



REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTRE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABOU-BEKR BELKAID - TLEMCCEN

MEMOIRE

Présenté à :

FACULTE DES SCIENCES – DEPARTEMENT DE CHIMIE

Pour l'obtention du diplôme de :

MASTER

Filière : **Chimie**

Option: Chimie Théorique et Computationnelle

Par :

Mme OUZLIFI FADIA

Sur le thème

ETUDE QSAR DE LA TOXICITE D'UNE SERIE DE COMPOSES PHENOLIQUES VIS-A-VIS DES ALGUES MARINES

Soutenu publiquement le 25 juin 2019 à Tlemcen devant le jury composé de :

Mme CHEMOURI Hafida	Maître de Conférences A	Université de Tlemcen	Président
Mr CHARIF Imad Eddine	Maître de Conférences A	Université de Tlemcen	Examineur
Mme KARA ZAITRI Lamia	Maître Assistant A	Université de Tlemcen	Examineur
Mme BELLIFA Khadidja	Maître de Conférences B	Université de Tlemcen	Encadreur

*Laboratoire de Thermodynamique Appliquée et Modélisation Moléculaire (LATA2M), N°53
BP 119, 13000 Tlemcen - Algérie*

Dédicaces

Un grand merci pour mes chers parents qui s'inquiétaient toujours pour moi, qui sans eux je n'aurais pas pu arriver jusque-là, à mes sœurs qui ont été toujours là pour le meilleur et pour le pire et dont je n'arrive pas à trouver les mots convenables pour exprimer mon admiration, à mon frère qui m'a toujours entouré par la bonne humeur.

Merci finalement à mon mari qui était toujours présent pour m'aider, pour me soulager la déchéance dont j'avais vécu de temps en temps et pour me donner un nouveau souffle qui me soutient de nouveau et me repousse un pas en avant, à toute sa famille qui sont d'ailleurs une deuxième famille pour moi, j'avais connu une seconde mère, un second père et même des grands frères et une sœur si gentilles et très compréhensifs.

Mes remerciements sont destinés à tous mes amis, mes collègues et à toute ma promotion d'étude.

A tous ceux qui me sont chers...

Remerciements

Le travail présenté dans ce mémoire a été réalisé dans le laboratoire de thermodynamique appliquée et modélisation moléculaire (LATA2M) de l'université A. Belkaid de Tlemcen dirigé par Mademoiselle la professeur NEGADI Latifa.

J'adresse mes chaleureux remerciements à Mme Errahoui-BellifaKhadidja pour m'avoir confié le sujet de ce travail. Je tiens vivement à la remercier pour son investissement et sa célérité pour son attention à se soumettre à mes questions pressantes, ses réponses sont désormais partie intégrante de ce mémoire. Je la remercie pour la grande autonomie qu'elle m'a accordée. J'espère m'être montré digne de sa confiance.

J'adresse ma plus grande et respectueuse reconnaissance à Mme CHEMOURI HAFIDA pour avoir accepté de présider le jury et pour le temps consacré à l'examen de ce mémoire.

Je tiens également à remercier Monsieur Imad Charif et Mme KARA ZAITRI Lamia pour m'avoir fait l'honneur d'examiner ce travail et d'avoir participé au jury.

Un grand remerciement à Monsieur le professeur Mekkeleche qui a inculqué en nous l'amour de la science. Grand respect pour votre personne.

Un autre grand merci à Mme BENCHOUK pour leur disponibilité à répondre à tout questionnement.

SOMMAIRE

INTRODUCTION GENERALE.....	1
CHAPITRE I : GENERALITE-TOXICITE-AQUATIQUE	
Introduction	5
I. GENERALITES	5
1. LES PHENOLS.....	5
-Quelques exemples des phénols	6
1-les phénols simples.....	6
1.1-les chlorophénols.....	6
2-les polyphénols	6
I. 2. Toxicité aquatique des phénols	7
I.3. les algues	7
I.3.1. définition des algues	7
- Types d'algues	7
1. les cyanobactéries.....	7
2. les algues eucaryotes.....	7
-croissance des algues	8
- importance des algues	10
I.3.2. mesure de la toxicité des produits chimiques vis -à -vis des algues	11
I.3.2.1. principe	11
II. Etudes QSAR de la toxicité des phénols vis-à-vis des algues	12
CHAPITRE II : BASES THEORIQUES – partie A- Méthodes QSAR	
Introduction.....	18
II.1-Principe des méthodes QSAR	18
II.2- méthodologie générale d'une étude QSAR.....	18
II.2.1- élaborer le modèle.....	18
II.2.2- validation du modèle.....	18
II.2.3 détermination du domaine d'applicabilité du modèle	19

1-Elaboration du modèle QSAR	19
a .Base de données	19
b .Descripteur moléculaire.....	19
.Descripteurs théoriques (1D ,2D, 3D).....	19
.Descripteurs empiriques (physico-chimique).....	19
-coefficient de partage	20
-mesure et estimation de log p.....	20
c. Méthodes d'analyse des données	21
2-Interpretation et validation d'un modèle QSAR/QSPR	21
a. Validation interne	21
-validation croisée.....	21
-Y-Randomisation	21
b. Validation externe	21
3- Domaine d'applicabilité.....	22
CHAPITRE II : BASES THEORIQUES – partie B- Méthodes QUANTIQUES	
II-Les méthodes quantiques	25
II.1.methodes semi-empiriques	25
- CNDO.....	25
- INDO.....	25
- MNDO/3.....	25
-MNDO.....	26
-AM1.....	26
-PM3.....	26
II.2.Théorie de la fonctionnelle de Densité (DFT)	26
II.2. 1.le premier théorème de Hohenberg et Kohn (1964).....	27
II.2.2.Deuxième théorème de Hohenberg et Kohn.....	27
II.2. 3.Méthode de Kohn et Sham.....	28
II.2.4. Terme d'échange –corrélation	28
II.2.5.L'approximation locale LDA.....	29
II.2. 6. L'approximation des gradients généralisés GGA	29

II.2.7. Fonctionnelle hybride B3LYP.....	30
--	----

CHAPITRE II : BASES THEORIQUES – partie C- Méthodes statistiques

II.3.Méthodes d’analyse statistiques.....	33
---	----

II.3.1-Méthodes de Régression linéaire (SLR, MLR).....	33
--	----

II.3.1.1. Régression linéaire simple (RLS).....	33
---	----

1- Méthode d’estimation des paramètres β_0 et β_1	34
---	----

- Critère des Moindres Carrés Ordinaires » (MCO, Ordinary Least Squares).....	34
---	----

2- Analyse de la variance.....	35
--------------------------------	----

II.3.1.2. Régression linéaire multiple (MLR).....	37
---	----

Estimation des paramètres statistiques du modèle.....	37
---	----

Ecart type des coefficients de régression (erreurs sur les coefficients de régression).....	38
---	----

