase separation of surfactant from water, taking into account the various polarities of the hydrophilic head group. The method provides accurate predictions of the critical micellar concentration within a logarithmic unit for various types of surfactants and provides first-principles-based predictions of formulation properties in more complex formulations.

The study by Qin et al. [141] (2021) demonstrates that graph convolutional neural networks (GCNs) can predict chemically modified compounds (CMCs) directly from the molecular structure of the surfactant. The GCN encodes the structure of the surfactant in a molecular graph and is trained using CMC experimental data. GCN can predict CMCs with higher accuracy than previous methods and can be generalized to anionic, cationic, zwitterionic, and nonionic surfactants. The proposed model has a $R^2 = 0.96$ and an RMSE of 0.23 for the test set. Molecular saliency maps reveal how atom types and surface-active molecular substructures contribute to CMCs, and indicate that this corresponds to the physical rules that link constitutional and topological information to CMCs. The study also proposes new surfactants without experimental CMCs, showing similar trends to molecular simulations. This indicates that GCNs can enable high-throughput screening of surfactants with desired self-assembly properties.

Abooali and Soleimani (2023) [142] examined the ability and feasibility of linear and non-linear models to predict the CMC of anionic surfactants based on molecular descriptors and external factors such as pH and temperature. To build and test the models, 488 data points corresponding to the experimental CMC values from the literature for 111 sodium-based

anionic surfactants, including sulfate types, sulfonate, benzene sulfonate, sulfosuccinate, and polyoxyethylene sulfate. They found that the non-linear model called stochastic gradient boosting trees ($R^2 = 0.999$) is more robust and reliable for predicting the CMC of these surfactants.

I.3.3 Critical review on the surface tension of surfactants

Research on developing QSPR models to predict surfactants' interfacial tension (γ), has been very limited.

Wang et al. [143–145] are the primary authors in this field. They built the first QSPR model in 2002 using a database that included experimental values of surface tension in CMC (γ^0), for 30 nonionic surfactants in an aqueous solution [143]. The model based on two descriptors may be defined as follows:

$$\gamma^0 = 12.66 - 0.01529\Delta H_f + 0.6310KH0 + 0.0006576\Delta H_f KH0 \qquad \text{I.44}$$

Where ΔH_f and $KH0$ represent the heat of formation, and Kier & Hall zeroeth-order index, respectively. Equation (I.44) has a R^2 value of 0.9877, as reported by the authors.

Wang et al.(2003) [144] have also established QSPR models for 20 anionic surfactants with different structures concerning their surface tension reduction effect expressed as surface pressure (Π) at different temperatures (T) and different counterion concentrations (C). This effect is normally expressed as maximum surface pressure measured at CMC condition Π_{CMC}. Besides C and T, the authors proposed 3 other descriptors which are the Kier & Hall index of zero order, $KH0$, of the hydrophobic fragment, the molar heat of formation ΔH_f, and the molecular dipole moment D. Equation (I.45) shows the suggested model with five descriptors.

$$\Pi_{CMC} = 27.71 - 0.005239\Delta H_f + 17.16C - 0.1520T - 0.2130D + 1.080KH0 \qquad \text{I.45}$$

The best QSPR between Π_{CMC} with $R^2 = 0.9844$, $F = 56.43$, and $s^2 = 0.0134$.

Wang et al. [145] suggested a new model in 2005 utilizing the same database and new descriptors, including a constitutional descriptor, the number of oxygen atoms of the hydrophilic segment (NO); a topological descriptor, the Kier and Hall index of zero order ($KH0$) of the hydrophobic segment; and a quantum chemical descriptor, the heat of formation (ΔH_f) of a mole of surfactant molecules. Equation (I.46) yielded statistical values of $R^2 = 0.9942$, $F = 141.0$, and $s^2 = 0.5337$.

$$\gamma^0 = 11.63 + 0.6750NO + 0.6857KH0 - 0.01261\Delta H_f - 0.1387NO\ KH0 \quad \text{I.46}$$

Fini et al (2012) [146] used a quantitative approach to determine the relationship between the structure of a surfactant compound and its effect on the interfacial tension of oil-water. They developed a new database to measure the interfacial tension between the n-decane as an oil model and various solutions of certain ionic compounds. Despite other reports, they selected surfactant compounds using a scientific method that considered all structural information. Twenty-four different compounds were selected by the principal component analysis technique and their interfacial tensions were measured based on their CMC. Geometric optimization was performed at B3LYP/6-311G**, and the best-suited descriptors were selected using the genetic algorithm (GA-MLR). A predictive test was performed for an external prediction ensemble of six compounds. The resulting GA-MLR model can reasonably predict interfacial tension using only three selected descriptors, with a deviation of 7% between predicted and measured values.

In 2018, Gaudin et al. [147] proposed a novel QSPR model for estimating the surface tension of sugar-based surfactants. The researchers employed a dataset of 70 γ_{CMC} for single surfactants at ambient temperature. Descriptors of the overall structure, polar heads, and alkyl chains of the 70 surfactants were determined and used in multilinear regressions to demonstrate the most predictive and physically significant structure-property relationships. The best model, based on quantum chemical descriptors, attained a standard error of 2.4 mN/m on external validation. Simpler models were also developed based solely on the number of H atoms in the polar head, with a prediction error of 2.9 mN/m. They also discovered that the size of the polar head exhibit had a distinct influence on all generated models.

I.3.4 Critical review on HLB of surfactants

Since the late 1950s, the HLB of surfactants may be determined using Equation (I.47), the Davies' group contribution approach [52]:

$$\text{HLB} = 7 + \sum(\text{hydrophilic group contribution}) - \sum(\text{lipophilic group contribution}) \quad \text{I.47}$$

where the last term is the sum of the group number multiplied by its related value. More recently, exactly in 2006, an improvement of this equation was proposed by Guo et al. [148] using the effective chain length (ECL) instead of the actual chain length of the straight alkyl chain, polyoxyethylene (EO) chain, and polyoxypropylene (PO) chain. They used linear equations for the effective chain length of straight alkyl chain and PO chain, and a Gamma

probability density function (PDF) to describe the contribution of each EO group to HLB and the effective chain length of the EO chain could be obtained by integrating Gamma PDF. The group contribution was utilized to calculate HLB values for 224 nonionic surfactants. Compared to experimental data, the average absolute error (AAE) is less than 1.5, significantly lower than the AAE of 7.3 obtained with the Davies' equation.

In 2007, Chen et al. [149] proposed two high-performance QSPR models to predict HLB values of anionic surfactants based on the quantum chemical. They used multiple linear regression for the developed models. The first model contains 46 anionic surfactants for the series of alkyl sulfates and alkyl sulfonates, with a correlation coefficient (R^2) being 1.000. the second model involved 73 structures including polyoxyethylene, acetate, propionate, and fluorinated anionic surfactants, with R^2 of 0.993.

Gad's 2008 [150] study examined the relationship between molecular structure and HLB of nonionic surfactants using the QSPR approach. Various descriptors having physical significances on the behavior of surfactants in portioning between oil and water, such as the number of oxygen per a nonionic surfactant molecular weight (χ); octanol/water partition coefficient (log P), maximum surface area (Γ), molar volume (V_M), dipole moment (μ), energy of hydration (E_H) and molecular polarizability (α), parachor (ξ), and others Wiener index (W), Randic's molecular connectivity index (R), Kier and Hall valence connectivity index (K & H), average information content (AIC) were selected as inputs of developed model. The link between the chosen descriptors and the HLB values of nonionic surfactants was investigated using principal component analysis (PCA) and multiple linear regression (MLR). The PCA findings illustrate the correlations between HLB and other factors. The linear association between the specified descriptors and HLB was simulated based on the best statistical results. The best model has a coefficient of determination (R^2) of 0.9825, statistical significance (F = 1301), and standard errors ($s^2 = 1.0521$). The resulting QSPR model enables the estimation of HLB for nonionic surfactants using theoretically computed descriptors.

Luan and co-authors [151] conducted a study in 2009 on the estimation of HLB for the family of 73 anionic surfactants, including linear and branched alkylsulfate, linear alkylsulfonate, alkyl benzenesulfonate, dodecyl polyoxy ethylene sulfate, alkyl acetate 2-sulfonates, alkyl propinoate 3-sulfonate, linear alkyl acetate, fluorinated linear alkylsulfonate, and fluorinated linear alkyl acetate. They applied a statistical learning method including stepwise multiple linear regression and a machine learning method including a nonlinear radial basis function neural network (RBFNN) artificial neural network to build the models. Four

parameters were used as input for these models, which are: two quantum chemical descriptors (N_C^{max} and N_O^{min}) belong to the class of descriptors linked to the charge distribution, an electrostatic descriptor (HDCA-2) represents the molecule's ability to give a hydrogen bond and is calculated as the sum of the solvent accessible surface areas of the hydrogen atoms as possible hydrogen donors, and a topological descriptor (IC0) defined based on Shan-non information theory. The results showed that the RBFNN model accurately predicted HLB of anionic surfactants with a quadratic correlation coefficient (R^2)of 0.997 and root mean square (RMS) error of 0.675 for the training set and R^2 = 0.991, RMS = 1.189 for the test set.

In the same year, Chen and his co-authors (2009) [152] used the quantitative structure-property relationship (QSPR) method to investigate the relationship between the hydrophilic-lipophilic balance (HLB) value and the chemical structure of nonionic surfactants derived from ethoxylated alcohols and ethoxylated alkylphenols with the number of EO groups is less than 20. The best QSPR model found has only two basic descriptors. One is the zero order of the Kier-Hall index (KH0), while the other is the number of oxygen atoms in the hydrophilic group (No). The correlation coefficient (R^2) is 0.984. The resulting QSPR model has much greater accuracy than the effective chain length technique.

The study by Beteringhe et al (2011) [153] correlated the HLB values of 14 commercial nonionic emulsifiers with molecular descriptors using the CODESSA program. Due to their high collinearity, they used PLS regression to create a QSPR model with high predictive power (R^2 = 0.908), incorporating orthogonal factors from linear combinations of original descriptors.

Wu et al (2021) [51] developed a quantitative structure-property relationship (QSPR) model to predict the HLB values of 237 non-ionic surfactants using the norm descriptor concept. The model showed good agreement with experimental values, with a squared correlation coefficient (R^2) of 0.9901 and an average absolute relative deviation (AARD) of 2.93% for the whole set. The $R^2$and AARD of the training set and the testing set are 0.9901 (R^2_{train}) and 0.9900 (R^2_{test}), 2.92 % and 3.47 %, respectively. Cross-validation, Y-randomized test, mean absolute error test, and application domain analysis confirmed the model's accuracy, robustness, and reliability. They noticed that the model's accuracy and stability were further validated, indicating its suitability for describing the HLB values of non-ionic surfactants.

I.3.5 Critical review on cloud point of nonionic surfactants and Krafft point of anionic surfactants

An empirical relationship was established by Huibers and co-workers (1997) [154] to predict the cloud point of pure nonionic alkyl ethoxy surfactants. Cloud points can be estimated with an accuracy of 6.37 °C (median error 3.77 °C) for a set of 62 structures, consisting of linear alkyl ethoxylates, branched alkyl, cyclic alkyl, and alkyl phenyl, using the logarithm of the number of ethylene oxide residues as well as three topological descriptors which take into consideration the variation of the hydrophobic domain. The most efficient regression with $R^2 = 0.937$ and $s^2 = 42.3$.

Bünz and colleagues (1998) [155] developed two models to predict the cloud point of a database of 23 surfactants, including a method based on MLR and ANN. The prediction results obtained by the MLR model were very satisfactory. Four descriptors were correlated with excellent correlation: two topological descriptors and two constitutional descriptors of which the squared correlation coefficient is 0.9948, with a mean square deviation of 2.75. The 20 surfactants have a median error of ± 1.2 K in the training set and ± 0.7 K in the three-member prediction set.

In 1998, Heravi and Konouz [156] conducted a study to predict the Krafft Point of surfactants. Linear models have been developed based on 32 linear alkyl sulfates and sulfonates, sulfates, and sulfunates with an ether or ester bond to the hydrophobic tail. Two equations were chosen from various models due to their satisfactory statistical results. The best model is specified using four descriptors: the Randiü index, the heat of formation, reciprocal of the dipole moment, and reciprocal of the volume of tail of the molecule. These descriptors play an essential role in predicting the Krafft point of anionic surfactants. The cross-validation results and statistics of the best models demonstrate that both models are capable of predicting the Krafft point of anionic surfactants. Several methods, including the Akaike Information Criterion (AIC), were employed to select the most appropriate model. They conclude that the electronic descriptors and topological features, including the compactness and branching of anionic surfactants, influence the Krafft point of these molecules.

Yuan et al (2003) [157]established another model to estimate the cloud point for a set of 49 nonionic surfactants with their several structural, electronic, spatial, and thermodynamic characteristics as descriptors. the best regression has five descriptors with an $R^2 = 0.844$.

Ren et al. built the QSPR models in 2006 [158] to predict cloud points and study the clouding phenomena of nonionic surfactants in aqueous solutions. Four descriptors were selected by the heuristic method as inputs to MLR and SVR models. The SVM model outperformed the MLR model in terms of prediction ability and adequacy. They provided a predictive correlation coefficient (*R*) of 0.9882, an absolute average relative deviation of 9.5490, and a root mean square error of 4.2727 for the test set. The suggested models can identify and provide some insight into what structural features are related to the cloud points of compounds, such as the molecular size, structure, and isomerism of the hydrocarbon moiety and the degree of oxyethylation.

The structure-property relationship of a series of pure nonionic surfactants containing linear alkyl, cyclic alkyl, and phenyl alkyl ethoxylates was quantitatively analyzed by Ghasemi and Ahmadi in 2007 [159]. The cloud point of these compounds was modeled using MLR and PLS regression. They used a genetic algorithm (GA) as a method of variable selection in QSPR analysis. According to the results, GA is a very effective variable selection method for QSPR analysis. The analysis of the two regression methods used revealed that PLS has a higher prediction capacity than MLR.

In 2011, Ren et al. [160] developed a new QSPR model specific to the cloud points of nonionic surfactants, using CODESSA descriptors. With a GA method, four descriptors were chosen in order to relate the structures of nonionic surfactants to their corresponding cloud point values. PCA was also used to analyze the descriptors. Two non-linear models were developed which are SVM and projection pursuit regression (PPR). The models were validated in two different ways: internal cross-validation (CV) and test set. The results were analyzed by taking into account the basic elements that affect the property studied and modeling it. In addition, an external database of 16 nonionic surfactants was used to verify the generalization capacity of the optimal model.

Gou et al. (2012) [161] conducted a study combining statistical modeling and experimental analysis to investigate the cloud point of 82 nonionic surfactants. The study characterized the compounds' structural profiles using the hydrophilic domain and the whole molecular. Hundreds of descriptors were computed using the CODESSA program, and the resulting variables were modeled using the Gaussian process (GP). Various regression techniques such as MLR, PLS, SVM, and LSSVM were used to compare the GP with statistical analysis and experimental properties. The most predictable model was constructed using the GP modeling combination of heuristic variable selection related to the hydrophilic domain, with a predictive

coefficient of determination (r^2_{pred}) and root-mean-square error of prediction (RMSP) of 0.962 and 5.200 for a test set, respectively. The model reveals that the cloud point phenomenon is a comprehensive interaction of relative molecular weight, moments of inertia A, and topological structure account for the hydrophilic part.

More recently, a prediction model for the Krafft point was created by Chanachichalermwong et al. (2019) [162] using the Marrero and Gani Group Contribution (GC) concept. The model's accuracy was enhanced by the analysis of literature data and the proposal of a new third-order level. 53 anionic surfactants, comprising alkyl sulfonates, alkyl benzene sulfonates, alkyl ethoxy sulfates, alkyl ester sulfonates, alkyl ester sulfates, alkyl disulfates, alkyldiphenylether disulfonates, alkyl diester sulfonates, alkyl naphthalene sulfonates, and fluorocarbon surfactants were collected along with their Krafft points. The coefficients of the fractional groups in a surfactant molecule reveal the dependence of the Krafft point on the internal structure. The GC model for the Krafft point created provides greater accuracy with a simpler computation when compared to the more complex models, such as those based on advanced statistical concepts.

I.3.6 Critical review on multi-property surfactants

An empirical relationship was established by Li et al. (2005) [163] for predicting the Krafft point of anionic surfactants and the cloud point of nonionic surfactants using the QSPR approach. Concerning the Krafft point, a six-parameter structure-property relationship ($R^2 = 0.941$) was obtained for a diverse set of 46 hydrocarbon anionic surfactants, as well as a three-parameter relationship ($R^2 = 0.864$) for 19 linear fluorocarbon anionic surfactants, using topological techniques, thermodynamic and structural descriptor techniques. The most effective multiple linear regression model for the cloud point of 64 nonionic surfactants with six representative structures was composed of five parameters and had a correlation coefficient of $R^2 = 0.921$. Based on these relationships, the molecular structure of surfactants may be used to predict their cloud point or Krafft point.

Based on the concept of norm index, Shi et al (2019) [164] proposed new norm descriptors to describe the HLB and CMC of anionic surfactants, and two QSPR models were constructed to calculate these two values. The results demonstrated that models using norm descriptors could have a satisfactory prediction effect for the HLB and CMC of anionic surfactants, with a R^2 of 0.9983 for HLB and a R^2 of 0.9130 for CMC. Several satisfactory validations have proven that these models are stable, reliable, and have very high predictive capacity.

Furthermore, for anionic surfactants without experimental values, the HLB for 19 anionic surfactants and the CMC for 101 anionic surfactants were calculated using the proposed models. According to the results, it was found that the norm index was adequate to assess HLB and CMC.

A surface tension (SFT) data set of 154 model hydrocarbon surfactants at 20-30 °C was fitted by Seddon and co-workers (2022) to the Szyszkowski equation to extract three characteristic parameters: surface excess concentration (Γ_m); Langmuir constant (K_L) and CMC which are linked to a series of 2D and 3D molecular descriptors [165]. Ten descriptors were chosen by eliminating co-correlation, using a gradient-boosted regression model to assess feature importance, and performing recursive feature elimination (RFE). They used a cross-validated randomized grid search to adjust the hyperparameters of each target variable model to reduce overfitting and increase predictive accuracy. The study showed that ML models are effective in predicting experimental data, with a correlation coefficient ranging from 0.69-0.87.

In 2024, Brozos et al. [166] built GNN models to predict the CMC and Γ_m. They also investigated several learning techniques, such as single- and multi-task learning, as well as different training strategies, namely ensemble and transfer learning. They created CMC database with 429 molecules and Γ_m database with 164 molecules. They found that a multi-task GNN with ensemble learning trained on all Γ_m and CMC data achieves the best performance. Thus, the results show that the simultaneous use of data from highly correlated properties can improve the ability to predict surfactant properties for which only a small amount of experimental data is available. Finally, they tested the ability of the developed CMC model to generalize on industrial grade surfactants. GNN produces highly accurate predictions of CMC, showing great potential for future industrial applications.

I.4 Conclusion

We have shown in the first part that surfactants play an important role in various areas of daily life. Also, these have applications, depending on their structures and their physicochemical properties.

The purpose of the second section is to provide an overview of some fundamental ideas associated with the topic and the chosen modeling techniques. As described above, each programming tool, approach, model, etc., has its power that is only apparent if it has been utilized appropriately within its operational bounds. Additionally, each model may provide a distinct set of outcomes based on the data that is used.

In the third part, we provide a critical review of the literature on modeling tools to predict surfactant properties. The abbreviation QSPR combines statistical methods useful for quickly predicting the properties of surfactants. A series of models already exist for properties such as CMC, surface tension, cloud point, Kraft point, and HLB, which have shown good predictive accuracy when applied to an external set of amphiphiles.

Bibliographic References

[1] J. Hu, X. Zhang, Z. Wang, A review on progress in QSPR studies for surfactants, Int. J. Mol. Sci. 11 (2010) 1020–1047. https://doi.org/10.3390/ijms11031020.

[2] Y. Nakama, Surfactants, Elsevier Inc., 2017. https://doi.org/10.1016/B978-0-12-802005-0.00015-X.

[3] SURFACTANTS Types and Uses, 2 (2002).

[4] Huiling Lu, Contribution à l'étude des relations structure-propriétés de molécules amphiphiles à tête sucre, (2016) 263.

[5] M.J. Rosen, Surfactants and interfacial phenomena, Third edition, John Wiley & Sons, (2004).

[6] Z. Han, X. Yang, Y. Liu, Micellar and Interfacial Behavior of Mixed Systems Containing Glycoside-Based Surfactant and Cationic Didecyldimethylammonium Chloride, J. Surfactants Deterg. 18 (2015) 873–880. https://doi.org/10.1007/s11743-015-1705-7.

[7] E. Fernández, J.M. Benito, C. Pazos, J. Coca, Ceramic membrane ultrafiltration of anionic and nonionic surfactant solutions, J. Memb. Sci. 246 (2005) 1–6. https://doi.org/10.1016/j.memsci.2004.04.007.

[8] T. Iwasaki, M. Ogawa, K. Esumi, K. Meguro, Interactions between Betaine-Type Zwitterionic and Anionic Surfactants in Mixed Micelles, Langmuir. 7 (1991) 30–35. https://doi.org/10.1021/la00049a008.

[9] C. Larpent, Tensioactifs. Techniques de l'ingénieur constates chimiques des solvants et produits base documentaire: TIB337DUO., (1995).

[10] C.C. West, J.H. Harwell, Surfactants and Subsurface Remediation, Environ. Sci. Technol. 26 (1992) 2324–2330. https://doi.org/10.1021/es00036a002.

[11] F.M. Menger, J.S. Keiper, Gemini surfactants, Angew. Chemie - Int. Ed. 39 (2000) 1906–1920. https://doi.org/10.1002/1521-3773(20000602)39:11<1906::AID-ANIE1906>3.0.CO;2-Q.

[12] M.H. Mondal, A. Roy, S. Malik, A. Ghosh, B. Saha, Review on chemically bonded geminis with cationic heads: Second-generation interfactants, Res. Chem. Intermed. 42

(2016) 1913–1928. https://doi.org/10.1007/s11164-015-2125-z.

[13] R. Zana, Dimeric and oligomeric surfactants. Behavior at interfaces and in aqueous solution: A review, Adv. Colloid Interface Sci. 97 (2002) 205–253. https://doi.org/10.1016/S0001-8686(01)00069-0.

[14] Z. Ye, F. Zhang, L. Han, P. Luo, J. Yang, H. Chen, The effect of temperature on the interfacial tension between crude oil and gemini surfactant solution, Colloids Surfaces A Physicochem. Eng. Asp. 322 (2008) 138–141. https://doi.org/10.1016/j.colsurfa.2008.02.043.

[15] X.F. Geng, X.Q. Hu, J.J. Xia, X.C. Jia, Synthesis and surface activities of a novel di-hydroxyl-sulfate-betaine- type zwitterionic gemini surfactants, Appl. Surf. Sci. 271 (2013) 284–290. https://doi.org/10.1016/j.apsusc.2013.01.185.

[16] M.H. Mondal, S. Malik, A. Roy, R. Saha, B. Saha, Modernization of surfactant chemistry in the age of gemini and bio-surfactants: A review, RSC Adv. 5 (2015) 92707–92718. https://doi.org/10.1039/c5ra18462b.

[17] D. Shukla, V.K. Tyagi, Anionic Gemini Surfactants: A Distinct Class of Surfactants, J. Oleo Sci. 55 (2006) 215–226. https://doi.org/10.5650/jos.55.215.

[18] N. Kumar, R. Tyagi, Industrial Applications of Dimeric Surfactants: A Review, J. Dispers. Sci. Technol. 35 (2014) 205–214. https://doi.org/10.1080/01932691.2013.780243.

[19] J. Yang, B. Guan, Y. Lu, W. Cui, X. Qiu, Z. Yang, W. Qin, Viscoelastic evaluation of gemini surfactant gel for hydraulic fracturing, SPE - Eur. Form. Damage Conf. Proceedings, EFDC. 2 (2013) 942–946. https://doi.org/10.2118/165177-ms.

[20] L. Brinchi, R. Germani, L. Goracci, G. Savelli, C.A. Bunton, Decarboxylation and dephosphorylation in new gemini surfactants. Changes in aggregate structures, Langmuir. 18 (2002) 7821–7825. https://doi.org/10.1021/la020250o.

[21] S. Bhattacharya, V.P. Kumar, Evidence of Enhanced Reactivity of DAAP Nucleophiles toward Dephosphorylation and Deacylation Reactions in Cationic Gemini Micellar Media, J. Org. Chem. 69 (2004) 559–562. https://doi.org/10.1021/jo034745+.

[22] Y. Han, Y. Wang, Aggregation behavior of gemini surfactants and their interaction with macromolecules in aqueous solution, Phys. Chem. Chem. Phys. 13 (2011) 1939–1956.

https://doi.org/10.1039/c0cp01196g.

