République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE DE MEDEA
Faculté des Sciences et de la technologie
Département de Génie des Procédés et Environnement

MEMOIRE DE FIN D'ETUDES

En vue de l'obtention du Diplôme de Master en Génie des Procédés
Option : Pollution Chimique et Gestion Environnementale

Thème

Modélisation des propriétés physico-chimiques des tensionactifs

La soutenance aura lieu le **24/05/2017**

Réalisé par :
Mr. KOURDALI Mohamed El Amine
Mr. LAROUCI Boualem

Proposé et dirigé par :
Dr: RAHAL Soufiane

Promotion : 2016/2017


بسم الله الرحمن الرحيم

Université Yahia Fares de Médéa

Faculté de Technologie

Département de Génie de Procédés et Environnement

Mémoire de Master

Spécialité: Génie de Procédés

Option: Pollution Chimique et Gestion Environnementale

Modélisation des propreté physico-chimique des tensioactifs

Réalisé par :
Mr. KOURDALI Mohamed El Amine
Mr. LAROUCI Bou alem

Proposé et dirigé par :
Dr: RAHAL Soufiane

Date de soutenance : 24/05/2017

Plan d’exposé

1- Introduction générale

2- Généralités sur les tensioactifs

3- Modélisation de la CMC par la méthode QSPR-RNA

- TAs non-ioniques
- TAs cationiques
- TAs anioniques
- TAs global

4- Conclusions générales

Introduction générale

Les agents de surface ou tensio-actifs (surfactant) sont utilisés depuis longtemps par l'homme.Les tensioactifs sont des produits polyvalents utilisés dans les industries variétales telles que les détergents, les émulsifiants, les boues de forage et les agents de flottation.

À une concentration critique appelée CMC, les tensioactifs se rassemblent pour former des micelles. Les micelles sont en équilibre avec les monomères .le point particulier important dans l'industrie est la détermination et la prédiction de la CMC.

L'une des méthodes pour prédire les propriétés chimiques telles que la CMC basée uniquement sur l'information structurelle moléculaire est la modélisation quantitative de la structure propriété (QSPR). Les modèles QSPR sont des équations mathématiques qui relient la structure chimique à une propriété.
La modélisation QSPR est extrêmement utile pour prédire la propriété de nouveaux composés chimiques.

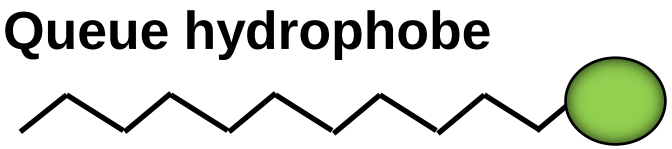
Dans le présent travail, les meilleures variables parmi les descripteurs électroniques et topologiques sont sélectionnées à l'aide des algorithmes génétiques, puis un modèle linéaire est développé pour prédire le Log CMC des tensioactifs à l'aide des RNA. Notre méthode montre une amélioration considérable pour les modèles de régression.

Généralités sur les tensioactifs: Définition

- **Les TAs sont des molécules amphiphiles qui modifient la tension superficielle entre deux surfaces.**



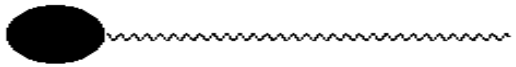
- **Elles contiennent à la fois :**
  1. **Une partie polaire (tête hydrophile)**
  2. **Une autre partie apolaire (partie lipophile)**

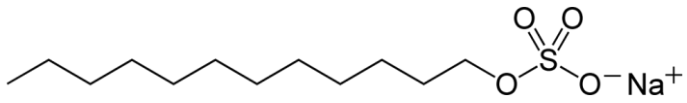


Généralités sur les tensioactifs: Types

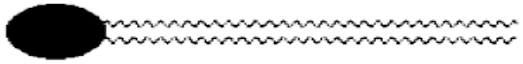
Les tensioactifs, quelle que soit la nature des têtes polaires, peuvent présenter des structures très diverses

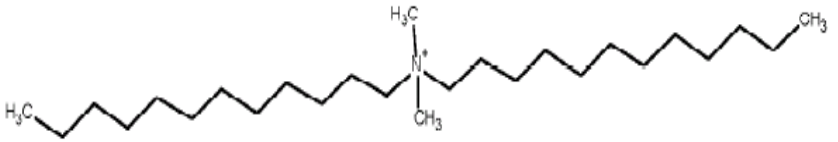
- **Monocaténaire**



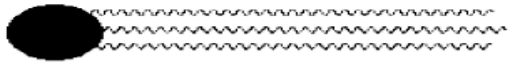


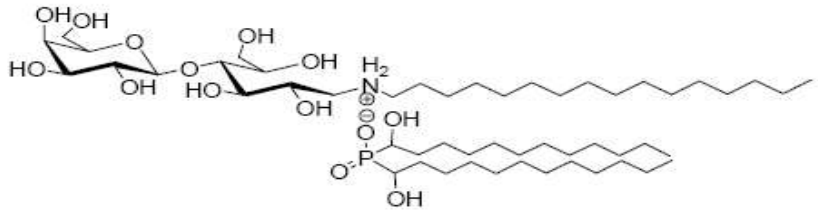
- **Bicaténaire**



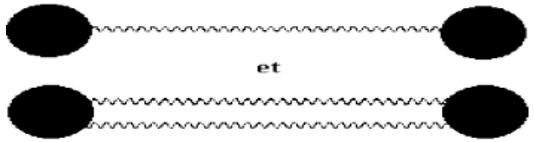


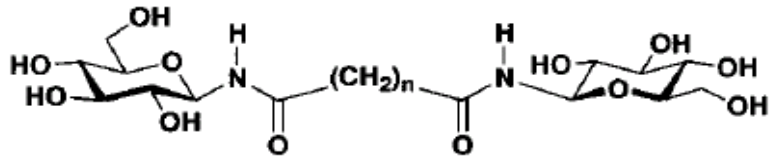
- **Tricaténaire**





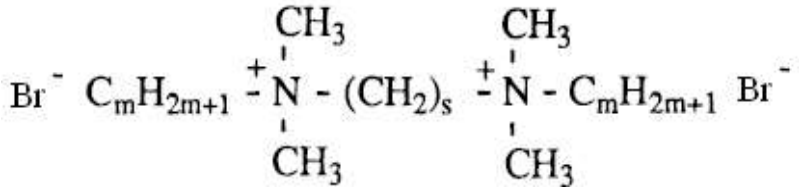
- **Bolaformes**





- **Gemini**





Structuration des tensioactifs en solution aqueuse

A faible concentration: Les tensioactifs moléculaires s'agrègent sous forme de micelles dispersées dans l'eau.

Ces micelles sont le plus souvent sphériques mais peuvent être aussi cylindriques

A concentration plus élevée: les agrégats deviennent des cylindres infinis, organisés suivant un réseau qu'on caractérise quantitativement par le paramètre d'empilement

$$p = v / a_0 lc$$

- a_0: Surface de la tête polaire du tensioactif
- v: Volume chaîne aliphatique
- lc: Longueur maximale de la chaîne aliphatique

- Micelle directe (sphère): P < 1/3



- Disque (cylindre): 1/3 < P <1/2



- Lamelle: 1/2 < P < 1



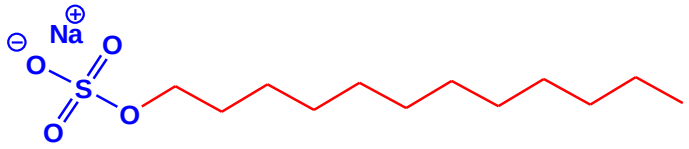
- Micelle inverse: P > 1



Classification des tensioactifs

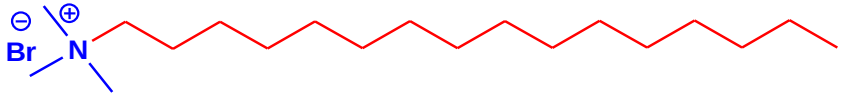
Il y a plusieurs types de classifications des tensioactifs selon :

1. **Longueur de la queue hydrophobe**
2. **Origine naturelle ou synthétique**
3. **Nature de la tête polaire (pôle hydrophile) en trouve quatre grandes familles:**



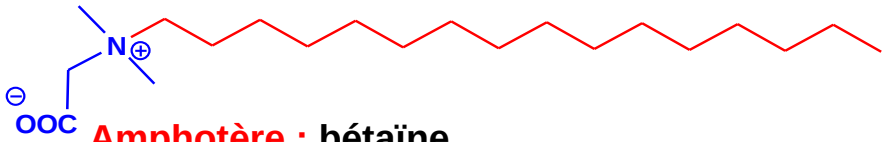
Anionique : Sodium Dodecyl Sulfate (SDS)

Anionique dont la partie hydrophobe de la molécule porte une charge négative comme les ions carboxylates (O-)



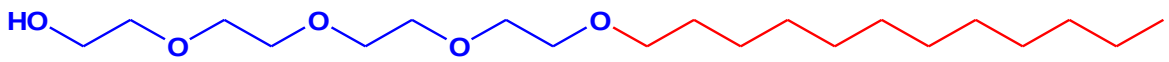
Cationique : CetylTrimethylAmmonium Bromide (CTAB)

Cationique dont la partie hydrophobe porte une charge positive résiduelle, comme les amines protonées



Amphotère : bétaïne

Amphotère au zwitterioniques aux centres positifs et négatifs de quel sont être trouvée dans la molécule



Non-ionique : polyoxyéthylène-4-lauryl éther

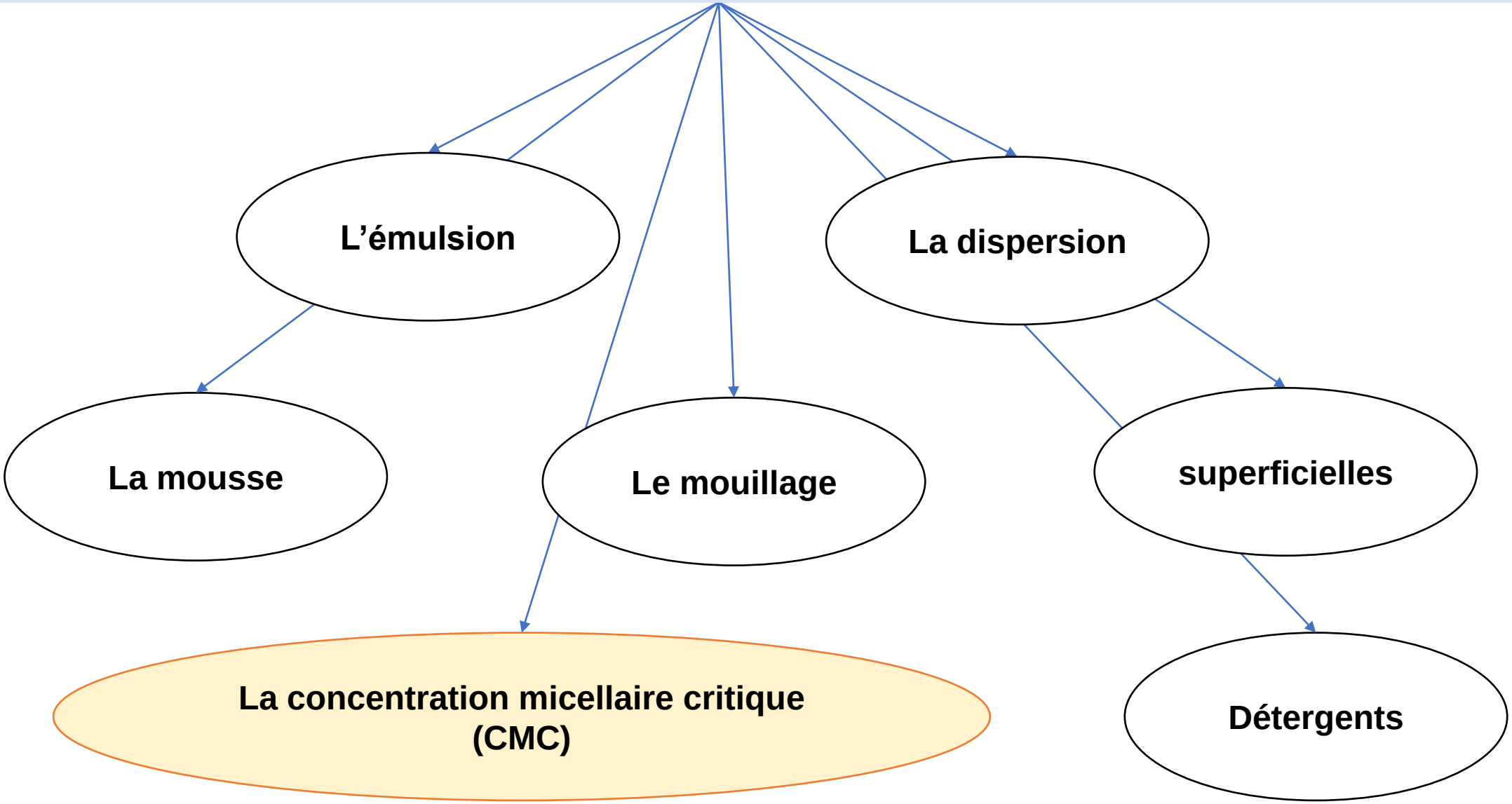
Non ionique n'est aucune charge électrique résiduelle, par exemple les alcools ethoxylés

Classification des tensioactifs en fonction de la longueur de la chaine hydrophobe

L'autre classification selon la longueur de la chaine hydrophobe. Concernant la queue hydrophobe, il est généralement admis que :

- **Les chaînes courtes, de 8 et 10 atomes de carbones (propriétés mouillantes).**
- **Les chaînes moyennes, de 12 à 16 atomes de carbone (propriétés mouillantes et détergentes).**
- **Les chaînes longues, de 18 à 22 atomes de carbone (pouvoirs émulsifiants et adoucissants).**

Propriétés du tensioactifs



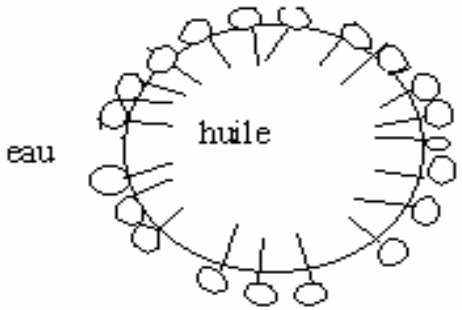
Détergents : Les détergents sont donc des produits capables d'éliminer les souillures grâce à leur aptitude à abaisser la tension superficielle du liquide dans lequel ils sont dissous

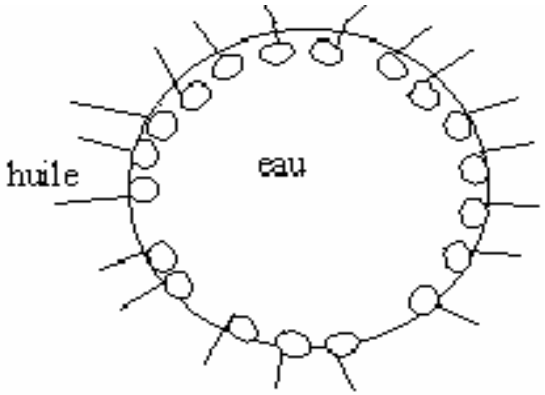
La tension Superficielles : est un phénomène physico-chimique lié aux interactions moléculaires d'un fluide. Elle résulte l'augmentation de l'énergie à sa surface qui en augmente localement la cohésion.

Le mouillage : Le pouvoir mouillant se définit par la propriété d'étalement du liquide sur un substrat donné et par son pouvoir de pénétration dans les pores du tissu.

La dispersion : Pour maintenir des particules dispersées, deux techniques sont possibles : on les enrobe d'un film colloïdal protecteur ou on utilise un électrolyte qui produit leur répulsion mutuelle.

L'émulsion : L'émulsion est définie comme étant une suspension de particules liquides au sein d'une autre phase liquide non miscible. Les molécules tensioactives stabilisent les gouttelettes d'huile dans l'eau ou d'eau dans l'huile

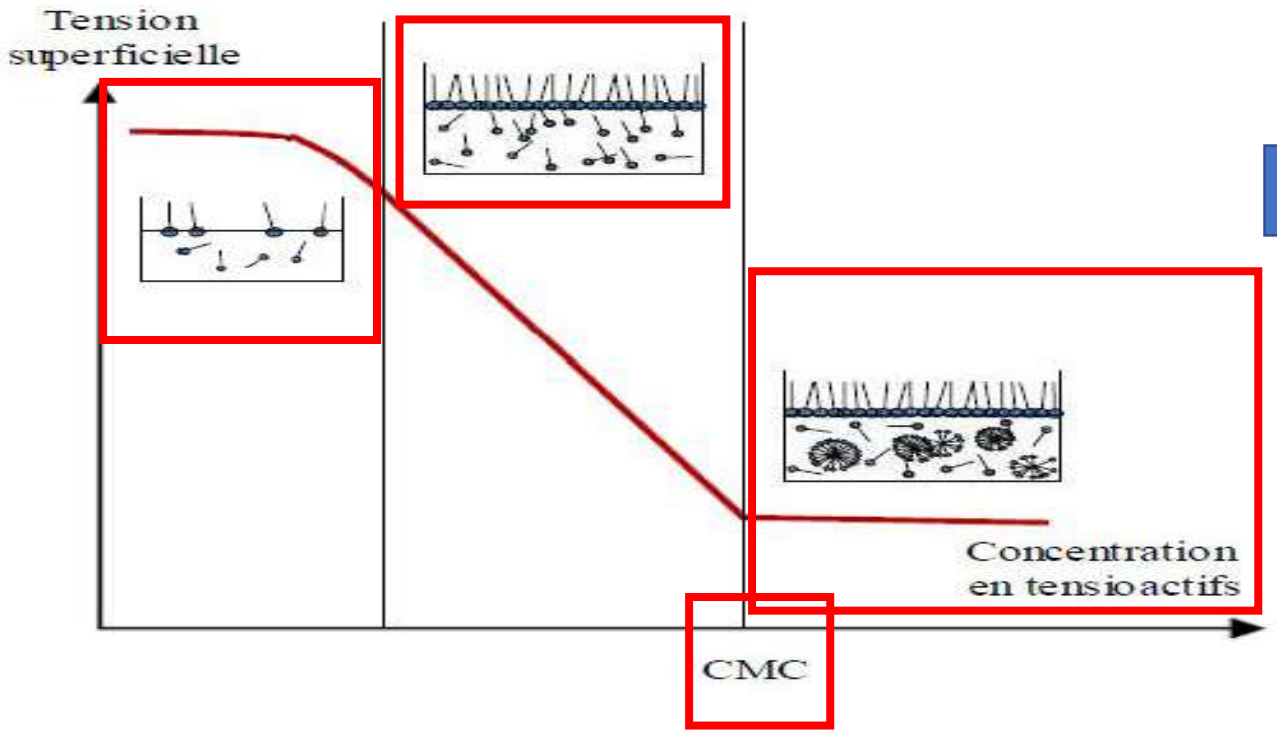




La concentration micellaire critique (CMC)

TAs = s'auto-organiser → CMC

Lorsque la concentration en tensioactif augmente, la tension superficielle diminue de manière proportionnelle jusqu'à atteindre une valeur limite: la CMC.



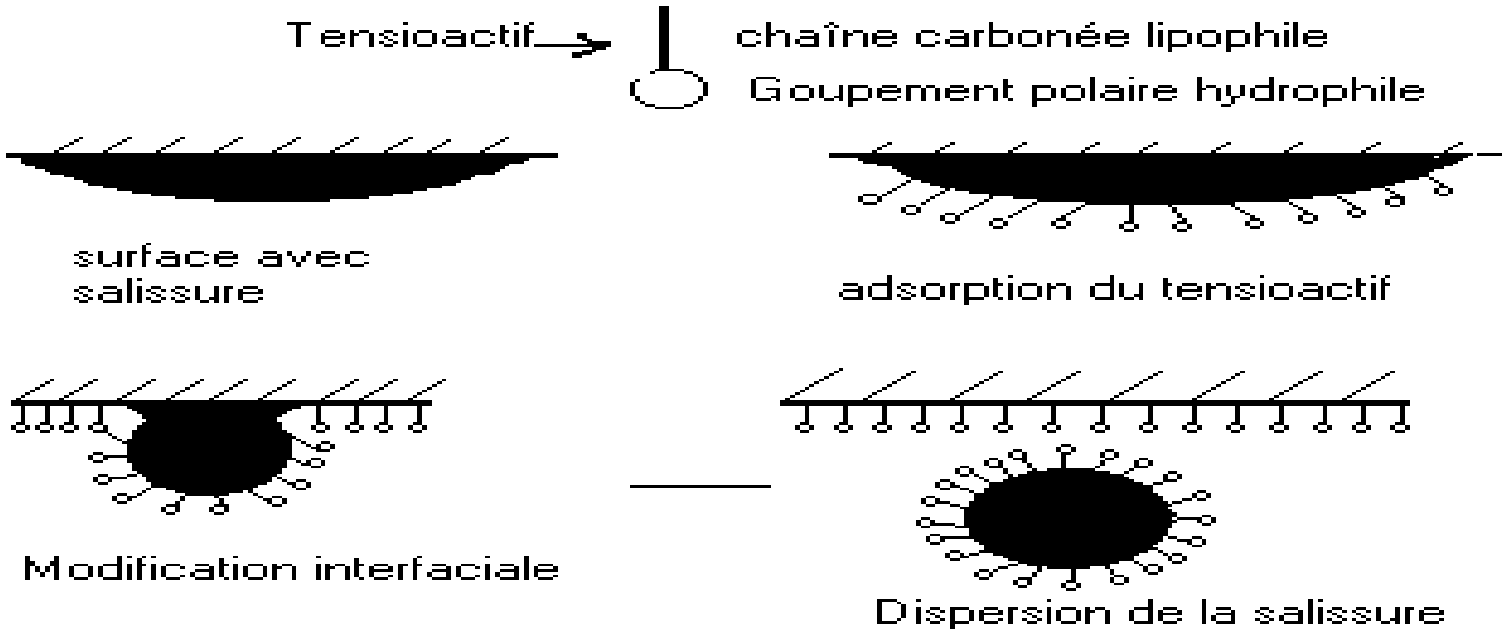
CMC

- **Effet de la longueur de la chaîne**
- **Effet de la tête polaire**
- **Effet du contre-ion**
- **Effet de la température**
- **Effet de sel**

Mode d’action des tensioactifs

Le mode d’action des tensioactifs repose sur un phénomène physique. Ils s’adsorbent à la surface de l’eau et s’y orientent, la partie hydrophile polaire est alors dirigée vers l’intérieur créant ainsi une discontinuité dans le film superficiel.

Cette dernière engendre une diminution de la tension superficielle du solvant.



Secteurs d’application des tensioactifs

- Détergence



- Cosmétique



- Les engrais



- Plastiques



- Pétrolière



Application des tensioactifs dans la remédiation environnementale

La contamination des sols par le polluant organique pose un vrai problème environnemental, cette contamination est due au :

- **Accidents de transport d'hydrocarbures**
- **Incendies ou d'explosions d'unités de production industrie**
- **Fuites industrielles**
- **Infiltrations de déchets industriels et urbains.**

La solubilisation des polluants organiques par les tensioactifs est un facteur clé dans la remédiation des sols contaminés par des hydrocarbures. Les tensioactifs peuvent augmenter la solubilité des hydrocarbures en les introduits dans les noyaux hydrophobes des micelles tensioactifs

Partie Modélisation

Objectif de notre travail est la modélisation de trois types de tensioactifs (anionique, non ionique et cationique) afin de trouver un modèle mathématique pour la détermination de log CMC dans une grande marge de concentration pour chaque type, puis en englobe les trois types au même temps.

Anionique

Cationique

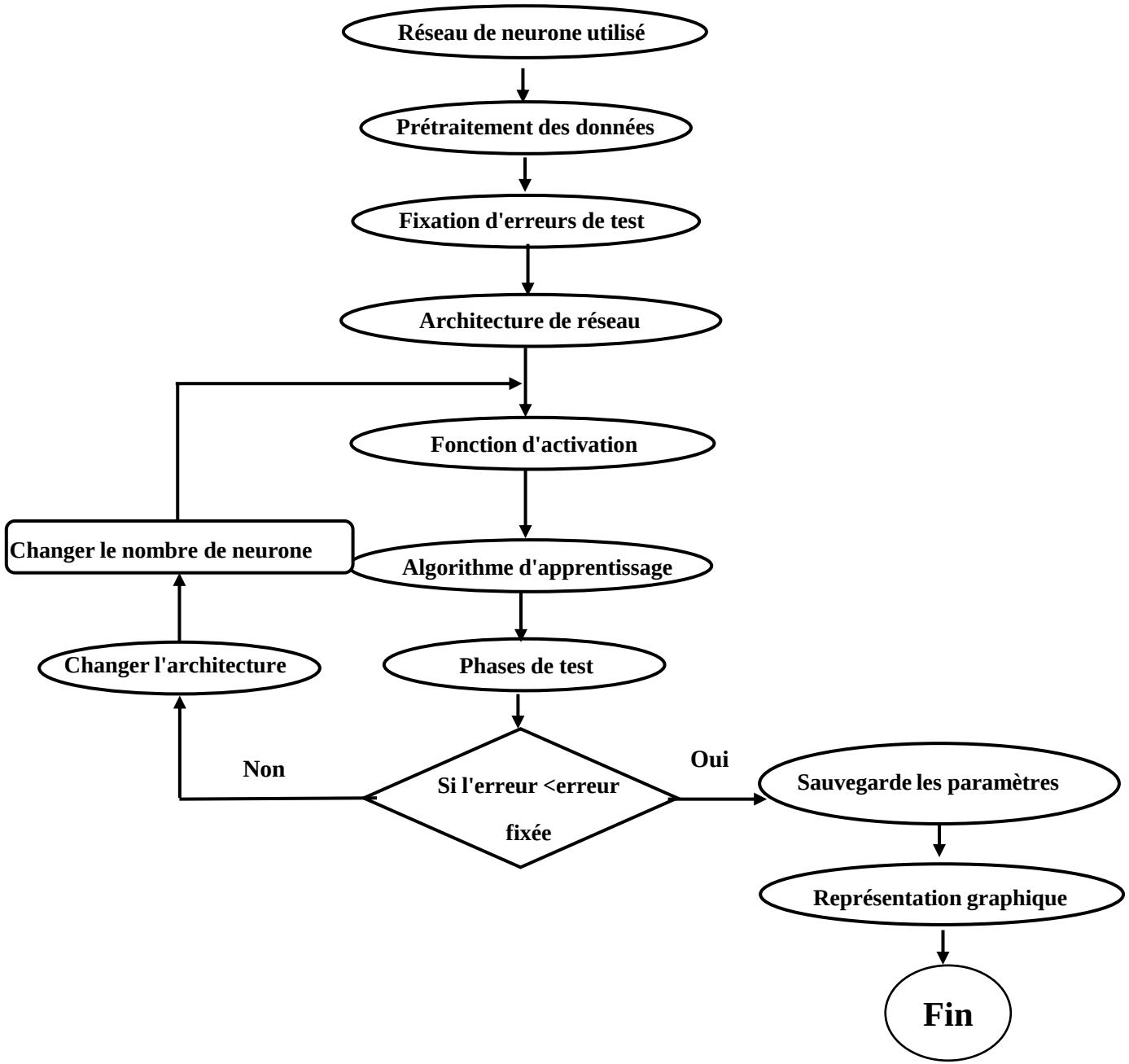
Non-ionique

globale

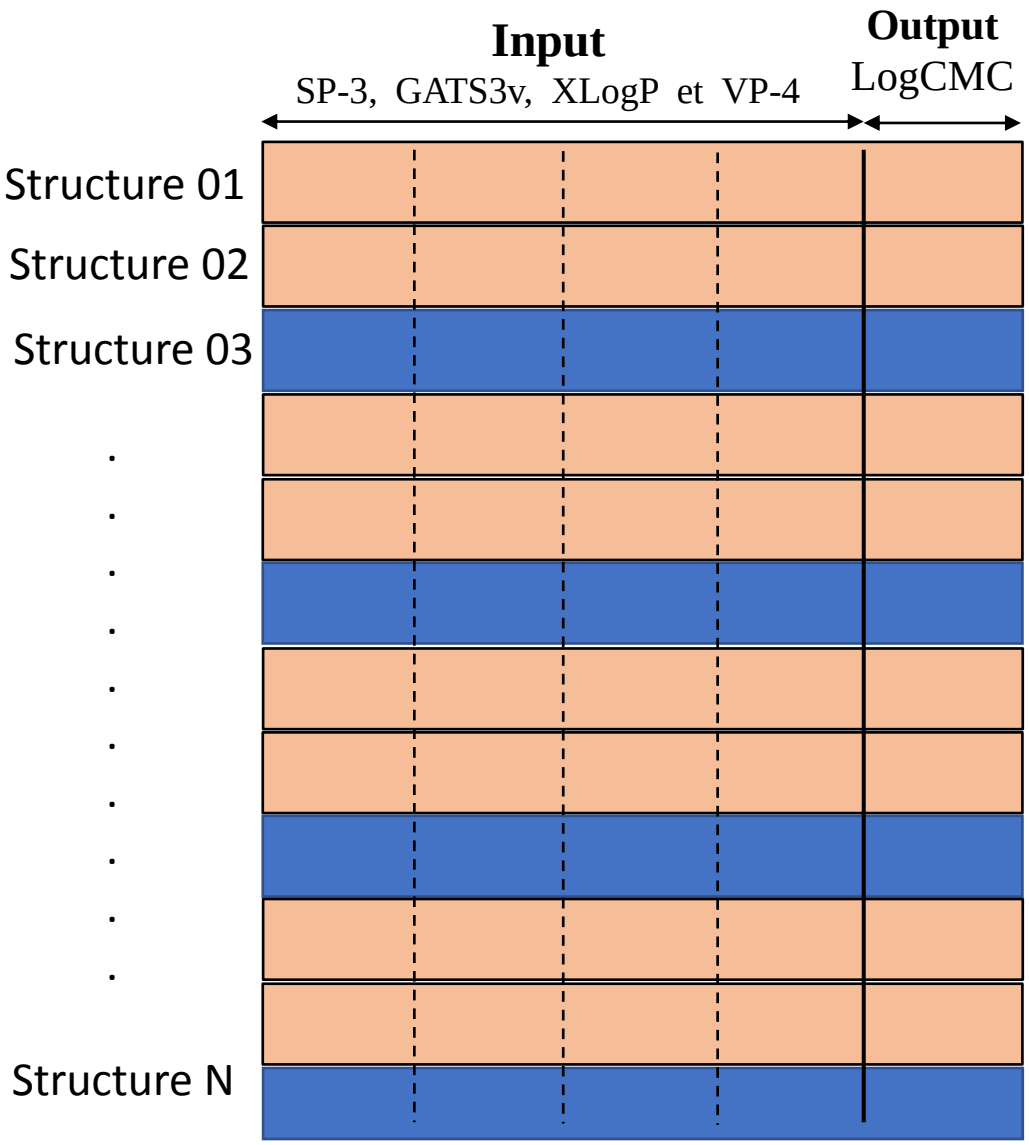
	A	B	C
1	C2SP3	minsCH3	SpAD_D
2	16	1,486892	830,739
3	20	1,480732	1537,1
4	24	1,478313	2462,11
5	16	1,748953	853,079
6	16	2,274828	629,854
7	20	1,742792	1561,06
8	24	1,740373	2487,10
9	18	1,433293	530,698
10	16	1,884021	1096,14
11	16	1,929539	1119,98
12	11	1,49479	471,237
13	17	1,480598	1409,54
14	16	2,276953	749,408
15	14	1,484514	877,91
16	20	1,478648	2066,35
17	13	1,502415	441,991
18	10	2,260414	621,003
19	18	1,482986	1156,77
20	15	1,504894	519,230
21	15	2,181694	426,173
22	15	1,490068	687,906

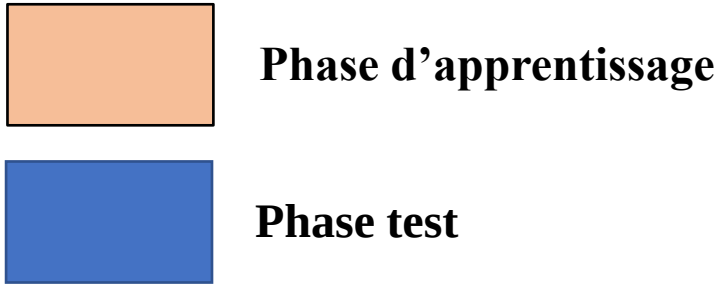
(A) Validation interne ou validation croisée

(B) Validation externe en utilisant la base d



Division de la phase d'apprentissage et de test





Modélisation par RNA de Log CMC

Codes	Base des données	Entrés (04)	Sortie
RNA1	Non-ionique	SP-3, GATS3v, XLogP , VP-4 et LogCMC	LogCMC
RNA2	Cationique	SpMAD_D, VR3_Dzs, ALogp2 et MATS3c	LogCMC
RNA3	Anionique	MATS8p, MATS7v, ATSC3e et nHBAcc_Lipinski	LogCMC
RNA4	Global	C2SP3, minsCH3, SpAD_DzZ et AATSC8p	LogCMC

Normalisation des données

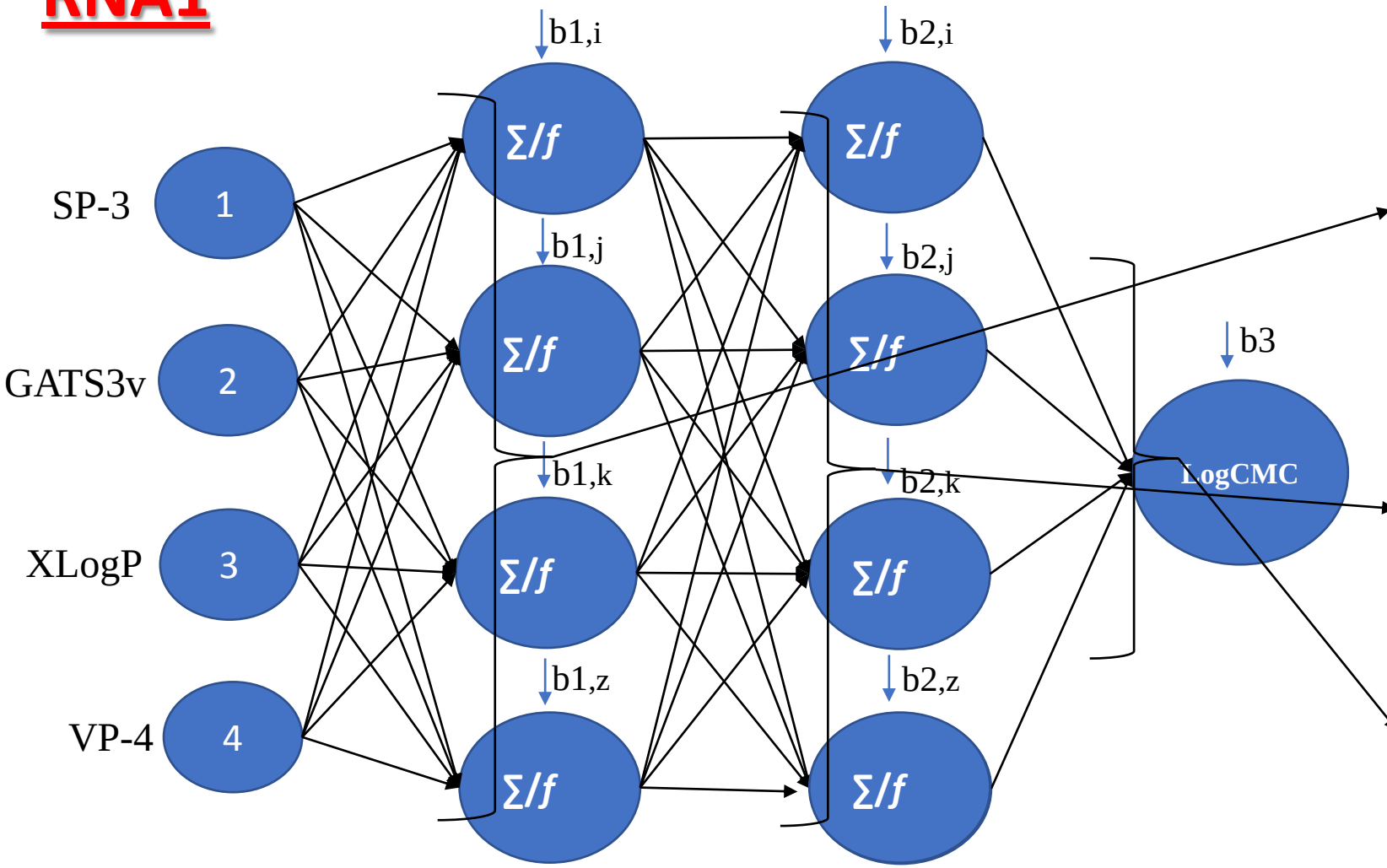
Les valeurs numériques ont été normalisées afin d'améliorer l'optimisation. La normalisation se fait par colonne (premnmx : MATLAB), chaque paramètre des entrées et de sortie (X1) est normalisé (XiN) par l'expression suivante :

$$X_{iN} = 2\frac{(X - X\,min)}{(X\,max - X\,min)} - 1$$

L'apprentissage supervisé est retenue dans cette étude, il consiste à fournir au réseau des jeux de données d'entrées pour lesquels on connaît les réponses souhaitées (LogCMC).