Confiance à accorder aux résultats	38
--	----

-Test de la signification globale de la régression	38
--	----

Test de Fisher –Snedecor.....	38
-------------------------------	----

Test de signification de chaque paramètre (chaque descripteur) t-student.....	39
---	----

- Identificateurs de qualité d’une régression linéaire..... 40
 - ❖ Coefficient de détermination R^2 40
 - ❖ Déviation standard (SD)..... 40

Sélection de variables et choix du modèle dans le cas d’une régression multiple	41
---	----

Critère de comparaison de modèle	41
--	----

Coefficient de détermination R^2	41
--	----

coefficient de détermination ajusté $R^2_{ajusté}$	41
--	----

II.3.2. Méthodes des moindres carrés partiels PLS	42
---	----

1- Définition.....	42
--------------------	----

2- Principe de la méthode PLS.....	42
------------------------------------	----

- Maximiser la corrélation
- Maximiser la variance

3-Construction de la première composante $h=1$	43
--	----

4-construction des composantes suivantes	43
--	----

5- choix de la dimension du modèle	44
--	----

6- critère de choix de nombre de composantes	44
➤ Critère de 'validation croisée'	44
• La validation croisé (LOO-CV).....	44

CHAPITRE III : RESULTATS ET DISCUSSION

Introduction.....	47
III.1 Méthodologie.....	48
III.1.1 Base de données.....	48
III.1.2 Calcul des descripteurs.....	49
III.1.3 Analyse statistique.....	55
III.2 Résultats et Discussion.....	55
III.2.1 Modèles QSAR avec un seul descripteur.....	56
III.2.2 Modèles QSAR avec deux descripteurs.....	60
III.2.4 Etude du meilleur modèle.....	61
III.2.4.1 Validation du meilleur modèle QSAR.....	62
III.2.4.2 Domaine d'applicabilité du meilleur modèle QSAR.....	64
Conclusion.....	65
Références Bibliographiques.....	
CONCLUSION GENERALE.....	67

LISTE DES ABREVIATIONS

AM1 Austin Model 1

B3LYP Becke3-ParameterLee-Yang-Parr

CGTO Contracted Gaussian Type Orbital Configuration

CLOA Combinaison Linéaire d'Orbitales Atomiques

DA Domaine d'applicabilité

DFT Density Functional Theory

DM Dipole Moment

GA Genetic Algorithm

GG Generalized Gradient Approximation

GTO Gaussian Type Orbital

HF Hartree-Fock

HOMO Highest Occupied Molecular Orbital

LDA Local Density Approximation

LMO Leave Many Out

LOO Leave One Out LUMO Lowest Unoccupied Molecular Orbital

MLR Multiple Linear Regression

MOA Mode of Action

OA Orbitale Atomique

OM Orbitale Moléculaire

PCA Principal Composante Analysis

PLS Partial Least Squares

QSPR Quantitative Structure-Property Relationships

QSAR Quantitative Structure Activity Relationship

SCF: Self Consistent Field STO Slater Type Orbital

SLR Simple Linear Regression

SSR Sum of Squares Regression

SSE Sum of Squares Error

Introduction générale

INTRODUCTION GENERALE

Avec l'avènement de la technologie scientifique moderne et de l'industrialisation, l'utilisation de produits chimiques s'accroît parallèlement à l'augmentation du nombre de nouveaux produits chimiques synthétisés chaque jour [1]. Beaucoup de ces composés sont libérés dans l'environnement et s'accumulent dans presque tous les milieux naturels, en particulier dans les systèmes aquatiques.

Il est donc bénéfique d'étudier sérieusement le danger potentiel de ces produits pour les organismes aquatiques.

L'analyse expérimentale de la toxicité chimique au moyen de tests sur ces espèces est généralement un processus long, coûteux et techniquement difficile [2-4]. Pour ces raisons, les méthodes QSAR [5] pourraient être utilisées comme solutions alternatives pour expliquer la toxicité des produits chimiques vis-à-vis les algues dans le but de conserver cette fortune naturelle.

Les phénols sont classés parmi les produits qui dépassent 1tonne par an (1tpa) et possédant une toxicité, a divers effets néfastes sur les organismes aquatiques, tels que l'inhibition de la croissance et de la reproduction, des modifications des intermédiaires du métabolisme et des lésions tissulaires. Actuellement, ces composés figurent sur la liste des polluants prioritaires détenu par la US Environmental Protection Agency (USEPA) [6]. Il a été rapporté que les dérivés du phénol sont parmi les produits chimiques les plus couramment impliqués dans les déversements accidentels en milieu marin [7].

Les micro-algues sont les représentants les plus communs des producteurs primaires dans un écosystème. Les effets nocifs qui s'accumulent chez ces organismes peuvent affecter les réseaux trophiques connexes et même les écosystèmes aquatiques entiers [8]. Leur sensibilité aux produits chimiques polluants est généralement égale ou supérieure à celle des représentants des consommateurs primaires et secondaires (tels que les daphnies et les poissons) [9]. En raison de ces caractéristiques, les micro-algues sont particulièrement bien adaptées aux organismes d'essai utilisés pour la recherche en toxicologie et l'évaluation des risques pour l'environnement liés aux produits chimiques.

La toxicité des produits phénoliques vis-à-vis des algues a été le sujet de recherche de plusieurs auteurs [10-12].

Ertürk et al. [11] ont étudié la toxicité d'une série de phénols vis-à-vis d'une algue marine, *Dunaliella tertiolecta*, en utilisant les techniques MLR et NN. Les descripteurs utilisés ont été calculés en utilisant les logiciels CODESSA [13] et DRAGON [14], les modèles obtenus ont montré des prédictions acceptables même si la signification physique des descripteurs concernés et leur corrélation avec la toxicité ne sont pas toujours clairement et rationnellement expliquée. Ertürk et al. [12] ont également modélisé la toxicité d'une série de phénols vis-à-vis l'algue *vulgaris Chlorella* en utilisant l'approche MLR, les résultats ont révélé que les modèles QSAR établis fournissent des prévisions acceptables ($R^2= 0,84$; $SD =0,20$) pour les narcotiques polaires et les découpleurs de la phosphorylation oxydative, mais ils sont incapables de prédire la toxicité des phénols réactifs présentant un mécanisme électrophile. Récemment, Bakire et al. [15] ont publié un article sur la toxicité (exprimée en EC_{50} pour une durée d'exposition de 72h) de 348 produits chimiques y compris des phénols vis-à-vis d'une algue marine *C. vulgaris* (EC_{50-72h}). Ils ont pu élaborer des modèles pour les différents modes d'action des composés étudiés. Les meilleurs modèles pour cette étude ont été obtenus en combinant des descripteurs topologiques (2D) et le coefficient de partage.

L'objectif ciblé pour ce présent travail est d'étudier la toxicité de 30 phénols vis-à-vis deux algues marines *Chlorella vulgaris* et *Dunaliella tertiolecta* en utilisant des descripteurs pertinents de la chimie quantique et en respectant toute la méthodologie QSAR et tous les critères d'OECD.