[23] M. In, V. Bec, O. Aguerre-Chariol, R. Zana, Quaternary ammonium bromide surfactant oligomers in aqueous solution: Self-association and microstructure, Langmuir. 16 (2000) 141–148. https://doi.org/10.1021/la990645g.

[24] L.G. Qiu, M.J. Cheng, A.J. Xie, Y.H. Shen, Study on the viscosity of cationic gemini surfactant-nonionic polymer complex in water, J. Colloid Interface Sci. 278 (2004) 40–43. https://doi.org/10.1016/j.jcis.2004.05.012.

[25] M.S. Kamal, A Review of Gemini Surfactants: Potential Application in Enhanced Oil Recovery, J. Surfactants Deterg. 19 (2016) 223–236. https://doi.org/10.1007/s11743-015-1776-5.

[26] M. Lukic, I. Pantelic, S. Savic, An overview of novel surfactants for formulation of cosmetics with certain emphasis on acidic active substances, Tenside, Surfactants, Deterg. 53 (2016) 7–19. https://doi.org/10.3139/113.110405.

[27] R. Sen, Biosurfactants. New york, spriger, (2010)

[28] M. Kjellin, I. Johansson, Surfactants from Renewable Resources, Surfactants from Renew. Resour. (2010). https://doi.org/10.1002/9780470686607.

[29] K. Shinoda, T. Yamanaka, K. Kinoshita, Surface chemical properties in aqueous solutions of non-ionic surfactants: Octyl glycol ether, α-octyl glyceryl ether and octyl glucoside, J. Phys. Chem. 63 (1959) 648–650. https://doi.org/10.1021/J150575A003.

[30] P. Liljekvist, M. Kjellin, J. Christer Eriksson, Surface pressure effect of pentaoxyethylene and maltoside surfactant head groups, Adv. Colloid Interface Sci. 89–90 (2001) 293–302. https://doi.org/10.1016/S0001-8686(00)00055-5.

[31] H.G. Bazin, T. Polat, R.J. Linhardt, Synthesis of sucrose-based surfactants through regioselective sulfonation of acylsucrose and the nucleophilic opening of a sucrose cyclic sulfate, Carbohydr. Res. 309 (1998) 189–205. https://doi.org/10.1016/S0008-6215(98)00121-9.

[32] K. Hill, O. Rhode, Sugar-based surfactants for consumer products and technical applications, Lipid - Fett. 101 (1999) 25–33. https://doi.org/10.1002/(sici)1521-4133(19991)101:1<25::aid-lipi25>3.0.co;2-n.

[33] S. Matsumura, K. Imai, S. Yoshikawa, K. Kawada, T. Uchibor, Surface activities, biodegradability and antimicrobial properties of n-alkyl glucosides, mannosides and galactosides, J. Am. Oil Chem. Soc. 1990 6712. 67 (1990) 996–1001. https://doi.org/10.1007/BF02541865.

[34] O.J. Rojas, C. Stubenrauch, L.A. Lucia, Y. Habibi, Interfacial Properties of Sugar-based Surfactants.

[35] R. J. Farn, Chemestry and technology of surfactants, John Wiley & Sons, (2008).

[36] C.C. Ruiz, Sugar-Based Surfactants: fundamental and application, CRC Press (2008).

[37] A.R. Katritzky, L.M. Pacureanu, S.H. Slavov, D.A. Dobchev, M. Karelson, QSPR study of critical micelle concentrations of nonionic surfactants, Ind. Eng. Chem. Res. 47 (2008) 9687–9695. https://doi.org/10.1021/IE800954K.

[38] S. Gabet, Remobilisation d'Hydrocarbures Aromatiques Polycycliques (HAP) présents dans les sols contaminés à l'aide d'un tensioactif d'origine biologique, Thèse Dr. Univ. LIMOGES. (2004) 177.

[39] K. Nesměrák, I. Němcová, Determination of critical micelle concentration by electrochemical means, Anal. Lett. 39 (2006) 1023–1040. https://doi.org/10.1080/00032710600620302.

[40] A.R. Glennie, M.M. Mohareb, R.M. Palepu, Thermodynamic and related properties of alkyl cationic surfactants based on dimethyl and diethyl ethanol amines, J. Dispers. Sci. Technol. 27 (2006) 731–738. https://doi.org/10.1080/01932690600660608.

[41] A. Bratovcic, S. Nazdrajic, Viscoelastic Behavior of Synthesized Liquid Soaps and Surface Activity Properties of Surfactants, J. Surfactants Deterg. 23 (2020) 1135–1143. https://doi.org/10.1002/jsde.12444.

[42] S. Thiruvengadam, M. Murphy, J.S. Tan, K. Miller, A Generalized Theoretical Model for the Relationship Between Critical Micelle Concentrations, Pressure, and Temperature for Surfactants, J. Surfactants Deterg. 23 (2020) 273–303. https://doi.org/10.1002/jsde.12360.

[43] L.R. Harutyunyan, R.S. Harutyunyan, Effect of Amino Acids on Micellization and Micellar Parameters of Anionic Surfactant Alpha Olefin Sulfonate C 14 -C 16 in Aqueous Solutions: Surface Tension, Conductometric, Volumetric, and Fluorescence

Studies, J. Chem. Eng. Data. 64 (2019) 640–650. https://doi.org/10.1021/acs.jced.8b00886.

[44] R. Nagarajan, Molecular packing parameter and surfactant self-assembly: The neglected role of the surfactant tail, Langmuir. 18 (2002) 31–38. https://doi.org/10.1021/la010831y.

[45] D. Lombardo, M.A. Kiselev, S. Magazù, P. Calandra, Amphiphiles self-assembly: Basic concepts and future perspectives of supramolecular approaches, Adv. Condens. Matter Phys. 2015 (2015). https://doi.org/10.1155/2015/151683.

[46] O. Massarweh, A.S. Abushaikha, The use of surfactants in enhanced oil recovery: A review of recent advances, Energy Reports. 6 (2020) 3150–3178. https://doi.org/10.1016/j.egyr.2020.11.009.

[47] R. Van Zanten, Stabilizing viscoelastic surfactants in high-density brines, SPE Drill. Complet. 26 (2011) 499–505. https://doi.org/10.2118/141447-PA.

[48] D. Lombardo, P. Calandra, L. Pasqua, S. Magazù, Self-assembly of organic nanomaterials and biomaterials: The bottom-up approach for functional nanostructures formation and advanced applications, Materials (Basel). 13 (2020) 1–43. https://doi.org/10.3390/ma13051048.

[49] S. Kondo, M. Shiohara, K. Maruyama, K. Fukaya, K. Yamada, S. Ogawa, S. Saito, Effect of the hydrophilic-lipophilic balance (HLB) of surfactants included in the post-CMP cleaning chemicals on porous SiOC direct CMP, Proc. IEEE 2007 Int. Interconnect Technol. Conf. - Dig. Tech. Pap. (2007) 172–174. https://doi.org/10.1109/iitc.2007.382381.

[50] S.S. Reham, H.H. Masjuki, M.A. Kalam, I. Shancita, I.M. Rizwanul Fattah, A.M. Ruhul, Study on stability, fuel properties, engine combustion, performance and emission characteristics of biofuel emulsion, Renew. Sustain. Energy Rev. 52 (2015) 1566–1579. https://doi.org/10.1016/j.rser.2015.08.013.

[51] J. Wu, F. Yan, Q. Jia, Q. Wang, QSPR for predicting the hydrophile-lipophile balance (HLB) of non-ionic surfactants, Colloids Surfaces A Physicochem. Eng. Asp. 611 (2021) 125812. https://doi.org/10.1016/j.colsurfa.2020.125812.

[52] J.T. Davies, A quantitative kinetic theory of emulsion type, I. Physical chemistry of the

emulsifying agent, Proc. 2nd Int. Congr. Surf. Act. 1 (1957) 426–438. http://www.firp.ula.ve/archivos/historicos/57_Chap_Davies.pdf.

[53] K. Holmberg, Handbook of applied surfaces and colloid chemestry, John Wiley & Sons, Ltd, chichester (2002).

[54] C. Dicharry, J. Diaz, J.P. Torré, M. Ricaurte, Influence of the carbon chain length of a sulfate-based surfactant on the formation of CO2, CH4 and CO2–CH4 gas hydrates, Chem. Eng. Sci. 152 (2016) 736–745. https://doi.org/10.1016/j.ces.2016.06.034.

[55] F. Makavipour, R.M. Pashley, A.F.M.M. Rahman, Low-Level Arsenic Removal from Drinking Water, Glob. Challenges. 3 (2019) 1–8. https://doi.org/10.1002/gch2.201700047.

[56] N.A. Malik, A. Ali, Krafft temperature and thermodynamic study of interaction of glycine, diglycine, and triglycine with hexadecylpyridinium chloride and hexadecylpyridinium bromide: A conductometric approach, J. Mol. Liq. 213 (2016) 213–220. https://doi.org/10.1016/j.molliq.2015.11.017.

[57] R. Bois, Étude des relations structures - propriétés - fonctionnalités de tensioactifs dérivés de sucres : caractérisation et anticipation des propriétés moussantes HAL Id : tel-03098133 Thèse présentée pour l ' obtention du grade de Docteur de l ' UTC, (2021).

[58] N. Parris, J.K. Weil, W.M. Linfield, Soap based detergent formulations. V. Amphoteric lime soap dispersing agents, J. Am. Oil Chem. Soc. 50 (1973) 509–512. https://doi.org/10.1007/BF02640521.

[59] D. Myers, Surfactant science and technology Third edition, John Wiley & sons, (2006).

[60] L.L. Schramm, E.N. Stasiuk, D.G. Marangoni, Surfactants and their applications, Annu. Reports Prog. Chem. - Sect. C. 99 (2003) 3–48. https://doi.org/10.1039/B208499F.

[61] M.J. Lawrence, Surfactant systems: Their use in drug delivery, Chem. Soc. Rev. 23 (1994) 417–424. https://doi.org/10.1039/CS9942300417.

[62] A. V. Kabanov, V.P. Chekhonin, V.Y. Alakhov, E. V. Batrakova, A.S. Lebedev, N.S. Melik-Nubarov, S.A. Arzhakov, A. V. Levashov, G. V. Morozov, E.S. Severin, V.A. Kabanov, The neuroleptic activity of haloperidol increases after its solubilization in surfactant micelles. Micelles as microcontainers for drug targeting, FEBS Lett. 258 (1989) 343–345. https://doi.org/10.1016/0014-5793(89)81689-8.

[63] A. ROGERSON, J. CUMMINGS, N. WILLMOTT, A.T. FLORENCE, The Distribution of Doxorubicin in Mice Following Administration in Niosomes, J. Pharm. Pharmacol. 40 (1988) 337–342. https://doi.org/10.1111/j.2042-7158.1988.tb05263.x.

[64] I.F. Uchegbu, J.A. Double, J.A. Turton, A.T. Florence, Distribution, Metabolism and Tumoricidal Activity of Doxorubicin Administered in Sorbitan Monostearate (Span 60) Niosomes in the Mouse, Pharm. Res. An Off. J. Am. Assoc. Pharm. Sci. 12 (1995) 1019–1024. https://doi.org/10.1023/A:1016210515134.

[65] F. K. Brown, Chemoinformatics: What is it and How does it Impact Drug Discovery. In ; Annual Reports in Medicinal Chemistry, Elsevier, (1998) 375–384. DOI: 10.1016/S0065- 7743(08)61100-8..

[66] V. Jakkula, Tutorial on Support Vector Machine (SVM), Sch. EECS, Washingt. State Univ. (2011) 1–13. http://www.ccs.neu.edu/course/cs5100f11/resources/jakkula.pdf.

[67] database_2.pdf,.

[68] S. Chtita, biologiquement actives par des méthodes QSAR / QSPR . To cite this version : THÈSE Présentée par : En vue de l ' obtention de Doctorat en Chimie Spécialité : Chimie Physique et Modélisation Modélisation de molécules organiques hétérocycliques biologiqueme, (2017).

[69] D. Weininger, J. Chem. Inj Comput., J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.

[70] A.R. Leach, V.J. Gillet, An introduction to chemoinformatics, 2007. https://doi.org/10.1007/978-1-4020-6291-9.

[71] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for Generation of Unique SMILES Notation, J. Chem. Inf. Comput. Sci. 29 (1989) 97–101. https://doi.org/10.1021/ci00062a008.

[72] D. Weininger, Smiles. 3. Depict. Graphical Depiction of Chemical Structures, J. Chem. Inf. Comput. Sci. 30 (1990) 237–243. https://doi.org/10.1021/ci00067a005.

[73] H. Moriwaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator, J. Cheminform. 10 (2018) 1–14. https://doi.org/10.1186/s13321-018-0258-y.

[74] J. Dong, D.S. Cao, H.Y. Miao, S. Liu, B.C. Deng, Y.H. Yun, N.N. Wang, A.P. Lu, W. Bin Zeng, A.F. Chen, ChemDes: An integrated web-based platform for molecular

descriptor and fingerprint computation, J. Cheminform. 7 (2015) 1–10. https://doi.org/10.1186/s13321-015-0109-z.

[75] J. Gasteiger, Chemoinformatics: Achievements and challenges, a personal view, Molecules. 21 (2016). https://doi.org/10.3390/molecules21020151.

[76] S.C. Peter, J.K. Dhanjal, V. Malik, N. Radhakrishnan, M. Jayakanthan, D. Sundar, D. Sundar, Quantitative structure-activity relationship (QSAR): Modeling approaches to biological applications, Elsevier Ltd., 2018. https://doi.org/10.1016/B978-0-12-809633-8.20197-0.

[77] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C.P. Varghese, R.K. Agrawal, Validation of QSAR Models - Strategies and Importance, Int. J. Drug Des. Disocovery. 2 (2011) 511–519.

[78] OECD, OECD PRINCIPLES FOR THE VALIDATION, FOR REGULATORY PURPOSES, OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP MODELS These principles were agreed by OECD member countries at the 37, Biotechnology. (2004) 3–4.

[79] K. Roy, R.N. Das, P. Ambure, R.B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, Chemom. Intell. Lab. Syst. 152 (2016) 18–33. https://doi.org/10.1016/J.CHEMOLAB.2016.01.008.

[80] J. LESZCZYNSKI, Recent Advances in QSAR Studies CHALLENGES AND ADVANCES IN, (2010).

[81] R. D. Tobias, An introduction to partial least squares regression, statistical analysis system institute Inc, cary, USA (2002).

[82] K. Gurney, An Introduction to neural networ, (1997).

[83] K. Suzuki, ARTIFICIAL NEURAL NETWORKS ⊺ METHODOLOGICAL ADVANCES AND Edited by Kenji Suzuki, n.d.

[84] K. Yamazaki, V.K. Vo-Ho, D. Bulsara, N. Le, Spiking Neural Networks and Their Applications: A Review, Brain Sci. 12 (2022) 1–30. https://doi.org/10.3390/brainsci12070863.

[85] A.K. Jain, J. Mao, K.M. Mohiuddin, Artificial neural networks: A tutorial, Computer

(Long. Beach. Calif). 29 (1996) 31–44. https://doi.org/10.1109/2.485891.

[86] C. SI-MOUSSA, Modélisation des équilibres de phases a pressions élevées par les réseaux de neurones artificiels, These Dr. (2003).

[87] Y. Ammi, L. Khaouane, S. Hanini, Stacked neural networks for predicting the membranes performance by treating the pharmaceutical active compounds, Neural Comput. Appl. 33 (2021) 12429–12444. https://doi.org/10.1007/s00521-021-05876-0.

[88] A. Baghban, M.A. Ahmadi, B.H. Shahraki, Prediction carbon dioxide solubility in presence of various ionic liquids using computational intelligence approaches, J. Supercrit. Fluids. 98 (2015) 50–64. https://doi.org/10.1016/j.supflu.2015.01.002.

[89] Shodhganga Chapter 5, Adaptive Neuro Fuzzy Inference System, Https://shodhganga. inflibnet. ac. in /bits tream/10603/93079/11/11_chapter%205.pdf.

[90] G. Chen, T.T. Pham, Introduction to fuzzy systems, Introd. to Fuzzy Syst. (2005) 1–313. https://doi.org/10.1201/b15888.

[91] A.Y. Sonmez, S. Kale, R.C. Ozdemir, A.E. Kadar, An adaptive neurofuzzy inference system (NFIS) to predic of cadmium (cd) concentration in the filyos river, Turkish J. Fish. Aquat. Sci. 18 (2018) 1333–1343.

[92] K. Zarei, F. Taheri, Prediction of solubility of some statin drugs in supercritical carbon dioxide using classification and regression tree analysis and adaptive neuro-fuzzy inference systems, Russ. Chem. Bull. 65 (2016) 1131–1138. https://doi.org/10.1007/s11172-016-1424-x.

[93] R. Roem, U. Padjadjaran, Prediction of Citarum River Basin Discharge Based on ANFIS (full paper), (2016).

[94] A. Shojaeian, M. Asadizadeh, Prediction of surface tension of the binary mixtures containing ionic liquid using heuristic approaches; an input parameters investigation, J. Mol. Liq. 298 (2020) 111976. https://doi.org/10.1016/j.molliq.2019.111976.

[95] X. Hu, Support Vector Machine and Its Application to Regression and Classification, Thesis. (2017). https://bearworks.missouristate.edu/theses/3177.

[96] M. Amroune, T. Bouktir, I. Musirin, Power System Voltage Stability Assessment Using a Hybrid Approach Combining Dragonfly Optimization Algorithm and Support Vector

Regression, Arab. J. Sci. Eng. 43 (2018) 3023–3036. https://doi.org/10.1007/s13369-017-3046-5.

[97] M. Mokhtarzad, F. Eskandari, N. Jamshidi Vanjani, A. Arabasadi, Drought forecasting by ANN, ANFIS, and SVM and comparison of the models, Environ. Earth Sci. 76 (2017) 0–9. https://doi.org/10.1007/s12665-017-7064-0.

[98] C.L. Giles, PhD_Thesis_LEARNING IN EXTREME CONDITIONS_PSU_SeydaErtekin, (2009).

[99] S. Mirjalili, Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems, Neural Comput. Appl. 27 (2016) 1053–1073. https://doi.org/10.1007/s00521-015-1920-1.

[100] D.S. Palmer, N.M. O'Boyle, R.C. Glen, J.B.O. Mitchell, Random forest models to predict aqueous solubility, J. Chem. Inf. Model. 47 (2007) 150–158. https://doi.org/10.1021/ci060164k.

[101] Y. Li, C. Zou, M. Berecibar, E. Nanini-Maury, J.C.W. Chan, P. van den Bossche, J. Van Mierlo, N. Omar, Random forest regression for online capacity estimation of lithium-ion batteries, Appl. Energy. 232 (2018) 197–210. https://doi.org/10.1016/j.apenergy.2018.09.182.

[102] P. Zahedi, S. Parvandeh, A. Asgharpour, B.S. McLaury, S.A. Shirazi, B.A. McKinney, Random forest regression prediction of solid particle Erosion in elbows, Powder Technol. 338 (2018) 983–992. https://doi.org/10.1016/j.powtec.2018.07.055.

[103] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R.P. Sheridan, B.P. Feuston, Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958. https://doi.org/10.1021/ci034160g.

[104] L. Breiman, Random forests, kluwer academic publishers, machine learning, 45(2001) 5-32..

[105] J.-L. Salager, R. Marquez, J. Bullon, A. Forgiarini, Formulation in Surfactant Systems: From-Winsor-to-HLDN, Encyclopedia. 2 (2022) 778–842. https://doi.org/10.3390/encyclopedia2020054.

[106] M.J. Rosen, The relationship of structure to properties in surfactants. IV. Effectiveness

in surface or interfacial tension reduction, J. Colloid Interface Sci. 56 (1976) 320–327. https://doi.org/10.1016/0021-9797(76)90257-5.

[107] P. Becher, Hydrophile-Lipopbile Balance: History and Recent Developments Langmnir Lecture – 1983∗, Journal of Dispersion Science and Technology, 5(1), (1984) 81–96. https://doi.org/10.1080/0193269840894321.

[108] J.C. Ravey, A. Gherbi, M.J. Stébé, Comparative study of fluorinated and hydrogenated nonionic surfactants. I. Surface activity properties and critical concentrations, Trends Colloid Interface Sci. II. 241 (2007) 234–241. https://doi.org/10.1007/bfb0114199.

[109] P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, § Dinesh, O. Shah, M. Karelson, Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 1. Nonionic Surfactants, Langmuir 12 (1996) 1462-1470.

[110] A. Surfactants, P.D.T.H. Uibers, V.S.L. Obanov, A.R.K. Atritzky, D.O.S. Hah, Prediction of Critical Micelle Concentration Using a Quantitative Structure – Property Relationship Approach, J. Colloid Interface Sci. 120 (1997) 113–120.

[111] M. Jalali-Heravi, E. Konouz, Prediction of critical micelle concentration of some anionic surfactants using multiple regression techniques: A quantitative structure-activity relationship study, J. Surfactants Deterg. 3 (2000) 47–52. https://doi.org/10.1007/s11743-000-0112-5.

[112] N. Anoune, M. Nouiri, Y. Berrah, J.Y. Gauvrit, P. Lanteri, Critical micelle concentrations of different classes of surfactants: A quantitative structure property relationship study, J. Surfactants Deterg. 5 (2002) 45–53. https://doi.org/10.1007/S11743-002-0204-2.

[113] S. Yuan, Z. Cai, G. Xu, Y. Jiang, Quantitative structure-property relationships of surfactants: Critical micelle concentration of anionic surfactants, J. Dispers. Sci. Technol. 23 (2002) 465–472. https://doi.org/10.1081/DIS-120014014.

[114] Z. Wang, G. Li, X. Zhang, R. Wang, A. Lou, A quantitative structure-property relationship study for the prediction of critical micelle concentration of nonionic surfactants, Colloids Surfaces A Physicochem. Eng. Asp. 197 (2002) 37–45. https://doi.org/10.1016/S0927-7757(01)00812-3.

[115] Z.W. Wang, G.Z. Li, X.Y. Zhang, and L. Li, Prediction on critical micelle concentration

of anionic surfactants in aqueous solution: quantitative structure-property relationship approach, Acta Chimica Sinica 60 (9) (2002) 1548-1552.

[116] Z.W. Wang, G.Z. Li, X.Y. Zhang, L. Li, Prediction on critical micelle concentration of anionic surfactants in aqueous solution: Quantitative structure-property relationship approach, Acta Chim. Sin. 60 (2002) 1548–1552. https://doi.org/10.1002/cjoc.20030211212.

[117] M. Jalali-Heravi, E. Konouz, Multiple linear regression modeling of the critical micelle concentration of alkyltrimethylammonium and alkylpyridinium salts, J. Surfactants Deterg. 6 (2003) 25–30. https://doi.org/10.1007/s11743-003-0244-7.

[118] R.A. Saunders, J.A. Platts, Correlation and prediction of critical micelle concentration using polar surface area and LFER methods, J. Phys. Org. Chem. 17 (2004) 431–438. https://doi.org/10.1002/poc.749.

[119] X. Li, G. Zhang, J. Dong, X. Zhou, X. Yan, M. Luo, Estimation of critical micelle concentration of anionic surfactants with QSPR approach, J. Mol. Struct. THEOCHEM. 710 (2004) 119–126. https://doi.org/10.1016/j.theochem.2004.08.039.

[120] G. Absalan, B. Hemmateenejad, M. Soleimani, M. Akhond, R. Miri, Quantitative structure - Micellization relationship study of gemini surfactants using genetic-PLS and genetic-MLR, QSAR Comb. Sci. 23 (2004) 416–425. https://doi.org/10.1002/qsar.200430872.

[121] Z. Kardanpour, B. Hemmateenejad, T. Khayamian, Wavelet neural network-based QSPR for prediction of critical micelle concentration of Gemini surfactants, Anal. Chim. Acta. 531 (2005) 285–291. https://doi.org/10.1016/j.aca.2004.10.028.

[122] E.A. Mahmoud Gad, QSPR for nonionic surfactants, J. Dispers. Sci. Technol. 28 (2007) 231–237. https://doi.org/10.1080/01932690601058810.

[123] A.R. Katritzky, L. Pacureanu, D. Dobchev, M. Karelson, QSPR study of critical micelle concentration of anionic surfactants using computational molecular descriptors, J. Chem. Inf. Model. 47 (2007) 782–793. https://doi.org/10.1021/ci600462d.

[124] A.R. Katritzky, L.M. Pacureanu, S.H. Slavov, D.A. Dobchev, D.O. Shah, M. Karelson, QSPR study of the first and second critical micelle concentrations of cationic surfactants, Comput. Chem. Eng. 33 (2009) 321–332.

https://doi.org/10.1016/j.compchemeng.2008.09.011.