Résultats

RNA1



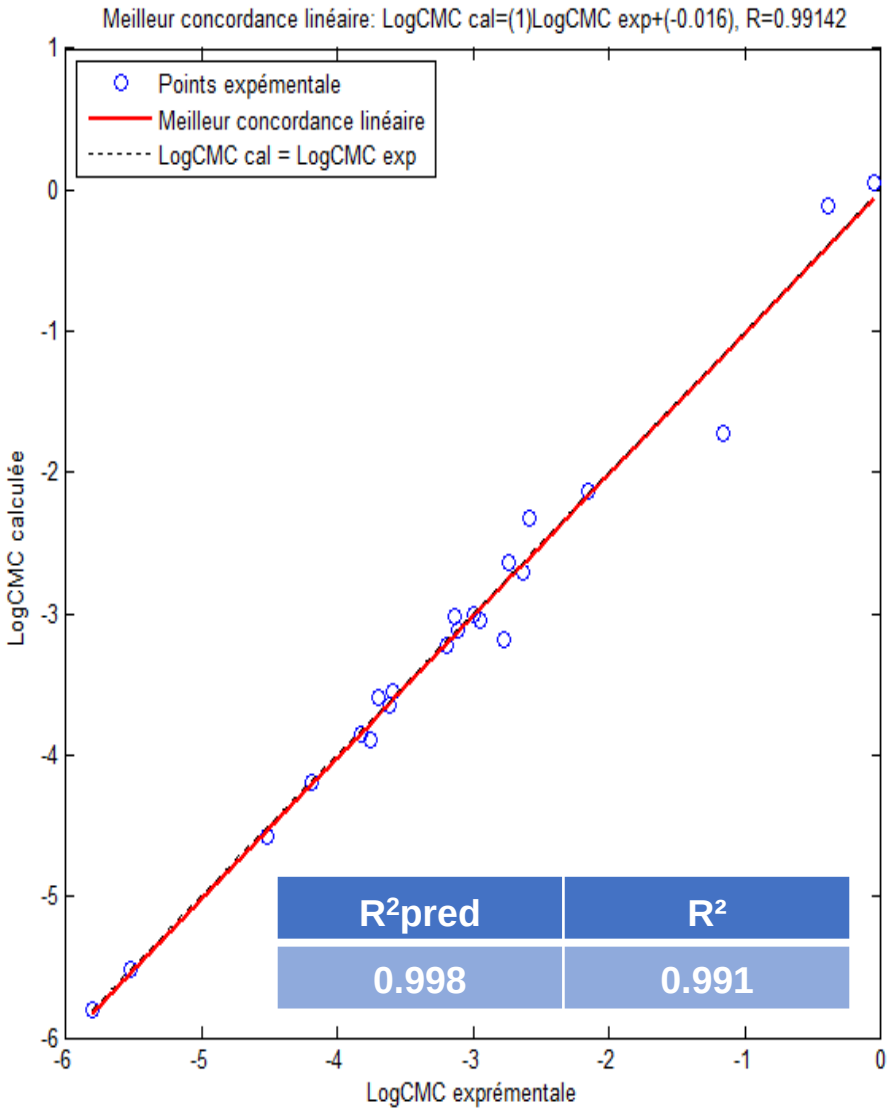
W1				b1
-1,053	0,356	0,591	-0,603	2,493
1,050	2,739	-0,047	0,132	0,233
-0,560	0,302	0,346	1,010	-0,061
-1,496	1,808	-1,287	-0,887	3,573

W2				b2
-0,840	-1,325	0,221	-2,312	1,208
-0,049	0,211	-1,122	-0,627	0,448
-1,829	0,115	-0,766	-0,916	-1,148
1,078	-1,604	-0,093	-0,514	1,893

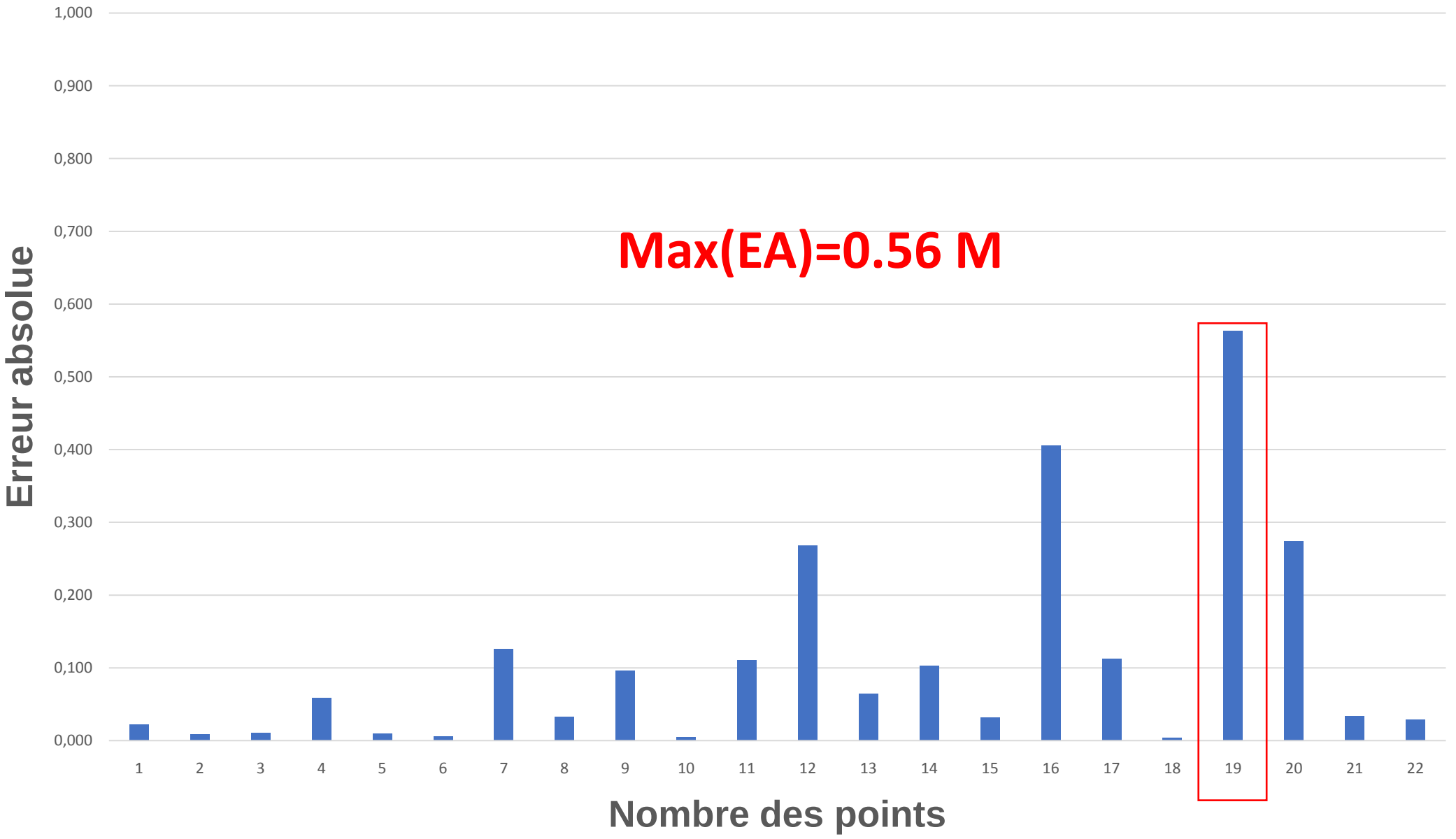
W3				b3
0,338	1,963	-0,553	0,509	-0,617

Test du RNA1

Nar	structure	LogCMC Expérimentales	LogCMC Calculés
03	C8H17(OC2H4)6OH	-2.1250	-2.1474
06	C10H21(OC2H4)8OH	-3.0000	-2.9917
09	C12H25(OC2H4)3OH	-4.1940	-4.1839
12	C12H25(OC2H4)7OH	-4.5690	-4.5103
15	C14H29(OC2H4)6OH	-5.7960	-5.8052
18	C16H33(OC2H4)6OH	-5.5090	-5.5152
21	C16H33(OC2H4)12OH	-3.8860	-3.7598
24	p-t-C8H17C6H4O(C2H4O)2H	-3.5530	-3.5861
27	p-t-C8H17C6H4O(C2H4O)5H	-2.6380	-2.7340
30	p-t-C8H17C6H4O(C2H4O)8H	-3.1190	-3.1141
33	C8H17CHOHCH2OH	-3.5850	-3.6952
36	C12H25CHOHCH2CH2OH	-2.3190	-2.5868
39	C13H27(OC2H4)5OH	-2.6990	-2.6345
42	C11H23CON(C2H4OH)2	-3.0490	-2.9465
45	C6F13C2H4SC2H4(OC2H4)5OH	-3.6510	-3.6197
48	n-Decyl-b-D-glucoside	-3.1800	-2.7745
51	Dodecyl-b-D-maltoside	-3.0180	-3.1303
53	(C4H9)2CHCH2(OC2H4)9OH	0.0490	-0.0450
56	C12H25(OC2H4)12OH	-1.7200	-1.1570
59	C11H23CON(CH3)CH2CHOHCH2OH	-0.1100	-0.3837
62	C12H25CON(CH3)CH2(CHOH)4CH2OH	-3.2220	-3.1883
65	C12H25N(CH3)CO(CHOH)4CH2OH	-3.8540	-3.8248



Présentation de l'erreur absolue de LogCMC



Test du RNA2

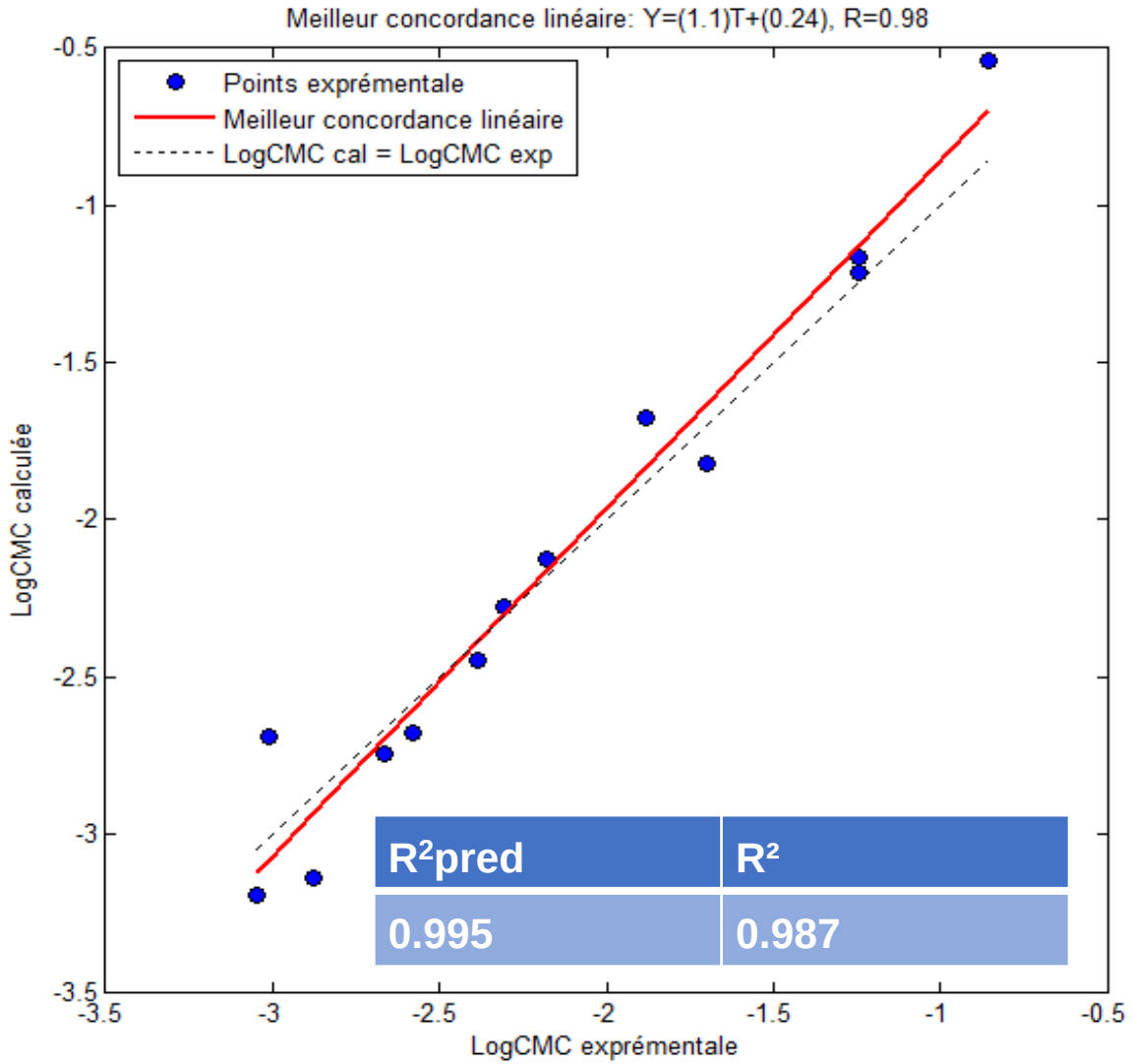
RNA2

N	Structure	logCMC expremental	LogCMC calculé
01	C8H17N⁺(CH3)3Br–	-0.854	-0,854
02	C10H21N⁺ (CH3)3Br–	-1.167	-1,242
03	C10H21N⁺ (CH3)3Cl–	-1.167	-1,242
04	C12H25N⁺ (CH3)3Br–	-1.796	-1,698
05	C12H25N⁺ (CH3)3Cl–	-1.699	-1,698
06	C14H29N⁺ (CH3)3Br–	-2.444	-2,383
07	C14H29N⁺ (CH3)3Cl–	-2.347	-2,383
08	C16H33N⁺ (CH3)3Br–	-3.009	-2,666
09	C10H21Pyr⁺Br–	-1.357	-1,356
10	C11H23Pyr⁺Br–	-1.678	-1,883
11	C12H25Pyr⁺Br–	-1.959	-2,117
12	C12H25Pyr⁺Cl–	-1.77	-2,117
13	C12H25Pyr⁺I–	-2.276	-2,117
14	C13H27Pyr⁺Br–	-2.276	-2,304
15	C14H29Pyr⁺Br–	-2.569	-2,568
16	C14H29Pyr⁺Cl–	-2.456	-2,568
17	C15H31Pyr⁺Br–	-2.886	-2,875
18	C16H33Pyr⁺Br–	-3.194	-3,045
19	C16H33Pyr⁺Cl⁻	-3.046	-3,045
20	C18H37Pyr⁺Cl⁻	-3.62	-2,463
21	C12H25N⁺ (C2H5)(CH3)2Br–	-1.854	-1,853
22	C12H25N⁺ (C4H9)(CH3)2Br–	-2.125	-2,181
23	C12H25N⁺ (C6H13)(CH3)2Br–	-2.509	-2,508
24	C12H25N⁺ (C8H17)(CH3)3Br–	-2.959	-2,781
25	C14H29N⁺ (C2H5)3Br–	-2.509	-2,508
26	C14H29N⁺ (C3H7)3Br–	-2.678	-2,578
27	C14H29N⁺ (C4H9)3Br–	-2.921	-2,920

N	Structure	LogCMC exprémental	LogCMC calcule
29	C12H25N⁺ (CH2C6H5)(CH3)2Cl–	-2.056	-2,061
30	C14H29N⁺ (CH2C6H5)(CH3)2Cl–	-2.69	-3,013
31	C12H25NH2CH2CH2OH_Cl–	-1.347	-1,346
32	C12H25N⁺H(CH2CH2OH)2Cl–	-1.444	-1,507
33	C12H25N⁺H(CH2CH2OH)3Cl–	-1.602	-1,601
34	(C10H21)2N⁺ (CH3)2Br–	-2.745	-2,664
35	(C12H25)2N⁺ (CH3)2Br–	-2.77	-2,769
36	1-tetradecylpyridinium bromide	-2.558	-2,568
37	1-pentadecylpyridinium bromide	-2.866	-2,875
38	1-hexadecylpyridinium bromide	-3.14	-2,875
39	4-dodecyl-1-ethylpyridinium iodide	-2.656	-2,655
40	N,N,N-trimethylhexadecan-1-aminium bromide	-3	-2,666
41	N,N,N-trimethyltetradecan-1-aminium bromide	-2.42	-2,383
42	N,N,N-trimethyldodecan-1-aminium bromide	-1.818	-1,698
43	N-benzyl-N,N-dimethylhexadecan-1-aminium chloride	-3.31	-3,309
44	N-benzyl-N,N-dimethyltetradecan-1-aminium chloride	-2.701	-3,013
45	N-benzyl-N,N-dimethyldodecan-1-aminium chloride	-2.066	-2,061
46	N,N,N-trimethyloctan-1-aminium bromide	-0.538	-0,854
47	N,N,N-trimethyldecan-1-aminium bromide	-1.319	-1,242
48	N-ethyl-N,N-dimethyldodecan-1-aminium bromide	-1.83	-1,853
49	N-ethyl-N,N-dimethyltetradecan-1-aminium bromide	-2.514	-2,513
50	N,N,N-trimethyldecan-1-aminium chloride	-1.215	-1,242
51	N,N,N-triethyltetradecan-1-aminium bromide	-2.509	-2,508
52	C10H21N⁺ (CH2C6H5)(CH3)2Cl–	-1.409	-1,213

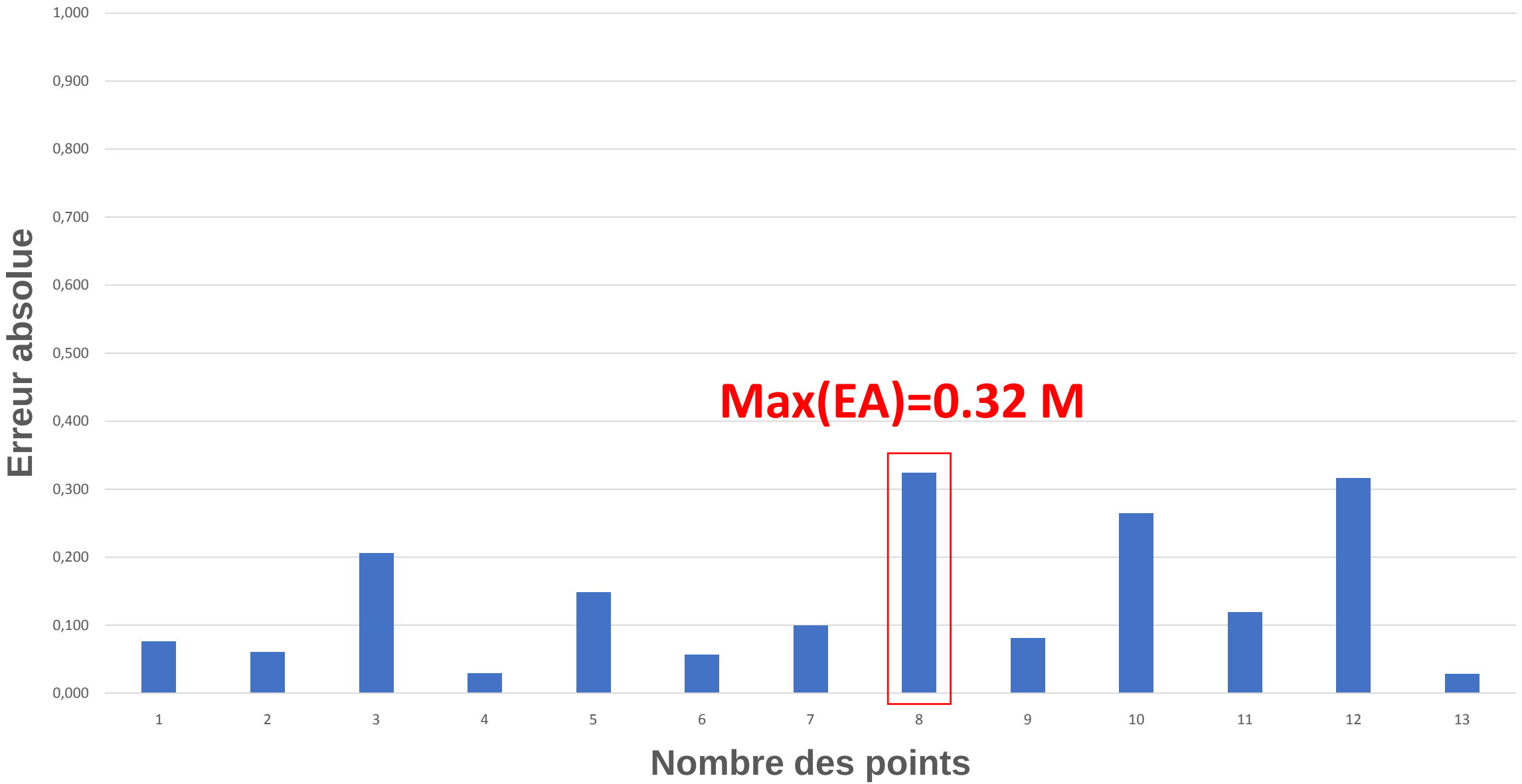
Test du RNA2

RNA2



R^2pred	R^2
0.995	0.987

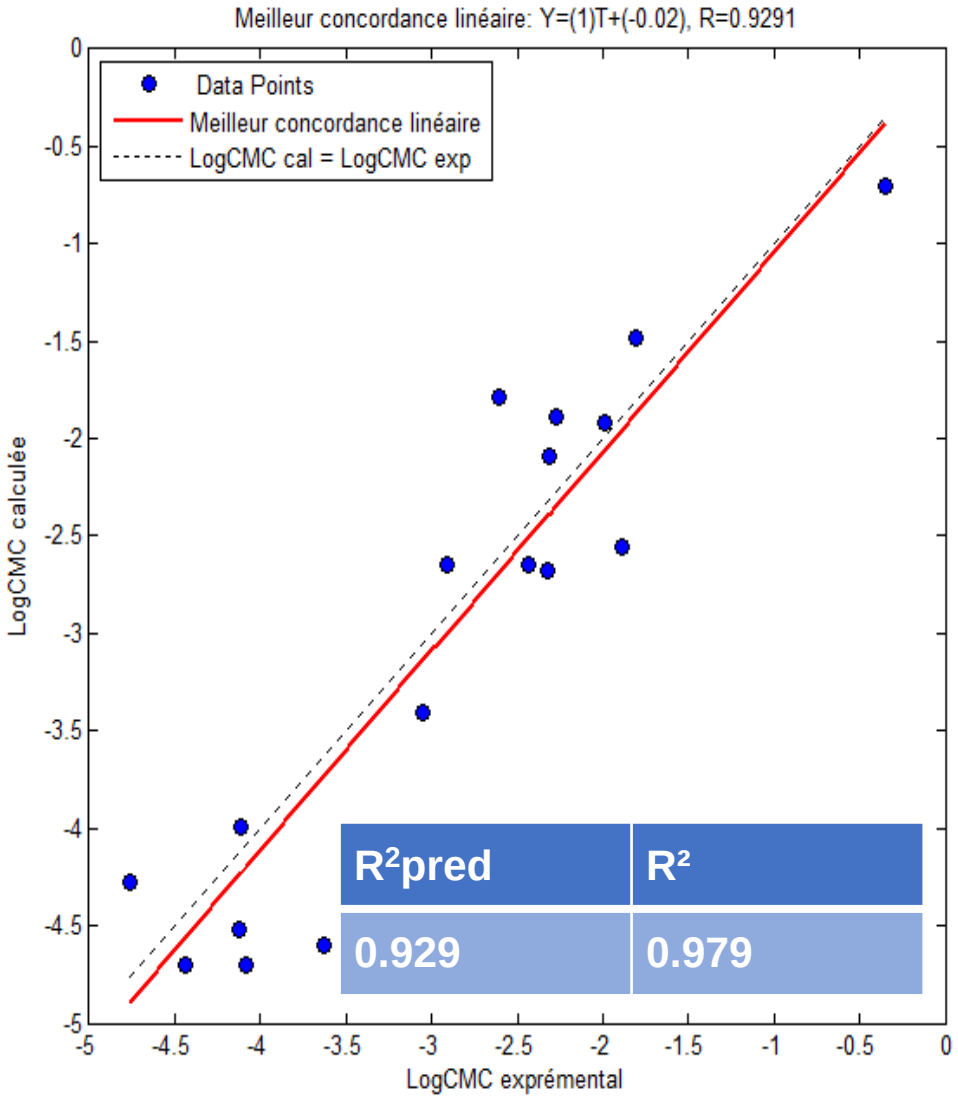
Présentation de l'erreur absolue de LogCMC



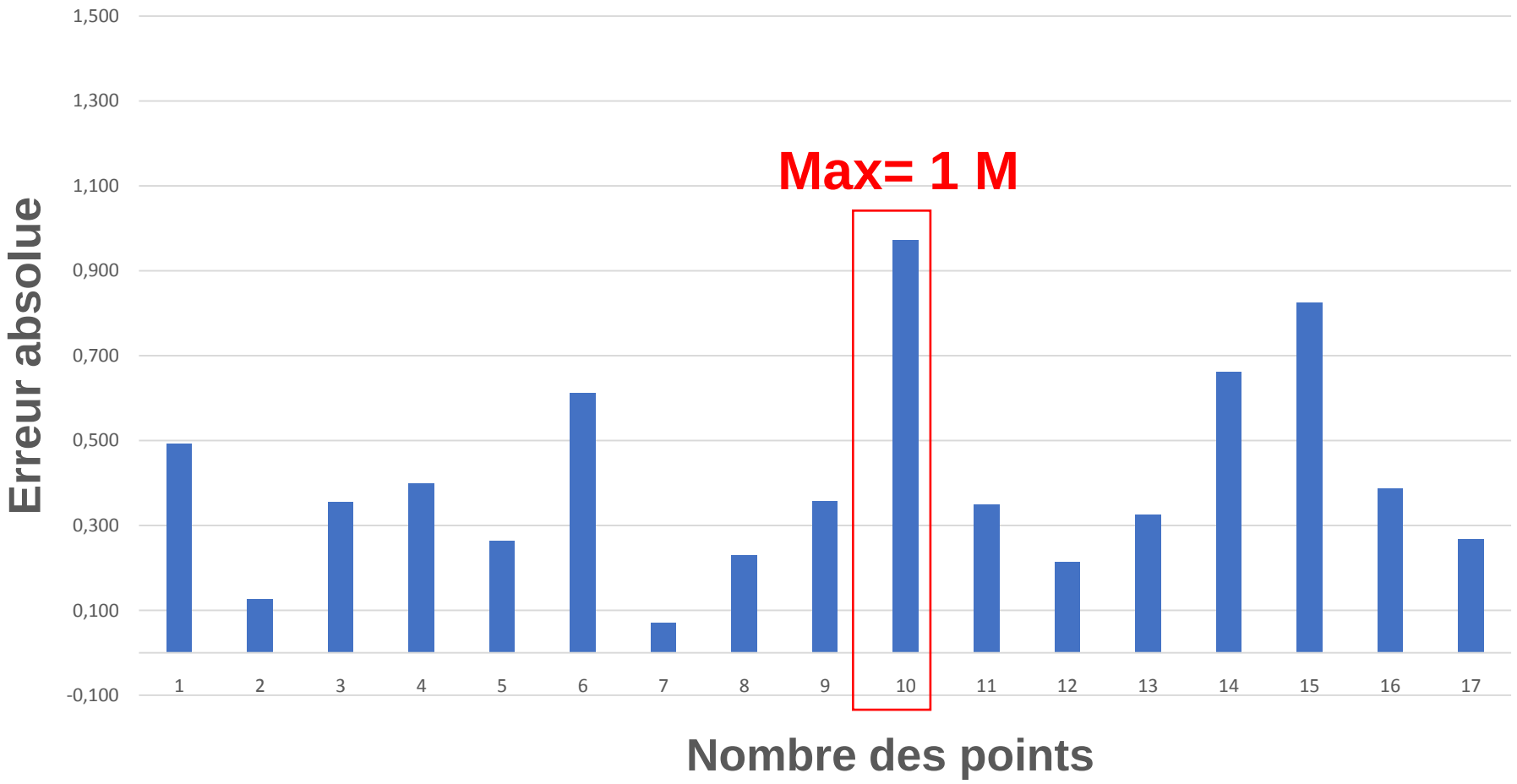
Test du RNA3

RNA3

N	Structure	LogCMC Expremental	LogCMC calculé
01	$C_{12}H_{25}O(CH_2CHCH_3O)_{13}CH_2CHCH_3SO_4$-Na+	-4.2750	-4.7666
02	$C_{12}H_{25}O(CH_2CHCH_3O)_{13}CH_2CHCH_3OCH_2CO_2$-Na+	-3.9910	-4.1169
03	$C_{12}H_{25}O(CH_2CHCH_3O)_5CH_2CHCH_3OCH_2OCH_2CO_2$-Na+	-3.4030	-3.0470
04	$C_{12}H_{25}O(CH_2CHCH_3O)_5CH_2CHCH_3OCH_2OCH_2CHCO_2$-Na+	-4.5220	-4.1227
05	sodium nonylphenoxy polypropyleneoxide sulfates n=9	-4.6980	-4.4339
06	sodium nonylphenoxy polypropyleneoxide sulfates n=13	-4.6980	-4.0867
07	C12,13–(PO)3–SO4Na	-1.9210	-1.9914
08	C14,15–(PO)3–SO4Na	-2.0860	-2.3156
09	$C_{15}H_{31}CONHCH[CH(CH_3)2]COO$-Na+	-2.6780	-2.3199
10	$C_{12}H_{25}SO_3$-Li+	-4.6020	-3.6294
11	$C_{12}H_{25}SO_4$-Na+	-0.6990	-0.3499
12	$C_{12}H_{25}SO_4$-N(C3H7)4+	-2.6480	-2.4334
13	$C_{12}H_{25}OC_2H_4SO_4$-Na+	-1.4810	-1.8067
14	$C_6H_{13}OOCCH_2SO_3$-Na+	-2.5530	-1.8906
15	$C_{10}H_{21}OOCCH_2SO_3$-Na+	-1.7830	-2.6082
16	$C_4H_9CH(C_2H_5)CH_2OOCCH_2CH(SO_3$-Na+)	-1.8830	-2.2696
17	$COOCH_2CH(C_2H_5)C_4H_9$	-2.6470	-2.9142

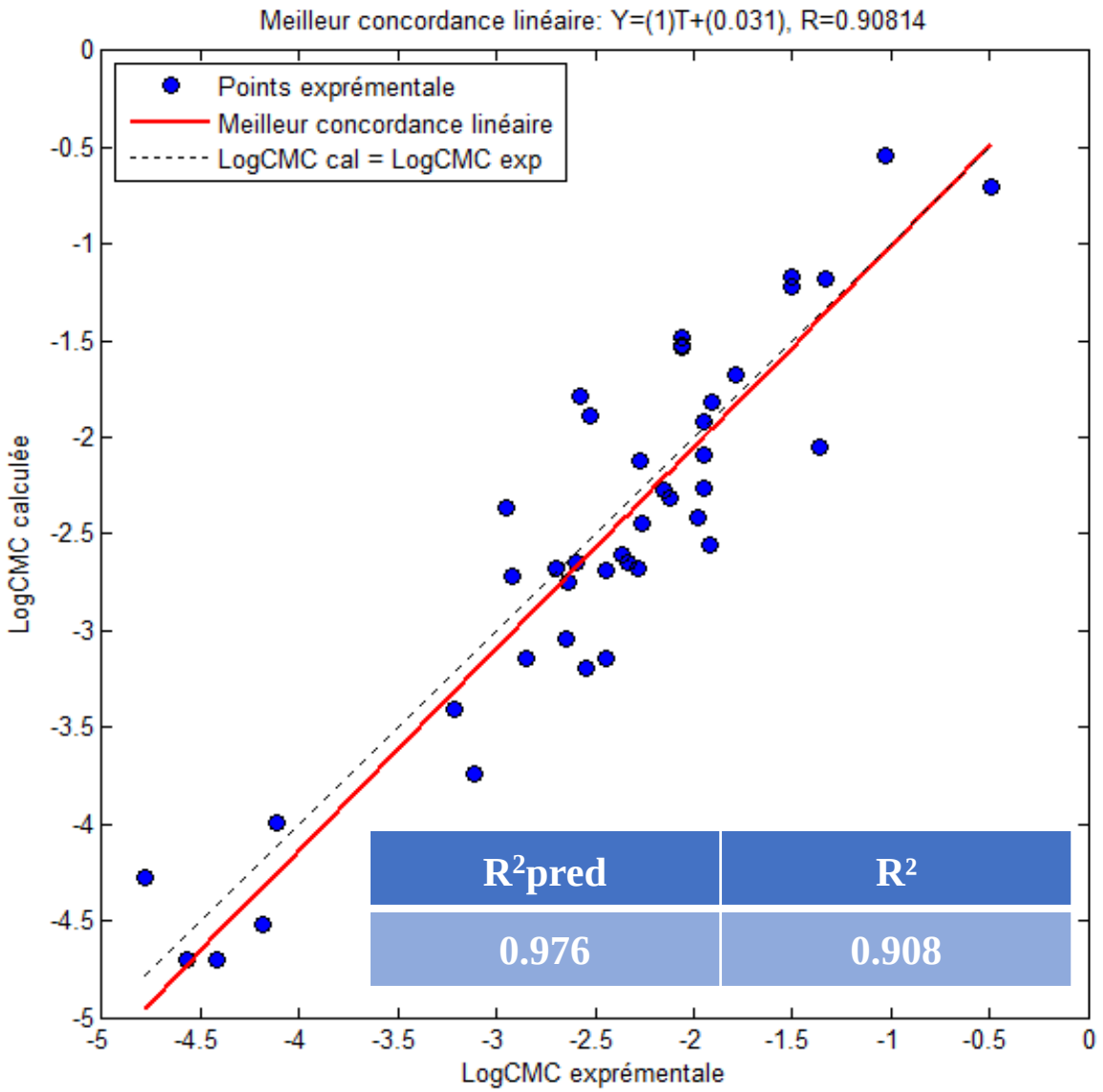


Présentation de l'erreur absolue de LogCMC



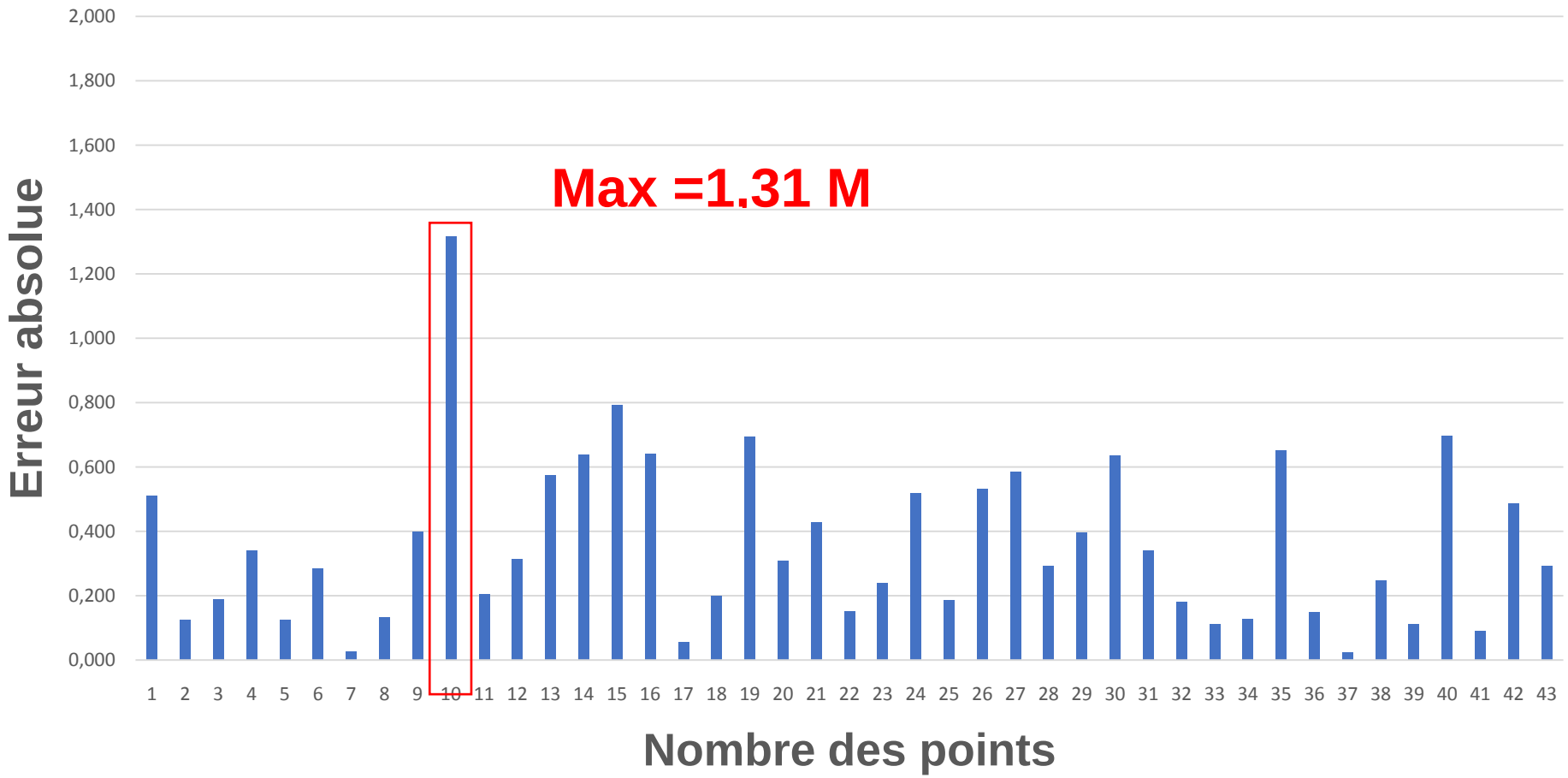
Test du RNA4

RNA4

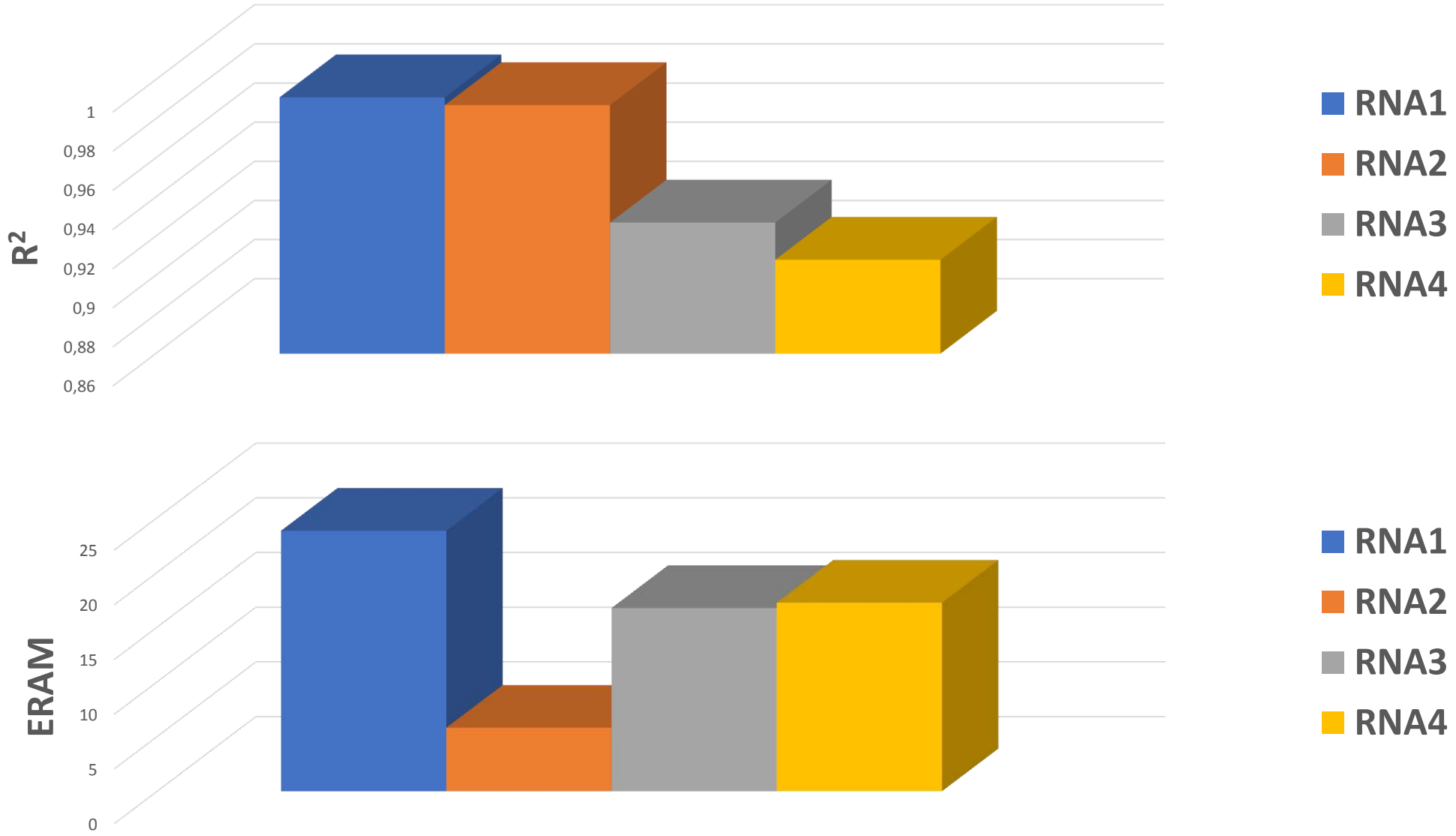


R^2pred	R^2
0.976	0.908

Présentation de l'erreur absolue de LogCMC



Comparaison des erreur



Conclusion générale

L'étude réalisée s'inscrit dans le cadre de la compréhension des mécanismes qui influe sur la concentration micellaire critique des tensioactifs par le changement structural d'une molécule.
Le nombre des entrées du modèle neuronale a été calculé par le logiciel Padel et optimisée par la méthode des algorithmes génétiques.

