PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
YAHIA FARES UNIVERSITY OF MEDEA



Faculty of Technology
Master's Thesis
Department of Process Engineering
Specialty: Environmental Process Engineering

QSPR MODELS FOR PREDICTING CRITICAL MICELLE CONCENTRATION OF GEMINI SURFACTANTS

Presented by

Miss. TOUMACHE Asma
Miss. YAHIAOUI Nour El Houda

Supervised by

Dr. LAIDI Maamar

Promotion 2019-2020

ACKNOWLEGMENTS

In the Name of Allah, the Most Merciful, the Most Compassionate all praise be to Allah, the Lord of the worlds; and prayers and peace be upon Mohamed his servant and messenger.

First and foremost, we must acknowledge our limitless thanks to Allah, the Ever-Magnificent; the Ever-Thankful, for his help and bless.

We owe a deep debt of gratitude to our University, Dr Yahia Fares, for giving us an opportunity to complete this work.

We are grateful to some people, who worked hard with us from the beginning till the completion of the present research particularly our supervisor Dr. Laidi Maamar.

Additionally we address our sincere thanks to the jury for having kindly examined and jugged this work.

We would like to express our wholehearted thanks to our families for their generous support they provided us throughout our entire life and particularly through the process of pursuing the master degree. Because of their unconditional love and prayers, we have the chance to complete this thesis.

Furthermore, we would like to take this opportunity to say warm thanks to all our Beloved friends, who have been so supportive along the way of doing our thesis.

Last but not least, deepest thanks go to all people who took part in making this thesis real.

Dedication

There is not anything impossible because yesterday dreams are today hopes and they can become in fact tomorrow.

I dedicate this thesis to my grandma, my mother and beloved brother .

My uncles. aunts and all my family YAHIAOUI and ALITURKI. For their unconditional love and their constant support which gave me enthusiasm to complete my proposal goal.

Without forgetting my loyal friends, especially MARYEM, SALMA, all close friends, and all my friends from elementary school to university.

Without forgetting a special thanks to Asma Toumache and her family.

Yahiaoui Nour Elhouda

Dedication

Every challenging work needs self-effort as well as guidance of elders especially those who are very close to our heart.

My humble effort I dedicate to my sweet and loving father and mother, whose affection, love, encouragement and prayers of day and night make me able to get such success and honor.

I dedicate my modest work to

My dearest brothers: Nassim, Hichem and Abdelmoumen for their constant encouragement, moral support and love.

To My brother's wife Nesrine and my beloving brother’s son Mohamed Assil.

To the whole Toumache family.

Along with hard working and respected teachers from my entire education sector.

To all my Faithful friends especially my best friend Besma, Lamia, Nawel and Selma and all my promotion 2019.

To my cooreasercher Nor El Houda and all her family.

To all my friends from elementary to university.

To all the people I love.

Toumache Asma

الملخص : نماذج QSPR للتنبؤ بالتركيز CMC لعوامل السطح ال Gemini.

الهدف الرئيسي من هذه الدراسة هو نمذجة لوغاريتم (CMC) لـ 120 عامل خافض للتوتر السطحي ال Gemini . عبر تطبيق علاقة البنية بالهيكل الكمي (QSPR).بحيث تم تطويرها باستخدام برنامج(AlvaDesc / AlvaModel) الذي يوفر بدوره العديد من أساليب التعلم الآلي ومجموعات من الواصفات الجزيئية كأدوات لإنشاء نماذج QSPR.

من اجل بناء نموذج QSPR ، قمنا باستخدام ثلاث تقنيات رئيسية مختلفة لنمذجة الانحدار الخطي والتي تم اختبارها من خلال ربط قيم CMC بخصائصها الهيكلية. وهذه التقنيات هي : نموذج المربعات الصغرى (OLS) ونموذج K-Nearest Neighbors (KNN) بالإضافة الى نموذج الإجماع ,المربعات الصغرى الجزئية (PLS) ولتنفيذ هذه العملية اعتمدنا منهجية مكونة من اربع خطوات رئيسية وهي :

- الخطوة الأولى :جمع وتطوير مجموعة البيانات المنشورة سابقا و التي تحتوي على 120 مركبًا من مركبات Gemini بما في ذلك قيم CMC الخاصة بها .
- الخطوة الثانية : نقل هذه المركبات الى برنامج HYPERCHEM. وتحميلها في برنامج AlvaDesc لحساب واصفات وfingerprint ذات الصلة
- الخطوة الثالثة : استخدام النماذج الأربعة لتطوير نموذج QSPR :

أ) نموذج المربعات الصغرى العادية (OLS).

ب) نموذج K-Nearest Neighbors (KNN).

ج) Consensus Form .

د) المربعات الجزئية الأقل (PLS).

- الخطوة الرابعة : مقارنة النتائج التي توضح أن نموذج OLS هو الأفضل للتنبؤ بـ CMC لخافضات التوتر السطحي ال Gemini .

الكلمات المفتاحية: علاقة البنية بالهيكل الكمي (QSPR) ، تركيز (CMC) ، عوامل التوتر السطحي ل Gemini.

Abstract: QSPR models for predicting critical micelle concentration of Gemini surfactants.

The aim of this work is to model the logarithm of the critical micelle concentration (CMC) of 120 diverse Gemini surfactants using quantitative structure-property relationship (QSPR) method. The QSPR was developed using the AlvaDesc/AlvaModel software which provides several machine learning methods and molecular descriptors sets as tools to build the QSPR models.

To build this QSPR model, we used three main different linear regression modeling techniques which have been tested by relating the CMC values to their structural characteristics. Least Squares model (OLS), K-Nearest Neighbors regression model (KNN), Consensus model and Partial least squares (PLS). To achieve this process, we adopted a methodology consisting of 4 main steps summarized as follows:

- Step 1: Collecting and developing data set from previous published literature containing 120 Gemini compounds including their CMC values.
- Step 2: Transferring these compounds into the HYPERCHEM software and then uploading them into the AlvaDesc software in order to calculate the relevant QSPR descriptors and fingerprints.
- Step 3: Using the four models to develop the QSPR model.
    a) Regular Minor Squares (OLS) model.
    b) Regression model K-nearest neighbor (KNN).
    c) Consensus Form.
    d) Less Partial Squares (PLS).
- Step 4: results comparison which demonstrates that the OLS model is the best for predicting the CMC of Gemini Surfactants.

Keywords: Quantitative Structure Property Relationship(QSPR), Critical Micelle Concentration (CMC), Gemini Surfactants.

Résumé : MODÈLES QSPR POUR PRÉVOIR LA CONCENTRATION DE MICELLE CRITIQUE DES TENSIOACTIFS GEMINI.

Le but de ce travail est de modéliser le logarithme de la concentration micellaire critique (CMC) de 120 tensioactifs Gemini divers en utilisant la méthode quantitative structure-propriété relation (QSPR). Le QSPR a été développé à l'aide du logiciel AlvaDesc / AlvaModel qui fournit plusieurs méthodes d'apprentissage automatique et des ensembles de descripteurs moléculaires comme outils pour construire les modèles QSPR.

Pour construire ce modèle QSPR, nous avons utilisé trois principales techniques de modélisation de régression linéaire qui ont été testées en reliant les valeurs CMC à leurs caractéristiques structurelles. Modèle des moindres carrés (OLS), modèle de régression K-Nearest Neighbours (KNN), modèle de consensus et moindres carrés partiels (PLS). Pour mener à bien ce processus, nous avons adopté une méthodologie composée de 4 étapes principales résumées comme suit:

 Étape 1: Collecte et développement d'un ensemble de données à partir de la littérature publiée précédente contenant 120 composés Gemini, y compris leurs valeurs CMC.

 Étape 2: Transférer ces composés dans le logiciel HYPERCHEM puis les télécharger dans le logiciel AlvaDesc afin de calculer les descripteurs et empreintes digitales QSPR pertinents.

 Étape 3: Utilisation des quatre modèles pour développer le modèle QSPR.

a) Modèle de carrés mineurs réguliers (OLS).

b) Modèle de régression K-plus proche voisin (KNN).

c) Formulaire de consensus.

d) Moins de carrés partiels (PLS).

 Étape 4: comparaison des résultats qui démontre que le modèle OLS est le meilleur pour prédire la CMC des tensioactifs Gemini.

Mots-clés: Relation de propriété de structure quantitative (QSPR), concentration micellaire critique (CMC), tensioactifs Gemini.

List of Abbreviations:

Abbreviation	Description
ANN	Artificial Neural Network
CMC	Critical Micelle Concentration
CNN	Convolutional Neural Network
CO_2	Carbon Dioxide
D_{qi}	Distance between the query molecule/ each i-the Neighbor
DATEM	Diacetyl Tartaric Acid Ester of Monoglyceride
HLB	Hydrophile-Lipophile Balance
K	A Ka Neighbors
KNN	K-Nearest Neighbors regression model
MM+	Molecular Mechanics
N2	Azote gaz
OLS	Least Squares model
PEO	polyethylene oxide
PLS	Partial least squares
QSAR	Quantitative Structure-Activity Relationship
QSBR	Quantitative Structure–Biodegradability Relationship
QSCR	Quantitative Structure–Chromatography Relationship
QSER	Quantitative Structure–Electrochemistry Relationship
QSPR	Quantitative Structure–Property Relationship
QSTR	Quantitative Structure–Toxicity Relationship
R^2	Coefficient of determination
RMSE	Root Mean Squared Error
w_i	weight
(W/O) / (O/W)	water-in-oil / oil-in-water
y_i	the target value of the i-the neighbor

CONTENTS

CHAPTER I

I.1.Introduction:.....1
I.2.Definition:.....5
I.3.Types of surfactants:.....6
I.3.1Anionic Surfactants.....7
I.3.2.Nonionic Surfactants:.....7
I.3.3Cationic Surfactants:.....7
I.3.4.Amphoteric Surfactants:.....7
I.4.Gemini surfactants:.....10
I.4.1. Characteristics of Gemini surfactant:.....10
I.4.2. Structure:.....11
I.4.3.Solubilization:.....11
I.4.4.Critical micelle concentration:.....12
I.5. Natural surfactants:.....13
I.6.Propreties:.....14
I.6.1.Physicochemical properties of surfactants:.....14
I.6.2. Technological properties:.....14
I.7. Emulsions:.....15
I.7.1. Definition:.....15
I.7.2. Types of emulsions:.....15
I.7.2.1.Water-in-oil emulsions (W/O):.....16
I.7.2.2. Oil-in-water emulsions (O/W):.....16
I.7.2.3.Multiples emulsions:.....17
I.7.3. Examples of Emulsions:.....17
I.7.4. Micro-emulsion:.....18
I.7.5. Emulsion formation :.....18
I.7.6. How Emulsifications works:.....19
I.7.7. Emulsifier Definition:https:.....19
I.7.8. Stability of emulsion :.....19
I.8.Micelles :.....20
I.8.1. Structure of the micelles formed by ionic surfactants consists of:.....20
I.8.2. Structure of Micelles formed by non-ionic surfactants consists of:.....20
I.8.3. MicellarSolubilisation:.....21

I.8.4. Micellisation: 21
I.9. Surfactants applications: 22
I.9.1.Somes exemples: 22
2-Detergency 23
I.10.Hydrophile-Lipophile Balance 26
I.11.Conclusion 27

CHAPTER II

Introduction: 29
II.1.Neural Network: 30
II.2.Definition of QSAR: 31
II.3.QSAR Work Flow: 33
II.4.Historical roots of QSAR: 33
II.5.Objectives of QSAR: 35
II.6.Importances of QSAR: 35
II.7.The development of QSPR models : 36
II.7.1.Generation of molecular descriptors from chemical structure: 36
II.7.1.1.Theoretical descriptors: 39
II.7.1.2. D2 descriptors are the most commonly used molecular descriptors: 41
II.7.2.Selection of most relevant molecular descriptors: 42
II.8.QSPR/QSAR METHODS : 45
II.9.Applications of QSAR: 47
II.10. CONCLUSION: 49

CHAPTER III

III.1.Introduction 51
III.1.1.QSAR calculation and filteringmethodology 52
III.2.Modellingsoftwares 53
III.3.Modelling techniques 53
III.3.1.Ordinary Least Squares model (OLS) 54
III.3.2.K-Nearest Neighbors regression model (KNN) 55
III.3.3.Consensus model 55
III.3.4.PLS model 55
III.3.5.Results and Discussion 64
III.3.5.1.Descriptors calculationand selection of relevant descriptors 64
III.3.5.2.OLS model development 65
III.3.5.3.KNNmodel development 66

III.3.5.4.Consensus 1Model 67
III.3.5.5.KNN2 model with 166 MACCS fingerprints 68
III.3.5.6.KNN3model with 1024 Hashed fingerprints 69
III.3.5.7.Consensus 2 model of KNN2+KNN3 70
III.3.5.8.PLS 71
III.4.Conclusion 74
General Conclusion……..……...……75
References 78

List of figures:

Chapter I

Figure I-1: Schematic representation of a typical surfactant molecule……...5

Figure I-2: A Non-ionic Surfactant nonylphenol ……....6

Figure I.3. Schematic of different types of surfactants…….....8

Figure I.4. Schematic Representation of Gemini Surfactant……...11

Figure I.5. Chemical-structure-of-the-gemini-surfactant-1-1-1-4-butan…….……..12

Figure I.6.classification of oil emulsion……..16

Figure I.7. W/O emulsion and emulsifiers on the surface of the dispersed phase (Water)..17

Figure I.8.O/W emulsion and structure of emulsifiers on the surface of the dispersed phase……17

Figure I.9. Droplet deformation during the emulsification process……..…....19

Figure I.10: structure of micelle…….…….21

Figure I.11.Structure of the micelles formed by ionic and non –ionic surfactants……...22

Figure I.12. Scheme various types of micelles…….23

Chapter II

Figure II-1:A processing element (PE) for a neural network……..……31

Figure II-2:A simple schematic overview of the formalism of QSPR……32

Figure II-3: Contribution of different kinds of documents recorded in Web of Science core collection on QSAR/QSPR subject (prepared on May 27, 2015)…….……32

Figure II-4: representation of a QSAR Work Flow…….33

Figure II-5: Growth pattern of publications on QSAR/QSPR subject (prepared on May27, 2015)…….……34

Figure II-6: A summary of the pioneering discoveries that led to the gradual evolution of the QSAR study……..……35

Figure II-7: Schematic illustration to show how chemical structures are employed to compute descriptors and QSPR model development……..…38

Figure II-8: Graphical example of different molecular representations of the same structure (ibuprofen)…….……40

Figure II-9: The broad application areas addressed by QSPR modeling studies…….....48

Figure II-10: Some representative endpoints addressed by QSPR analysis in the field of material science……..…….…….....48

Chapter III

Figure III.1: Basic molecular structure of (a) a Gemini surfactant and (b) a monomeric surfactant[1]…….54

Figure III.2: Example of the drawn structures of the investigated Gemini surfactants by HYPERCHEM software and optimized using molecular mechanics (MM+) at PM3 using the Polak-Ribiere algorithm……...…….55

Figure III.3: Scatter plot representing the descriptor (X_{AXIS} – VAR001 /Y_{AXIS} –$TDB09_V$) against the CMC values in the training set…….....65

Figure III. 4: Scatter plot representing the descriptor (X_{AXIS}-VAR001/Y_{AXIS}-RBF) against the CMC values in the training set…….…...65

Figure III.5: Scatter plot of descriptor representing the predicted values against CMC values(Xaxis-VAR001/Yaxis –Predicted) by the OLS model for the train/test set……..…..66

Figure III.6: Scatter plot of two descriptor representing the residuals values against the CMC (Xaxis –VAR001/Yaxis –Residuals) by OLS model for the train/test set…….....66

Figure III.7: Scatter plot of the descriptor representing the predicted values against the CMC values (Xaxis-VAR001/Yaxis –Predicted) by the KNN model for the train/test set……..…..67

Figure III.8: Scatter plot of the descriptor representing the residuals values against the CMC values (Xaxis-VAR001/Yaxis –Predicted) by the KNN model for the train/test set……..…..67

Figure III.9: Scatter plot of the descriptor representing the predicted values against the CMC values (Xaxis-VAR001/Yaxis –Predicted) by the Consensus1 model for the train/test set….68

Figure III.10: Scatter plot of the descriptor representing the Residual value against the CMC values(Xaxis-VAR001/Yaxis–Residuals)by the Consensus1 model for the train/test set……68

Figure III.11:Scatter plot of the descriptor representing the Predicted values against the CMC values (X axis-VAR001/Yaxis –Predicted) by the KNN2 model for the train/test set……69

Figure III 12. Scatter plot of the descriptor representing the Residuals values against the CMC values (X axis-VAR001/Yaxis–Reseduals) by the KNN2 model for the train/test set…….69

Figure III.13: Scatter plot of the descriptor representing the predicted values against the CMC values (X axis-VAR001/Yaxis –Predicted) by the KNN3 model for the train/test set……70

Figure III.14: Scatter plot of the descriptor representing the Residual value against the CMC values (Xaxis-VAR001/Yaxis-Residuals) by the KNN3 model for the train/test se…...…….70

Figure III.15: Scatter plot of the descriptor representing the predicted values against the CMC values (Xaxis-VAR001/Yaxis-Predicted) by the Consensus 2 model for the train/test set…..71
Figure III.16: Scatter plot of the descriptor representing the Residual value against the CMC values (X axis-VAR001/Yaxis-Residuals) by the Consensus 2 model for the train/test set...71
Figure III.17: Scatter plot of the descriptor representing the Predicted values against the CMC values by the PLS model for the train/test set……...……72
Figure III.18: Scatter plot of the descriptor representing the Residual value against the CMC values by the PLS model for the train/test set……...72
Figure III.19: Comparison of the performance of different models using different statistical parameters when predicting CMC values during the training stage (84 systems)……73
Figure III.20: Comparison of the performance of different models using different statistical parameters when predicting CMC values during the test stage (36 systems)……...74

List of tables:

Chapter I

Table I.1: Table representing various types of surfactants……8

Chapter II

Table II.1: represent different descriptors type……...…39

Table II.2: Representing important 2D descriptors……..…41

Chapter III

Table III.1: The structures and CMC values of 120 Gemini surfactants according to Figure 1……..……...58

Table III.2: Selected list of molecular descriptors calculated by AlvaDesc with their signification and type……64

INTRODUCTION:

Chemistry is a sub discipline of science that deals with the study of matter and the substances that constitute it. It focuses primarily on the atoms, ions, and molecules that in turn constitute elements and compounds. It branches into several types, including surface chemistry, which in turn deals with the study of phenomena that occur on the surfaces or interfaces of materials, such as adsorption, heterogeneous catalysis, colloidal formation, corrosion, crystallization, solubility, and electrode processes. Surface chemistry has been used by humans for a long time, in the Sumerian and Egyptian civilizations, through soap and washing. Then it manifested itself in various scientific fields, not to mention its applications in industry as well as in daily life.

Surfactants are a broad group of chemicals that play important and vital roles in a great variety of fields. Applications as diverse as cleaning, food, metallurgy, pharmacy, medicine, paints, mining, and many others utilize the characteristic properties provided by the surfactants .This mode of action is directly related to their chemical structure. All surfactants have in common an asymmetric skeleton with a hydrophobic and a hydrophilic moiety. The hydrophobic part generally consists of a linear or a branched alkyl chain, which is then linked to a hydrophilic group.

In addition to, there is Gemini surfactants sometimes called dimeric surfactants, are a kind of novel surfactants containing two hydrophobic tail-groups and two hydrophilic head-groups, which are covalently connected through a spacer at or near the two head-groups. In contrast to conventional surfactants that generally have a single hydrophilic head group and a single hydrophobic group in the molecule. Gemini surfactants can be ten to a thousand times more surface active than conventional surfactants with similar but single, hydrophilic and hydrophobic groups in the molecules.

These surfactants have attracted considerable attention from researchers on account of their attractive properties in contrast to their single-chain analogues, such as high aqueous solubility and viscosity, low critical micelle concentration (CMC), unique micelle structure

and aggregation behavior, high efficiency in decreasing the surface tension of water solutions and so forth.

However, critical Micelle Concentration (CMC) is the key point for surface chemistry as well as chemist and systematic data collection and day-to-day development in this regard is very essential. The CMC concentration is defined as the concentration of surfactants on which the micelles form and all the surfactants added to the system go to the micelles. So CMC is an important surface property. Moreover, determining the CMC of Gemini surfactants is of great importance and has attracted many interests; their measuring with traditional experimental methods is often a laborious and money-consuming task. For that, it is needed to predict the CMC of numerous Gemini surfactants by theoretical chemistry methodologies.

Consequently, studies on quantitative structure-property relationships (QSPRs) of surfactants and understanding of the effects of molecular structures on their functions and properties are becoming increasingly important. The QSPR is a successful strategy for prediction of surfactant properties based on modeling between calculated descriptors from molecular structures of the surfactants and chemical or physical properties of the solution. The calculated descriptors are such as topological indices, geometrical descriptors, some theoretical indices, single structural descriptors like constitutional descriptors, functional groups, aromaticity indices...etc. After calculating the descriptors, the most feasible ones were selected and a linear or nonlinear method can be used to develop a model between the selected descriptors and one or more particular properties.

In QSPR modeling different computational techniques have been used, such as multiple linear regression (MLR), PLS, ANN, SVM and Anfis. To the best to our knowledge, a few research papers have investigated the modeling techniques of Gemini surfactants based of relevant descriptors.

In this investigation, we discuss the role of QSPR model to identify the relation between the molecular structure and algorithm critical micelle concentration (CMC) of 120 Gemini surfactants.

For this reason we relied on the following chapters:

- Chapter One: Definition of surfactants in chemistry, industry and microbiology. Their types, classification, applications property and micelles.
- Chapter Two: We provide a theoretical overview of the different structural modeling methods used in our work. In addition to the softwares used in QSPR and their applications.
- Chapter three: result of modeling the CMC of 120 Gemini surfactant aided QSPR descriptors and fingerprint with the tested models (PLS - KNN –OLS and consensus)
- And finally a summarized conclusion which demonstrates that the OLS is the best for predicting the CMC of Gemini surfactants.

Chapter I:

SURFACTANTS

I.1.Introduction:

Surfactants form a unique class of chemical compounds. This chapter provides an introduction to the nature and physical properties of surfactants, emphasizing their ability to radically alter surface and interfacial properties and to self-associate and solubilize themselves in micelles. These properties provide the means to apply surfactants in wettability modification, detergency, and the displacement of liquid phases through porous media on one hand, and to stabilize dispersions (including foams, froths and emulsions), or to destabilize dispersions (again including foams and emulsions) on the other hand. These in turn lead to a vast array of practical application areas which are illustrated in terms of mineral and petroleum processing, biological systems, health and personal care products, foods, and crop protection.[1]

I.2.Definition:

Surfactant is shorthand for surface active agent. Surfactants are organic chemicals that, when added to two liquids, a gas and a liquid,or between a liquid and solid change their interfacial properties. They may also act as detergents, wetting agents, emulsifiers, foaming agents and dispersants.[2]

Surfactants are usually amphiphilic organic compounds, meaning they contain both hydrophobic groups (their "tails") and hydrophilic groups (their "heads"). Therefore, they are soluble in both organic solvents and water. Due to their unique amphiphilic structures, surfactants have been widely used in traditional industries.