[125] D.T. Stanton, On the importance of topological descriptors in understanding structure-property relationships, J. Comput. Aided. Mol. Des. 22 (2008) 441–460. https://doi.org/10.1007/s10822-008-9204-9.

[126] A. Mozrzymas, B. Rozycka-Roszak, Prediction of Critical Micelle Concentration of Nonionic Surfactants by a Quantitative Structure – Property Relationship, Comb. Chem. High Throughput Screen. 13 (2009) 39–44. https://doi.org/10.2174/138620710790218195.

[127] C. Guo, P. Zhou, J. Shao, X. Yang, Z. Shang, Integrating statistical and experimental protocols to model and design novel Gemini surfactants with promising critical micelle concentration and low environmental risk, Chemosphere. 84 (2011) 1608–1616. https://doi.org/10.1016/j.chemosphere.2011.05.031.

[128] Z.C. Zhu, Q. Wang, Q.Z. Jia, H.M. Tang, P.S. Ma, Quantitative structure-property relationship of the critical micelle concentration of different classes of surfactants, Wuli Huaxue Xuebao/ Acta Phys. - Chim. Sin. 29 (2013) 30–34. https://doi.org/10.3866/PKU.WHXB201210265.

[129] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants, Chem. Eng. Sci. 73 (2012) 86–98. https://doi.org/10.1016/J.CES.2012.01.005.

[130] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices, 3: Modeling of critical micelle concentration of cationic surfactants, Chem. Eng. Sci. 81 (2012) 169–178. https://doi.org/10.1016/j.ces.2012.07.008.

[131] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices: Exploring effects of hydrophobicity, branching and electronic parameters on logCMC values of anionic surfactants, Chem. Eng. Sci. 87 (2013) 141–151. https://doi.org/10.1016/j.ces.2012.10.002.

[132] M. Mattei, G.M. Kontogeorgis, R. Gani, Modeling of the critical micelle concentration (CMC) of nonionic surfactants with an extended group-contribution method, Ind. Eng. Chem. Res. 52 (2013) 12236–12246. https://doi.org/10.1021/IE4016232.

[133] T. Gaudin, P. Rotureau, I. Pezron, G. Fayet, New QSPR models to predict the critical

micelle concentration of sugar-based surfactants, Ind. Eng. Chem. Res. 55 (2016) 11716–11726. https://doi.org/10.1021/acs.iecr.6b02890.

[134] A. Baghban, J. Sasanipour, M. Sarafbidabad, A. Piri, R. Razavi, On the prediction of critical micelle concentration for sugar-based non-ionic surfactants, Chem. Phys. Lipids. 214 (2018) 46–57. https://doi.org/10.1016/J.CHEMPHYSLIP.2018.05.008.

[135] Y. Wang, F. Yan, Q. Jia, Q. Wang, Quantitative structure-property relationship for critical micelles concentration of sugar-based surfactants using norm indexes, J. Mol. Liq. 253 (2018) 205–210. https://doi.org/10.1016/j.molliq.2018.01.037.

[136] A. Mozrzymas, On the Head Group Effect on Critical Micelle Concentration of Cationic Surfactants Using Molecular Connectivity Indices and Atomic Partial Charges, J. Solution Chem. 48 (2019) 875–890. https://doi.org/10.1007/s10953-019-00887-x.

[137] L. Jiao, Y. Wang, L. Qu, Z. Xue, Y. Ge, H. Liu, B. Lei, Q. Gao, M. Li, Hologram QSAR study on the critical micelle concentration of Gemini surfactants, Colloids Surfaces A Physicochem. Eng. Asp. 586 (2020). https://doi.org/10.1016/j.colsurfa.2019.124226.

[138] E. Setiawan, M. Wijaya, K. Wijaya, QSPR models for predicting critical micelle concentration of gemini cationic surfactants combining machine-learning methods and molecular descriptors.

[139] S. Rahal, N. Hadidi, M. Hamadache, In Silico Prediction of Critical Micelle Concentration (CMC) of Classic and Extended Anionic Surfactants from Their Molecular Structural Descriptors, Arab. J. Sci. Eng. 45 (2020) 7445–7454. https://doi.org/10.1007/S13369-020-04598-0.

[140] M. Turchi, A.P. Karcz, M.P. Andersson, First-principles prediction of critical micellar concentrations for ionic and nonionic surfactants, J. Colloid Interface Sci. 606 (2022) 618–627. https://doi.org/10.1016/j.jcis.2021.08.044.

[141] S. Qin, T. Jin, R.C. Van Lehn, V.M. Zavala, Predicting Critical Micelle Concentrations for Surfactants Using Graph Convolutional Neural Networks, J. Phys. Chem. B. 125 (2021) 10610–10620. https://doi.org/10.1021/acs.jpcb.1c05264.

[142] D. Abooali, R. Soleimani, Structure-based modeling of critical micelle concentration (CMC) of anionic surfactants in brine using intelligent methods, Sci. Rep. 13 (2023) 1–16. https://doi.org/10.1038/s41598-023-40466-1.

[143] Z. Wang, G.Z. Li, J.H. Mu, X.Y. Zhang, A.J. Lou, Quantitative structure-property relationships on prediction of surface tension of nonionic surfactants, Chinese chemical letters. 13 (2002) 363-366.

[144] Z.W. Wang, D.Y. Huang, G.Z. Li, X.Y. Zhang, L.L. Liao, Effectiveness of surface tension reduction by anionic surfactants - Quantitative structure-property relationships, J. Dispers. Sci. Technol. 24 (2003) 653–658. https://doi.org/10.1081/DIS-120023810.

[145] Z.W. Wang, J.L. Feng, H.J. Wang, Z.G. Cui, G.Z. Li, Effectiveness of surface tension reduction by nonionic surfactants with quantitative structure-property relationship approach, J. Dispers. Sci. Technol. 26 (2005) 441–447. https://doi.org/10.1081/DIS-200054572.

[146] M.F. Fini, S. Riahi, A. Bahramian, Experimental and QSPR studies on the effect of ionic surfactants on n-decane-water interfacial tension, J. Surfactants Deterg. 15 (2012) 477–484. https://doi.org/10.1007/s11743-012-1330-7.

[147] T. Gaudin, P. Rotureau, I. Pezron, G. Fayet, Investigating the impact of sugar-based surfactants structure on surface tension at critical micelle concentration with structure-property relationships, J. Colloid Interface Sci. 516 (2018) 162–171. https://doi.org/10.1016/j.jcis.2018.01.051.

[148] X. Guo, Z. Rong, X. Ying, Calculation of hydrophile-lipophile balance for polyethoxylated surfactants by group contribution method, J. Colloid Interface Sci. 298 (2006) 441–450. https://doi.org/10.1016/j.jcis.2005.12.009.

[149] Chen M.-L., Wang Z.-W., Zhang G.-X., Wang W.-D. (2007) Prediction on hydrophile-lipophile balance values of anionic surfactants with QSPR method, Acta Chim. Sinica 65, 1265-1272, (2007) 2007.

[150] E.A.M. Gad, K.S. Khairou, QSPR for HLB of nonionic surfactants based on polyoxyethylene group, J. Dispers. Sci. Technol. 29 (2008) 940–947. https://doi.org/10.1080/01932690701808353.

[151] F. Luan, H. Liu, Y. Gao, Q. Li, X. Zhang, Y. Guo, Prediction of hydrophile-lipophile balance values of anionic surfactants using a quantitative structure-property relationship, J. Colloid Interface Sci. 336 (2009) 773–779. https://doi.org/10.1016/j.jcis.2009.04.002.

[152] M.L. Chen, Z.W. Wang, H.J. Duan, QSPR for HLB values of nonionic surfactants using

two simple descriptors, J. Dispers. Sci. Technol. 30 (2009) 1481–1485. https://doi.org/10.1080/01932690903123338.

[153] A. Beteringhe, A.C. Răduţiu, M. Bem, R.D. Bărăţoiu, E.N. Hristea, T. Constantinescu, Partial least squares (PLS) analysis for hydrophile-lipophile balance (HLB) of some nonionic surfactants. QSPR study, Rev. Roum. Chim. 56 (2011) 347–350.

[154] P.D.T. Huibers, D.O. Shah, A.R. Katritzky, Predicting surfactant cloud point from molecular structure, J. Colloid Interface Sci. 193 (1997) 132–136. https://doi.org/10.1006/jcis.1997.5053.

[155] A.P. Bünz, B. Braun, R. Janowsky, Application of quantitative structure-performance relationship and neural network models for the prediction of physical properties from molecular structure, Ind. Eng. Chem. Res. 37 (1998) 3043–3051. https://doi.org/10.1021/ie970910y.

[156] M. Jalali-Heravi, E. Konouz, Use of Quantitative Structure-Property Relationships in Predicting the Krafft Point of Anionic Surfactants, Internet Electron. J. Mol. Des. 1 (2002) 410–417. http://www.biochempress.comhttp//www.biochempress.com.http://www.biochempress.comhttp://www.biochempress.com.

[157] Yuan, S.L.; Cai, Z.T.; Xu, G.Y.; Wang, W. Prediction of cloud poiont of nonionic surfactants using quantitative structure-property relationship method. Acta Phys. Chim. Sin. 2003, 19, 334– 337, (2003) 2003.

[158] Y.Y. Ren, H.X. Liu, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, The accurate QSPR models for the prediction of nonionic surfactant cloud point, J. Colloid Interface Sci. 302 (2006) 669–672. https://doi.org/10.1016/j.jcis.2006.06.072.

[159] J. Ghasemi, S. Ahmadi, Combination of genetic algorithm and partial least squares for cloud point prediction of nonionic surfactants from molecular structures, Annali di Chimica,. 97 (2007) 69-83.

[160] Y. Ren, B. Zhao, Q. Chang, X. Yao, QSPR modeling of nonionic surfactant cloud points: An update, J. Colloid Interface Sci. 358 (2011) 202–207. https://doi.org/10.1016/j.jcis.2011.02.042.

[161] C. Guo, Y. Ren, P. Zhou, J. Shao, X. Yang, Z. Shang, Toward a Quantitative Model and

Prediction of the Cloud Point of Normal Nonionic Surfactants and Novel Gemini Surfactants with Heuristic Method and Gaussian Process, J. Dispers. Sci. Technol. 33 (2012) 1401–1410. https://doi.org/10.1080/01932691.2011.620873.

[162] W. Chanachichalermwong, A. Charoensaeng, U. Suriyapraphadilok, Krafft Point Prediction of Anionic Surfactants Using Group Contribution Method: First-Order and Higher-Order Groups, J. Surfactants Deterg. 22 (2019) 907–919. https://doi.org/10.1002/jsde.12290.

[163] Y. Li, G. Xu, Y. Luan, S. Yuan, X. Xin, Property prediction on surfactant by quantitative structure-property relationship: Krafft point and cloud point, J. Dispers. Sci. Technol. 26 (2005) 799–808. https://doi.org/10.1081/DIS-200063127.

[164] Y. Shi, F. Yan, Q. Jia, Q. Wang, Norm descriptors for predicting the hydrophile-lipophile balance (HLB) and critical micelle concentration (CMC) of anionic surfactants, Colloids Surfaces A Physicochem. Eng. Asp. 583 (2019) 123967. https://doi.org/10.1016/j.colsurfa.2019.123967.

[165] D. Seddon, E.A. Müller, J.T. Cabral, Machine learning hybrid approach for the prediction of surface tension profiles of hydrocarbon surfactants in aqueous solution, J. Colloid Interface Sci. 625 (2022) 328–339. https://doi.org/10.1016/j.jcis.2022.06.034.

[166] C. Brozos, J.G. Rittig, S. Bhattacharya, E. Akanny, C. Kohlmann, A. Mitsos, Graph Neural Networks for Surfactant Multi-Property Prediction, Colloids Surfaces A Physicochem. Eng. Asp. 694 (2024) 134133. https://doi.org/10.1016/j.colsurfa.2024.134133.

Chater II Computational QSPR model to predict the CMC of classic and extended sugar-based surfactants

Chapter II Computational QSPR model to predict the CMC of classic and extended sugar-based surfactants

II.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 dusted: anionic (with a negative charge), cationic (with a positive charge), nonionic (uncharged), and amphoteric (zwitterions carrying both positive and negative charges) [1].

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, 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 squares, 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.

II.2 Methodology

II.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 (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 A.1) collected from previously mentioned literature [1,4,28–30], was used to create the QSPR model (70% of the data comprise the training set, whereas 30% 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.

II.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].

II.2.3 Molecular descriptor selection

An important step in the QSPR model is decreasing the number of descriptors. This decrease serves two objectives: it prevents overfitting and 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/) 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.

II.2.4 QSPR model's development and validation

Multiple linear regression (MLR), Partial least squares (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/). 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 (II.1) – (II.17) were employed, and the terms in these equations are specified appropriately:

$$Q^2_{Loo} = 1 - \frac{\sum_{i=1}^{n_{INT}} (y^{obs}_{i\,(training)} - y^{pred}_{i\,(training)})^2}{\sum_{i=1}^{n} (y^{obs}_{i\,(training)} - \bar{y}^{obs}_{(training)})^2} \quad \text{II.1}$$

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y^{obs}_{i\,(test)} - y^{pred}_{i\,(test)})^2}{\sum_{i=1}^{n} (y^{obs}_{i\,(test)} - \bar{y}^{obs}_{(training)})^2} \quad \text{II.2}$$

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y^{obs}_{i\,(test)} - y^{pred}_{i\,(test)})^2}{\sum_{i=1}^{n} (y^{obs}_{i\,(test)} - \bar{y}^{obs}_{(test)})^2} \quad \text{II.3}$$

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

$$k = \frac{\sum_{i=1}^{n} (y^{obs}_{i} \cdot y^{pred}_{i})}{\sum_{i=1}^{n} (y^{pred}_{i})^2} \quad \text{II.5}$$

$$k' = \frac{\sum_{i=1}^{n} (y^{obs}_{i} \cdot y^{pred}_{i})}{\sum_{i=1}^{n} (y^{obs}_{i})^2} \quad \text{II.6}$$

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

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

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

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

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

$$\Delta r_m^2 = |r_m^2 - r_m'^2| \quad \text{II.12}$$

$$\bar{y}^{obs} = \frac{\sum_{i=1}^{n} y_i^{obs}}{n} \quad \text{II.13}$$

$$\bar{y}^{pred} = \frac{\sum_{i=1}^{n} y_i^{pred}}{n} \quad \text{II.14}$$

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

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

$$MAE = \frac{\sum_{i=1}^{n}|y_i^{obs} - y_i^{pred}|}{n} \quad \text{II.17}$$

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, $\bar{r}_m^2$ & Δr_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 Table II.1 [37].

Table II.1 Acceptance criterion (A.C) of a model [37].

Parameter	Q^2_{ext}	R^2_{ext}	k	k'	$\bar{r}^2_m$	Δr^2_m	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 +3σ) after omitting 5% high residual compounds. An ExternalValidationPlus (http://teqip.jdvu.ac.in/QSAR_Tools/) is an online tool that may be used to compute these MAE-based external validation requirements [39].

II.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.

II.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 ±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:

$$H = X(X^tX)^{-1}X^t \qquad \text{II.18}$$

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:

$$h^* = \frac{3(p+1)}{n} \qquad \text{II.19}$$

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

II.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/). The basic theory, algorithm, and methodology, as well as the benefits of the suggested approach, are all available in the literature [44].

II.3 Results and discussion

II.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.

II.3.2 MLR model

The MLR model, generated by using 6 variables (descriptors) is illustrated by a linear equation (Eq II.20) and the statistical parameters represented in Table II.2.

pCMC = 12,09731 (± 5,88554) + 0,16059 (± 0,0293) SssCH2 + 0,17617 (± 0,06241) ATSC5e + 0,25775 (± 0,06659) AATSC4m - 7,58599 (± 2,16343) BIC3 - 6,20729 (± 2,10092) MATS1c - 4,55738 (±2,92592) SpMin1_Bhe II.20

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^2_{loo} should be less than 0.3. As part of this study, the R^2-Q^2_{loo} = 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 II.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 II.2 Validation parameters for MLR model's train, validation, test, and global set.

Statistical values	Train	Validation	Test	Global
n	83	18	18	119
$\mathbf{R^2}$	0.834	0.902	0.945	0.868
$\mathbf{Q^2_{Loo}}$	0.834	--	--	--
$\mathbf{Q^2_{F1}}$	--	0.887	0.934	--
$\mathbf{Q^2_{F2}}$	--	0.834	0.930	--
K	1.000	0.903	1.042	0.997
k'	0.979	1.089	0.949	0.983
$\mathbf{r_0^2}$	0.834	0.902	0.940	0.866
$\mathbf{r_0'^2}$	0.806	0.888	0.928	0.839
$\mathbf{r_m^2}$	0.834	0.882	0.879	0.838
$\mathbf{r_m'^2}$	0.693	0.795	0.821	0.719
$\mathbf{\bar{r}_m^2}$	0.764	0.838	0.850	0.779
$\mathbf{\Delta r_m^2}$	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 +3σ (95%data)	--	--	0.786	--

II.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 II.3 shows the statistical parameters for the prediction set.

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

Statistical values	Train	Validation	Test	Global
n	83	18	18	119
$\mathbf{R^2}$	0.783	0.832	0.973	0.832
$\mathbf{Q^2_{Loo}}$	0.782	--	--	--
$\mathbf{Q^2_{F1}}$	--	0.836	0.956	--
$\mathbf{Q^2_{F2}}$	--	0.758	0.953	--
K	0.999	0.900	1.054	0.998
k'	0.973	1.078	0.944	0.977
$\mathbf{r_0^2}$	0.783	0.828	0.968	0.832
$\mathbf{r_0'^2}$	0.737	0.821	0.963	0.797
$\mathbf{r_m^2}$	0.773	0.783	0.909	0.820
$\mathbf{r_m'^2}$	0.614	0.746	0.879	0.676

Statistical values	Train	Validation	Test	Global
$\bar{r}_m^2$	0.694	0.764	0.894	0.749
Δr_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 +3σ (95%data)	--	--	0.704	--

II.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 A.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 Figure II.1. A close correlation was found between the predicted and observed values of pCMC, except for some outlier points, considered as accepted out of range.

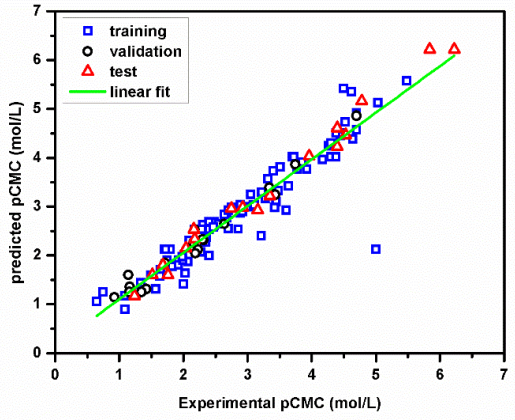


Figure II.1. Scatter plot of the predicted values of pCMC versus the experimental values by ANN model for the training, validation, and test set.

The statistical parameters for validation results in Table II.4 meet the required standards, indicating that the ANN model is robust and provides excellent predictive performance.

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

Statistical value	Train	Validation	Test	Global
n	83	18	18	119
$\mathbf{R^2}$	0.861	0.972	0.985	0.908
$\mathbf{Q^2_{Loo}}$	0.856	--	--	--
$\mathbf{Q^2_{F1}}$	--	0.978	0.982	--
$\mathbf{Q^2_{F2}}$	--	0.967	0.981	--
K	0.991	0.974	0.975	0.986
K'	0.990	1.022	1.023	0.999
$\mathbf{r^2_0}$	0.857	0.971	0.985	0.906
$\mathbf{r'^2_0}$	0.856	0.969	0.985	0.906
$\mathbf{r^2_m}$	0.804	0.939	0.976	0.873
$\mathbf{r'^2_m}$	0.802	0.913	0.982	0.864
$\mathbf{\bar{r}^2_m}$	0.803	0.926	0.979	0.868
$\mathbf{\Delta r^2_m}$	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 Figure II.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).

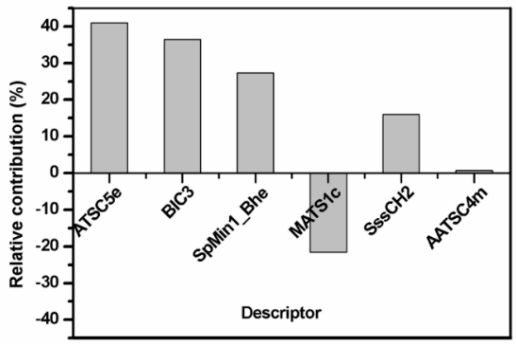


Figure II.2. Plot of the fraction contribution of the descriptors to the pCMC of sugar-based surfactants.

ATSC5e is centred broto-moreau autocorrelation of lag 5weighted 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 $-CH_2$ 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), electro-topology, and molecule symmetry are important in the micellization of sugar-based surfactants.

II.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 II.5 reports a value for various parameters used in the ANFIS model.

Table II.5 Validation parameters for ANFIS model's train, test, and global set.

Statistical value	Train	Test	Global
N	101	18	119
$\mathbf{R^2}$	0.935	0.916	0.932
$\mathbf{Q^2_{Loo}}$	0.935	--	--
$\mathbf{Q^2_{F1}}$	--	0.917	--
$\mathbf{Q^2_{F2}}$	--	0.894	--
K	1.000	0.972	0.994
K'	0.990	1.018	0.996
$\mathbf{r_0^2}$	0.935	0.902	0.931
$\mathbf{r_0'^2}$	0.931	0.915	0.930
$\mathbf{r_m^2}$	0.935	0.806	0.911
$\mathbf{r_m'^2}$	0.877	0.884	0.893
$\mathbf{\bar{r}_m^2}$	0.906	0.845	0.902
$\mathbf{\Delta r_m^2}$	0.058	0.079	0.019
CCC	--	0.952	0.965
MAE	0.089	0.270	0.116
RMSE	0.299	0.384	0.314
MAE (95%data)	--	0.227	--
MAE+3 σ (95%data)	--	0.882	--

II.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 II.2, II.3, II.4, and II.5), 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.

II.3.7 Applicability domain

The Williams plot and standardization technique were used to examine the applicability domain of the ANN model. In Figure II.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.

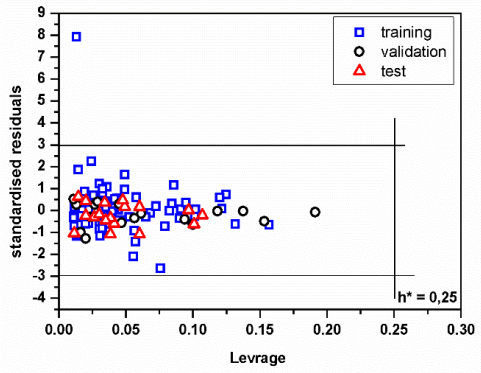


Figure II.3. Projection of the training, validation, and test set of sugar-based surfactants in the Williams plot.

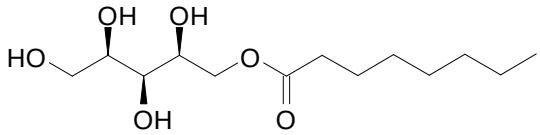
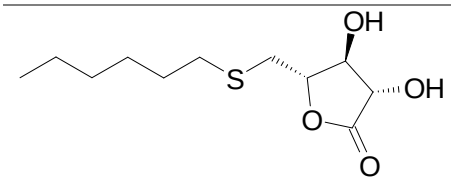
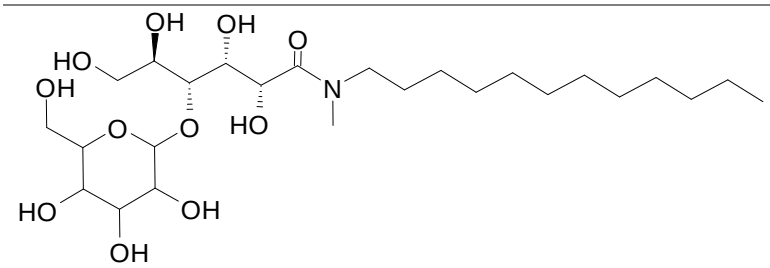
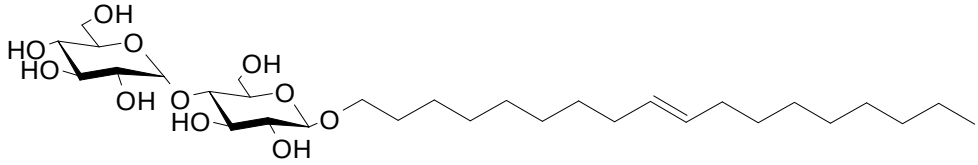
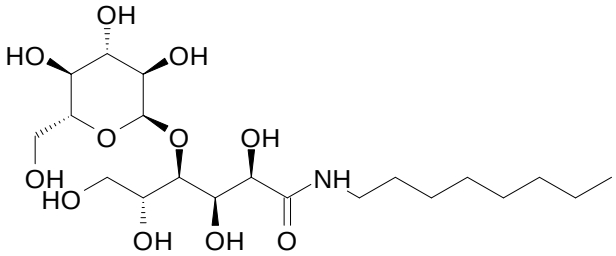
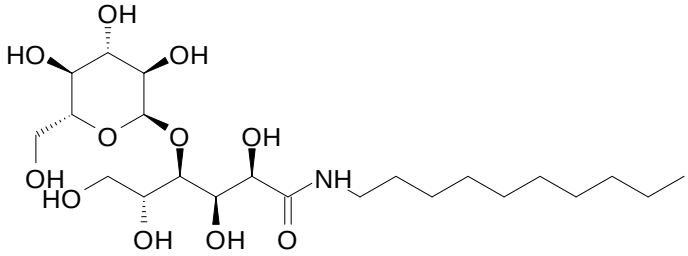
II.3.8 Comparison with previously reported models

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 II.6). 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 II.7. 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.