➢**Le log CMC de 51 tensioactifs non-ioniques avec un modèle neuronal qui a une architecture de {4-4-4-1}, ce dernier été trouvé avec un coefficient de corrélation égale R=0.991, une Erreur Relative Absolue Moyenne ERAM=23.815 et R^2_{pred}=0.998. Ces résultats vérifient la capacité prédictive externe de notre modèleen confirmant la robustesse du modèle établi.**

Conclusion générale

- **Le second modèle comporte une architecture de {4-7-7-1}, a été optimisé afin de calculer les agents de surface de 88 structures anioniques avec un R=0.987, ARDM=5.806 et R^2_{pred}=0.995.Ces résultats montrent la robustesse du modèle établi.**

- **Le troisième réseau de neurones à une structure de {4-5-5-1}, ce dernier montre son aptitude à prédire logCMC de 71 composés anioniques avec un R=0.927, ARDM=16.744 et R^2_{pred}=0.979.**

- **Le dernier, c'est un réseau global d'architecture {4-6-6-1} englobant un mélange de 171 tensioactifs de différents types (anionique, cationique, non-ionique), le réseau a été trouvé avec un R=0.908, ARDM=17.235 et R^2_{pred}=0.976.**

En conséquence les réseaux de neurones peuvent être utilisés comme des estimateurs robustes et de concurrence pour la prédiction des paramètres de la CMC.

Merci pour votre attention

Remerciements

Au terme de ce Mémoire nous tenons à remercier tout naturellement en premier lieu ***Dieu le tout Puissant*** *qui nous a donné la force, le courage et la patience de bien mener ce travail.*

Ce travail a été réalisé sous la direction de Monsieur ***S. RAHAL****, Maître de Conférences "B" à l'Universitaire Yahia Farès de Médéa, qu'il trouve ainsi l'expression de notre profonde reconnaissance pour sa compétence, ses encouragements et ses précieux conseils durant le déroulement de ce travail. Si ce travail, a pu faire l'objet de communications nationales et international c'est bien grâce à ses directives et son soutien.*

Nous remercions Monsieur N. HADIDI, qui a bien voulu assurer la Présidence de notre jury ainsi que Monsieur Y KERCHICH pour avoir accepté de juger ce modeste travail.

Nos sentiments vont également à toutes les personnes qui ont contribué de prés ou de loin à la réalisation de ce travail.

Merci du fond de mon cœur à mes meilleures amis DAHLOUK Zakaria,

BOUKACHKACH Adel et KHALIL CHERFI Abdenour.

Merci

Dédicace

Je dédie ce travail à :

Mes parents

Mes frères et mes sœurs

Toute ma famille

Men binôme Boualem

Toute mes amies

Tous mes collègues

KOURDALI Mohamed El Amine

Dédicace

Je dédie ce travail à :

Mes parents

Mes frères et ma sœur et ma femme

Toute ma famille

Men binôme Amine

Toute mes amies

Tous mes collègues

LAROUCI Boualem

Sommaire

Lite des tableaux

Liste des figures

Nomenclature

Introduction générale **1**

I. Généralité sur la tensioactif **3**

I.1. Introduction **3**

I.2. Définition **3**

I.3. Structuration des tensioactifs en solution aqueuse **5**

I.4. Classification **6**

I.4.1. Les tensioactifs anioniques **7**

I.4.2. Les tensioactifs non ioniques **8**

I.4.3. Les tensioactifs cationiques **9**

I.4.4. Les tensioactifs zwitterioniques (amphotères) **9**

I.4.5. Classification des tensioactifs en fonction de la longueur de la chaine hydrophobe **10**

I.5. Propriétés **10**

I.5.1. Détergents **10**

I.5.2. Propriétés superficielles **13**

I.5.3. La concentration micellaire critique (CMC) **13**

I.5.4. Le mouillage **14**

I.5.5. La mousse **15**

1.1.1. La dispersion **15**

I.5.6. L'émulsion **16**

I.6. Mode d'action **16**

I.7. Secteurs d'application des tensioactifs **17**

I.8. Application des tensioactifs dans la remédiation environnementale **19**

I.9. Conclusion **23**

II.1. Introduction **24**

II.2. Historique **25**

II.3. Généralités sur la modélisation QSPR **26**

II.3. Principe des méthodes QSPR **27**

II.4. Méthodologie générale d'une étude QSPR **28**

II.5. Descripteurs moléculaires **29**

II.6. Modèles statistiques **29**

II.7. Descripteurs moléculaires **32**

II.7.1. Les descripteurs moléculaires théoriques **32**

II.7.1.1. Les descripteurs 1D **32**

II.7.1.2. Les descripteurs 2D **32**

II.7.1.3. Les descripteurs 3D **33**

II.8. Méthodes d'analyse des données **33**

II.9. Interprétation et validation d'un modèle QSPR **33**

II.9.1. Validation interne **34**

II.9.2. Y-Randomization **34**

II.9.3. Validation externe **34**

II.10. Domaine d'applicabilité **35**

II.11. Applications des méthodes QSPR **35**

II.11.1 Propriétés physico-chimiques **35**

II.11.3. Autres propriétés/activités **35**

II.12. Régression linéaire multiple (MLR) **36**

II.12.1. Description de la méthode **36**

II.13. La régression des moindres carrés partiels (PLS) **40**

II.13.1. L'Algorithme de la régression des moindres carrés partiels (PLS) **40**

II.13.2. Caractéristiques **41**

II.13.3. La régression des moindres carrés partiels (PLS) **41**

II.14. Réseaux de neurones artificiels (RNA) **43**

II.14.1. Structure des réseaux de neurones artificiels **43**

II.14.2. Fonction d'activation **44**

II.14.2. La fonction non linéaire sigmoïde **44**

II.14.3. Apprentissage d'un réseau de neurone **46**

II.14.3.1. Apprentissage supervisé **46**

II.14.3.2. L'algorithme d'apprentissage **47**

II14.3.3. L'algorithme de rétro-propagation **48**

II.15. Conclusion **50**

III. Modélisation de la CMC des tensioactifs **51**

III.1. Logiciel utilisé pour la génération de descripteurs **51**

III.2. Développement d'un modèle **51**

III.2.1. Analyse statistique effectuée **51**

III.2.2. Validation de modèle **51**

III.2.3. Mesure de validation interne et externe **52**

III.2.3.1. Q^2 (ou Q^2_{Int}) **52**

III.2.3.2. Q^2_{ext} ou R^2_{pred} **52**

III.2.3.3. Paramètres de r^2_m **53**

III.2.3.4. La randomisation des données **54**

III.2.4. Domaine d'application (DA) **55**

III.2.5. Méthodologie de calcul par RNA **56**

III.2.6. Préparation de la base de donnée **57**

III.2.7. Normalisation des données **58**

III.2.8. Organigramme de la méthodologie de calcul de RNSO **59**

Résultat et discussion **61**

III.3. Modélisation des structures Non-ionique **61**

III.3.1. Base des donnes **61**

III.3.2. Conception de la topologie de RNSO **63**

III.3.2.1. Choix des entrées et des sorties **63**

III.3.2. Expression de calcul **64**

III.3.3. Modélisation de LogCMC **66**

III.4. Modélisation des structures Cationique **69**

III.4.1. Choix des entrées et des sorties **69**

III.4.2. Expression de calcul **69**

III.4.3. Modélisation de LogCMC **71**

III.5. Modélisation des structures Anionique **74**

III.5.1. Choix des entrées et des sorties **74**

III.5.2. Expression de calcul **74**

III.5.3. Modélisation de LogCMC **76**

III.6.2. Expression de calcul **79**

III.6.3. Modélisation de LogCMC **81**

Liste des figures

- **Figure I-1** : Schéma d'un tensioactif 3
- **Figure I.2** Exemple de tensioactif anionique : le dodécyl sulfate benzène de sodium (SDBS) 8
- **Figure I.3.** Exemple de polyoxyéthylène POE (noté R12POE6) 8
- **Figure I.4** Exemples de tensioactifs cationiques : le HTAB et le HPCl 9
- **Figure I.5** Exemples de tensioactifs zwitterioniques : la sulfobétaïne et la carboxybétaïne 10
- **Figure I.6 :** la forme des micelles dans liquide 12
- **Figure I.7 :** Comportement des tensioactifs en milieu aqueux– phénomène de micellisation 13
- **Figure I.8** : Courbe de la tension de surface en fonction de la concentration de tensioactifs en solution 14
- **Figure I.9 :** structure de la mousse dans le milieux air/eau 15
- **Figure I.10 :** Exemple de La dispersion d'un agent surfactant de la salissure 16
- **Figure I.11 :** structure d'une émulsion 16
- **Figure I.12 :** Mécanisme de la détergence 17
- **Figure I.13 :** Mécanisme de transfert des HCS après un déversement à la surface du sol 20
- **Figure II.1**: Modèle d'étude quantitative des relations structure/activité (QSAR) 25
- **Figure.II.2.** Schéma d'élaboration et validation d'un modèle QSPR 31
- **Figure II.3** : Structure d'un neurone artificiel 44
- **Figure III.1 :** Méthodologie de calcul du (Log CMC) par les réseaux de neurones 56
- **Figure III.2 :** Illustration de la phase d'apprentissage et de test 58
- **Figure III.3 :** L'organigramme de la méthodologie de calcul du réseau de neurones statique optimale RNSO 60
- **Figure III.4 :** Architecture de réseau de neurones 65
- **Figure III.5** : Réseau de neurones multicouche pour logCMC 65

- **Figure III.6 :** Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales 66
- **Figure III.7 :**Présentation de l'erreur absolue 67
- **Figure III.8:** Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales 72
- **Figure III.9 :**Presentation de l'erreur absolue 73
- **Figure III.10 :** Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales 77
- **Figure III.11 :**Présentation de l'erreur absolue 78
- **Figure III.12:** Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales 81
- **Figure III.12 :** Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales 82
- **Figure III.13 :**Présentation de l'erreur absolue 83

Lite des tableaux

- **Tableau I-1** : Différents types de tensioactifs 4
- **Tableau I.2**: Différentes structures de micelles 6
- **Tableau I.3 :** exemples de structures de certains tensioactifs représentatifs pour chaque classe **7**
- **Tableau I.4 :** Mécanisme d'action des tensio-actifs 17
- **Tableau I.5** Récapitulatif des différents types de tensioactifs et leurs applications 22
- **Tableau II.1** : Sodium DodecylSulfate en différant démentions33
- **Tableau II.2** : Analogie entre le neurone biologique et le neurone formel 37
- **Tableau II.3**: Fonctions de transfert usuelles, *a=f(n)* 45
- **Tableau II.4** : Exemple de différents types d'apprentissages en mode supervisé 47
- **Tableau III.1 :** base des données des structure Non-ioniques de ce étude 61
- **Tableau III.2** : Structure du réseau de neurones optimisé 64
- **Tableau III.3 :** Matrice des poids et des biais du couch d'entrée 65
- **Tableau III.4 :** Matrice des poids et des biais du couch caché 66
- **Tableau III.5 :** Matrice des poids et des biais du couch sortie 66
- **Tableau III.6 :** Valeurs des erreurs calculées pour LogCMC 67
- **Tableau III.7 :** les valeur min et max du Log CMC 68
- **Tableau III.8** : Structure du réseau de neurones optimisées structure cationique 70
- **Tableau III.9 :** Matrice des poids et des biais du couch d'entrée 70
- **Tableau III.10 :** Matrice des poids et des biais du couch caché 70
- **Tableau III.11 :** Matrice des poids et des biais du couch sortie 71
- **Tableau III.12 :** Valeurs des erreurs calculées pour LogCMC 72
- **Tableau III.13 :** les valeur min et max du Log CMC 73
- **Tableau III.14** : Structure du réseau de neurones optimisées structure cationique75
- **Tableau III.15 :** Matrice des poids et des biais du couch d'entrée 75
- **Tableau III.16 :** Matrice des poids et des biais du couch caché 76
- **Tableau III.17 :** Matrice des poids et des biais du couch sortie 76

- **Tableau III.18 :** Valeurs des erreurs calculées pour LogCMC 77
- **Tableau III.19 :** les valeur min et max du Log CMC 78
- **Tableau III.20** : Structure du réseau de neurones optimisées structure cationique80
- **Tableau III.21 :** Matrice des poids et des biais du couch d'entrée 80
- **Tableau III.22 :** Matrice des poids et des biais du couch caché 80
- **Tableau III.23 :** Matrice des poids et des biais du couch sortie 81
- **Tableau III.24 :** Valeurs des erreurs calculées pour LogCMC 82
- **Tableau III.25 :** les valeur min et max du Log CMC 83

الملخص: نمذجة الخصائص الفيز وكيميائية لبعض السطحيات

والهدف من هذه الدراسة هو نمذجة الخصائص الفيز وكيميائية المتكونة من 171 أنواع مختلفة من السطحي (أنيوني، الموجبة، غير أيوني وشاملة) من خلال أربعة أنواع من نموذج هجينة بين QSPR-الشبكة العصبية الاصطناعية (RNA). تم حساب عدد الإدخالات نموذج الخلايا العصبية بواسطة برنامج المجذاف والأمثل من خلال طريقة الخوارزمية الجينية.

وتعرض خصائص الشبكات العصبية الاصطناعية الأربع على النحو التالي:

1. logCMC لشبكة الاولي متكونة من 51 السطحي غير أيوني مع نموذج العصبية الاصطناعية التي لديها بنية {1-4-4-4}، وجد هذا الأخير مع معامل الارتباط يساوي R = 0.991، خطأ المطلق النسبي متوسط AARDM=23815 و R^2_{pred} = 0.998 . هذه النتائج بالتحقق من القوة التنبؤية الخارجية لنموذجنا من خلال التأكيد على متانة نموذج المعمول بها.
2. النموذج الثاني يحتوي على بنية {1-7-7-4}، وقد تم تحسين من أجل حساب السطحي لي 88 هيكل أنيوني وجد هذا الأخير مع R = 0.987، ARDM= 5.806 وR^2_{pred} = 0.995. هذه النتائج تثبت متانة نموذج المعمول بها.
3. الشبكة العصبية الثالثة التي لديها بنية {4-5-5-1}، وتظهر هذه الأخير قدرته على التنبؤ ب logCMC لي 71 مركبات الأيونية مع R = 0.929، ARDM =17.944 و R^2_{pred} =0.979.
4. الآخير، هو الشبكة شاملة لديها بنية {1-6-6-4} تضم خليطا من 171 أنواع مختلفة من السطحي (أنيوني، الموجبة، غير أيوني)، تم العثور على شبكة مع R = 0.908، ARDM=17.232 وR^2_{pred}= 0.976.

الكلمات المفتاحية: السطحي، الشبكات العصبية الاصطناعية، والخوارزميات الجينية، وتركيز مذيلة حرجة

Résumé : Modélisation des propriétés physico-chimique des tensioactifs

Le but de ce travail est de modéliser les propriétés physico-chimiques de 171 tensioactifs de différents types (anionique, cationique, non-ionique et globale) par quatre modèle hybrides de type QSPR- réseau de neurones artificiel (RNA). Le nombre des entrées du modèle neuronale a été calculé par le logiciel Padel et optimisée par la méthode des algorithmes génétiques.

Les caractéristiques des quatre réseaux de neurones artificiels sont présentées comme suit :

1. Le logCMC de 51 tensioactifs non-ioniques avec un modèle neuronal qui a une architecture de {4-4-4-1}, ce dernier a été trouvé avec un coefficient de corrélation égale R=0.991, une Erreur Relative Absolue Moyenne ERAM=23.815 et R^2_{pred}=0.998. Ces résultats vérifient la capacité prédictive externe de notre modèle en confirmant la robustesse du modèle établi.
2. Le second modèle comporte une architecture de {4-7-7-1}, a été optimisé afin de calculer les agents de surface de 88 structures cationiques avec un R=0.987, ARDM=5.806 et R^2_{pred}=0.995. Ces résultats montrent la robustesse du modèle établi.
3. Le troisième réseau de neurones à une structure de {4-5-5-1}, ce dernier montre son aptitude à prédire logCMC de 71 composés anioniques avec un R=0.927, ARDM=16.744 et R^2_{pred}=0.979.
4. Le dernier, c'est un réseau global d'architecture {4-6-6-1} englobant un mélange de 171 tensioactifs de différents types (anionique, cationique, non-ionique), le réseau a été trouvé avec un R=0.908, ARDM=17.235 et R^2_{pred}=0.976.

Mots clés : Tensioactif, réseaux de neurones artificiels, algorithmes génétiques, concentration micellaire critique

Abstract: Modeling of physicochemical properties of some surfactants

The aim of this work is to model using four hybrid models qualitative structure-property relationship and artificial neural networks QSPR-ANN the physico-chemical properties of 171 different surfactants (anionic, cationic, non-ionic and global). All inputs of our model are descriptors, these ones were generated using Padel software and optimized using genetic algorithm interface.

The characteristic of the four designed models are illustrated into the following points:

1. The first model has the architecture of {4-4-4-1} was designed to correlate 51 non-ionic surfactants, this model was found with a very low Absolute Average Relative Deviation AARD=23.815, a correlation coefficient R=0.991 and R^2_{pred}=0.998. Results confirm the external prediction attitude and robustness of our optimized model.
2. The second model has an architecture of {4-7-7-1}, this one was optimized to calculate the surfactants of 88 cationic structure with R=0.987, AARD=5.806 and R^2_{pred}=0.995. which confirm the robustness of our model.
3. The third ANN was elaborated with an architecture of {4-8-8-1}, this model present its aptitude to predict logCMC of 71 anionic compounds with a R=0.927, AARD=16.744 and R^2_{pred}=0.979.
4. And the last model has an architecture of {4-6-6-1} involving the three different surfactant types in one model. This ANN was found with R=0.908, AARD=17.235 And R^2_{pred}=0.976.

Keywords: Surfactants, Artificial Neural Networks, genetic algorithm, **critical** micelle concentration

Introduction générale

Les agents de surface ou tensio-actifs (surfactant) sont utilisés depuis longtemps par l'homme. L'usage du savon remonte aux civilisations sumériennes et égyptienne. Pendant plusieurs siècles, il a satisfait la demande en ce qui concerne l'hygiène et la propreté. Avec l'avènement de la société industrielle, de nouveaux agents tensio-actifs ont été élaborés, tandis que de nouvelles techniques de lavage et de nouvelles exigences en matière d'hygiène ont été adoptées. Initialement réalisée à partir de matières premières issues d'organismes vivants (huiles végétales, graisses animales), la fabrication industrielle des tensio-actifs fait aujourd'hui largement appel aux produits dérivés du charbon et du pétrole. Actuellement, les agents de surface sont largement utilisés que ce soit à des fins domestiques ou industrielles, notamment dans les préparations détergentes.

Les tensioactifs sont habituellement des composés organiques amphiphiles, ce qui signifie qu'ils contiennent à la fois des groupes hydrophobes (leurs « queue ») et des groupes hydrophiles (leurs « têtes »). Par conséquent, ils sont solubles dans les solvants organiques et l'eau. En raison de leurs structures amphiphiles uniques **[1].**

Les tensioactifs sont des produits polyvalents utilisés dans les industries variétales telles que les détergents, les émulsifiants, les boues de forage et les agents de flottation. De nos jours, ceux-ci sont appliqués dans des domaines de haute technologie tels que la micro-électronique **[2, 3]** et la nanotechnologie **[4].** À une concentration critique appelée CMC, les tensioactifs se rassemblent pour former des micelles. Les micelles sont en équilibre avec les monomères **[2,3].** le point particulier important dans l'industrie est la détermination et la prédiction de la CMC.

L'une des méthodes pour prédire les propriétés chimiques telles que la CMC basée uniquement sur l'information structurelle moléculaire est la modélisation quantitative de la structure propriété (QSPR). Les modèles QSPR sont des équations mathématiques qui relient la structure chimique à une propriété. La modélisation QSPR est extrêmement utile pour prédire la propriété de nouveaux composés chimiques. Pour une modélisation QSPR, les éléments suivants doivent être effectués:

(1) génération de descripteurs,

(2) fractionnement de données à des ensembles de formation et de prédiction (ou validation),

(3) sélection de variables,

(4) recherche d'un modèle approprié entre variables sélectionnées et propriété

(5) validation de modèle.

Un grand nombre de descripteurs peuvent être générés par des codes existants (e-dragon, Padel). Ces paramètres classés comme des descripteurs géométriques, topologiques, physico-chimiques et électroniques **[5,6].** Choisir des descripteurs adéquats pour les études QSPR est difficile. Pour surmonter ce problème, une technique de sélection de variable puissante est nécessaire **[7, 8].**

L'objectif de notre travail est la modélisation de trois types de tensioactifs (anionique, non ionique et cationique) afin de trouver un modèle mathématique pour la détermination de log CMC dans une grande marge de concentration pour chaque type, puis un autre modèle qui englobe les trois types au même temps.

Dans le présent travail, les meilleures variables parmi les descripteurs électroniques et topologiques sont sélectionnées à l'aide des algorithmes génétiques, puis un modèle linéaire est développé pour prédire le Log CMC des tensioactifs à l'aide des RNA. Notre méthode montre une amélioration considérable pour les modèles de régression.

Ce mémoire s'articule sur les chapitres suivants :

- Chapitre I : cette partie sera consacré à une mise au point sur le noyau des agents du surface. Nous aborderons une généralité sur le tensioactif et leur caractéristique
- Chapitre II : nous présentons un aperçu général théorique sur les différentes méthodes de la modélisation structurales utilisées dans notre travail (Quantitative Structure- Propriétés Relationship (QSPR) et réseaux neuronaux artificiels (RNA)). Nous présenterons aussi leur domaine d'application et les programmes utilisés.
- Chapitre I : rassemblant les différents résultats.

Et on termine par une conclusion générale.

I. Généralité sur la tensioactif

I.1. Introduction

Les molécules amphiphiles sont constituées de groupements ayant des affinités opposées vis-à-vis d'un même solvant, très souvent l'eau. Plus généralement cette propriété est la conséquence de fortes différences de polarité entre ces groupements. Ces molécules possèdent donc une partie hydrophile et une partie hydrophobe. Les tensioactifs ainsi que certains copolymères présentent ce caractère amphiphile et possèdent donc la propriété de s'auto-associer en solution aqueuse.

Dans ce chapitre nous allons décrire les différents types de tensioactifs et ces applications : les tensioactifs monomères possédant une seule tête polaire (monocaténaire, bicaténaire et tricaténaire), les tensioactifs dimères possédant deux têtes polaires (bolaformes et gémini), les oligomères avec plus de deux têtes polaires (trimères et tétramères) et enfin les polysavons. Compte tenu des systèmes polymères au centre de ce travail, nous allons surtout nous intéresser à des tensioactifs cationiques.

I.2. Définition

Les molécules tensioactives sont aussi appelées agents de surface ou encore surfactants en anglais. Leur toute première propriété est leur tendance très marquée à migrer vers les interfaces, en particulier l'interface air-liquide et en conséquence d'abaisser la tension superficielle de l'eau [9].

En outre les molécules sont amphiphiles en Grass appeler amphiphilos (Amphi : les 2, philos : aimant). Elles contiennent à la fois une partie polaire possède une grande affinité pour l'eau (tête hydrophile) et une autre partie apolaire hydrophobes présentant un comportement tout à fait opposé qui est soluble dans les huiles, on les schématise souvent comme montre la **figure I-1.**

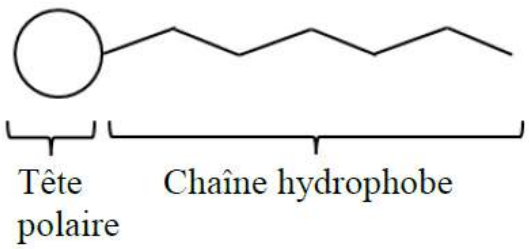
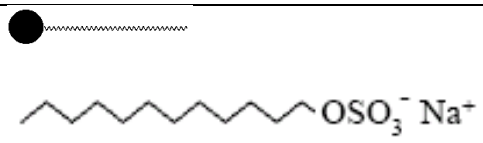
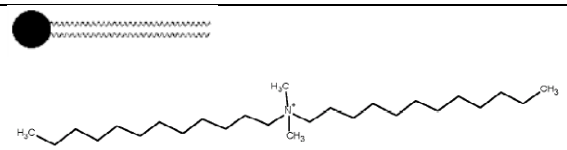
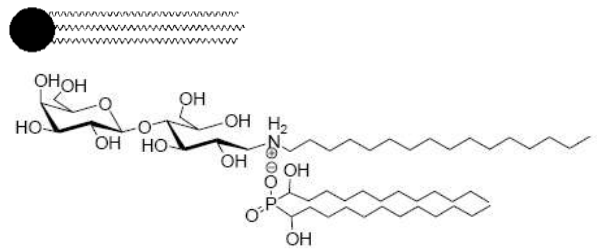
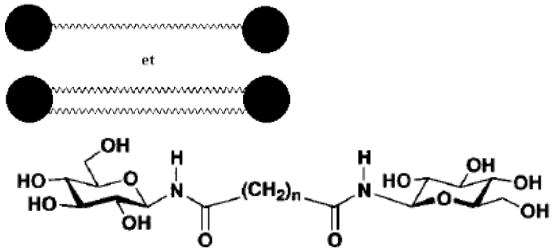
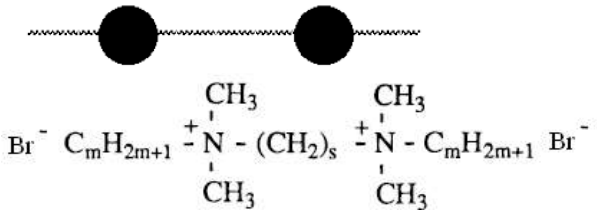


Figure I-1 : Schéma d'un tensioactif [9]

Les tensioactifs, quelle que soit la nature des têtes polaires, peuvent présenter des structures très diverses (Tableau I-2). Dans le cas des tensioactifs monocaténaires, bicaténaires et tricaténaires, les têtes polaires portent respectivement une, deux et trois chaînes carbonées. Les bolaformes, quant à eux, sont constitués de deux têtes polaires reliées par un ou deux segments hydrophobes. Enfin, les tensioactifs gémini sont constitués de deux têtes polaires portant chacune une chaîne alkyle et reliées par un segment hydrophile ou hydrophobe appelé espaceur [9].

Tableau I-1 : Différents types de tensioactifs.

Type de tensioactif avec exemple	**Structure du tensioactif**
Monocaténaire SDS: Sodium Dodecyl Sulfate (tensioactif anionique) $C_{12}H_{25}$-O-S $O^-_3Na^+$ **[9]**	
Bicaténaire BDDA : Bromure de Dimethyl-Dodecyl-Ammonium (C12H 25)2-N^+-(CH3)2 Br^- **[10]**	
Tricaténaire N-hexadécylammonium-1-Déoxylactitol bis-(α-hydroxy dodecyl) phosphonate **[11]**	
Bolaformes [12]	
Gemini Alkanediyl-α,ω-bis(Alkyl Dimethylammonium)Bromide **[13]**	

= tête polaire = chaîne apolaire

En solution à faible concentration, un tensioactif est présent sous forme de monomères. Au-delà d'une concentration critique qui s'appelle la concentration micellaire critique (CMC), les molécules de monomères s'organisent pour former des agrégats moléculaires appelés micelles, de taille inferieur a 10 nm.

Les micelles sont dites "directes" dans l'eau où l'assemblage de tensioactifs se constitue avec les têtes polaires orientées vers l'extérieur et les chaînes hydrophobes rassemblées au cœur de la micelle. L'autre type est dit les micelles inverses, dont la tête polaire hydrophile à l'intérieur de la micelle et les chaînes hydrocarbonées se situant dans le milieu organique.

I.3. Structuration des tensioactifs en solution aqueuse

A faible concentration, les tensioactifs moléculaires s'agrègent sous forme de micelles dispersées dans l'eau. Ces micelles sont le plus souvent sphériques mais peuvent être aussi cylindriques. A concentration en tensioactif plus élevée, les agrégats deviennent des cylindres infinis, organisés suivant un réseau bidimensionnel hexagonal dans la phase hexagonale, puis des lamelles infinies empilées périodiquement de façon alternée avec des couches d'eau dans la phase lamellaire. Ces structures se distinguent les unes des autres par la courbure inter faciale **[14]**.

En l'absence de contraintes extérieures, la courbure de l'interface et plus spécifiquement la géométrie des agrégats dilués dépend de la géométrie du tensioactif que l'on caractérise quantitativement par le paramètre d'empilement $\mathbf{p = v/a_0 l_c}$ où $\mathbf{a_0}$ est la surface de la tête polaire du tensioactif et **v** et $\mathbf{l_c}$ le volume et la longueur maximale de la chaîne aliphatique **[15,16]**. Ainsi, le rayon de courbure spontané est défini par la structure chimique du tensioactif via les paramètres moléculaires v et l_c mais aussi par les interactions locales entre molécules voisines qui fixent la valeur optimale de $\mathbf{a_0}$. Cette valeur résulte en effet de la compétition entre l'attraction hydrophobe entre les chaînes et les interactions répulsives d'origine stérique ou électrostatique entre les têtes polaires.

Tableau I.2: Différentes structures de micelles

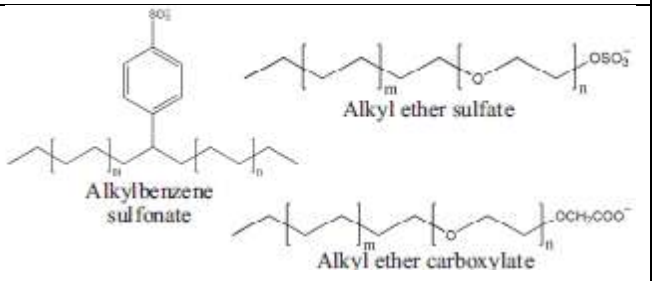
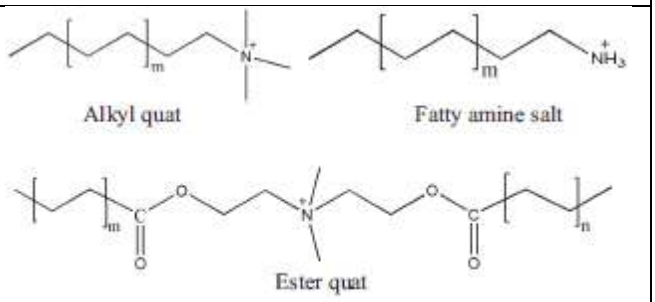
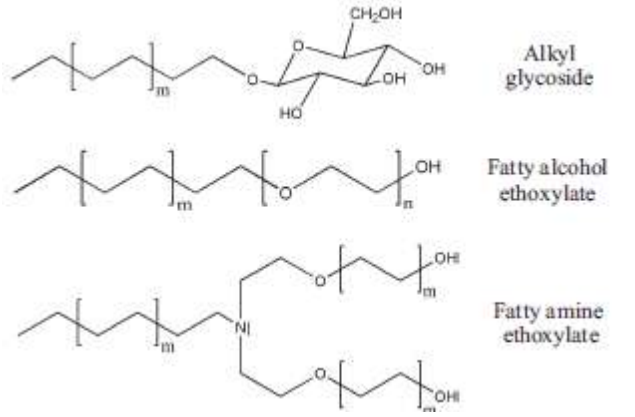
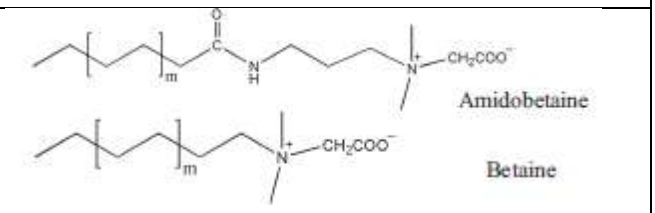
Pour former une micelle directe (sphère), les molécules devront adopter une forme de cône pour laquelle $P < 1/3$	
Pour former un disque ou un cylindre, elles constitueront une portion de disque ou de cylindre pour laquelle $1/3 < P < ½$	
Pour former une lamelle, elles devront adopter une forme cylindrique pour laquelle $1/2 < P < 1$	
Enfin, si $P > 1$, les molécules constitueront une micelle inverse.	

De nombreux facteurs physiques et physico-chimiques peuvent modifier les interactions entre les têtes polaires et les chaînes aliphatiques et par conséquent induire des variations des paramètres a_0 et l_c et de la courbure spontanée. Il en est ainsi de l'addition d'un co-tensioactif, de la variation de la force ionique lorsque le tensioactif est chargé ou encore de celle de la température. L'ajout de sel à un tensioactif ionique, en écrantant les interactions électrostatiques entre les têtes polaires permet une augmentation rapide de la taille des micelles.

I.4. Classification

Il y a plusieurs types de classifications des tensioactifs selon la nature de la tête polaire, la longueur de la queue hydrophobe et de leur origine naturelle ou synthétique , la plus connues mondialement est selon la nature de leur pôle hydrophile en trouve quatre grandes familles : les tensioactifs anioniques chargés négativement, les cationiques chargés positivement, les zwitterioniques possédant à la fois une charge positive et une charge négative et les non ioniques, non chargés.

Tableau I.3 : exemples de structures de certains tensioactifs représentatifs pour chaque classe [17,18].

Classe de tensioactifs	Exemple de leur structure
Tensioactifs anioniques	
Tensioactifs Cationiques	
Tensioactifs non ioniques	
Tensioactifs Zwitterioniques	

I.4.1. Les tensioactifs anioniques

Les tensioactifs anioniques sont historiquement les premiers tensioactifs utilisés sous la forme de savons. Sont constitués d'un groupe de tête anionique comme l'exemple de SDS ou un phosphate. Résultant de la saponification de triglycérides d'origine végétale (huile de palme,

d'olive,…) ou animale. Peu après la première guerre mondiale, les alkyls sulfatent R-O-SO3– M+ et alkyléthers sulfates

R-(O-CH2-CH2)n-O-SO3–M+ ont été introduits sur le marché. Leur intérêt résulte dans la possibilité de synthétiser une large gamme de molécules à l'hydrophobicité variable en fonction de la longueur de la chaîne, de 12 à 18 carbones. La plus connue est le dodécyl sulfate benzène de sodium (SDBS).

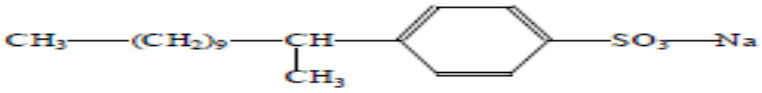


Figure I.2 Exemple de tensioactif anionique : le dodécyl sulfate benzène de sodium (SDBS)

Les détergents synthétiques à base d'ABS (alkyl benzène sulfonâtes) ont remplacé les savons pour le lavage domestique du linge car plus économiques et présentant de meilleures propriétés détergentes, même dans une eau dure. Ils furent eux-mêmes abandonnés vers les années 50 au profit de LAS (alkyl benzène sulfonates linéaires) plus biodégradables du fait de la linéarité de la chaîne. Outre cela ils ne produisent pas de mousses persistantes lors de leur élimination des eaux usées sans affecter leur qualité de détergents.