The lipophilic part of the structure usually consists of hydrocarbon groups, in a simple case, it is one long alkyl chain, typically with more than ten carbon atoms. The hydrophilic part of the structure may be represented by non-ionic polar groups or ionic groups.[3]

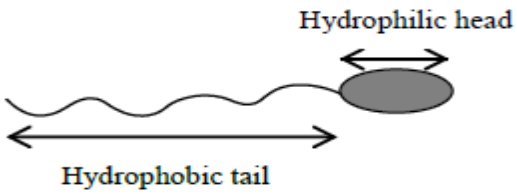


Figure I-1: Schematic representation of a typical surfactant molecule.**[4]**

I.3.Types of surfactants:

SAA`S compounds are classified into the following different classes:

1- Ionic surfactants: can be categorized into two main groups, such as cationic and anionic. Cationic surfactants after dissociation in environment give positive charged ions. They exert a strong toxic effect on environment as they are adsorbed strongly on the soil surface and lead to biodegradability. Anionic surfactants cause the major water pollution as they are widely used in the detergents and soap industries. They discharge negatively charged ions into the environment.

2- Non-ionic surfactants include those surfactants, which cannot undergo dissociation as they contain head and tail groups only and not containing the counter-ions, the example of non-ionic surfactant is nonylphenol shown in **Figure I-2.**[5]

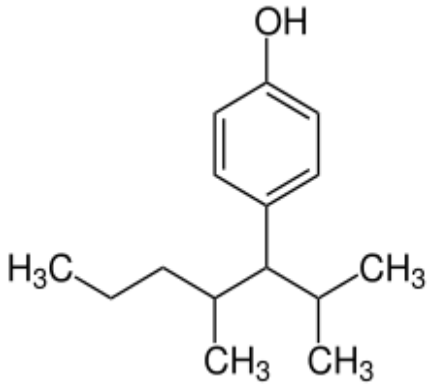


Figure I-2:A Non-ionic Surfactant nonylphenol**.**[6]

Surfactants are also classified depending on their solubility, such as hydrophilic surfactants that are soluble in water or hydrophobic (lipophilic) surfactants that are soluble in lipids. Ionic surfactants are generally hydrophilic surfactants, but nonionic surfactants can be either hydrophilic or lipophilic, depending on the balance of the hydrophilic group and lipophilic group. In other words, the solubility of nonionic surfactants depends on the balance between the hydrophilic group's capacity of attracting water and the lipophilic group's capacity of attracting oil**.**[7]

From the commercial point of view surfactants are often classified according to their use. However, this is not very useful because many surfactants have several uses, and confusions may arise from that. The most accepted and scientifically sound classification of surfactants is based on their dissociation in water or on electric charge in relative hydrophilic heads.[8]

I.3.1Anionic Surfactants

Anionic surfactants carry a negative charge on their hydrophilic head, thereby providing excellent detergency properties due to their uncanny ability to bind to positively charged particles, such as dirt and oils, lifting and suspending these particles in a micelle-like structure. Once in the micelle, the dirt and grime can be swept away and down a drain via an aqueous solution, leaving your home and work place spick and span.[9]

I.3.2.Nonionic Surfactants:

Nonionic surfactants are one of the more commonly used surfactants in industry. A nonionic surfactant is a type of surfactant that does not carry a charge on its hydrophilic head group and is therefore milder in nature. Due to the mildness associated with nonionic surfactants, they are commonly used throughout the home and personal care markets, as well as the agrochemical industry.[9]

I.3.3Cationic Surfactants:

Unlike anionic surfactants, cationic surfactants carry a positive charge on their hydrophilic head group, which makes them useful in antistatic products such as fabric softeners and hair conditioners. The positively charged polar head group is deposited onto the slightly negatively charged surfaces of hair or of fibers of textiles. Furthermore, the positive nature of cationic surfactants is known to aid in disrupting the bacteria cell membrane, making cationic surfactants useful as antimicrobial agents and are commonly found in hard-surface cleaners and disinfectants.[9]

I.3.4.Amphoteric Surfactants:

Amphoteric surfactants are zwitterionic, meaning they carry both a positive and a negative charge on their polar head-group, thereby making the overall net charge zero. Generally speaking, amphoteric surfactants have low toxicity, low eye and skin irritation, are tolerant to hard water, and exhibit excellent foaming and compatibility with other surfactants.Due to this unique characteristic of amphoteric surfactants, they behave differently depending on the pH of the final formulation – in an acidic environment, amphoteric adopt more cationic characteristics, while in an alkaline environment, they behave more like an anionic surfactant. Because an amphoteric to an amphoteric surfactant’s structure, this class of molecules are compatible will all other classes of surfactants.[9]

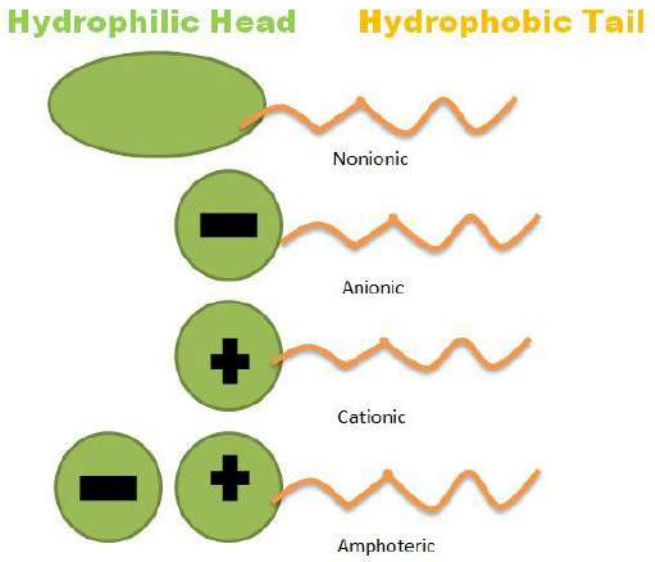
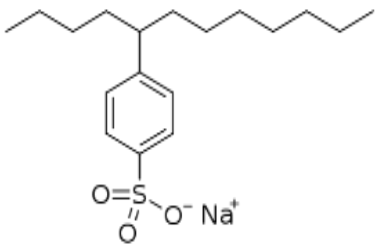
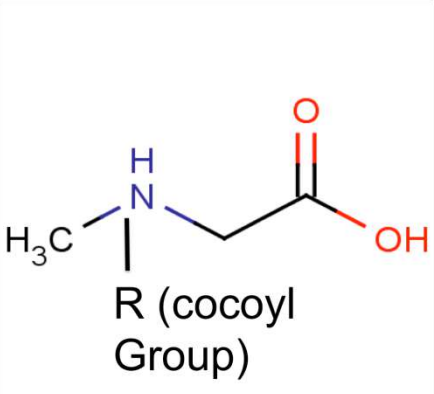
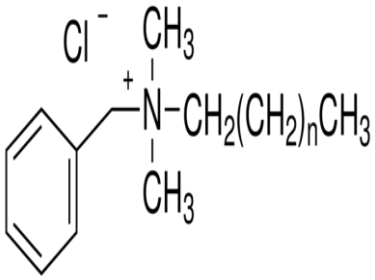
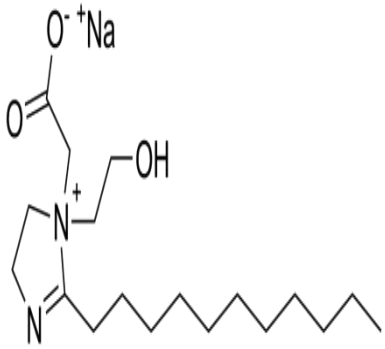


Figure I.3. Schematic of different types of surfactants.[10]

Table I.1: Table representing various types of surfactants.[11]

Types	examples	structure	properties	use
Anionic	-alkyl benzene sulfonate		- Detergents. -foaming agent. - Dispersants. -Higher incidence of irritation. -textiles, agriculture.	Account for about 50 % of the world production.
Nonionic	-Cocamide		-cleaning ability. -do not have good foam stability. -have better emulsifying qualities.	About 45% of the overall industrial productions of all Surfactants are Non-ionic.

Cationic	Benzalkonium chloride		-Have antimicrobial characteristic -More expensive. -detergents. -they are the ideal surfactant for products such as fabric and softeners.	Cationic emulsifiers are much more common in beauty products and Detergents. - Account for only 5-6% of the total surfactant production.
Amphoteric	-sodium lauroamphoacetat		-Milder and less irritating. -Expensive -Biodegradable -Resistant to hard water. -Emulsifier.	Are fairly new to the market and are still undergoing development pharmaceuticals and cosmetics.

I.4.Gemini surfactants:

Surfactants show interesting interfacial and bulk properties and have a wide variety of uses, which are mostly met by conventional representatives. Changes in the molecular structure and type to improve upon their properties have attracted the attention of chemists. This has led to the preparation of new generation surfactants such as Geminis.

In 1990s the search of newer surfactants and preparation started with *bis*-surfactants, which were later termed as "Gemini surfactants". Geminis is a class of surfactants which possess more than one hydrophobic tail and hydrophilic head group. Conventional surfactants have a single hydrophobic tail connected to an ionic or polar head group whereas Gemini surfactants has in sequence a long hydrocarbon chain, an ionic group, a spacer, a second ionic group and another hydrocarbon tail.

In Geminis the hydrocarbon tail can be short or long, the polar group can be anionic, cationic, nonionic and zwitterionic and the spacer group can be polar or non- polar, rigid or flexible. On the basis of polar head group Gemini are classified in cationic, anionic, non-ionic and zwitterionic Gemini surfactants, out of which anionic Gemini surfactants have an important place.[12]

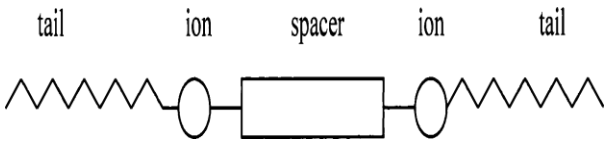


Figure I.4. Schematic Representation of Gemini Surfactant.[12]

I.4.1. Characteristics of Gemini surfactant:

Gemini surfactants are a new class of surfactants that comprise of two anionic head groups and two hydrophobic tails separated by a spacer group. They are characterized by:

- Gemini surfactants have low CMC values and show extraordinary surface activity.
- These surfactants have good water solubility and their ability to form micelles and lowering surface tension characteristics are fairly good as compared to conventional anionic surfactants.
- These surfactants have excellent lime soap dispersing ability (LSDA) too.
- Because of their higher surface activity these surfactants are used as promising surfactants in industrial detergency and emulsification.

- These surfactants are cost effective in area such as laundry detergent compositions, soil clean up, enhanced oil recovery and hydrotropic agents.[12]

I.4.2. Structure:

Some structural features of anionic Gemini surfactants are given below:

-All Gemini surfactants have at least two hydrophobic chains and two anionic groups(carboxylate, sulfonate, sulfate,phosphateetc.).

- Spacer can be polar(polyether)or nonpolar (aliphatic or aromatic), rigid or flexible
- Spacer length can be short or long.[12]

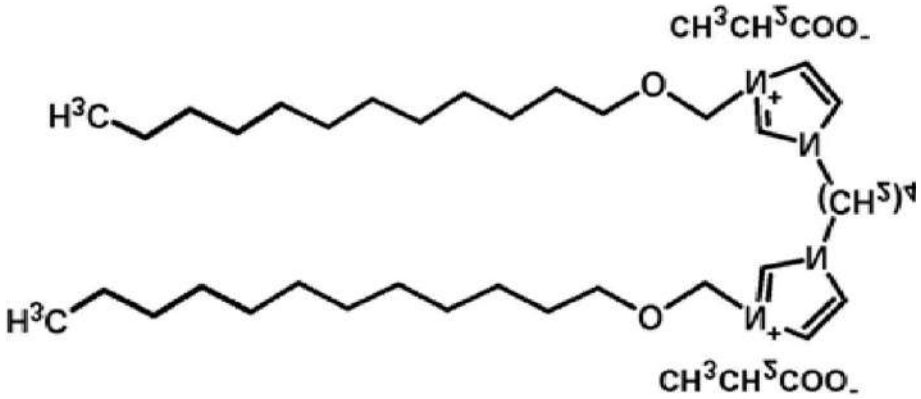


Figure I.5. Chemical-structure-of-the-gemini-surfactant-1-1-1-4-butan.[12]

I.4.3.Solubilization:

Solubilization is an important phenomenon required in tertiary oil recovery and detergency, as both are involved phenomenon of homogenizing.Geminis are better solubilizers due to very low CMC values. Many organic compounds that are normally insoluble in water, or only slightly soluble, dissolved to a greater extent in the presence of micellar surfactant. The increased solubility (to form thermodynamically stable and isotropic solutions) is called solubilization. A few aspects of solubilization are discussed here:

a. The location within the micelles of a solubilized compound depends upon its structure. Saturated hydrocarbons concentrate at the micellar core. A single functional group, however, shifts the additives solubilization site to the interfacial region made up of head groups, the first few methylene groups of the chain and water.

b. Even nonpolar aromatic moieties prefer to the interfacial region.

c. Increase in the surfactant chain length increases the solubilization power.

d. Branching of the surfactant chain length tends to diminish solubilization power.

e. For a given chain length solubilization by micelles of different charge types usually follows:

-The sequence is nonionic > cationic > anionic

-The water solubility of some dicarboxylate Gemini surfactant in hard water have been reported to be greater than that of comparable sodium carboxylates.

Krafft point temperature is the characteristic temperature at which the solubility of the surfactant becomes equal to CMC. For surfactant to be considered soluble their Krafft point temperature has to be below room temperature. [12]

I.4.4.Critical micelle concentration:

Critical Micelle Concentration (CMC) is the key point for surface chemistry as well as chemist and systematic data collection and day-to-day development in this regard is very essential.[13]

In chemistry, the critical micelle concentration (CMC) is defined as the concentration of surfactants above which micelles are spontaneously formed. Upon addition of surfactants into a system, they will initially partition into the interface, reducing the system free energy by lowering the energy of the interface and by removing the hydrophobic parts of the surfactant from contacts with water. When the surface coverage by the surfactants increases and the surface free energy (surface tension) has decreased, the surfactants start aggregating into micelles, thus again decreasing the system free energy by decreasing the contact area of hydrophobic parts of the surfactant with water.[3]

The most important parameter for measuring the surface activity of a surfactant is the determination of critical micelle concentration (CMC) values. The CMC is considered the ideal surfactant; the CMC values should be as low as possible. The CMC is the most important parameter of surfactant in solution; because it can be correlated with industrial characteristics of surfactant such as viscosity, foam stability, detergency, and dispersion ability. The CMC value depends on many factors including temperature, pH, pressure, additive presence, and molecular structure of surfactants. Usually, the CMC of surfactants is determined using experimental techniques, such as conductometry, fluorescence emission spectroscopy, cyclic voltammetry, NMR spectroscopy,tensiometry, etc…[14]

I.5. Natural surfactants:

Natural surfactants or bio surfactants are amphiphilic biological compounds, usually Extracellular, produced by variety of microorganisms from various substances including waste materials. There is an increasing interest on this topic because of their unique properties such as:

- low toxicity
- functionality under extreme conditions, based on renewable Substances and biologically degradable nature

The diversity of these molecules supports them for their potential application in the field of petroleum, medicine, agriculture, food, cosmetics etc. They are also effective in curtailing green-house effect by reducing the emission of CO_2.

They can be termed as 'green' because of their low toxicity, biodegradability and relative stability under a wide range of physicochemical environments.

The source may be of either plant or animal origin and the product should be obtained by some kind of separation procedure such as extraction, precipitation or distillation. No organic synthesis should be involved, not even as an after-treatment.

In spite of possessing diverse structures and better physicochemical properties than chemical surfactants, bio surfactants are not able to compete with their synthetic counterparts because of their high production and downstream costs.

The commercial realization of these ecofriendly biomolecules are restricted by low productivity, expensive downstream processing and lack of appropriate understanding of the bioreactor systems for their production. But we expect that in future better reactor design and product recovery technology would be developed and overproducer microbial strain would be screened. Then production cost would be decreased and yield would be increased (the production would be both ecologically and economically favored).[15]

Many examples for this type of surfactants such as:

- Cholesterol: Of animal origin, used in cosmetics, it is used in the composition of liposomes
- Lecithin: Of vegetable origin (glycerophospho lipid); used in cosmetics as an agent emulsifiers, it is part of the composition of liposomes
- Saponin: Plant-based, used in cosmetics as foaming agents instead synthetic surfactants
- Proteins: May be of different origins, used mainly as stabilizers of dispersion. [11]

I.6.Propreties:

I.6.1.Physicochemical properties of surfactants:

A surfactant is characterized by its tendency to adsorb at surfaces and interfaces. Examples of interfaces involving a liquid phase include suspension (solid-liquid), emulsion (liquid-liquid) and foam (liquid-vapor). In many formulated products several types of interfaces are present at the same time.

Another general and fundamental property of surface active agents is that monomers in solutions tend to form aggregates, called micelles.

Micelles form already at very low surfactant concentrations in water .The concentration at which micelles start to form is called critical micelle concentration (CMC). [16]

The physical properties include surface tension, osmotic pressure, electrical conductivity and ability to solubilize non-polar organics in the surfactant solution.[17]

I.6.2. Technological properties:

Surfactants are widely used and many topical pharmaceutical formulations, cosmetics, antiseptics, shampoos, detergents, creams and lotions contain surfactants. Surface active products are for to their amphiphilic properties used as emulsifiers, suspending, wetting, solubilizing and stabilizing exposure. The exposed groups and risk groups often worked in different service occupations, such as cleaning, hairdressing, home-help service, kitchen, restaurant work and hospital care but also in different types of industrial production. The most harmful exposure turned out to be to unspecied chemicals, water, detergents, dust and dry dirt. The risk of hand eczema is probably further increased if the exposure also involves substances that can cause allergic contact dermatitis.[16]

However surfactants may be also as :

- Wetting agents: they reduce the tension between a liquid and a solid; so they allow spread the product better on the skin.
- Detergents: they facilitate the elimination of impurities by unchecking them.
- Foaming: they promote the dispersion of a gas in a liquid; and thus allow form foam.
- Dispersants or dispersives: they allow the solubilization of the substance by forming micelles.
- Dispersant; or they limit the meeting of droplets by creating a repulsion between them.[11]

I.7. Emulsions:

Emulsions have been widely used in different industrial processes. The development and production of good quality emulsions depend on the knowledge of emulsion preparation, stability mechanisms and rheological studies.[18]

I.7.1. Definition:

An emulsion is a mixture of two immiscible liquid phases where one phase is dispersed into another. Basically, an emulsion is consisting of a continuous phase which is known as external phase, where the droplets are dispersed, and a dispersed phase which is defined as internal or discontinuous phase. Emulsions are found in different industries such as food, pharmaceutical, agricultural, cosmetics, and petroleum. An emulsion is not contained only water and oil; it may also contain some solid particles and even gas.

Inherently, an emulsion is an unstable system due to the unfavorable contact between oil and water phases .Although, some emulsions are stable as a result of the small drop sizes and the existence of interfacial film, which surrounds the drops.[18]

I.7.2. Types of emulsions:

Emulsions are classified depending upon the nature of the dispersed phase. Essentially, there are three types of emulsions which are water-in-oil (W/O), oil-in-water (O/W), and complex emulsions such as water-in-oil-in-water (W/O/W). The complex emulsions also known as a multiple emulsion. The three types of emulsion are displayed in Figure I.6.[19]

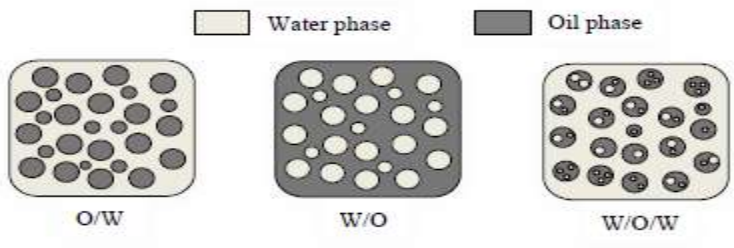


Figure I.6.classification of oil emulsion.[18]

In the case of simple emulsions, droplets of one liquid phase are dispersed in another immiscible liquid phase. Simple emulsions could be of two types. [20]

I.7.2.1.Water-in-oil emulsions (W/O):

A water-in-oil emulsion is the type of emulsions which the continuous phase is usually hydrophobic materials such as oil and the dispersed phase is water.

More than 95% of the crude oil emulsion formed in the oil field are the W/O type . The W/O emulsions contain three substances such as; a solvent, a surfactant, and water as demonstrated in **Figure I.7.**[18]

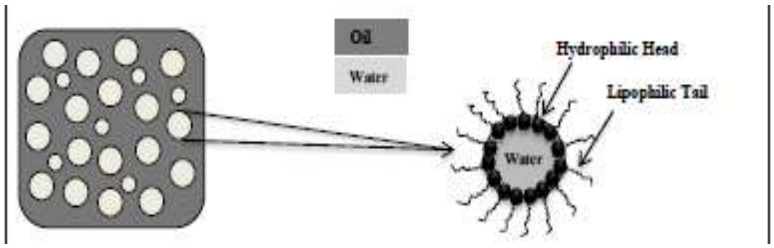


Figure I.7. W/O emulsion and emulsifiers on the surface of the dispersed phase (Water).[18]

In this type, the surfactant is usually soluble in water. A small concentration of soluble surfactant exists in water in the form of monomers.[21]

I.7.2.2. Oil-in-water emulsions (O/W):

An oil-in-water emulsion is the emulsion in which the oil exists as the dispersed phase and water as the dispersion medium or continuous phase as shown in Figure I.8.[18]



Figure I.8. O/W emulsion and structure of emulsifiers on the surface of the dispersed phase (Oil).[18]

However, the emulsions of W/O are more common than O/W, thus O/W emulsions are often pointed as reverse emulsions.Thus, in this class, the surfactant is preferentially soluble in the oil phase. The aqueous phase is present along with an oil phase rich in the surfactant (Winsor II). [21]

I.7.2.3.Multiples emulsions:

Multiple emulsions are double emulsions, that is, “emulsions of emulsions”, The globule of a multiple emulsion itself contains one or more smaller droplets inside. The size of a multiple emulsion globule is significantly larger than the size of the internal droplets especially when there are present a number of internal droplets.

Multiple emulsions are often classified into three groups depending on the number of internal droplets present in the multiple emulsion globules:

1. Type A multiple emulsion: where the multiple emulsion globule consists of only one large internal droplet, that is, the multiple emulsion globule is of “coreshell”.
2. Type B multiple emulsion: where the multiple emulsion globule consists of several small internal droplets.
3. Type C multiple emulsion: where the multiple emulsion globule consists of a large number of internal droplets.