Les travaux décrits dans ce mémoire font l'objet de trois chapitres, auxquels s'ajoutent la présente introduction et une conclusion générale.

- Dans le premier chapitre nous exposons des généralités sur les phénols et les algues ainsi le problème de toxicité des phénols et leur impact sur le milieu aquatique.
- Le deuxième chapitre regroupe toutes les bases théoriques nécessaires pour aboutir à un modèle QSAR, ce chapitre est donné en trois parties :
 - Les méthodes QSAR

Introduction générale

- Les méthodes de la chimie quantique
 - Les méthodes statistiques d'analyse des données
- Le troisième chapitre est consacré aux résultats obtenus et leur discussion

Références bibliographiques

- [1] J.R. Seward, E. Hamblen, T.W. Schultz, Regression comparisons of *Tetrahymena pyriformis* and *Poecilia reticulata* toxicity, *Chemosphere*, 47 (2002) 93-101.
- [2] J. Michałowicz, W. Duda, Phenols--Sources and Toxicity, *Polish Journal of Environmental Studies*, 16 (2007).
- [3] K. Khan, E. Benfenati, K. Roy, Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: Ranking and prioritization of the DrugBank database compounds, *Ecotoxicology and environmental safety*, 168 (2019) 287-297.
- [4] K.E. Hevener, D.M. Ball, J.K. Buolamwini, R.E. Lee, Quantitative structure-activity relationship studies on nitrofuranyl anti-tubercular agents, *Bioorganic & medicinal chemistry*, 16 (2008) 8042-8053.
- [5] A.A. Toropov, A.P. Toropova, I. Raska Jr, D. Leszczynska, J. Leszczynski, Comprehension of drug toxicity: Software and databases, *Computers in Biology and Medicine*, 45 (2014) 20-25.
- [6] W. Du, F. Zhao, B. Zeng, Novel multiwalled carbon nanotubes-polyaniline composite film coated platinum wire for headspace solid-phase microextraction and gas chromatographic determination of phenolic compounds, *Journal of Chromatography A*, 1216 (2009) 3751-3757.
- [7] I. Cunha, S. Moreira, M.M. Santos, Review on hazardous and noxious substances (HNS) involved in marine spill incidents—An online database, *Journal of hazardous materials*, 285 (2015) 509-516.
- [8] A. Gerofke, P. Kömp, M.S. McLachlan, Bioconcentration of persistent organic pollutants in four species of marine phytoplankton, *Environmental Toxicology and Chemistry: An International Journal*, 24 (2005) 2908-2917.
- [9] A. Weyers, G. Vollmer, Algal growth inhibition: effect of the choice of growth rate or biomass as endpoint on the classification and labelling of new substances notified in the EU, *Chemosphere*, 41 (2000) 1007-1010.
- [10] S. Gökçe, M.T. Saçan, Assessments of Algal Toxicity and PBT Behaviour of Pesticides with No Eco-toxicological Data: Predictive Ability of QSA/(T) R Models, *Molecular informatics*, (2019).
- [11] M.D. Ertürk, M.T. Saçan, M. Novic, N. Minovski, Quantitative structure-activity relationships (QSARs) using the novel marine algal toxicity data of phenols, *Journal of Molecular Graphics and Modelling*, 38 (2012) 90-100.
- [12] M.D.Erturk, M.T.saçan, Assessment and modeling of the novel toxicity data set of phenols to *Chlorella vulgaris*, *Ecotoxicology and Environmental Safety*, 90 (2013) 61-68.
- [13] A.R. Katritzky, R. Petrukhin, H. Yang, M. Karelson, CODESSA PRO, User's manual. University of Florida, (2002).
- [14] R. Todeschini, V. Consonni, DRAGON professional 5.4 program, TALETE srl, Milano, Italy, http://www.talete.mi.it/products/dragon_description.htm, (2006).
- [15] S. Tisseron, *La résilience*, Presses universitaires de France(2007).

Chapitre I

Généralités-Toxicité aquatique

Introduction

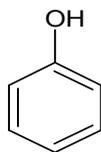
L'utilisation industrielle, agricole et domestique des produits chimiques est en augmentation dans le monde entier. Cependant une évaluation et une caractérisation sont nécessaires pour estimer les éventuels effets néfastes sur la santé et l'environnement [1-5]. Parmi ces produits qui suscitent beaucoup d'attention à cause de leur toxicité sont les produits en grande production >1tpa tels que le phénol et ses dérivés.

Les phénols sont en production depuis les années 1860 [6]. Ils ont été utilisés comme intermédiaires dans la synthèse de colorants, de pigments, de composés phénoliques d'acides aminés et d'acides aminés de pesticides et d'herbicides, ainsi que dans de nombreuses industries et produits de consommation [7, 8]. Ils sont devenus de plus en plus répandus et distribués dans la nature [8, 9]. En raison de leur persistance dans l'environnement et de leurs effets néfastes sur les organismes vivants, on s'intéresse à la détermination de leurs dangers potentiels pour divers organismes aquatiques, y compris les bactéries, les daphnies, les protozoaires les poissons et les algues [10].

1- Généralités

1.1- Les phénols :

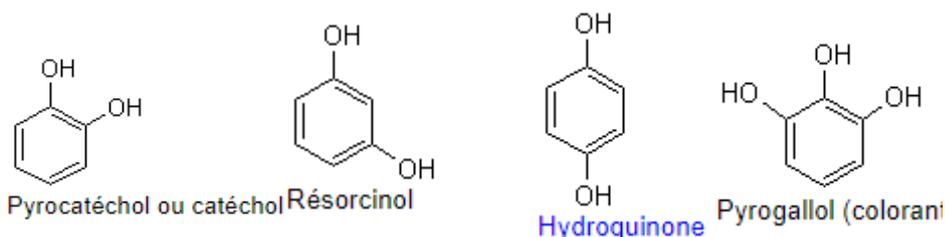
Les phénols sont des composés chimiques ont un cycle aromatiques monosubstitué portant une fonction hydroxyle -OH. Le phénol est présent dans de nombreux végétaux, dans les laboratoires, il est couramment utilisé comme un produit chimique pour diverses activités. Il peut être trouvé directement en tant que réactif commercial ou bien préparé dans le cadre d'un mélange dans le laboratoire.



Phénol

Figure 1: Structure moléculaire du phénol

- Quelques exemples de phénols :



Le terme composé phénolique englobe une grande variété de substances possédant un ou plusieurs noyaux aromatiques, substitués par un ou plusieurs groupements hydroxyles et / ou méthoxyles [11], on distingue :

1- Les phénols simples

Les phénols simples englobent toutes les molécules hydroxylées diversement substituées de noyau du phénol. Selon cette définition plusieurs phénols appartiennent à ce groupe, par exemple: le phénol, les chlorophénols, les alkylphénols, les acides hydroxybenzoïque, les nitrophénols...