I.4.2. Les tensioactifs non ioniques

Les tensioactifs non ioniques sont la deuxième classe d'agents tensioactifs les plus utilisés. Le groupe de tête de tensioactif ne supporte ni une charge positive ni une charge négative. Les polyoxyéthylènes (POE) issus de la polymérisation d'unités oxydes d'éthylène sur des alcools gras (Figure I.3). De faible coût, ils sont rapidement utilisés pour leurs performances tensioactives à faible concentration. Peu moussants ils rentrent dans la composition des lessives. Les tensioactifs non ioniques sont souvent utilisés pour les détergents en poudre ou liquides et leur application peut être trouvée dans un domaine industriel étendu

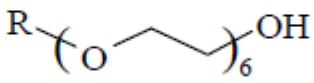


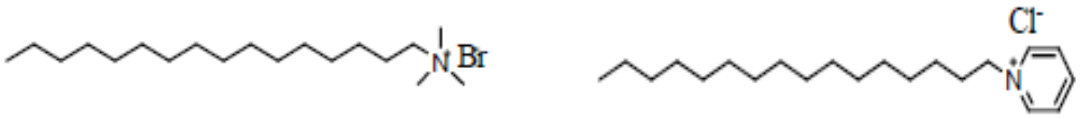
Figure I.3. Exemple de polyoxyéthylène POE (noté R12POE6).

I.4.3. Les tensioactifs cationiques

Les tensioactifs cationiques contiennent habituellement un atome d'azote chargé positivement. Les amines ainsi que les tensioactifs à base d'ammonium quaternaire sont fréquents (tableau I.3). Le plus souvent ce sont des sels d'ammonium quaternaires triméthylés ou des sels de pyridinium [19].

Ils se dissocient en solution aqueuse en un cation organique, et un anion généralement de type halogénure. La grande majorité de ces surfactifs sont des composés azotés de type sel d'amine grasse ou d'ammonium quaternaire. Ces surfactifs sont beaucoup plus chers à fabriquer que les antérieurs, et de ce fait on ne les utilise que dans des applications particulières qui font appel à leur propriété bactéricide, ou à leur capacité de s'adsorber facilement sur des substrats biologiques ou inertes possédant une charge négative. Cette adsorption en fait d'excellents agents antistatiques, hydrophobants et mêmes inhibiteurs de corrosion, utilisés dans les produits tant industriels que d'usage domestique.

Les tensioactifs cationiques sont généralement plus coûteux que les tensioactifs anioniques. Ce type d'amphiphiles est la troisième classe d'agents tensioactifs utilisés habituellement pour les adoucisseurs de tissus, dans les détergents pour lessive (principalement pour améliorer l'emballage des molécules tensioactives anioniques) ou dans des produits ménagers domestiques (en raison de leur propriété d'être désinfectant). [17,18]



HTAB (bromure d'exadecenyltrimethilammonium) HPCL (chlorure d'éxadécyl pyridinium)

Figure I.4 Exemples de tensioactifs cationiques : le HTAB et le HPCl.

I.4.4. Les tensioactifs zwitterioniques (amphotères)

Les tensioactifs zwitterioniques, ou amphotères présentent à la fois une charge cationique et anionique (tableau I.3). Possèdent sur la partie hydrophile à la fois une charge positive et une charge négative (Figure I.5). Ce type de tensioactifs peut alors aisément devenir cationique ou anionique selon le pH de la solution dans laquelle ils sont solubilisés. Cette structure dipolaire s'apparente à celle des phospholipides naturels et conduit à une famille de produits généralement

non irritants et peu agressifs sur le plan biologique. Ces surfactants sont en général très peu irritants, compatibles avec les autres surfactants et dans la plupart des cas, ils peuvent s'utiliser dans des formules pharmaceutiques ou cosmétiques.

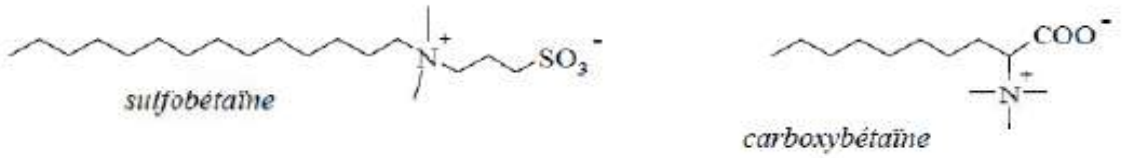


Figure I.5 : Exemples de tensioactifs zwitterioniques : la sulfobétaïne et la carboxybétaïne.

I.4.5. Classification des tensioactifs en fonction de la longueur de la chaine hydrophobe

L'autre classification Solon la longueur de la chaine hydrophobe. Concernant la queue hydrophobe, il est généralement admis que :

- Les chaînes courtes, de 8 et 10 atomes de carbones, confèrent à la molécule des propriétés mouillantes.
- Les chaînes moyennes, de 12 à 16 atomes de carbone, confèrent à la molécule des propriétés mouillantes et détergentes.
- Les chaînes longues, de 18 à 22 atomes de carbone, confèrent à la molécule des pouvoirs émulsifiants et adoucissants.

Donc un tensioactif peut avoir différentes fonctions : Détergents, Agents de solubilisation, Agents moussants, Agents mouillants, Agents dispersants, Agents émulsifiants [20].

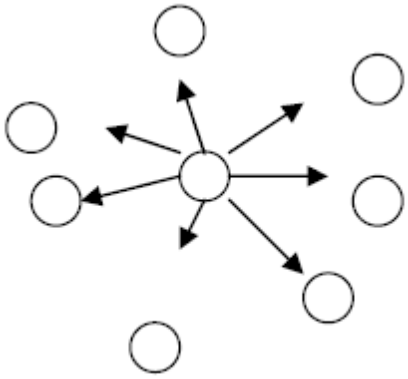
I.5. Propriétés

I.5.1. Détergents

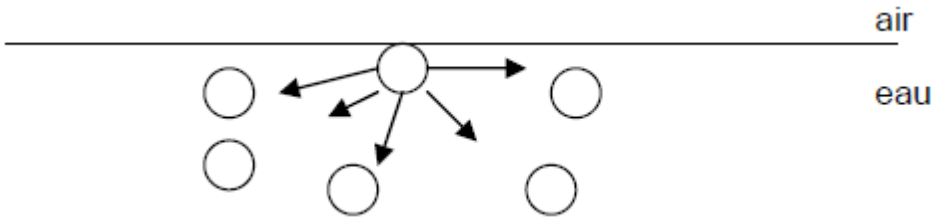
Le mot détergent provient du latin « déterger » qui signifie nettoyé. Les détergents sont donc des produits capables d'éliminer les souillures grâce à leur aptitude à abaisser la tension superficielle du liquide dans lequel ils sont dissous [21,22]. Ils sont composés généralement d'un tensioactif qui représente la matière active et d'un ensemble de produits chimiques appelés adjuvants.

I.5.2. Propriétés superficielles

Dans les liquides, les molécules sont en interactions entre elles (forces de Van der Waals). Au sein du liquide, la résultante des forces appliquées à une molécule est nulle :



Au voisinage de la surface, par contre, apparaît une dissymétrie : la résultante des forces appliquées à une molécule n'est plus nulle et s'exerce vers l'intérieur. La tendance de la surface du liquide est donc de se restreindre.

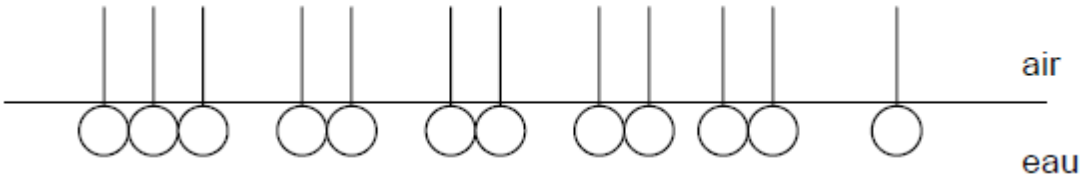


Pour étirer la surface du liquide, il faut donc appliquer une force et donc fournir du travail. On appelle tension superficielle γ , le travail d_W qu'il faut fournir par unité de surface d_S pour étirer celle-ci ou la force par unité de longueur pour étirer celle-ci.

$$\gamma = d_W / d_S = dF / dx \qquad (1)$$

γ s'exprime en J.m-2 ou en N/m.

L'ajout d'un produit tensio-actif diminue considérablement la tension superficielle de l'eau par la disposition que les molécules prennent dans le liquide :



C'est cette disposition qui est à la base des pouvoirs mouillant, moussant et émulsionnant.

Toutefois à partir d'une certaine concentration en tensio-actif, appelée concentration micellaire critique (CMC), la surface du liquide étant saturée de molécules tensio-actives, celles – ci se regroupent entre elles dans le liquide en formant des micelles voire des vésicules comme les liposomes, les niosomes ou les nanocapsules :

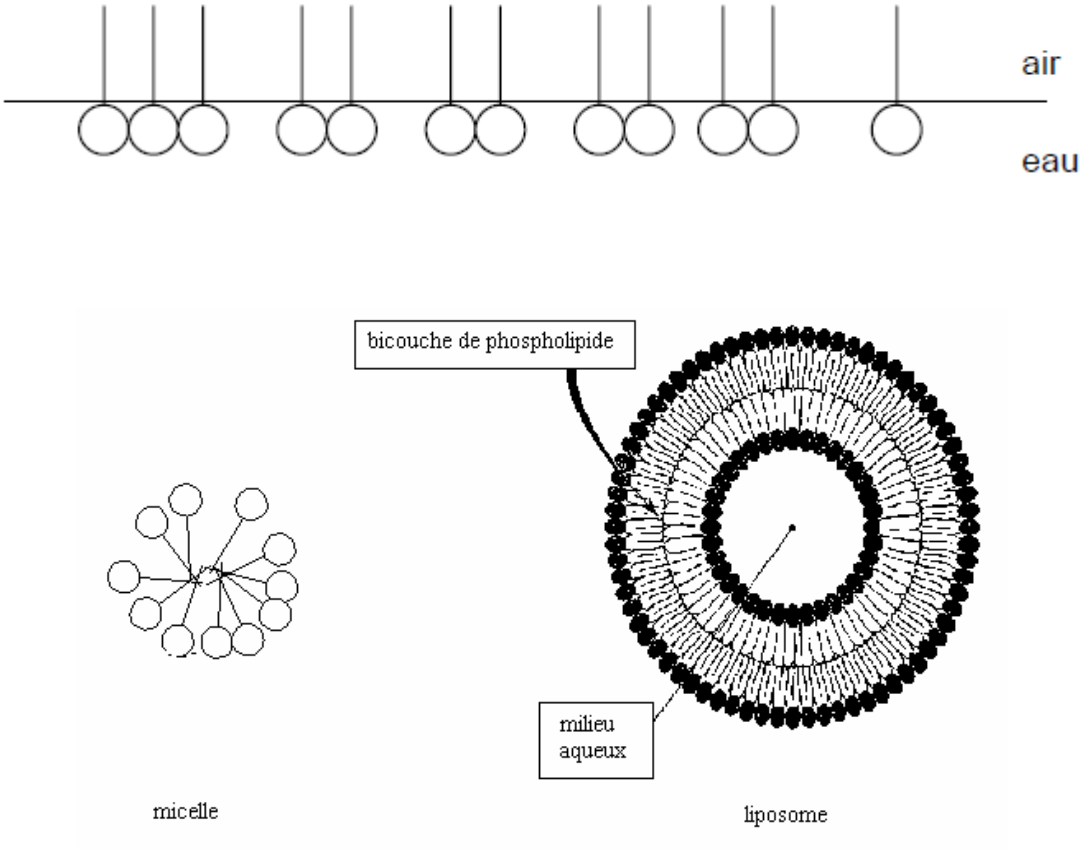


Figure I.6 : la forme des micelles dans liquide

C'est la formation de micelles qui est à la base des pouvoirs de détergence.

Les liposomes et autres vésicules sont de plus en plus utilisés en cosmétologie et pharmacie pour véhiculer les agents actifs au travers des membranes cellulaires.

L'état thermodynamique des molécules superficielles d'une surface condensée, est différent de celui des molécules se trouvant à l'intérieur du liquide ; de même l'attraction des molécules superficielles vers l'intérieur est souvent plus intense que celle existant dans un liquide, ce qui entraîne leur soumission à des forces non équilibrées. On dit alors que le liquide est soumis à une tension superficielle**.**

I.5.3. La concentration micellaire critique (CMC)

A faible concentration de tensioactif, la solution est constituée de monomere situées à l'interface eau-air formant une couche mono moléculaire continue. Une fois l'interface saturée, les nouvelles molécules ajoutées se regroupent au sein du liquide soit par leurs parties hydrophiles soit par leurs parties hydrophobes; on assiste alors à la formation de micelles.

On distingue différents types de micelles : les sphériques qui existent en solution diluée et les lamellaires qui sont présentes en solution concentrée. La présence de micelles a pour conséquence une augmentation importante de la solubilité et de la solubilisation [22,23].

C'est une caractéristique essentielle du tensioactif, elle dépend de sa nature chimique ainsi que de certains facteurs externes tels que la concentration en électrolytes et la température. Les CMC des tensioactifs non ioniques (de l'ordre de 10^{-5} mol. L^{-1}) sont inférieures aux CMC des tensioactifs ioniques conventionnel (de l'ordre de 10^{-3} mol. L^{-1}) [24,25].

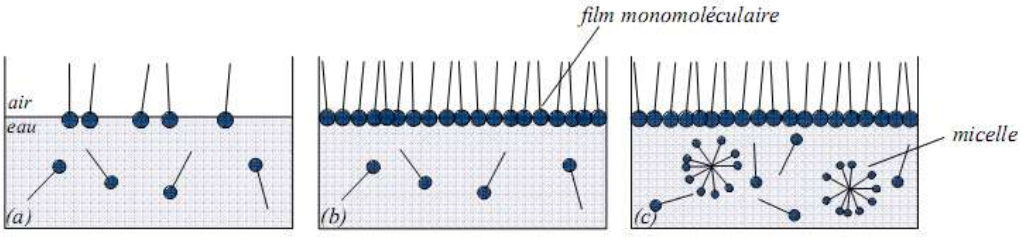


Figure I.7 : Comportement des tensioactifs en milieu aqueux– phénomène de micellisation.

La surface du liquide devient saturée en molécules de tensioactif et la tension superficielle du liquide reste constante. L'addition supplémentaire de tensioactifs provoque un auto assemblage spontané des monomères en agrégats qui peuvent être divers, définis par leur taille et leur nombre d'agrégation (nombre de molécules composant l'agrégat). Quand la concentration totale de tensioactifs est supérieure à la CMC, la concentration en monomères dans la solution reste égale à la CMC **(Figure I.7).**

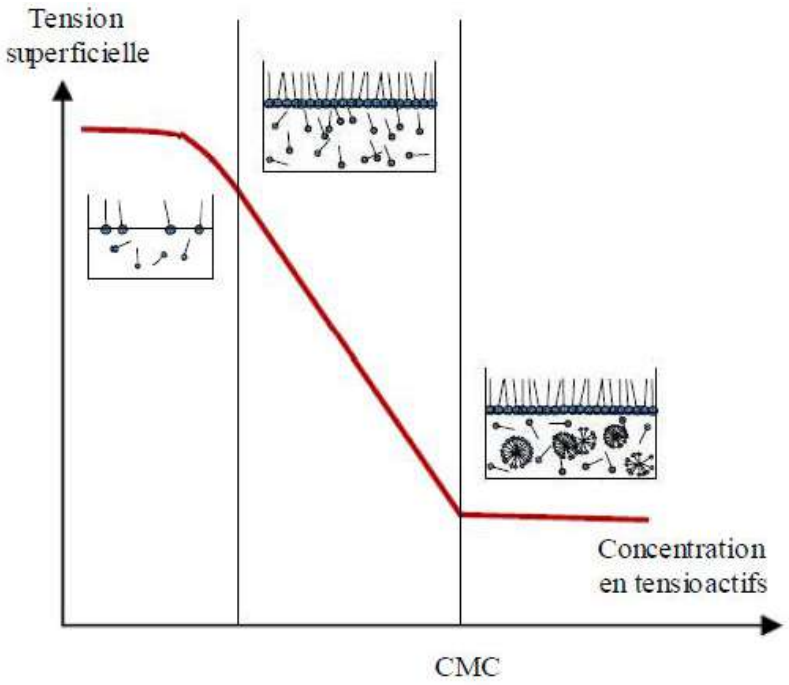


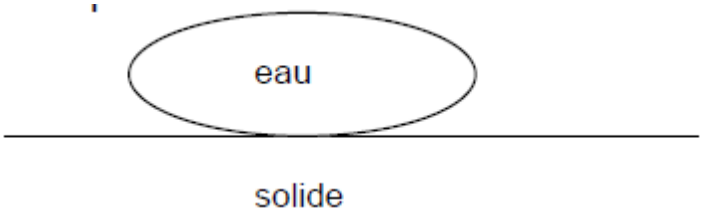
Figure I.8 : Courbe de la tension de surface en fonction de la concentration de tensioactifs en solution.

I.5.4. Le mouillage

Le pouvoir mouillant se définit par la propriété d'étalement du liquide sur un substrat donné et par son pouvoir de pénétration dans les pores du tissu.

Eau seule : γ élevée

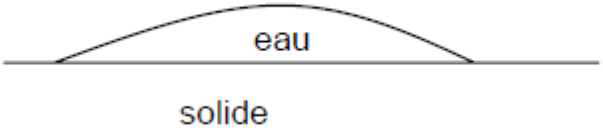
- La cohésion du liquide l'emporte sur les interactions avec le solide
- Le liquide s'étale peu



Moins de cohésion

Eau + tensio-actif => la tension superficielle démunie → Le liquide s'étale

Mouillage meilleur



I.5.5. La mousse

La mousse est un film dans lequel le tensioactif est adsorbé aux deux interfaces eau/air formant une double couche.

Les molécules tensio-actives renforcent la mince pellicule d'eau qui forme les bulles.

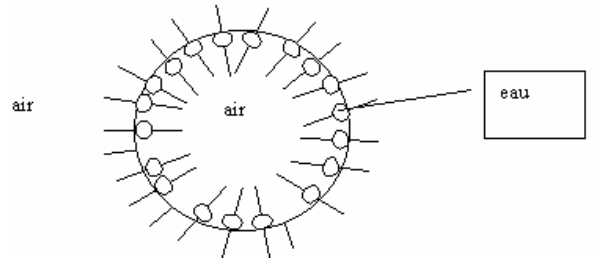


Figure I.9 : structure de la mousse dans le milieux air/eau

1.1.1 La dispersion

Pour maintenir des particules dispersées, deux techniques sont possibles : on les enrobe d'un film colloïdal protecteur ou on utilise un électrolyte qui produit leur répulsion mutuelle.

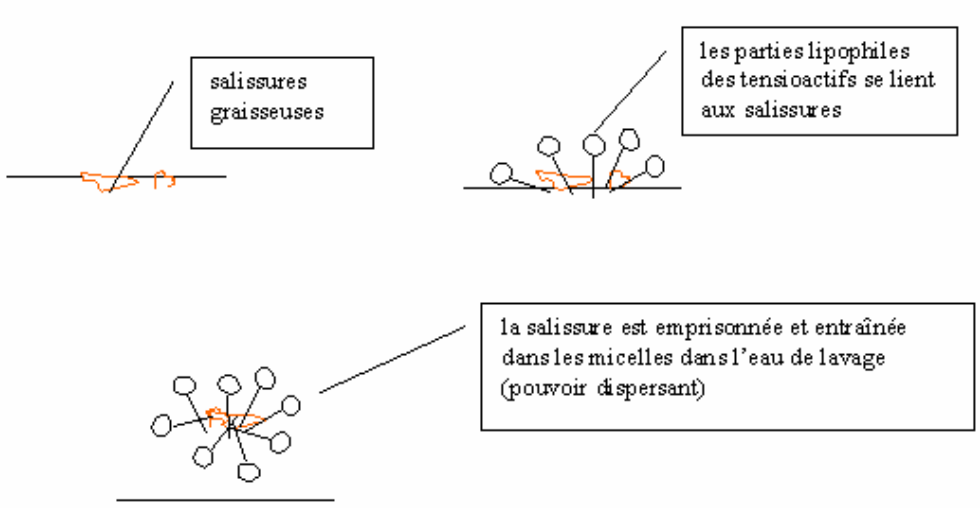


Figure I.10 : Exemple de La dispersion d'un agent surfactant de la salissure

I.5.6. L'émulsion

L'émulsion est définie comme étant une suspension de particules liquides au sein d'une autre phase liquide non miscible.

Les molécules tensio-actives stabilisent les gouttelettes d'huile dans l'eau ou d'eau dans l'huile

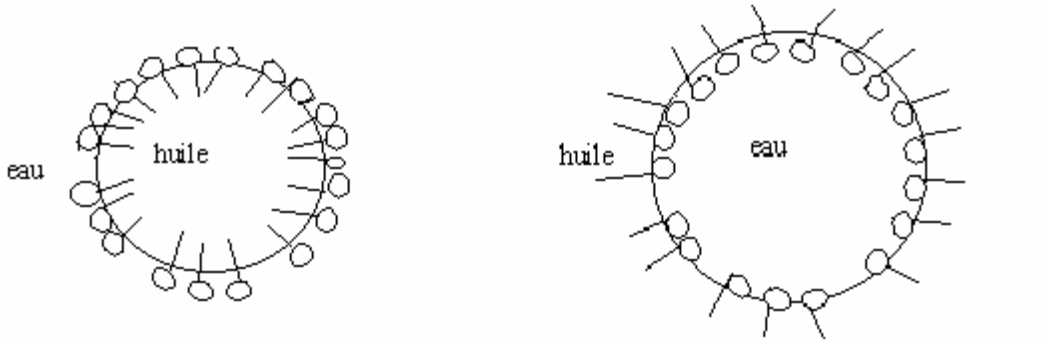


Figure I.11 : structure d'une émulsion

I.6. Mode d'action

Le mode d'action des tensioactifs repose sur un phénomène physique. Leur action superficielle correspond à un véritable phénomène d'adsorption. En effet ils s'adsorbent à la surface de l'eau et s'y orientent, la partie hydrophile polaire est alors dirigée vers l'intérieur créant ainsi une discontinuité dans le film superficiel. Cette dernière engendre une diminution de la tension superficielle du solvant. (**La Figure I.11)** schématise le mécanisme de la détergence.

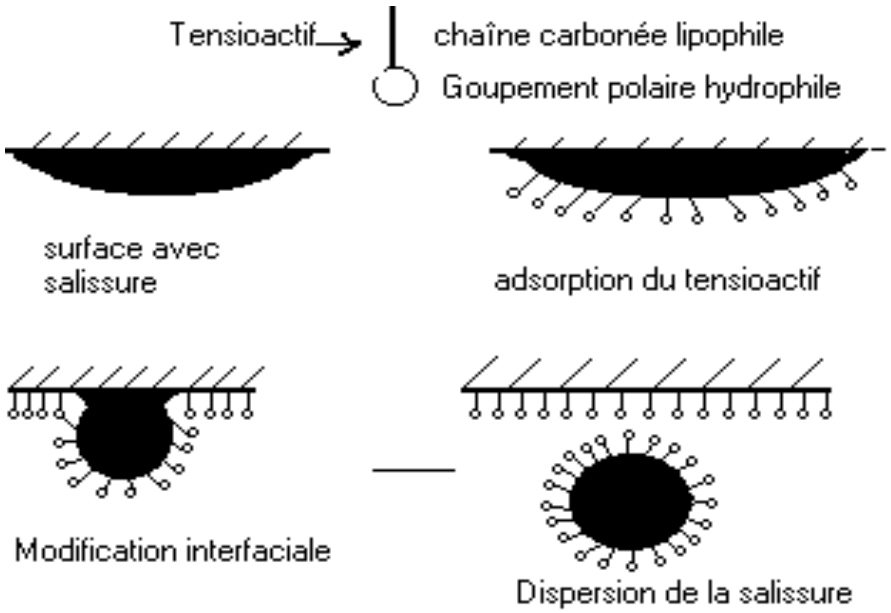


Figure I.12 : Mécanisme de la détergence

Les différents pouvoirs tels que le mouillage qui permet à la solution détersive de pénétrer dans les pores jusqu'au contact avec les impuretés, le pouvoir émulsionnant qui maintient les impuretés en solution, le pouvoir moussant qui entraîne la salissure, le pouvoir dispersant qui empêche l'agrégation de la salissure et le pouvoir solubilisant qui dissout les liquides organiques, jouent chacun un rôle important dans ce que l'on appelle le pouvoir détersif.

Tableau I.4 : Mécanisme d'action des tensio-actifs [21]

Parties	Partie **hydrophile**	Partie **lipophile**
Polarité	Polaire	Apolaire
Composition	Groupe carboxylate $-COO^-$, sulfonate $-S(O_2)\,O^-$, ...	Longue chaîne carbonée saturée $-CH_2-CH_2-$ ou partiellement insaturée -CH=CH-
Attirée par des composés	polaires, comme l'eau et les solutions aqueuses	apolaires, comme les graisses ou les solvants organiques

I.7. Secteurs d'application des tensioactifs

De nombreux domaines sont concernés : les industries des additifs alimentaires, des céramiques, du cuir, des matières plastiques, métallurgique, minière, du papier, des peintures, du

pétrole, routière, textile… Les produits phytosanitaires et les engrais sont également de gros consommateurs de tensioactifs. Ci-après sont donnés quelques exemples d'utilisation technique dans les industries :

- **La détergence**

Les tensioactifs entrent dans la formulation de tous les détergents, ils représentent en moyenne 20% de leur composition. Le tensioactif est adapté à l'usage et à l'aspect souhaité pour le détergent, il peut se présenter sous forme de poudre ou de liquide.

Il faut distinguer la détergence ménagère et la détergence professionnelle. Cette dernière concerne spécifiquement les industries et les institutions (écoles, hôpitaux, etc.). L'industrie agroalimentaire très préoccupée par les problèmes d'hygiène et de désinfection, représente le plus gros marché. La détergence ménagère, quant à elle, regroupe plusieurs grandes classes de produits : les lessives, les adoucissants, les produits vaisselle et les produits d'entretien. Les secteurs des lessives est le plus gros consommateur de tensioactifs, il représente plus de 50 % du marché.

- **Cosmétique**

Il faut distinguer deux grands domaines d'application : le cosmétique rincé et le cosmétique blanche. Le cosmétique rincé englobe les produits d'hygiène lavant tels que les shampooings, gels douches et dentifrices. Dans ce domaine la propriété détergente du tensioactif est recherchée. Le cosmétique blanc concerne les produits de beauté en général, c'est-à-dire, les crèmes, les laits, les maquillages et autres produits de soins. Les tensioactifs sont principalement utilisés pour générer des émulsions, leur pouvoir adoucissant est également fortement sollicité.

- **Cuir** : les diverses opérations effectuées au cours du traitement du cuir nécessitent l'utilisation des différentes propriétés des tensioactifs (mouillante et détergente), notamment lors de la préparation des peaux au tannage pour mouiller le cuir.
- **Plastiques** : les tensioactifs interviennent dans différentes opérations, notamment au cours des synthèses réalisées par polymérisation en émulsion, dans la formulation des plastisols PVC pour en modifier les caractéristiques rhéologiques, et enfin, comme antistatiques.
- **Métallurgique** : les tensioactifs sont utilisés pour le nettoyage et le dégraissage des matériaux mais également dans les formulations mises en œuvre pour les opérations de formage, d'emboutissage et de découpage fin.

- **Minière** : les tensioactifs sont employés pour la flottation des minerais, technique utilisée lors de l'enrichissement du minerai. La partie noble du minerai à extraire est concentrée dans les mousses de surface obtenues en présence de tensioactifs.
- **Peintures :** les tensioactifs permettent de stabiliser les formulations, de favoriser le mouillage des pigments et de les maintenir en suspension. Ils donnent également leur caractère hydrophobe aux antirouilles.
- **Pétrolière :** la plupart des opérations de l'industrie pétrolière, du forage au raffinage nécessite l'utilisation d'agents de surface de toute nature (anioniques, non ioniques, cationiques).
- **Des produits phytosanitaires** Ils sont également utilisés pour mettre en suspension les principes actifs phytosanitaires en faisant appel à leurs propriétés mouillantes et dispersantes.
- **Engrais**

Les tensioactifs permettent de conditionner les engrais sous forme de granulés pour éviter leur agglomération.

I.8. Application des tensioactifs dans la remédiation environnementale

La contamination des sols par des polluants organiques tels que les hydrocarbures est le résultat d'une accumulation quotidienne de la pollution générée par les activités industrielles sur une longue période (voir la figure I.12**)** [26,27]. Ces sols contaminés présentent un intérêt particulier pour les autorités sanitaires en raison du danger qu'ils présentent pour la santé humaine [26 -28]. En effet, en tant que zone source par nature difficilement biodégradable, ils constituent des foyers de dissémination à long terme [27-30]. Ils sont une menace sérieuse pour l'homme et les animaux à travers la contamination de la chaîne alimentaire, la pollution atmosphérique et celle des eaux souterraines. La sauvegarde et la réutilisation de ces ressources nécessitent donc de traiter prioritairement les sources de contamination, [26,31].

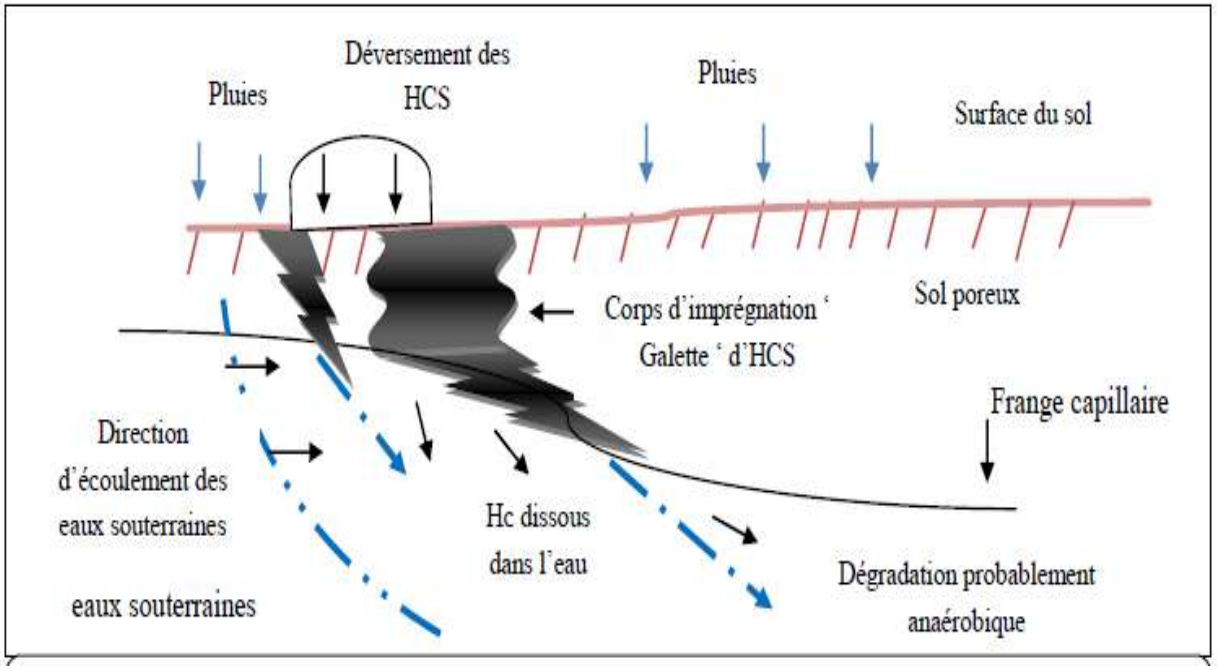


Figure I.13 : Mécanisme de transfert des HCS après un déversement à la surface du sol [32]

Les sols contaminés par les hydrocarbures présentent un danger lors d'un contact direct avec l'homme ou lors de leur transfert dans la chaîne alimentaire par le phénomène de bioaccumulation [26].

La toxicité des hydrocarbures vis-à-vis des personnes et de l'environnement dépend de leur structure et de leur concentration, [29,30]. Les problèmes liés au site contaminé par les polluants organiques dans les milieux environnementaux ont fait l'objet d'une attention croissante des chercheurs. Pour résoudre ces problèmes, une technologie innovante pour le nettoyage de cette contamination.

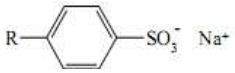
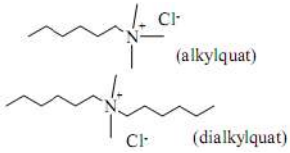
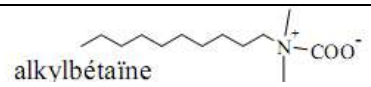
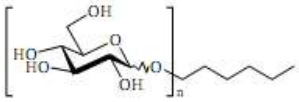
La contamination des sols par le polluant organique pose un vrai problème environnemental, cette contamination est due au :

- Déversements plus ou moins massifs lors d'accidents de transport d'hydrocarbures par route de fer, lors de ruptures accidentelles d'oléoducs, lors d'incendies ou d'explosions d'unités de production industrie ou de cuves de stockage ou des centres de distribution, lors des opérations de transfert et de distribution ;
- Fuites d'oléoducs et de cuves enterrées, industrielles ;

- Infiltrations à partir de dépôt de déchets industriels et urbains contenant des hydrocarbures, mal stockés et mal contrôlés ;
- Déversements et enfouissements volontaires de déchets d'hydrocarbures, liés à des activités domestiques, commerciales, agricoles ou industrielles [33].

La solubilisation des polluants organiques par les tensioactifs est un facteur clé dans la remédiation des sols contaminés par des hydrocarbures, [34,35]. Les tensioactifs peuvent augmenter la solubilité des hydrocarbures en les introduits dans les noyaux hydrophobes des micelles tensioactifs [36,37]. Le tableau I.5 récapitule quelques types de tensioactifs avec leurs utilisations

Tableau I.5 Récapitulatif des différents types de tensioactifs et leurs applications.

famille de tensioactifs	principaux caractères	Utilisation
ANIONIQUE		
savons $R\text{-}COO^-Na^+$		Détergent
alkylsulfate $R\text{-}O\text{-}SO_3^-Na^+$	- les plus courants - peu onéreux - n'irritent pas la peau bonne biodégradabilité	agent moussant (shampooing dentifrice), humectant et détergent (liquide vaisselle)
alkyl benzene suffocates linéaire (LAS) 		agent émulsifiant, dispersant et détergent
CATIONIQUE		
ammonium quaternaire 	Charge positive qui leur permet de s'adsorber sur des substrats charges négativement	agent antistatique et adoucissant (produits de rinçage linge et cheveux) bactericide
Amphotères		
 alkylsulfobétaïne derives d'acidesamines	peu agressifs pour les tissus vivants compatibles avec toutes les autres classes de tensioactifs	adoucissant pour textile et cheveux additif inhibiteur de corrosion
NON IONIQUE		
Alcoolséthoxylés $R\text{-}(O\text{-}CH_2\text{-}CH_2)_n\text{-}OH$ Alkylphénolséthoxylés	peu biodégradables	Détergent (6 à 10 OE) dispersant, humectant et émulsifiant (liquide vaisselle et linge)
alkylpolyglycosides (APG) 	- très bas niveau de toxicité. - pas sensible au pH. - non irritant. - bonne synergie avec les autres tensioactifs.	C8-C10 : agent moussant dans l'agrochimie. C12-C16 : Co-tensioactif en détergence soins corporels.

Conclusion

Dans cette étude bibliographique, on a montré le rôle important que jouent les tensioactifs dans le domaine industriel tels que les domaines agroalimentaire, pétrolier et pharmaceutique. Les applications pratiques se rattachent donc à la structure moléculaire de ces derniers. Les propriétés hydrophile et hydrophobe que possèdent les agents de surface leur permettent de jouer un rôle important dans ces domaines. Par ailleurs, le rôle des tensioactifs se manifeste d'abord à l'échelle macroscopique par une forte diminution de la tension superficielle (ou interfaciale), de plus, au-delà de seuils de concentrations caractéristiques (CMC), les solutions de tensioactifs sont capables d'incorporer des quantités importantes de composés normalement insolubles dans le milieu considéré.

II.1. Introduction

La modélisation par apprentissage statistique consiste à construire, à partir d'un échantillon d'individus, des modèles mathématiques qui reproduisent le comportement d'un système, afin de pouvoir prédire – pour un ensemble plus grand d'individus – une ou plusieurs réponses du système à partir de ses variables d'entrée [38].

Dans de nombreux domaines, comme les sciences sociales, la chimie moléculaire ou le traitement de données textuelles, il arrive que les entrées du système se présentent sous forme de structures (réseaux sociaux, arrangements d'atomes, constructions grammaticales des phrases, ...). Ii serait alors avantageux d'utiliser ces structures pour la modélisation des réponses étudiées.

Ceci est souvent le cas dans le domaine de la chimie où, dans de nombreuses applications, les entités en entrée d'un procédé peuvent être des molécules dont on cherche à prédire les propriétés physico-chimiques (réponses du procédé), pour des réactions particulières, à l'aide de modèles construits à partir de données expérimentales.

Il existe un certain nombre de méthodes, dans le domaine de la chimiométrie, qui s'appuient sur le principe que les propriétés physico-chimiques des molécules dépendent fortement de leur structure.

Regroupées sous l'acronyme QSPR (pour Quantitative Structure- Propriétés Relationship), ce sont principalement des méthodes de régression linéaire ou non linéaires qui ont pour objectif de modéliser les propriétés (ou activités) physico-chimiques à partir de caractéristiques décrivant la structure des molécules [39,40].

Le point de départ de telles méthodes se construit sur la définition des descripteurs moléculaires empiriques ou théoriques. Ces dernières prennent en compte des informations sur la structure et les caractéristiques physico-chimiques des molécules.

Le choix de la base des données expérimentales de référence est décisif dans une étude QSPR. Elle doit être composée de données expérimentales fiables obtenues en suivant un protocole expérimental unique. En effet, la robustesse du modèle dépend fortement de la base sur lequel il se fonde.

Enfin, le lien entre les descripteurs et la base de données est déterminé grâce à des outils d'analyse comme les régressions multilinéaires (MLR), les régressions aux moindres carrés partiels (PLS), les arbres de décisions, les réseaux de neurones (RNA), et les algorithmes génétiques.

Les informations extraites à partir des résultats d'étude de QSPR peuvent être utilisées pour obtenir une meilleure connaissance des structures moléculaires et probablement le mode d'action au niveau moléculaire.