Multiple emulsions (types A, B or C) can be further classified as either oil-in-water-in-oil (O/W/O) multiple emulsion or water-in-oil-in-water (W/O/W) multiple emulsion. In the case of an O/W/O multiple emulsion.[20]

ConsequentlyMultiple emulsions (or double emulsions) are very complex dispersion systems which are characterized by a low thermodynamic stability**.** [22]

I.7.3. Examples of Emulsions :

- Oil and water mixtures are emulsions when shaken together. The oil will form drops and disperse throughout the water.
- Egg yolk is an emulsion containing the emulsifying agent lecithin.
- Crema on espresso is an emulsion consisting of water and coffee oil.
- Butter is an emulsion of water in fat.
- Mayonnaise is oil in water emulsion that is stabilized by the lecithin in egg yolk.
- The photosensitive side of photographic film is coated with an emulsion of silver halide in gelatin.[23]

I.7.4. Micro-emulsion:

Micro-emulsion IUPAC defines micro-emulsion as dispersion made of water, oil, and surfactants that is an isotropic and thermodynamically stable system with dispersed domain diameter varying approximately from 1 to 100 nm, usually 10 to 50 nm.[24]

I.7.5. Emulsion formation:

The process of emulsion formation is called “emulsification”. Emulsification is a dynamic and nonspontaneous process, and energy is required to produce the droplets. The emulsification process usually requires a considerable amount of mechanical energy, to disperse one of the liquids in the form of small droplets in the continuous phase. There are several methods of emulsification including:

- Simple shaking, mixing with rotor-stator systems
- Liquid injection through porous membranes, or high pressure homogenizers and ultrasound generators. To break up a droplet into smaller sizes, the form of the droplet must be changed and this deformation occurs during the shearing or stirring of the emulsion. The process of droplet deformation and breakup is shown in**Figure I.9.**[18]

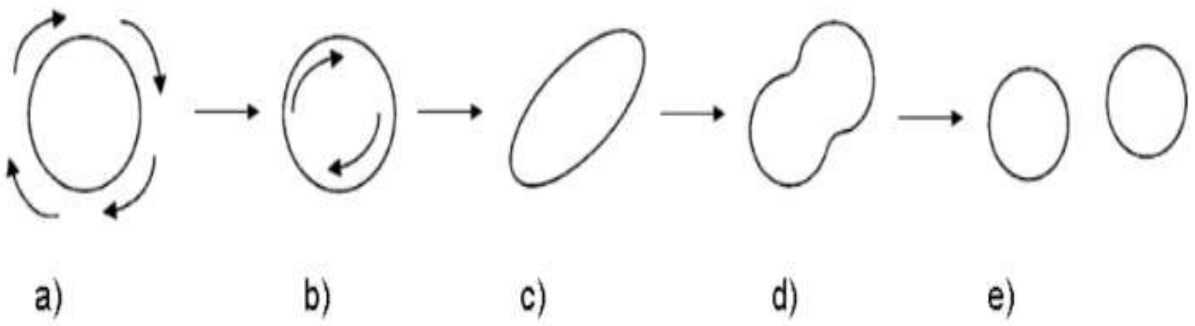


Figure I.9. Droplet deformation during the emulsification process.[18]

For the formation of an emulsion three conditions are necessary:

- Immiscibility between phases of the emulsion.
- Agitation to disperse one liquid in another.
- Existence of surfactants.[18]

I.7.6. How Emulsifications works:

There are a few mechanisms that may be involved in emulsification:

- Emulsification may occur when the interfacial surface tension between two liquids is reduced. This is how surfactants work.
- An emulsifier may form a film over one phase in a mixture to form globules that repel each other, allowing them to remain evenly dispersed or suspended.
- Certain emulgents increase the viscosity of the medium, making it easier for the globules to remain suspended. Examples include the hydrocolloids acacia and tragacanth, glycerine, and the polymer carboxymethyl cellulose.[23]

I.7.7. Emulsifier Definition :

A substance that stabilizes an emulsion is called an emulsifier or emulgent**.** Emulsifiers work by increasing the kinetic stability of a mixture. Surfactants or surface active agents are one type of emulsifiers. Detergents are an example of a surfactant. Other examples of emulsifiers include lecithin, mustard, soy lecithin, sodium phosphates, diacetyl tartaric acid ester of monoglyceride (DATEM), and sodium stearoyllactylate.[23]

I.7.8. Stability of emulsion :

The emulsion stability related to the type and amount of surfactants. These surfactants promote the stability of the emulsion by forming films around the water drops in water/oil interfaces. The formation of films increase the stability of the emulsion by increasing interfacial viscosity and decreasing interfacial tension. It has been investigated that the parameters such as water concentration, mixing speed, temperature can affect to the emulsion stability. By adding energy to the system the particles will break down into smaller sizes and finally the emulsion becomes more stable which will not be easy to treat.[18]

I.8.Micelles:

One important property of surfactants is the formation of colloidal-sized clusters in solutions, known as micelles. When surfactants are added to a liquid at low concentrations, they tend to adsorbs at surface or interface. As additional surfactant is added, the interface becomes fully occupied, and the excess molecules are forced into the bulk of the liquid.

At still higher concentrations, the molecules of surfactants in the bulk of liquid begin to form oriented colloidal size aggregates or micelles.

In a micelle, the hydrophobic tails flock to the interior in order to minimize their contact with water, and the hydrophilic heads remain on the outer surface in order to maximize their contact with water. Most micelles are spherical and contain between 60 and 100 surfactant molecules.[25]

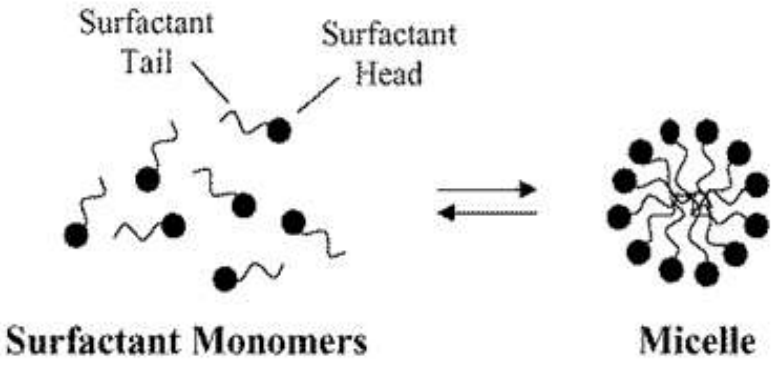


Figure I.10: structure of micelle.[25]

I.8.1. Structure of the micelles formed by ionic surfactants consists of:

- A hydrophobic core composed of the hydrocarbon chains of the surfactant molecule.
- A Stern layer surrounding the core, which is a concentric shell of hydrophilic head groups with counterions.
- A Gouy–Chapman electrical double layer surrounding the Stern layer, which is a diffuse layer containing the counterions required to neutralise the charge on the kinetic micelle.
- The thickness of the double layer is dependent on the ionic strength of the solution and is greatly compressed in the presence of electrolyte.[25]

I.8.2. Structure of Micelles formed by non-ionic surfactants consists of:

- For the non-ionic surfactants having a polyethylene oxide (PEO) head group, the structure is essentially the same, except that the counter ions are not present in the outer region, but rather coils of hydrated polyethylene oxide chains.
- Micelles formed in non-aqueous solution (reverse or inverted micelles) have a core composed of the hydrophilic groups surrounded by a shell of the hydrocarbon chains.[25]

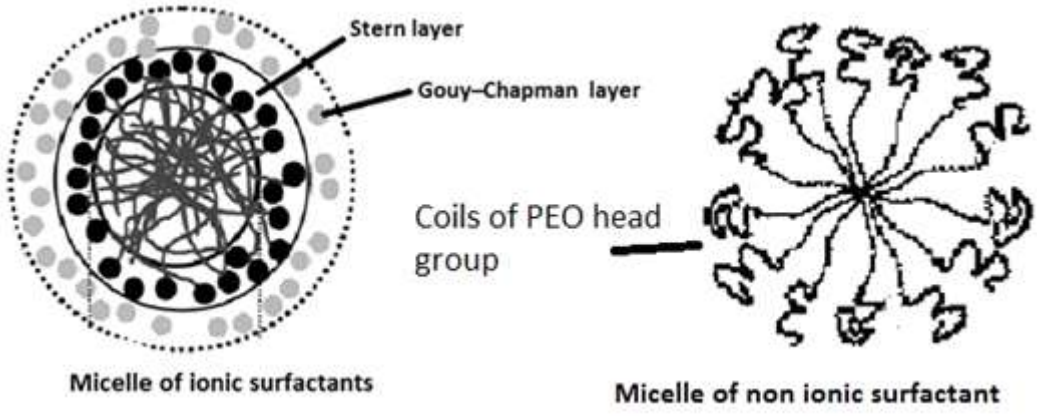


Figure I.11.Structure of the micelles formed by ionic and non –ionic surfactants.[25]

I.8.3.Micellar Solubilization:

Solubilization can be defined as the preparation of a thermodynamically stable isotropic solution of a substance normally insoluble or very slightly soluble in a given solvent by the introduction of an additional amphiphilic component or components.

If we plot the solubility of a poorly soluble compound as a function of the concentration of surfactant, usually what happens is that the solubility is very low until the surfactant concentration reaches the CMC. At surfactant concentrations above the CMC the solubility increases linearly with the concentration of surfactant, indicating that solubilization is related to micellization.

1. The location of a solubilized molecule in a micelle is determined primarily by the chemical structure of the solubilizate. Solubilization can occur at a number of different sites in a micelle Onthe surface, at the micelle–solvent interface.
2. Between the hydrophilic head groups.
3. In the palisades layer, i.e., between the hydrophilic groups and the first few carbon atoms of the hydrophobic groups that comprises the outer regions of the micelle core.[25]

I.8.4. Micellisation:

Micelles are formed at the critical micelle concentration (CMC), which is detected as an inflection point when physicochemical properties such as surface tension are plotted as a function of concentration. The main reason for micelle formation is the attainment of a minimum free energy state. The main driving force for the formation of micelles is the increase of entropy that occurs when the hydrophobic regions of the surfactant are removed

from water and the ordered structure of the water molecules around this region of the molecule is lost. Most micelles are spherical and contain between 60 and 100 surfactant molecules.[26]

The concentration at which micelles begin to appear in the bulk is called the critical micelle concentration (CMC) and can be determined from the discontinuity point in the plot of a certain physical property as a function of the surfactant concentration in the system.[17]

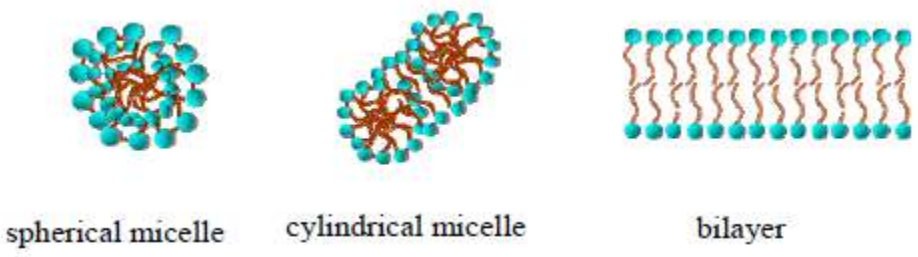


Figure I.12. Scheme various types of micelles.[17]

I.9. Surfactants applications:

The diversity of the surfactants is determined by a wide range of their applications. These compounds are used, for example , as ingredients, cosmetics, including personal hygiene products, such as liquid soaps, shower gels, shampoos or toothpaste .Thanks to their unique properties surfactant are also widely used in paper production ,food production ,paints and varnishes, as well as the metallurgical , agrochemical ,pharmaceutical, mining, textile, construction etc..[27]

It was found that without surfactant cleaning is not possible because it seems impossible to remove soils from cloths using water as solvent or we can say that cleaning formulations is not possible without surfactant. The aggregation of particles is known as micellization and aggregate known as Micelles.Micelles formation take place at very low concentration of surfactants called as CMC.[28]

I.9.1.Somes examples:

1- Wetting

The wetting can be defined as spreading of liquid droplets on solid surfaces, e.g. cotton material and the ability to penetrate into its pores. When the liquid without the addition of a surfactant is applied onto the surface of the material, it remains thereon in the form of a "drop". If you take the same amount of pure liquid and add a surfactant, you can notice the difference in the area occupied by this liquid. Thanks to the surfactant**s'** ability to lower the surface tension, it is possible to quickly soak the fabric in the water and wash it.[27]

2-Detergency

Detergency is the process of removing dirt. Non-polar portions, i.e. hydrocarbon chains of surfactants, combine with dirt particles to form spherical solution called "micelle". Micelles form a boundary layer that separates particles of dirt from the surface to be cleaned. This results in the formation of emulsion, which is insoluble in water, so that it is easily removed by washing with water, such as rinsing your hands under the stream of water after using soap.[27]

3-Household chemicals:

Surfactants are used as :

- Detergents, washing, cleaning and softening agents.
- Anti-microbial agents.
- Bleaching agents.[5]

4-Cosmetics industry:

They are used as:

- Emulsifiers.
- Foaming agents.
- Dispersing agents.
- Anti-microbes agents.[5]

5-Food industry:

Surfactants are involved in the production of many common food items and can be found in the extraction of cholesterol, solubilization of oils, liquor emulsification, prevention of component separation, and solubilization of essential nutrients.[1]

6-Textile industry:

- Wetting and washing agents.
- Dispersing agents.
- Additives for fabric dyeing.
- Supporting agents to produce synthetic textiles.[1]

7-Petroleum, mining and metallurgical industry :

A number of petroleum industry use surfactant for oil recovery which include micellar alkali surfactant, polymer and gas (hydrocarbon, N2, CO2 or steam flooding). The concentration of surfactant is more than CMC (critical micelle concentration) to remove oil from industry**.**

Also they work as:

- Components of drilling fluids.
- Corrosion inhibitors.
- Anti-icing additives.
- Demulsifying agents for the desalination of crude oil.
- Additives to petroleum products.
- Cleaning agents of metal surfaces.[28]

8-Surfactants And pharmacological industry :

Surfactants are widely used in pharmaceutical industries because of their antibacterial, antifungal and antiviral properties. These combating properties can be used in many diseases. In addition to this ,their role as anti adhesive agents against several disease causing pathogens makes their utility as suitable in various pharmaceutical, biomedical and therapeutic perspectives. Most of these surfactant are amphiphilic compounds which produced mostly by microbial cell surfaces and can be excreted extra cellular. The resultant contains both the hydrophobic as well as hydrophilic groups which accumulate the fluid phases together. Most of the bio surfactant are used in bioremediation of pollutants in low cost. [28]

9- Surfactants in microbiology:

The activity of a drug or drug action may be affected by surfactant. The penetration of hexyl resorcinol into the pinworm Ascaris is increased by the using a low concentration of surfactant. This activity is due to a reduction of interfacial tension between the liquid phase and the cell wall of the organism. It result, the adsorption and spreading of hexyl resorcinol over the surface of the organism is facilitated.

And excess concentration form micelles, the rate of penetration of the anthelmentic decreases to zero, because the drug is partitioned between the micelles and the aqueous phase. In biochemistry, the practical as well as theoretical importance of surfactants may be illustrated with the following examples: Surfactants have allowed the investigation of molecular properties of membrane proteins and lipoproteins, acting as solubilizing agents and as probes for hydrophobic binding sites.[28]

10-Surfactant in Nanotechnology:

Surfactant in Nanotechnology Having amphiphilic character allows surfactant molecules self- assembles into varieties of nanostructures from spherical rod like micelles to lamellar sheets. Using surfactant nanotechnology, it is possible to engineer a surfactant package capable of solubilizing large volumes of oil while aggressively water wetting the casing. By using surfactant mixtures, it is possible to form in-sitution oil-in-water micro emulsions, allowing surfaces to be water wet. This can allow the elimination of toxic solvents.[29]

11-Surfactants in the environment:

Surfactants are a diverse group of chemicals that are best known for their wide use in detergents and other cleaning products. After use, residual surfactants are discharged into sewage systems or directly into surface waters, and most of them end up dispersed in different environmental compartments such as soil, water or sediment. The toxic effects of surfactants on various aquatic organisms are well known.

In general, surfactants are present in the environment at levels below toxicity and in Croatia below the national limit. Most surfactants are readily biodegradable and their amount is greatly reduced with secondary treatment in wastewater treatment plants. The highest concern is the release of untreated wastewater or wastewater that has undergone primary treatment alone.

The discharge of wastewater polluted with massive quantities of surfactants could have serious effects on the ecosystem. Future studies of surfactant toxicities and biodegradation are necessary to withdraw highly toxic and non-biodegradable compounds from commercial use and replace them with more environmentally friendly ones.[30]

I.10.Hydrophile-Lipophile Balance:

The term hydrophile-lipophile balance (HLB), first suggested by Clayton, refers to the balance in size and strength between the hydrophilic and hydrophobic parts of a surfactant molecule. Griffin later developed the concept of the HLB for emulsifiers on the basis of their aqueous solubility. The HLB value is an empirical number assigned to nonionic surfactants on the basis of a wide variety of emulsion experiments carried out on surfactants at the Atlas Powder Company. These HLB values range from 1 to 40, the low numbers generally indicating solubility in oil and the high numbers solubility in water. Nevertheless, emulsifiers with the same HLB value may differ in solubility.

An emulsifier has two different actions: it promotes the formation of an emulsion, and it determines whether an oil/water (O/W) or a water/oil (W/O) emulsion will be formed. The second action is closely connected with the HLB value. On the basis of systematic emulsion experiments, Griffin found that the HLB values of mixtures of two or more emulsifiers are additive. The HLB value of the mixture is equal to the sum of the HLB values of the constituents multiplied by their weight fractions in the mixture x7:[31]

$$HLB = L\ x7(HLB)i$$

I.11.Conclusion

This chapter gives an introduction to the subject. Characteristic features of surfactant, classification and role of surfactant in surface science, characterization of surfactant and surfactant solutions, their performance properties, applications of surfactants and the scope of present work is discussed in detail.

Surfactants are surface-active compounds, which are used in industrial processes as well as in trade and household products. They have one of the highest production rates of all organic chemicals. Commercial mixtures of surfactants consist of several tens to hundreds of homologues, oligomers and isomers of anionic, non-ionic, cationic and amphoteric compounds. Therefore, their identification and quantification in the environment is complicated and cumbersome. Detection, identification and quantification of these compounds in aqueous solutions, even in the form of matrix-free standards, still poses the analyst considerable problems.[32]

Gemini surfactants have good commercial utilization potential; their preparation should thus be cost effective. As lower amount of surfactant is needed for a particular performance. Cost effectiveness may not be come in this way. Many more detergent and oil-based companies have yet to come forward to make newer cost effective Geminis. It will thus be challenge for new generation scientists to synthesize newer demand oriented geminis.[12]

Chapter II:

QUANTITATIVE STRUCTURE-ACTIVITY REATIONSHIP

Chapter II QUANTITATIVE STRUCTURE-ACTIVITY REATIONSHIP

Introduction:

In our fast moving society, the use of computer will become more and more extensive. The material things will be more and more supported by the virtual situations, which will facilitate our life.

However, science is the first actor in this direction, and thus of course, the use of computer methods affect all scientific sectors ,within this internal dialogue between computer science and other disciplines, looking for possible applications. The Nobel Prize for chemistry in 2013 recognized the formidable progress in chemistry related to the use of advanced, complex modeling approaches chemistry.[33]

The critical micelle concentration (CMC) is the single most useful quantity for characterizing surfactants. This parameter alone captures much of the surface activity of the molecule. The CMC, above which micelles are present in solution, is also key to the industrial use or biological activity of the surfactant. [34]

Chemistry plays an important role in defining the behavioral manifestations of chemical compounds. Development of suitable techniques, which allow modification of the chemical features of molecules, is very useful not only in the field of chemistry but also in other branches of natural sciences. Quantitative structure activity relationship (QSAR) modeling is one such technique that allows the interdisciplinary exploration of knowledge on compounds covering the aspects of chemistry, physics, biology, and toxicology. It provides a formalism for developing mathematical correlation between the chemical features and the behavioral manifestations of similar compounds. The entire technique is defined on the basis of a strong mathematical algorithm, and it provides a reasonable basis for establishing a predictive correlation model. Apart from providing a mathematical correlation, QSAR technique also enables the exploration of chemical features encoded within descriptors.[35]

This chapter presents a review on recent progress in quantitative structure-property relationship (QSPR) studies of surfactants and applications of descriptors. QSPR studies on critical micelle concentration (CMC)of surfactants areintroduced.By quantum chemical calculations and its effects on the structures and properties of the colloids of surfactants are also reviewed**.**[3]

II.1.Neural Network:

A neural network is a collection of processing elements, interconnected in a specified way that takes a group of inputs and transforms them into a group of outputs. Such a network of elements, when given certain characteristics, can be used as a computational means of answering problems that are less tractable with other computer methods. Presently, most neural network research involves software simulation of these devices and networks. Hardware implementations of neural networks that can be programmed and reconfigured dynamically are still mainly in the developmental stage.

Neural networks are referred to by-many different names, including connectionist models, neuro computers, artificial neural systems, parallel associative networks, and parallel distributed processors. While there are nuances to the use of each of these synonyms, broadly speaking they all refer to neural network technology. The term "artificial" is frequently used to distinguish neural networks from their biological counterparts. Artificial neural networks are inspired by biological networks, but for the purposes of applications, the emphasis is on computational power rather than close similarity to true biological systems. Instead of a highly interconnected network of neurons, an artificial neural network is made up of a number of highly interconnected processing elements. The network is a computing system that stores information by changing its state (the properties of the network) in response to external inputs**.**[36]

Artificial neural networks (ANNs) are a family of machine learning algorithms, inspired by the operations of neurons in the brain. Each neuron in an ANN receives numerous input signals, performs a weighted sum of the inputs, generates an activation response through a nonlinear activation function and passes the output signals to subsequent connected neurons . Multilayer ANNs can be constructed by organizing neurons into different layers and connecting neurons in consecutive layers. The combination of nonlinear units enables ANNs to learn highly complex functions of the inputs. ANNs have been widely applied to all branches of chemo informatics, including modeling QSAR/QSPR properties of small molecules as well as performing pharmacokinetic and pharmacodynamics analysis.