1.1. Les chlorophénols

Les chlorophénols sont des composés organiques dans lequel le noyau phénolique est substitué en différentes manières par un ou plusieurs atomes de chlore remplaçant l'hydrogène du cycle aromatique.

Les chlorophénols servent pour produire des intermédiaires dans la synthèse des phénols plus chlorés ou de dérivés des chlorophénols comme les herbicides, les colorants, les pigments et les résines phénoliques.

2- Les polyphénols

Les polyphénols sont des composés phénoliques à haut poids moléculaires. Ils se composent d'un ou plusieurs cycles benzéniques portant un ou plusieurs groupements hydroxyles et autres constituants [12].

Ils sont présents dans diverses substances naturelles: sous forme d'anthocyanine dans les fruits rouges, de proanthocyanidines dans le chocolat, d'acides cafeoylquinique et

feruloylquinique dans le café, de flavonoïdes dans les agrumes, de catéchines comme l'épigallo-catéchine gallate du thé vert, de quercitrine dans les pommes, ...etc [13]. Le polyphénol se présente comme un composé de base, constituant plusieurs produits synthétiques simples possédant une fonction alcool supplémentaire comme l'hydroquinone, catéchol et sous forme de polyphénols polymériques comme les colorants, les plastiques et les résines à base de bisphénol.

I.2. Toxicité aquatique des phénols

Au cours des dernières années, de nombreux groupes de recherche se sont concentrés sur le groupe hydroxyle phénolique en raison de son large rayon d'activité. La toxicité élevée de ces composés est dangereuse pour la santé humaine ainsi que pour la faune et la flore. L'élimination inappropriée de ces substances dans le système aquatique peut entraîner l'extinction de nombreuses espèces.

Le processus d'évaluation des produits chimiques synthétisés ou importés plus d'1tpa/h nécessite des informations écotoxicologiques de base, notamment des données de toxicité à court terme sur les algues.

Les algues jouent un rôle crucial dans l'écosystème et dans l'évaluation réglementaire des risques car elles fournissent de la nourriture aux niveaux trophiques supérieurs et constituent donc la base des réseaux trophiques. Malgré leur distribution dans les écosystèmes aquatiques et de leurs avantages pour les tests en laboratoire, les données de toxicité fiable sur les algues sont limitées.

La détermination de la toxicité vis-à-vis d'espèces non ciblées telles que les algues est bénéfique pour comprendre l'impact des phénols sur les écosystèmes.

I.3. Les algues

1- Définition des algues

Le mot algue vient du latin Alga, elles font partie des premières espèces vivantes, notamment des premiers végétaux apparus après les débuts de la vie [14]. Les algues

Chapitre I : Généralités-Toxicité aquatique des phénols

constituent un groupe très diversifié de plantes sans floraison qui vivent principalement dans l'eau.

Une algue est un organisme ayant la capacité d'effectuer la photosynthèse en obtenir son énergie du soleil, mais elle est différente d'une embryophyte ou d'une plante terrestre.

Types d'algues

Les algues sont classées en comparant leurs pigments, leurs formes de réserve cellulaire, leurs parois cellulaires, leurs flagelles et les organites intracellulaires. On distingue deux grandes familles d'algues :

1- **Les Cyanobactéries** (algues bleues) : sont des organismes procaryotes (dont le matériel génétique est contenu dans la zone cytoplasmique). Elles appartiennent au règne des *Eubactériae*.

2- Les algues eucaryotes :

- Les algues rouges : sont pour la plupart pluricellulaires et vivent surtout dans les eaux marines.

- *Les algues vertes* : peuvent être uni- ou pluricellulaire. Elles appartiennent à un vaste ensemble, les Chlorobiontae (organismes verts).

- *Les Glaucophytae* : sont de petits flagellés unicellulaires d'eau douce, dont les plastes sont colorés en bleu-vert par des pigments surnuméraires. Ils appartiennent au règne des *Plantae* ou *Primpoplastobiotae* caractérisé par des chloroplastes à deux membranes.

- *Les Euglénophyceae* : sont de petites algues unicellulaires flagellées, incolore ou possédant des chloroplastes, vivant en eau douce essentiellement.

- Les Cryptophytae : sont des algues unicellulaires flagellées d'eau douce ou marine, incolores ou aux plastes rouges ou bleus. Elles ont un appareil flagellaire particulier, un ensemble de plaques ceinturant la cellule d'une invagination antérieure.

- Les Chlorarachniophyceae : sont des amiboflagellés pourvus de chloroplastes.

Chapitre I : Généralités-Toxicité aquatique des phénols

- Les Haptophytae : possèdent des plastes colorés en brun-doré par les caroténoïdes et un organe particulier l'haptonème*.
- Les algues brun-dorées (Ochromyphytae) dont les plus connues sont les algues brunes (Phaeophyceae). Ces dernières sont des algues pluricellulaires marines.
- Les Diatomées : sont des algues unicellulaires avec une coque de silice que l'on trouve dans tous les milieux.
- Les Dinophytae : constituent le plancton quelquefois luminescent grâce aux bactéries qu'elles contiennent.

Croissance des algues

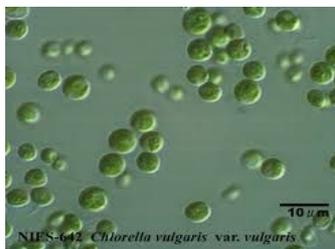
Les micro-algues se multiplient par division cellulaire et ont donc un taux de croissance très rapide. Ce taux de croissance, de dédoublement varie d'une espèce à l'autre et certaines se reproduisent plus rapidement. Le tableau ci-dessous regroupe le taux de croissance et le temps de doublement pour les micro-algues les plus connues et les plus utilisées pour les tests toxicologiques.

Tableau.1. Taux de croissance et de doublement des algues marines

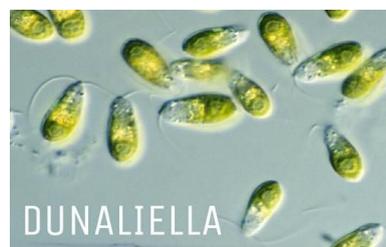
Algue	Taux de croissance (J ⁻¹)	Temps de doublement (Jour)
<i>Botriococcusbraunii</i>	0,2	3,4
<i>Chlamydomonas reinhardii</i>	3,8	0,18
<i>Chlorellavulgaris</i>	1,84	0,37
<i>Dunatiellatertiolecta</i>	3,5	0,2
<i>Isochrysisgalbana</i>	2	0,34
<i>Navicula muralis</i>	2,63	0,26
<i>Pleurochrysisarterae</i>	0,65	1,1
<i>Rhodomonas salina</i>	0,6	1,15
<i>Spirulinaplatansis</i>	0,5	1,38
<i>Thalassiosirapseudonana</i>	2,48	0,27