Ces informations peuvent alors être utilisées pour prévoir les propriétés physico-chimiques et les activités biologiques de nouveaux composés ainsi que pour concevoir de nouvelles structures (figure II.1) [41].

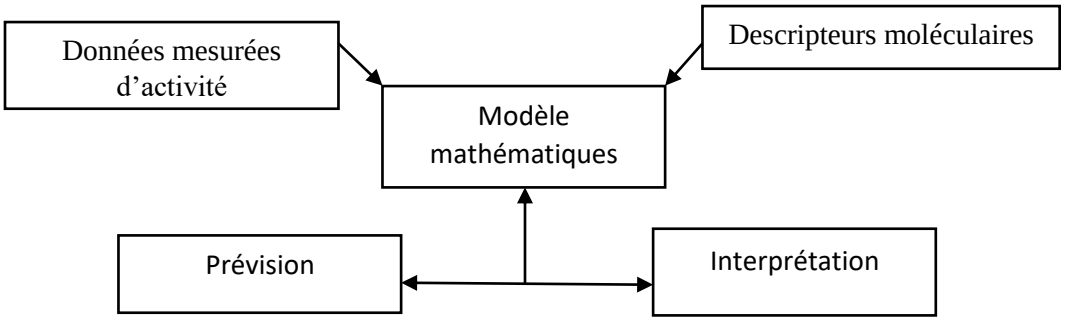


Figure II.1: Modèle d'étude quantitative des relations structure/propriété (QSPR).

II.2. Historique

L'exemple le plus ancien de RSA remonte à 1868, lorsque deux scientifiques britanniques, Alexander Crum-Brown et Thomas Fraser, publièrent une étude sur la relation entre la structure chimique et l'activité physiologique. Dans les décennies qui suivirent, Richet établit une relation entre la toxicité et les propriétés physicochimiques d'une série de composés organiques polaires (éthers, alcools, cétones), indiquant que « plus ils sont solubles, moins ils sont toxiques » [42,43].

De manière indépendante, Meyer et Overton ont décrit une corrélation linéaire entre la lipophile (coefficient de partage huile-eau) et les effets biologiques (narcotiques), la « théorie lipoïde de la narcose » [44,45].

En 1937, Hammett introduisit une méthode pour séparer les effets des substituants (propriétés donneuses ou capteuses d'électrons) sur les mécanismes des réactions de chimie organique [46]. Quelques années plus tard, Taft améliora le modèle de Hammett en proposant le premier paramètre stérique, ES [47].

En 1962, Hansch et coll publièrent leur étude princeps sur la corrélation de l'activité biologique de substances régulatrices de la croissance végétale (acides phénoxyacétiques monosubstitués sur le cycle benzénique) avec les constantes de substitution de Hammett et

l'hydrophobicité [48]. Deux ans plus tard, le même groupe de recherche introduisit un nouveau paramètre, π (hydrophobicité relative d'un substituant), qui peut être défini comme suit :

$$\pi X = \log Px - \log PH \quad (1)$$

Où πX représente l'hydrophobicité relative d'un substituant, et Px et PH les coefficients de partage respectifs de la molécule dérivée et de la molécule parente [49]. Cette méthode est à présent connue sous le terme de QSAR bidimensionnelle de Hansch-Fujita ou (2D-QSAR) [50,51].

En 1964, Free et Wilson ont suggéré l'approche dite "*de novo* " qui postule que chaque substituant apporte un effet additif et constant à l'activité biologique, quels que soient les autres substituants présents dans le reste de la molécule [52,53] :

$$BA = \Sigma\, ai\, xi + u \quad (2)$$

Où "BA" représente l'activité biologique, "u" la contribution moyenne de la molécule parente, ai la contribution de chaque motif structural et xi la présence (xi = 1) ou l'absence (**xi**= 0) d'un motif structural particulier.

Quelques années plus tard, les indices topologiques décrivant l'intégralité de la structure moléculaire ont été introduits. L'origine de cette terminologie remonte au début des années 1970 [54], bien que les premiers travaux sur les descripteurs constitutionnels numériques aient démarré pratiquement 25 ans en arrière avec Wiener [55]. Ces indices topologiques de première génération étaient des nombres entiers basés sur des invariants vertex locaux entiers, et intéressaient essentiellement les propriétés physiques des hydrocarbures saturés (alcanes). Parmi d'autres travaux importants, les indices topologiques de Randić [56], de Balaban [57], la différence topologique minimale de Simon [58] et l'approche de connectivité moléculaire selon Kier et Hall [59] constituèrent le socle nécessaire à l'avènement de l'ère moderne des QSAR / 3D-QSAR.

II.3. Généralités sur la modélisation QSPR

Une relation QSPR est un modèle ou une formule mathématique qui permet de relier, d'une manière quantitative, la structure d'une molécule à une propriété ou à une activité donnée. Les méthodes QSPR sont de plus en plus utilisées, du fait de la croissance des moyens de calculs (machines, logiciels, ...). Récemment, on assiste à la mise en place d'un nouveau règlement

REACH [60] (*Registration, Evaluation, Authorisation and Restriction of Chemicals)* qui recommande l'utilisation des méthodes alternatives pour limiter le recours à l'expérimentation.

En fait, les premiers travaux QSPR remontent au 19ème siècle. En effet, dès 1868, Crum-Brown et Fraser [61] ont postulé l'existence de relations entre les activités physiologiques et les structures chimiques en reliant les variations de l'activité biologique à des modifications structurales. Cependant, à cette époque, les structures moléculaires n'étaient pas encore connues.

Une avancée importante vers les modèles QSPR proprement dits a été réalisée grâce au développement des équations de Hammett [56] dans lesquelles les constantes σ caractérisent de manière quantitative les vitesses de réactions pour les composés organiques.

$$\log \frac{K}{K0} = \sigma\rho \qquad (3)$$

Ou K et K0 sont les constantes respectives de la réaction étudiée et de celle d'une référence et ρ une constante de réaction dépendant du type de réaction [62,63].

Les premiers travaux utilisant la méthodologie QSPR telle qu'employée actuellement sont dus à Hansch [64] et Free et Wilson [65]. D'un côté, Hansch a proposé des modèles reliant directement l'activité biologique des composés avec les propriétés hydrophobes, électroniques et stériques à l'échelle moléculaire. D'un autre côté, Free et Wilson ont développé des modèles empiriques, dits de contributions de groupes, pour l'étude de l'activité biologique.

Au cours de ces dernières décennies, l'utilisation des méthodes QSPR n'a pas cessé de progresser. Elle est même devenue indispensable en chimie pharmaceutique, en toxicologie et pour la conception de médicaments [66-67].

II.3. Principe des méthodes QSPR

Le principe des méthodes QSPR est d'établir une relation mathématique reliant de manière quantitative des propriétés moléculaires, appelées descripteurs, avec une observable macroscopique (activité biologique, toxicité, propriété physico-chimique, etc.), pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données. La forme générale d'un tel modèle est la suivante :

$$\text{Propriété/ Activité} = f(D1, D2, \ldots Dn, \ldots) \qquad (4)$$

D1, D2, …Dn sont des descripteurs des structures moléculaires.

L'objectif d'une telle méthode est d'analyser les données structurales afin de détecter les facteurs déterminants pour la propriété /activité mesurée. Pour ce faire, différents types d'outils statistiques peuvent être employés :

- Régressions linéaires simples et multiples [68].
- Régressions aux moindres carrées partielles (PLS) [69].
- Arbres de décision [70].
- Réseaux de neurones artificiels (RNA) [71,72].
- Algorithmes génétiques [73].
- Vecteurs Machines [74].

Une fois cette relation est établie et validée, elle peut alors être employée pour la prédiction de la propriété /activité de nouvelles molécules, pour lesquelles les valeurs expérimentales ne sont pas disponibles. De tels modèles peuvent être également utilisés pour mieux comprendre les mécanismes et les modes d'action.

II.4. Méthodologie générale d'une étude QSPR

La méthodologie générale d'une étude QSPR est la suivante :

- Constituer une base de données à partir des mesures expérimentales fiables de la propriété ou de l'activité de chaque composé.
- Sélectionner les descripteurs en relation avec la propriété ou l'activité étudiée.
- Diviser cette base de données, aléatoirement, en une série d'apprentissage (training set) qui contient généralement les 2/3 de la base de données et une série de test (test set) constituée par le 1/3 restant.
- Etablir des modèles mathématiques en utilisant la série d'apprentissage.
- Caractériser les modèles élaborés par leurs indices de validation internes et vérifier leur robustesse par un test d'hasardisation (Randomization) de la variable dépendante Y (réponse).
- Valider les modèles élaborés en utilisant la série de test et calculer leurs paramètres statistiques de validation externe.
- Elaborer le domaine d'applicabilité du modèle retenu.
- Explorer et exploiter les modèles validés pour comprendre les mécanismes et les modes d'action.

II.5. Descripteurs moléculaires

Les descripteurs moléculaires sont des propriétés numériques qui peuvent être calculées à partir de la table de connectivité d'une molécule ou d'une représentation planaire (2D) ou tridimensionnelle (3D) de la structure. Ils sont basés sur les éléments présents, les charges partielles, la nature des liaisons [75]. On peut calculer ainsi :

- Des propriétés physiques : polarisabilité, charge totale, réfractivité moléculaire, masse, densité, coefficient de partage eau/octanol (log P) …
- Des approximations d'aires de surfaces : surface de Van der Waals, surface accessible au solvant…
- Le dénombrement des atomes, liaisons et pivots ;
- Des indices de formes et de connectivité : indices de Kier et Hall ;
- Des descripteurs topologiques : matrices de distance et de connectivité ;
- Les dénombrements de fonctions pharmacophoriques : donneurs, accepteurs de liaison hydrogène, polaire (donneur et accepteur), positif, négatif, hydrophobe et autres … ;
- Les descripteurs de charges partielles : charge partielle positive (négative) totale, aire de la surface de Van der Waals positive (négative), aire de la surface de Vander Waals polaire, … ;
- Les indices de Hansch : r, s et p

II.6. Modèles statistiques

L'objectif de l'analyse statistique est justement de « démêler » ces descripteurs et d'identifier ceux qui sont corrélés à la variable cible, qui produisent du signal, de ceux qui ne le sont pas, qui produisent du bruit. L'analyse statistique permet également d'identifier les descripteurs qui sont corrélés entre eux pour ne garder que les principaux et réduire ainsi la redondance d'informations [76].

Dans la terminologie des statistiques, la propriété étudiée est appelée variable cible ou variable dépendante ou encore variable Y. C'est la caractéristique que l'on cherche à expliquer et à optimiser.

Les descripteurs, en l'occurrence chimiques, sont appelés variables explicatives, variables indépendantes ou encore variables X.

L'analyse statistique détermine et quantifie les corrélations entre les descripteurs et la variable cible. Elle indique également la contribution relative de chaque descripteur dans l'explication globale de l'activité. Le modèle statistique est une équation donnant la valeur de la variable cible en fonction de la somme des valeurs pondérées des descripteurs.

Les principaux outils statistiques pour obtenir un modèle sont :

- La régression linéaire multivariée (*Multivariate Linear Regression* -MLR) ;
- La régression en composante principale (*Principale Componant Regression* - PCR) ;
- La régression des moindre carrés partiels (*Partial Least Squares* - PLS) ;
- Les réseaux de neurones artificiels (*Artificial Neural Network* - ANN).

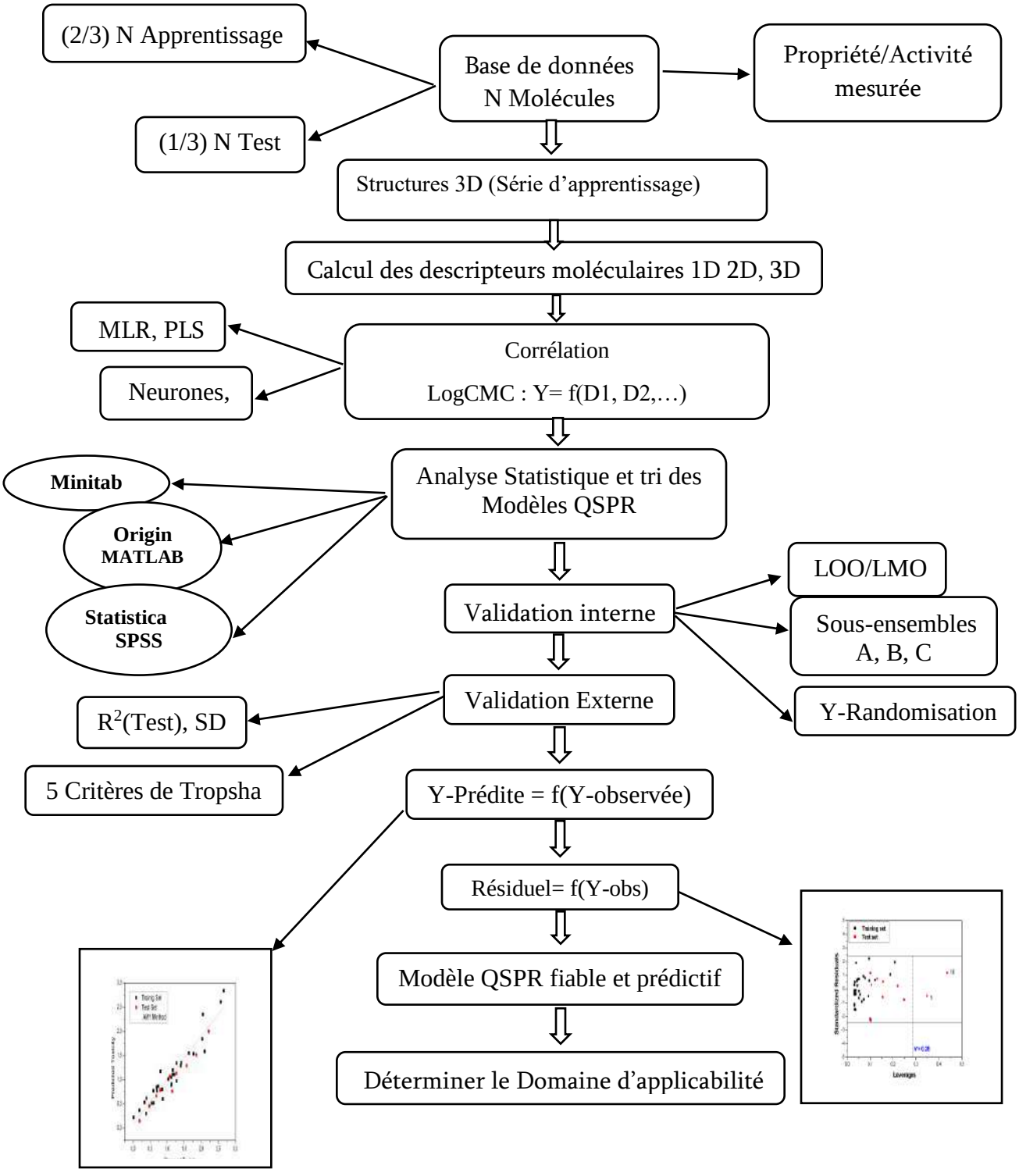
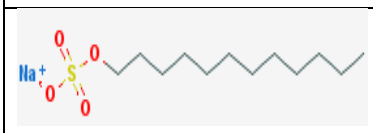
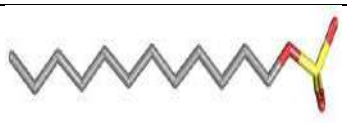
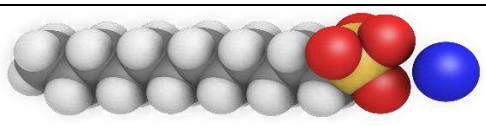


Figure.II.2. Schéma d'élaboration et validation d'un modèle QSPR [77]

II.7. Descripteurs moléculaires

De nombreuses recherches ont été menées, au cours de ces dernières décennies, pour trouver la meilleure façon de représenter l'information contenue dans la structure des molécules et ces structures elles-mêmes en un ensemble de nombres réels appelés descripteurs qui sont soit théoriques ou empiriques :

Tableau II.1 : Sodium Dodecyl Sulfate en différant démentions

1D	2D	3D
		

II.7.1. Les descripteurs moléculaires théoriques

II.7.1.1. Les descripteurs 1D

Sont accessibles à partir de la formule brute de la molécule et décrivent des propriétés globales du composé comme le nombre d'atomes et la masse moléculaire, …etc.

Ces descripteurs sont couramment utilisés du fait de leur extrême simplicité. Cependant, ils peuvent poser problème pour une bonne interprétation des mécanismes d'interaction du fait qu'ils ne permettent pas de tenir en compte des effets stériques et d'isomérie.

II.7.1.2. Les descripteurs 2D

Sont calculés à partir de la formule développée de la molécule. On distingue :

- **Les indices 2D constitutionnels** : qui caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles, …etc.
- **Les indices 2D topologiques** : peuvent être obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Exemples : indice de Wiener [78], indice de Randić [79], indice de connectivité de valence de Kier-Hall [80], indice de Balaban [81], …etc.

Ces descripteurs 2D permettent de prédire les propriétés physiques mais sont insuffisantes pour expliquer certaines propriétés et activités biologiques comme la toxicité.

II.7.1.3. Les descripteurs 3D

Sont évalués à partir des positions relatives des atomes dans l'espace, et décrivent des caractéristiques plus complexes. Leurs calculs nécessitent donc de connaître la géométrie 3D de la molécule.

- **Les descripteurs 3D géométriques** : les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie.
- **Les descripteurs 3D électroniques** : permettent de quantifier les différents types d'interactions inter- et intramoléculaires, de grande influence sur l'activité biologique des molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie est minimale, et fait souvent appel à la chimie quantique.

II.8. Méthodes d'analyse des données :

Pour élaborer un modèle QSPR nous avons besoin d'une méthode d'analyse de données, cette méthode permet de quantifier la relation qui existe entre la propriété/Activité et la Structure (descripteurs).

Il existe plusieurs méthodes pour construire un modèle et analyser les données statistiques de ce dernier, certaines sont linéaires telles que la régression linéaire multiple (MLR), la régression aux moindres carrés partiels (PLS), d'autres sont non linéaires comme les arbres de décisions, les réseaux de neurones… ces méthodes sont disponibles dans des logiciels tels que, MATLAB, Excel, Origin Microcal, Minitab, Statistica, SPSS, R, …

La méthode utilisée dans nos études est la méthode des réseaux de neurone artificielle (RNA).

II.9. Interprétation et validation d'un modèle QSPR

Une fois développé, le modèle doit être interprété en analysant tous les paramètres statistiques de ce modèle, sa qualité doit être aussi étudiée, cette qualité est vérifiée par ce que l'on appelle validation. Sa robustesse, c'est-à-dire l'influence des composés de la série d'apprentissage sur le modèle, est estimée par des méthodes de validation interne. Afin d'estimer son pouvoir prédictif, des données expérimentales supplémentaires sont nécessaires afin de déterminer la capacité du modèle à prédire ces valeurs c'est ce que l'on appelle validation externe. Enfin, il est important de savoir quel type de molécules utilisées avec quel modèle. On parle alors de domaine d'applicabilité.

II.9.1. Validation interne

Dans le passé, la validation interne d'un modèle QSPR a été réalisée en utilisant la validation croisée LOO (leave-one out) ou LMO (leave-many out). Ce processus consiste à extraire un certain nombre k de molécules du jeu initial à N molécules et à construire un nouveau modèle avec les (N-k) molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent). Ce processus est ensuite réitéré pour retirer et prédire les valeurs de toutes les molécules de la série d'apprentissage. En fonction du nombre de molécules retirées à chaque itération, on parlera de Leave-One-Out (LOO) ou de Leave-Many-Out (LMO) selon qu'une ou plusieurs molécules est (sont) retirée(s) [82]. Dans ces dernières années, d'autres méthodes sont utilisées pour faire la validation interne, tel que l'hasardisation de la réponse (Y-Randomization).

II.9.2. Y-Randomization

Afin de s'assurer qu'un modèle QSPR est fiable, les tests de Y-randomization [82] sont une des techniques les plus employées. En effet, il n'est pas rare d'obtenir des corrélations fortuites (ou « chance corrélation »),

Ces modèles aléatoires peuvent être détectés par la procédure Y-randomization. Elle consiste à mélanger aléatoirement les propriétés/activités expérimentales pour la série d'apprentissage en utilisant les mêmes descripteurs, de nouveaux modèles sont obtenus. Ces derniers doivent avoir des performances très faibles. Cependant, la validation interne est insuffisante pour étudier le pouvoir prédictif d'un modèle. Pour cette raison la validation externe du modèle est devenue une norme et une partie obligatoire dans la modélisation QSPR [83, 84].

II.9.3. Validation externe

Cette méthode consiste à prédire la propriété/activité d'une série de molécules appelée généralement série de test qui ne sont pas dans la série de développement du modèle, cette validation est caractérisée par les paramètres R^2 (test) R^2cv (test). Récemment plusieurs études [85, 86] ont montré l'insuffisance des paramètres R^2, R^2_{CV} pour vérifier le pouvoir prédictif des modèles QSPR. Par conséquent, d'autres paramètres doivent être vérifiés pour cet objectif. Ces paramètres sont connus sous le nom "critères de validation externe" ou souvent appelés "critères de Trophsa" [78].

II.10. Domaine d'applicabilité

Comme les modèles QSPR sont utilisés pour la prise de décisions relatives aux augmentations de gestion des produits chimiques, la fiabilité des prévisions du modèle est une préoccupation croissante.

Un modèle QSPR ne peut pas être considéré comme un modèle universel, parce qu'il est développé sur un nombre limité de composés qui ne couvrent pas tout l'espace chimique. Par conséquent l'activité/propriété prédite d'un composé, chimiquement dissimilaire au jeu d'apprentissage, ne pourra pas être considérée fiable [87, 88]. Le domaine d'applicabilité (DA) permet de définir la zone dans laquelle un composé pourra être prédit avec confiance. Le DA correspond donc à la région de l'espace chimique incluant les composés du jeu d'apprentissage et les composés similaires, proches dans ce même espace **[**89]. Ils existent plusieurs méthodes pour la détermination de domaine d'applicabilité d'un modèle QSPR parmi ces méthodes on trouve la méthode de « leverage ».

II.11. Applications des méthodes QSPR

Les applications des méthodes QSPR sont très nombreuses, elles touchent tous les domaines ou la structure chimique intervient, entre autres on peut citer :

II.11.1 Propriétés physico-chimiques

Point d'ébullition, Point de fusion, densité, indice de réfraction, température critique, viscosité, solubilité, pression de vapeur, tension superficielle, Coefficients de partition : eau/octanol, air/eau, huile/air, lait/plasma [90].

II.11.2. Activités biologiques

Anti VIH, Anti malaria, Anti Diabète, Anti Cancer, Anti-oxydante, Antiinflammatoire…

II.11.3. Autres propriétés/activités

- Prédiction de la toxicité aquatique des composés chimiques vis-à-vis des espèces environnementales.
- Toxicité des nanoparticules
- Toxicité des pesticides et des colorants
- Propriétés inhibitrices de corrosion
- **Concentration micellaire critique**

- Prédiction de plusieurs propriétés dangereuses telle que l'explosibilité et l'inflammabilité de certaines familles de molécules chimiques.
- Conception des médicaments et de nombreux autres produits tels que les agents tensio-actifs, parfums, les colorants et les produits de la chimie fine.…

II.12. Régression linéaire multiple (MLR)

Une analyse de régression où la variable dépendante *Y* dépend linéairement de plusieurs variables indépendantes *X*1, *X*2, . . ., *Xj* est appelée régression linéaire multiple.

L'équation de régression linéaire multiple est de la forme :

Y = *f* (*X*1, *X*2, . . ., *Xj*) où *f (X1, X2, . . ., Xj)* est une fonction linéaire de *X*1, *X*2, . . ., *Xj* [80].

Le modèle mathématique peut se présenter sous la forme :

$$\boldsymbol{Y}\mathbf{i} = \boldsymbol{a} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{j=P}} \boldsymbol{bj}\, \boldsymbol{x}_{\boldsymbol{j}}^{\boldsymbol{i}}$$

Où (b_j) sont les coefficients de régression et (a) la constante de régression.

II.12.1. Description de la méthode

La régression multilinéaire est la méthode la plus simple et la plus employée pour le développement de modèles prédictifs [65]. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante Y (ici, la propriété) et une série de p variables indépendantes Xi (ici, les descripteurs). L'objectif est d'obtenir une équation de la forme suivante :

$$Y = \beta 0 + \beta 1\ X1 + \beta 2\ X2 + \beta 3\ X3 + \ldots + \beta p\ Xp + \varepsilon \qquad (1)$$

Où **βi** sont les coefficients de la régression.

La détermination de l'équation (1) se fait alors à partir d'une base de données de n échantillons pour laquelle a la fois les variables dépendantes et la variable indépendante sont connues. Il s'agit donc de considérer un système de p équations.

$$Y1 = \beta0 + \beta1\ X1,1 + \beta2\ X2,1 + \ldots + \beta p\ Xn,1 + \varepsilon1$$

$$\begin{array}{ccccccc} . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{array}$$

$$Y2 = \beta0 + \beta1\ X1,2 + \beta2\ X2,2 + \ldots + \beta p\ Xn,2 + \varepsilon2$$

$$Yn = \beta0 + \beta1\ X1,p + \beta2\ X2,p + \ldots + \beta p\ Xn,p + \varepsilon n$$

Où les résidus εi représentent l'erreur du modèle, constituée par l'incertitude sur la variable dépendante Yi d'une part, sur les variables indépendantes Xi d'autre part, mais aussi par les informations contenues dans les variables indépendantes mais non exprimées via les variables dépendantes.

Ce système d'équations peut être écrit sous la forme matricielle suivante :

$$\begin{pmatrix} Y1 \\ Y2 \\ . \\ . \\ . \\ . \\ . \\ Yn \end{pmatrix} = \begin{Bmatrix} 1 & X1,1 & X1,2 & \ldots & \ldots & \ldots & X1,P \\ . & . & . & & & & . \\ . & . & . & . & & & . \\ . & . & . & & . & & . \\ . & . & . & & & . & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ 1 & Xn,1 & Xn,2 & \ldots & \ldots & \ldots & Xn,p \end{Bmatrix} \begin{Bmatrix} \beta1 \\ \beta2 \\ . \\ . \\ . \\ . \\ . \\ \beta n \end{Bmatrix} + \begin{Bmatrix} \varepsilon1 \\ \varepsilon2 \\ . \\ . \\ . \\ \varepsilon n \end{Bmatrix}$$

la matrice X*(n,p)* comporte dans sa première colonne un vecteur constitue uniquement de 1.

Ce vecteur correspond à la constante *X0*. La matrice X est alors de dimension *(n, p+1)*. Le modèle s'écrit alors sous forme $Y = X\ b + \varepsilon$.

La méthode consiste alors à choisir les coefficients du vecteur β en faisant en sorte de minimiser la somme des carrés des écarts entre les valeurs prédites et les valeurs réelles sur l'intégralité de la base de données et ceci sous couvert de certaines hypothèses de départ.

En premier lieu, les variables indépendantes Xi, comme leur nom l'indique, sont supposées indépendantes entre elles et leur incertitude est négligeable. Ensuite, les différents échantillons Yi sont supposés indépendants entre eux. Enfin, par nature, la dépendance de Y vis-à-vis des Xi est supposée linéaire.

La valeur prédite de la variable dépendante est alors :

$$\hat{Y} = \beta 0 + \hat{\beta} 0\, X1, i + \cdots + \hat{\beta} p\, Xn, i \qquad (2)$$

Les résidus peuvent donc être définis comme la différence entre les valeurs prédites et observées de Y.

$$\varepsilon = Yi - \hat{Y}i \qquad (3)$$

Il s'agit alors de trouver les coefficients $\widehat{\boldsymbol{Bi}}$ afin de minimiser la somme des carrés de ces résidus (*RSS*) pour l'intégralité de la base de données.

II.13. La régression des moindres carrés partiels (PLS)

La régression PLS (*Partial Least Squares régression*) est une technique d'analyse et de prédiction relativement récente. Elle a été conçue pour faire face aux problèmes résultants de l'insuffisance de l'utilisation de la régression linéaire classique, qui trouve ses limites dès lors que l'on cherche à modéliser des relations entre des variables pour lesquelles il y a peu d'individus, ou beaucoup de variables explicatives en comparaison au nombre d'individus (le nombre de variables explicatives pouvant excéder très largement le nombre d'individus), ou encore lorsque les variables explicatives sont fortement corrélées entre la régression PLS.

Hypothèse préliminaire : Variables explicatives X continues et qualitatives. Variable réponse Y continue, binaire ou polychromique $K \geq 2$.

II.13.1. L'Algorithme de la régression des moindres carrés partiels (PLS)

De manière générale, les données continues doivent être centrées-réduites. L'Algorithme présenté ci-dessous, inspiré de l'algorithme NIPALS, implémente le calcul étape par étape des composantes D_h. En fonction de la version de la régression (logistique ou linéaire) PLS, la méthode de modélisation varie. A noter que la régression PLS nécessite le paramétrage du nombre H de composantes à calculer [90]

A. **Étape 1** : Calcul de D_1

- Calcul du coefficient ω_1^j, associé à la variable X^j, par régression logistique ou linéaire variée de Y. $\forall j \in \{1, \ldots \ldots, P\}$
- Normalisation du vecteur des coefficients ω_1 en.

$$\omega_{1,N} = \frac{\omega 1}{\sqrt{\sum_{j=1}^{P}\left(\omega_1^j\right)^2}} \quad (4)$$

- Calcul de la composante.

$$D_1 = \frac{X.\omega_{1,N}}{\omega_{1,N}^t.\omega_{1,N}} \quad (5)$$

- Détermination du vecteur des coefficients $\omega_{1,*}$ par régression linéaire de D_1 sur X sous contrainte que $\left|\omega_{1,*}\right| = 1$ et sans estimation du paramètre constant.
- Calcul de la matrice résiduelle X_1 via la formule :

$$X_1 = X - D_1.\left(\frac{D_1^t.X}{D_1^t.D_1}\right) \quad (6) \ .$$

- Pour la version linéaire de la régression PLS exclusivement, nous devons également calculer le vecteur résiduel Y_1 via la formule :

$$Y_1 = Y - D_1.\left(\frac{D_1^t.Y}{D_1^t.D_1}\right) \quad (7)$$

B. **Étape $h \in [2, H]$:** Calcul de D_h

- Calcul du coefficient D_h^j associé à la variable $X_{h-1}^{\ j}$ dans la régression multiple logistique de Y ou liéaire de Y_{h-1} sur les variables $D_1, \ldots\ldots, D_{h-1}, X_{h-1}^{\ j}$, $\forall j \in \{1, \ldots\ldots, P\}$.
- Normalisation du vecteur des coefficients ω_h en.

$$\omega_{h,N} = \frac{\omega h}{\sqrt{\sum_{j=1}^{P}\left(\omega_h^j\right)^2}} \quad (8)$$

- Calcul la composante.

$$D_h = \frac{X_{h-1}.\omega_{h,N}}{\omega_{h,N}^t.\omega_{h,N}} \quad (9)$$

- Détermination du vecteur des coefficients $\omega_{h,*}$ par régression linéaire de D_h sur X sous contrainte que $\left|\omega_{1,*}\right| = 1$ et sans estimation du paramètre constant.
- Calcul de la matrice résiduelle X_h via la formule :

$$X_h = X_{h-1} - D_{h-1} \cdot \left(\frac{D_1^t . X_{h-1}}{D_h^t . D_h} \right) \qquad (10)$$

- Pour la version linéaire de la régression PLS exclusivement, nous devons également calculer le vecteur résiduel Y_h via la formule :

$$Y_h = Y_{h-1} - D_h \cdot \left(\frac{D_h^t . Y_{h-1}}{D_h^t . D_h} \right) \qquad (11)$$

C. **Étape finale** : Calcul des coefficients de la règle décisionnelle

Nous nous retrouvons à la fin de la H étape de l'algorithme avec nos composantes $D_1, \ldots, D_H$ et nos vecteurs de coefficients $\omega_{1,*}, \ldots, \omega_{H,*}$.

- Dans un premier temps nous estimons les coefficients $\alpha_0, \ldots, \alpha_h$ par régression logistique ou linéaire standard de Y sur $D_1, \ldots, D_H$.

Nous avons alors le coefficient $\beta_0 = \alpha_0$ et sera conservé comme le coefficient constant du modèle final.

- Enfin, la détermination des coefficients finaux $\beta_1, \ldots, \beta_P$ se fait en développant l'équation suivante :

$$\left(\beta_1, \ldots, \beta_P\right) = \alpha_1 \dot{\omega}_{1,*}, \ldots, \alpha_h \omega_{h,*} \qquad (12)$$

La règle décisionnelle est la même que celle de la régression logistique ou linéaire, en fonction de la version choisie, et s'exprime en fonction du vecteur des coefficients $\left(\beta_0, \ldots, \beta_P\right)$.

II.13.2. Caractéristiques

- Le calcul des composantes $D_h, h \in \{1, H\}$ est analogue à celui des composantes principales de l'ACP à ceci près que dans le cas de la régression PLS elles sont déterminées sous contrainte de maximiser la covariance entre les variables explicatives et la variable réponse.
- La problématique de la multi colinéarité est traitée en deux temps : lors de la phase de régression uni variée qui revient à calculer les coefficients « individuellement » de chaque variable et donc à retenir l'information la plus brut qui soit. Et lors de la régression sur les composantes $D_h, h \in \{1, H\}$ qui synthétisent l'information maximisée par projection orthogonale et la fusionne.
- La gestion du problème de type P > N est traitée lorsque de la phase de régression uni variée et limite ainsi les risques de sur-apprentissage du modèle [91].
- En pratique, le nombre de composantes D_h nécessaires et suffisantes est rarement supérieur à trois ou quatre. Au-delà les performances n'évoluent quasiment plus. Ce dernier est choisi soit par validation-croisée soit par maximisation des performances.
- L'Algorithme PLS est très rapide étant donné qu'il consiste en une succession de régressions d'où son efficacité sur de grandes bases de données et ne nécessite aucune opération matricielle complexe [92].
- La régression PLS tend à devenir la modélisation dominante et reste généralement plus efficace que la régression logistique, la régression linéaire, la régression sur composantes principales et la régression Ridge.
- Dans le cas de la régression logistique, les coefficients β donne la valeur des odds ratios lorsque nous leur appliquons la fonction exponentielle. Le calcul des intervalles de confiance pour le rejet de l'hypothèse $\beta_j = 0$ se fait par bootstrap [93, 94].

II.13.3. La régression des moindres carrés partiels (PLS)

Nous présentons ici la démonstration de la recherche de la première composante D_1.

Nous avons vu que la composante D_1 peut s'écrire de la manière suivante [95],

$$D_1 = X\omega_{1,*} = X\omega_1$$

Elle est alors obtenue en maximisant le critère de Tucker,

$$\mathrm{cov}^2\left(Y, X\omega_1\right) = r^2\left(Y, X\omega_1\right)\mathrm{Var}\left(X\omega_1\right)\mathrm{Var}(Y) \qquad (13)$$

Sous contrainte que $\left\|\omega_1\right\| = 1$.

L'idée est d'essayer de maximiser parallèlement la variance expliquée par D_1 et la corrélation entre $D_{1,Y}$ tel que,

$$< D_1, Y >==< X\omega_1, Y \geq \left\|X\omega_1\right\| . \|Y\|. \mathrm{cor}(X\omega_1, Y) \qquad (14)$$

Le vecteur des coefficients ω_1 est déterminé à l'aide des multiplicateurs Lagrangien,

$$\lambda \in \mathbf{R}$$

$$L\left(\omega_1, \lambda\right) = \mathrm{cov}\left(Y, X\omega_1\right) - \lambda\left(\boldsymbol{\omega}_1^t\omega_1 - 1\right) = \boldsymbol{\omega}_1^t X^t Y - \lambda\left(\boldsymbol{\omega}_1^t\omega_1 - 1\right) \qquad (15)$$

Il faut ensuite annuler les dérivées,

$$\frac{\partial L}{\partial \lambda} = -\left(\boldsymbol{\omega}_1^t\omega_1 - 1\right) = 0 \Rightarrow \boldsymbol{\omega}_1^t\omega_1 = 1 \qquad (16)$$

$$\frac{\partial L}{\partial \omega_1} = X^t Y - 2Y\omega_1 = 0 \Rightarrow X^t Y = 2Y\omega_1 \qquad (17)$$

Nous avons alors,

$$\boldsymbol{\omega}_1^t X^t Y = \boldsymbol{\omega}_1^t \mathbf{2}\lambda\,\omega_1 = \mathbf{2}\lambda\,\boldsymbol{\omega}_1^t\omega_1 = \mathbf{2}\lambda \qquad (18)$$

Maintenant, soit $\theta \in R$ tq :

$$\theta = \mathbf{2}\lambda = \boldsymbol{\omega}_1 Y^t X \Rightarrow \left(X^t Y\right)\left(\boldsymbol{\omega}_1 Y^t X = \theta\boldsymbol{\omega}_1\right)\theta = \theta^2\boldsymbol{\omega}_1 \qquad (20)$$

Nous en déduisons que $\boldsymbol{\omega}_1$ est le vecteur propre de $X^t YY^t X$ associé à la valeur propre θ^2

Nous pouvons déduire de ce résultat que,

$$\lambda_{\mathbf{1}} =< \omega_1 XY >^2 = YX^t XY^t \qquad (21)$$

Par conséquent,

$$\Rightarrow \boldsymbol{\omega}_1 = X^t Y \qquad (22)$$

La division par la norme est alors obligatoire sous contrainte fixée sur ω_1.

II.14. Réseaux de neurones artificiels (RNA)

II.14.1. Structure des réseaux de neurones artificiels

McCulloch et Pitts en 1943 ont implémenté un système de réseaux neuronaux artificiels, qui est analogue aux neurones biologiques fondée sur une structure complexe. Le système des RNA est considéré comme un arrangement d'éléments de structure identique appelés neurones interconnectés par analogie avec cellules du système nerveux humain. Il est composé également d'une succession de couches connectées de manière à ce que chaque neurone tienne son entrée de la sortie du neurone précédant. Chaque neurone dans ce cas fonctionne indépendamment par rapport aux autres afin que 1 'ensemble forme un système compact. L'information est emmagasinée de façon répartie dans le réseau sous forme de coefficients synaptiques. Le neurone formel calcule régulièrement un résultat qu'il transmet ensuite aux neurones suivant, chaque calcul est associé à un poids qui définit la force de la connexion [96].