Deep learning networks are a recent extension of ANNs, which utilize deep and specialized architectures to learn useful features from raw data. The recent success of deep learning provides an opportunity to develop tools for automatically extracting task specific representations of chemical structures. Deep convolutional neural networks (CNNs) comprise a subclass of deep learning networks. In CNNs, local filters scan through the input space to

search for recurring local spatial patterns that are useful for the classification performance. Owing to unique local spatial properties of images, CNNs have achieved great success in the computer vision community, and have recently been applied to the biomedical field. For example, Torng and Altman viewed protein structures as '3D images' with four different atom type channels, and used 3D-CNNs to analyze amino acid microenvironment similarities and to predict effects of mutations in proteins**.** [37]

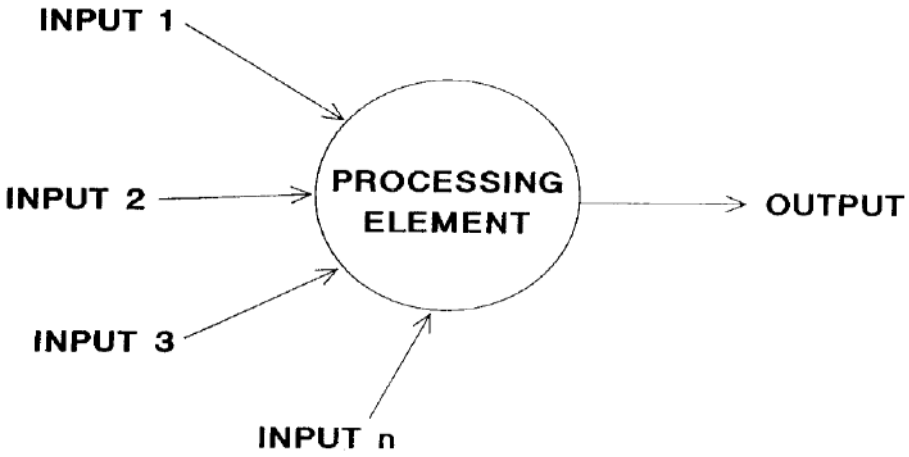


Figure II-1:A processing element (PE) for a neural network**.**[37]

II.2.Definition of QSAR:

Quantitative structure–activity relationship (QSAR) is a quantitative method that deals with finding a model to relate chemical structural features of compounds (descriptors) to their definite biological activity. Modeling the biological activity was extended to other physical/ chemical properties, which is called quantitative structure–property relationship (QSPR). Different properties or behaviors of chemical molecules have been investigated in the field of QSPR. Some examples are quantitative structure–reactivity relationships (QSRRs), quantitative structure–chromatography relationships (QSCRs) ,quantitative structure–toxicity relationships (QSTRs), quantitative structure– electrochemistry relationships (QSERs) , and quantitative structure–biodegradability relationships (QSBRs**).** [38]

In a QSAR/QSPR study, collecting or designing a subset of chemical or biological compounds, generation of potent descriptors capable to reflect the structure of compounds, selection of reverent descriptors to include in model, construction of a regression model and checking the validity and stability of the suggested model are five essential steps.[35]

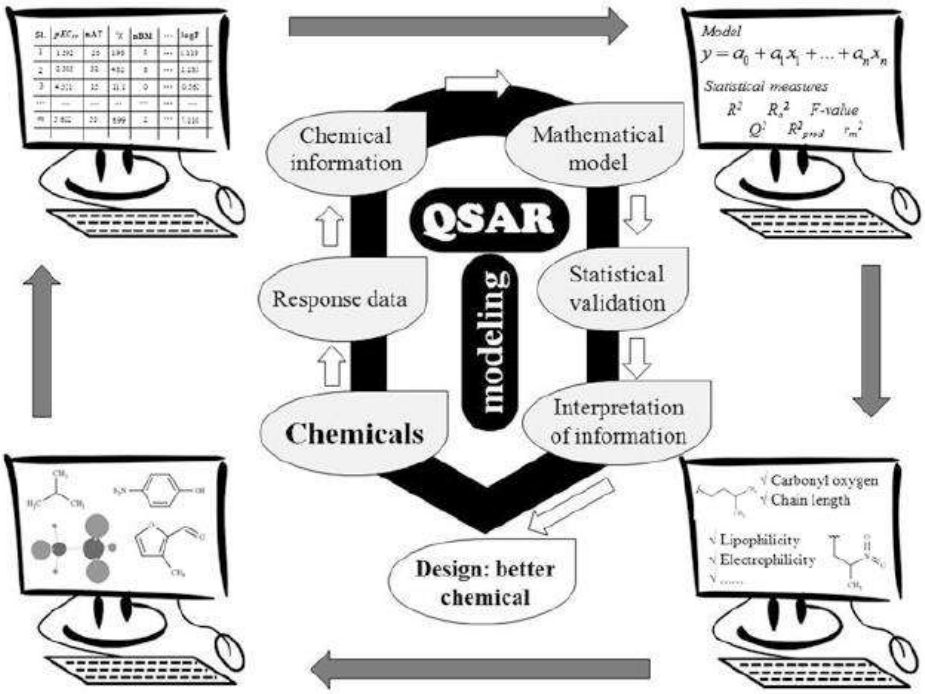


Figure II-2:A simple schematic overview of the formalism of QSAR.[35]

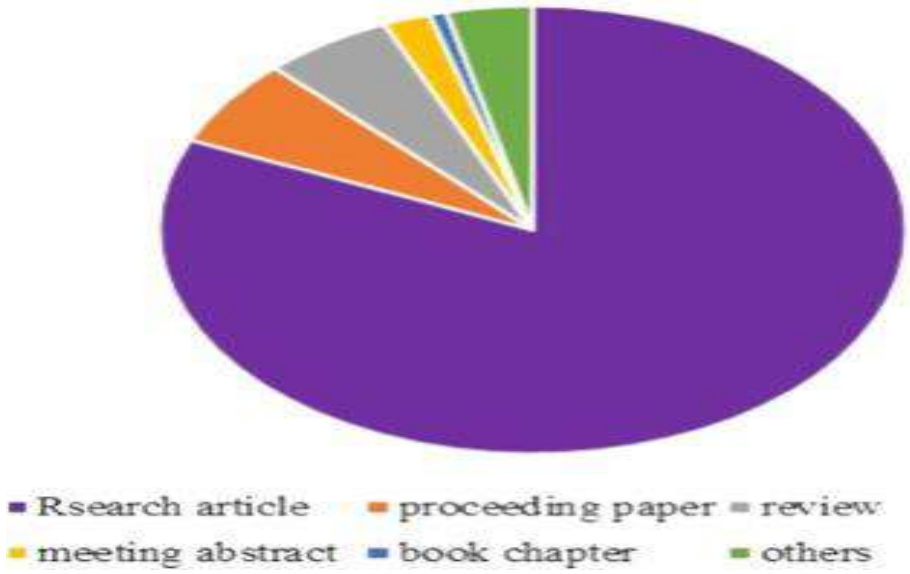


Figure II-3: Contribution of different kinds of documents recorded inWeb of Science core collection on QSAR/QSPR subject (prepared on May 27, 2015).[38]

II.3.QSAR Work Flow:

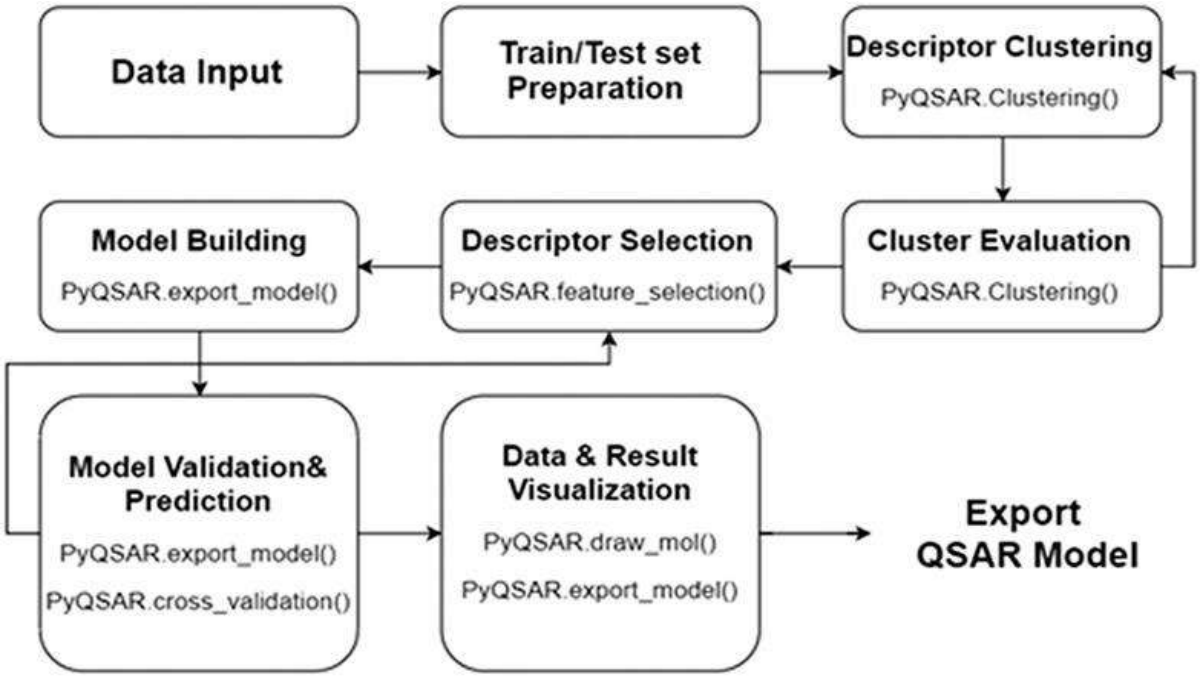


Figure II-4: representation of a QSAR Work Flow.[39]

II.4.Historical roots of QSAR:

By searching in literature, the root of huge QSAR tree could be found in the thesis of Cros, entitled "Action de l'alcoholamyliquesurl'organisme" (1863, Faculty of Medicine, University of Strasbourg, France) which noticed the relationship between toxicity of primary aliphatic alcohols and their water solubility. Without doubt, focusing on the concept of "molecular structure" in the years between 1860 and 1880 has led to the development of QSAR researches. Among the works on the molecular structure especially "three-dimensional" concept, the attempts of Butlerov (1861–65), Wislicenus (1869–73) and Van't Hoff (1874–75), are notable.

Maybe the first primary QSAR study could be related to Brown and Fraser , who proposed the existence of a correlation between molecular constitution and biological activity of different alkaloids. In 1884, Mills explained more on the hypothesis of structure–property correlation in his article entitled "On melting-point and boiling-point as related to chemical composition". After that, at the end of the 19th to beginning of the 20th centuries, other attempts were made to clarify this hypothesis (e.g.). Somebody believes that the works of Hammet on substituent effects in organic reactions, 1935–1938 , had an outstanding role in developing QSAR models. Some years later, when the first theoretical QSAR/QSPR studies

were proposed in the middle of the 20th century based on the descriptors obtained from graph theory, it was not expected that QSAR can become an inseparable part of molecular and drug design.

After the development of some other categories of structural descriptors, the world came ready to see the revolution of QSAR/QSPR. To reach this step, attempts of Pauling and also Coulson on the chemical bond concept, Sanderson on the atomic electronegativity, and the researches on electronic distribution and quantum-chemical descriptors were very important and determinant. After that, some specific QSAR/QSPR approaches were proposed step by step which is prominent from the historical point of view. After the suggestion of Linear Free Energy Relationships (LFER) by Hammet, one of the most principal approaches was linear solvation energy relationship (LSER) which had been proposed in 1952–3 by Taft Someone considers the start of the modern period of QSARin 1962.Other work of Hansch and coworkers attracted a lot of interests to this field and caused the QSAR explosion.[38]

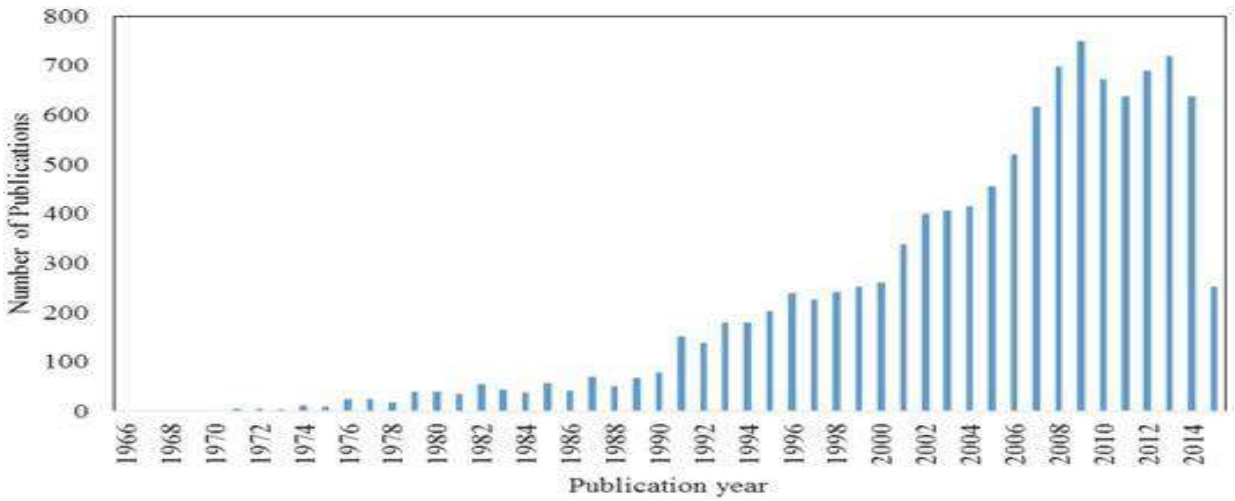


Figure II-5:Growth pattern of publications on QSAR/QSPR subject (prepared on May27, 2015).[38]

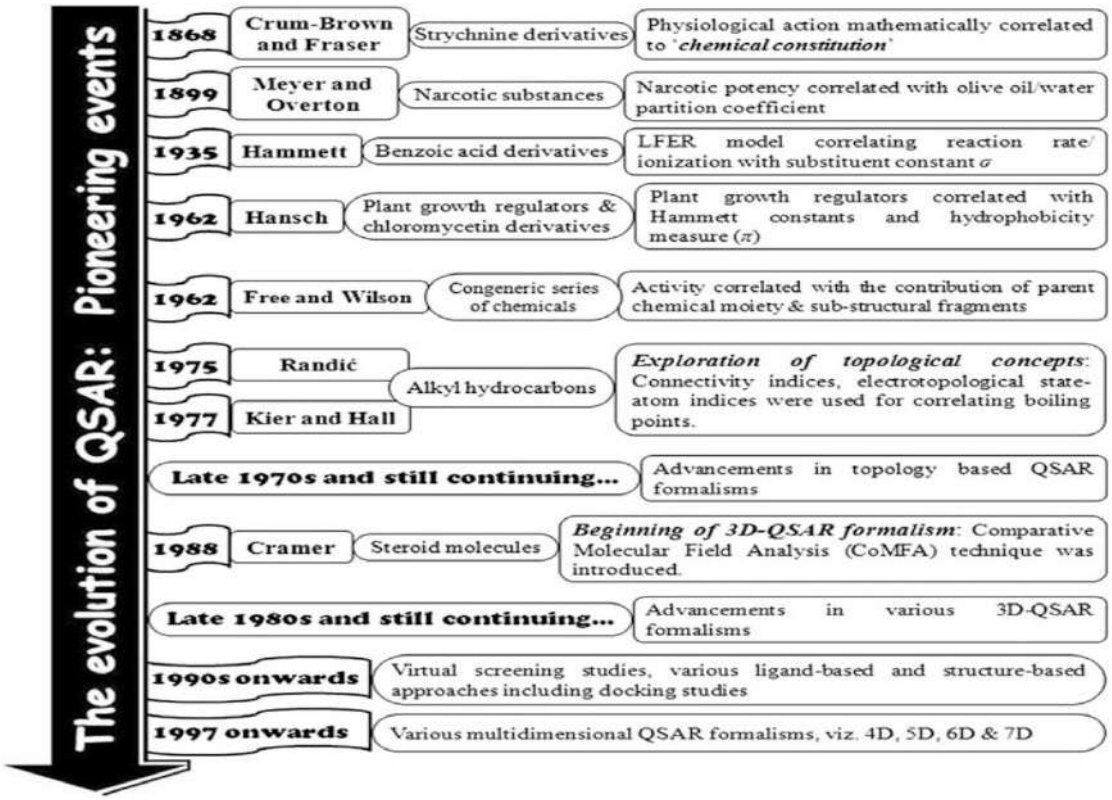


Figure II-6: A summary of the pioneering discoveries that led to the gradual evolution of the QSAR study.[35]

II.5.Objectives of QSAR:

The principal objective of any QSAR analysis lies in the rational development of a mathematical model accompanied with the exploration of the chemical information involved therein. Such modeling always uses comparatively less amount of data of chemical response and allows the prediction for a relatively large number of compounds. This provides an opportunity for this technique to be utilized in various fields**.**[35]

II.6.Importances of QSAR:

Although the development of predictive QSAR/QSPR/QSTR models appears to be a relatively simple task, it has got enormous applications in serving the need of scientific fraternity. It has always been a matter of curiosity that how it is possible for different chemical agents to exert different response profile, and sometimes it is rather astonishing that even the same chemical can elicit different biological actions. Hence, the chemical features appear to be very crucial in determining behavior of chemicals. QSAR techniques can provide

several advantages in terms of model predictivity and utilization of limited experimental resources, employing less computational time. Such features encourage the use of QSAR and related techniques in costly research programs such as drug-discovery and development where it can provide valuable information by aiding rational designing strategy with minimal cost involvement. Furthermore, since the QSAR technique can allow the prediction of a chemical response of relatively large number of compounds (within the chemical domain) by using response data of limited number of chemicals, it is widely employed in predictive toxicology analysis for the assessment of chemical hazards..[35]

II.7.The development of QSPR models:

The first step in each QSAR/QSPR analysis is collecting a set of compounds with relatively the same structural backbone. QSAR/QSPR models use quantitative activities or properties, which either may be obtained experimentally or from the literature.[40]

The development of QSPR models consists of four steps :

- Generation of molecular descriptors from chemical structure.
- Selection of most relevant molecular descriptors.
- Model validation.
- Model application.[41]

II.7.1.Generation of molecular descriptors from chemical structure:

Molecular descriptors are formal mathematical representations of a molecule, obtained by a well-specified algorithm, and applied to a defined molecular representation or a well-specified experimental procedure. The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.

Although, molecular descriptors are numerical indexes encoding some information related to the molecular structure. They can be both experimental physico-chemical properties of molecules and theoretical indexes calculated by mathematical formulas or computational algorithms.[42]

Chapter II QUANTITATIVE STRUCTURE-ACTIVITY REATIONSHIP

An ideal descriptor should possess the following features for the construction of a reliable QSAR model:

1. A descriptor should be relevant to a broad class of compounds.
2. A descriptor must be correlated with the studied biological responses while illustrating insignificant correlation with other descriptors.
3. Calculation of the descriptor should be fast and independent of experimental properties.
4. A descriptor should produce different values for structurally dissimilar molecules, even if the structural differences are little.
5. A descriptor should possess physical interpretability to determine the query features for the studied compounds.[35]

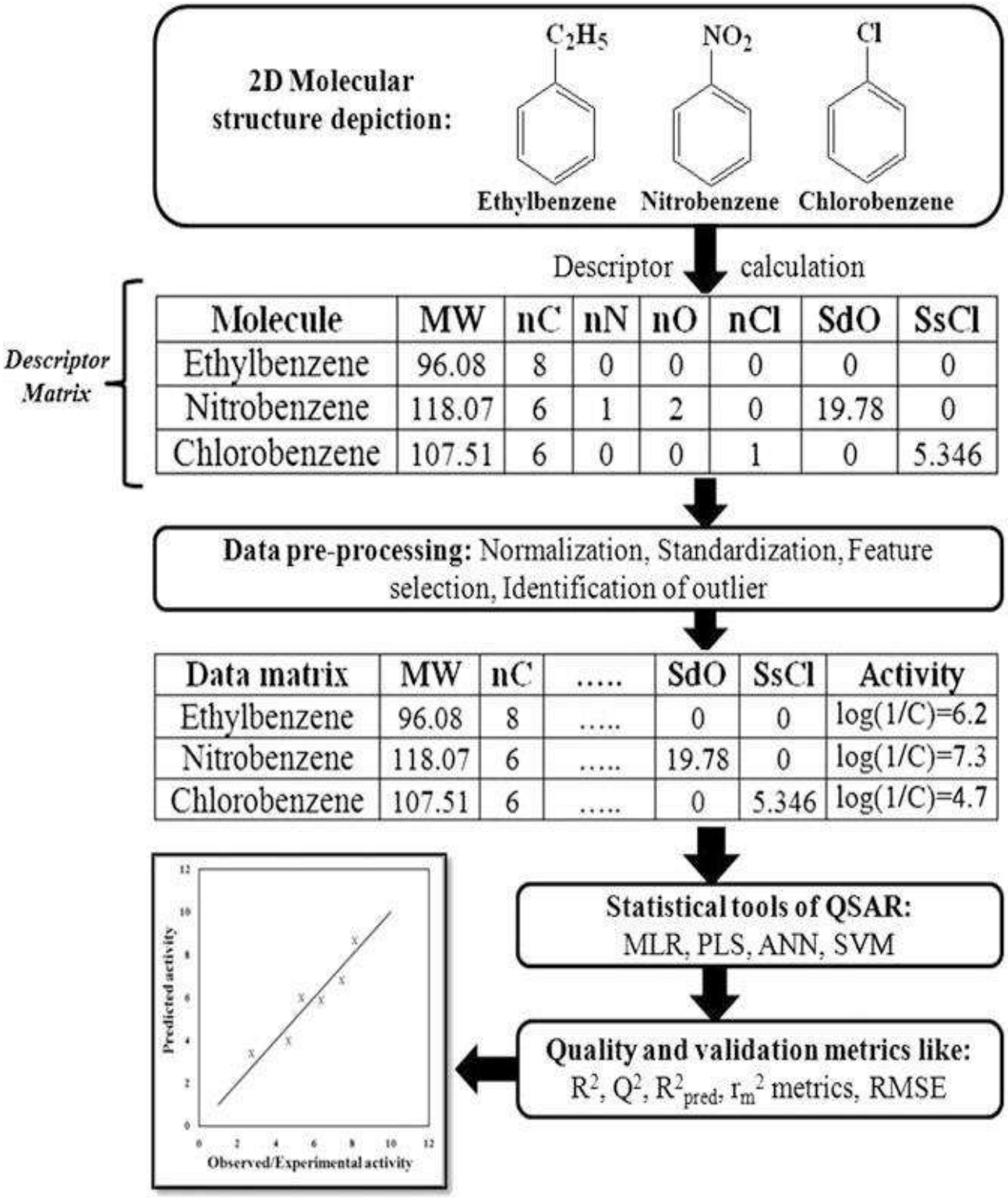


Figure II-7: Schematic illustration to show how chemical structures are employed to compute descriptors and QSAR model development.[35]

Table II.1: represent different descriptors type .[40]

Descriptor type	Description
1D	They consider properties inferred only the chemical formula of a chemical
2D	They consider properties inferred about the structure of the chemical based on the 2 dimensional structural formula
3D	They consider properties inferred from the spatial shape of the chemical for one conformation
4D	They are similar to 3D descriptors extended to multiple conformations

II.7.1.1.Theoretical descriptors:

The simplest molecule representation is the chemical formula that codifies the atom types and their occurrences within a molecule. Chemical formula does not carry any information about atom connections, and molecular descriptors obtained by this kind of representation are referred to as 0D molecular descriptors.

- **1D:** The substructure list consists of a partial list of structural fragments included in a molecule. Molecular descriptors derived from this representation are usually cited as 1D molecular descriptors. They cannot represent the whole topology of the chemical structure, can be easily calculated and interpreted, and are typically used for similarity/diversity analysis and virtual screening of large chemical databases.
- **2D:** The topological representation is the most popular one. This is a two-dimensional representation that, in addition to the atomic composition, includes the information about connectivity of atoms in the molecule in terms of a molecular graph. Molecular descriptors derived from two-dimensional representation of molecule are referred to as 2D molecular descriptors.
- **3D:**molecular descriptors derive the molecular representation where the spatial information about atom positions is added to atom connectivities; in other words, every atom of the molecule is associated with a set of three-dimensional coordinates

determining its own position in the three-dimensional space. This representation is referred to as geometrical representation of a molecule where the molecule is considered as a rigid geometrical structure.