Importance des algues

- Les algues marines sont une source importante de substances naturelles bioactives.
- Elles ont été utilisées comme produits alimentaires dans l'alimentation asiatique depuis des siècles car elles contiennent des caroténoïdes, des fibres alimentaires, des protéines, des acides gras essentiels, des vitamines et des minéraux
- Les algues ont reçu beaucoup d'attention pour leur potentiel antioxydant naturel et la plupart de leurs composés ont montré des activités antibactériennes. Antifongique, antivieillessement, diététique, anti-inflammatoire, cytotoxiques, antipaludéens, antiprolifératives et des propriétés anticancéreuses
- Les algues sont des organismes écologiquement importants dans les écosystèmes aquatiques et sont fréquemment utilisées dans les études environnementales pour évaluer la toxicité relative de divers produits chimiques organiques
- D'un point de vue écologique, les algues constituent la base des réseaux trophiques (c. -à-d. les producteurs primaires); ils fournissent de la nourriture et produisent de l'oxygène qui est vital pour la durabilité de la vie dans les milieux aquatiques. Elles jouent un rôle crucial dans les écosystèmes, leur répartition omniprésente dans le monde entier, leur facilité de collecte et de culture, et leur croissance rapide les rendent idéales pour les tests de laboratoire [15-17].
- Le phytoplancton est la base de la chaîne alimentaire aquatique. A ce titre, les organismes phytoplanctoniques sont sensibles aux contaminants et polluants de toutes natures, mais ils sont aussi, à leur insu, le point de départ du transfert des contaminants le long de la chaîne trophique.



Chlorella vulgaris



Dunaliella tertiolecta

Figure.1. Images de micro-algues

Mesure de la toxicité des produits chimiques vis-à-vis des algues

La toxicité d'un produit chimique vis-à-vis d'une algue est quantifiée par la concentration inhibitrice (EC_{50}) ce cette dernière durant une période d'exposition.

Principe

Une culture d'une algue en croissance exponentielle est exposée, dans un système de microtitration sans renouvellement des solutions, à différentes concentrations d'une substance d'essai ou à une série de dilutions d'un effluent ou d'un mélange, pendant plusieurs générations, dans des conditions définies. On compare la croissance des algues exposées à celle d'algues témoins **durant une période déterminée** (48h, 72h, 96h). Une substance d'essai est considérée comme toxique lorsqu'on observe une inhibition statistiquement significative de la croissance des algues en fonction de la concentration.

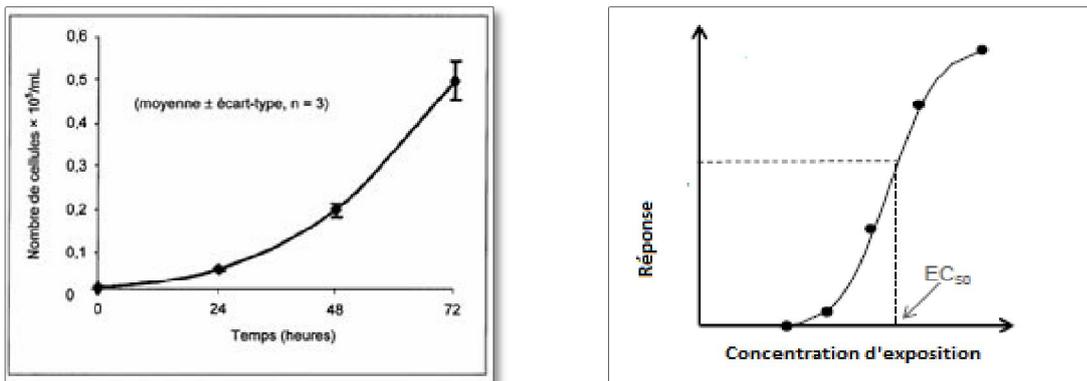


Figure.2. Mesure de la concentration inhibitrice de la croissance d'une algue pour une durée d'exposition de 72h

2. Etudes QSAR de la toxicité des phénols vis-à-vis des algues

Aujourd'hui, les QSAR s'établissent et sont acceptés pour l'évaluation des risques des produits chimiques[18].

Certaines études récentes décrivent l'utilisation des QSAR pour le risque environnemental [19] et l'utilisation réglementaire des QSAR aux États-Unis, au Canada et en Europe.

Un grand nombre de travaux sur la toxicité aquatique des phénols en utilisant la méthode QSAR est publié.

Erturk et al. [20] ont déterminé la toxicité d'une série de composés phénoliques pour l'algue marine *Dunaliella tertiolecta*. De plus, ils ont établi plusieurs relations interspécifiques en utilisant les données sur la toxicité des algues marines et les données de la littérature sur la toxicité des chlorophénols pour des organismes de différents niveaux trophiques, y compris une bactérie (*Vibriofischeri*), un protozoaire (*Tetrahymena pyriformis*), une daphnie (*Daphnia magna*), une algue dulcicole (*Pseudokirchneriella subcapitata*), et un poisson (*Pimephales promelas*).

Les résultats suggèrent que la toxicité des chlorophénols augmente avec l'augmentation de nombre du chlore. Cependant, la position de l'atome de chlore sur le cycle aromatique est également importante dans la toxicité des chlorophénols. Le chlore -ortho-substitué a entraîné une toxicité moindre que les congénères substitués en -méta- et -en para-substitués en plus il a été constaté que la toxicité des chlorophénols ayant tendance à diminuer entre 48h et 96h. Cela résulte probablement d'une augmentation du pH du milieu d'essai, qui a entraîné une formation accrue de chlorophénols ionisés, ce qui a à son tour contribué à une toxicité inférieure à celle de la forme neutre correspondante. Un autre facteur qui pourrait être lié à la diminution de la toxicité est l'acclimatation des algues aux produits chimiques testés. Malgré les changements dans la toxicité selon la durée d'exposition, les données sur la toxicité des algues marines se sont avérées bien corrélées avec celles de cinq espèces aquatiques, quelle que soit la durée des essais.

Chapitre I : Généralités-Toxicité aquatique des phénols

Ces chercheurs dans une autre étude [21] ont étudié la toxicité de 30 phénols sur *Chlorella vulgaris* pour une durée d'exposition de 96h. L'étude a révélé que la toxicité des algues des narcotiques polaires et des découpleurs respiratoires était fortement corrélée au paramètre d'hydrophobicité à un pH donné, $\log D$, ce qui montre l'importance de l'ionisation dans le système de test de *C. vulgaris* tandis que les composés réactifs nécessitent la présence d'un autre descripteur comme l'indice d'énergie d'activation (AEI). Les résultats ont montré que le modèle de surface-réponse était hautement interprétable et fournissait des prévisions acceptables pour les narcotiques polaires et les découpleurs respiratoires, bien qu'il n'était pas assez fiable pour prédire la toxicité des phénols réactifs. Dans deux autres modèles QSAR utilisant des descripteurs topologiques, WHIM (Weighted Holistic Invariant Molecular), et Tm (T total size index/weighted by atomic masses), ont révélé des résultats prometteurs qui pourraient être utilisés pour prévoir la toxicité algale de composés autres que les phénols comme les anilines.