Tableau II.2 : Analogie entre le neurone biologique et le neurone formel

Neurone artificiel	**Neurone biologique**
Poids de connexion	Synapses
Signal de sortie	Axones
Signal d'entrée	Dendrite
Fonction d'activation	Soma

Mathématiquement, tel qu'illustré par la figure II.3, chaque neurone reçoit des entrées sous forme vectorielle puis il calcule une somme pondérée de ses entrées pour que le résultat passe ensuite par la fonction d'activation afin de créer une sortie.

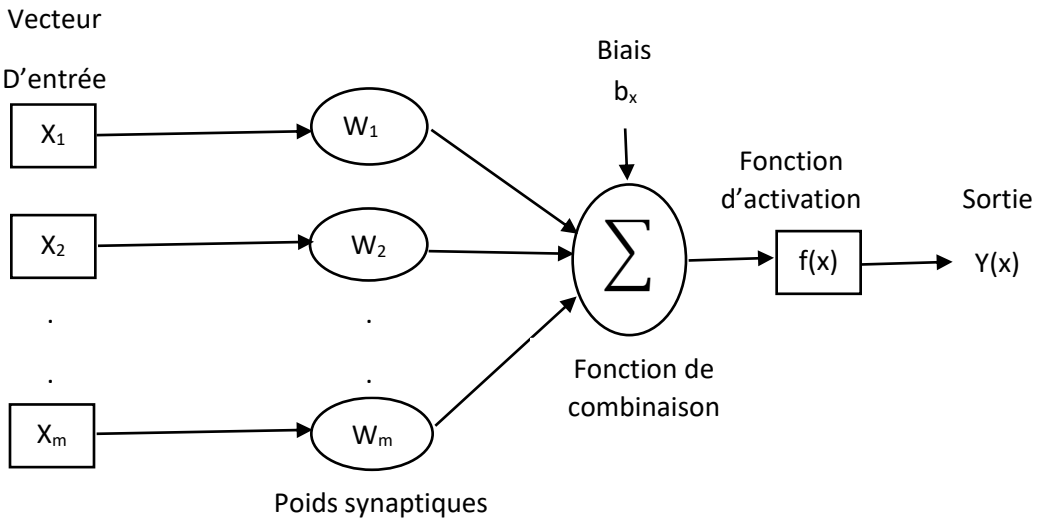


Figure II.3 : Structure d'un neurone artificiel [97]

II.14.2. Fonction d'activation

La fonction d'activation (ou fonction de transfert) sert à convertir le résultat de la somme pondérée des entrées d'un neurone en une valeur de sortie, cette conversion s'effectue par un calcul de l'état du neurone en introduisant une non-linéarité dans le fonctionnement du neurone [95]

Le biais b joue un rôle de seuil, quand le résultat de la somme pondérée dépasse ce seuil, l'argument de la fonction de transfert devient positif ou nul ; dans le cas contraire, il est considéré négatif. Finalement si le résultat de la somme pondérée est :

1. En dessous du seuil, le neurone est considéré comme non-actif
2. Aux alentours du seuil, le neurone est considéré en phase de transition.
3. Au-dessus du seuil, le neurone est considéré comme actif

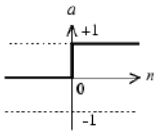
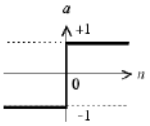
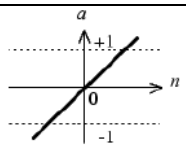
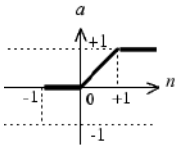
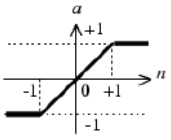
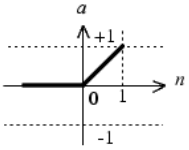
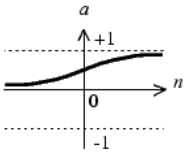
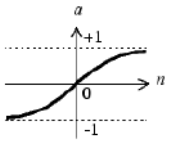
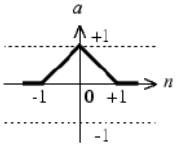
Il y a plusieurs types de fonctions de transfert qui peuvent être utilisées dans les RNA, les fonctions d'activation souvent utilisées sont représentées ci-dessous :

II.14.2. La fonction non linéaire sigmoïde

La fonction non linéaire sigmoïde est fréquemmt utilisée dans les RNA, particulièrement dans les réseaux utilisant 1 'algorithme de rétro-propagation **[98]** car contrairement à la fonction sigmoïde, les antre fonctions donnent seulement une sortie binaire ce qui rend plus difficile à estimer les poids optimaux.

La fonction sigmoïde est définie dans le tableau II.3

Tableau II.3: Fonctions de transfert usuelles, *a=f(n)*

Nom de la fonction	**Relationentrée / sortie**	**Graphe de la function**	**Fonction MATLAB® correspondante**
Seuil	$a = 0$ si $n < 0$ $a = 1$ si $n \geq 0$		hardlim
Seuil symétrique	$a = -1$ si $n < 0$ $a = 1$ si $n \geq 0$		hardlims
Linéaire	$a = n$		purelin
Linéaire saturée	$a = 0$ si $n < 0$ $a = n$ si $0 \leq n \leq 1$ $a = 1$ si $n > 1$		Satlin
Linéaire saturéesymétrique	$a = -1$ si $n < -1$ $a = n$ si $-1 \leq n \leq 1$ $a = 1$ si $n > 1$		Satlins
Linéaire positive	$a = 0$ si $n < 0$ $a = n$ si $n \geq 0$		Poslin
Logistique (Sigmoïde logarithmique)	$a = \frac{1}{1+e^{-n}}$		Logsig
Tangente hyperbolique	$a = \frac{e^{+n} - e^{-n}}{e^{+n} + e^{-n}}$		Tansig
Triangulaire de base	$a = 1 - \lvert n \rvert,\ -1 \leq n \leq +1$ $a = 0,\ autrement$		Tribas

II.14.3. Apprentissage d'un réseau de neurone

L'apprentissage est une étape très importante du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié itérativement jusqu'à l'obtention du comportement désiré, et ce par l'ajustement des poids (connexion ou synapse) des neurones à une source d'information bien définit [99,100]. L'apprentissage consiste également à extraire des conformités parmi les données utilisées pour 1' entrainement du réseau mais l'objectif essentiel de l'apprentissage est la résolution du problème par la prévision, la classification ... etc [101].

Pour un RNA, l'apprentissage peut être regardé également comme étant le processus de la mise à jour des poids (connexion) au sein du réseau dans le but d'ajuster la réponse du réseau à l'expérience et aux exemples.

Il existe plusieurs types de règles d'apprentissage qui peuvent être rassemblées en deux catégories qui sont l'apprentissage supervisé et l'apprentissage non supervisé.

II.14.3.1. Apprentissage supervisé

Dans ce type d'apprentissage, on cherche à imposer au réseau un fonctionnement donné en forçant les sorties des réseaux à prendre des valeurs bien données (choisie par l'opérateur) et ce en modifiant les poids synaptiques.

Le réseau se comporte alors comme un filtre dont les paramètres de transfert sont ajustés à partir des couples entrée-sortie présentés [102].

L'adaptation des paramètres du réseau s'effectue à partir d'un algorithme d'optimisation, l'initiation des poids synaptiques étant le plus souvent aléatoire.

Des exemples d'utilisations de différentes règles d'apprentissage, architectures et algorithme qui sont relié à ce type supervisé sont illustrés dans le tableau II.4

Tableau II.4 : Exemple de différents types d'apprentissages en mode supervisé

Paradigme	Architecture	Règle d'apprentissage	Algorithme	Taches
Supervisé	Perception Simple où Multicouches	Correction d'erreur	Perceptron, Rétro-Propagation, Adaline, Madaline	Classification, Approximation de fonctions, Prédiction, Contrôle
	Récurrente	Bolzmann	Classification	Apprentissage de Boltzmann
	Multicouches non bouclés	Hebb	Analyse de Discriminants Linéaires	Analyse de données Classification
	à compétition	Par compétition	LVQ	Catégorisation au sein d'une classe, Compression de Données
	ART		ARTMap	Classification, Catégorisation au sein d'une classe

II.14.3.2. L'algorithme d'apprentissage

Définir l'architecture d'un RNA pour un problème quelconque revient à faire un choix judicieux de la taille du réseau, du nombre total de couches et de neurones, distribution des données et des fonctions de transfert [103]. Le choix de ces paramètres dépend de l'utilisateur. Il n'existe pas dans la littérature pour le moment des données suffisantes qui peuvent déterminer clairement les paramètres à adopter pour résoudre un problème donné [104-107]. Les différentes techniques présentées par différents auteurs sont destinées à résoudre des cas particuliers [108-113].

Afin de répondre à la problématique de ce mémoire, telle que la prédiction des paramètres géométriques des conducteurs géologiques recherchés (profondeur, pendage, conductance, longueur, largeurs d'un conducteur) nous avons choisi l'architecture de perceptrons multicouches PMC **[**114] en utilisant un algorithme de rétro-propagation avec fonction sigmoïde comme fonction de transfert.

Ce choix d'utiliser le PMC et l'algorithme rétro-propagation est dû principalement à la popularité de ces deux derniers dans la gestion des relations non linéaire entre les variables, ce qui est le cas pour notre étude du problème électromagnétique.

II14.3.3. L'algorithme de rétro-propagation

L'algorithme d'apprentissage de rétro-propagation consiste dans un premier temps à circuler vers l'avant les données d'entrées jusqu'à l'obtention d'une entrée calculée par le réseau, puis la seconde étape est de comparer la sortie calculée à la sortie réelle connue **[115].** Les poids sont modifiés de telle sorte qu'à la prochaine itération, l'erreur commise entre la sortie calculée est minimisée, en prenant en considération la présence des couches cachées, l'erreur est rétro-propagé vers l'arrière jusqu'à la couche d'entrée tout en modifiant la pondération.

Le processus est répété sur tous les exemples jusqu'au temps où l'on obtienne une erreur de sortie considérée comme négligeable.

Mathématiquement on peut résumer l'algorithme de rétro-propagation comme suit : on considère un ensemble d'apprentissage composé de K paire de vecteur

$$\left(X_1, O_1\right), \left(X_2, O_2\right), \ldots\ldots\ldots, (X_k, O_k)$$

Avec un vecteur d'entrée :

$$X_p = \left(X_{p,o,1}, X_{p,o,2} \ldots\ldots\ldots\ldots X_{p,o,n}\right)^t \in R^n \qquad (23)$$

Et un vecteur de sortie désiré :

$$O_p = \left(O_{p,1}, O_{p,2} \ldots\ldots\ldots\ldots O_{p,m}\right)^t \in R^m \qquad (24)$$

et un vecteur de sortie réel du réseau :

$$y_p = \left(y_{p,1,1}, X_{p,1,2} \ldots\ldots\ldots\ldots y_{p,l,m}\right)^t \in R^m \qquad (25)$$

On définit 1 'entrée totale du K nœud pour la couche j par :

$$h_{p,j,k} = \sum_{i=0}^{n} W_{j,k,i} X_{p,j-1,i} \quad (26)$$

Alors la sortie de ce nœud sera :

$$F\left(h_{p,j,k}\right) = \sum_{i=0}^{n} W_{j,k,i} X_{p,j-1,i} + \theta_{j,k} \quad (27)$$

Où

$W_{j,k,i}$: la connexion entre le neurone k de la couche j-1 et le neurone i de la couche j.

$h_{p,j,k}$: l'entrée totale du neurone k pour l'échantillon p de la couche j.

$\theta_{j,k} = W_{j,k,o}$: le poids fictif du neurone k de la couche j

On pourra alors définir le terme d'erreur total commit sur tous les nœuds de sortie par :

$$E_p = \frac{1}{2}\sum_{i=0}^{m} \delta_{p,k}^{2} = \frac{1}{2}\sum_{i=0}^{m} (O_{p,k} - x_{p,l,k})^2 \quad (28)$$

Où $\delta_{p,k} = (O_{p,k} - x_{p,l,k})$ est l'erreur commise sur le K^{ème} nœud de sortie.

Afin de minimiser l'erreur Ep, le gradient est calculé par rapport à chaque poids w, puis à leurs tour les poids sont modifiés dans le sens inverse du gradient.

La mise à jour des poids de la couche de sortie est décrite par les équations suivantes :

$$\nabla E_p = \frac{\partial E_p}{\partial w_{l,k,j}} = -\left(O_{p,k} - x_{p,l,k}\right)\frac{\partial\left[f\left(y_{p,l,k}\right)\right]}{\partial w_{l,k,j}} \quad (29)$$

$$\nabla E_p == -\left(O_{p,k} - x_{p,l,k}\right)\frac{\partial f\left(y_{p,l,k}\right)\partial\left(y_{p,l,k}\right)}{\partial w_{l,k,j}\,\partial\left(y_{p,l,k}\right)} \quad (30)$$

$$\begin{cases} \frac{\partial y_{p,l,k}}{\partial w_{l,k,j}} = \frac{\partial}{\partial w_{l,k,j}} \sum_{j=0}^{m} W_{l,k,j} - X_{l,k,j} = X_{p,l-1,j} \\ \frac{\partial f(y_{p,l,k})}{\partial y_{l,k,j}} = \acute{f}(y_{p,l,k}) \end{cases} \quad (31)$$

$$\nabla E_p = \frac{\partial E_p}{\partial w_{l,k,j}} = -\left(O_{p,k} - x_{p,l,k}\right) \acute{f}\left(y_{p,l,k}\right) x_{p,l-1,k,j} \quad (32)$$

$$\nabla E_p = \delta_{p,k} x_{p,l,k} \left(1 - x_{p,l,k}\right) x_{p,l-1,j} \quad (33)$$

Ainsi la mise à jour des poids peut être décrite comme par les équations suivantes

$$w_{l,k,j}(t+1) = w_{l,k,j}(t) + \Delta_p w_{l,k,j}(t) \quad (34)$$

$$\Delta_p w_{l,k,j}(t) = \lambda \left(O_{p,k} - x_{p,l,k}\right) \acute{f}\left(y_{p,l,k}\right) x_{p,l-1,j} \quad (35)$$

Où λ est le pas $0 < \lambda < 1$

II.15. Conclusion

L'objectif de ce chapitre est de définir les différentes méthodes utilisées pour la modélisation des propriétés physico-chimiques des tensioactifs en employant une approche basée sur la théorie QSPR (relation quantitative structures-propriétés). Afin de trouver un modèle mathématique qui décrit les propriétés des tensioactifs en fonction des descripteurs, un ensemble des méthodes a été proposé dans la littérature, parmi ces derniers on a : la régression linéaire multiple (MLR), la régression des moindres carrés partiels (PLS) et les réseaux de neurones artificiels (RNA).

III. Modélisation de la CMC des tensioactifs :

III.1. Logiciel utilisé pour la génération de descripteurs

Le logiciel qu'on a utilisé dans cette étude est le PaDEL-Descriptor, c'est un logiciel pour calculer les descripteurs moléculaires. Le logiciel détermine actuellement 1875 descripteurs (1444 descripteurs 1D, 2D et 431 descripteurs 3D (http://www.scbdd.com/padel_desc/index/).

III.2. Développement d'un modèle

III.2.1. Analyse statistique effectuée

Dans cette étude, nous avons effectué une régression linéaire multiple (MLR) comme outil statistique pour le développement de modèles à l'aide d'un logiciel (GFA). Dans cette étude, nous avons également tenté de développer des modèles non linéaires à l'aide de réseaux neuronaux artificiels (RNA) avec l'algorithme rétro-propagation dans la 1 ère phase et Conjuguée dans la 2 ème phase pour former les réseaux, le réseau développé a été utilisé pour prédire la propriété des structures testés.

III.2.2. Validation de modèle

Pour l'évaluation de la fiabilité et de la robustesse des modèles développés, différents types de critères de validation ont été utilisés. Pour la validation des modèles, trois stratégies [116] ont été adoptées :

(A) Validation interne ou validation croisée à l'aide de la base d'apprentissage.

(B) Validation externe en utilisant la base de teste.

(C) La randomisation des données.

L'objectif principal de la modélisation QSPR est de développer un modèle robuste qui devrait être capable de prédire la propriété d'un nouveau composé d'une manière totalement impartiale, précise et fiable [116]. Le modèle QSPR développé à l'aide de la base d'apprentissage doivent être validés à l'aide d'une base de teste pour vérifier la capacité prédictive des modèles développés. L'étape de validation évaluent donc la fiabilité des modèles développés pour leur possible application potentielle sur un nouveau tensioactif de la même famille et permettent de juger de la confiance de la prédiction [117].

III.2.3. Mesure de validation interne et externe

III.2.3.1. Q^2 (ou Q^2_{Int}).

Les modèles développés ont été soumis à une validation interne croisée (leave-one-out : LOO). La validation croisée de R^2 (Q^2) est déterminée sur la base de la capacité prédictive du modèle, pour une meilleure prédiction du modèle la valeur R^2 (Q^2) devrait être entre 0.5 et 1. Le R^2 est exprimé comme :

$$Q^2 = 1 - \frac{\sum(Y_{obs} - Y_{pred})^2}{\sum(Y_{obs} - \bar{Y}_{training})^2} \qquad (1)$$

Dans l'équation (1) Y_{obs} est la propriété observée des structures dans la base d'apprentissage, Y_{pred} est la propriété LOO-prédite des structures de la base d'apprentissage, $\bar{Y}_{training}$est la propriété observée moyenne des structures de la base d'apprentissage. La capacité prédictive du modèle est acceptable lorsque Q^2 est supérieure à 0,5

III.2.3.2. Q^2_{ext} ou R^2_{pred}

Est appelé R^2 prédictif d'un modèle développé et il sert à vérifier la capacité prédictive externe d'un modèle QSPR. En utilisant les structure de la base de donnée, différents modèles ont été développés et les propriétés des structure de test ont été prédites en utilisant les modèles développés. La capacité prédictive d'un modèle a été déterminée en calculant la métrique statistique appelée R^2_{pred} ou Q^2_{ext} qui est définie par l'équation suivant :

$$R^2_{pred} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{training(test)})^2} \qquad (2)$$

Où $Y_{obs\ (test)}$ est la propriété observée des structures de la base de test, Y_{pred} (test) est la propriété prédite des structures de la base de test et $\bar{Y}_{training(test)}$ correspond à la moyenne

des propriétés observées des composés de formation. La valeur de R^2_{pred} pour un modèle acceptable devrait être plus forte de 0,5.

III.2.3.3. Paramètres de r^2_m

Cette nouvelle métrique statistique (r^2_m) a été proposée par Roy et al [117, 118] pour obtenir Un véritable modèle prédictif de QSPR. Cette métrique est calculée en fonction des corrélations entre les valeurs observées et prédites avec (r^2) et hors (r^2_0) dans le domaine des lignes de régression des moindres carrés, la formule générale pour le paramètre r^2_m peut être présentée par l'équation suivante :

$$r_m^2 = r^2\left(1 - \sqrt{r^2 - r_0^2}\right) \qquad (3)$$

Initialement, les valeurs observées étaient considérées dans l'axe des y alors que les valeurs prédites étaient considérées dans l'axe des x et que le paramètre était déterminé pour la validation de la base d'apprentissage, Test et les deux au même temps, et les notations correspondantes étaient respectivement r^2_m (LOO), r^2_m (test) et r^2_m (global) [118]. Récemment, Une autre variante de r^2_m, appelée r'^2_m, a été proposée par Roy et al [119] où les axes sont inversés, les valeurs prédictives étant considérées dans l'axe des y et les valeurs observées sont considérées dans l'axe des x donnant le paramètre r'^2_m comme suit :

$$r_m'^2 = r^2\left(1 - \sqrt{r^2 - r_0'^2}\right) \qquad (4)$$

Où r'^2_0 ont la même signification que r^2_0, mais compte tenu des axes inversés. On constate que la valeur de r^2 reste identique pendant le changement d'axes, alors que ce n'est pas vrai pour le cas de r^2_0. Lorsque les valeurs observées des structures de la base de test (axe des y) sont Tracé contre les valeurs prédites des structures (axe x) intercepté à zéro, la pente de la ligne ajustée donne la valeur De k. L'échange des axes donne la valeur de k '. Les équations suivantes sont utilisées pour le calcul de r^2, r'^2, k et k ':

$$r_0^2 = 1 - \frac{\sum(Y_{obs} - KY_{pred})^2}{\sum(Y_{obs} - \bar{Y}_{obs})^2} \quad (5)$$

$$r'^2_0 = 1 - \frac{\sum(Y_{pred} - K'Y_{obs})^2}{\sum(Y_{pred} - \bar{Y}_{obs})^2} \quad (6)$$

$$K = \frac{\sum(Y_{obs} Y_{pred})}{\sum(Y_{pred})^2} \quad (7)$$

$$K' = \frac{\sum(Y_{obs} Y_{pred})}{\sum(Y_{obs})^2} \quad (8)$$

Dans les cas où la gamme Y n'est pas très large, l'une ou l'autre des mesures r^2_m et r'^2_m peut pénaliser fortement la qualité des prédictions. Deux autres paramètres ont été développés [25]. À savoir $\overline{r^2_m}$, la valeur moyenne de r^2_m et r'^2_m, et Δr^2_m, la différence absolue entre r^2_m et r'^2_m. La moyenne de r^2_m et r'^2_m et $\overline{r^2_m}$ s'est révélée mieux métrique que les mesures r^2m originales. En général, la différence entre les valeurs r^2_m et r'^2_m devrait être faible pour les bons modèles. Il a été démontré que la valeur de Δr^2_m devrait de préférence être inférieure à 0,2, à condition que la valeur de $\overline{r^2_m}$ soit supérieure à 0,5 comme les paramètres de r^2_m d'origine, les paramètres de moyenne et de différence peuvent être déployés pour juger de la qualité prédictive du modèle pour entraînement

$\left(\bar{r}^2_{m(LOO)} et\ \Delta r^2_{m(LOO)}\right)$, test $\left(\bar{r}^2_{m(test)} et\ \Delta r^2_{m(test)}\right)$ aussi bien $\left(\bar{r}^2_{m(overall)} et\ \Delta r^2_{m(overall)}\right)$ [120].

III.2.3.4. La randomisation des données

Les relations entre LogCMC et les descripteurs ont encore été vérifiées pour la robustesse des modèles par le test de randomisation en utilisant le logiciel Cerius 2 (version 4.10). Deux types de tests de randomisation peuvent être effectués : randomisation du processus et randomisation du modèle. Dans le processus de test de randomisation, le

brouillage aléatoire de la variable de réponse s'accompagne d'une nouvelle sélection de variables de l'ensemble du pool de descripteurs lorsque, dans le cas du test de randomisation du modèle, seuls les descripteurs présents dans le modèle d'origine sont utilisés. Pour un modèle QSPR approprié, le coefficient de corrélation moyen (Rr) des modèles randomisés devrait être inférieur au coefficient de corrélation (R) du modèle non randomisé. Un paramètre CR_p^2, utilisé par Roy's et al, Pénalise le modèle R^2 pour une petite différence entre le coefficient de corrélation moyen carré (R^2r) des modèles aléatoires et le coefficient de corrélation carré (R^2) du modèle non randomisé. Le paramètre CR_p^2 est défini comme

$$CR_p^2 = R\sqrt{R^2 - R_r^2} \qquad (9)$$

Pour un modèle QSPR acceptable, la valeur de CR_p^2 devrait être supérieure à 0,5.

III.2.4. Domaine d'application (DA)

Le domaine d'applicabilité (DA) est une autre mesure de validation pour vérifier les modèles QSPR. Les prédictions d'un modèle QSPR sont plus fiables si les composés prédits sont dans le domaine d'applicabilité du modèle. L'DA d'un modèle QSPR est l'espace, la connaissance ou l'information biologique, structurelle ou physico-chimique sur laquelle le modèle de la d'application d'apprentissage a été développé et pour lequel il est applicable pour faire des prédictions pour de nouveaux composés. Lorsqu'un composé est très différent à tous les composés de l'ensemble de modélisation, une prédiction fiable de la séparabilité est incertaine. L'idée de DA [121] est utilisée pour éviter une extrapolation aussi infondée des prédictions d'activité. Pour étudier le domaine d'application des produits chimiques, nous avons appliqué l'approche de levier [122] dans la présente étude. Pour observer domaine d'application a d'un modèle QSPR, on a utilisé un graphique de valeurs résiduelles normalisées prédites (R) contre le levier (Hatdiagonal) (h) (la parcelle de Williams) pour la détection animale et simple des deux réponses (c.-à-d. Les composés avec des résidus normalisés prévus supérieurs à trois unités de déviation standard,> 3 σ) Et des produits chimiques influents sur le plan structurel (h> h *) où h * est la valeur du chapeau critique. Le paramètre h * est calculé comme $h^* = 3p'/n$, où p 'est le nombre de variables de modèle plus un, et n est le nombre d'objets utilisés pour calculer le modèle.

III.2.5. Méthodologie de calcul par RNA

La méthodologie de calcul par les réseaux de neurones du (LogCMC) est illustrée par l'organigramme de la figure III.1.

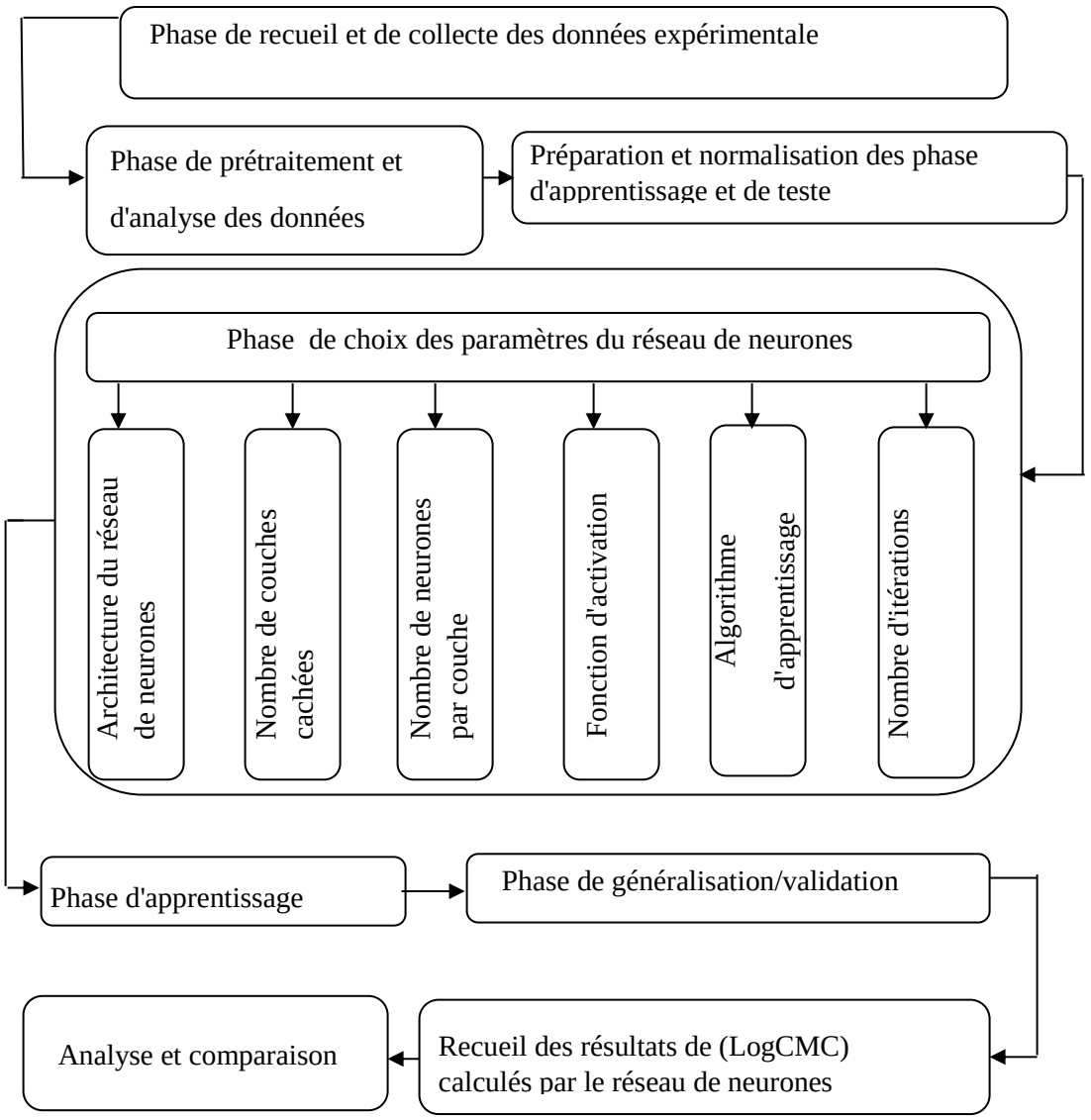


Figure III.1 : Méthodologie de calcul du (Log CMC) par les réseaux de neurones.

Les poids de connexion sont initialisés de façon aléatoire et la procédure d'apprentissage est menée en utilisant l'algorithme de rétro-propagation [123] et l'algorithme du second ordre de levenberg-marquard [124,125].

III.2.6. Préparation de la base de donnée

Les base des données d'entrée est réalisée par une matrice rectangulaire de dimension

- Non-ionique (4, 88).
- Cationique (4, 51).
- Anionique (4, 70).
- Globale (Anionique + Cationique + Non-ionique) (4, 171).

Les base des données de sortie est réalisée par une matrice rectangulaire de dimension

- Non-ionique (1, 88).
- Cationique (1, 51).
- Anionique (1, 70).
- Globale (Anionique + Cationique + Non-ionique) (1,171).

La propriété émergente d'un RNA est sa capacité d'apprentissage [126]. A partir d'un état quelconque, il va, en s'appuyant sur un algorithme d'apprentissage, adapter la valeur des poids synaptiques pour fournir une réponse vraisemblable vis-à-vis du problème traité. Il existe différentes techniques d'apprentissage pour déterminer la valeur des poids synaptiques optimaux [127].

La base de données constituée de l'ensemble des calculs analytiques est divisée en deux sous-bases une base d'apprentissage/validation qui permet l'optimisation de l'architecture du RN et une base test utilisée pour déterminer ses performances prédictives. Les deux sous bases de la base des données sont sélectionnées selon le schéma illustratif de figure III.2.

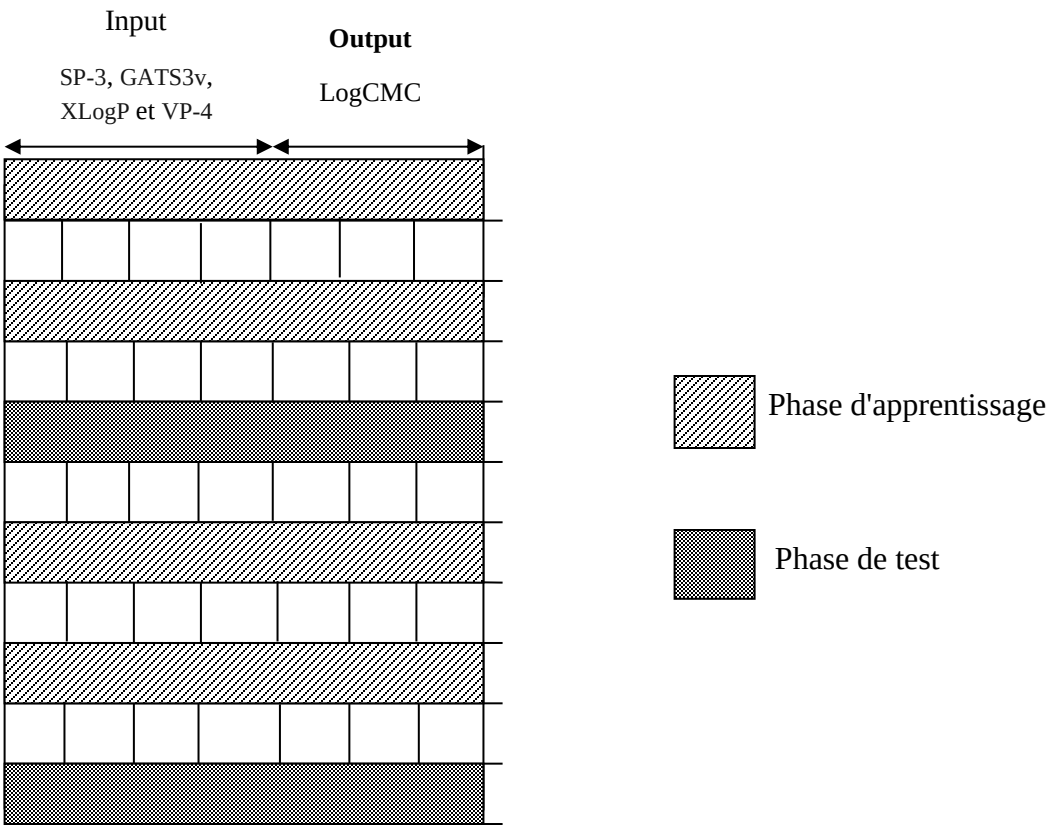


Figure III.2 : Illustration de la phase d'apprentissage et de test

III.2.7. Normalisation des données

Les données sont correspondant respectivement aux

- Non-ionique (SP-3, GATS3v, XLogP et VP-4 et LogCMC).
- Cationique (4, 51).
- Anionique (4, 70).
- Globale (Anionique + Cationique + Non-ionique)

(SP-3, GATS3v, XLogP et VP-4 et LogCMC). Les valeurs numériques ont été normalisées afin d'améliorer l'optimisation. La normalisation se fait par colonne (premnmx : MATLAB), chaque paramètre des entrées et de sortie (X1) est normalisé (XiN) par l'expression suivante :

$$X_{iN} = 2\frac{(X - X\,min)}{(X\,max - X\,min)} - 1 \qquad (10)$$

L'apprentissage supervisé est retenue dans cette étude, il consiste à fournir au réseau des jeux de données d'entrées (SP-3, GATS3v, XLogP et VP-4) pour lesquels on connaît les réponses souhaitées (LogCMC). On construit alors un critère d'erreur en comparant la réponse du RN à celle attendue (LogCMC) expérimentalement. Un algorithme de correction des poids synaptiques va permettre de diminuer cette erreur et d'améliorer ainsi la réponse du RN.

La mise en œuvre de cette technique impose la constitution d'une base d'apprentissage formée de l'ensemble de jeux de données et des sorties associées.

L'apprentissage du RN et son architecture sont optimisés par une méthode de validation croisée [128] appliquée à la base d'apprentissage validation : le leave-one-out [129].

Le nombre optimal de neurones cachés est défini comme le nombre pour lequel l'erreur quadratique moyenne et le coût de généralisation sont minimaux.

III.2.8. Organigramme de la méthodologie de calcul de RNSO

L'organigramme de la méthodologie de calcul du réseau de neurones statique optimale RNSO est illustré par la figure 3.

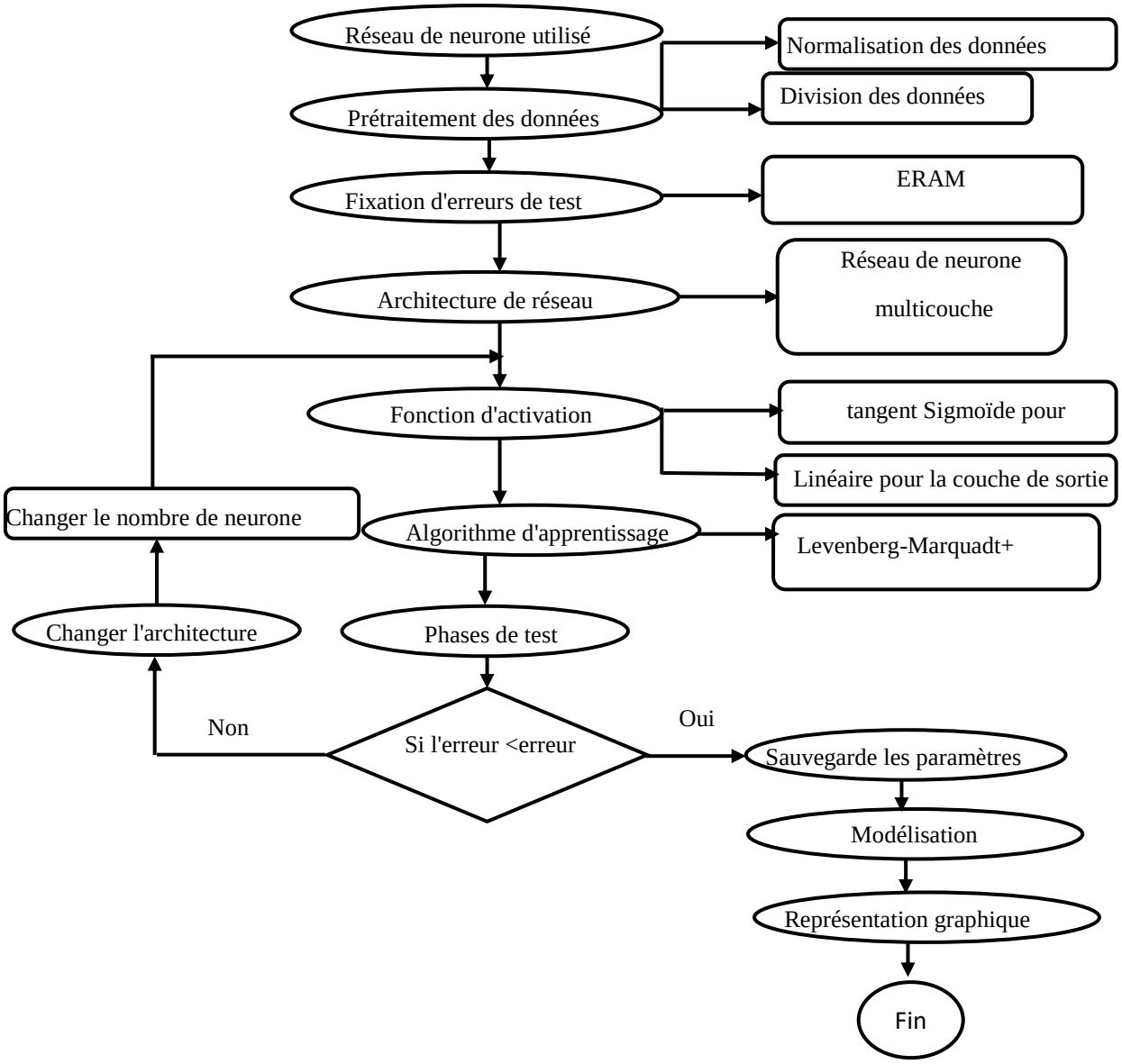


Figure III.3 : L'organigramme de la méthodologie de calcul du réseau de neurones statique optimale RNSO

Résultat et discussion

III.3. Modélisation des structures Non-ionique

III.3.1. Base des donnes

Les valeurs de CMC sélectionnées pour 88 structures des tensioactifs non-ioniques (Tableau 1) a été recueilli à partir la littérature [130]. Tous les 88 composés ont été sélectionnés dont les valeurs de CMC ont été mesurées à 25 ° C dans de l'eau purifiée sans aucun additif.