- **4D:**representation takes also into account molecular properties deriving from electron distribution and the interactions of the molecules with probes.[43]

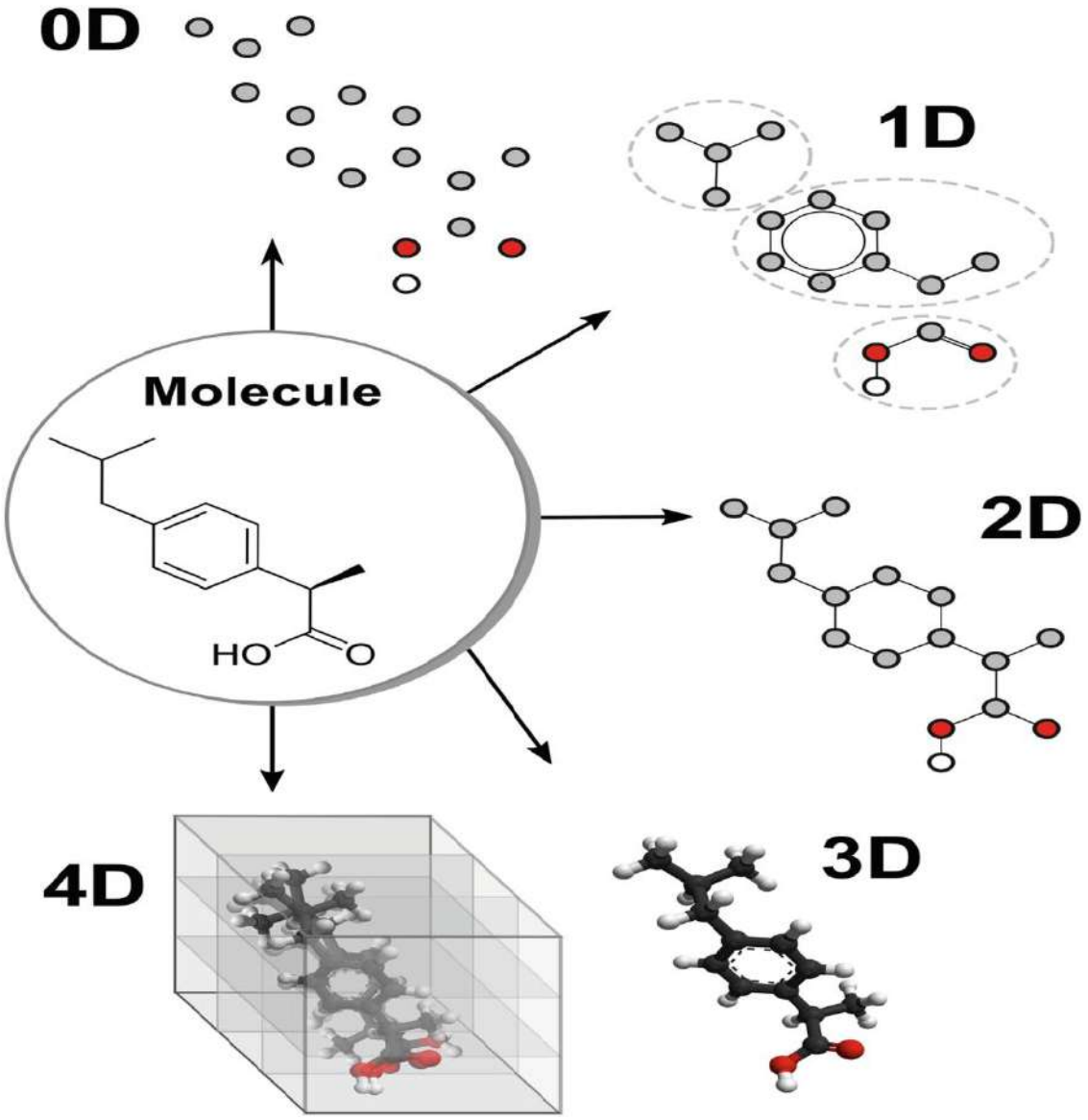


Figure II-8: Graphical example of different molecular representations of the same structure (ibuprofen) [44]

II.7.1.2. D2 descriptors are the most commonly used molecular descriptors:

Table II.2: Representing important 2D descriptors.[40]

2D Descriptor Types	Description	Example
Constitutional Descriptors	They represent properties related to molecular structure	molecular weight, total number of atoms in the molecule, number of aromatic rings
Electrostatic	They represent properties related to the electronic nature of the compound	atomic net and partial charges
Topological Descriptors	They represent properties which can be inferred by treating the structure of the compound as a graph, with atoms as vertices and covalent bonds as edges	total number of bonds in shortest paths between all pairs of non-hydrogen atoms
Geometrical Descriptors	They represent properties related to spatial arrangement of atoms constituting the compound	Vander Waals Area
Fragment based Descriptors	They represent properties related to sub structural motifs	MDL Keys and Molecular Fingerprints

1-Constitutional descriptors :

Constitutional descriptors depend only on the chemical composition of the molecule and describe very simple dependences such as the additivity of molecular properties from constant fragment contributions.[45]

2-Electronic Descriptors:

These descriptors reflect the electronic structure of the molecule and, basically, some overall characteristics of the partial charge distribution.[46]

3-Topological descriptors:

As the name suggests, topological descriptors consider the topology of a molecule. That is, in the most general case, only the connections between the atoms in a hydrogen-suppressed molecule, effectively converting it into a mathematical graph. Certain topological descriptors consider the type or certain properties of atoms involved in the connections as weights. Topological descriptors characterize features such as path lengths and connectivity.[47]

4-Geometric Descriptors :

Geometric descriptors characterize the shape and extent of the molecule in terms of its 3-D coordinates. As a result, accurate coordinates are required and so the structure must be geometry optimized before these descriptors can be calculated. [47]

II.7.2.Selection of most relevant molecular descriptors:

1) Improves Interpretation
   - Less features, simpler models.
   - Expert-driven feature selection enhances the mechanistic interpretation of the models.
2) Reduces Overfitting.
   - Less redundant data means lesser decisions based on noise.
3) Reduces Training Time
   - Less data to learn from ensures quicker model development.[40]

1-KNN: k-nearest Neighbor

K-nearest Neighbor is a non-parametric method used in classification and regression problems.

Principle: The property of an instance (chemical) is similar to instances close to them, where closeness is defined by the appropriate distance function using the feature space .

Highlights.

- Different distance functions available: Euclidean, Manhattan, Minkowski
- Simple to implement
- Easy to interpret (conceptually similar to read-across)[40]

K-Nearest Neighbor algorithm:

The KNN algorithm is a supervised non-parametric classification algorithm .Using this algorithm, pattern classification could be carried out by a comparison of this pattern with an example set. One by one, this pattern is compared with each of the example set. Out of these, the items selected finally are those whose distance to the pattern to be classified is smaller. Specifically, the k patterns with the shortest distance are chosen. Eventually, the output class of the pattern is assigned as the majority class of the k patterns closest to the former pattern. Since a set of example patterns is needed, as well as the class to which each pattern belongs, this is a supervised classification algorithm [48]. The KNN algorithm has great applicability and has been successfully used in a lot of different problems.

The k-nearest neighbors (KNN) estimate the response of a vector of features, according to:

Where y_i is the target value of the i-th neighbor and w_i is the weight:

Where D_{qi} is the distance between the query molecule and each i-th neighbor

Let W = {x_1, x_2,….., x_n} be a set of n labeled samples. The algorithm is as follows:

BEGIN

Input y, of unknown classification.

Set ,$1 \leq K \leq n$.

Initialize i= 1.

DO UNTIL (k-nearest neighbors found)

Compute distance from y to xi.

IF ($i \leq k$) THEN

Include xi ,in the set of k-nearest neighbors

ELSE IF (xi is closer to y than any previous nearest neighbor) THEN

Delete farthest in the set of k-nearest neighbors

Include x, in the set of k-nearest neighbors.

END IF

Increment i.

END DO UNTIL

Determine the majority class represented in the set of K-nearest neighbors.

IF (a tie exists) THEN

Compute sum of distances of neighbors in each class which tied.

IF (no tie occurs) THEN

Classify y in the class of minimum sum

ELSE

Classify y in the class of last minimum found.

END IF

ELSE

Classify y in the majority class.

END IF

END [49]

3-Partial least squares (PLS:

(PLS) regression is a statistical method that bears some relation to principal components regression. PLS finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the X and Y data are projected to new spaces, the PLS family of methods are known as bi-linear factor models. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multi-collinearity among X values – by contrast, standard regression will fail in these cases.[50]

Algorithms for PLS Regression

With the algorithms used for performing PLS Regression a distinction has to be made between the PLS_1 and PLS_2 regression algorithms. The cases must be differentiated where the explanatory variable Y is univariant and where it is multivariant because there are two PLS Regression methods (PLS_1 and PLS_2) which adapt to said circumstances. In the PLS_1 Regression (PLS univariate regression), there is only one variable to explain and there are p explanatory variables, whilst in the PLS_2 Regression (PLS multivariate regression) there are q variables to explain $q > 1$ and p explanatory variables. PLS_1 is the simplest algorithm. With this algorithm the first component is extracted, from which the rest of components are extracted, in such a way that they are uncorrelated (orthogonal). How this algorithm functions will now be described, although not comprehensively, just to show how the partial least squares method works (Valencia and Diaz-Llanos, 2003). The first component is defined thus:

$$t_i = w_{i1}X_1 + W_{i2}X_2 + \ \ldots + W_{ip}X_p = \sum_{J=1}^{P} W_{ij}X_j$$

whereX_j are the explanatory variables and y the variable to be explained The w_{ij} coefficients are:

$$w_{ij} = \frac{cov(x_j, y)}{\sqrt{\sum_{j=1}^{p} cov^2\,(x_j, y)}} = \frac{(x_j, y)}{\sqrt{\sum_{j=1}^{p}(x_j, y)^2}} \quad , j = 1, \ldots, p$$

From which it can be deduced that in order to obtain w1j the scalar products (x_j, y) must be calculated for each j = 1,…..,p. [51]

OLS algorithm:

Equations for the Ordinary Least Squares regression

Ordinary Least Squares regression (OLS) is more commonly named linear regression (simple or multiple depending on the number of explanatory variables).

In the case of a model with p explanatory variables, the OLS regression model writes:

$$Y = \beta 0 + \Sigma j=1..p\ \beta jXj + \varepsilon$$

where Y is the dependent variable, β0, is the intercept of the model, X j corresponds to the jth explanatory variable of the model (j= 1 to p), and e is the random error with expectation 0 and variance σ².[3]

In the training phase a set of values (called beta) is calculated as follow:

$$\hat{\beta} = (X^{\intercal}X)^{-1}X^{\intercal}Y$$

Where Y is the one-column matrix carrying the target values and X is the training dataset matrix.[51]

II.8.QSPR/QSAR Methods:

QSPR and QSAR consist of a body of statistical methods for relating physical and activity properties of a chemical product to its chemical structure. Important applications can be found today in many fields of chemistry, biology and engineering. The central problem in QSPR and QSAR is to convert the chemical structures into molecular descriptors that are relevant to a certain physicalproperty or bioactivity.

Various physico-chemical properties may be subjected toQSPR/QSAR studies, but the major application is drug discovery although their predictive modeling capabilities have extended also to other areas.

Obtaining a good quality QSAR model depends on many factors, such as the quality of input data, the choice of descriptors and statistical methods for modelling and for validation. Any QSAR modelling should ultimately lead to statistically robust and predictive models capable of making accurate and reliable predictions of the modelled response of new compounds. Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose. For QSAR models, validation must be mainly for robustness, prediction performances and applicability domain. Various strategies are adopted:[50]

1-Internal validation or cross-validation.

To validate the model, the dataset is divided into a training set (calibration set) used to develop the model and a test set (prediction set). In this way, the performance of different models can be compared because they are all trained with the same training set. The predictive performance of the model is measured by a cross-validation squared correlation coefficient q^2 as shown in the equation below – where y_i, $y_{i,p}$ and $y_{i,m}$ are the measured, predicted and mean values of y, respectively. Note that although a high q^2 value is a necessary condition for good predictions, it is not a sufficient condition.[50]

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_{i,p})^2}{\sum_{i=1}^{n}(y_i - y_{i,m})^2}$$

2-Data randomisation or Y-scrambling:

Is used for verifying the absence of chance correlation between the response and the modeling descriptors.Randomising the y variables (permutation test) is a common method for testing robustness of a model. A simpler variant of this method is testing the model on completely random data. By generating a random series of numbers for the figure of merit (y) and running the model, one should get only noise. If meaningful results are obtained (high R^2, q^2 values) using random data, then there is something wrong with the model.[50]

3-External validation:

Using only internal validation is not a reliable approach to evaluate the validity and predictability of a QSAR/QSPR model. To further check, several authors have suggested a more rigorous way to estimate the true predictive ability of a constructed model by comparing

the predicted and observed activity/property of an external test set of compounds that were not utilized in the model development process.

Therefore, it is necessary to split the data into two separate sets of "training" and "test". The metrics R_p, R^2 p, $RMSE_P$ and $R_m(test)^2$ are the common statistics to check the quality of external validation.[50]

II.9.Applications of QSAR:

Chemicals represent an indispensable part of human necessity considering varying applications spanning from laboratory to industrial processes as well as household usage. QSAR presents a suitable option in the rational monitoring of activity/ property/toxicity of chemicals and hence is useful in a wide variety of applications.

Since fine-tuning of the behavioral nature of chemicals gives fruitful results for a significantly large class of chemicals involving pharmaceuticals, agrochemicals, perfumeries, analytical reagents, solvents, surface modifying agents, etc., the application area and possibility of the QSAR technique is enormous.

In a global perspective, the chemicals modeled using the QSAR method can be overviewed in three major types:

1. Health benefits (drugs, pharmaceuticals, food ingredients, etc.)
2. Chemicals involved in industrial/laboratory processes (solvents, reagents, etc.)
3. And the chemicals posing hazardous outcome [persistent organic pollutants (POPs), toxins, xenobiotic, carcinogents, volatile organic compounds.

We have attempted to divide the employment of QSAR application in three broad areas, namely drug designing, material science, and predictive toxicology.[35]

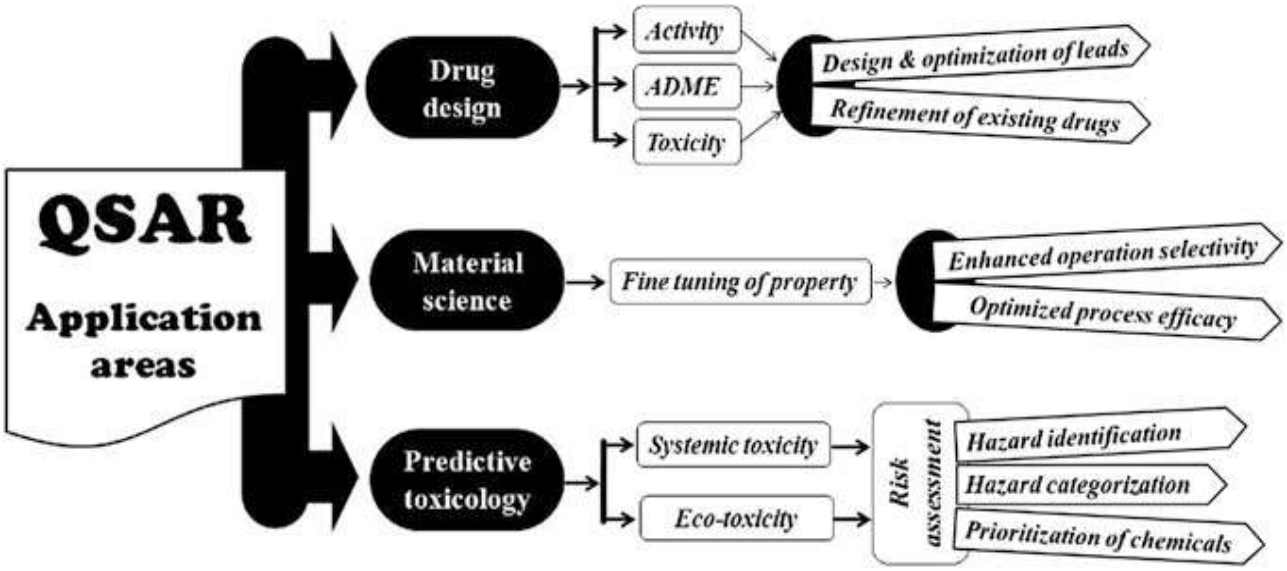


Figure II-10: The broad application areas addressed by QSAR modeling studies
Some potential areas of material science, which can be addressed by employing predictive QSAR modeling, have been depicted in **Figure II-11.**[35]



Figure II-11: Some representative endpoints addressed by QSAR analysis in the field of material science.[35]

II.10.CONCLUSION

The quantitative structure-activity relationship (QSAR)is an important research field in theoretical medicinal chemistry, which deals with the prediction of the biological activities of new compounds using mathematical relationships based on structural, physicochemical, and conformational properties of previously tested potential agents. QSAR relationships are helpful in understanding and explaining the mechanism of drug action at the molecular level and allow the design and development of new compounds presenting desirable biological properties. [52]

The aim of this chapter is to define the different methods used to model the physico-chemical substance suitable for active tension by using a theory-based approach QSAR. In order to find a mathematical model that describes the characteristics of active voltage according to the descriptions, we used Three main different linear regression modeling techniques, Least Squares model (OLS), K-Nearest Neighbors regression model (KNN) . Partial least square (PLS).

Chapter III:

DEVELOPMENT OF KNN-PLS-OLS AND CONSENSUS AIDED QSAR MODELS TO ESTIMATE CRITICAL MICELLE CONCENTRATION OF GEMINI SURFACTANTS.

III.1.Introduction

Surfactants are essential interfacial active compounds, which have been commonly used in chemical industry, environmental industry, detergent industry, petroleum industry and many other industrial fields. Gemini surfactants, also known as dimeric surfactants, are a kind of novel surfactants containing two hydrophobic tail-groups and two hydrophilic head-groups, which are covalently connected through a spacer at or near the two head-groups (Fig. 1). These surfactants have attracted considerable attention from researchers on account of their attractive properties in contrast to their single-chain analogues, such as high aqueous solubility and viscosity, low critical micelle concentration (CMC), unique micelle structure and aggregation behavior, high efficiency in decreasing the surface tension of water solutions and so forth[53], [54], [63], [55]–[62].

In colloid and surface chemistry, critical micelle concentration is defined as the concentration of surfactants when the micelles spontaneously begin to emerge. CMC is an important characteristic parameter of surfactants, because a number of properties of surfactant solution change sharply at this point, such as osmotic pressure, foaming, interfacial tension, surface tension, emulsification, detergency, conductivity, etc.. Thus, determining the CMC of Gemini surfactants is of great importance and has attracted many interests from surfactants researchers. Because measuring the CMC with traditional experimental methods is often a laborious and money-consuming task, it is needed to predict the CMC of numerous Gemini surfactants by theoretical chemistry methodologies[41], [64]–[71].

Quantitative structure-property/activity relationship (QSPR/QSAR) is an important type of chemometric technique which tries to relate the properties of chemicals to their molecular structures[72]–[80]. No matter the properties of molecules have been determined or not, they could be calculated by QSPR models, which could ultimately accelerate the development of new chemicals. QSPR technique has been extensively and successfully utilized to predict the CMC of many surfactants as an effective alternative of experimental methods [41], [53], [78], [63]–[70].

Hence, this work set up a new QSRR and diverse fingerprint as an accurate and convenient QSPR model to study the underlying relationship between the CMC value and structure of Gemini surfactants.

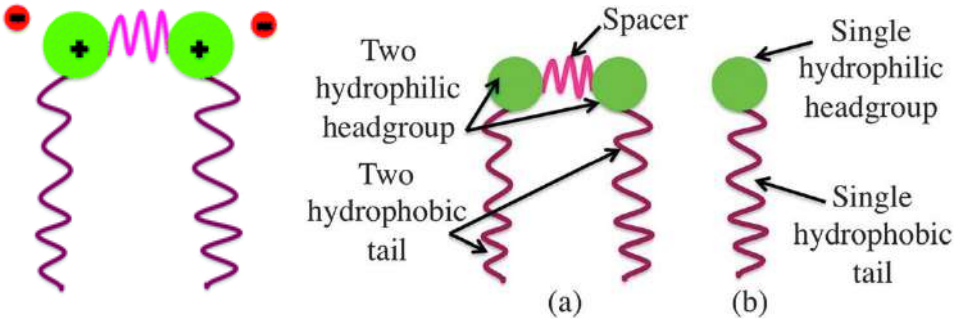


Figure III.1: Basic molecular structure of (a) a Gemini surfactant and (b) a monomeric surfactant[81]

III.1.1.QSAR calculation and filtering methodology

The methodology adopted for this work consisted of five main steps that are summarized below:

1. Dataset gathering (The chemical structures of 120 different Gemini surfactants molecules drawn and optimized by HyperChem software (Fig .2), and entered to AlvaDesc software to compute relevant QSAR descriptors and fingerprints). The database has been collected from previously published papers in literature[41], [63], [90]–[96], [82]–[89] and presented in table 1. The experimental CMC values of 120 diverse dibromides Gemini surfactants were obtained by conductmeter at 20−30 °C.

2. To avoid the phenomenon of over fitting, the number of variables (descriptors) must be reduced based on the following steps and using AlvaDesc software:

a. Constant and near constant descriptor values were removed

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

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

d. Descriptors with pair correlation larger or equal to ≥ 0.95 were removed.

3. Relevant descriptors were selected after performing stepwise MLR with F value method. The training dataset has *84* rows which represent the molecules used to train/test the model and *22* columns (also known as the models' *features*).

4. QSAR model was developed using four models.

5. Results comparison.

The 2D-structures of the surfactants were drawn usingHyperchem software (www.Hyper.com); geometries were pre- optimized using the molecular mechanics force field

(MM+) available in the Hyperchem package. The molecular geometries energy minima were then obtained using the semiempiricalAM1 using the Polack–Rabiere algorithm until the root mean square gradient limit as stopping criterion at 0.01 kcal/Å.

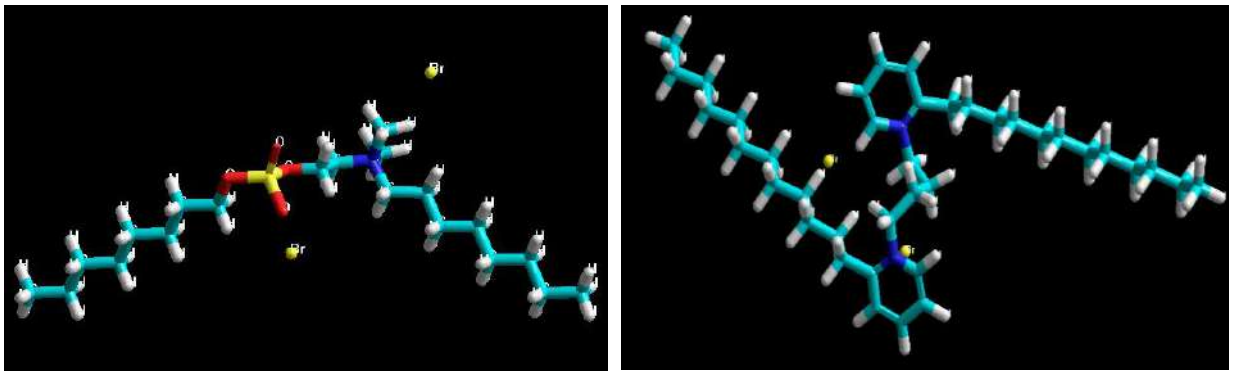


Figure III.2: Example of the drawn structures of the investigated Gemini surfactants by HyperChem software and optimized using molecular mechanics (MM+) at PM3 using the Polak-Ribiere algorithm.

III.2.Modelling software

AlvaDesc is a tool for the calculation of molecular descriptors and fingerprints. It calculates almost **53051, 2, 3-dimensional descriptors** splitted in 30 logical blocks. Together with the calculation of molecular descriptors, it carries out the calculation of three different molecular fingerprints:

- MACCS166 fingerprint**(166)**
- Hashedfingerprints**(1024)**
    - Extended ConnectivityFingerprint
    - PathFingerprint

AlvaModel is a software tool to create Quantitative Structure Activity/Property Relationship (QSAR/QSPR) models using descriptors and fingerprints previously calculated in AlvaDesc. These models can be used to predict the biological, physicochemical and environmental properties of chemicals.

III.3.Modelling techniques

Quantitative Structure Activity/Property Relationship (QSAR/QSPR) models are at the center of interest for different sectors (e.g., pharmacological, agricultural, electronic). QSAR/QSPR models are mathematical models that can be used to determine the behaviour of chemical compounds by predicting physicochemical, biological and environmental fate properties from the knowledge of their chemical structure.