Table II.6 Results of our best model's (ANN) internal and external validation compared with those from previous research.

MODELS	N_{total}	Training				Validation				Test				
		R^2	Q^2_{Loo}	$\overline{r_m^2}$	Δr_m^2	R^2	Q^2_{F1}	$\overline{r_m^2}$	Δr_m^2	R^2	Q^2_{F1}	Δr_m^2	*MAE* (95%data)	*MAE*+3σ (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
Baghban et al.	83	0.999	--	--	--	--	---	--	--	0.997	--	--	--	--

Table II.7 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

II.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 Δr_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 was not contained in the data set.

Bibliographic References

[1] T. Gaudin, P. Rotureau, I. Pezron, G. Fayet, New QSPR models to predict the critical micelle concentration of sugar-based surfactants, Ind. Eng. Chem. Res. 55 (2016) 11716–11726. https://doi.org/10.1021/acs.iecr.6b02890.

[2] M.J. Rosen, Surfactants and interfacial phenomena Third edition, John Wiley & sons (2004).

[3] Y. Wang, F. Yan, Q. Jia, Q. Wang, Quantitative structure-property relationship for critical micelles concentration of sugar-based surfactants using norm indexes, J. Mol. Liq. 253 (2018) 205–210. https://doi.org/10.1016/j.molliq.2018.01.037.

[4] S. Iglauer, Y. Wu, P. Shuler, Y. Tang, W.A. Goddard, Analysis of the influence of alkyl polyglycoside surfactant and cosolvent structure on interfacial tension in aqueous formulations versus n-octane, Tenside, Surfactants, Deterg. 47 (2010) 87–97. https://doi.org/10.3139/113.110056.

[5] K. Shinoda, T. Yamanaka, K. Kinoshita, Surface chemical properties in aqueous solutions of non-ionic surfactants: Octyl glycol ether, α-octyl glyceryl ether and octyl glucoside, J. Phys. Chem. 63 (1959) 648–650. https://doi.org/10.1021/J150575A003.

[6] P. Liljekvist, M. Kjellin, J. Christer Eriksson, Surface pressure effect of pentaoxyethylene and maltoside surfactant head groups, Adv. Colloid Interface Sci. 89–90 (2001) 293–302. https://doi.org/10.1016/S0001-8686(00)00055-5.

[7] H.G. Bazin, T. Polat, R.J. Linhardt, Synthesis of sucrose-based surfactants through regioselective sulfonation of acylsucrose and the nucleophilic opening of a sucrose cyclic sulfate, Carbohydr. Res. 309 (1998) 189–205. https://doi.org/10.1016/S0008-6215(98)00121-9.

[8] M. Kjellin, I. Johansson, Surfactants from Renewable Resources, Surfactants from Renew. Resour. (2010). https://doi.org/10.1002/9780470686607.

[9] S. Matsumura, K. Imai, S. Yoshikawa, K. Kawada, T. Uchibor, Surface activities, biodegradability and antimicrobial properties of n-alkyl glucosides, mannosides and galactosides, J. Am. Oil Chem. Soc. 1990 6712. 67 (1990) 996–1001.

https://doi.org/10.1007/BF02541865.

[10] O.J. Rojas, C. Stubenrauch, L.A. Lucia, Y. Habibi, Interfacial Properties of Sugar-based Surfactants, (2009) 475-480.

[11] P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, § Dinesh, O. Shah, M. Karelson, Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 1. Nonionic Surfactants, Langmuir 12 (1996) 1462-1470.

[12] J.N. Israelachvili, D.J. Mitchell, B.W. Ninham, Theory of self-assembly of hydrocarbon amphiphiles into micelles and bilayers, J. Chem. Soc. Faraday Trans. 2 Mol. Chem. Phys. 72 (1976) 1525–1568. https://doi.org/10.1039/F29767201525.

[13] A. Baghban, J. Sasanipour, M. Sarafbidabad, A. Piri, R. Razavi, On the prediction of critical micelle concentration for sugar-based non-ionic surfactants, Chem. Phys. Lipids. 214 (2018) 46–57. https://doi.org/10.1016/J.CHEMPHYSLIP.2018.05.008.

[14] J. Hu, X. Zhang, Z. Wang, A review on progress in QSPR studies for surfactants, Int. J. Mol. Sci. 11 (2010) 1020–1047. https://doi.org/10.3390/ijms11031020.

[15] D. Myers, Surfactant science and technology, Third edition, John Wiley & sons (2006).

[16] P. Lecomte Du Noüy, AN INTERFACIAL TENSIOMETER FOR UNIVERSAL USE, J. Gen. Physiol. 7 (1925) 625. https://doi.org/10.1085/JGP.7.5.625.

[17] G. Fayet, P. Rotureau, How to use QSPR-type approaches to predict properties in the context of Green Chemistry, Biofuels, Bioprod. Biorefining. 10 (2016) 738–752. https://doi.org/10.1002/BBB.1723.

[18] B. Creton, C. Nieto-Draghi, N. Pannacci, Prediction of Surfactants' Properties using Multiscale Molecular Modeling Tools: A Review, Oil Gas Sci. Technol. – Rev. d'IFP Energies Nouv. 67 (2012) 969–982. https://doi.org/10.2516/OGST/2012040.

[19] C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.C. De Hemptinne, P. Ungerer, B. Rousseau, C. Adamo, A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes, Chem. Rev. 115 (2015) 13093–13164. https://doi.org/10.1021/ACS.CHEMREV.5B00215.

[20] A.R. Katritzky, L.M. Pacureanu, S.H. Slavov, D.A. Dobchev, M. Karelson, QSPR study of critical micelle concentrations of nonionic surfactants, Ind. Eng. Chem. Res. 47

(2008) 9687–9695. https://doi.org/10.1021/IE800954K.

[21] R.A. Saunders, J.A. Platts, Correlation and prediction of critical micelle concentration using polar surface area and LFER methods, J. Phys. Org. Chem. 17 (2004) 431–438. https://doi.org/10.1002/poc.749.

[22] A. Mozrzymas, B. Rozycka-Roszak, Prediction of Critical Micelle Concentration of Nonionic Surfactants by a Quantitative Structure – Property Relationship, Comb. Chem. High Throughput Screen. 13 (2009) 39–44. https://doi.org/10.2174/138620710790218195.

[23] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants, Chem. Eng. Sci. 73 (2012) 86–98. https://doi.org/10.1016/J.CES.2012.01.005.

[24] M. Mattei, G.M. Kontogeorgis, R. Gani, Modeling of the critical micelle concentration (CMC) of nonionic surfactants with an extended group-contribution method, Ind. Eng. Chem. Res. 52 (2013) 12236–12246. https://doi.org/10.1021/IE4016232.

[25] N. Anoune, M. Nouiri, Y. Berrah, J.Y. Gauvrit, P. Lanteri, Critical micelle concentrations of different classes of surfactants: A quantitative structure property relationship study, J. Surfactants Deterg. 5 (2002) 45–53. https://doi.org/10.1007/S11743-002-0204-2.

[26] M. Hamadache, O. Benkortbi, S. Hanini, A. Amrane, QSAR modeling in ecotoxicological risk assessment: application to the prediction of acute contact toxicity of pesticides on bees (Apis mellifera L.), Environ. Sci. Pollut. Res. 25 (2018) 896–907. https://doi.org/10.1007/S11356-017-0498-9.

[27] S. Rahal, N. Hadidi, M. Hamadache, In Silico Prediction of Critical Micelle Concentration (CMC) of Classic and Extended Anionic Surfactants from Their Molecular Structural Descriptors, Arab. J. Sci. Eng. 45 (2020) 7445–7454. https://doi.org/10.1007/S13369-020-04598-0.

[28] G. Lemahieu, J. Aguilhon, H. Strub, V. Molinier, J.F. Ontiveros, J.M. Aubry, Hexahydrofarnesyl as an original bio-sourced alkyl chain for the preparation of glycosides surfactants with enhanced physicochemical properties, RSC Adv. 10 (2020) 16377–16389. https://doi.org/10.1039/d0ra02326d.

[29] T. Gaudin, H. Lu, G. Fayet, A. Berthauld-Drelich, P. Rotureau, G. Pourceau, A. Wadouachi, E. Van Hecke, A. Nesterenko, I. Pezron, Impact of the chemical structure on amphiphilic properties of sugar-based surfactants: A literature overview, Adv. Colloid Interface Sci. 270 (2019) 87–100. https://doi.org/10.1016/j.cis.2019.06.003.

[30] C. Scorzza, P. Godé, G. Goethals, P. Martin, M. Miñana-Pérez, J.L. Salager, A. Usubillaga, P. Villa, Another new family of "extended" glucidoamphiphiles. Synthesis and surfactant properties for different sugar head groups and spacer arm lengths, J. Surfactants Deterg. 5 (2002) 337–343. https://doi.org/10.1007/S11743-002-0231-Z.

[31] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474. https://doi.org/10.1002/jcc.21707.

[32] K. Roy, On some aspects of validation of predictive quantitative structure-activity relationship models, Expert Opin. Drug Discov. 2 (2007) 1567–1577. https://doi.org/10.1517/17460441.2.12.1567.

[33] N. Chirico, P. Gramatica, Real external predictivity of QSAR models: How to evaluate It? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, J. Chem. Inf. Model. 51 (2011) 2320–2335. https://doi.org/10.1021/CI200211N.

[34] P. Gramatica, A. Sangion, A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology, J. Chem. Inf. Model. 56 (2016) 1127–1131. https://doi.org/10.1021/ACS.JCIM.6B00088.

[35] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring rm2 metrics for validation of QSPR models, Chemom. Intell. Lab. Syst. 107 (2011) 194–205. https://doi.org/10.1016/J.CHEMOLAB.2011.03.011.

[36] R. Ahmadi, B. Sepehri, R. Ghavami, 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. 39 (2019) 264–275. https://doi.org/10.1080/10799893.2019.1660898.

[37] N. Chirico, P. Gramatica, Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot

inspection, J. Chem. Inf. Model. 52 (2012) 2044–2058. https://doi.org/10.1021/CI300084J.

[38] K. Roy, R.N. Das, P. Ambure, R.B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, Chemom. Intell. Lab. Syst. 152 (2016) 18–33. https://doi.org/10.1016/J.CHEMOLAB.2016.01.008.

[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/.

[40] J.S. N'dri, B. Ouattara, M.G.-R. Koné, A.L.C. Kablan, G.S. Dembélé, C.G. Kodjo, N. Ziao, Quantitative Structure-Activity Study against Plasmodium falciparum of a Series of Derivatives of Azetidine-2-Carbonitriles by the Method of Density Functional Theory, Mediterr. J. Chem. 11 (2021) 162. https://doi.org/10.13171/MJC02103241572MGRK.

[41] OECD: Principles for the validation, for regulatory purposes, of (quantitative) structure–activity relationship models (2009).

[42] A. Rybinska, A. Sosnowska, M. Barycki, T. Puzyn, Geometry optimization method versus predictive ability in QSPR modeling for ionic liquids, J. Comput. Aided. Mol. Des. 30 (2016) 165–176. https://doi.org/10.1007/s10822-016-9894-3.

[43] T. Asadollahi, S. Dadfarnia, A.M.H. Shabani, J.B. Ghasemi, M. Sarkhosh, 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. 16 (2011) 1928–1955. https://doi.org/10.3390/molecules16031928.

[44] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemom. Intell. Lab. Syst. 145 (2015) 22–29. https://doi.org/10.1016/J.CHEMOLAB.2015.04.013.

[45] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77. https://doi.org/10.1002/QSAR.200390007.

[46] A. Golbraikh, M. Shen, Z. Xiao, Y. De Xiao, K.H. Lee, A. Tropsha, Rational selection

of training and test sets for the development of validated QSAR models, J. Comput. Mol. Des. 2003 172. 17 (2003) 241–253. https://doi.org/10.1023/A:1025386326946.

[47] S.D. Sawant, A.G. Nerkar, N.D. Pawar, A. V Velapure, design, synthesis, qsar studies and biological evaluation of novel triazolopiperazine based β-amino amides as dipeptidyl peptidase-iv (dpp-iv) inhibitors: part-II, Int. J. Pharmacy. Pharmaceutical. Sci. 6 (2014) 812-817.

[48] M. Hamadache, O. Benkortbi, S. Hanini, A. Amrane, L. Khaouane, C. Si Moussa, A Quantitative Structure Activity Relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction, J. Hazard. Mater. 303 (2016) 28–40. https://doi.org/10.1016/J.JHAZMAT.2015.09.021.

[49] J. OLDEN, An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data, Ecol. Modell. (2004). https://doi.org/10.1016/S0304-3800(04)00156-5.

[50] E. Buyukbingol, A. Sisman, M. Akyildiz, F.N. Alparslan, A. Adejare, 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. 15 (2007) 4265–4282. https://doi.org/10.1016/J.BMC.2007.03.065.

[51] G.J. Lavado, D. Baderna, E. Carnesecchi, A.P. Toropova, A.A. Toropov, J.L.C.M. Dorne, E. Benfenati, 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. 423 (2022) 127236. https://doi.org/10.1016/j.jhazmat.2021.127236.

Chapter III
QSPR for the prediction of CMC of different classes of surfactants using machine learning algorithms

Chapter III QSPR for the prediction of CMC of different classes of surfactants using machine learning algorithms

III.1 Introduction

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

Within the wide range of methodologies found in the literature for predicting substance properties by understanding the chemical structure, QSPR modeling is a significant field of research in computational chemistry [9] because it can provide a faster, more accurate, and less expensive method for understanding and measuring structural characteristics that affect

physical property. In the QSPR framework, a wide range of molecular descriptors effectively capture the intricate details of the molecular structure, such as topological, constitutional, geometrical, electronic, and more [8]. Therefore, a chosen set of descriptors is statistically associated with the studied experimental property, producing a mathematical model that can be used for discovering valuable correlations between structure and property.

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

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

III.2 Methodology

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

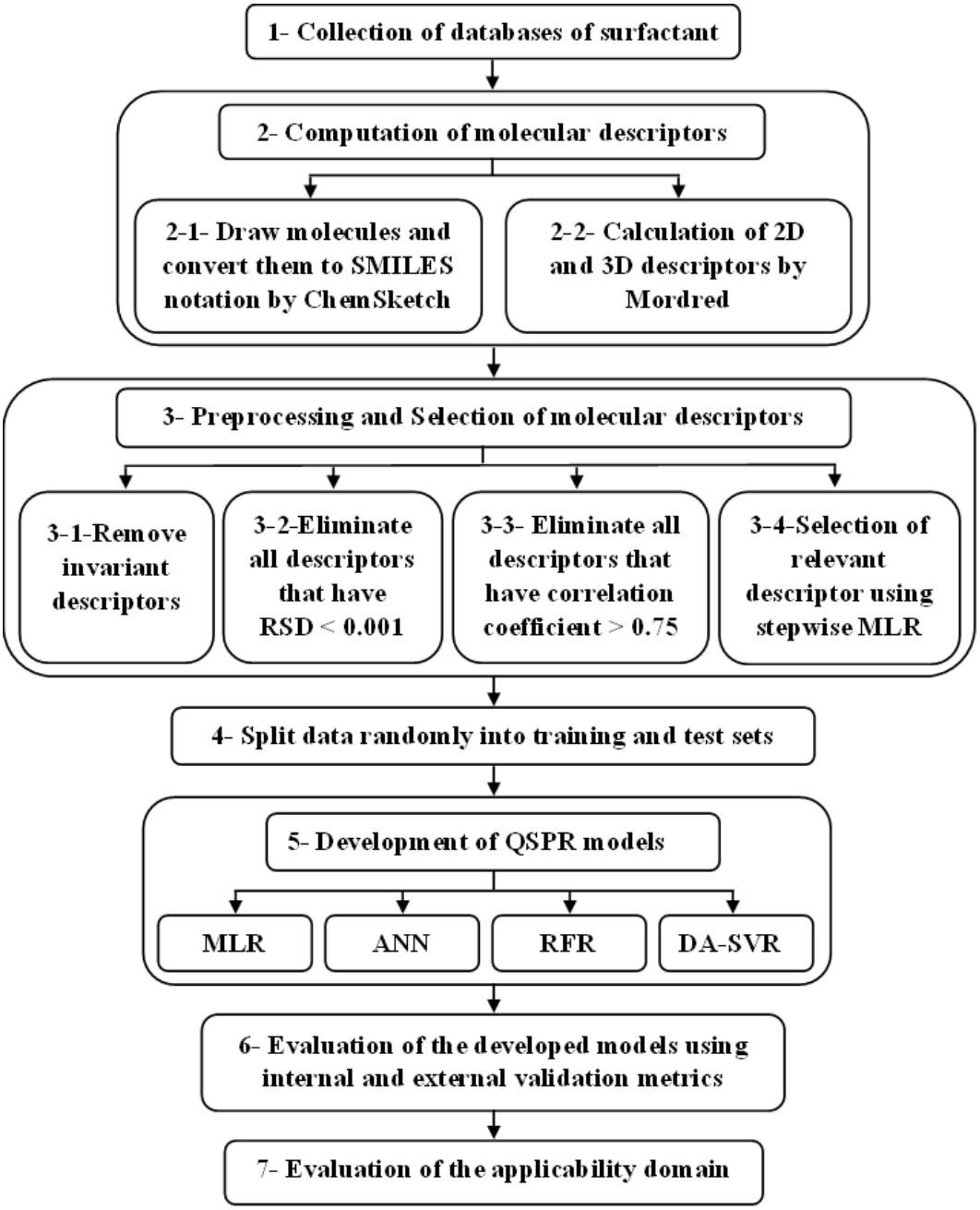


Figure III.1. General steps for the generation of QSPR models in the current study.

III.2.1 Data collection

A total of 593 distinct classes of surfactants, including anionic, cationic, nonionic, zwitterionic, and Gemini surfactants, were collected from a previous publication [2,10,25–29,17–24]. These surfactant CMCs were measured at temperatures ranging from 10°C to 60 °C. To ensure a normal distribution, CMC values were transformed into the negative logarithmic form of pCMC (in mol/L), as shown in Figure III.2. The surfactant compounds and their associated experimental pCMC values are listed in the Annex as Table A2.

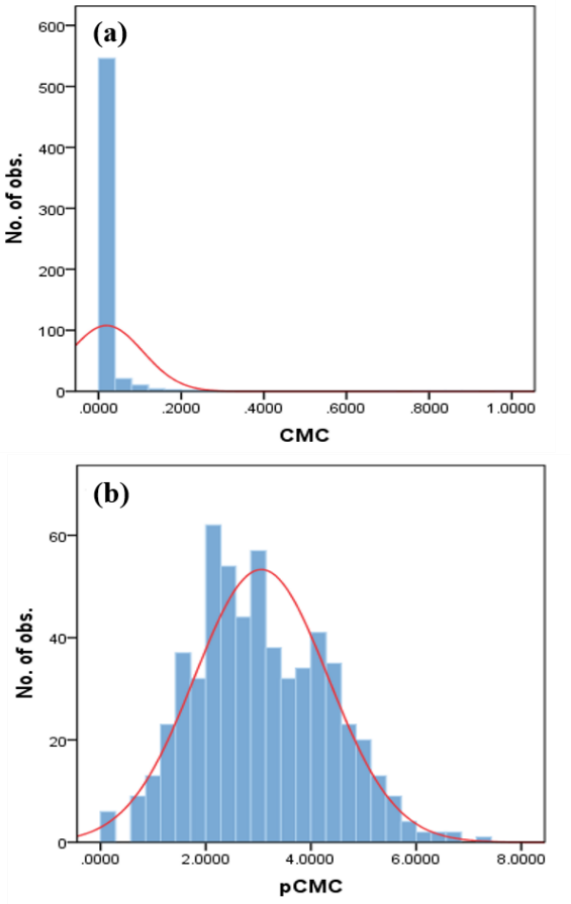


Figure III.2. Normality distribution plot of CMC data: (a) before transformation, (b) after log transformation (pCMC).

III.2.2 Molecular descriptors calculation

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

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

III.2.3 Selection of relevant descriptors

To ensure the relevance of molecular descriptors, it is necessary to implement a feature selection method. First, we excluded invariant descriptors that had missing values. Any descriptors with zero values are also removed. Then, descriptors with a relative standard

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

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

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

III.2.4 Model development

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

learning were used namely: MLR, ANN, RFR, and DA-SVR to build models. Both DA-SVR and ANN were implemented by MATLAB® R 2018a. RFR was implemented by Scikit_learn Python toolkit. While MLR was implemented by the MLRPlusValidation tool (http://dtclab.webs.com/).

III.2.5 Model validation

Various significant statistical parameters and applicability domains were applied to evaluate the efficacity of the model. The metrics assessed were the coefficient of determination (R^2), different regression metrics (Q^2, Q^2_{F1}, Q^2_{F2}), average and delta of r^2_m ($\bar{r}^2_m, \Delta r^2_m$), concordance correlation coefficient (*CCC*) to evaluate reproducibility, root mean squared error (*RMSE*), mean absolute error (*MAE*), average absolute relative deviation (*AARD*) and Akaike's information criterion (*AIC*) [33], which is an additional statistical parameter that is utilized to compare different models and determine which model best represents the experimental results. A low *AIC* value suggests a better-fitting model. The statistical parameters' formulas are available in the literature [34–37].

III.3 Results and discussion

III.3.1 MLR predictive model

The MLR model's performance depends on the size and quality of the dataset. For a reliable model, the ratio of datasets to variables must exceed 5 [38]. The linear equation representing the generated MLR model is given below.

pCMC = 2.12274(± 1.0075) + 0.33915(± 0.02052) SLogP + 3.45516(± 0.38652) GATS6d - 0.17197(± 0.06904) nAcid - 1.86614(± 0.40426) GATS8dv - 0.25316(± 0.37152) GATS4dv + 0.02768(± 0.01231) PEOE_VSA11 + 1.69493(± 0.35779) GATS8d + 0.00014(± 0.00014) ATS0p + 0.07065(± 0.42478) GATS1d + 1.03934(± 1.5928) MATS5p - 1.34056(± 0.43476) GATS3d - 0.10528(± 0.08413) NdssC - 1.35554(± 0.33952) GATS6dv + 0.02183(± 0.00494) EState_VSA4 - 0.01049(± 0.00714) T III.1

Where

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

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

III.3.2 ANN predictive model

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

III.3.2.1 Mathematical equation application for the QSPR-ANN model

The network architecture developed in this study is a multilayer perceptron with a structure of {15,10,1}. So, the network includes fifteen inputs (X_i, *i*= 1 to 15), one output (Z), and a hidden layer of ten neurons. We used hyperbolic tangent and linear functions as activation functions, which are mathematically defined in Eqs. (III.2) and (III.3), respectively.

$$f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \text{III.2}$$

$$f(x)=x \quad \text{III.3}$$

The output signal of each hidden neuron (Y_j) is calculated by Eq. (III.4), and the output of the network is given by Eq. (III.5).

$$Y_j=f\left[\sum_{i=1}^{15} w_{i,j}X_i+b_j\right]=\frac{exp\left(\sum_{i=1}^{15} w_{i,j}X_i+b_j\right)-exp\left(-\sum_{i=1}^{15} w_{i,j}X_i+b_j\right)}{exp\left(\sum_{i=1}^{15} w_{i,j}X_i+b_j\right)+exp\left(-\sum_{i+1}^{15} w_{i,j}X_i+b_j\right)} \quad \text{III.4}$$

$$Z=f\left[\sum_{j=1}^{10} w_{1,j}Y_j+b_1\right]=\sum_{j=1}^{10} w_{1,j}Y_j+b_1 \quad \text{III.5}$$

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

The final mathematical formula for predicting the critical micelle concentration of different classes of surfactants derived from the ANN approach is given in Eq. (III.5). The values of the weights and bias for each layer are given in the Annex as Table A3 and Table A4, respectively. Table III.5 shows the predicted values for 11 surfactants that were not included in the construction of the QSPR models. These predictions were based on Eq. (III.5).