Tableau III.1 : base des données des structure Non-ioniques de ce étude

Nbr	Structure	logCMC	Reference
01	$C_8H_{17}OC_2H_4OH$	-2.310	[130]
02	$C_8H_{17}OC_2H_4OH$	-2.125	[130]
03	$C_8H_{17}(OC_2H_4)_6OH$	-2.004	[130]
04	$C_{10}H_{21}(OC_2H_4)_4OH$	-3.167	[130]
05	$C_{10}H_{21}(OC_2H_4)_6OH$	-3.046	[130]
06	$C_{10}H_{21}(OC_2H_4)_8OH$	-3.000	[130]
07	$C_{11}H_{23}(OC_2H_4)_8OH$	-3.523	[130]
08	$C_{12}H_{25}(OC_2H_4)_2OH$	-4.481	[130]
09	$C_{12}H_{25}(OC_2H_4)_3OH$	-4.284	[130]
10	$C_{12}H_{25}(OC_2H_4)_4OH$	-4.194	[130]
11	$C_{12}H_{25}(OC_2H_4)_5OH$	-4.194	[130]
12	$C_{12}H_{25}(OC_2H_4)_7OH$	-4.086	[130]
13	$C_{12}H_{25}(OC_2H_4)_8OH$	-4.000	[130]
14	$C_{13}H_{27}(OC_2H_4)_8OH$	-4.569	[130]
15	$C_{14}H_{29}(OC_2H_4)_6OH$	-5.000	[130]
16	$C_{14}H_{29}(OC_2H_4)_8OH$	-5.046	[130]
17	$C_{15}H_{31}(OC_2H_4)_8OH$	-5.456	[130]
18	$C_{16}H_{33}(OC_2H_4)_6OH$	-5.796	[130]
19	$C_{16}H_{33}(OC_2H_4)_7OH$	-5.770	[130]
20	$C_{16}H_{33}(OC_2H_4)_9OH$	-5.678	[130]
21	$C_{16}H_{33}(OC_2H_4)_{12}OH$	-5.638	[130]
22	$C_{16}H_{33}O(C_2H_4O)_{15}H$	-5.509	[130]
23	$C_{16}H_{33}O(C_2H_4O)_21H$	-5.409	[130]
24	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_2H$	-3.886	[130]
25	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_3H$	-4.013	[130]
26	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_4H$	-3.886	[130]

27	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_5H$	-3.824	[130]
28	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_6H$	-3.678	[130]
29	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_7H$	-3.602	[130]
30	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_8H$	-3.553	[130]
31	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_9H$	-3.523	[130]
32	p-t-$C_8H_{17}C_6H_4O(C_2H_4O)_{10}H$	-3.481	[130]
33	$C_8H_{17}CHOHCH_2OH$	-2.638	[130]
34	$C_8H_{17}CHOHCH_2CH_2OH$	-2.638	[130]
35	$C_{10}H_{21}CHOHCH_2OH$	-3.745	[130]
36	$C_{12}H_{25}CHOHCH_2CH_2OH$	-4.886	[130]
37	$C_8H_{17}(OC_2H_4)_5OH$	-2.036	[130]
38	$C_{10}H_{21}(OC_2H_4)_5OH$	-3.119	[130]
39	$C_{13}H_{27}(OC_2H_4)_5OH$	-4.310	[130]
40	$C_9H_{19}C_6H_4(OC_2H_4)_{10}OH$	-4.125	[130]
41	$C_9H_{19}C_6H_4(OC_2H_4)_{31}OH$	-3.745	[130]
42	$C_{11}H_{23}CON(C_2H_4OH)_2$	-3.585	[130]
43	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_2OH$	-2.602	[130]
44	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_3OH$	-2.553	[130]
45	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_5OH$	-2.432	[130]
46	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_7OH$	-2.319	[130]
47	n-Octyl-b-D-glucoside	-1.602	[130]
48	n-Decyl-b-D-glucoside	-2.658	[130]
49	Dodecyl-b-D-glucoside	-3.721	[130]
50	Decyl-b-D-maltoside	-2.699	[130]
51	Dodecyl-b-D-maltoside	-3.824	[130]
52	Sucrose monolaurate	-3.469	[130]
53	$(C_4H_9)_2CHCH_2(OC_2H_4)_6OH$	-2.977	[130]
54	$(C_4H_9)_2CHCH_2(OC_2H_4)_9OH$	-3.049	[130]
55	$C_{12}H_{25}(OC_2H_4)_6OH$	-4.144	[130]
56	$C_{12}H_{25}(OC_2H_4)_9OH$	-3.954	[130]
57	$C_{12}H_{25}(OC_2H_4)_{12}OH$	-3.788	[130]
58	$C_{11}H_{23}CONH(C_2H_4O)_4H$	-3.651	[130]
59	$C_{10}H_{21}CON(CH_3)(CHOH)_4CH_2OH$	-2.054	[130]
60	$C_{11}H_{23}CON(CH_3)CH_2CHOHCH_2OH$	-2.951	[130]
61	$C_{11}H_{23}CON(CH_3)CH_2(CHOH)_3CH_2OH$	-3.06	[130]
62	$C_{11}H_{23}CON(CH_3)CH_2(CHOH)_4CH_2OH$	-3.18	[130]
63	$C_{12}H_{25}CON(CH_3)CH_2(CHOH)_4CH_2OH$	-3.518	[130]
64	$C_{13}H_{27}CON(CH_3)CH_2(CHOH)_4CH_2OH$	-3.982	[130]
65	$C_{10}H_{21}N(CH_3)CO(CHOH)_4CH2OH$	-2.007	[130]
66	$C_{12}H_{25}N(CH_3)CO(CHOH)_4CH_2OH$	-3.018	[130]

67	$C_{14}H_{29}N(CH_3)CO(CHOH)_4CH_2OH$	-4.095	[130]
68	$C_{16}H_{33}N(CH_3)CO(CHOH)_4CH_2OH$	-4.971	[130]
69	$C_{18}H_{37}N(CH_3)CO(CHOH)_4CH_2OH$	-5.97	[130]
70	$C_6F_{13}CH_2CH_2(OC_2H_4)_{14}OH$	0.049	[130]
71	$C_8F_{17}CH_2CH_2N(C_2H_4OH)2$	-0.279	[130]
72	$n\text{-}C_4H_9(OC_2H_4)_6OH$	-0.927	[130]
73	$(CH_3)_2CHCH_2(OC_2H_4)_6OH$	-0.649	[130]
74	$n\text{-}C_6H_{13}(OC_2H_4)_6OH$	-1.72	[130]
75	$(C_2H_5)_2CHCH_2(OC_2H_4)_6OH$	-1.153	[130]
76	$C_{15}H_{31}CON(C_2H_4OH)_2$	-5.218	[130]
77	$C_4H_9O(C_2H_4O)H$	-0.009	[131]
78	$C_4H_9O(C_2H_4O)_6H$	-0.110	[131]
79	$C_6H_{13}O(C_2H_4O)_3H$	-1.000	[131]
80	$C_6H_{13}O(C_2H_4O)_6H$	-1.164	[131]
81	$C_8H_{17}O(C_2H_4O)_9H$	-1.886	[131]
82	$C_{10}H_{21}O(C_2H_4O)_3H$	-3.222	[131]
83	$C_{10}H_{21}O(C_2H_4O)_9H$	-2.886	[131]
84	$C_{12}H_{25}O(C_2H_4O)_6H$	-4.060	[131]
85	$C_{12}H_{25}O(C_2H_4O)_9H$	-4.000	[131]
86	$C_{12}H_{25}O(C_2H_4O)_{12}H$	-3.854	[131]
87	$C_8H_{17}(C_6H_4)O(C_2H_4O)H$	-3.678	[131]
89	$C_6F_{13}C_2H_4SC_2H_4OC_2H_4SC_2H_4(OC_2H_4)_2OH$	-4.553	

III.3.2. Conception de la topologie de RNSO

III.3.2.1. Choix des entrées et des sorties

Afin de rassembles les resultas expérimntaux on a choisit comme:

Entrées (SP-3, GATS3v, XLogP et VP-4) tel que:

- (SP-3) : Chemin simple, ordre 3 [132].
- (GATS3v) : Autocorrélation Geary - retard 3 / pondéré par Van der Waals volumes [133].
- (XLogP) : XLogP [133].
- (VP-4) : Valence chemin, ordre 4 [132]

Ce qui donne une matrice de 4 colonnes et 88 lignes.

Sorties (LogCMC) tel que :

La CMC : c'est la concentration critique missel

Ce qui donne une matrice de 1 colonnes et 88 lignes.

III.3.2. Expression de calcul

Sur la base des résultats déterminer à l'aide des logiciels PADEL et les algorithmes génétiques, un programme a été développé sous MATLAB afin de prédire LogCMC.

Le RNA est construit et testé en utilisant le logiciel MATLAB (nntool : version 7.5). L'optimisation de l'architecture des RNA consiste en la détermination : du nombre de neurones cachés, de la fonction d'activation de la couche cachée, de la fonction d'activation de la couche de sortie et de l'algorithme d'apprentissage, par comparaison des performances des RN sur la base d'apprentissage/validation en terme d'erreur relative absolue et l'erreur absolue moyenne (ERAM, EAM) et de coût de généralisation.

Tableau III.2 : Structure du réseau de neurones optimisé.

Type de réseau	Couche d'entrée	Couche cachée		Couche cachée		Couche de sortie	
	Nb. de neurones	Nb. de neurones	Fonction d'activation	Nb. de neurones	Fonction d'activation	Nb .de neurones	Fonction d'activation
RNS MLP	4	4	Tangente - sigmoïde hyperbolique « tansig »	4	Tangente - sigmoïde hyperbolique « tansig »	1	Identité Purelin

La représentation schématique du perceptron multicouches (4 :4 :4 :1) pour notre réseau est donnée par les figure III.3 et III.4

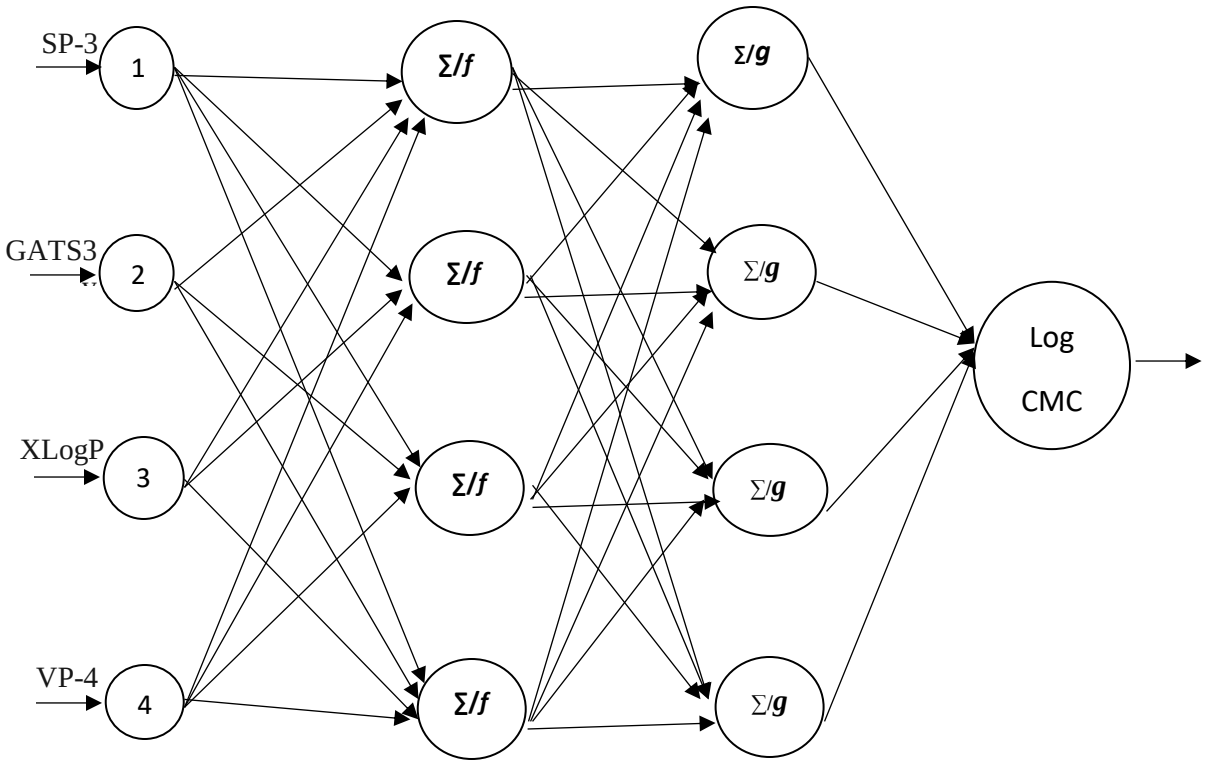


Figure III.4 : Architecture de réseau de neurones.

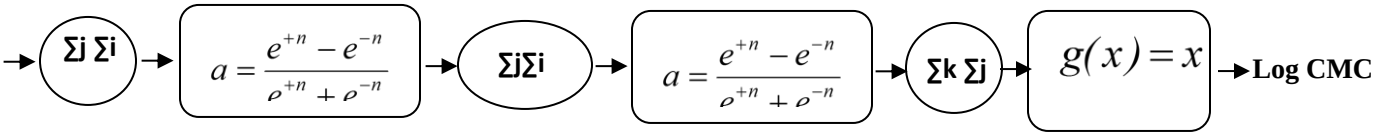


Figure III.5 : Réseau de neurones multicouche pour logCMC.

Tableau III.3 : Matrice des poids et des biais du couch d'entrée

W1				b1
-1,053	0,356	0,591	-0,603	2,493
1,050	2,739	-0,047	0,132	0,233
-0,560	0,302	0,346	1,010	-0,061
-1,496	1,808	-1,287	-0,887	3,573

Tableau III.4 : Matrice des poids et des biais du couch caché

W2				b2
-0,840	-1,325	0,221	-2,312	1,208
-0,049	0,211	-1,122	-0,627	0,448
-1,829	0,115	-0,766	-0,916	-1,148
1,078	-1,604	-0,093	-0,514	1,893

Tableau III.5 : Matrice des poids et des biais du couch sortie

W3				b3
0,338	1,963	-0,553	0,509	-0,617

III.3.3. Modélisation de LogCMC

La figure III.6 présente la validation du modèle de réseau de neurone, on remarque bien que les points expérimentaux se rapprochent de la première bissectrice.

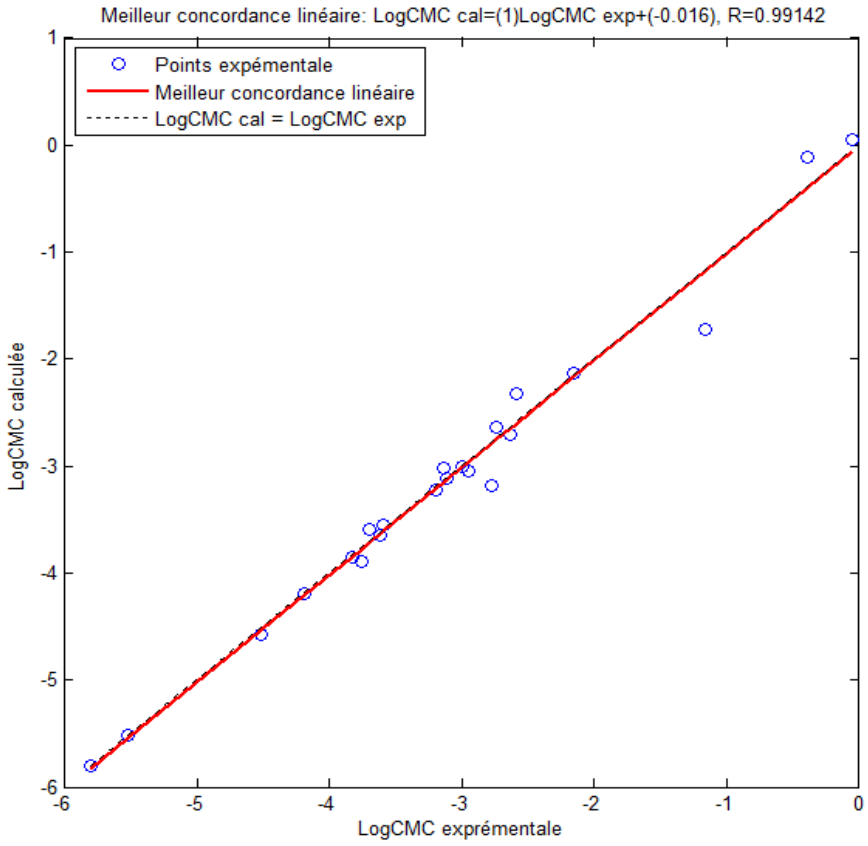


Figure III.6 : Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales

Nous constatons que le modèle de réseau de neurones valide parfaitement les résultats expérimentaux figure III.5.

Tableau III.6 : Valeurs des erreurs calculées pour LogCMC

ERAM%	R^2pred	R^2
23.815	0.998	0.991

Du tableau 4, on constate que le modèle rapporté est d'une qualité acceptable. Avec un coefficient de régression R^2= 0,991 et de R^2pred =0.998.

L'histogramme de la figure III.6, présente l'erreur absolue entre les points expérimentaux et le modèle de réseau de neurone.

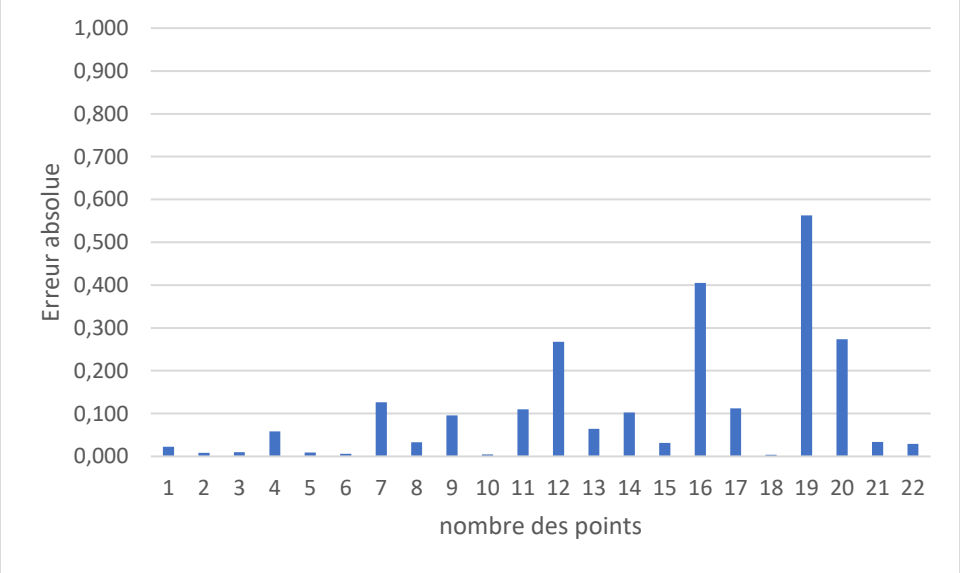


Figure III.7 : Présentation de l'erreur absolue

D’après la figures III.7 qui représente l’erreur absolue entre les résultats prédit par RNA et les résultats expérimentaux de la base de données de teste, on constate que parmi 22 points expérimentaux on a uniquement quatre points qui possèdent des erreurs très élevées.

Le tableau III.7 illustre les valeurs minimale et maximale de log CMC qui représente le domaine d'application de modèle neuronal, on contrepartie on constate que les minimum et maximum entre les valeurs calculée et expérimentales sont très proches

Tableau III.7 : les valeur min et max du Log CMC

	Min	Max
Log CMC calculée	-5,796	-0,049
Log CMC expérimentale	-5,805	-0,045

III.4. Modélisation des structures Cationique

III.4.1. Choix des entrées et des sorties

Afin de rassembles les résulta expérimentaux on a choisi comme :

Entrées (SpMAD_D, VR3_Dzs, ALogp2 et MATS3c) tel que :

- (SpMA_D) : Déviation absolue moyenne spectrale de la matrice de Barysz / pondérée par le nombre atomique [132].
- (VR3_Dzs) : Logarithmique de Randic-like auto-vecteur indice basé de Barysz matrice / pondéré par l'état initiale [132].
- (ALogp2) : Logarithme carré de l'apprentissage [133].
- (MATS3c) : Autocorrélation de Moran - décalage 3 / pondéré par des charges [132].

Ce qui donne une matrice de 4 colonnes et 51 lignes.

Sorties (Log CMC) tel que :

La CMC : c'est la concentration critique missel

Ce qui donne une matrice de 1 colonnes et 51 lignes.

III.4.2. Expression de calcul

Sur la base des résultats déterminer à l'aide des logiciels PADEL et les algorithmes génétiques, un programme a été développé sous MATLAB afin de prédire LogCMC.

Le RNA est construit et testé en utilisant le logiciel MATLAB (nntool : version 7.5). L'optimisation de l'architecture des RNA consiste en la détermination : du nombre de neurones cachés, de la fonction d'activation de la couche cachée, de la fonction d'activation de la couche de sortie et de l'algorithme d'apprentissage, par comparaison des performances des RN sur la base d'apprentissage/validation en terme d'erreur relative absolue et l'erreur absolue moyenne (ERAM, EAM) et de coût de généralisation.

Tableau III.8 : Structure du réseau de neurones optimisées structure cationique.

Type de réseau	Couche d'entrée	Couche cachée		Couche cachée		Couche de sortie	
	Nb. de neurones	Nb. de neurones	Fonction d'activation	Nb. de neurones	Fonction d'activation	Nb .de neurones	Fonction d'activation
RNS MLP	4	7	Tangente - sigmoïde hyperbolique « tansig »	7	Tangente - sigmoïde hyperbolique « tansig »	1	Identité Purelin

La représentation schématique du perceptron multicouches (4 : 7 : 7 : 1) pour notre réseau

Tableau III.9 : Matrice des poids et des biais du couch d'entrée

w1				b1
-1,117	0,309	-1,733	0,959	2,032
2,155	0,518	0,219	0,067	-1,818
-0,030	-1,519	2,001	1,189	-0,849
0,678	-1,326	0,144	-0,915	-0,251
1,803	-0,921	0,072	0,180	0,489
1,497	-0,813	-1,199	1,063	0,999
-0,499	1,136	0,686	-1,688	-2,578

Tableau III.10 : Matrice des poids et des biais du couch caché

w2							b2
0,215	-0,853	-1,117	-0,505	-0,155	-0,373	0,927	-2,301
-0,735	-0,891	0,058	-0,672	-0,616	-0,138	-0,843	2,135
0,161	-1,275	1,203	0,011	0,640	-1,185	0,087	-1,426
-0,582	1,191	-0,402	0,558	1,237	-0,234	-0,487	1,212
0,650	0,972	-0,401	0,599	-1,209	-0,781	0,611	-0,862
0,276	-0,866	-0,535	-0,053	-0,972	1,147	1,035	-0,587
0,713	-0,663	-0,589	-0,365	-1,089	-0,291	1,015	-0,390
1,252	-0,308	1,195	0,293	1,004	-0,294	-0,327	0,013
0,795	0,686	-2,104	0,398	-0,477	-0,544	0,065	0,233

0,400	-0,893	1,164	1,194	0,310	-1,202	-0,025	0,785
0,683	-0,814	-1,006	-1,294	-1,167	-0,139	0,468	0,851
0,289	-1,588	0,340	0,724	0,109	0,100	1,050	1,229
-0,306	0,994	0,515	0,748	1,237	-0,683	-0,172	-1,589
-0,847	0,080	1,196	-0,924	1,082	0,382	0,559	-1,796
-1,007	0,431	-0,859	1,099	-0,239	0,474	-0,853	-2,028

Tableau III.11 : Matrice des poids et des biais du couch sortie

w3							
-0,475	-0,364	-1,185	-0,661	0,400	0,452	-0,313	-0,385

w3							b3
-1,006	0,799	0,917	-0,680	0,380	0,762	-0,510	0,886

III.4.3. Modélisation de LogCMC

La figure III.8 présente la validation du modèle de réseau de neurone, on remarque bien que les points expérimentaux se rapprochent de la première bissectrice.

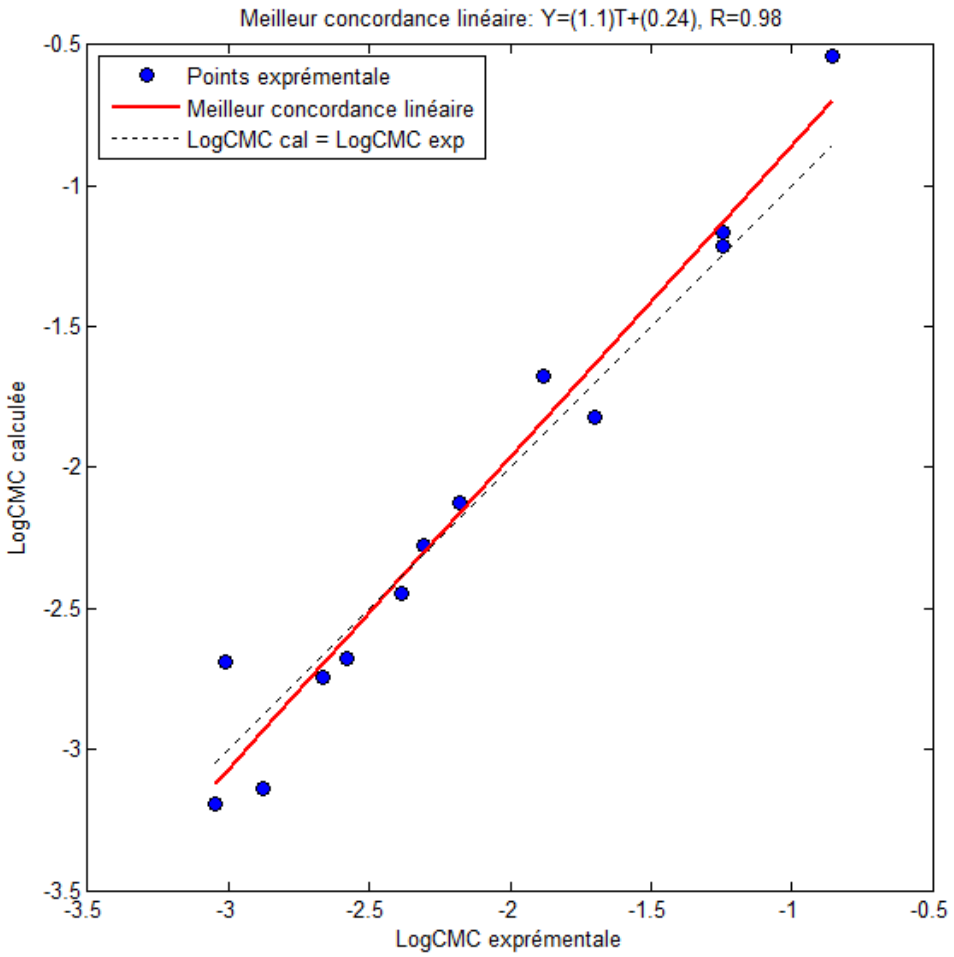


Figure III.8 : Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales

Nous constatons que le modèle de réseau de neurones valide parfaitement les résultats expérimentaux figure III.8.

Tableau III.12 : Valeurs des erreurs calculées pour LogCMC

ERAM%	R^2pred	R^2
5.806	0.995	0.987

Du tableau III.12, on constate que le modèle rapporté est d'une qualité acceptable. Avec un coefficient de régression R^2= 0,987 et de R^2pred =0.995.

L'histogramme de la figure III.9, présente l'erreur absolue entre les points expérimentaux et le modèle de réseau de neurone.

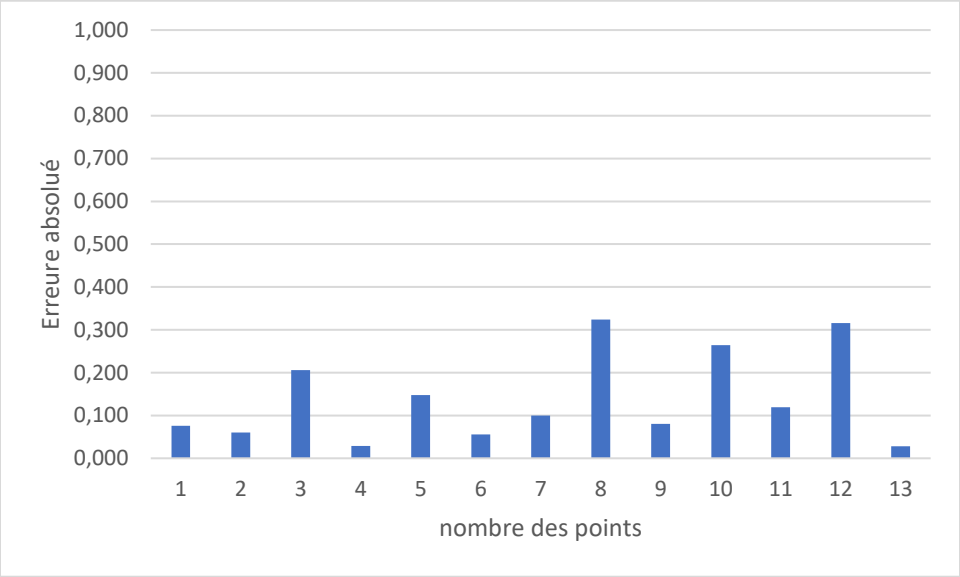


Figure III.9 : Présentation de l'erreur absolue

D'après la figures III.9 qui représente l'erreur absolue entre les résultats prédit par RNA et les résultats expérimentaux de la base de données de teste, on constate que parmi 13 points expérimentaux on a uniquement deux points qui possèdent des erreurs très élevées.

Le tableau III.13 illustre les valeurs minimale et maximale de log CMC qui représente le domaine d'application de modèle neuronal, on contrepartie on constate que les minimum et maximum entre les valeurs calculée et expérimentales sont très proches

Tableau III.13 : les valeur min et max du Log CMC

	Min	Max
Log CMC calculée	-3,194	-0,538
Log CMC expérimentale	-3,046	-0,854

III.5. Modélisation des structures Anionique

III.5.1. Choix des entrées et des sorties

Afin de rassembles les résulta expérimentaux on a choisi comme :

Entrées (MATS8p, MATS7v, ATSC3e et nHBAcc_Lipinski) tel que :

- (MATS8p) : L'autocorrélation de Moran - lag 8 / pondérée par Van der Waals volumes [132].
- (MATS7v) : L'autocorrélation de Moran - lag 7 / pondérée par Van der Waals volumes [132].
- (ATSC3e) Autocorrélation centrée de Broto-Moreau - retard 3 / pondérée par les électronégativités de Sanderson [132].
- (nHBAcc_Lipinski) : Nombre d'accepteurs de liaison hydrogène (en utilisant la définition de Lipinski: tout azote et oxygène) [133].

Ce qui donne une matrice de 4 colonnes et 71 lignes.

Sorties (Log CMC) tel que :

La CMC : c'est la concentration critique missel

Ce qui donne une matrice de 1 colonnes et 71 lignes.

III.5.2. Expression de calcul

Sur la base des résultats déterminer à l'aide des logiciels PADEL et les algorithmes génétiques, un programme a été développé sous MATLAB afin de prédire LogCMC.

Le RNA est construit et testé en utilisant le logiciel MATLAB (nntool : version 7.5). L'optimisation de l'architecture des RNA consiste en la détermination : du nombre de neurones cachés, de la fonction d'activation de la couche cachée, de la fonction d'activation de la couche de sortie et de l'algorithme d'apprentissage, par comparaison des performances des RN sur la base d'apprentissage/validation en terme d'erreur relative absolue et l'erreur absolue moyenne (ERAM, EAM) et de coût de généralisation.

Tableau III.14 : Structure du réseau de neurones optimisées structure cationique.

Type de réseau	Couche d'entrée	Couche cachée		Couche cachée		Couche de sortie	
	Nb. de neurones	Nb. de neurones	Fonction d'activation	Nb. de neurones	Fonction d'activation	Nb .de neurones	Fonction d'activation
RNS MLP	4	5	Tangente - sigmoïde hyperbolique « tansig »	5	Tangente - sigmoïde hyperbolique « tansig »	1	Identité Purelin

La représentation schématique du perceptron multicouches (4 :5 :5 :1) pour notre réseau

Tableau III.15 : Matrice des poids et des biais du couch d'entrée

w1				b1
0,993	-1,068	2,176	0,976	-1,873
1,542	-0,652	1,154	0,790	0,216
-2,682	-0,272	-0,519	0,386	0,671
-1,347	-0,895	1,222	-0,794	-0,924
-0,454	-1,093	-1,965	-0,498	-1,790

Tableau III.16 : Matrice des poids et des biais du couch caché

w2					b2
-0,453	-0,601	-0,793	1,161	-1,051	1,884
-1,104	-1,207	-0,948	-0,706	0,063	0,934
-1,859	-0,346	-1,923	-0,098	0,947	0,204
0,575	-1,796	0,280	-0,818	0,078	1,252
-0,711	0,591	-1,257	-0,554	-1,152	-1,773

Tableau III.17 : Matrice des poids et des biais du couch sortie

w3					b3
-0,961	0,403	0,888	-1,166	-1,633	0,551

III.5.3. Modélisation de LogCMC

La figure III.10 présente la validation du modèle de réseau de neurone, on remarque bien que les points expérimentaux se rapprochent de la première bissectrice.

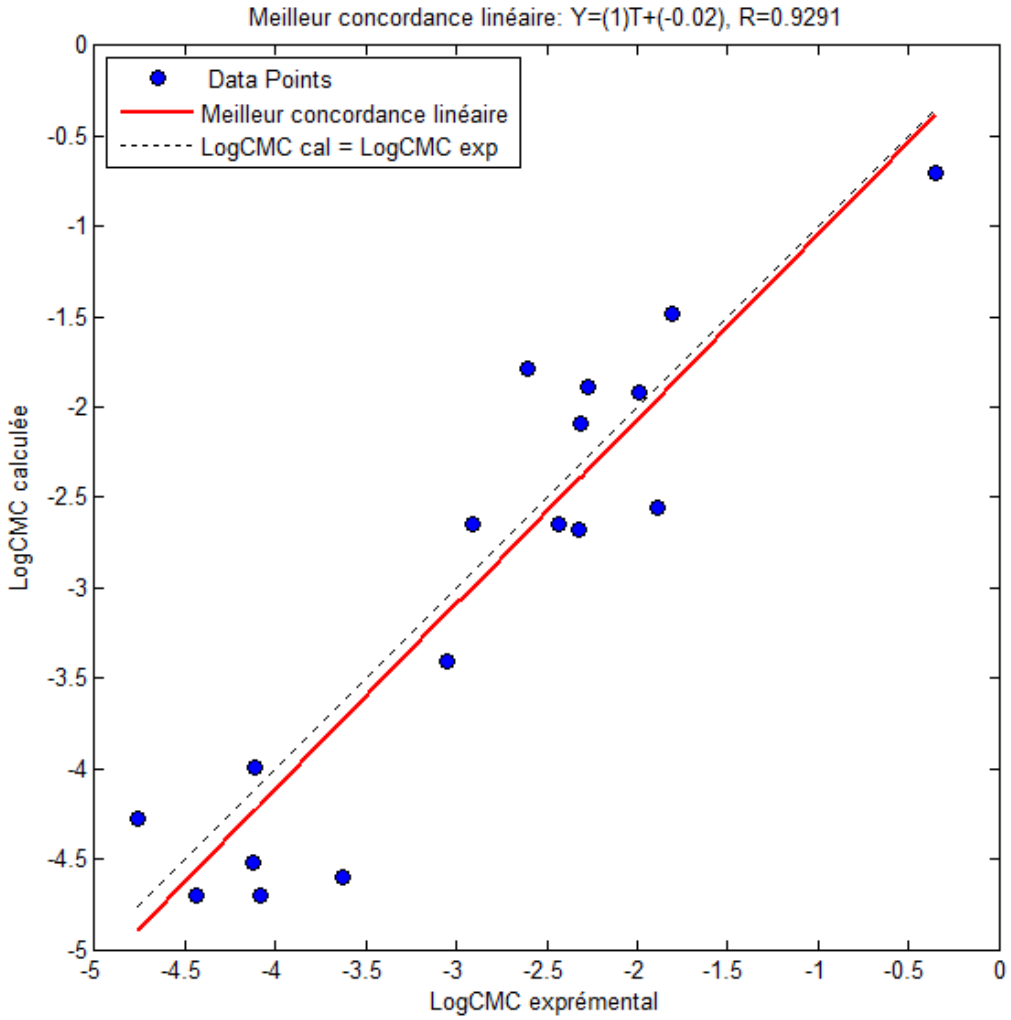


Figure III.10 : Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales

Nous constatons que le modèle de réseau de neurones valide parfaitement les résultats expérimentaux figure III.10.

Tableau III.18 : Valeurs des erreurs calculées pour LogCMC

ERAM%	R^2pred	R^2
16.744	0.979	0.929

Du tableau 18, on constate que le modèle rapporté est d'une qualité acceptable. Avec un coefficient de régression R^2= 0,987 et de R^2pred =0.995.

L'histogramme de la figure III.11, présente l'erreur absolue entre les points expérimentaux et le modèle de réseau de neurone.