A regression model is a mathematical tool for estimating the relationships between a target variable and one or more independent variables (features). This study deals with four different modeling techniques cited and discussed below:

- OLS1: an Ordinary Least Squares model (OLS) with 22 relevant descriptors
- KNN1: a K-Nearest Neighbors regression model (KNN)with 22 relevant descriptors
- Consensus 1: a Consensus model defined as the arithmetic mean of the values predicted by OLS1+KNN1 models
- KNN2: KNN with 166 MACCS fingerprints.
- KNN3: KNN with 1024 Hashedfingerprints.
- Consensus 2 : a Consensus model defined as the arithmetic mean of the values predicted by KNN2+KNN3 models.
- PLS: Partial least squares method with 22 relevant descriptors

III.3.1.Ordinary Least Squares model (OLS)

The Ordinary Least Squares (also referred to as multiple linear regression or just linear regression) is a way to modelling the relationship between a target value and one or more features. These features can be selected among the descriptors present in the project.The OLS model cannot make use of fingerprints as features.

In the training phase a set of values (called *beta*) is calculated as follow:

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$$

Where Y is the one-column matrix carrying the target values and *X* is the training dataset matrix containing the selected features and a constant value of 1 which represent the intercept. Both matrices refer to the training dataset.

The resulting matrix can then be used to predict the values:

$$\hat{Y} = W\hat{\beta}$$

Where *W* can either be *X* or a matrix with the same features of *X* but related to a new set of samples.

III.3.2.K-Nearest Neighbors regression model (KNN)

The K-Nearest Neighbors algorithm is a non-parametric method often used for classification, but it can also be used for regression. In the training phase the features and target values of the training dataset are stored inside the model. Whenused to predict, the model:

1. finds the *k* closest elements (aka neighbors) of the training dataset and
2. calculates the weighted average of the target values of these *k* neighbors (neighbors)

When defining this type of model you can select the number of neighbors (*k*) and the distance that must be used to identify the closest ones.Once the closest neighbors are identified, the weighted average is calculated as follow:

$$\hat{y_q} = \sum_{i=1}^{k} w_i y_i$$

Where y_i is the target value of the i-th neighbor and w_i is the weight:

$$w_i = \frac{\frac{1}{0.1+D_{qi}}}{\sum_{j=1}^{k} \frac{1}{0.1+D_{qj}}}$$

Where D_{qi} is the distance between the query molecule and each i-neighbor. When applying the model to the training dataset, the KNN algorithm finds *k* neighbors leaving out the element it is predicting.

III.3.3.Consensus model

The Consensus model contains a collection of models with the same target variable. It predicts by using the contained models and then averaging their predicted values.

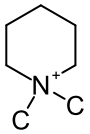
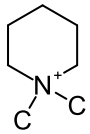
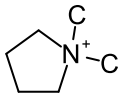
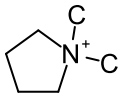
III.3.4.PLS model

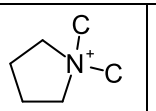
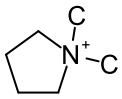
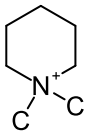
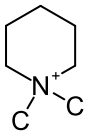
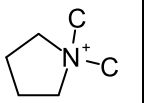
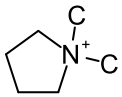
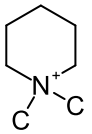
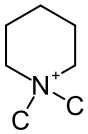
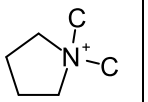
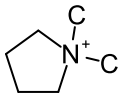
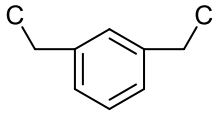
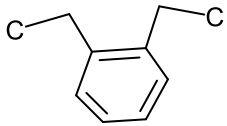
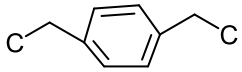
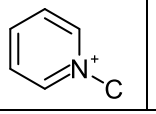
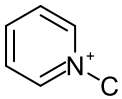
PLS method as an extension of multiple regression analysis (linear regression method) was used in this study for the model building.

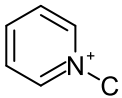
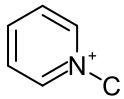
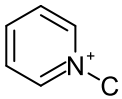
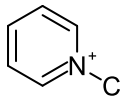
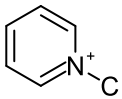
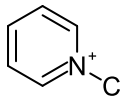
Table III.1: The structures and CMC values of 120 Gemini surfactants according to Figure 1

No.	Tail_1	Tail_2	Ion_1	Ion_2	Spacer	-logCMC		
						Exptl.	OLS	PLS
1	C_8H_{17}	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	3	3.391	3.49038
2	C_8H_{17}	$C_{10}H_{21}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	3.9586	4.129	3.94151
3	$C_{10}H_{21}$	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	4	4.018	4.02224
4	C_8H_{17}	$C_{12}H_{25}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	4.8539	4.664	4.166517
5	$C_{12}H_{25}$	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	4.7447	4.583	4.368887
6	$C_{10}H_{21}$	$C_{10}H_{21}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	4.8861	4.747	4.20792
7	C_8H_{17}	$C_{14}H_{29}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.2147	5.058	4.737763
8	$C_{14}H_{29}$	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.0969	4.958	4.57894
9	$C_{10}H_{21}$	$C_{12}H_{25}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	4.8861	4.665	4.59211
10	$C_{10}H_{21}$	$C_{14}H_{29}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.2757	5.129	5.25808
11	$C_{14}H_{29}$	$C_{10}H_{21}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.2596	5.288	4.79465
12	$C_{12}H_{25}$	$C_{12}H_{25}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.0458	5.362	4.78981
13	C_8H_{17}	$C_{18}H_{36}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.3188	5.424	5.279743
14	$C_{18}H_{36}$	C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.2291	5.148	5.40514
15	$C_{10}H_{21}$	$C_{16}H_{32}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.301	5.52	5.38421
16	$C_{14}H_{29}$	$C_{12}H_{25}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.1549	5.227	4.92519
17	$C_{12}H_{25}$	$C_{14}H_{29}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	5.3665	5.374	5.107463
18	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_2$	3.0969	3.071	3.06685
19	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	3.0269	3.321	3.14504
20	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	2.9318	2.883	2.83805
21	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	2.9872	2.812	2.694977
22	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_8$	3.0809	3.334	3.054113
23	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{10}$	3.2007	3.374	3.22409
24	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{12}$	3.4318	3.331	3.419767
25	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{14}$	3.699	3.662	3.63916
26	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{16}$	3.9208	4.104	4.128773
27	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	1.8601	1.734	1.490317

28	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	1.5817	1.381	0.88335
29	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	1.58	1.617	1.51417
30	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	1.5986	1.485	1.83891
31	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_2$	4.6778	4.602	4.406683
32	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	4.5686	4.381	4.086283
33	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	4.3665	4.235	4.404073
34	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	4.6021	4.573	4.26526
35	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	4.4437	4.559	4.626067
36	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_8$	4.4815	4.298	4.31047
37	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{10}$	4.5686	4.494	4.41108
38	$C_{16}H_{32}$	$C_{16}H_{32}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{12}$	4.699	4.729	4.76229
39	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	2.0565	2.17	2.24049
40	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	3.8697	3.739	3.68771
41	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_2$	2.2076	2.387	2.071753
42	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	2.1871	2.095	2.05336
43	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	2.0605	2.001	1.78503
44	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	2.0362	2.147	2.25515
45	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_8$	2.1249	2.053	2.24603
46	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{10}$	2.3768	2.351	2.704407
47	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_{12}$	2.6576	2.765	2.903457
48	C_6H_{13}	C_6H_{13}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	0.7959	1.201	3.15151
49	C_9H_{19}	C_9H_{19}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	1.6021	1.551	2.09548
50	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	2.0506	2.104	2.35528
51	$C_{11}H_{23}$	$C_{11}H_{23}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	2.5229	2.392	2.61474
52	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	2.9586	2.885	3.21729
53	$C_{13}H_{27}$	$C_{13}H_{27}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	3.3872	3.536	3.44529
54	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_5$	3.7447	3.677	3.753763
55	$C_{13}H_{27}$	$C_{11}H_{23}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	3.0555	2.911	2.890263
56	$C_{14}H_{29}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	3.0757	3.193	3.246
57	$C_{16}H_{32}$	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	3.1249	3.051	3.004357
58	$C_{18}H_{37}$	C_6H_{13}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	3.2291	3.352	3.32193
59	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	1.5086	1.301	1.69468

60	C_9H_{19}	C_9H_{19}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	1.7447	1.893	2.1064
61	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	2.0757	2.323	2.53885
62	$C_{11}H_{23}$	$C_{11}H_{23}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	2.5229	2.972	2.813887
63	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	2.9586	3.239	3.32548
64	$C_{13}H_{27}$	$C_{13}H_{27}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	3.4815	3.473	3.5884
65	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	3.9208	3.995	3.853107
66	$C_{15}H_{31}$	$C_{15}H_{31}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	4.3665	4.23	4.041673
67	$C_{16}H_{33}$	$C_{16}H_{33}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2N(CH_3)$	4.9208	4.923	4.746
68	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	1.3565	1.337	1.38468
69	C_9H_{19}	C_9H_{19}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	1.6576	1.819	1.86129
70	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	2.1135	2.037	1.98021
71	$C_{11}H_{23}$	$C_{11}H_{23}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	2.585	2.356	2.74297
72	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	3	2.978	3.210367
73	$C_{13}H_{27}$	$C_{13}H_{27}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	3.4202	3.277	3.41162
74	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	3.7212	3.365	3.669773
75	$C_{16}H_{33}$	$C_{16}H_{33}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2O$	3.1938	3.817	4.115813
76	C_6H_{13}	C_6H_{13}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	0.8861	0.902	0.873973
77	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	1.4318	1.512	2.01008
78	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	2.2076	2.445	2.6857
79	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	3.0177	3.263	3.5478
80	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	4	3.935	4.17439
81	$C_{16}H_{33}$	$C_{16}H_{33}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	5.0132	4.784	4.704243
82	$C_{18}H_{37}$	$C_{18}H_{37}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4)_2S$	6	5.707	5.660977
83	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$CH_2C{\equiv}CCH_2$	2.857	2.683	2.195753
84	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$CH_2C{\equiv}CCH_2$	1.5935	1.731	1.476227
85	$C_{12}H_{25}$	$C_{12}H_{25}$			$CH_2C{\equiv}CCH_2$	3.0506	3.183	3.146907
86	$C_{12}H_{25}$	$C_{12}H_{25}$			$CH_2C{\equiv}CCH_2$	3.041	2.85	2.85001

87	C_8H_{17}	C_8H_{17}			$CH_2C{\equiv}CCH_2$	1.4634	1.476	1.515667
88	C_8H_{17}	C_8H_{17}			$CH_2C{\equiv}CCH_2$	1.3206	1.335	1.58367
89	$C_{10}H_{21}$	$C_{10}H_{21}$			$CH_2C{\equiv}CCH_2$	2.2757	2.357	2.2752
90	$C_{10}H_{21}$	$C_{10}H_{21}$			$CH_2C{\equiv}CCH_2$	2.2248	2.127	2.58049
91	$C_{14}H_{29}$	$C_{14}H_{29}$			$CH_2C{\equiv}CCH_2$	3.8861	3.948	3.552023
92	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$		1.6459	1.969	2.378227
93	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$		1.6289	1.795	1.664357
94	C_8H_{17}	C_8H_{17}	$(CH_3)_2N^+$	$(CH_3)_2N^+$		1.6108	1.402	1.74164
95	$C_{10}H_{21}$	$C_{10}H_{21}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2CH(OH))_2$	2.4318	2.518	2.76719
96	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2CH(OH))_2$	3.1549	3.124	2.979077
97	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2CH(OH))_2$	4.0706	3.98	3.76745
98	$C_{16}H_{33}$	$C_{16}H_{33}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2CH(OH))_2$	4.301	4.246	4.476
99	$C_{10}H_{21}$	$C_{10}H_{21}$			$(CH_2)_4$	2.5702	2.6	2.67814

100	$C_{12}H_{25}$	$C_{12}H_{25}$			$(CH_2)_4$	3.2518	3.194	3.704773
101	$C_{10}H_{21}$	$C_{10}H_{21}$			$(CH_2)_6$	2.699	2.632	3.002397
102	$C_{12}H_{25}$	$C_{12}H_{25}$			$(CH_2)_6$	3.2676	3.428	3.514247
103	B-C_8H_{17}	B-C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	3.7212	3.545	3.0617
104	B-C_8H_{17}	B-C_8H_{17}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	3.1549	3.052	3.003983
105	B-C_8H_{17}	C_9H_{19}	PO_4^-	$(CH_3)_2N^+$	C_2H_4	3.699	3.576	3.054547
106	B-C_8H_{17}	$C_{11}H_{23}$	PO_4^-	$(CH_3)_2N^+$	C_2H_4	3.8239	3.911	3.821183
107	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4O)_2C_2H_4$	3.0362	3.098	3.186617
108	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4O)_3C_2H_4$	3.0655	2.765	2.607
109	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4O)_4C_2H_4$	3.0458	2.624	2.12586
110	$C_{12}H_{25}$	$C_{12}H_{25}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(C_2H_4O)_5C_2H_4$	2.9101	2.805	2.23469
111	E-$C_{12}O$	E-$C_{12}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	3.2596	3.485	3.735947
112	E-$C_{12}O$	E-$C_{12}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	3.2218	3.034	3.192703
113	E-$C_{12}O$	E-$C_{12}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	3.2403	3.33	3.85064
114	E-$C_{14}O$	E-$C_{14}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	3.58	3.59	3.981573
115	E-$C_{14}O$	E-$C_{14}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	3.3799	3.333	3.496483
116	E-$C_{14}O$	E-$C_{14}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	3.4802	3.772	4.102373
117	E-$C_{16}O$	E-$C_{16}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_3$	4.5406	4.422	4.704307

118	E-$C_{16}O$	E-$C_{16}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_4$	4.3507	4.593	4.30039
119	E-$C_{16}O$	E-$C_{16}O$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	4.4101	4.536	4.67236
120	$C_{14}H_{29}$	$C_{14}H_{29}$	$(CH_3)_2N^+$	$(CH_3)_2N^+$	$(CH_2)_6$	3.8013	3.739	3.79092

Table III.2: Selected list of molecular descriptors calculated by AlvaDesc with their signification and type

Descriptors code	Descriptors signification	Descriptors Type
VE2_B(i)	Average coefficient of the last eigenvector (absolute values) from Burden matrix weighted by ionization potential	2D matrix-baseddescriptors
RBF	Rotatable bond fraction	Constitutionnel indices
nROR	Number of ethers (aliphatic)	Randicmolecular profiles
CATS2D_02_AL	CATS2D Acceptor-Lipophilic at lag 02	Pharmacophoredescriptors
GATS4m	Geary autocorrelation of lag 4 weighted by mass	2D autocorrelations
Eig05_AEA(dm)	Eigenvalue n. 5 from augmented edge adjacency mat. weighted by dipole moment	ETA indices
ITH	Total information content on the leverage equality	GETAWAY descriptors
RDF050m	Radial Distribution Function - 050 / weighted by mass	RDF descriptors
G1u	1st component symmetry directional WHIM index / unweighted	WHIM descriptors
Mor28m	Signal 28 / weighted by mass	3D-MoRSE descriptors
Mor25m	Signal 25 / weighted by mass	3D-MoRSE descriptors
CATS3D_12_AA	CATS3D Acceptor-Acceptor BIN 12 (12.000 - 13.000 Å)	CATS 3D descriptors
Mor09u	Signal 09 / unweighted	3D-MoRSE descriptors
PJI3	3D Petitjeanshape index	Geometricaldescriptors
E2m	2nd component accessibility directional WHIM index / weighted by mass	WHIM descriptors
VE2_RG	Average coefficient of the last eigenvector (absolute values) from reciprocal squared geometrical matrix	3D matrix-baseddescriptors
TDB09v	3D Topological distance-based descriptors - lag 9 weighted by van der Waals volume	3D autocorrelations
JGI10	Mean topological charge index of order 10	2D autocorrelations
RDF080m	Radial Distribution Function - 080 / weighted by mass	RDF descriptors

RDF060m	Radial Distribution Function - 060 / weighted by mass	RDF descriptors
H-052	H attached to C^0_{sp3} with 1X attached to next C	Atom-centred fragments
RDF075m	Radial Distribution Function - 075 / weighted by mass	RDF descriptors

III.3.5.Results and Discussion

III.3.5.1.Descriptors calculation and selection of relevant descriptors

Using AlvaDesc, 5305 descriptors were calculated for each surfactant. The method used to select the most significant descriptors was described previously. 22descriptors were selected with stepwise regression. For each model, we used the same type and number of descriptors, as well as the same composition of the training and test sets. A set of the selected descriptors is summarized in table 2 with their signification and types.

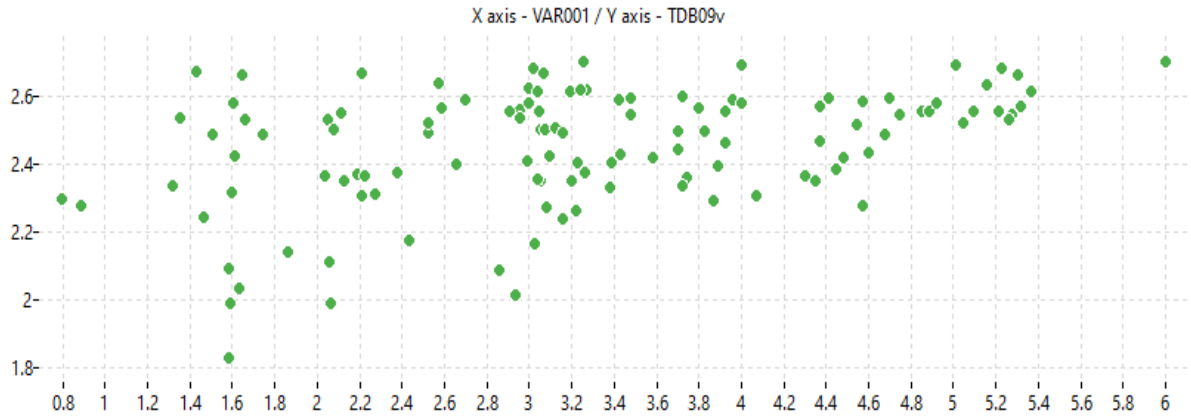


Figure III.3: Scatter plot representing the descriptor (X_{AXIS} – VAR001 /Y_{AXIS} – TDB09V) against the CMC values in the training set.

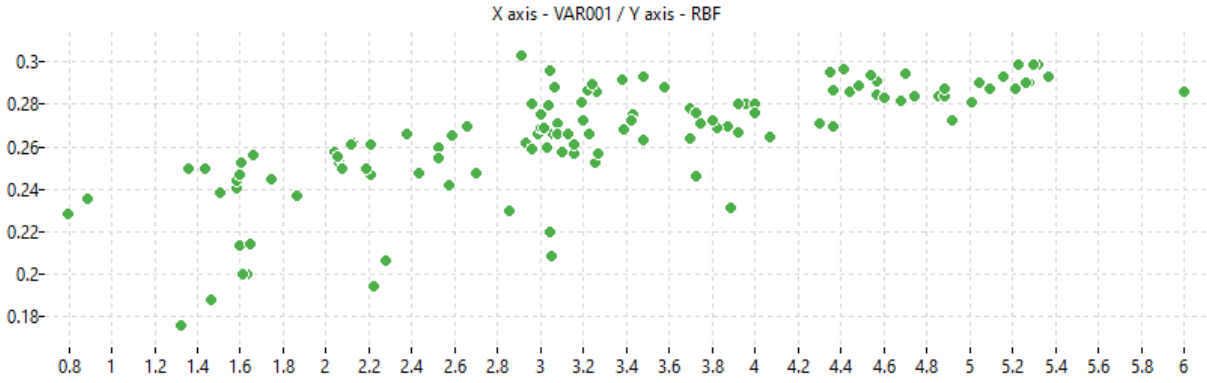


Figure III. 4: Scatter plot representing the descriptor (X_{AXIS}-VAR001/Y_{AXIS}-RBF) against the CMC values in the training set.

III.3.5.2.OLS model development:

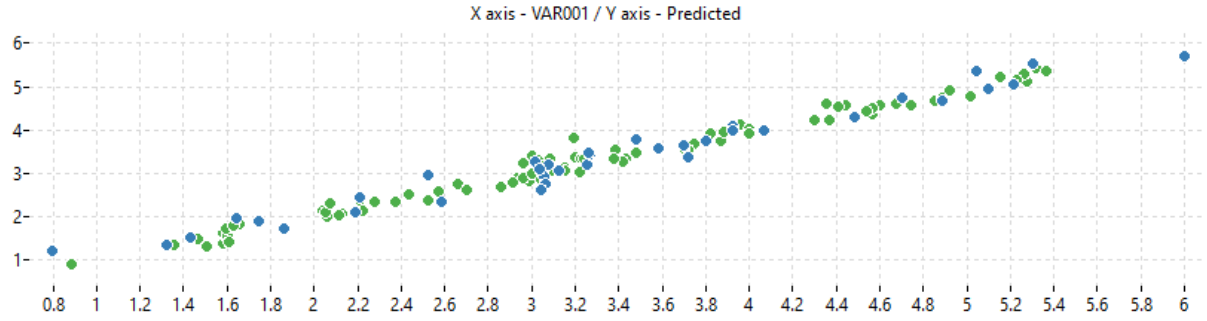


Figure III.5: Scatter plot of descriptor representing the predicted values against CMC values(X axis-VAR001/Yaxis –Predicted) by the OLS model for the train/test set.

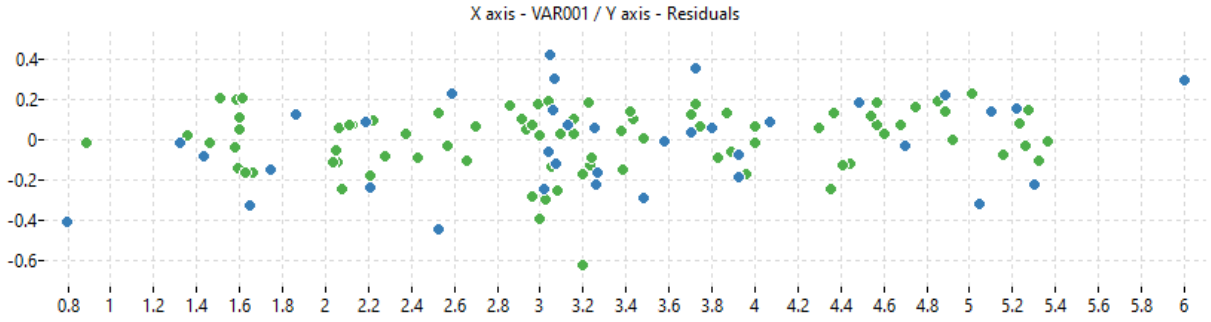


Figure III.6: Scatter plot of two descriptor representing the residuals values against the CMC (X_{axis} –VAR001/Y_{axis} –Residuals) by OLS model for the train/test set.

This model (OLS) shows a good correlation between the experimental and predicted values of the CMC .with a determination coefficient (R^2 =0.98) and (RMSE= 0.15) and confirmed by a range of residuals with vary between (-0.6- 0.4) for the global data set.