La toxicité de 30 composés phénoliques [21] a été déterminée dans le même laboratoire par le même chercheur en utilisant *Chlorella vulgaris* comme organisme d'essai.

L'analyse de la composition mécanistique de l'ensemble de données présentée dans cette étude, fondée sur les règles établies par Schultz et ses collaborateurs [22, 23], a révélé que 19 composés sur les 30 devraient présenter une toxicité algale par **narcose polaire**. De plus, quatre découpleurs respiratoires et sept produits chimiques réactifs (c. -à-d. ceux qui agissent par des mécanismes d'action électrophiles) ont été identifiés.

Erturk et al. [20] ont utilisé l'algue marine *Dunaliella tertiolecta* pour étudier la toxicité de 30 composés phénoliques. Ils ont déduit que la toxicité des narcotiques polaires et des découpleurs respiratoires était fortement corrélée à l'hydrophobicité décrite par le logarithme du coefficient de partage octanol/eau ($\log P$). Il a été démontré que les composés censés agir par des mécanismes plus réactifs, en particulier les hydroquinones, présentaient une toxicité supérieure à celle prévue par le $\log P$. Un modèle QSAR a été obtenue avec le $\log P$ et un descripteur topologique d'autocorrélation 2D assimilé à la

polarisabilité atomique (MATS3p) après élimination des hydroquinones de l'ensemble des données. Ils ont constaté que la toxicité des algues marines des narcotiques polaires et des découpleurs respiratoires était fortement corrélée à l'hydrophobicité décrite par le log P. Les composés qui devraient agir par des mécanismes plus réactifs, en particulier les hydroquinones, ont montré une toxicité supérieure à celle prévue dans le log P.

Dans une tentative de modéliser de la toxicité des phénols pour *D. tertiolecta* en utilisant uniquement des descripteurs 2D et logP, un QSAR de qualité utilisant le paramètre d'hydrophobicité (log P) et MATS3p, ce dernier étant un descripteur d'autocorrélation 2D pondéré par la polarisation atomique, a été obtenu seulement après retrait des hydroquinones des données. Suite à l'inclusion de descripteurs 3D dans le processus de modélisation, trois QSAR globaux ont été développés en utilisant MLR. Les analyses effectuées sur des sous-ensembles basés sur la nature des mécanismes impliqués dans l'ensemble des données (c. -à-d. les mécanismes non covalents et les mécanismes à médiation covalente) ont révélé que les descripteurs 3D-MoRSE, particulièrement celui pondéré par la polarisabilité atomique (c. -à-d. Mor24p), étaient fortement corrélés à la toxicité des phénols, sauf les hydroquinones pour les algues marines. Cette découverte suggère que la toxicité des phénols pour les algues marines *D. tertiolecta* est principalement contrôlée par la polarisabilité atomique et la structure 3D de la molécule. Les hydroquinones, dont on croit qu'elles agissent par des mécanismes réactifs d'action toxique sur le système d'essai de *D. tertiolecta*, ont été présentées comme des composés influents dans les QSAR développés.

Maran et al. [24], ont étudié la toxicité de 50 produits chimiques narcotiques non polaires en utilisant l'algue *Pseudokirchneriella subcapitata*. La plupart des composés testés sont des produits chimiques industriels à grand volume. Un modèle QSAR pour les composés narcotiques non polaires a été établi et comparé à des modèles analogues antérieurs. Les descripteurs du modèle ont été analysés et le domaine d'applicabilité a été évalué.

Chapitre I : Généralités-Toxicité aquatique des phénols

Bakire et al. [25] ont développé une série de modèles QSAR pour prédire l'inhibition de la croissance des algues après 72 h d'exposition pour une grande base de données de divers produits chimiques. Au total, 348 composés organiques, y compris 30 phénols, ont été classés en cinq modes d'action toxique selon la classification de Verhaar [26]. Ces travaux ont étudié l'inhibition de la croissance des algues et les modèles développés couvrent les cinq modes d'action.

Références bibliographiques

- [1] L. Gianfreda, G. Iamarino, R. Scelza, M.A. Rao, Oxidative catalysts for the transformation of phenolic pollutants: a brief review, *Biocatalysis and biotransformation*, 24 (2006) 177-187.
- [2] E.L. Cavalieri, K.-M. Li, N. Balu, M. Saeed, P. Devanesan, S. Higginbotham, J. Zhao, M.L. Gross, E.G. Rogan, Catechol ortho-quinones: the electrophilic compounds that form depurinating DNA adducts and could initiate cancer and other diseases, *Carcinogenesis*, 23 (2002) 1071-1077.
- [3] M. Megharaj, H. Pearson, K. Venkateswarlu, Toxicity of phenol and three nitrophenols towards growth and metabolic activities of *Nostoc linckia*, isolated from soil, *Archives of environmental contamination and toxicology*, 21 (1991) 578-584.
- [4] L.J. Kennedy, J.J. Vijaya, K. Kayalvizhi, G. Sekaran, Adsorption of phenol from aqueous solutions using mesoporous carbon prepared by two-stage process, *Chemical Engineering Journal*, 132 (2007) 279-287.
- [5] H. Razani, K. Nanba, S. Murachi, Acute toxic effect of phenol on zebrafish *Branchydanio rerio*, *Bulletin of the Japanese Society of Scientific Fisheries (Japan)*, (1986).
- [6] P.R. Duchowicz, A.G. Mercader, F.M. Fernández, E.A. Castro, Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR, *Chemometrics and Intelligent Laboratory Systems*, 90 (2008) 97-107.
- [7] C.J. Schmitt, J.L. Zajicek, T.W. May, D.F. Cowman, Organochlorine residues and elemental contaminants in US freshwater fish, (1976-1986), National Contaminant Biomonitoring Program, *Reviews of environmental contamination and toxicology*, Springer(1999) 43-104.
- [8] A.O. Aptula, D.W. Roberts, M.T. Cronin, T.W. Schultz, Chemistry– toxicity relationships for the effects of di- and trihydroxybenzenes to *Tetrahymena pyriformis*, *Chemical research in toxicology*, 18 (2005) 844-854.
- [9] J. Jensen, Chlorophenols in the terrestrial environment, *Reviews of environmental contamination and toxicology*, Springer(1996) 25-51.
- [10] K. Kaiser, Evolution of the international workshops on quantitative structure-activity relationships (QSARs) in environmental toxicology, SAR and QSAR in Environmental Research, 18 (2007) 3-20.
- [11] B. Charrière, Les composés phénoliques marqueurs de la matière organique terrestre dans deux écosystèmes marins: le delta du Rhône et le prodelta de la Têt, Toulouse, INPT, (1991).
- [12] L.A.P. Palencia, Chemical characterization, bioactive properties, and pigment stability of polyphenolics in açai (*Euterpe oleracea* Mart.), Texas A&M University(2009).
- [13] J. Holt-Lunstad, T.B. Smith, J.B. Layton, Social relationships and mortality risk: a