III.3.3 RFR predictive model

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

III.3.4 DA-SVR predictive model

The optimal value of the SVR hyperparameter was found using the dragonfly algorithm. Our model used the radial basis function (rbf) as a kernel function. The highest performance of the DA-SVR model was obtained with the following optimal hyperparameter values: regularization constant (C = 200), insensitivity function (ε = 1.9939), and Gaussian function parameter (γ = 0.0080). The predicted pCMC values for all 593 surfactants are presented in the Annex as Table A1. The scatter plot shown in Figure III.3, shows a strong correlation between the experimental values and the predicted results obtained by the DA-SVR. In the training set, this model achieved statistical parameters of R^2=0.9712, Q^2=0.9710, $\bar{r}_m^2$=0.9581, Δr_m^2=0.0222, *MAE*=0.0871, *RMSE*=0.2171. Table III.2 provides an overview of the external criteria of the DA-SVR model.

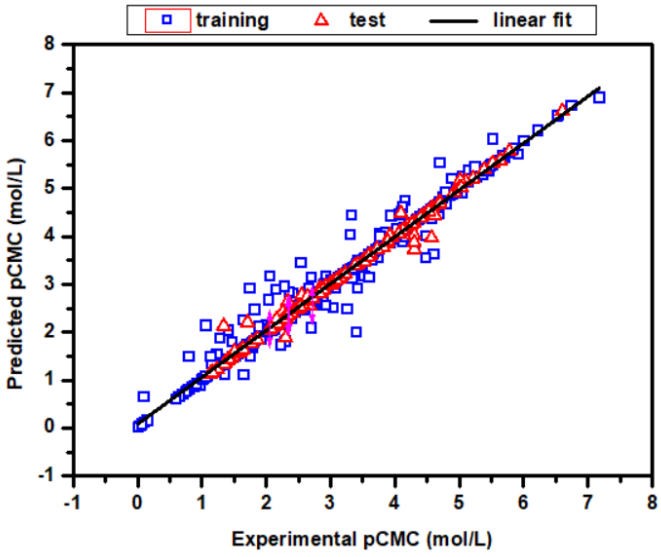


Figure III.3. Scatter plots Predicted vs. Observed values of pCMC for the DA-SVR for training and test sets.

III.3.5 Comparison of four statistical models

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

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

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

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

Validation metrics	MLR	ANN	RFR	DA-SVR
$\boldsymbol{R^2}$	0.5024	0.9247	0.9418	0.9740
$\boldsymbol{Q^2}$	0.5021	0.9247	0.9365	0.9739
$\boldsymbol{r^2_0}$	0.5021	0.9247	0.9366	0.9740
$\boldsymbol{r^2_m}$	0.4937	0.9247	0.8739	0.9740
$\boldsymbol{\bar{r}^2_m}$	0.3478	0.8898	0.8819	0.9627
$\boldsymbol{\Delta r^2_m}$	0.3085	0.0700	0.1095	0.0244
CCC	0.6743	0.9607	0.9649	0.9868
MAE	0.6991	0.2229	0.2022	0.0810
RMSE	0.8939	0.3476	0.3192	0.2047
AARD (%)	81.1891	10.7150	24.2263	4.4447
AIC	-102.2222	-1222.3	-1323.3	-1850.05

III.3.6 Applicability domain

Once the model is validated, it is crucial to determine its applicability domain as defined by the third OECD principle. This step ensures the reliability of the model and defines the range of samples whose predictions can be considered accurate. The Williams plot [39,40] was used in this study to establish the applicability domain of the DA-SVR model. As shown in Figure III.4, The threshold leverage computed, denoted by h^*, equals 0.101. A total of 569 compounds were found in both the training and test sets within the domain, and only 24 compounds were identified as outliers, where 8 compounds are outside the applicability domain (h> h*) and 16 compounds are outside the range of the ± standard deviation unit. Thus, an accuracy rate of 95.95% was achieved through predictions within the applicability domain. Fortunately, the DA-SVR model produced accurate predictions for these compounds.

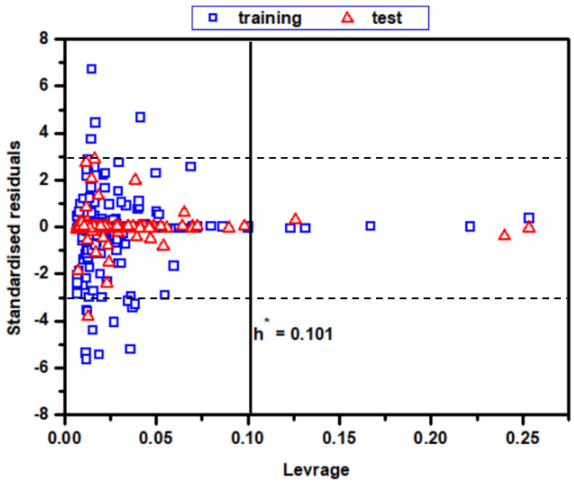


Figure III.4. William's plot for the DA-SVR model.

III.3.7 Contribution of the selected descriptors

The set of 15 features obtained from feature pre-screening was used to build our predictive models. According to Table III.1, the selected descriptors have been categorized into five distinct groups according to their properties and characteristics. The relevance factor [41] method was utilized to assess the impact of each of the 15 descriptors of the DA-SVR model in predicting surfactant CMC. Figure III.5 demonstrates that the CMC is directly influenced by descriptors known as SLogP, GATS6d, GATS8d, MATS5p, GATS3d, GATS6dv, and EState_VSA4. Conversely, the temperature and other descriptors exhibit an inverse relationship with CMC. Moreover, the greatest relevance input variables were SLogP and

GATS4dv, with relevance factors of +0.57 and -0.18, respectively. GATS3d and NdssC showed the least effect, with a relevance factor of +0.01 and -0.02, respectively.

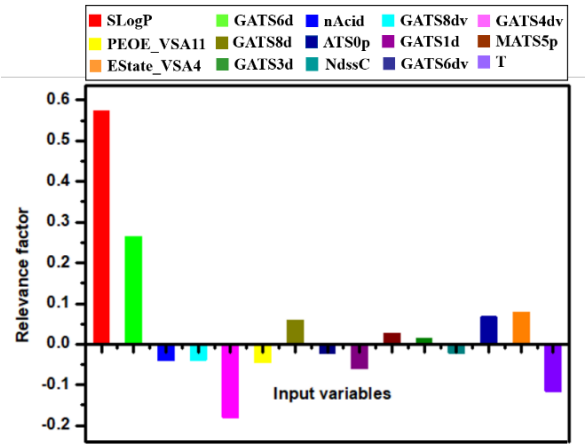


Figure III.5. The results of the relevance factor analysis carried out on the DA-SVR model.

III.3.8 Comparison with previously reported models

As part of model validation, our DA-SVR model for predicting the pCMC of 593 surfactants has been compared to a previously published model in the literature. Comparing the current study with previous ones is difficult due to differences in data sets and modeling methods involving molecular descriptors and algorithms. It is important to note that the majority of these QSPR models have exclusively used a single class of surfactant. For instance, Wang et al. [42] considered nonionic surfactants, Rahal et al. [43] focused on anionic surfactants, Guo et al. [21] considered Gemini surfactants, and Roy et al. [44] explored cationic surfactants. However, Qin et al.'s [15] and Anoune et al.'s [16] studies were exceptional as they encompassed anionic, cationic, nonionic, and zwitterionic surfactants. Table III.4 displays the statistical parameters for the results obtained in this study and previous studies found in the literature. All of these models have the potential to provide excellent prediction abilities. However, our model surpasses all previously published models in all statistical metrics available for comparison, with a higher coefficient correlation and a lower root mean squared error. Furthermore, we used our ANN-based equation and DA-SVR models to predict the pCMC of 11 surfactants. These surfactants were not included in the construction of the QSPR models. The results are shown in Table III.5. The findings indicate that the DA-SVR model is more consistent with the actual data compared to the ANN model.

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

Models	QSPR method	Data points	Internal validation				External validation			
			R^2	Q^2	$\bar{r}_m^2$	Δr_m^2	R^2	Q_{ext}^2	*MAE*	*RMSE*
This study	**DA-SVR**	**593**	**0.971**	**0.971**	**0.958**	***0.087***	**0.986**	**0.986**	**0.056**	**0.144**
Qin et al. [15]	GCN	202	-	-	-	-	0.920	-	-	0.300
Anoune et al [16]	PLS	49	0.900	-	-	-	-	-	-	-
wang et al.[42]	MLR	83	0.959	0.947	-	-	0.946	-	-	-
Rahal et al.[43]	MLP-ANN	50	0.940	0.930	0.890	0.060	-	0.950	-	-
Guo et al. [21]	GA-LSSVM	120	0.964	0.917	-	-	0.855	-	-	0.500
Roy et al.[44]	GFA	35	0.949	0.893	0.851	0.020	0.904	-	-	-

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

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

III.4 Conclusion

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

Bibliographic References

[1] K. Holmberg, B. Jönsson, B. Kronberg, B. Lindman, Surfactants and Polymers in Aqueous Solution, 2002. https://doi.org/10.1002/0470856424.

[2] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices: Exploring effects of hydrophobicity, branching and electronic parameters on logCMC values of anionic surfactants, Chem. Eng. Sci. 87 (2013) 141–151. https://doi.org/10.1016/j.ces.2012.10.002.

[3] L. Jiao, Y. Wang, L. Qu, Z. Xue, Y. Ge, H. Liu, B. Lei, Q. Gao, M. Li, Hologram QSAR study on the critical micelle concentration of Gemini surfactants, Colloids Surfaces A Physicochem. Eng. Asp. 586 (2020). https://doi.org/10.1016/j.colsurfa.2019.124226.

[4] D. Myers, Surfactant science and technology, Third edition, John Wiley & sons, (2006).

[5] M.J.L. Castro, C. Ojeda, A.F. Cirelli, Advances in surfactants for agrochemicals, Environ. Chem. Lett. 12 (2014) 85–95. https://doi.org/10.1007/s10311-013-0432-4.

[6] K. Hill, O. Rhode, Sugar-based surfactants for consumer products and technical applications, Lipid - Fett. 101 (1999) 25–33. https://doi.org/10.1002/(sici)1521-4133(19991)101:1<25::aid-lipi25>3.0.co;2-n.

[7] A.R. Katritzky, L.M. Pacureanu, S.H. Slavov, D.A. Dobchev, M. Karelson, QSPR study of critical micelle concentrations of nonionic surfactants, Ind. Eng. Chem. Res. 47 (2008) 9687–9695. https://doi.org/10.1021/IE800954K.

[8] X. Li, G. Zhang, J. Dong, X. Zhou, X. Yan, M. Luo, Estimation of critical micelle concentration of anionic surfactants with QSPR approach, J. Mol. Struct. THEOCHEM. 710 (2004) 119–126. https://doi.org/10.1016/j.theochem.2004.08.039.

[9] A.R. Katritzky, U. Maran, V.S. Lobanov, M. Karelson, Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties, J. Chem. Inf. Comput. Sci. 40 (2000) 1–18. https://doi.org/10.1021/ci9903206.

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

11716–11726. https://doi.org/10.1021/acs.iecr.6b02890.

[11] A. Baghban, J. Sasanipour, M. Sarafbidabad, A. Piri, R. Razavi, On the prediction of critical micelle concentration for sugar-based non-ionic surfactants, Chem. Phys. Lipids. 214 (2018) 46–57. https://doi.org/10.1016/J.CHEMPHYSLIP.2018.05.008.

[12] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants, Chem. Eng. Sci. 73 (2012) 86–98. https://doi.org/10.1016/J.CES.2012.01.005.

[13] P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, § Dinesh, O. Shah, M. Karelson, Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 1. Nonionic Surfactants, Langmuir 12 (1996) 1462-1470.

[14] A.R. Katritzky, L. Pacureanu, D. Dobchev, M. Karelson, QSPR study of critical micelle concentration of anionic surfactants using computational molecular descriptors, J. Chem. Inf. Model. 47 (2007) 782–793. https://doi.org/10.1021/ci600462d.

[15] S. Qin, T. Jin, R.C. Van Lehn, V.M. Zavala, Predicting Critical Micelle Concentrations for Surfactants Using Graph Convolutional Neural Networks, J. Phys. Chem. B. 125 (2021) 10610–10620. https://doi.org/10.1021/acs.jpcb.1c05264.

[16] N. Anoune, M. Nouiri, Y. Berrah, J.Y. Gauvrit, P. Lanteri, Critical micelle concentrations of different classes of surfactants: A quantitative structure property relationship study, J. Surfactants Deterg. 5 (2002) 45–53. https://doi.org/10.1007/S11743-002-0204-2.

[17] Z. qiang He, M. jun Zhang, Y. Fang, G. yong Jin, J. Chen, Extended surfactants: A well-designed spacer to improve interfacial performance through a gradual polarity transition, Colloids Surfaces A Physicochem. Eng. Asp. 450 (2014) 83–92. https://doi.org/10.1016/j.colsurfa.2014.03.012.

[18] A. Fernández, C. Scorzza, A. Usubillaga, J.L. Salager, Synthesis of new extended surfactants containing a carboxylate or sulfate polar group, J. Surfactants Deterg. 8 (2005) 187–191. https://doi.org/10.1007/s11743-005-346-2.

[19] I.J. Lin, L. Marszall, CMC, HLB, and effective chain length of surface-active anionic and cationic substances containing oxyethylene groups, J. Colloid Interface Sci. 57 (1976) 85–93. https://doi.org/10.1016/0021-9797(76)90178-8.

[20] L.H. Lin, Y.C. Lai, K.M. Chen, H.M. Chang, Oxyethylene chain length affects the physicochemical properties of sugar-based anionic surfactants with phosphates groups, Colloids Surfaces A Physicochem. Eng. Asp. 485 (2015) 118–124. https://doi.org/10.1016/j.colsurfa.2015.09.021.

[21] C. Guo, P. Zhou, J. Shao, X. Yang, Z. Shang, Integrating statistical and experimental protocols to model and design novel Gemini surfactants with promising critical micelle concentration and low environmental risk, Chemosphere. 84 (2011) 1608–1616. https://doi.org/10.1016/j.chemosphere.2011.05.031.

[22] M.G. Rosen, Surfactants and Interfacial Phenomena, Third edition, Jhon Wiley & sons, (2004).

[23] M. Mattei, G.M. Kontogeorgis, R. Gani, Modeling of the critical micelle concentration (CMC) of nonionic surfactants with an extended group-contribution method, Ind. Eng. Chem. Res. 52 (2013) 12236–12246. https://doi.org/10.1021/IE4016232.

[24] S.K. Hait, S.P. Moulik, Determination of Critical Micelle Concentration (CMC) of Nonionic Surfactants by Donor-Acceptor Interaction with Iodine and Correlation of CMC with Hydrophile-Lipophile Balance and Other Parameters of the Surfactants, J. Surfactants Deterg. 4 (2001) 303–309. https://doi.org/10.1007/s11743-001-0184-2.

[25] M.T. Lima, V.J. Spiering, S.N. Kurt-Zerdeli, D.C. Brüggemann, M. Gradzielski, R. Schomäcker, The hydrophilic-lipophilic balance of carboxylate and carbonate modified nonionic surfactants, Colloids Surfaces A Physicochem. Eng. Asp. 569 (2019) 156–163. https://doi.org/10.1016/j.colsurfa.2019.03.001.

[26] T. Gaudin, H. Lu, G. Fayet, A. Berthauld-Drelich, P. Rotureau, G. Pourceau, A. Wadouachi, E. Van Hecke, A. Nesterenko, I. Pezron, Impact of the chemical structure on amphiphilic properties of sugar-based surfactants: A literature overview, Adv. Colloid Interface Sci. 270 (2019) 87–100. https://doi.org/10.1016/j.cis.2019.06.003.

[27] S. Iglauer, Y. Wu, P. Shuler, Y. Tang, W.A. Goddard, Analysis of the influence of alkyl polyglycoside surfactant and cosolvent structure on interfacial tension in aqueous formulations versus n-octane, Tenside, Surfactants, Deterg. 47 (2010) 87–97. https://doi.org/10.3139/113.110056.

[28] G. Lemahieu, J. Aguilhon, H. Strub, V. Molinier, J.F. Ontiveros, J.M. Aubry,

Hexahydrofarnesyl as an original bio-sourced alkyl chain for the preparation of glycosides surfactants with enhanced physicochemical properties, RSC Adv. 10 (2020) 16377–16389. https://doi.org/10.1039/d0ra02326d.

[29] C. Scorzza, P. Godé, G. Goethals, P. Martin, M. Miñana-Pérez, J.L. Salager, A. Usubillaga, P. Villa, Another new family of "extended" glucidoamphiphiles. Synthesis and surfactant properties for different sugar head groups and spacer arm lengths, J. Surfactants Deterg. 5 (2002) 337–343. https://doi.org/10.1007/S11743-002-0231-Z.

[30] F. Grisoni, V. Consonni, R. Todeschini, Impact of Molecular Descriptors on Computational Models, 2018. https://doi.org/10.1007/978-1-4939-8639-2_5.

[31] H. Moriwaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator, J. Cheminform. 10 (2018) 1–14. https://doi.org/10.1186/s13321-018-0258-y.

[32] D. Wang, Y. Yuan, S. Duan, R. Liu, S. Gu, S. Zhao, L. Liu, J. Xu, QSPR study on melting point of carbocyclic nitroaromatic compounds by multiple linear regression and artificial neural network, Chemom. Intell. Lab. Syst. 143 (2015) 7–15. https://doi.org/10.1016/j.chemolab.2015.02.009.

[33] O. Falyouna, O. Eljamal, I. Maamoun, A. Tahara, Y. Sugihara, Magnetic zeolite synthesis for efficient removal of cesium in a lab-scale continuous treatment system, J. Colloid Interface Sci. 571 (2020) 66–79. https://doi.org/10.1016/j.jcis.2020.03.028.

[34] K. Roy, R.N. Das, P. Ambure, R.B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, Chemom. Intell. Lab. Syst. 152 (2016) 18–33. https://doi.org/10.1016/J.CHEMOLAB.2016.01.008.

[35] K. Roy, I. Mitra, On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design, Comb. Chem. High Throughput Screen. 14 (2011) 450–474. https://doi.org/10.2174/138620711795767893.

[36] N. Chirico, P. Gramatica, Real external predictivity of QSAR models: How to evaluate It? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, J. Chem. Inf. Model. 51 (2011) 2320–2335. https://doi.org/10.1021/CI200211N.

[37] G. Schüürmann, R.U. Ebert, J. Chen, B. Wang, R. Kühne, External validation and prediction employing the predictive squared correlation coefficient - Test set activity

mean vs training set activity mean, J. Chem. Inf. Model. 48 (2008) 2140–2145. https://doi.org/10.1021/ci800253u.

[38] A. Tropsha, P. Gramatica, V.K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR Comb. Sci. 22(2003) 69-77.

[39] H.F. Chen, In silico log p prediction for a large data set with support vector machines, radial basis neural networks and multiple linear regression, Chem. Biol. Drug Des. 74 (2009) 142–147. https://doi.org/10.1111/j.1747-0285.2009.00840.x.

[40] Z. Ghomisheh, A.E. Gorji, M.A. Sobati, Journal of Molecular Graphics and Modelling Prediction of critical properties of sulfur-containing compounds : New QSPR models, J. Mol. Graph. Model. 101 (2020) 107700. https://doi.org/10.1016/j.jmgm.2020.107700.

[41] A. Baghban, A. Jalali, M. Shafiee, M.H. Ahmadi, K. wing Chau, Developing an ANFIS-based swarm concept model for estimating the relative viscosity of nanofluids, Eng. Appl. Comput. Fluid Mech. 13 (2019) 26–39. https://doi.org/10.1080/19942060.2018.1542345.

[42] Y. Wang, F. Yan, Q. Jia, Q. Wang, Quantitative structure-property relationship for critical micelles concentration of sugar-based surfactants using norm indexes, J. Mol. Liq. 253 (2018) 205–210. https://doi.org/10.1016/j.molliq.2018.01.037.

[43] S. Rahal, N. Hadidi, M. Hamadache, In Silico Prediction of Critical Micelle Concentration (CMC) of Classic and Extended Anionic Surfactants from Their Molecular Structural Descriptors, Arab. J. Sci. Eng. 45 (2020) 7445–7454. https://doi.org/10.1007/S13369-020-04598-0.

[44] K. Roy, H. Kabir, QSPR with extended topochemical atom (ETA) indices, 3: Modeling of critical micelle concentration of cationic surfactants, Chem. Eng. Sci. 81 (2012) 169–178. https://doi.org/10.1016/j.ces.2012.07.008.

General conclusion and Outlook

General Conclusion and Outlook

Surfactants are attracting growing interest due to their varied exceptional characteristics, such as wetting action, foam formation, emulsion formation and stabilization, dispersion, and detergent action.

This project was set up to highlight the study of the properties of surfactants, notably CMC, which is a fundamental characteristic that provides information on molecular interactions, surfactant behavior, formulation optimization, pharmaceutical research, and so on, with this in mind generally measured using experimental devices, which represents a cost and can be time-consuming. Their prediction therefore represents a significant benefit, potentially making it possible to screen large databases of molecules at a low cost and before any synthesis.

In order to promote molecules of interest, this thesis aimed to develop different QSPR models linking the physicochemical property to the molecular structure using computational intelligence methods. A large number of molecular descriptors have been calculated (constitutional, electronic, topological, geometric, physicochemical descriptors, and more). The main validation techniques were used (internal validation, external validation, areas of applicability, etc.). These models were developed following the five OECD principles for validating QSPR models.

Firstly, the CMC of sugar-based surfactants was evaluated using experimental data taken from the literature in a dataset containing 119 experimental data points divided into three random sets as follows: a training set, an internal validation set, and a test set where 6 significant PaDEL descriptors were used as model input variables. In this study four modeling methods, MLR, PLS, ANN, and ANFIS were used, the results obtained show that:

- The developed models are all statistically valid with an overall RMSE between 0.314 to 0.490.
- ANN has significantly better predictive ability than PLS, ANFIS, and MLR, but the latter gives easily interpretable results.
- The ANN model using the trainbr training algorithm and tansig and purelin as activation functions showed very significant results with a value of R^2=0.98 for the test set.
- In-silico models demonstrate the importance of a molecule's electrical characteristics, electro-topology, and symmetry in the micellization process.

Secondly, we used MLR, RFR, ANN, and DA-SVR as computational intelligence methods to construct QSPR models of different surfactant classes for their CMC properties. 80% of the data, chosen randomly, was used as the training set to build the model. The remaining data were used as a test set to evaluate the model by considering the 14 significant Mordred descriptors (SlogP, GATS6d, nAcid, GATS8dv, GATS4dv, PEOE_VSA11, GATS8d, ATS0p, GATS1d, MATS5p, GATS3d, NdssC, GATS6dv et EState_VSA4) and temperature as model inputs. the results obtained show that:

- The methods used were compared, and among them, the DA-SVR based on descriptors proposed by the S-MLR presented the best predictive capacity than the others with an overall R^2 of 0.974.
- The poor results of internal and external validation for MLR show that the model proposed by this method is not capable of predicting the property satisfactorily.
- The predictions in the application domain achieved an accuracy of 95.95%. As a result, the DA-SVR model provided appropriate predictions for these substances.

To this end, we hope that we will improve our work in the future. As perspectives we propose:

- Modeling the CMC by hybrid algorithms.
- Modify the QSPR model construction such as selecting different types of descriptors.
- Promote these methods for other physicochemical properties of surfactants.
- Use ongoing improvements to computational intelligence methods, such as “deep learning” to predict/correlate the physicochemical properties of surfactants.