III.3.5.3.KNN model development

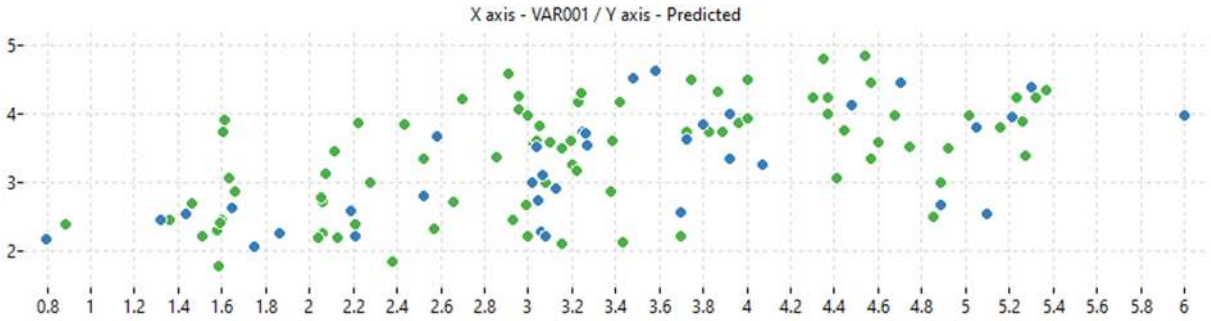


Figure III.7:Scatter plot of the descriptor representing the predicted values against the CMC values (X axis-VAR001/Yaxis –Predicted) by the KNN model for the train/test set.

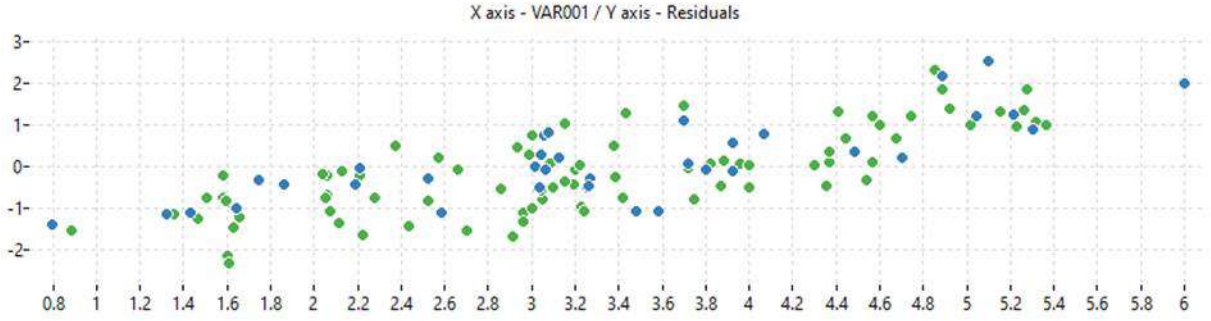


Figure III.8: Scatter plot of the descriptor representing the residuals values against the CMC values (X axis-VAR001/Yaxis –Predicted) by the KNN model for the train/test set.

This model (KNN) shows a bad correlation between the experimental and predicted values of the CMC .with a determination coefficient (R^2 = 0.275) and (RMSE=0.99) and confirmed by a range of residuals with vary between (-2- 3) for the global data set.

III.3.5.4.Consensus 1model

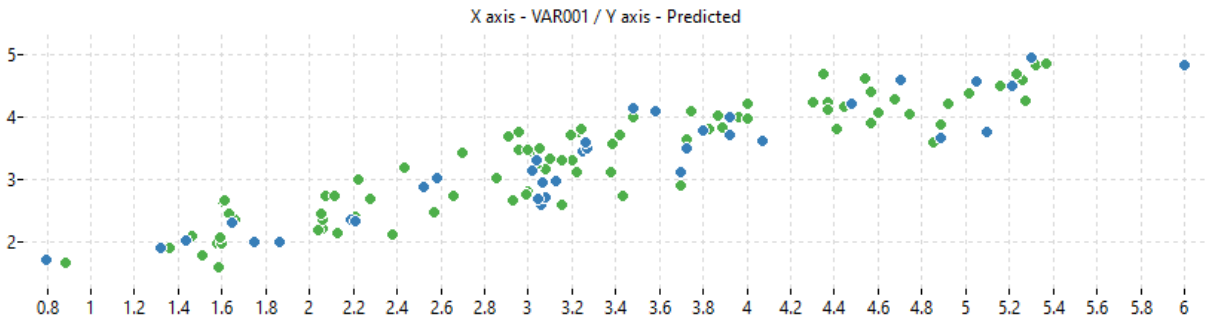


Figure III.9:Scatter plot of the descriptor representing the predicted values against the CMC values (X axis-VAR001/Yaxis –Predicted) by the Consensus 1 model for the train/test set.

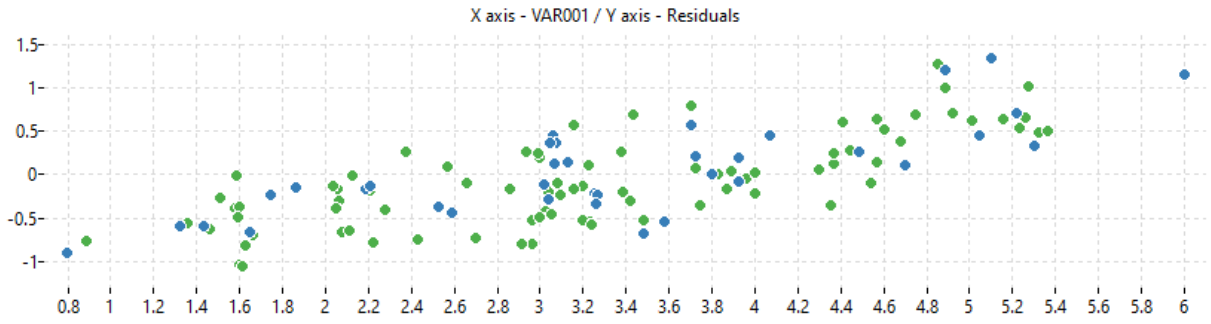


Figure III.10:Scatter plot of the descriptor representing the Residuals values against the CMC values (X axis-VAR001/Yaxis –Residuals) by the Consensus 1 model for the train/test set.

This model (**Consensus 1**) shows an acceptable correlation(in contrast to the KNN model), between the experimental and predicted values of the CMC .with a determination coefficient (R^2 =0.80) and (RMSE=0.51) and confirmed by a range of residuals with vary between (-1- 1.5) for the global data set.

III.3.5.5.KNN2 model with 166 MACCS fingerprints

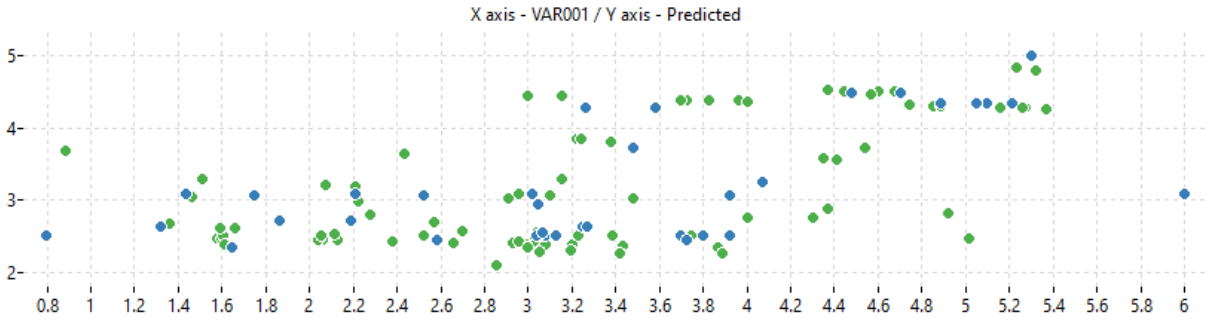


Figure III.11:Scatter plot of the descriptor representing the predicted values against the CMC values (X axis-VAR001/Yaxis –Predicted) by the KNN2 model for the train/test set.

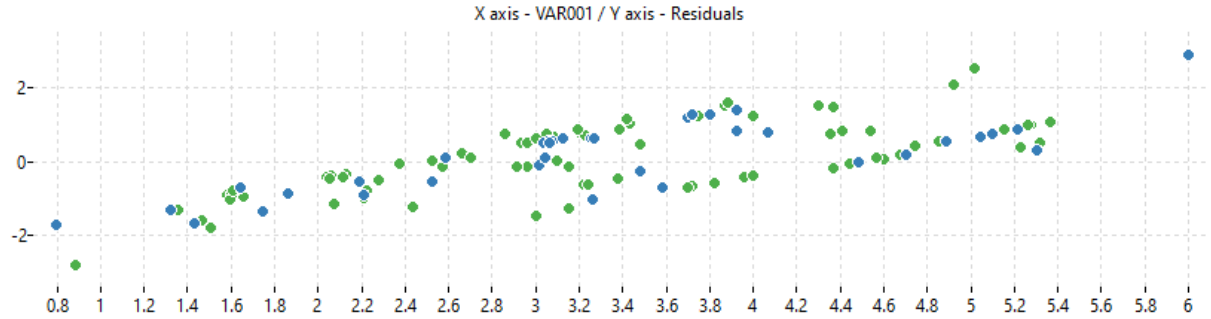


Figure III 12. Scatter plot of the descriptor representing the Residuals values against the CMC values (X axis-VAR001/Yaxis –Reseduals) by the KNN2 model for the train/test set.

This model (**KNN 2**) shows a bad correlation between the experimental and predicted values of the CMC .with a determination coefficient (R^2 =0.33) and (RMSE=0.93) and confirmed by a range of residuals with vary between (-2 - 2) for the global data set.

III.3.5.6.KNN3model with 1024 Hashed fingerprints

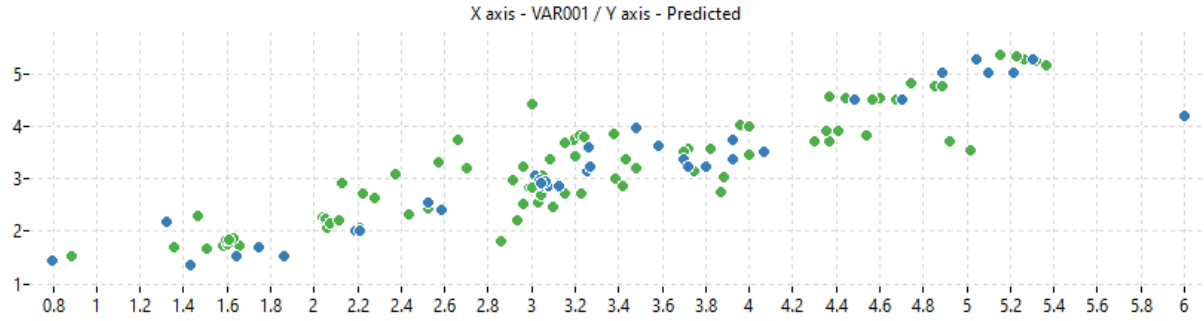


Figure III.13:Scatter plot of the descriptor representing the predicted values against the CMC values (X axis-VAR001/Yaxis –Predicted) by the KNN3 model for the train/test set.

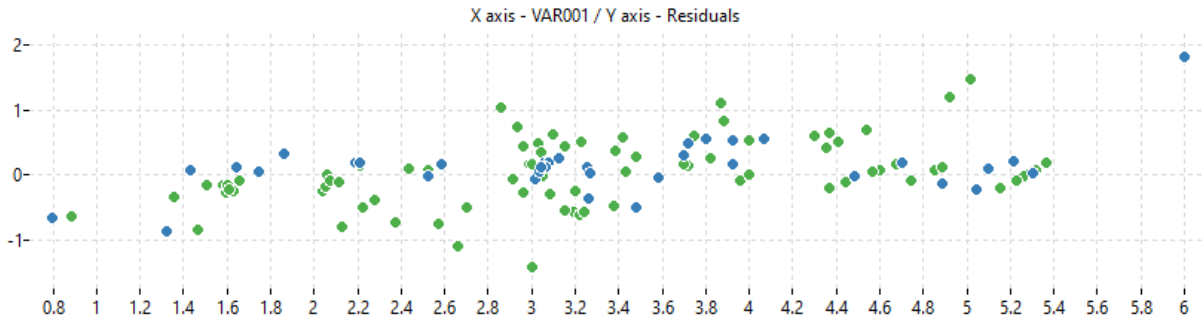


Figure III.14: Scatter plot of the descriptor representing the Residuals values against the CMC values (X axis-VAR001/Yaxis –Residuals) by the KNN3 model for the train/test set.

This model (**KNN 3**) shows an acceptable correlation between the experimental and predicted values of the CMC .with a determination coefficient (R^2 =0.81) and (RMSE=0.49) and confirmed by a range of residuals with vary between (-1 - 2) for the global data set.

III.3.5.7.Consensus 2 model of KNN2+KNN3

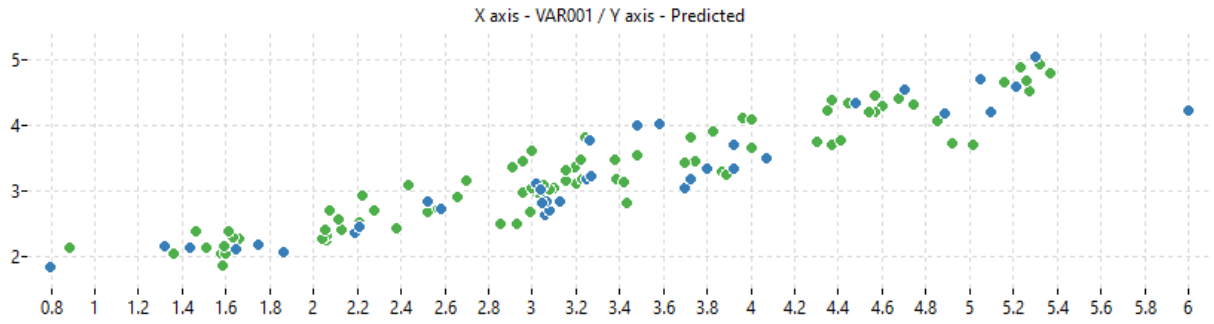


Figure III.15: Scatter plot of the descriptor representing the predicted values against the CMC values (X axis-VAR001/Yaxis –Predicted) by the Consensus 2 model for the train/test set.

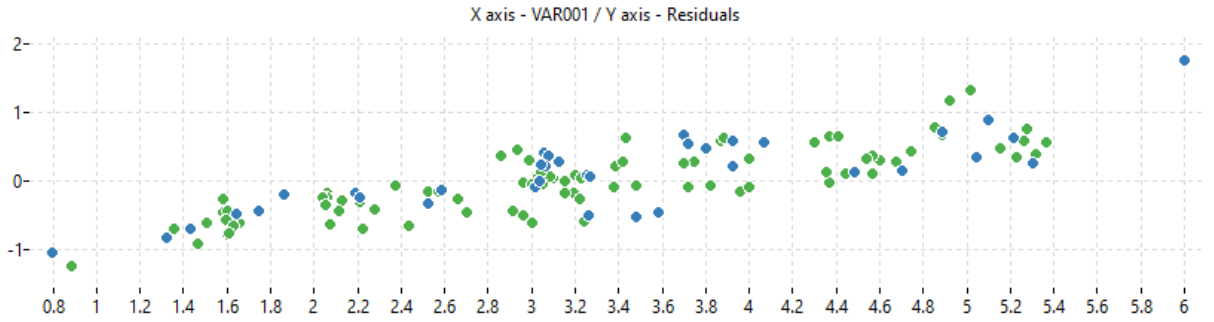


Figure III.16: Scatter plot of the descriptor representing the Residuals values against the CMC values (X axis-VAR001/Yaxis – Residuals) by the Consensus 2 model for the train/test set.

This model (**Consensus 2)** shows an acceptable correlation between the experimental and predicted values of the CMC .with a determination coefficient (R^2 =0.82) and (RMSE=0.48) and confirmed by a range of residuals with vary between (-1 - 2) for the global data set.

III.3.5.8.PLS

The results obtained with PLS model for the prediction of CMC of Gemini surfactants, using 120 compounds, are summarized below:

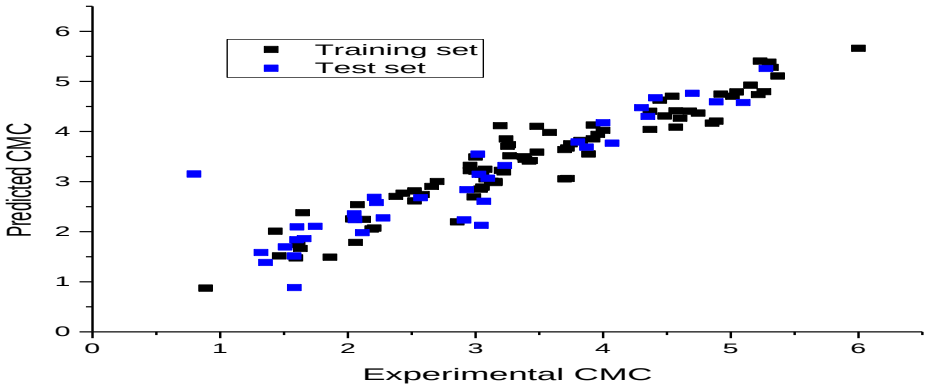


Figure III.17:Scatter plot of the descriptor representing the Predicted values against the CMC values by the PLS model for the train/test set.

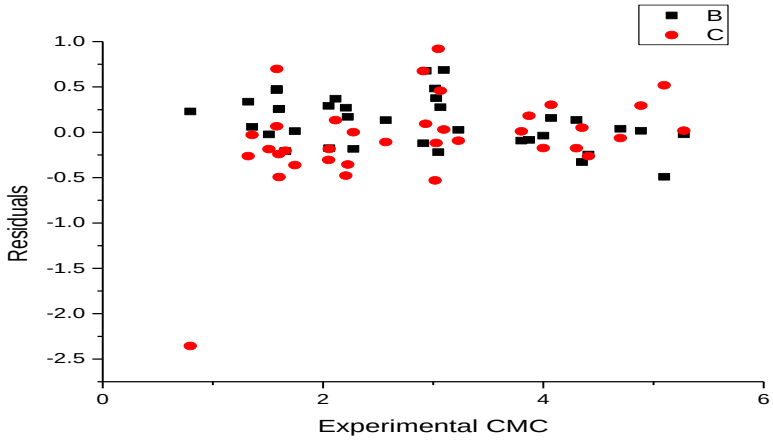
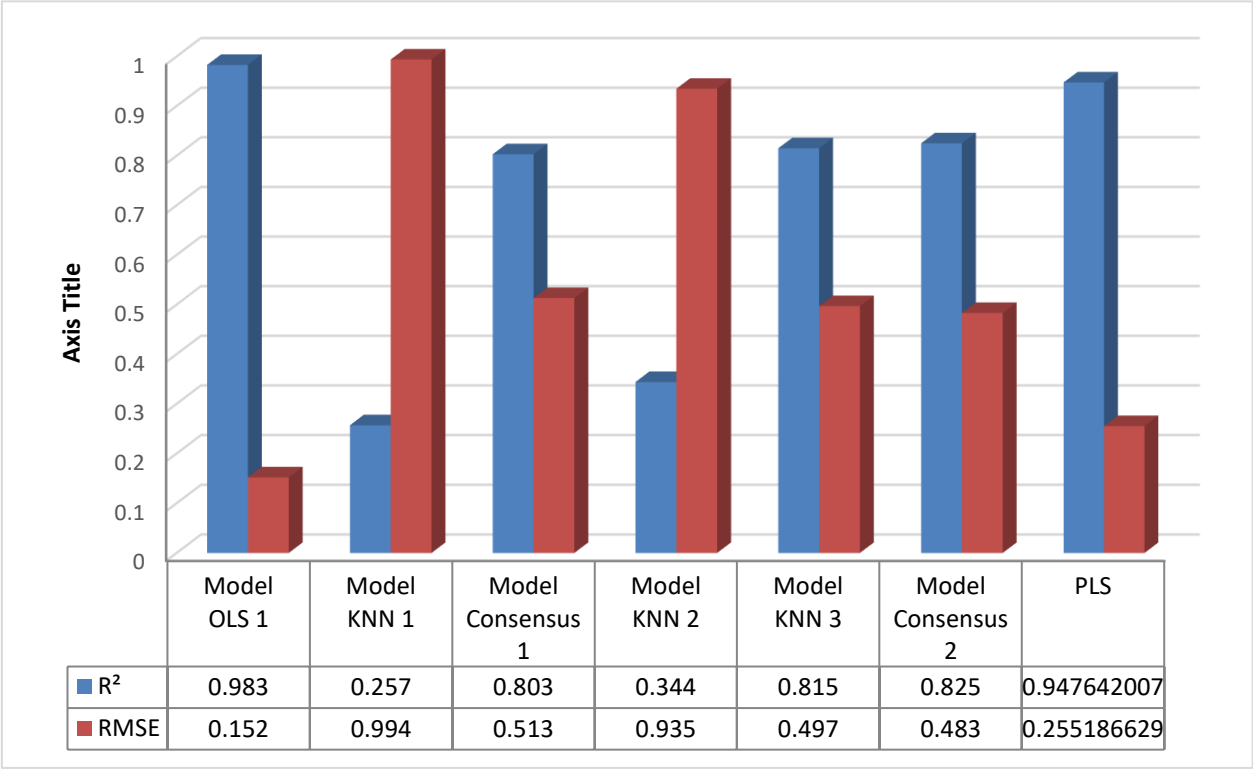


Figure III.18:Scatter plot of the descriptor representing the Residuals values against the CMC values by the PLS model for the train/test set

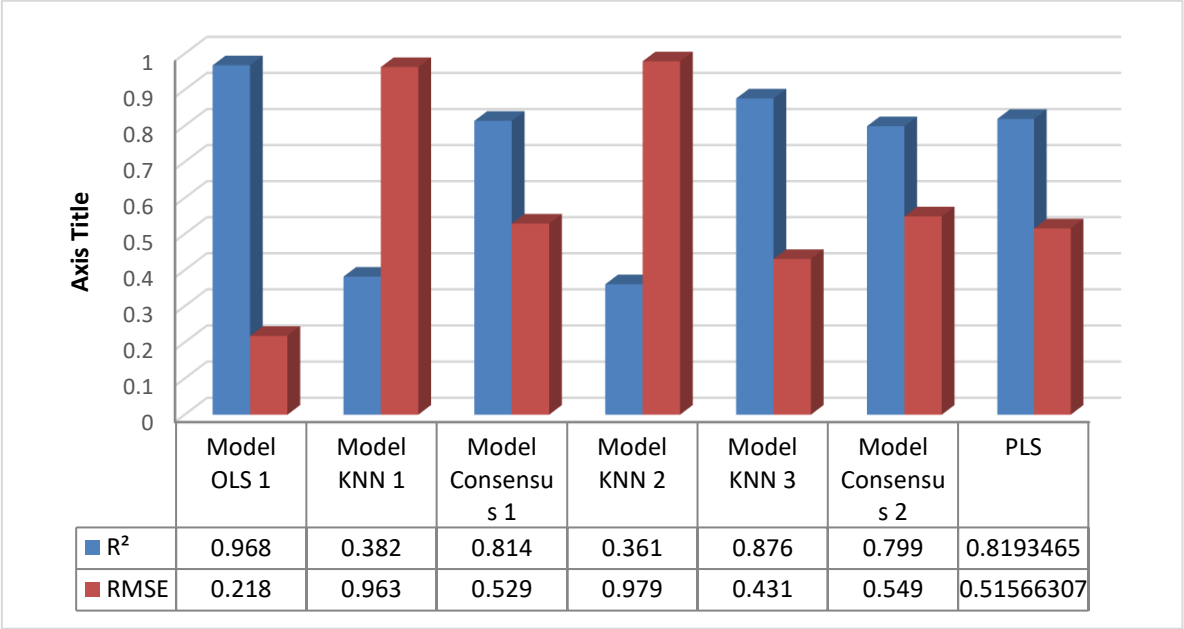
This model (**PLS)** shows an acceptable correlation between the experimental and predicted values of the CMC .with a determination coefficient (R^2 =0.94) and (RMSE=0.25) and confirmed by a range of residuals with vary between (-2.5 - 1) for the global data set.