- meta-analytic review, PLoS medicine, 7 (2010) .
- [14] E. DE LA THESE, Hortense FALLER, UNIVERSITE DE LIMOGES, (1987).
- [15] N. Basant, S. Gupta, K.P. Singh, Predicting aquatic toxicities of chemical pesticides in multiple test species using nonlinear QSTR modeling approaches, Chemosphere, 139 (2015) 246-255.
- [16] E. Furusjö, A. Svenson, M. Rahmberg, M. Andersson, The importance of outlier detection and training set selection for reliable environmental QSAR predictions, Chemosphere, 63 (2006) 99-108.
- [17] M.T. Cronin, T.I. Netzeva, J.C. Dearden, R. Edwards, A.D. Worgan, Assessment and modeling of the toxicity of organic chemicals to *Chlorella vulgaris*: development of a novel database, Chemical research in toxicology, 17 (2004) 545-554.
- [18] T.W. Schultz, M.T. Cronin, J.D. Walker, A.O. Aptula, Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective, Journal of Molecular structure: THEOCHEM, 622 (2003) 1-22.
- [19] T. Öberg, A QSAR for the hydroxyl radical reaction rate constant: validation, domain of application, and prediction, Atmospheric Environment, 39 (2005) 2189-2200.
- [20] M.D. Ertürk, M.T. Saçan, First toxicity data of chlorophenols on marine alga *Dunaliella tertiolecta*: correlation of marine algal toxicity with hydrophobicity and interspecies toxicity relationships, Environmental toxicology and chemistry, 31 (2012) 1113-1120.
- [21] M.D. Erturk, M.T. saçan, Assessment and modeling of the novel toxicity data set of phenols to *Chlorella vulgaris*, Ecotoxicology and Environmental Safety, 90 (2013) 61-68.
- [22] W. Schultz, P. Dayan, P.R. Montague, A neural substrate of prediction and reward, Science, 275 (1997) 1593-1599.
- [23] B.F. Cook, Selected Roman and Italian acquisitions in the Department of Greek and Roman Antiquities, British Museum, (1980-1992), Journal of Roman Archaeology, 8 (1995) 225-234.
- [24] A.T. Peterson, J. Soberón, R.G. Pearson, R.P. Anderson, E. Martínez-Meyer, M. Nakamura, M.B. Araújo, Ecological niches and geographic distributions (MPB-49), Princeton University Press(2011).
- [25] S. Lozano, M.P. Halm-Lemeille, A. Lepaillieur, S. Rault, R. Bureau, Consensus QSAR related to global or MOA models: Application to acute toxicity for fish, Molecular informatics, 29 (2010) 803-813.
- [26] S. Bakire, X. Yang, G. Ma, X. Wei, H. Yu, J. Chen, H. Lin, Developing predictive models for toxicity of organic chemicals to green algae based on mode of action, Chemosphere, 190 (2018) 463-470.
- [27] H.J.M. Verhaar, J.L.M. Hermens, W. Mulder, QSAR for Ecotoxicity. EV5V-CT92-0211, RITOX, Utrecht, The Netherlands, (1995).

Chapitre II
Bases Théoriques

Chapitre II

Partie A-Méthodes QSAR

1. Introduction

Depuis le milieu du 19^{ème} siècle et à partir de la fin de 20^{ème} siècle, une appréciation que l'activité biologique est liée à la structure chimique surtout aux propriétés physico-chimiques est développée [1]. Les Relations Quantitatives Structure-Activité (QSAR) se rapportent à la construction de modèles prédictifs d'activités biologiques, en fonction de l'information de structure moléculaire d'une série de composés. C'est un procédé qui relie de manière quantitative les propriétés moléculaires appelées descripteurs, avec une activité biologique [2], suivant un modèle mathématique :

$$[\text{Activité biologique}] = f [\text{Descripteurs}]$$

Ces méthodes peuvent expliquer les données expérimentales disponibles et de prédire les activités pour des nouveaux composés ou des composés pour lesquels les données expérimentales ne sont pas disponibles

2. Principe des méthodes QSAR

Le principe des méthodes QSAR est de mettre en place une relation mathématique reliant de manière quantitative des propriétés moléculaires aussi bien électroniques que géométriques, appelées descripteurs, avec une activité biologique/toxicité, pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données en l'occurrence les régressions linéaires simple et multiple (SLR, MLR), régressions aux moindres carrés partiels (PLS) [3] méthode d'analyse en composantes principales (PCA) [4].

3. Méthodologie générale d'une étude QSAR

Pour aboutir à un modèle QSAR explicatif et prédictif on doit

1) *Elaborer le modèle*

- a- Collecter une base de données fiable
- b- Rechercher et calculer des descripteurs adéquats pour l'activité étudiée.
- c- Utiliser une méthode d'analyse des données pour relier l'activité aux descripteurs.

2) *Validation du modèle*

- a- Validation interne : on utilise la série d'apprentissage constituée de 2/3 de la base de données. Ce type de validation a pour but de vérifier la stabilité et le pouvoir explicatif du modèle QSAR retenu.
- b- Validation externe : on utilise la série de test constituée généralement de 1/3 de la base de données. Dans le but de vérifier le pouvoir prédictif du modèle élaboré.

3) Détermination du domaine d'applicabilité du modèle

1) Elaboration du modèle QSAR

a- Base de données :

Le choix de la base de données expérimentale initiale est une étape critique pour le développement des modèles QSAR, les données expérimentales doivent être obtenues en suivant le même protocole [5].

b- Descripteur moléculaire :

Un descripteur moléculaire est toute propriété (une valeur numérique) propre à une structure chimique donnée. Ces valeurs peuvent être obtenues expérimentalement ou calculées à partir de la structure de la molécule.

Les descripteurs moléculaires jouent un rôle fondamental dans les études de la relation quantitative structure activité. Ils sont utilisés en tant que variables indépendantes pour prédire une variable dépendante (activité) [6].

Historiquement, deux grands schémas pour la classification des descripteurs moléculaires ont été établis : l'un en fonction de leur origine (constitutionnel, topologique, géométrique, quantique, thermodynamique...), et un autre sur leur dimensionnalité (1D, 2D, 3D) [7].

- Descripteurs théoriques :

Les descripteurs 1-D :

Ces descripteurs sont calculés à partir de la formule brute de la molécule à l'aide de la composition moléculaire, c'est-à-dire les atomes qui la constituent, et ils représentent des propriétés générales telles que : les pourcentages massiques des atomes, la masse molaire, le poids moléculaire...

Les descripteurs 2-D :

Les descripteurs 2D sont obtenus à partir de la structure plane de la molécule. Ils sont classés en deux catégories, topologiques et constitutionnels [8].