Annex

Annex

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

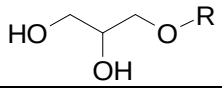
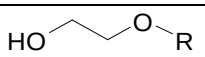
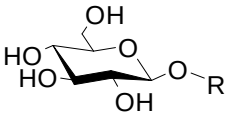
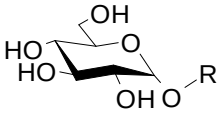
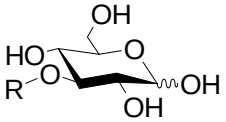
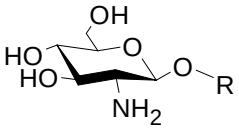
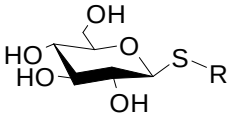
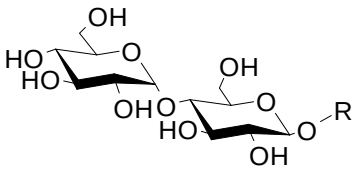
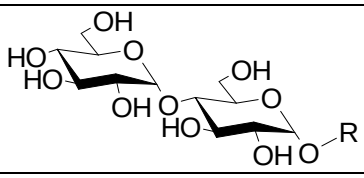
N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
1		Octyl	2.237	2.394
2		Octyl	2.310	2.319
3		Octyl	1.669	2.125
4		1-butylhexyl	1.824	1.776
5		1-propylheptyl	1.921	1.806
6		1-ethyloctyl	2.071	1.877
7		1-mehylnonyl	2.347	2.276
8		3,7-dimithyloctyl	2.398	1.995
9		Decyl	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		3,7,11trimethyldodecyl	4.526	4.467
33		Dodecyl	3.921	3.770

Table A.1 (*Continued*)

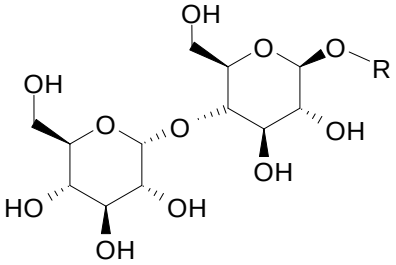
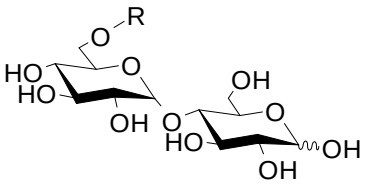
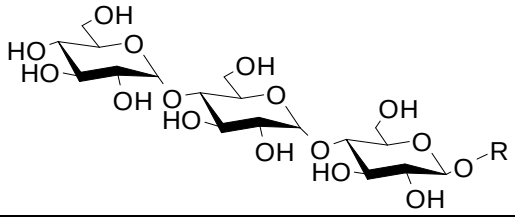
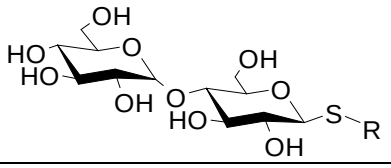
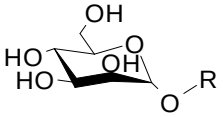
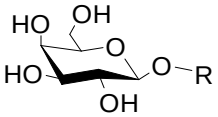
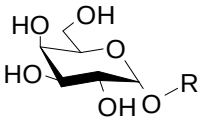
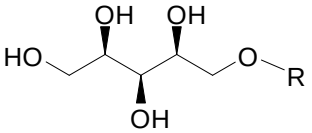
N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
34		Hexadecyl	6.222	6.219
35		Dodecanoyl	3.481	3.327
36		3,7-dimethyloctyl	2.301	2.327
37		Oleyl	4.377	4.503
38		Octyl	2.086	2.307
39		Decyl	3.222	3.294
40		Octyl	2.222	2.125
41		Decyl	3.602	2.928
42		Dodecyl	4.301	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 A.1 (*Continued*)

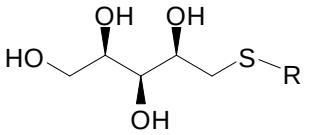
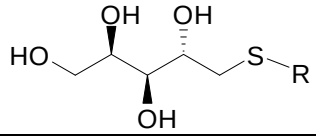
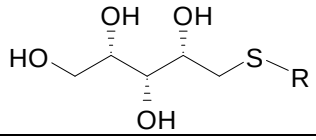
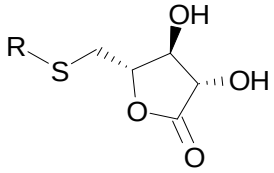
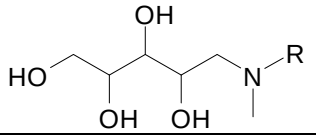
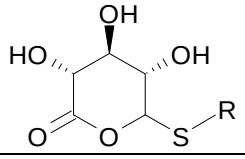
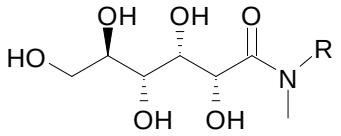
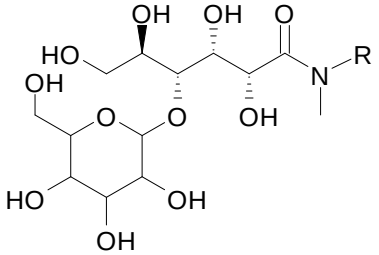
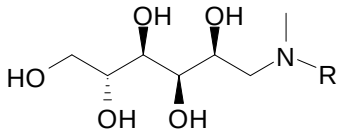
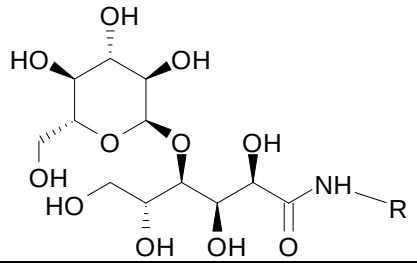
N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
63		Butyl	0.745	1.249
64		Pentyl	1.337	1.442
65		Hexyl	1.745	1.908
66		Octyl	2.745	2.980
67		Hexyl	1.921	1.908
68		Octyl	2.921	2.980
69		Decyl	3.398	3.732
70		Octyl	3.319	3.569
71		Decyl	4.481	4.505
72		Dodecanoyl	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		Oleyl	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		Oleyl	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 A.1 (*Continued*)

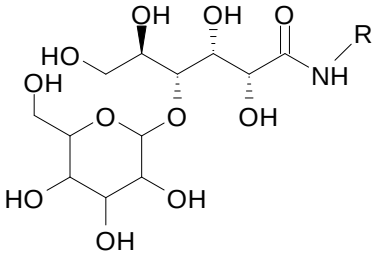
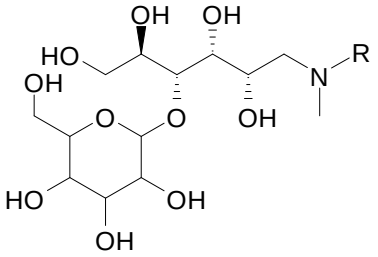
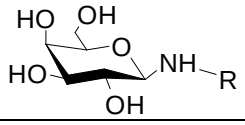
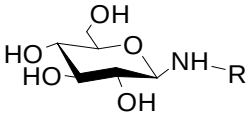
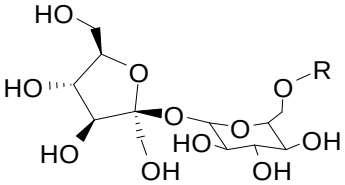
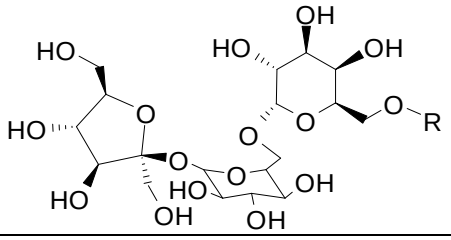
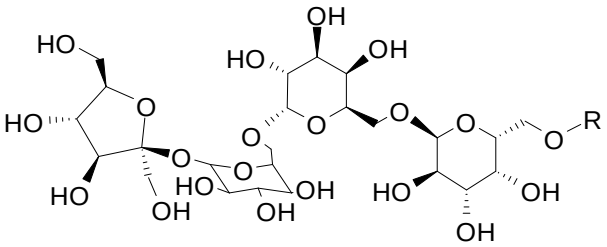
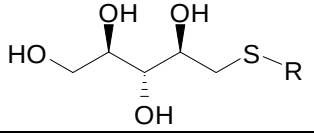
N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	**Predicted****
90		Decyl	2.886	2.869
91		Decanoyl	2.481	2.622
92		Dodecanoyl	3.347	3.219
93		Tetradecanoyl	4.167	3.969
94		Octanoyl	1.347	1.250
95		Octanoyl	1.155	1.250
96		Dodecanoyl	3.337	3.389
97		Dodecanoyl	3.022	2.997
98		Dodecanoyl	2.638	2.647
99		Hexyl	1.991	1.908
100		Octyl	3.420	2.980

Table A.1 (*Continued*)

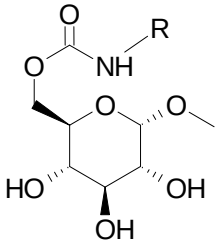
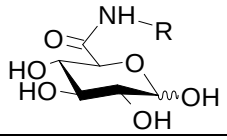
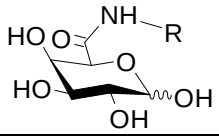
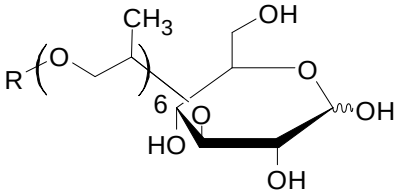
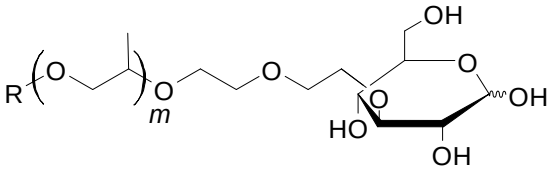
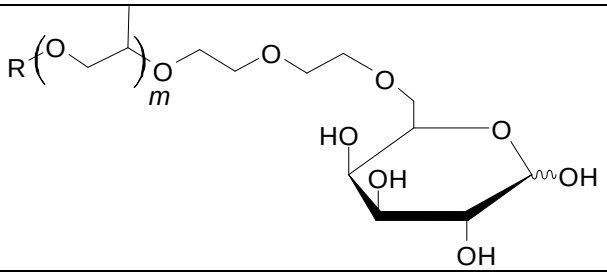
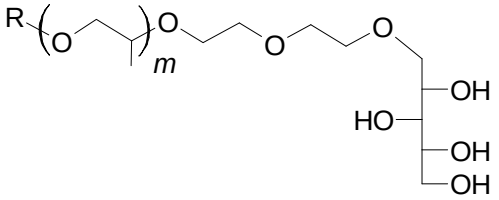
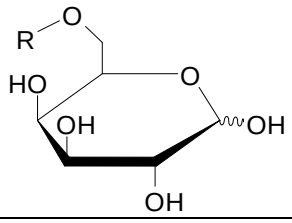
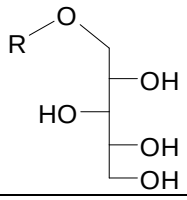
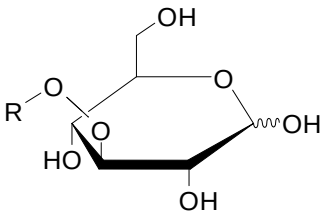
N°	Sugar surfactants structures	Substituent (R)	pCMC (mol/L)	
			Observed*	Predicted**
101		heptyl	1.710	1.851
102		Octyl	2.481	2.685
103		Octyl	2.398	2.685
104		Dodecyl	4.398	4.230
		Dodecyl		
105		m=6	3.959	4.032
106		m=10	4.398	4.610
107		m=14	4.699	4.862
		Dodecyl		
108		m=6	4.301	4.303
109		m=10	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 A.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

Table A.2 List of 593 surfactants with experimental and predicted pCMC values

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
1	$C_{15}H_{31}CONHCH[CH(CH_3)_2]COO^-Na^+$	25	2.7212	2.5623
2	$C_{15}H_{31}CONHCH[CH_2CH(CH_3)_2]COO^-Na^+$	25	2.8239	2.8164
3	$C_{10}H_{21}SO_3^-Na^+$	25	1.3665	1.3747
4	$C_{12}H_{25}SO_3^-Na^+$	25	1.9066	1.9145
5	$C_{12}H_{25}SO_3^-Li^+$	25	1.9586	1.9505
6	$C_{12}H_{25}SO_3^-NH_4^+$	25	2.0506	3.1758
7	$C_{12}H_{25}SO_3^-K^+$	25	2.0315	2.6789
8	$C_{11}H_{23}SO_4^-Na^+$	21	1.7959	1.6901
9	$C_{12}H_{25}SO_4^-Na^+$	25	2.0862	2.0354
10[t]	$C_{12}H_{25}SO_4^-Li^+$	25	2.0506	2.0711
11[t]	$C_{12}H_{25}SO_4^-N^+(CH_3)_4$	25	2.2596	2.4261
12[t]	$C_{12}H_{25}SO_4^-N^+(C_3H_7)_4$	25	2.6576	2.5960
13	$C_{14}H_{29}SO_4^-Na^+$	25	2.6778	2.7363
14	$C_{10}H_{21}OC_2H_4SO_3^-Na^+$	25	1.7986	1.7906
15	$C_{12}H_{24}OC_2H_4SO_4^-Na^+$	25	2.4089	2.8228
16	$C_{12}H_{24}(OC_2H_4)_2SO_4^-Na^+$	25	2.5376	3.4484
17	$C_{16}H_{33}(OC_2H_4)_5SO_4^-Na^+$	25	4.6021	4.5939
18[t]	$C_8H_{17}CH(C_6H_{13})CH_2(OC_2H_4)_5SO_4^-Na^+$	25	4.0655	4.0736
19	$C_6H_{13}OOCCH_2SO_3^-Na^+$	25	0.7696	0.7779
20	$C_8H_{17}OOCCH_2SO_3^-Na^+$	25	1.1805	1.1885
21	$C_{10}H_{21}OOCCH_2SO_3^-Na^+$	25	1.6576	1.6661
22	$C_4H_9OOCCH_2CH(SO_3^-Na^+)COOC_4H_9$	25	0.6990	0.7068
23	$C_5H_{11}OOCCH_2CH(SO_3^-Na^+)COOC_5H_{11}$	25	1.2757	1.8769
24	$C_6H_{13}OOCCH_2CH(SO_3^-Na^+)COOC_6H_{13}$	25	1.8539	1.8617
25	$C_4H_9CH(C_2H_5)CH_2OOCCH_2CH(SO_3^-Na^+)COOCH_2CH(C_2H_5)C_4H_9$	25	2.6021	2.4658
26	$C_8H_{17}OOCCH_2CH(SO_3^-Na^+)COOC_8H_{17}$	25	3.0410	2.5063
27	$C_{12}H_{25}CH(SO_3^-Na^+)COOC_2H_5$	25	2.6478	2.6556

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
28	$C_{12}H_{25}CH(SO_3^-Na^+)COOC_4H_9$	25	2.8697	2.8619
29	$C_{14}H_{29}CH(SO_3^-Na^+)COOCH_3$	25	3.1367	3.1284
30	$C_7F_{15}COO^-K^+$	25	1.5376	1.5460
31	$C_7F_{15}COO^-Na^+$	25	1.5229	1.4929
32[t]	$C_7F_{15}COO^-Li^+$	25	1.4815	1.4893
33	$(CF_3)_2CF(CF_2)_4COO^-Na^+$	25	1.5229	1.5305
34[t]	$C_8F_{17}COO^-Li^+$	25	2.3098	2.3014
35	$C_8F_{17}SO_3^-Li^+$	25	2.2007	2.2085
36	$C_{11}H_{23}COO^-Na^+$	25	2.5243	2.5163
37	$C_{12}H_{25}(OC_2H_4)_5OCH_2COO^-Na^+$	25	2.4572	2.4489
38	$C_{12}H_{25}(OC_2H_4)_9OCH_2COO^-Na^+$	25	2.2291	2.2368
39	$C_9H_{19}(C_6H_4)[OCH_2CH(CH_3)]_3SO_4^-Na^+$	25	4.2291	4.2375
40	$C_9H_{19}(C_6H_4)[OCH_2CH(CH_3)]_6SO_4^-Na^+$	25	4.5229	4.5146
41	$C_9H_{19}(C_6H_4)[OCH_2CH(CH_3)]_9SO_4^-Na^+$	25	4.7959	4.6858
42[t]	$C_9H_{19}(C_6H_4)[OCH_2CH(CH_3)]_{12}SO_4^-Na^+$	25	4.9208	4.9289
43	$C_{12}H_{25}O[CH_2CH(CH_3)O]_5CH_2CH(CH_3)SO_4^-Na^+$	25	3.8297	3.8215
44	$C_{12}H_{25}O[CH_2CH(CH_3)O]_9CH_2CH(CH_3)SO_4^-Na^+$	25	4.2757	4.2838
45	$C_{12}H_{25}O[CH_2CH(CH_3)O]_{13}CH_2CH(CH_3)SO_4^-Na^+$	25	4.8861	4.8781
46	$C_{12}H_{25}O[CH_2CH(CH_3)O]_5CH_2CH(CH_3)OCH_2COO^-Na^+$	25	3.5421	3.5339
47	$C_{12}H_{25}O[CH_2CH(CH_3)O]_9CH_2CH(CH_3)OCH_2COO^-Na^+$	25	3.9914	3.9995
48[t]	$C_{12}H_{25}O[CH_2CH(CH_3)O]_{13}CH_2CH(CH_3)OCH_2COO^-Na^+$	25	4.4437	4.4355
49	$C_{12}H_{25}O[CH_2CH(CH_3)O]_5CH_2CH(CH_3)(OC_2H_4)_2SO_4^-Na^+$	25	3.7696	4.0557
50	$C_{12}H_{25}O[CH_2CH(CH_3)O]_5CH_2CH(CH_3)(OC_2H_4)_2OCH_2COO^-Na^+$	25	3.4034	3.4115
51	$C_{12}H_{25}(OC_2H_4)_3SO_4^-Na^+$	25	2.5528	2.5449
52	$C_{12}H_{25}(OC_2H_4)_5SO_4^-Na^+$	25	2.7212	2.7297
53	$C_{16}H_{33}SO_4^-Na^+$	25	3.3979	3.4061
54	$C_{16}H_{33}OC_2H_4SO_4^-Na^+$	25	3.6576	3.6652

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
55	$C_{16}H_{33}(OC_2H_4)_2SO_4^-Na^+$	25	3.8539	4.0791
56	$C_{16}H_{33}(OC_2H_4)_3SO_4^-Na^+$	25	4.0000	3.9921
57	$C_{16}H_{33}(OC_2H_4)_4SO_4^-Na^+$	25	4.0000	4.0078
58	$C_{18}H_{37}SO_4^-Na^+$	25	3.9586	3.9504
59	$C_{18}H_{37}OC_2H_4SO_4^-Na^+$	25	3.9586	3.9504
60	$C_{18}H_{37}(OC_2H_4)_2SO_4^-Na^+$	25	4.1549	4.1630
61[t]	$C_{18}H_{37}(OC_2H_4)_3SO_4^-Na^+$	25	4.3010	4.2927
62	$C_{18}H_{37}(OC_2H_4)_4SO_4^-Na^+$	25	4.3979	4.4061
63[t]	$C_{24}H_{48}O(C_2H_4O)_7O(PO_4Na_2)_4$	25	2.8356	2.9167
64[t]	$C_{24}H_{48}O(C_2H_4O)_{20}O(PO_4Na_2)_4$	25	2.9872	2.9955
65	$C_{24}H_{48}O(C_2H_4O)_{30}O(PO_4Na_2)_4$	25	3.0757	2.9955
66	$C_{12}H_{25}COO^-K^+$	30	1.9208	1.9289
67	$C_9H_{19}CONHCH_2COO^-Na^+$	40	1.4202	1.4281
68	$C_{11}H_{23}CONHCH_2COO^-Na^+$	40	2.0000	1.9919
69	$C_{11}H_{23}CON(CH_3)CH_2COO^-Na^+$	30	2.0000	2.1493
70	$C_{11}H_{23}CON(CH_3)C_2H_4COO^-Na^+$	30	2.1192	2.1268
71	$C_{13}H_{27}CONHCH_2COO^-Na^+$	40	2.3768	2.3688
72	$C_8H_{17}SO_3^-Na^+$	40	0.7959	1.5022
73	$C_{10}H_{21}SO_3^-Na^+$	10	1.3188	1.3267
74[t]	$C_{10}H_{21}SO_3^-Na^+$	40	1.3979	1.4061
75	$C_{12}H_{25}SO_3^-Na^+$	40	1.9431	1.9512
76	$C_{14}H_{29}SO_3^-Na^+$	40	2.6021	2.5290
77	$C_{16}H_{33}SO_3^-Na^+$	50	3.1549	3.1628
78	$C_8H_{17}SO_4^-Na^+$	40	0.8539	0.8616
79	$C_{10}H_{21}SO_4^-Na^+$	40	1.4815	1.4854
80	$C_{12}H_{25}SO_4^-Na^+$	40	2.0655	2.0574
81[t]	$C_{12}H_{25}SO_4^-K^+$	40	2.1079	2.1155

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
82	$C_{12}H_{25}SO_4^-N^+(C_2H_5)_4$	30	2.3468	2.3545
83	$C_{12}H_{25}SO_4^-N^+(C_4H_9)_4$	30	2.8861	2.8780
84	$C_{13}H_{27}SO_4^-Na^+$	40	2.3665	2.3606
85	$C_{14}H_{29}SO_4^-Na^+$	40	2.6576	2.6656
86[t]	$C_{15}H_{31}SO_4^-Na^+$	40	2.9208	2.9582
87	$C_{16}H_{33}SO_4^-Na^+$	40	3.2366	3.2282
88	$C_{13}H_{27}CH(CH_3)CH_2SO_4^-Na^+$	40	3.0969	3.0892
89	$C_{12}H_{25}CH(C_2H_5)CH_2SO_4^-Na^+$	40	3.0458	3.0380
90	$C_{11}H_{23}CH(C_3H_7)CH_2SO_4^-Na^+$	40	2.9586	2.9505
91	$C_{10}H_{21}CH(C_4H_9)CH_2SO_4^-Na^+$	40	2.8239	2.8156
92	$C_9H_{19}CH(C_5H_{11})CH_2SO_4^-Na^+$	40	2.6990	2.7074
93	$C_8H_{17}CH(C_6H_{13})CH_2SO_4^-Na^+$	40	2.6383	2.6459
94	$C_7H_{15}CH(C_7H_{15})CH_2SO_4^-Na^+$	40	2.5229	2.4849
95	$C_{12}H_{25}CH(SO_4^-Na^+)C_3H_7$	40	2.7696	2.7779
96	$C_{10}H_{21}CH(SO_4^-Na^+)C_5H_{11}$	40	2.6198	2.6120
97	$C_8H_{17}CH(SO_4^-Na^+)C_7H_{15}$	40	2.3665	2.3745
98	$C_{18}H_{37}SO_4^-Na^+$	50	3.6383	3.6304
99	$C_{12}H_{25}(OC_2H_4)_2SO_4^-Na^+$	10	2.5086	2.5165
100	$C_{12}H_{25}(OC_2H_4)_2SO_4^-Na^+$	40	2.5528	2.5605
101	$C_{12}H_{25}(OC_2H_4)_3SO_4^-Na^+$	50	2.6990	2.7066
102	$C_{12}H_{25}(OC_2H_4)_4SO_4^-Na^+$	50	2.8861	2.6902
103	$C_8H_{17}OOC(CH_2)_2SO_3^-Na^+$	30	1.3372	1.3457
104	$C_{10}H_{21}OOC(CH_2)_2SO_3^-Na^+$	30	1.9586	2.0402
105	$C_{12}H_{25}OOC(CH_2)_2SO_3^-Na^+$	30	2.6576	2.6498
106	$C_{14}H_{29}OOC(CH_2)_2SO_3^-Na^+$	40	3.0458	3.0379
107	$C_{12}H_{25}CH(SO_3^-Na^+)COOCH_3$	13	2.5528	2.5613
108	$C_{16}H_{33}CH(SO_3^-Na^+)COOCH_3$	33	3.7447	3.7369