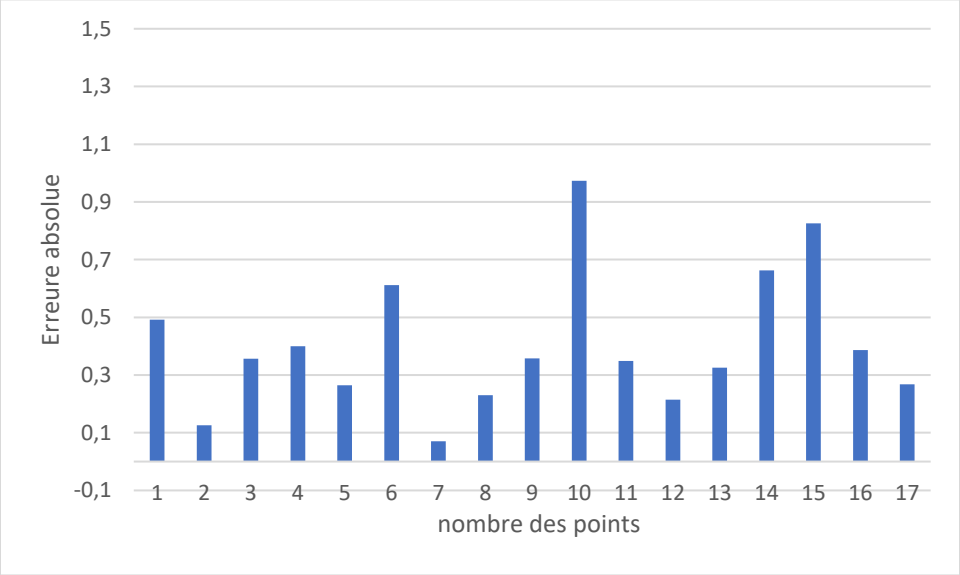


Figure III.11 : Présentation de l'erreur absolue

D'après la figures III.11 qui représente l'erreur absolue entre les résultats prédit par RNA et les résultats expérimentaux de la base de données de teste, on constate que parmi 17 points expérimentaux on a uniquement deux points qui possèdent des erreurs très élevées.

Le tableau III.19 illustre les valeurs minimale et maximale de log CMC qui représente le domaine d'application de modèle neuronal, on contrepartie on constate que les minimum et maximum entre les valeurs calculée et expérimentales sont très proches

Tableau III.19 : les valeur min et max du Log CMC

	Min	Max
Log CMC calculée	-4,698	-0,699
Log CMC expérimentale	-4,77E+00	-0,349

III.6. Modélisation des structures globale (anionique, cationique et non-ionique)

III.6.1. Choix des entrées et des sorties

Afin de rassembles les résulta expérimentaux on a choisi comme :

Entrées (C2SP3, minsCH3, SpAD_DzZ et AATSC8p) tel que :

- (C2SP3) : Le carbone à liaison unique est lié à deux autres carbones [132].
- (minsCH3) : E-État de type atome minimum -CH3 [133]
- (SpAD_DzZ) : Déviation absolue spectrale de la matrice de Barysz / pondérée par le nombre atomique [132].
- (AATSC8p) : Autocorrélation moyenne Broto-Moreau centrée - lag 8 / pondérée par les polarisations [132].

Ce qui donne une matrice de 4 colonnes et 171 lignes.

Sorties (Log CMC) tel que :

La CMC : c'est la concentration critique missel

Ce qui donne une matrice de 1 colonnes et 171 lignes.

III.6.2. Expression de calcul

Sur la base des résultats déterminer à l'aide des logiciels PADEL et les algorithmes génétiques, un programme a été développé sous MATLAB afin de prédire LogCMC.

Le RNA est construit et testé en utilisant le logiciel MATLAB (nntool : version 7.5). L'optimisation de l'architecture des RNA consiste en la détermination : du nombre de neurones cachés, de la fonction d'activation de la couche cachée, de la fonction d'activation de la couche de sortie et de l'algorithme d'apprentissage, par comparaison des performances des RN sur la base d'apprentissage/validation en terme d'erreur relative absolue et l'erreur absolue moyenne (ERAM, EAM) et de coût de généralisation.

Tableau III.20 : Structure du réseau de neurones optimisées structure cationique.

Type de réseau	Couche d'entrée	Couche cachée		Couche cachée		Couche de sortie	
	Nb. de neurones	Nb. de neurones	Fonction d'activation	Nb. de neurones	Fonction d'activation	Nb .de neurones	Fonction d'activation
RNS MLP	4	6	Tangente - sigmoïde hyperbolique « tansig »	6	Tangente - sigmoïde hyperbolique « tansig »	1	Identité Purelin

La représentation schématique du perceptron multicouches (4 :6 :6 :1) pour notre réseau

Tableau III.21 : Matrice des poids et des biais du couch d'entrée

w1				b1
-0,908	-1,914	0,212	-0,869	2,189
-1,823	-0,292	-0,359	-1,407	0,815
0,548	1,487	-0,371	-2,029	-0,387
1,756	0,673	0,405	0,796	0,668
-0,856	-1,024	1,345	-1,090	-1,266
1,018	-1,729	-0,432	1,033	2,033

Tableau III.22 : Matrice des poids et des biais du couch caché

w2						b2
1,073	0,611	-0,612	1,003	-0,005	-1,045	-1,852
0,987	-1,058	-0,891	0,351	-0,558	0,940	-1,370
-0,286	0,940	0,813	0,958	0,221	0,823	0,546
-0,612	-0,059	1,433	-1,481	0,057	0,685	0,004
0,899	-0,849	-0,102	0,716	-0,567	-0,915	0,898
0,499	-1,176	0,565	0,542	-0,643	1,092	1,116
0,683	0,696	-0,911	-0,954	-0,611	0,787	1,955

Tableau III.23 : Matrice des poids et des biais du couch sortie

w3							b3
0,279	-0,298	-0,357	0,538	-0,188	-0,816	0,695	0,514

III.6.3. Modélisation de LogCMC

La figure III.12 présente la validation du modèle de réseau de neurone, on remarque bien que les points expérimentaux se rapprochent de la première bissectrice.

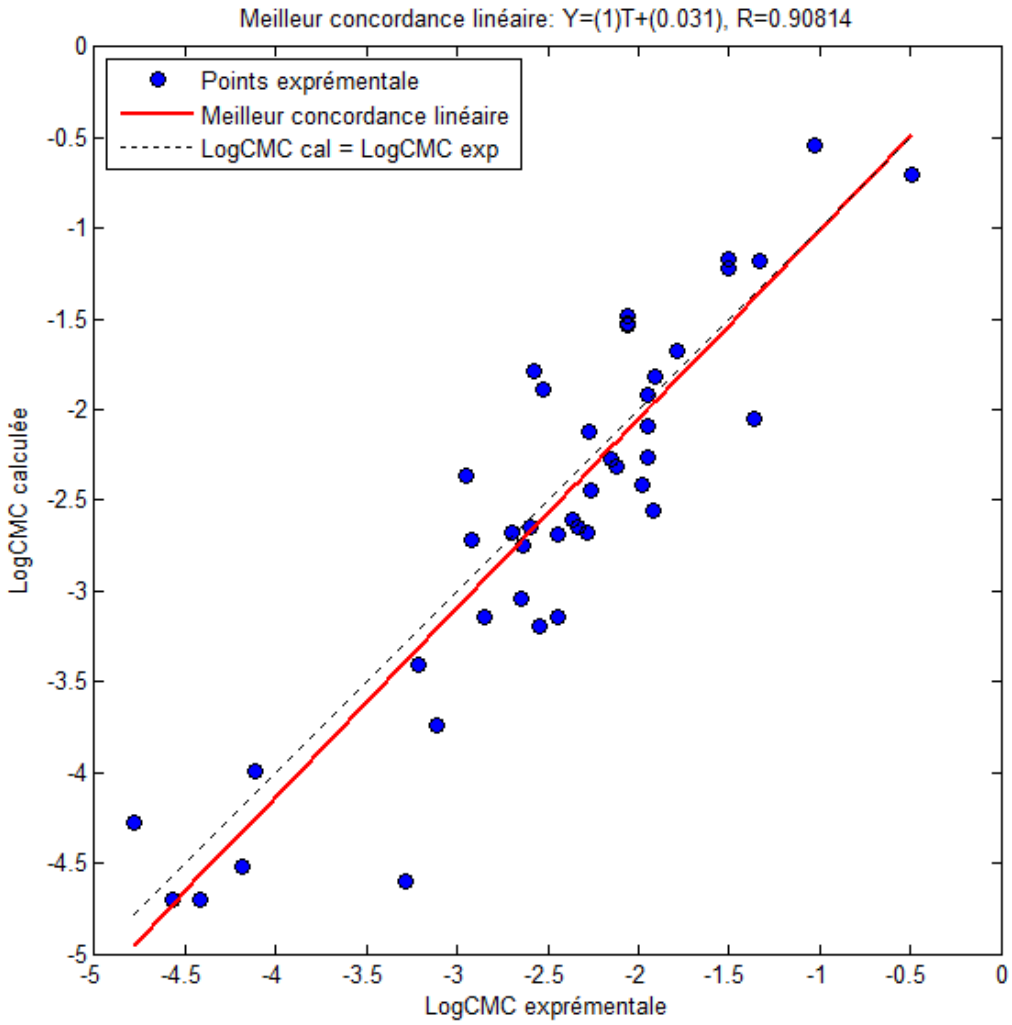


Figure III.12 : Validation de Log CMC déterminé par les réseaux de neurones avec les valeurs expérimentales

Nous constatons que le modèle de réseau de neurones valide parfaitement les résultats expérimentaux figure III.12.

Tableau III.24 : Valeurs des erreurs calculées pour LogCMC

ERAM%	R^2pred	R^2
17.235	0.976	0.908

Du tableau 24, on constate que le modèle rapporté est d'une qualité acceptable. Avec un coefficient de régression R^2= 0,908 et de R^2pred =0.976.

L'histogramme de la figure III.13, présente l'erreur absolue entre les points expérimentaux et le modèle de réseau de neurone.

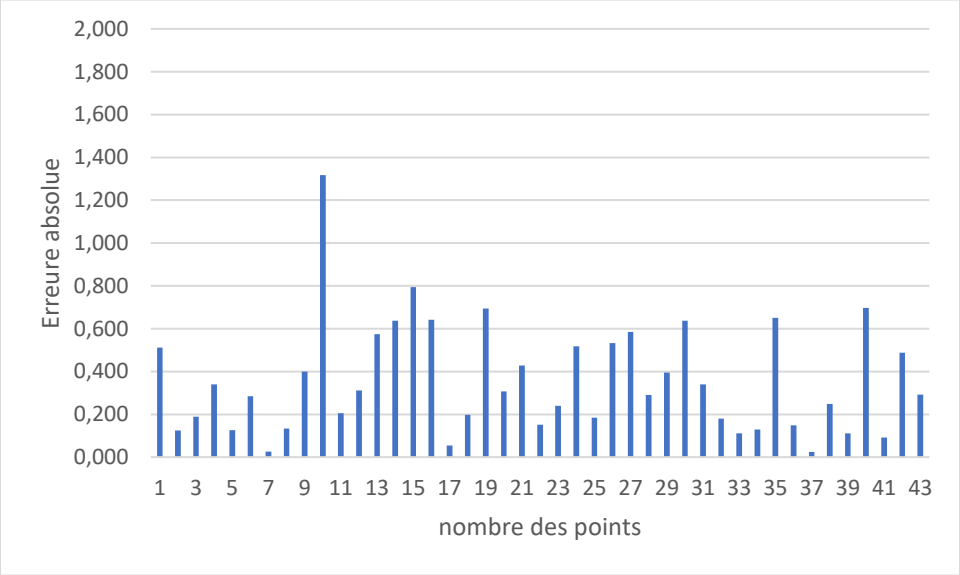


Figure III.13 : Présentation de l'erreur absolue

D'après la figures III.13 qui représente l'erreur absolue entre les résultats prédit par RNA et les résultats expérimentaux de la base de données de teste, on constate que parmi 171 points expérimentaux on a uniquement un sel points qui possèdent des erreurs très élevées.

Le tableau III.25 illustre les valeurs minimale et maximale de log CMC qui représente le domaine d'application de modèle neuronal, on contrepartie on constate que les minimum et maximum entre les valeurs calculée et expérimentales sont très proches

Tableau III.25 : les valeur min et max du Log CMC

	Min	Max
Log CMC calculée	-4,698	-0,538
Log CMC expérimentale	-4,786	-1,025

Conclusion

Le but de ce travail est de modéliser les propriétés physico-chimiques de 171 tensioactifs de différents types (anionique, cationique, non-ionique et globale) par quatre modèle hybrides de type QSPR- réseau de neurones artificiel (RNA). Le nombre des entrées du modèle neuronale a été calculé par le logiciel Padel et optimisée par la méthode des algorithmes génétiques. Les quatre modèles neuronaux conçus présentent un triple avantage :

- Prévoir avec une très bonne concordance (exactitude) les mesures expérimentales
- Un seul réseau permet de modéliser à la fois trois tensioactifs de différents types (anionique, cationique, non-ionique)
- Concurrence les modèles physiques.

Les caractéristiques des quatre réseaux de neurones artificiels sont présentées comme suit :

1. Le logCMC de 51 tensioactifs non-ioniques avec un modèle neuronal qui a une architecture de {4-4-4-1}, ce dernier a été trouvé avec un coefficient de corrélation égale R=0.991, une Erreur Relative Absolue Moyenne ERAM=23.815 et R^2_{pred}=0.998. Ces résultats vérifient la capacité prédictive externe de notre modèle en confirmant la robustesse du modèle établi.
2. Le second modèle comporte une architecture de {4-7-7-1}, a été optimisé afin de calculer les agents de surface de 88 structures cationique avec un R=0.987, ARDM=5.806 et R^2_{pred}=0.995. Ces résultats montrent la robustesse du modèle établi.
3. Le troisième réseau de neurones à une structure de {4-5-5-1}, ce dernier montre son aptitude à prédire logCMC de 71 composés anioniques avec un R=0.927, ARDM=16.744 et R^2_{pred}=0.979.

4. Le dernier, c'est un réseau global d'architecture {4-6-6-1} englobant un mélange de 171 tensioactifs de différents types (anionique, cationique, non-ionique), le réseau a été trouvé avec un R=0.908, ARDM=17.235 et R^2_{pred}=0.976.

En conséquence les réseaux de neurones peuvent être utilisés comme des estimateurs robustes et de concurrence pour la prédiction des paramètres de la CMC.

Conclusion générale

L'étude réalisée s'inscrit dans le cadre de la compréhension des mécanismes qui influe sur la concentration micellaire critique des tensioactifs par le changement structural d'une molécule. L'objectif principal était de modéliser les propriétés physico-chimiques de 171 tensioactifs de différents types (anionique, cationique, non-ionique et globale) par quatre modèle hybrides de type QSPR- réseau de neurones artificiel (RNA).

Le nombre des entrées du modèle neuronale a été calculé par le logiciel Padel et optimisée par la méthode des algorithmes génétiques. Les quatre modèles neuronaux conçus présentent un triple avantage :

- Prévoir avec une très bonne concordance (exactitude) les mesures expérimentales
- Un seul réseau permet de modéliser à la fois trois tensioactifs de différents types (anionique, cationique, non-ionique)
- Concurrence les modèles physiques.

Les caractéristiques des quatre réseaux de neurones artificiels sont présentées comme suit :

1. Le logCMC de 51 tensioactifs non-ioniques avec un modèle neuronal qui a une architecture de {4-4-4-1}, ce dernier a été trouvé avec un coefficient de corrélation égale R=0.991, une Erreur Relative Absolue Moyenne ERAM=23.815 et R^2_{pred}=0.998. Ces résultats vérifient la capacité prédictive externe de notre modèle en confirmant la robustesse du modèle établi.
2. Le second modèle comporte une architecture de {4-7-7-1}, a été optimisé afin de calculer les agents de surface de 88 structures anioniques avec un R=0.987, ARDM=5.806 et R^2_{pred}=0.995. Ces résultats montrent la robustesse du modèle établi.
3. Le troisième réseau de neurones à une structure de {4-5-5-1}, ce dernier montre son aptitude à prédire logCMC de 71 composés anioniques avec un R=0.927, ARDM=16.744 et R^2_{pred}=0.979.
4. Le dernier, c'est un réseau global d'architecture {4-6-6-1} englobant un mélange de 171 tensioactifs de différents types (anionique, cationique, non-ionique), le réseau a été trouvé avec un R=0.908, ARDM=17.235 et R^2_{pred}=0.976.

En conséquence les réseaux de neurones peuvent être utilisés comme des estimateurs robustes et de concurrence pour la prédiction des paramètres de la CMC.

Ce modeste travail est loin d'être achevé. Les perspectives de recherche qui se dégagent de cette étude sont résumées dans ce qui suit :

- Modélisation des propriétés thermodynamique des trois types des tensioactifs synthétiques
- Modélisation des points de Krafft et de trouble pour les tensioactifs anioniques et non ioniques.
- Modélisation des HLB des tensioactifs

Nomenclature

CMC. Concentration micelle critique

QSPR : quantitative structure propriété relationship

MLR : les régressions multilinéaires

PLS : les régressions aux moindres carrés partiels

ERA : erreur relative absolue, (%): $ERA = \left(\frac{|x_{mes} - x_{cal}|}{x_{mes}} \right).100$

EA : erreur absolue : $EA = |x_{mes} - x_{cal}|$

RNA : réseau de neurones artificie.l

RN : réseau de neurones.

RNSO : réseau de neurones statique optimale

TANSIG : Tangent sigmoïde hyperbolique, $a = \frac{e^{+n} - e^{-n}}{e^{+n} + e^{-n}}$

XiN : grandeurs normalisées des valeurs d'entées et de sortie

GA : génétique algorithme

Reference bibliographies

[1] Li, G.Z.; Sui, H.; Zhu, W.Z. Progress of surfactant study [in Chinese]. China Surfact. Deterg. Cosm. **1999**, 1, 24–26.

[2] M. J. Rosen, Surfactants and Interfacial Phenomena, Wiley, New Jersey, 2004.

[3] B. Jonsson, B. Lindman, K. Holmberg, B. Kronberg, Surfactants and Polymers in Aqueous Solution, Wiley, New York, 1999.

[4] V. Uskoković, M. Drofenik, Reverse micelles: Inert nano-reactors or physicochemically active guides of the capped reactions, Adv. Colloid. Interfac. **133 (**2007) 23–34.

[5] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Willy–VCH, Weinheim, 2000.

[6] T. Isaksson, T. Fearn, T. Davies, T. Næs, A User-Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Charlton Mill, 2004.

[7] R. G. Brereton, Chemometrics — Data Analysis for the Laboratory and Chemical Plant, Wiley, Atrium, 2003.

[8] B. Hemmateenejad, M. Yazdani, QSPR models for half-wave reduction potential of steroids: A comparative study between feature selection and feature extraction from subsets of or entire set of descriptors, Anal. Chim. Acta. **634** (2008) 27–35.

[9] Menger, F.M. and C.A. Littau, J. Am. Chem. Soc., 1991. 113: p. 1451-1453.

[10] Marques, E.F., et al., Adv. in colloïde. et Inter. Sci., 2003. 100-102 : p. 83-104.

[11] Soussan, E., thèse de l'Université Paul Sabatier, 2007.

[12] Masuda, M., V. Vill, and T. Shimizu, J. Am. Chem. Soc., 2000. 122: p. 12327-12333.

[13] Zhao , J., S.D. Christian, and B.M. Fung, J. Phys. Chem. B, 1998. 102: p. 7613-7618.

[14] Bellocq, A.M., J. Phys. IV France, 1999. 9.

[15] Mitchell, D.J., et al., J. Chem. Soc. Faraday Trans. 1 1983. 79: p. 975-979.

[16] Israelachvili, J.N., Intermolecular & Surface Forces, Academc Press 2nd edition,

[17] Holmberg, K., J¨onsson, B., Kronberg, B., Lindman, B., Eds. Surfactants and Polymers in Aqueous Solution 2nd edition; JohnWiley & Sons Ltd, Chichester, 2003.

[18] Evans, D. F., Wennerstr¨om, H., Eds. The Colloidal Domain, 2nd edition; Wiley-VCH, Weinheim, 1999.

[19] G.N. Kustova, E.B. Burgina, V.A. Sadykov, and S.G. Poryvaev, Vibrational Spectroscopic Investigation of the Goethite Thermal Decomposition Products, PhysChem Minerals 1992, 18:379-382

[20] Rahal S, Khalladi R, Moulai-Mostefa N (2016) Solubilization of Crude Oil by Extended and Other Anionic Surfactants. Arab J Sci Eng 41:111–117.

[21] Schwertmann U., Cornel R.M., (1991) « Iron Oxides in the laboratory preparation and Characterization », edition VCH.benzènesulfonate de sodium par diffusion centrale des neutrons » physics abstracts 61.12 page 432

[22] Jolivet J.P., (1994). « L'interface oxyde-solution »', chapitre 6 dans 'De la solution à l'oxyde', Inter Editions, CNRS Editions, Paris p 254.

[23] Belattar S., Debbache N., (2009) « réactivités thermiques et photochimiques du fer en solution aqueuse vis-à-vis de molécules organiques en phase homogène et hétérogène». Mémoire de Magister Uinv Mentouri Départ de Chimie Constantine (Algérie).

[24] Ming-Chun L., Chia N., (2000) « Oxidation of chlorophenols with hydrogen peroxide in the presence of goethite » Chemosphere 40 125-130

[25] Healy T.W., Somasundaran P., Fuerstenau D.W. (2003) « The adsorption of alkyl and alkylbenzene sulfonates at mineral oxide-water interfaces » Int. J. Miner. Process. 72

[26] Karaborni S., Esselink K. et Smit B. 1996

[27] Scriban, R. : Biotechnologie : restauration par voie biologique des solscontaminés par les polluantes organiques. 5èmes éditions. Ed Lavoisier, (1999).

[28] Paria, S.: Surfactant-enhanced remediation of organic contaminated soil and water, *Advan. Colloid Interface Sci*, 138, 24-58, (2008).

[29] Lee, M. H.; Kang, W. D.: Application of nonionic surfactant-enhanced in situ: flushing to a diesel-contaminated site, *Water Res*., 39, 139-146, (2005).

[30] Pearl M., Pruijn M., Bovendeur J, The application of soil washing to the remediation of contaminated soils, Land Contamination& Reclamation, 14 (3), 713-726(2006).

[31] Pennell K.D., Abriola L.M., Weber W.J., Surfactant-enhanced solubilization of residual dodecane in soil columns : 1. Experimental Investigation. Environ. Sci. Technol. 27, 2332-2340(1993).

[32] La pollution des sols par les hydrocarbures par Jean-Miche l BERGU E nov.-déc 1986 Réf. 3141

[33] Khalladi R., Benhabiles O., Bentahar F., Moulai-Mostefa N., Surfactant remediation of diesel fuel polluted soil. j. hazard. Mater. 164, 1179-1184, (2009).

[34] Marée noires et sols polluées par les hydrocarbures enjeux environnementaux et traitement des pollutions par Christian Bocard Editions technip, Paris 2006.

[35] Bera, A. ; Ojha, K. ; Kumar, T . ; Mandal. ; A. : Phase Behavior and Physicochemical Properties of (Sodium DodecylSulfate + Brine + Propan-1-ol + Heptane) Microemulsions"*J. Chem. Eng. Data*, 57 (3) 1000–1006 (2012).

[36] Hoar, T. P.; Schulman, J. H.: Transparent Water-in-Oil Dispersions: the Oleopathic Hydro-Micelle. Nature. 152, 102–103, (1943)

[37] Rosen M.J., Surfactants and interfacial phenomena, 3rd ed., Wiley, New Jersey, (2004).

[38] Dreyfus, G., Martinez, J.-M., Samuelides, M., Gordon, M.B., Badran, F., et Thiria, S. Apprentissage statistique, Eyrolles, (2008).

[39] Bevan D.R. QSAR and Drug Design, http://www.netsci.org/Science/Compchem/feature12.html

[40] Hansch, C., Leo, A., Hoekman, D. (1995) Exploring QSAR - Hydrophobic, Electronic, and Steric Constants. American Chemical Society, Washington, D.C.

[41] Wold S. (1991). Validation of QSARs, QSAR, 10, 191-193

[42] H.T. Phuong, Thèse de doctorat, Synthèse et étude des relations structure/activité quantitatives (QSAR/2D) d'analogues Benzo[c]phénanthridiniques, Université d'Angers, 2007

[43] A. Crum-Brown, T.R. Fraser, Trans. R. Soc. Edinburgh, 25, 151 (1868). C. Richet, C. R. Séances Soc. Biol. Ses. Fil., 9, 775 (1893).

[44] H. Meyer Zur Theorie der Alkoholnarkose, Arch. Exp. Pathol. Pharmakol., 42, 109 (1899).

[45] E. Overton, Studien über die Narkose. Zugleichnein Beitrag zur allgemeinen Pharmakologie. Jena: Gustav Fischer, Germany, 1901.

[46] L.P. Hammett, J. Am. Chem. Soc., 59, 96 (1937).

[47] R.W. Taft, J. Am. Chem. Soc., 74, 3120 (1952).

[48] C. Hansch, P.P. Maloney, T. Fujita, R.M. Muir, Nature, 194, 178 (1962).

[49] T. Fujita, J. Iwasa, C. Hansch, J. Am. Chem. Soc., 86, 5175 (1964).

[50] C. Hansch, T. Fujita, J. Am. Chem. Soc., 86, 1616 (1964).

[51] Y.C. Martin, Quantitative drug design, Marcel Dekker, . New-York, 1978.

[52] S.M. Free, J. Wilson, J. Med. Chem., 7, 395 (1964).

[53] C.D. Selassie. In: Abraham DJ, Ed. Burger's medicinal chemistry and drug discovery. 6th ed, John Wiley & Sons Inc., New-York, 2003.

[54] H. Hosoya, Bull. Chem. Soc. Jpn., 44, 2332 (1971).

[55] H. Wiener, J. Am. Chem. Soc., 69, 17 (1947).

[56] M. Randić, J. Am. Chem. Soc., 97, 6609 (1975).

[57] A.T. Balaban, Chem. Phys. Lett., 89, 399 (1982).

[58] Z. Simon, Angew. Chem. Int. Ed. Eng., 13, 719 (1974).

[59] L.B. Kier, L.H. Hall, Molecular connectivity in chemistry and drug research, Academic Press, New-York, 1976.

[60] N. Margossian, Le règlement REACH - La règlementation européenne sur les produits chimiques, Dunod / L'usine Nouvelle, Paris, 2007.

[61] A. Crum Brown, T.R. Fraser, On the connection between chemical constitution and physiological action. Part 1. On the physiological action of salts of the ammonium bases, derived from strychnia, brucia, thebia, codeia, morphia and nicotia.Trans. Roy.Soc. Edinburgh, 1868, 25, 151-203.

[62] L.P. Hammett, The effect of structure upon the reactions of organic compounds. Benzene derivatives. J. Am. Chem. Soc. 1937, 59, 96-103.

[63] A.C. Brown, and T.R. Fraser, On the connection between chemical constitution and Physiological Action; with special reference to the physiological action of the salts of the ammonium bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia, J. Anat. Physiol . 1868, 2, 224-242.

[64] C. Hansch, A. Leo, D. Hoekmann, Exploring QSAR: hydrophobic, electronic and steric constants.Washington, DC: Am. Chem. Soc.1995.

[65] S.M. Free, J.W. Wilson, A mathematical contribution to structure–activity studies. J. Med. Chem.1964, 7, 395-399.

[66] M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure- Property relationships in pharmaceutical research Part2. Pharm. Sci. Tech. Today, 2000, 3, 50-57.

[67] A.R.D. Katritzky, C. Fara, R.O. Petrukhin, D.B. Tatham, U. Maran, A. Lomaka, M.Karelson, The present utility and future Potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors, Curr. Top. Med. Chem. 2002, 2,1333-1356.

[68] J. Ghasemi, S. Saaidpour, S.D. Brown, QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. J. Mol. Struct. (Theochem). 2007, 805, 27-32.

[69] P. Geladi, B.R. Kowalski, Partial Least Squares Regression: a Tutorial, Anal. Chim. Acta. 1986, 185, 1-17.

[70] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, An introduction to decision tree modeling, J. Chemom. 2004, 18, 275-285.

[71] A.F. Duprat, T. Huynh, G. Dreyfus, Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of LogPJ. Chem. Inf. Comput. Sci. 1998, 38, 586-594.

[72] J. Gasteiger, J. Zupan, Neural Networks in Chemistry, Angew. Chem. Int. Ed. Engl. 1993, 32, 503-527.

[73] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review
J. Chemometr. 2001, 15, 559-569.

[74] A. Fortuné, Thèse de doctorat, Techniques de Modélisation Moléculaire Appliquées à l'Etude et à l'Optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance, Université Joseph Fourier – Grenoble I, 2006.

[75] Y. Dodge, Statistique Dictionnaire encyclopédique, Springer-Verlag France, Paris, 2007.

[76] I. Charif, Elaboration des corrélations quantitatives structure-activité des acides carbonés. Etude théorique des effets de solvants sur les équilibres libres cétoénoliques des composés b-dicarbonyles cycliques, Thèse de doctorat, Université de Tlemcen, 2012.

[77] H. Wiener, Structural determination of paraffin boiling points. J. Am. Chem. Soc. 1947, 69, 17–20.

[78] M. Randic, Characterization of molecular branching. J. Am. Chem. Soc. 97, 6609–6615, J. Am. Chem. Soc. 1975, 97, 6609-6615.

[79] L.B. Kier, L.H. Hall, W.J. Murray, Molecular connectivity I: Relationship to nonspecific local anesthesia. J. Pharm. Sci. 1975, 64, 1971–1974

[80] A.T. Balaban, Highly Discriminating Distance-Based Topological Index, Chem.Phys. Lett. 1982, 89, 399-404.

[81] L. Zhang, H. Zhu, T.I. Oprea, A. Golbraikh, A. Tropsha, QSAR Modeling of the Blood-Brain Barrier Permeability for Diverse Organic Compounds, Pharm. Res.2008, 25, 1902–1914.

[82] L. He, P.C. Jurs, Assessing the reliability of a QSAR model's predictions. J. Mol.Graph. Model. 2005, 23, 503-523.

[83] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb.Sci. 2003, 22, 69-77.

[84] P.P. Roy, S. Paul, I. Mitra, K. Roy, On Two Novel Parameters for Validation of Predictive QSAR models, Molecules, 2009, 14, 1660-1701.

[85] A. Golbraikh, A. Tropsha, Beware of q2! J. Mol. Graph. Model. 2002, 20, 269–276.

[86] T. M. Martin, P. Harten, D. M. Young, E.N. Muratov, A. Golbraikh, H. Zhu and A.Tropsha, Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? J. Chem. Inf. Model . 2012, 52, 2570−2578.

[87] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R.Todeschini, D. Fourches, Varnek A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena-pyriformis: focusing on applicability domain and overfitting by variable selection. J. Chem.Inf. Model. 2008, 48, 1733

[88] J. Jaworska, N.N. Jeliazkova, T. Aldenberg: QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. Altern. Lab. Anim 2005, 33, 445-459.

[89] T. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M. Cronin, P. Gramatica, J. Jaworska, S. G. Kahn, C. Klopman, G. Marchant, N.N. Myatt, G.Jeliazkova, R.

Patlewicz, D. Roberts, T. Schultz, D. Stanton, J. Sandt, W. Tong, G. Veith, C. Yang, Current status of methods for defining the applicability domain of (Quantitative) Structure–Activity Relationships. Altern. Lab. Anim. 2005, 33, 1-19.

[90] A. R. Katritzky, M. Kuanar, S. Slavov, and C. D. Hall, Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction, Chem. Rev. 2010, 110, 5714–5789

[91] PLS-regression: a basic tool fo chemometrics de Svante Wold, Michael Sjöström, Lennart Eriksson

[92] Data mining et statistique décisionnelle de Séphane Tufféry

[93] Probabilité, analyse des données et Statistique de Gilbert Saporta

[94] La régression PLS: théorie et pratique, M. Tenenhaus

[95] Introduction à la régression PLS de Carole BINARD

[96] The Elements of Statistical Learning de Trevor Hastie, Robert Tibshirani et Jerome Friedman

[97] Hagan, M.T., Demuth, H.B. et Beale, M. (1996). Neural network design. PWS Publishing Company, Boston, Massachusetts.

[98] Haykin, S. (1999). Neural networks A comprehensive foundation, 2nd ed.: Prentice Hall.

[99] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function: Math. Control, Signais and Systems, 2, 303-314.

[100] Van der Baan and Jutten, 2000

[101] Hebb, D. (1949). The organization of the behavior. Wiley, New York.

[102] Grossberg, S. (1982). Studies ofmind and brain. D. Reidel Publishing Co., Boston.

[103] Rumelhart, D.E., Hinton, E. et Williams, J. (1986). Learning internai representation by error propagation. Dans Parallel distributed processing. Vol.l. MIT Press, Cambridge, Massachusetts. pp. 318-362.

[104] Weiss, S.M. et Kulikowski, C.A. (1991). Computer systems that leam. Morgan Kaufmann Publishers Inc., San Mateo, Californie.

[105] Hassoum, M.H. (1995). Fundamentals of artificial neural networks. MIT Press, Cambridge.

[106] Baum, E. et Haussier, D. (1989). What size net gives valid generalization? Neural Computation, 1(1): 151-160.

[107] Coulibaly, P., Anctil, F. et Bobée, B. (1998). Real time neural network- based forecasting system for hydropower reservoirs. Proceedings of the First International Conference on New Information Technologies for Decision Making in Civil Engineering, (Montreal, 1998), Vol. 2. École de Technologie Supérieure, Montréal. pp. 1001-1011.

[108] Zurada, J.M. (1992). Introduction to artificial neural systems. West Publishing Co., Saint Paul, Minnesota.

[109] Widrow, B. et Lehr, M.A. (1990). 30 years of adaptive neural networks: perceptron, madaline and backpropagation. Proceedings ofthe IEEE, 78(9): 1415-1442.

[110] Murata, N.M., Yoshizawa, S. et Amari, S. (1994). Network information criterion determining the number of hidden units for an artificial neural network mode!. IEEE Transactions on Neural Networks, 5(6): 865-872.

[111] Jacobs, R.A., Peng, F. et Tanner, M.A. (1997). A Bayesian approach to mode! selection in hierarchical mixtures-of-experts architectures. Neural Networks, 10(2): 231-241.

[112] Rudolph, S. (1997). On topology, size and generalization of non-linear feed-forward neural networks. Neurocomputing, 16: 1-22.

[113] Hecht-Nielsen, R. (1987). Komogorov's mapping neural network existence theorem. Dans Proceedings of the International Conference on Neural Networks. Vol. 3. IEEE Press, New York. pp. 11-13.

[114] Haykin, S. (1994). Neural Networks: A Comprehensive Foundation, Upper Saddle River, NJ: Prentice Hall.

[115] Rumelhart, D.E., Hinton, E. et Williams, J. (1986). Learning internai representation by error propagation. Dans Parallel distributed processing. Vol.l. MIT Press, Cambridge, Massachusetts. pp. 318-362.

[116] Roy, P.P., Leonard, J.T., Roy, K., 2008. Exploring the impact of the size of training sets for the development of predictive QSAR models. Chemom. Intell. Lab. Syst. 90,31–42.

[117] Roy, K., Mandal, A.S., 2008. Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones. J. Enzyme Inhibition Med. Chem. 23 (6), 980–995.

[118] Roy, P.P., Roy, K., 2008. On some aspects of variable selection for partial least squares regression models. QSAR Comb. Sci. 27, 302–313.

[119] Roy, P.P., Paul, S., Mitra, I., Roy, K., 2009. On two novel parameters for validation of predictive QSAR models. Molecules 14, 1660–1701.

[120] Ojha, P.K., Mitra, I., Das, R.N., Roy, K., 2011. Further exploring rm2 metrics for validation of QSPR models. Chemom. Intell. Lab. Syst. 107, 194–205.

[121] Zhang, L., Zhu, H., Oprea, T., Golbraikh, A., Tropsha, A., 2008. QSAR modeling of the blood–brain barrier permeability for diverse organic compounds. Pharm. Res. 25, 1902–1914.

[122] Atkinson, A.C., 1985. Plots, Transformations and Regression. Clarendon Press, Oxford.

[123] Hadjazi, S. et Hadjazi. S, "Etude et modélisation des paramètres physico-chimique de la ciniétique de fermentation de l'oxytétracycline, PFE", Centre universitaire de Médéa .Promotion 2003.

[124] A.DEMEYER, F.JACOB, M.JAY, G.MENGUY, J.PERIER**,** "La conversion bioénergétique du rayonnement solaire et les biotechnologies", Technique et Documentation – Lavoisier, 1982.

[125] D.Hassani, " Méthode des Réseaux Neurones pour le Calcul de la Dispersion Axiale dans les Llits Fixes", Mémoire de magister, Université de Médéa, Algérie, (2002).

[126] C.Touzet, "Les Réseaux de Neurones Artificiels: Introduction au Connexionnisme", (1992).

[127] L. Mokhnache, "Application de RN dans le Diagnostique et la Prédiction des Isolations HT", Thèse de Doctorat, Ecole Nationale Polytechnique d'Alger, Janvier (2004).

[128] M.R.J. Fombellida, M.J.M. Minsoul et J, L.O. Destine "Perceptrons Multi-Couches et fonctions d'activation non-monotones", In. Proc. Neuronimes'90, EC2 et Cie, Nîmes, (1990).

[129] E. Davalo, P. Naïrn,"Des Réseaux de Neurones", Eyrolles, 2ème édition, Mars. (1993).

[130] Rosen, M.J., 1978. Surfactants and Interfacial Phenomena, third edition Wiley, New York.

[131] Huibers, P.D.T., Lobanov, V.S., Katritzky, A.R., Shah, D.O., Karelson, M., 1996. Prediction of critical micelle concentration using a quantitative structure– property relationship approach. Langmuir 12, 1462–1470.

[132] Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 27-37

[133] Ghose, A.K. and Crippen, G.M. , Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity, Journal of Computational Chemistry, 1986, 7:565-577;

Ghose, A.K. and Crippen, G.M. , Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions, Journal of Chemical Information and Computer Science, 1987, 27:21-35