A comparison between the tested models is represented



	Model OLS 1	Model KNN 1	Model Consensus 1	Model KNN 2	Model KNN 3	Model Consensus 2	PLS
R^2	0.983	0.257	0.803	0.344	0.815	0.825	0.947642007
RMSE	0.152	0.994	0.513	0.935	0.497	0.483	0.255186629

Figure III.19: Comparison of the performance of different models using different statistical parameters when predicting CMC values during the training stage (84 systems).



	Model OLS 1	Model KNN 1	Model Consensus 1	Model KNN 2	Model KNN 3	Model Consensus 2	PLS
R^2	0.968	0.382	0.814	0.361	0.876	0.799	0.8193465
RMSE	0.218	0.963	0.529	0.979	0.431	0.549	0.51566307

Figure III.20: Comparison of the performance of different models using different statistical parameters when predicting CMC values during the test stage (36 systems).

As can be seen in Figures (**III. 19** and **III.20)**, the comparison of the performance of different linear models (KNN PLS OLS Consensus) which review the R^2 and RMSE values to estimate the CMC of Gemini surfactant ,as we find out that the best(model) is the OLS for the two error optimized (R^2= 0.968)(RMES =0.218) in 36 systems.

Such as in 84 systems ($R^{2=}$0.983) (RMES =0.152) a quince to other model values .

III.4.Conclusion

The Ordinary Least Squares model developed in this study can be used to estimate the CMC of di-bromides Gemini surfactants based on 22 selected descriptors with high performance in comparison to the other tested models. The obtained deviation against the experimental values can be explained by the fact that the selected descriptors for QSRR analysis could not capture some relevant structural features that may be present in these compounds. The obtained OLS model can be used to estimate CMC of new Gemini surfactants if these later locate within the AD of the model. With

(R^2= 0.968) (RMES =0.218) in the test set.

General Conclusion:

The objective set for this work was to study the modeling of critical micelle concentration (CMC) of Gemini surfactants using QSPR model based on five regression models (KNN-PLS-OLS and Consensus, With defining the different methods used to model the physicochemical substance adapted to the surfactant, Building a theoretical approach QSPR to predict the best model to use.

In this study, we built the QSPR model which is a mathematical model accompanied with the exploration of the chemical information involved therein,and allow the prediction of a chemical response of relatively large number of compounds, and employing less computational time, where it can provide valuable information by aiding rational designing strategy with minimal cost involvement.

QSPR models were established with a successful strategy for predicting the CMC of surfactant.The data set was collected from previous published literature containing 120 Gemini surfactants with their CMC which represent the input and the output was considered as the descriptors calculated.

To build the QSPR model five regression methods(PLS,KNN AND OLS) were used and modeled using the AlvaDesc software and for the PLS model was model using a free network application.

As a consequence a comparison was conducted between all the models,and we concluded that the OLS was the best model for predicting the CMC of Gemini Surfactants. With (R2= 0.968)(RMES =0.218) in the test set.

Reference chapter I

[1] L. L. Schramm, E. N. Stasiuk, and D. G. Marangoni, “Surfactants and their applications,” Annu. Reports Prog. Chem. - Sect. C, vol. 99, no. 2, pp. 3–48, 2003, doi: 10.1039/B208499F.

[2] R. A. Llenado" and R. A. Jamieson Procter and Gamble Company, Ivorydale Technical Center, Cincinnati, Ohio 452 17 Surfactants.

[3] J. Hu, X. Zhang, and Z. Wang, “A review on progress in QSPR studies for surfactants,” Int. J. Mol. Sci., vol. 11, no. 3, pp. 1020–1047, 2010, doi: 10.3390/ijms11031020.

[4] K. Holmberg, B. Jonsson, B. Kronberg, B. Lindman, Surfactants and Polymers in Aqueous Solution, 2nd Ed., John Wiley and Sons, New York, 2002

[5] R. Kurrey, M. Mahilang, M. K. Deb, and K. Shrivas, “Analytical approach on surface active agents in the environment and challenges,” Trends Environ. Anal. Chem., vol. 21, no. January, p. e00061, 2019, doi: 10.1016/j.teac.2019.e00061.

[6] “Nonylphenol Ethoxylate | Supplier Bahan Kimia, Supplier Fosfat, Jual Amine Indonesia, Laju Usaha Gemilang.” https://lugchemical.co.id/product/nonylphenol-ethoxylate/ (accessed Aug. 14, 2020).

[7] Y. Nakama, Surfactants. Elsevier Inc., 2017.

[8] “Surfactants Types and Uses,” vol. 2, 2002.

[9] “Primary & Secondary Surfactant | Nonionic, Anionic, Cationic, Amphoteric.” https://www.oxiteno.us/what-is-nonionic-amphoteric-cationic-anionic-surfactant-types-primary-secondary/ (accessed Aug. 21, 2020).

[10] “Understanding Surfactants: A Guide - International Products Corporation.” https://www.ipcol.com/blog/an-easy-guide-to-understanding-surfactants/ (accessed Aug. 21, 2020).

[11] L. Tensioactifs, “Chapitre II Généralités sur les tensioactifs Chapitre II Généralités sur les tensioactifs,” pp. 10–19.

[12] D. Shukla and V. K. Tyagi, “Anionic Gemini Surfactants: A Distinct Class of Surfactants,” J. Oleo Sci., vol. 55, no. 5, pp. 215–226, 2006, doi: 10.5650/jos.55.215.

[13] Rosen M. J., “kinetic studies of some esters and amides in presence of micelles Chapter - 2 Critical Micelle Concentration of Surfactant, Mixed Surfactant and Polymer By Different Methods at Room Temperature And Its Importance,” Kinet. Stud. Some Esters Amides Presence Micelles Kinet. Stud. Some Ester, pp. 55–105, 2004.

[14] E. Setiawan, M. Mudasir, and K. Wijaya, “QSPR Models for Predicting Critical Micelle

Concentration of Gemini Cationic Surfactants Combining Machine-Learning Methods and Molecular Descriptors," no. 1, 2020, doi: 10.26434/chemrxiv.11626068.

[15] S. De, S. Malik, A. Ghosh, R. Saha, and B. Saha, "A review on natural surfactants," RSC Adv., vol. 5, no. 81, pp. 65757–65767, 2015, doi: 10.1039/c5ra11101c.

[16] A. D. Venereol, "Allergenic Oxidation Products INTRODUCTION," Scand. Univ. Press. ISSN 0001-5555, no. 18, 1999.

[17] Yunfei He, "Study on the interfacial properties of surfactants and their interactions with DNA," 2013, [Online]. Available: https://tel.archives-ouvertes.fr/tel-00852186/document.

[18] Sweeta Akbari and Abdurahman Hamid Nour, "Emulsion types, stability mechanisms and rheology: A review," Int. J. Innov. Res. Sci. Stud., vol. 1, no. 1, pp. 14–21, 2018, doi: 10.1016/j.colsurfa.2007.03.036.

[19] B. Zunic and S. Peter, "World ' s largest Science , Technology & Medicine Open Access book publisher," no. 2016, pp. 267–322, 2018.

[20] R. Pal, "Rheology of simple and multiple emulsions," Curr. Opin. Colloid Interface Sci., vol. 16, no. 1, pp. 41–60, 2011, doi: 10.1016/j.cocis.2010.10.001.

[21] F. Goodarzi and S. Zendehboudi, "A Comprehensive Review on Emulsions and Emulsion Stability in Chemical and Energy Industries," Can. J. Chem. Eng., vol. 97, no. 1, pp. 281–309, 2019, doi: 10.1002/cjce.23336.

[22] G. Muschiolik, "Multiple emulsions for food use," Curr. Opin. Colloid Interface Sci., vol. 12, no. 4–5, pp. 213–220, 2007, doi: 10.1016/j.cocis.2007.07.006.

[23] "What Is an Emulsion? Definition and Examples." https://www.thoughtco.com/definition-of-emulsion-605086 (accessed Aug. 21, 2020).

[24] S. N. Kale and S. L. Deore, "Emulsion micro emulsion and nano emulsion: A review," Syst. Rev. Pharm., vol. 8, no. 1, pp. 39–47, 2016, doi: 10.5530/srp.2017.1.8.

[25] R. Pedada, E. Vanka, and S. AMS, "Enhancement of Solubility; an Over View," PharmaTutor, vol. 1, no. 2, pp. 60–74, 2013.

[26] B. Robertson and J. Johansson, "Surfactant," Neonatal Respir. Disord. Second Ed., pp. 12–25, 2003, doi: 10.4011/shikizai1937.65.634.

[27] "Surfactants - PCC Group Product Portal." https://www.products.pcc.eu/en/k/surfactants/ (accessed Aug. 21, 2020).

[28] S. Sonawane, S. Pal, and N. Bisht, "Application of Surfactant in various Fields," Int. J. Sci. Eng. Res., vol. 6, no. 12, pp. 30–32, 2015, [Online]. Available: https://www.ijser.org/researchpaper/Application-of-Surfactant-in-various-Fields.pdf.

[29] S. K. Shah, A. Bhattarai, and S. K. Chatterjee, “Applications of Surfactants in modern Science and Technology,” Mod. Trends Sci. Technol., no. May 2014, pp. 147–158, 2013.

[30] T. Ivanković and J. Hrenović, “Surfactants in the environment,” Arh. Hig. Rada Toksikol., vol. 61, no. 1, pp. 95–110, 2010, doi: 10.2478/10004-1254-61-2010-1943.

[31] Y. Moroi, “Surface-Active Agents 2. 1.,” pp. 7–24, 1992.

[32] S.S.Borse, Patil T.J. and Borse M.S., Tenside Surfactants Detergents, 51 (3) (2014) 267-273. “Effect of tuned head polarity of cetyl trimethyl ammonium bromide on their physicochemical properties”.

Reference chapiter II

[1] V. Mandlik, P. R. Bejugam, and S. Singh, Application of Artificial Neural Networks in Modern Drug Discovery. Elsevier Inc., 2016.

[2] P. D. T. Huibers, V. S. Lobanov, A. R. Katritzky, D. O. Shah, and M. Karelson, “Prediction of critical micelle concentration using a quantitative structure-property relationship approach. 1. Nonionic surfactants,” Langmuir, vol. 12, no. 6, pp. 1462–1470, 1996, doi: 10.1021/la950581j.

[3] K. Roy, S. Kar, and R. N. Das, A Primer on QSAR/QSPR Modeling: Fundamental Concepts, vol. 1. 2015.

[4] J. Hu, X. Zhang, and Z. Wang, “A review on progress in QSPR studies for surfactants,” Int. J. Mol. Sci., vol. 11, no. 3, pp. 1020–1047, 2010, doi: 10.3390/ijms11031020.

[5] M. E. Lacy, “Neural network technology and its application in chemical research,” Tetrahedron Comput. Methodol., vol. 3, no. 3–4, pp. 119–128, 1990, doi: 10.1016/0898-5529(90)90047-C.

[6] Y. C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, “Machine learning in chemoinformatics and drug discovery,” Drug Discov. Today, vol. 23, no. 8, pp. 1538–1546, 2018, doi: 10.1016/j.drudis.2018.05.010.

[7] S. Yousefinejad and B. Hemmateenejad, “Chemometrics tools in QSAR/QSPR studies: A historical perspective,” Chemom. Intell. Lab. Syst., vol. 149, pp. 177–204, 2015, doi: 10.1016/j.chemolab.2015.06.016.

[8] S. Kim and K. Cho, “PyQSAR: A Fast QSAR Modeling Platform Using Machine Learning and Jupyter Notebook,” Bull. Korean Chem. Soc., vol. 40, no. 1, p. bkcs.11638, Dec. 2018, doi: 10.1002/bkcs.11638.

[9] H. Schmidli, “Quantitative Structure Activity Relationships (QSAR),” pp. 5–15, 1995, doi: 10.1007/978-3-642-50015-2_2.

[10] G. Absalan, B. Hemmateenejad, M. Soleimani, M. Akhond, and R. Miri, “Quantitative structure - Micellization relationship study of gemini surfactants using genetic-PLS and genetic-MLR,” QSAR Comb. Sci., vol. 23, no. 6, pp. 416–425, 2004, doi: 10.1002/qsar.200430872.

[11] T. Puzyn, J. Leszczyński, and M. Cronin, Recent Advances in QSAR Studies: Methods and Applications. 2010.

[12] J. Leszczynski, A. Kaczmarek-Kedziera, T. Puzyn, M. G. Papadopoulos, H. Reis, and M. K. Shukla, “Handbook of computational chemistry,” Handb. Comput. Chem., pp. 1–2381, 2017, doi: 10.1007/978-3-319-27282-5.

[13] D.-D. Varsou, S. Nikolakopoulos, A. Tsoumanis, G. Melagraki, and A. Afantitis, Chapter 14 Discovery and Computational Toxicology, vol. 1800, no. C. .

[14] M. K. Arelson, U. M. Aran, Y. W. Ang, and A. R. K. Atritzky, “1551 QSPR AND QSAR MODELS DERIVED USING LARGE MOLECULAR DESCRIPTOR SPACES . A REVIEW OF CODESSA APPLICATIONS The fast progress in modern computer technology has created an entirely new environment for the efficient use of the theoretical constructions of,” vol. 64, pp. 1551–1571, 1999.

[15] A. R. Katritzky and E. V. Gordeeva, “Traditional Topological Indices vs Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research,” J. Chem. Inf. Comput. Sci., vol. 33, no. 6, pp. 835–857, 1993, doi: 10.1021/ci00016a005.

[16] “QSAR Methodology and ADAPT The previous chapter described a number of modeling techniques with which,” pp. 56–88.

[17] D. Rivero, E. Fernandez-Blanco, J. Dorado, and A. Pazos, “A new signal classification technique by means of Genetic Algorithms and kNN,” 2011 IEEE Congr. Evol. Comput. CEC 2011, pp. 581–586, 2011, doi: 10.1109/CEC.2011.5949671.

[18] J. M. Keller and M. R. Gray, “A Fuzzy K-Nearest Neighbor Algorithm,” IEEE Trans. Syst. Man Cybern., vol. SMC-15, no. 4, pp. 580–585, 1985, doi: 10.1109/TSMC.1985.6313426.

[19] A. C. Dimian, C. S. Bildea, and A. A. Kiss, Chemical Product Design, vol. 35. 2014.

[20] G. Mateos-Aparicio, “Partial least squares (PLS) methods: Origins, evolution, and application to social sciences,” Commun. Stat. - Theory Methods, vol. 40, no. 13, pp. 2305–2317, 2011, doi: 10.1080/03610921003778225.

[21] P. A. Martins, G. Barbosa, K. F. M. Pasqualoto, and M. C. Ferreira, “LQTA-QSAR : A

New 4D-QSAR Methodology," pp. 1428–1436, 2009.

Reference chapter III

[1] SK Yadav, Parikh K, Kumar S (2017) Solubilization potentials of single and mixed oppositely charged gemini surfactants: A case of polycyclic aromatic hydrocarbons. Colloids Surfaces A Physicochem Eng Asp 514:47–55. https://doi.org/10.1016/j.colsurfa.2016.11.042

[2] Zhu HL, Hu ZY, Wang JL, Cao DL (2014) Micellization of gemini surfactant with dibenzyl ether spacer in water-organic mixed media. J Mol Liq 195:54–58. https://doi.org/10.1016/j.molliq.2014.01.030

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

[4] Sharma R, Kamal A, Abdinejad M, et al (2017) Advances in the synthesis, molecular architectures and potential applications of gemini surfactants. Adv Colloid Interface Sci 248:35–68. https://doi.org/10.1016/j.cis.2017.07.032

[5] Shakil Hussain SM, Kamal MS, Murtaza M (2019) Effect of aromatic spacer groups and counterions on aqueous micellar and thermal properties of the synthesized quaternary ammonium gemini surfactants. J Mol Liq 296:111837. https://doi.org/10.1016/j.molliq.2019.111837

[6] Yi M, Huang Z, Hao J (2019) Magnetic Gemini Surfactants. Langmuir 35:9538–9545. https://doi.org/10.1021/acs.langmuir.9b01478

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

[8] Ma T, Feng H, Wu H, et al (2019) Property evaluation of synthesized anionic-nonionic gemini surfactants for chemical enhanced oil recovery. Colloids Surfaces A Physicochem Eng Asp 581:123800. https://doi.org/10.1016/j.colsurfa.2019.123800

[9] Cardoso AM, Morais CM, Cruz AR, et al (2015) Gemini Surfactants Mediate Efficient Mitochondrial Gene Delivery and Expression. Mol Pharm 12:716–730. https://doi.org/10.1021/mp5005349

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

https://doi.org/10.1016/j.chemosphere.2011.05.031

[11] Jiao L, Wang Y, Qu L, et al (2020) Hologram QSAR study on the critical micelle concentration of Gemini surfactants. Colloids Surfaces A Physicochem Eng Asp 586:. https://doi.org/10.1016/j.colsurfa.2019.124226

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

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

[14] Roberts DW (2002) Application of octanol/water partition coefficients in surfactant science: A quantitative structure-property relationship for micellization of anionic surfactants. Langmuir 18:345–352. https://doi.org/10.1021/la0108050

[15] Katritzky AR, Pacureanu LM, Slavov SH, et al (2009) QSPR study of the first and second critical micelle concentrations of cationic surfactants. Comput Chem Eng 33:321–332. https://doi.org/10.1016/j.compchemeng.2008.09.011

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

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

[18] Bhhatarai B, Gramatica P (2011) Prediction of aqueous solubility, vapor pressure and critical micelle concentration for aquatic partitioning of perfluorinated chemicals. Environ Sci Technol 45:8120–8128. https://doi.org/10.1021/es101181g

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

[20] Poša M (2011) QSPR study of the effect of steroidal hydroxy and oxo substituents on the critical micellar concentration of bile acids. Steroids 76:85–93. https://doi.org/10.1016/j.steroids.2010.09.003

[21] Jiao L, Wang X, Bing S, et al (2015) Quim. Nova,. 38:510–517

[22] Martens HA, Dardenne P (1998) Validation and verification of regression in small data sets. Chemom Intell Lab Syst 44:99–121. https://doi.org/10.1016/S0169-7439(98)00167-1

[23] Pal N, Saxena N, Mandal A (2017) Equilibrium and dynamic adsorption of gemini surfactants with different spacer lengths at oil/aqueous interfaces. Colloids Surfaces A Physicochem Eng Asp 533:20–32. https://doi.org/10.1016/j.colsurfa.2017.08.020

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

[25] Jiao L, Wang X, Bing S, et al (2015) QSPR study on the photolysis half-life of PCDD/Fs adsorbed on spruce (Picea abies (L.) Karst.) needle surfaces under sunlight irradiation by using a molecular distance-edge vector index. RSC Adv 5:6617–6624. https://doi.org/10.1039/c4ra14178d

[26] Jiao L, Zhang X, Qin Y, et al (2016) Hologram QSAR study on the electrophoretic mobility of aromatic acids. Chemom Intell Lab Syst 157:202–207. https://doi.org/10.1016/j.chemolab.2016.06.020

[27] Parikh K, Singh S, Desai A, Kumar S (2019) An interplay between spacer nature and alkyl chain length on aqueous micellar properties of cationic Gemini surfactants: A multi-technique approach. J Mol Liq 278:290–298. https://doi.org/10.1016/j.molliq.2019.01.044

[28] Jiao L, Xue Z, Wang G, et al (2014) QSPR study on the relative retention time of polybrominated diphenyl ethers (PBDEs) by using molecular distance-edge vector index. Chemom Intell Lab Syst 137:91–96. https://doi.org/10.1016/j.chemolab.2014.06.015

[29] Kulkarni AS, Kasabe AJ, Bhatia MS, et al (2019) Quantitative Structure–Property Relationship Approach in Formulation Development: an Overview. AAPS PharmSciTech 20:. https://doi.org/10.1208/s12249-019-1480-2

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

[31] Hu CY, Hua SJ, Lin YL, et al (2020) Kinetics and formation of disinfection byproducts during iohexol chlor(am)ination. Sep Purif Technol 243:116797. https://doi.org/10.1016/j.seppur.2020.116797

[32] Dreja M, Gramberg S, Tieke B (1998) Cationic amphitropic gemini surfactants with hydrophilic oligo(oxyethylene) spacer chains. Chem Commun 1371–1372.

https://doi.org/10.1039/a801448e

[33] Kumar A, Alami E, Holmberg K, et al (2003) Branched zwitterionic gemini surfactants micellization and interaction with ionic surfactants. Colloids Surfaces A Physicochem Eng Asp 228:197–207. https://doi.org/10.1016/S0927-7757(03)00300-5

[34] Koya PA (2010) Gemini Surfactant in Water + Organic Solvent Mixed Media †. 1921–1929

[35] Dreja M, Pyckhout-Hintzen W, Tieke B (1998) Copolymerization behaviour and structure of styrene and polymerizable surfactants in three-component cationic microemulsion. Macromolecules 31:272–280. https://doi.org/10.1021/ma971153i

[36] Jiang X, Zhou L, Li Y, et al (2007) Synthesis and properties of a novel class of gemini pyridinium surfactants. Langmuir 23:11404–11408. https://doi.org/10.1021/la701154w

[37] Khalaf MM, Tantawy AH, Soliman KA, Abd El-Lateef HM (2020) Cationic gemini-surfactants based on waste cooking oil as new 'green' inhibitors for N80-steel corrosion in sulphuric acid: A combined empirical and theoretical approaches. J Mol Struct 1203:127442. https://doi.org/10.1016/j.molstruc.2019.127442

[38] Dreja M, Tieke B (1998) Polymerization of styrene in ternary microemulsion using cationic gemini surfactants. Langmuir 14:800–807. https://doi.org/10.1021/la9710738

[39] preview.pdf

[40] Wu J, Gao H, Shi D, et al (2020) Cationic gemini surfactants containing both amide and ester groups: Synthesis, surface properties and antibacterial activity. J Mol Liq 299:112248. https://doi.org/10.1016/j.molliq.2019.112248

[41] Rosen MJ, Liu L (1996) Surface activity and premicellar aggregation of some novel diquaternary gemini surfactants. JAOCS, J Am Oil Chem Soc 73:885–890. https://doi.org/10.1007/BF02517990

[42] Alami E, Beinert G, Marie P, Zana R (1993) Alkanediyl-α,ω-bis(dimethylalkylammonium bromide) Surfactants. 3. Behavior at the Air-Water Interface. Langmuir 9:1465–1467. https://doi.org/10.1021/la00030a006

[43] Dreja M, Heine K, Tieke B, Junkers G (1997) Effects of functionalized latex particles and anionic surfactants on the flow behavior of aqueous gelatin dispersions. J Colloid Interface Sci 191:131–140. https://doi.org/10.1006/jcis.1997.4936

[44] Savelli G, Brinchi L, Germani R (2001) Reactivity Control by Aqueous Amphiphilic Self-Assembling Systems

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

Q

[46] Dreja M, Pyckhout-Hintzen W, Mays H, Tieke B (1999) Cationic Gemini Surfactants with Oligo(oxyethylene) Spacer Groups and Their Use in the Polymerization of Styrene in Ternary Microemulsion. Langmuir 15:391–399. https://doi.org/10.1021/la981354v.