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
106	$C_{11}H_{23}CON(CH_3)CH_2CH_2SO_4^-Na^+$	30	2.0506	2.0589
110[t]	$C_{12}H_{25}NHCOCH_2SO_4^-Na^+$	35	2.2840	2.2924
111	$C_{12}H_{25}NHCO(CH_2)_2SO_4^-Na^+$	35	2.3565	2.8396
112	$p-C_8H_{17}C_6H_4SO_3^-Na^+$	35	1.8239	2.4737
113	$p-C_{10}H_{21}C_6H_4SO_3^-Na^+$	50	2.5086	2.5163
114	$p-C_{12}H_{25}C_6H_4SO_3^-Na^+$	60	2.9208	2.9126
115	$C_8F_{17}COO^-Na^+$	35	1.9586	1.9663
116	$C_4F_9CH_2OOCCH(SO_3^-Na^+)CH_2COOCH_2C_4H_9$	30	2.7959	2.7874
117	$C_{16}H_{33}O(C_2H_4O)_3CH_2COO^-Na^+$	30	4.0259	4.0181
118	$C_{16}H_{33}O(C_2H_4O)_5CH_2COO^-Na^+$	30	4.3429	4.3513
119	$C_{16}H_{33}O(C_2H_4O)_7CH_2COO^-Na^+$	30	4.9957	4.9874
120	$C_{14}H_{29}(OC_2H_4)_5OCOC_2H_4COO^-Na^+$	30	2.8894	2.5675
121	$(CH_3)_2CH(C_3H_6)CH(CH_3)C_2H_4CH(CH_3)CH_2(OCH_2CH(CH_3))_7OSO_3^-Na^+$	30	4.3372	4.3452
122	$(CH_3)_2CH(C_3H_6)CH(CH_3)C_2H_4CH(CH_3)CH_2(OCH_2CH(CH_3))_{13}OSO_3^-Na^+$	30	4.9208	4.9128
123	$C_8H_{17}N^+(CH_3)_3Br^-$	25	0.8539	0.8597
124[t]	$C_{10}H_{21}N^+(CH_3)_3Br^-$	25	1.1675	1.1759
125	$C_{10}H_{21}N^+(CH_3)_3Cl^-$	25	1.1675	1.1598
126	$C_{12}H_{25}N^+(CH_3)_3Br^-$	25	1.7959	1.7878
127	$C_{12}H_{25}N^+(CH_3)_3Cl^-$	25	1.6990	1.7262
128	$C_{14}H_{29}N^+(CH_3)_3Br^-$	25	2.4437	2.4198
129[t]	$C_{14}H_{29}N^+(CH_3)_3Cl^-$	25	2.3468	2.3385
130	$C_{16}H_{32}N^+(CH_3)_3Br^-$	25	3.0088	3.0167
131[t]	$C_{10}H_{21}Pyr^+Br^-$	25	1.3565	1.3649
132	$C_{11}H_{23}Pyr^+Br^-$	25	1.6778	1.6439
133	$C_{12}H_{25}Pyr^+Br^-$	25	1.9431	1.9511
134	$C_{12}H_{25}Pyr^+Cl^-$	25	1.7696	1.7773
135	$C_{12}H_{25}Pyr^+I^-$	25	2.2757	2.2678

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
136	$C_{13}H_{27}Pyr^{+}Br^{-}$	25	2.2757	2.2673
137	$C_{14}H_{29}Pyr^{+}Br^{-}$	25	2.5686	2.5824
138	$C_{14}H_{29}Pyr^{+}Cl^{-}$	25	2.4559	2.4295
139[t]	$C_{15}H_{31}Pyr^{+}Br^{-}$	25	2.8861	2.8902
140	$C_{16}H_{33}Pyr^{+}Br^{-}$	25	3.1938	3.1861
141	$C_{16}H_{33}Pyr^{+}Cl^{-}$	25	3.0458	3.0541
142[t]	$C_{18}H_{37}Pyr^{+}Cl^{-}$	25	3.6198	3.6114
143	$C_{12}H_{25}N^{+}(C_2H_5)(CH_3)_2Br^{-}$	25	1.8539	1.8618
144	$C_{12}H_{25}N^{+}(C_4H_9)(CH_3)_2Br^{-}$	25	2.1249	2.1330
145[t]	$C_{12}H_{25}N^{+}(C_6H_{13})(CH_3)_2Br^{-}$	25	2.5086	2.5165
146	$C_{12}H_{25}N^{+}(C_8H_{17})(CH_3)_2Br^{-}$	25	2.9586	2.9510
147	$C_{14}H_{29}N^{+}(C_2H_5)_3Br^{-}$	25	2.5086	2.5164
148[t]	$C_{14}H_{29}N^{+}(C_3H_7)_3Br^{-}$	25	2.6778	2.5506
149	$C_{14}H_{29}N^{+}(C_4H_9)_3Br^{-}$	25	2.9208	2.9291
150	$C_{10}H_{21}N^{+}(CH_2C_6H_5)(CH_3)_2Cl^{-}$	25	1.4089	2.0394
151	$C_{12}H_{25}N^{+}(CH_2C_6H_5)(CH_3)_2Cl^{-}$	25	2.0555	2.0475
152	$C_{14}H_{29}N^{+}(CH_2C_6H_5)(CH_3)_2Cl^{-}$	25	2.6990	2.0956
153[t]	$C_{12}H_{25}NH_2^{+}CH_2CH_2OHCl^{-}$	25	1.3468	1.3387
154[t]	$C_{12}H_{25}NH^{+}(CH_2CH_2OH)_2Cl^{-}$	25	1.4437	1.4516
155[t]	$C_{12}H_{25}N^{+}(CH_2CH_2OH)_3Cl^{-}$	25	1.6021	1.6102
156	$(C_{10}H_{21})_2N^{+}(CH_3)_2Br^{-}$	25	2.7328	2.7409
157	$(C_{12}H_{25})_2N^{+}(CH_3)_2Br^{-}$	25	3.7545	3.7467
158	$C_{10}H_{21}(OC_2H_4)N^{+}(CH_3)_3Cl^{-}$	25	1.5143	1.5224
159	$C_{10}H_{21}(OC_2H_4)_2N^{+}(CH_3)_3Cl^{-}$	25	1.6383	1.6466
160[t]	$C_{12}H_{25}(OC_2H_4)N^{+}(CH_3)_3Cl^{-}$	25	2.1487	2.1409
161	$C_{12}H_{25}(OC_2H_4)_2N^{+}(CH_3)_3Cl^{-}$	25	2.2840	1.8029
162	$C_{18}H_{37}N^{+}(CH_3)_3Cl^{-}$	25	3.4685	3.4602

Table A.2 Continued

Nos.	Compounds					T(°C)	pCMC	
							Observed	Predicted [a]
163	$C_{10}H_{21}Pyr^+Cl^-$					25	1.1871	1.1796
164	$C_{10}H_{21}(OC_2H_4)Pyr^+Cl^-$					25	1.5918	1.5834
165	$C_{10}H_{21}(OC_2H_4)_2Pyr^+Cl^-$					25	1.6778	1.6856
166	$C_{12}H_{25}(OC_2H_4)Pyr^+Cl^-$					25	2.1612	2.1693
167	$C_{12}H_{25}(OC_2H_4)_2Pyr^+Cl^-$					25	2.3098	2.3015
168	$C_{16}H_{33}(OC_2H_4)Pyr^+Cl^-$					25	3.5376	3.5293
169	$C_{16}H_{33}(OC_2H_4)_2Pyr^+Cl^-$					25	3.7447	3.5626
170	$C_{14}H_{29}N^+(CH_3)_3Br^-$					40	2.3768	2.3848
171[t]	$C_{14}H_{29}N^+(CH_3)_3Br^-$					60	2.2596	2.2681
172[t]	$C_{16}H_{33}N^+(CH_3)_3Cl^-$					30	2.8861	2.8940
173	$C_{18}H_{37}N^+(CH_3)_3Br^-$					40	3.4685	3.4605
174	$C_{12}H_{25}Pyr^+Br^-$					10	1.9318	1.9240
175	$C_{12}H_{25}Pyr^+Br^-$					40	1.9508	1.9429
176[t]	$C_{12}H_{25}Pyr^+Cl^-$					10	1.7570	1.7647
177[t]	$C_{12}H_{25}Pyr^+Cl^-$					40	1.7696	1.7774
	Tail 1	Tail 2	Ion-1	Ion-2	Spacer			
178[t]	C_8H_{17}	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	3.0000	2.9923
179[t]	C_8H_{17}	$C_{10}H_{21}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	3.9586	3.9669
180	$C_{10}H_{21}$	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	4.0000	4.0083
181	C_8H_{17}	$C_{12}H_{25}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	4.8539	4.8458
182	$C_{12}H_{25}$	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	4.7447	4.8145
183	$C_{10}H_{21}$	$C_{10}H_{21}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	4.8861	4.8780
184[t]	C_8H_{17}	$C_{14}H_{29}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.2147	5.2096
185[t]	$C_{14}H_{29}$	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.0969	5.1455
186	$C_{10}H_{21}$	$C_{12}H_{25}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	4.8861	5.2038
187	$C_{10}H_{21}$	$C_{14}H_{29}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.2757	5.2517
188	$C_{14}H_{29}$	$C_{10}H_{21}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.2596	5.2517

Table A.2 Continued

Nos.	Compounds					T(°C)	pCMC	
							Observed	Predicted [a]
	Tail 1	Tail 2	Ion-1	Ion-2	Spacer			
189	$C_{12}H_{25}$	$C_{12}H_{25}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.0458	5.2517
190	C_8H_{17}	$C_{18}H_{37}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.3188	5.3174
191	$C_{18}H_{37}$	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.2291	5.2370
192	$C_{10}H_{21}$	$C_{16}H_{33}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.3010	5.2928
193	$C_{14}H_{29}$	$C_{12}H_{25}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.1549	5.2928
194	$C_{12}H_{25}$	$C_{14}H_{29}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	25	5.3665	5.2928
195	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_2$	25	3.0969	3.0891
196[t]	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	25	3.0269	3.0190
197	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	25	2.9318	3.1688
198	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	2.9872	3.0227
199	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_8$	25	3.0809	2.9228
200	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{10}$	25	3.2007	3.2087
201	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{12}$	25	3.4318	3.4514
202	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{14}$	25	3.6990	3.7072
203[t]	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{16}$	25	3.9208	4.0145
204[t]	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	25	1.8601	1.8519
205[t]	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	25	1.5817	1.5739
206	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	1.5800	1.5719
207	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	1.5986	1.5908
208[t]	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_2$	25	4.6778	4.6856
209[t]	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	25	4.5686	3.9675
210	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	4.3665	4.3586
211	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	25	4.6021	3.6287
212	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	4.4437	4.4516
213	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_8$	25	4.4815	4.0145
214	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{10}$	25	4.5686	4.3666

Table A.2 Continued

Nos.	Compounds					T(°C)	pCMC	
							Observed	Predicted [a]
	Tail 1	Tail 2	Ion-1	Ion-2	Spacer			
215[t]	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{12}$	25	4.6990	4.7073
216	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	25	2.0565	2.0649
217[t]	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	25	3.8697	3.8617
218	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_2$	25	2.2076	2.2156
219	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	25	2.1871	2.1955
220	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	25	2.0605	2.0649
221	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	2.0362	2.0343
222	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_8$	25	2.1249	2.1171
223	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{10}$	25	2.3768	2.5565
224	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{12}$	25	2.6576	2.9228
225	C_6H_{13}	C_6H_{13}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	0.7959	0.8042
226[t]	C_9H_{19}	C_9H_{19}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	1.6021	1.6097
227	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	2.0506	2.0426
228	$C_{11}H_{23}$	$C_{11}H_{23}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	2.5229	2.5145
229	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	2.9586	2.9682
230	$C_{13}H_{27}$	$C_{13}H_{27}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	3.3872	3.3795
231	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	25	3.7447	3.7521
232	$C_{13}H_{27}$	$C_{11}H_{23}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	3.0555	3.0227
233	$C_{14}H_{29}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	3.0757	3.0227
234[t]	$C_{16}H_{32}$	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	3.1249	3.1168
235[t]	$C_{18}H_{37}$	C_6H_{13}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	3.2291	3.2210
236[t]	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	1.5086	1.5983
237	C_9H_{19}	C_9H_{19}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	1.7447	1.7368
238	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	2.0757	2.0835
239	$C_{11}H_{23}$	$C_{11}H_{23}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	2.5229	2.5146
240	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	2.9586	2.9639

Table A.2 Continued

Nos.	Compounds					T(°C)	pCMC	
							Observed	Predicted [a]
	Tail 1	Tail 2	Ion-1	Ion-2	Spacer			
241	$C_{13}H_{27}$	$C_{13}H_{27}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	3.4815	3.4116
242	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	3.9208	3.8744
243	$C_{15}H_{31}$	$C_{15}H_{31}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	4.3665	4.3748
244	$C_{16}H_{33}$	$C_{16}H_{33}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2NCH_3$	25	4.9208	4.9128
245	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	25	1.3565	1.1276
246	C_9H_{19}	C_9H_{19}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	25	1.6576	1.6656
247	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	25	2.1135	2.0940
248	$C_{11}H_{23}$	$C_{11}H_{23}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	25	2.5850	2.5766
249	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	25	3.0000	3.0828
250	$C_{13}H_{27}$	$C_{13}H_{27}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	25	3.4203	3.5072
251	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	25	3.7212	3.7131
252	$C_{16}H_{33}$	$C_{16}H_{33}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	25	3.1938	3.2014
253	C_6H_{13}	C_6H_{13}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	25	0.8861	0.8781
254	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	25	1.4318	1.4394
255	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	25	2.2076	2.1992
256	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	25	3.0177	3.0254
257	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	25	4.0000	4.0887
258[t]	$C_{16}H_{33}$	$C_{16}H_{33}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	25	5.0132	5.1763
259	$C_{18}H_{37}$	$C_{18}H_{37}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	25	6.0000	5.9921
260	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$CH_2C \equiv CCH_2$	25	2.8570	2.8487
261	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$CH_2C \equiv CCH_2$	25	1.5935	1.6012
262	$C_{12}H_{25}$	$C_{12}H_{25}$			$CH_2C \equiv CCH_2$	25	3.0506	3.0589
263[t]	$C_{12}H_{25}$	$C_{12}H_{25}$			$CH_2C \equiv CCH_2$	25	3.0410	3.0494
264	C_8H_{17}	C_8H_{17}			$CH_2C \equiv CCH_2$	25	1.4634	1.8067

Table A.2 Continued

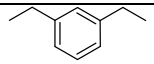
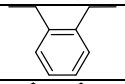
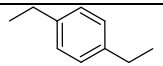
Nos.	Compounds					T(°C)	pCMC	
							Observed	Predicted [a]
	Tail 1	Tail 2	Ion-1	Ion-2	Spacer			
265	C_8H_{17}	C_8H_{17}			$CH_2C \equiv CCH_2$	25	1.3206	1.3289
266	$C_{10}H_{21}$	$C_{10}H_{21}$			$CH_2C \equiv CCH_2$	25	2.2757	2.1636
267	$C_{10}H_{21}$	$C_{10}H_{21}$			$CH_2C \equiv CCH_2$	25	2.2248	2.2168
268	$C_{14}H_{29}$	$C_{14}H_{29}$			$CH_2C \equiv CCH_2$	25	3.8861	3.8782
269	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$		25	1.6459	1.1140
270	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$		25	1.6289	1.6370
271	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$		25	1.6108	1.6187
272	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$((CH_2CH(OH))_2$	25	2.4318	2.4237
273[t]	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$((CH_2CH(OH))_2$	25	3.1549	3.1631
274	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$((CH_2CH(OH))_2$	25	4.0706	4.0624
275[t]	$C_{16}H_{33}$	$C_{16}H_{33}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$((CH_2CH(OH))_2$	25	4.3010	4.0225
276[t]	$C_{10}H_{21}$	$C_{10}H_{21}$			$(CH_2)_4$	25	2.5702	2.5785
277	$C_{12}H_{25}$	$C_{12}H_{25}$			$(CH_2)_4$	25	3.2518	3.2437
278	$C_{10}H_{21}$	$C_{10}H_{21}$			$(CH_2)_6$	25	2.6990	2.7071
279	$C_{12}H_{25}$	$C_{12}H_{25}$			$(CH_2)_6$	25	3.2676	3.2597
280	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4O)_2C_2H_4$	25	3.0362	3.0282
281	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4O)_3C_2H_4$	25	3.0655	3.0734
282	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4O)_4C_2H_4$	25	3.0458	3.0470

Table A.2 Continued

Nos.	Compounds					T(°C)	pCMC	
							Observed	Predicted [a]
	Tail 1	Tail 2	Ion-1	Ion-2	Spacer			
283	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4O)_5C_2H_4$	25	2.9101	3.0226
284	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	25	3.8013	3.7937
285[t]	$C_8H_{17}(OC_2H_4)_5OH$					25	2.0362	2.0400
286	$C_{12}H_{25}OC_2H_4OH$					25	4.6380	4.6298
287	$C_{12}H_{25}(OC_2H_4)_2OH$					25	4.2596	4.2675
288	$C_{12}H_{25}(OC_2H_4)_3OH$					25	4.2840	4.2763
289[t]	$C_{12}H_{25}(OC_2H_4)_4OH$					25	4.1938	4.2104
290	$C_{12}H_{25}(OC_2H_4)_5OH$					25	4.1938	4.0989
291	$C_{12}H_{25}(OC_2H_4)_7OH$					25	4.0862	3.9851
292	$C_{12}H_{25}(OC_2H_4)_8OH$					25	3.9626	3.9654
293	$C_{12}H_{25}(OC_2H_4)_9OH$					25	4.0000	3.9577
294	$C_{12}H_{25}(OC_2H_4)_{12}OH$					25	3.8539	3.9743
295	$C_{12}H_{25}(OC_2H_4)_{23}OH$					25	4.2218	4.2297
296	$C_{13}H_{27}(OC_2H_4)_6OH$					25	4.0993	4.5256
297	$C_{13}H_{27}(OC_2H_4)_{10}OH$					25	3.9370	4.4380
298	$C_{16}H_{33}(OC_2H_4)_2OH$					25	7.1739	6.8950
299[t]	$C_{16}H_{33}(OC_2H_4)_6OH$					25	5.7799	5.7719
300	$C_{16}H_{33}(OC_2H_4)_7OH$					25	5.7696	5.7415
301	$C_{16}H_{33}(OC_2H_4)_8OH$					25	5.9210	5.7142
302	$C_{16}H_{33}(OC_2H_4)_{10}OH$					25	5.6990	5.6547
303	$C_{16}H_{33}(OC_2H_4)_{15}OH$					25	5.5086	5.5008
304	$C_{16}H_{33}(OC_2H_4)_{20}OH$					25	5.1549	5.3866
305[t]	$C_{18}H_{37}(OC_2H_4)_2OH$					25	6.6021	6.6100
306	$C_{18}H_{37}(OC_2H_4)_{10}OH$					25	5.5229	6.0381
307	$C_{18}H_{37}(OC_2H_4)_{20}OH$					25	5.2441	5.4513
308	$C_{18}H_{37}(OC_2H_4)_{21}OH$					25	5.4089	5.4008

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
309	$C_8H_{17}CHCHC_8H_{16}(OC_2H_4)_2OH$	25	1.6048	1.6127
310	$C_8H_{17}CHCHC_8H_{16}(OC_2H_4)_{10}OH$	25	3.0269	3.0348
311[t]	$C_{17}H_{34}OC(OC_2H_4)_8OH$	25	3.4283	3.4203
312	$C_8H_{17}C_6H_4(OC_2H_4)_1OH$	25	4.3050	3.7319
313	$C_8H_{17}C_6H_4(OC_2H_4)_3OH$	25	4.0132	4.0053
314	$C_8H_{17}C_6H_4(OC_2H_4)_4OH$	25	3.8861	3.8778
315[t]	$C_8H_{17}C_6H_4(OC_2H_4)_5OH$	25	3.8239	3.7694
316	$C_8H_{17}C_6H_4(OC_2H_4)_6OH$	25	3.6778	3.6861
317	$C_8H_{17}C_6H_4(OC_2H_4)_7OH$	25	3.6021	3.6157
318	$C_8H_{17}C_6H_4(OC_2H_4)_8OH$	25	3.5528	3.5579
319[t]	$C_8H_{17}C_6H_4(OC_2H_4)_9OH$	25	3.5229	3.5147
320	$C_8H_{17}C_6H_4(OC_2H_4)_{10}OH$	25	3.4815	3.4867
321	$C_8H_{17}C_6H_4(OC_2H_4)_{30}OH$	25	3.9590	3.9510
322[t]	$C_8H_{17}C_6H_4(OC_2H_4)_{40}OH$	25	4.1190	4.1270
323[t]	$C_9H_{19}C_6H_4(OC_2H_4)_2OH$	25	3.3770	3.3852
324	$C_9H_{19}C_6H_4(OC_2H_4)_5OH$	25	3.3280	4.4401
325	$C_9H_{19}C_6H_4(OC_2H_4)_{10}OH$	25	4.1249	4.1169
326	$C_9H_{19}C_6H_4(OC_2H_4)_{12}OH$	25	3.3010	4.0327
327	Tween-65	25	6.7447	6.7366
328	Tween-85	25	6.5376	6.5297
329	Tween-60	25	4.6778	4.4795
330	Tween-80	25	5.0000	4.9922
331	Tween-40	25	4.6383	4.6467
332	Tween-20	25	4.3010	4.2929
333	$C_{13}H_{27}(OC_2H_4)_4OH$	25	4.1602	4.7446
334	$C_{13}H_{27}(OC_2H_4)_5OH$	25	4.1191	4.6107
335[t]	$C_{13}H_{27}(OC_2H_4)_7OH$	25	4.0931	4.4785

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
336	$C_{13}H_{27}(OC_2H_4)_8OH$	25	4.0706	4.4539
337	$C_6H_{13}(OC_2H_4)_3OH$	25	0.9800	0.8956
338	$C_6H_{13}(OC_2H_4)_4OH$	25	1.0320	1.0401
339	$C_6H_{13}(OC_2H_4)_5OH$	25	1.0170	1.0249
340[t]	$C_6H_{13}(OC_2H_4)_6OH$	20	1.1308	1.1389
341[t]	$C_8H_{17}(OC_2H_4)_4OH$	25	2.0630	2.0712
342	$C_8H_{17}(OC_2H_4)_6OH$	25	2.0044	1.9962
343	$C_{10}H_{21}(OC_2H_4)_4OH$	25	3.1675	3.1753
344[t]	$C_{10}H_{21}(OC_2H_4)_5OH$	25	3.1192	3.1079
345	$C_{10}H_{21}(OC_2H_4)_6OH$	25	3.0458	3.0540
346[t]	$C_{12}H_{25}(OC_2H_4)_6OH$	25	4.0605	4.0522
347[t]	$C_{14}H_{29}(OC_2H_4)_6OH$	25	5.0000	5.0077
348	$C_{16}H_{33}(OC_2H_4)_9OH$	25	5.6778	5.6855
349	$C_4H_9(OC_2H_4)_6OH$	20	0.0969	0.6587
350	$C_8H_{17}OC_2H_4OH$	25	2.3098	2.3018
351	$C_8H_{17}(OC_2H_4)_3OH$	25	2.1249	2.0660
352	$C_{10}H_{21}(OC_2H_4)_8OH$	25	3.0000	2.9917
353[t]	$C_{11}H_{23}(OC_2H_4)_8OH$	25	3.5229	3.4768
354	$C_{14}H_{29}(OC_2H_4)_8OH$	25	5.0458	4.9264
355	$C_{15}H_{31}(OC_2H_4)_8OH$	25	5.4559	5.3569
356	$C_{16}H_{33}(OC_2H_4)_{12}OH$	25	5.6383	5.5906
357	$C_{16}H_{33}(OC_2H_4)_{21}OH$	25	5.4089	5.3682
358[t]	$C_8H_{17}C_6H_4(C_2H_4O)_2OH$	25	3.8861	3.8942
359	$(CH_3)_2CHCH_2(OC_2H_4)_6OH$	20	0.0410	0.0491
360	$(C_2H_5)_2CHCH_2(OC_2H_4)_6OH$	20	1.0000	1.0079
361[t]	$(C_3H_7)_2CHCH_2(OC_2H_4)_6OH$	20	1.6383	1.6463
362[t]	$(C_4H_9)_2CHCH_2(OC_2H_4)_6OH$	20	2.5086	2.5006

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
363	$(C_4H_9)_2CHCH_2(OC_2H_4)_9OH$	20	2.4949	2.5027
364	$C_6F_{13}C_2H_4(OC_2H_4)_4OH$	20	3.2147	3.2069
365	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_2OH$	25	2.6021	2.6097
366[t]	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_3OH$	25	2.5528	2.7796
367	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_5OH$	25	2.4318	2.4235
368	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_7OH$	25	2.3188	2.3268
369	$C_4H_9(OC_2H_4)_1OH$	25	0.0090	0.0169
370	$C_4H_9(OC_2H_4)_6OH$	25	0.1100	0.1181
371[t]	$C_8H_{17}(OC_2H_4)_1OH$	25	2.3100	2.3018
372	$C_8H_{17}(OC_2H_4)_3OH$	25	2.1250	2.0660
373	$C_8H_{17}(OC_2H_4)_9OH$	25	1.8860	1.8943
374	$C_9H_{19}(OC_2H_4)_8OH$	25	2.5200	2.5017
375	$C_{10}H_{21}(OC_2H_4)_3OH$	25	3.2220	3.1203
376	$C_{10}H_{21}(OC_2H_4)_7OH$	25	3.0150	3.0168
377	$C_{10}H_{21}(OC_2H_4)_9OH$	25	2.8860	2.9767
378	$C_{12}H_{25}(OC_2H_4)_{14}OH$	25	4.2600	4.0064
379	$C_{14}H_{25}(OC_2H_4)_9OH$	25	5.0460	4.9076
380[t]	$C_{16}H_{33}(OC_2H_4)_{12}OH$	25	5.6380	5.5906
381	$C_8H_{17}C_6H_4(OC_2H_4)_2OH$	25	4.1160	3.8942
382	$C_8H_{17}OCH_2CHOHCH_2OH$	25	2.2370	2.2289
383	$C_8H_{17}CHOHCH_2OH$	25	2.6380	2.6298
384	$C_{10}H_{21}CHOHCH_2OH$	25	3.7450	3.7528
385	$C_8H_{17}CHOHCH_2CH_2OH$	25	2.6380	2.6456
386	$C_{10}H_{21}CHOHCH_2CH_2OH$	25	4.8860	4.8777
387[t]	$C_{11}H_{23}C(O)N(C_2H_4OH)_2$	25	3.5850	3.5772
388	$C_9H_{19}C(O)N((C_2H_4O)_3NCH_3)_2$	25	2.2990	1.8184
389[t]	$C_9H_{19}C(O)N((C_2H_4O)_4NCH_3)_2$	25	2.1930	2.2013

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
390	$C_{11}H_{23}C(O)N((C_2H_4O)_2NCH_3)_2$	25	3.3980	3.3898
391	$C_{11}H_{23}C(O)N((C_2H_4O)_3NCH_3)_2$	25	3.2920	3.3002
392	$C_{11}H_{23}C(O)N((C_2H_4O)_4NCH_3)_2$	25	3.6110	3.6028
393	$C_{12}H_{25}NHCH(CH_3)COON(C_2H_4O)_4H$	25	3.4130	3.4051
394	$C_{12}H_{25}NHCH_2COON(C_2H_4O)_4H$	25	3.4740	3.4820
395	$C_{12}H_{25}N(CH_3)CH_2COON(C_2H_4O)_4H$	25	3.5330	3.5409
396	$C_{12}H_{25}C(O)N(CH_3)C_2H_4O(C_2H_4O)_3H$	25	3.2920	3.2999
397	$C_{12}H_{25}C(O)N(CH_3)C_2H_4O(C_2H_4O)_6H$	25	3.1870	3.1788
398	$C_{12}H_{25}C(O)N(CH_3)C_2H_4O(C_2H_4O)_9H$	25	3.1250	3.1331
399	$C_6F_{13}CH_2C(O)N((C_2H_4O)_3CH_3)_2$	25	3.2600	3.2519
400	$C_8F_{17}CH_2C(O)N((C_2H_4O)_3CH_3)_2$	25	4.9210	4.9295
401	$C_{10}F_{21}CH_2C(O)N((C_2H_4O)_3CH_3)_2$	25	6.5230	6.5150
402	$C_6F_{13}C_2H_4(SC_2H_4OC_2H_4)_1(SC_2H_4OC_2H_4)_2OH$	25	4.6380	4.5928
403[t]	$C_6F_{13}C_2H_4(SC_2H_4OC_2H_4)_2(SC_2H_4OC_2H_4)_1OH$	25	4.5850	4.5928
404	$C_6F_{13}C_2H_4(SC_2H_4OC_2H_4)_3(SC_2H_4OC_2H_4)_1OH$	25	4.4690	4.4611
405	$H(CF_2)_4CH_2(C_2H_4O)_3CH_3$	25	2.0970	2.1051
406	$H(CF_2)_6CH_2(C_2H_4O)_3CH_3$	25	3.5230	3.1740
407	$F(CF_2)_4CH_2(C_2H_4O)_3CH_3$	25	2.6990	2.6909
408	$F(CF_2)_6CH_2(C_2H_4O)_3CH_3$	25	4.0970	4.0894
409	$F(CF_2)_4C_3H_6NHC(O)(OC_2H_4)_2CH_3$	25	2.0090	2.0170
410[t]	$F(CF_2)_4C_3H_6NHC(O)(OC_2H_4)_3CH_3$	25	2.8540	2.8463
411	$F(CF_2)_6C_3H_6NHC(O)(OC_2H_4)_2CH_3$	25	3.8240	3.8160
412	$F(CF_2)_6C_3H_6NHC(O)(OC_2H_4)_3CH_3$	25	4.0460	4.0539
413	$F(CF_2)_8C_3H_6NHC(O)(OC_2H_4)_2CH_3$	25	4.6200	4.6280
414	$F(CF_2)_8C_3H_6NHC(O)(OC_2H_4)_3CH_3$	25	4.9590	4.9672
415	$C_{17}H_{35}COO(CH_2CHOHCH_2O)_4H$	25	4.4840	4.4764
416	$C_{17}H_{35}COO(CH_2CHOHCH_2O)_6H$	25	4.5620	4.5702

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
417[t]	$C_{17}H_{35}COO(CH_2CHOHCH_2O)_{10}H$	25	4.6760	4.6679
418	$C_{11}H_{23}COO(CH_2CHOHCH_2O)_4H$	25	4.4020	4.3938
419	$C_{11}H_{23}COO(CH_2CHOHCH_2O)_6H$	25	4.4460	4.4541
420[t]	$C_{11}H_{23}COO(CH_2CHOHCH_2O)_{10}H$	25	4.5490	4.5409
421	$C_4H_9(OC_2H_4)_6OH$	40	0.1487	0.1569
422	$(CH_3)_2CHCH_2(OC_2H_4)_6OH$	40	0.0706	0.0784
423	$C_6H_{13}(OC_2H_4)_6OH$	40	1.2840	1.2919
424	$(C_2H_5)_2CHCH_2(OC_2H_4)_6OH$	40	1.0605	2.1393
425	$C_{10}H_{21}(OC_2H_4)_8OH$	15	2.8539	2.8456
426	$C_{10}H_{21}(OC_2H_4)_8OH$	40	3.1192	3.1130
427	$C_{11}H_{23}(OC_2H_4)_8OH$	15	3.3979	3.3938
428	$C_{11}H_{23}(OC_2H_4)_8OH$	40	3.6383	3.6300
429	$C_{12}H_{25}(OC_2H_4)_2OH$	10	4.4202	4.4121
430	$C_{12}H_{25}(OC_2H_4)_2OH$	40	4.4949	4.4873
431	$C_{12}H_{25}(OC_2H_4)_3OH$	10	4.2007	4.1683
432[t]	$C_{12}H_{25}(OC_2H_4)_3OH$	40	4.2518	4.2437
433	$C_{12}H_{25}(OC_2H_4)_4OH$	10	4.0862	4.0940
434	$C_{12}H_{25}(OC_2H_4)_4OH$	40	4.2291	4.2374
435	$C_{12}H_{25}(OC_2H_4)_5OH$	10	4.0458	4.0532
436	$C_{12}H_{25}(OC_2H_4)_5OH$	40	4.2291	4.2208
437	$C_{12}H_{25}(OC_2H_4)_7OH$	10	3.9172	3.9088
438	$C_{12}H_{25}(OC_2H_4)_7OH$	40	4.1367	4.1718
439	$C_{12}H_{25}(OC_2H_4)_8OH$	10	3.8069	3.8147
440	$C_{12}H_{25}(OC_2H_4)_8OH$	40	4.0315	4.1575
441[t]	$C_{13}H_{27}(OC_2H_4)_8OH$	10	4.4949	4.4866
442	$C_{13}H_{27}(OC_2H_4)_8OH$	40	4.6990	4.6683
443	$C_{14}H_{29}(OC_2H_4)_8OH$	15	4.9586	4.9669

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
444	$C_{14}H_{29}(OC_2H_4)_8OH$	40	5.1427	5.1345
445[t]	$C_{15}H_{31}(OC_2H_4)_8OH$	15	5.3872	5.3952
446[t]	$C_{15}H_{31}(OC_2H_4)_8OH$	40	5.5229	5.5307
447	Octyl-D,L-Glycerol	25	2.2366	2.2289
448[t]	Octyl Glycol	25	2.3098	2.3018
449	Octyl-β-D-Glucoside	25	1.6990	1.7343
450	Octyl-α-D-Mannoside	25	2.2218	1.7343
451	Octyl-β-D-Galactoside	25	1.7959	1.7343
452	Decyl-α-D-Mannoside	25	3.6021	3.1467
453[t]	Decyl-β-D-Galactoside	25	3.1549	3.1467
454[t]	Dodecyl-α-D-Mannoside	25	4.3010	3.7293
455	Dodecyl-β-D-Galactoside	25	3.6990	3.7293
456[t]	1-O-Butyl-D,L-Xylitol	25	1.2366	1.2272
457[t]	1-O-Pentyl-D,L-Xylitol	25	1.4202	1.4326
458	1-O-Hexyl-D,L-Xylitol	25	2.0269	2.0346
459	1-O-Heptyl-D,L-Xylitol	25	2.0362	2.0284
460[t]	1-O-Octyl-D,L-Xylitol	25	2.1739	2.1818
461	1-O-Nonyl-D,L-Xylitol	25	2.6778	2.6860
462	1-O-Pentanoyl-D,L-Xylitol	25	0.9208	0.9289
463	1-O-Hexanoyl-D,L-Xylitol	25	1.2366	1.5528
464	1-O-Heptanoyl-D,L-Xylitol	25	2.0000	1.9922
465	1-O-Octanoyl-D,L-Xylitol	25	1.7447	1.7527
466	1-O-Nonanoyl-D,L-Xylitol	25	2.3565	2.3641
467	1-O-Decanoyl-D,L-Xylitol	25	2.7447	2.7366
468	S-Butyl-1-Thio-D,L-Xylitol	25	0.7447	0.7531
469[t]	S-Pentyl-1-Thio-D,L-Xylitol	25	1.3372	2.1280
470	S-Hexyl-1-Thio-L-Xylitol	20	1.9208	1.9992

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
471	S-Octyl-1-Thio-L-Xylitol	20	2.9208	2.9127
472	S-Decyl-1-Thio-L-Xylitol	20	3.3979	1.9992
473	S-Hexyl-1-Thio-D-Lyxitol	20	1.7447	2.9127
474	S-Octyl-1-Thio-D-Lyxitol	20	2.7447	2.7527
475[t]	S-Hexyl 5-Thio-D-Arabinonolactone	20	2.1739	2.1821
476	S-Octyl 5-Thio-D-Arabinonolactone	20	3.3188	3.3269
477	S-Decyl 5-Thio-D-Arabinonolactone	20	4.4815	3.5576
478	S-Hexyl 5-Thio-D-Xylonolactone	20	2.3010	2.3512
479	S- Octyl 5-Thio-D-Xylonolactone	20	3.2757	3.2839
480	S- Decyl 5-Thio-D-Xylonolactone	20	4.6383	4.6303
481	1-Butylhexyl- β-D-Glucoside	25	1.8239	1.8314
482	1-Propylheptyl- β-D-Glucoside	25	1.9208	1.8686
483[t]	1-Ehyloctyl- β-D-Glucoside	25	2.0706	2.0787
484	1-Methynonyl- β-D-Glucoside	25	2.3468	2.3548
485[t]	3,7-Dimethyloctyl- β-D-Glucoside	25	2.3979	2.4064
486	3,7-Dimethyloctyl- β-D-Maltoside	25	2.2757	2.2833
487[t]	3,7-Dimethyloctyl- β-D-Maltotrioside	25	2.3010	1.8860
488	2-Amino-2-Deoxy-Octyl- β-D-Glucoside	25	1.6383	2.2585
489	2-Amino-2-Deoxy-Nonyl- β-D-Glucoside	25	2.1549	2.8943
490[t]	N-Decyl-N-Methyl Gluconamide	25	2.8861	2.8943
491	N-Dodecyl-N-Methyl Gluconamide	20	3.8539	3.8847
492[t]	N-Tetradecyl-N-Methyl Gluconamide	20	4.6198	4.4459
493	N-Oleyl-N-Methyl Gluconamide	20	4.4949	4.4869
494[t]	N-Decyl-N-Methyl Lactobionamide	20	2.6383	2.6459
495	N-Dodecyl-N-Methyl Lactobionamide	20	3.6021	3.5937
496	N-Tetradecyl-N-Methyl Lactobionamide	20	4.4437	4.3953
497[t]	N-Hexadecyl-N-Methyl Lactobionamide	20	5.0315	5.0238

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
498	N-Octadecyl-N-Methyl Lactobionamide	20	5.4815	5.4765
499[t]	N-Oleyl-N-Methyl Lactobionamide	20	4.2676	4.2594
500	Dodecanoyl-N-Methylxylamine	25	3.4437	3.4521
501[t]	Octanoyl-N-Methylglucamine	25	1.1612	1.1691
502	Nonanoyl-N-Methylglucamine	25	1.6778	1.6860
503	Decanoyl-N-Methylglucamine	25	2.1739	2.1660
504	Decyl- β-D-Maltoside	25	2.6990	2.7066
505	Decyl-β-D-Glucoside	25	2.6990	3.1467
506	Dodecyl- β-D-Maltoside	25	3.7696	3.9965
507	Oleyl- β-D-Maltoside	25	4.6990	4.7069
508	Oleyl- β-D-Maltotrioside	25	4.3768	4.3690
509	Hexyl-D-Maltonamide	25	1.0809	1.0885
510	Octyl-D-Maltonamide	25	2.2441	2.3263
511	Decyl-D-Maltonamide	25	2.8861	2.8776
512	Dodecyl-D-Maltonamide	25	3.5086	3.5167
513	Decyl-D-Lactobionamide	25	2.8861	2.8776
514	6’-O-Dodecanoylmaltose	25	3.4815	3.4893
515	N-Decanoyl-N-Methyl Lactitolamine	25	2.4815	2.4893
516	N-Dodecanoyl-N-Methyl Lactitolamine	25	3.3468	3.3385
517	N-Tetradecanoyl-N-Methyl Lactitolamine	25	4.1675	3.9850
518[t]	Nonyl- β-D-Glucoside	25	2.1612	2.2759
519	6-O-Dodecanoylsucrose	20	3.3372	3.1220
520	6-O-Dodecanoylraffinose	20	3.0223	3.0138
521[t]	6-O-Dodecanoylstachyose	20	2.6383	2.7421
522	Octanoyl- β-D-Galactosylamine	25	1.3468	1.3548
523	Octanoyl- β-D-Glucosylamine	25	1.1549	1.3548
524	S-Hexyl-1-Thio-L-Ribitol	20	1.9914	1.9992

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
525	S-Octyl-1-Thio-L-Ribitol	20	3.4202	2.9127
526[t]	Methyl-6-O-(N-Heptylcarbamoyl)- α-D-Glucopyranoside	25	1.7100	2.2065
527[t]	n-Dodecyl- β-D- Glucopyranoside	25	3.7212	3.7293
528	n-Hexadecyl- β-D- Maltopyranoside	25	6.2218	6.2140
529[t]	Hexahedrofarnesyl maltoside	20	4.5258	4.5179
530	Dodecyl- α-D- Glucoside	20	4.3768	4.3686
531	Octyl- β-D- Maltoside	20	1.6345	1.6424
532	Nonyl- β-D- Maltoside	20	2.1871	2.1948
533	Tetradecyl- β-D- Maltoside	20	4.7773	4.9315
534	Hexadecyl- β-D- Maltoside	20	5.8386	5.8311
535	Dodecyl- α-D- Maltoside	20	3.9208	3.9125
536	Heptyl- α-D- Galagtoside	25	1.4881	1.4961
537	Heptyl- β-D- Galagtoside	25	1.7570	1.4961
538	Octyl- α-D- Galagtoside	25	1.7258	1.7343
539	Heptyl- α-D- Mannoside	25	1.5143	1.4961
540[t]	Nonyl- α-D- Galagtoside	25	2.2840	2.2759
541	Nonyl - α-D- Mannoside	25	2.3979	2.2759
542	Heptyl- β-D- Glucoside	25	1.1409	1.4961
543	N-Octyl-D-Glucuronamide	25	2.4815	2.4736
544	N-Octyl-D-Galacturonamide	25	2.3979	2.4736
545	Hexyl- β-D- Glucoside	25	0.6445	0.6526
546	Hexyl - β-D- Thioglucoside	25	1.0857	1.0941
547	Heptyl - β-D- Thioglucoside	25	1.5654	1.5576
548	Octyl - β-D- Thioglucoside	25	1.6737	1.6820
549	Nonyl - β-D- Thioglucoside	25	2.5528	2.4840
550	Decyl - β-D- Thioglucoside	25	3.0458	3.0538
551	Octyl - β-D- Thiomaltoside	25	2.0862	2.0943

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
552[t]	Decyl - β-D- Thiomaltoside	25	3.2218	3.2142
553	3-O-Octyl-D-Glucose	25	2.8539	2.8462
554	3-O-Octanoyl-D-Glucose	25	3.2147	3.2067
555[t]	3-O-n-$C_{12}H_{25}$-O-(α-PP-O)$_6$-D-Glucopyranose	25	4.3979	4.3899
556	3-O-n-$C_{12}H_{25}$-O-(α-PP-O)$_6$-C_2H_4-O-C_2H_4-D-Glucopyranose	25	3.9586	3.9670
557	3-O-n-$C_{12}H_{25}$-O-(α-PP-O)$_{10}$-C_2H_4-O-C_2H_4-D-Glucopyranose	25	4.3979	4.4062
558[t]	3-O-n-$C_{12}H_{25}$-O-(α-PP-O)$_{14}$-C_2H_4-O-C_2H_4-D-Glucopyranose	25	4.6990	4.6912
559[t]	6-O-n-$C_{12}H_{25}$-O-(α-PP-O)$_6$-C_2H_4-O-C_2H_4-D-Glucopyranose	25	4.3010	3.8738
560	6-O-n-$C_{12}H_{25}$-O-(α-PP-O)$_{10}$-C_2H_4-O-C_2H_4-D-Glucopyranose	25	4.5229	4.5151
561	$C_{12}H_{25}$O(α-PP-O)$_6C_2H_4OC_2H_4OCH_2$-butane-1,2,3,4 tetrol	25	4.6990	4.6910
562	$C_{12}H_{25}$O(α-PP-O)$_{10}C_2H_4OC_2H_4OCH_2$-butane-1,2,3,4 tetrol	25	4.6990	5.5349
563[t]	6-O-n-$C_8H_{17}(C_6O_5H_{11})$	25	2.4685	2.4769
564	6-O-n-$C_{10}H_{21}(C_6O_5H_{11})$	25	3.4318	3.4237
565	$C_8H_{17}OCH_2$-butane-1,2,3,4 tetrol	25	2.1739	2.1818
566[t]	$C_{10}H_{21}OCH_2$-butane-1,2,3,4 tetrol	25	3.0915	3.0834
567	3-O-n-C_8H_{17}- D-Glucopyranose	25	2.9208	2.5668
568	3-O-n-$C_{10}H_{21}$- D-Glucopyranose	25	3.1079	3.1004
569	3-O-n-$C_{12}H_{25}$- D-Glucopyranose	25	3.6383	3.5114
570	3-O-n-$C_{14}H_{29}$- D-Glucopyranose	25	3.7447	3.7528
571	$C_{10}H_{21}N^+(CH_3)_2CH_2COO^-$	23	1.7447	1.7528
572	$C_{12}H_{25}N^+(CH_3)_2CH_2COO^-$	25	2.6990	2.6907
573	$C_{14}H_{29}N^+(CH_3)_2CH_2COO^-$	25	3.6576	3.6659
574	$C_{16}H_{32}N^+(CH_3)_2CH_2COO^-$	23	4.6990	4.6911
575[t]	$C_{12}H_{25}N^+(CH_3)_2(CH_2)_3COO^-$	25	2.3372	2.6470
576	$C_{12}H_{25}N^+(CH_3)_2(CH_2)_5COO^-$	25	2.5850	2.5767
577	$C_{12}H_{25}N^+(CH_3)_2(CH_2)_7COO^-$	25	2.8239	2.8318
578	$C_{10}H_{21}CH(Pyr^+)COO^-$	25	2.2840	2.9619

Table A.2 Continued

Nos.	Compounds	T(°C)	pCMC	
			Observed	Predicted [a]
579	$C_{12}H_{25}CH(Pyr^+)COO^-$	25	3.2218	3.2297
580	$C_{10}H_{21}N^+(CH_3)(CH_2C_6H_5)CH_2COO^-$	25	2.2757	2.2839
581	$C_{12}H_{25}N^+(CH_3)(CH_2C_6H_5)CH_2COO^-$	25	3.2596	2.4810
582[t]	$C_{12}H_{25}N^+(CH_3)_2(CH_2)_3SO_3^-$	25	2.5229	2.5309
583	$C_{14}H_{29}N^+(CH_3)_2(CH_2)_3SO_3^-$	25	3.4949	3.4869
584	$C_8H_{17}N^+(CH_3)_2CH_2COO^-$	27	0.6021	0.6099
585	$C_8H_{17}CH(COO^-)N^+(CH_3)_3$	27	1.0132	1.0212
586	$C_8H_{17}CH(COO^-)N^+(CH_3)_3$	60	1.0655	1.0732
587	$C_{10}H_{21}CH(COO^-)N^+(CH_3)_3$	27	1.8861	2.1187
588[t]	$C_{12}H_{25}CH(COO^-)N^+(CH_3)_3$	27	2.8861	2.8781
589[t]	$p-C_{12}H_{25}Pyr^+COO^-$	50	2.7212	2.7131
590[t]	$m-C_{12}H_{25}Pyr^+COO^-$	50	2.8239	2.8156
591	$C_{14}H_{29}CH(Pyr^+)COO^-$	40	4.1308	4.1232
592	$C_{10}H_{21}N^+(CH_3)(CH_2C_6H_5)CH_2COO^-$	40	2.3565	2.3649
593[t]	$C_{10}H_{21}N^+(CH_3)(CH_2C_6H_5)C_2H_4SO_3^-$	40	2.3372	2.3451

178-284 di-bromide Gemini surfactants, Pyr^+ :pyridinium, [t] the compounds were in the test set, and the rest were in the training set, [a]pCMC values predicted by DA-SVR model

Table A.3 Weights and bias between input and hidden layer.

j	$W_{j.1}$	$W_{j.2}$	$W_{j.3}$	$W_{j.4}$	$W_{j.5}$	$W_{j.6}$	$W_{j.7}$	$W_{j.8}$	$W_{j.9}$	$W_{j.10}$	$W_{j.11}$	$W_{j.12}$	$W_{j.13}$	$W_{j.14}$	$W_{j.15}$	b_j
1	-0.0436	-1.1112	1.0772	-0.9539	-0.0225	-0.7530	-0.4353	-0.4585	-0.0968	0.5681	-0.2310	-1.0391	0.3588	0.0579	0.5066	-0.0404
2	-0.1581	1.4168	-0.8784	0.3794	-0.4771	0.4291	0.6731	0.1263	0.2820	-0.0257	0.0512	-0.3597	-0.3589	0.3470	0.1897	0.2412
3	-0.0239	0.6647	0.3799	-0.6923	-0.8288	0.2211	0.3899	0.5898	-0.2002	-0.7846	0.3648	-0.5412	-0.3917	-0.0086	0.0653	0.0394
4	1.1252	-1.2401	0.7080	0.1303	0.1198	-0.1617	-0.3487	-0.5382	-0.3663	-0.1821	0.6153	1.2523	-0.5850	-0.2897	0.1449	0.7050
5	-0.1115	0.5510	-0.0649	0.4721	0.2652	0.2758	0.5349	-1.2269	-0.5439	0.3821	0.2202	0.9440	-0.1439	0.2535	-0.0187	0.6070
6	-0.2956	-0.1130	-1.3200	-1.0503	0.8023	0.1090	0.3248	0.4289	0.3880	0.7905	0.3967	-0.0613	-0.5818	0.1274	-0.0391	0.0957
7	1.0659	-0.1079	-0.7503	-0.6031	0.5220	-0.0536	0.2135	1.7801	-0.0637	0.8271	-0.5463	0.2731	0.2886	-0.4789	0.1935	1.5688
8	0.3835	-0.5331	-0.2412	-0.0669	0.5873	0.0430	-0.0706	0.3147	0.4101	1.1515	0.6632	-0.4847	-0.8299	1.1094	0.1172	0.6817
9	0.8836	0.4952	0.4958	0.0552	0.7458	-0.0497	-0.0421	-0.2660	-0.4012	0.6598	-0.5374	-0.4473	0.2037	1.1221	0.0408	0.3083
10	-1.5633	-0.3284	-0.3712	-0.0242	-0.0330	0.0934	0.1552	0.9712	0.0858	0.6320	0.1800	0.0928	0.2864	-0.7698	-0.0119	-0.8114

Table A.4 Weights and bias between hidden and output layers.

$W_{1,1}$	$W_{1,2}$	$W_{1,3}$	$W_{1,4}$	$W_{1,5}$	$W_{1,6}$	$W_{1,7}$	$W_{1,8}$	$W_{1,9}$	$W_{1,10}$	b_1
-1.1056	-1.5484	1.4467	-1.5142	1.5129	-1.4354	1.6972	1.6063	-1.6555	-1.5340	-0.5226