Les descripteurs 3-D

Ce type de descripteurs nécessitent une conformation 3D de la molécule ; Ils sont évalués à partir des positions relatives de leurs atomes dans l'espace et décrivent des caractéristiques plus complexes leurs calculs nécessitent donc de connaître, le plus souvent par « modélisation moléculaire empirique » ou « ab-initio », la géométrie 3D de la molécule. Ces descripteurs sont très importants pour l'interprétation mécanistique des modèles QSAR élaborés.

- **Descripteurs empiriques (physico-chimique) :**

Les descripteurs physicochimiques, (ou indices physicochimiques) certains d'entre eux reflètent la composition moléculaire du composé (le nombre et le type d'atomes et de liaisons présents dans la molécule, le nombre de cycle, les propriétés donneur/accepteur de liaison H, cation, anion, etc....) [9]. D'autres représentent le caractère hydrophile ou lipophile de la molécule généralement évalué à partir du coefficient de partage Octanol/eau représenté par le log P [10].

-**coefficient de partage** : Le logP, aussi appelé log Kow, est une mesure de la solubilité différentielle de composés chimiques dans deux solvants (coefficient de partage octanol/eau).

Le logP est égal au logarithme du rapport des concentrations de la substance étudiée dans l'octanol et dans l'eau. $\log P = \log(C_{\text{oct}}/C_{\text{eau}})$

Le logP rend compte de la tendance d'une molécule à s'accumuler dans les membranes biologiques des organismes vivants. S'il est élevé, le risque de bioaccumulation est important. Le partage d'une molécule entre une phase aqueuse et une phase lipidique conditionne en partie ses propriétés biologiques telles que **le transport, le passage à travers les membranes,**

l'affinité pour un récepteur et la fixation par une protéine, **l'activité pharmacologique** ou encore **la toxicité**.

Mesure et Estimation de logP :

La méthode la plus classique et la plus fiable des méthodes de détermination du log P est la méthode dite du « flacon agité » [11]. Ce descripteur peut être obtenu également par les méthodes d'estimation en l'occurrence : méthode de Hansch [12], méthode de Rekker [13], méthode de Ghose et Viswanathan [14].

Plusieurs logiciels sont utilisés pour l'estimation de logP tels que ACD, chemdraw, Chemaxon, chemspider (online), Hyperchem

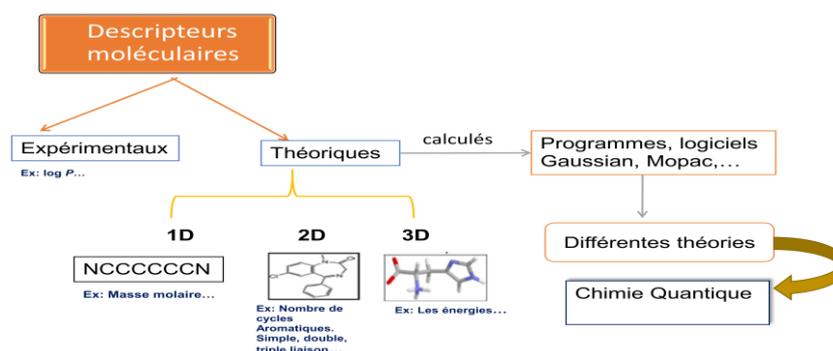


Figure.1. Les descripteurs moléculaires

c- Méthodes d'analyse des données :

Pour élaborer un modèle QSAR il faut une méthode d'analyse des données. Il existe plusieurs méthodes pour élaborer un modèle et analyser les données statistiques de ce dernier, on peut obtenir ces méthodes dans des logiciels tels que, Excel, OriginMicrocal, Minitab, Statistica...(Chapitre II.Partie C)

2) Interprétation et validation d'un modèle QSAR/QSPR

a- Validation interne :

La stabilité et la fiabilité d'un modèle QSAR peut être vérifiée par

- **Validation croisée** : elle consiste à retirer à chaque fois une ou plusieurs molécules de la série d'apprentissage et élaborer le modèle avec les molécules restantes, ensuite faire la prédiction de l'activité de la molécule retirée. ce procédé est répété pour retirer et prédire toutes les molécules de la série d'apprentissage. Cette méthode est quantifiée par R^2_{cv} (LOO/LMO). Si R^2_{cv} est proche de R^2 , le modèle est jugé stable.

- **Y-Randomization** : c'est une méthode de vérification de la crédibilité du modèle QSAR obtenu [15]. Le principe de Y-randomisation est de mélanger la l'activité expérimentale d'une manière aléatoirement pour la série d'apprentissage en utilisant les mêmes descripteurs et d'obtenir des modèles QSAR de mauvaise qualité, c'est-à-dire avec des performances plus faibles à celles obtenues avec le vrai modèle $R^2_r \ll R^2$.

R^2_r : coefficient de détermination pour les modèles obtenus avec les y-randomisés

R^2 : coefficient de détermination pour le modèle obtenu avec les vraies activités.

b- Validation externe

La validation externe consiste à estimer l'activité d'une série de composés dite série de test qui ne sont pas inclus dans la série d'apprentissage. Le principe de cette technique est basée sur le calcul du coefficient de corrélation R^2 (test) et vérifier les critères de Tropsha (Tropshacriteria) [16].

D'après Tropsha

$R^2 > 0,7$ Critère 1

$R^2 > 0,6$ Critère 2

$\frac{R^2 - R^2_0}{R^2} < 0,1$ et $0,85 \leq K \leq 1,15$ critère 3

$$\frac{R^2 - R_0^2}{R^2} < 0,1 \text{ et } 0,85 \leq K' \leq 1,15 \quad \text{critère 4}$$

$$|R^2 - R_0^2| \leq 0,3 \quad \text{critère 5}$$

Avec

R^2 Coefficient de corrélation pour la série de test.

R_0^2 coefficient de corrélation entre les valeurs prédites et expérimentales pour la série de test.

R_0^2 coefficient de corrélation entre les valeurs expérimentales et prédites pour la série de test.

K : est la constante de la droite (à l'origine) de corrélation (valeurs prédites en fonction des valeurs expérimentales)

K' : est la constante de la droite (à l'origine) de corrélation (valeurs expérimentales en fonction des valeurs prédites)

3) Domaine d'applicabilité

La méthode de « leverage » est la méthode la plus utilisée pour la détermination du domaine d'applicabilité d'un modèle QSAR [17, 18] ou l'intervalle dans lequel nos composés pourront être prédits avec confiance. Cette technique est basée sur la variation des résiduels standardisés de la variable dépendance avec « leverage » (la distance entre les valeurs des descripteurs et leurs moyennes). Le principe est que le leverage d'un composé doit être inférieur à la valeur de seuil $h^* = 3p/n$ (ou p est le nombre de descripteurs plus 1 et n le nombre de points). Dans le cas contraire le composé sera considéré en dehors du domaine d'applicabilité du modèle obtenu.

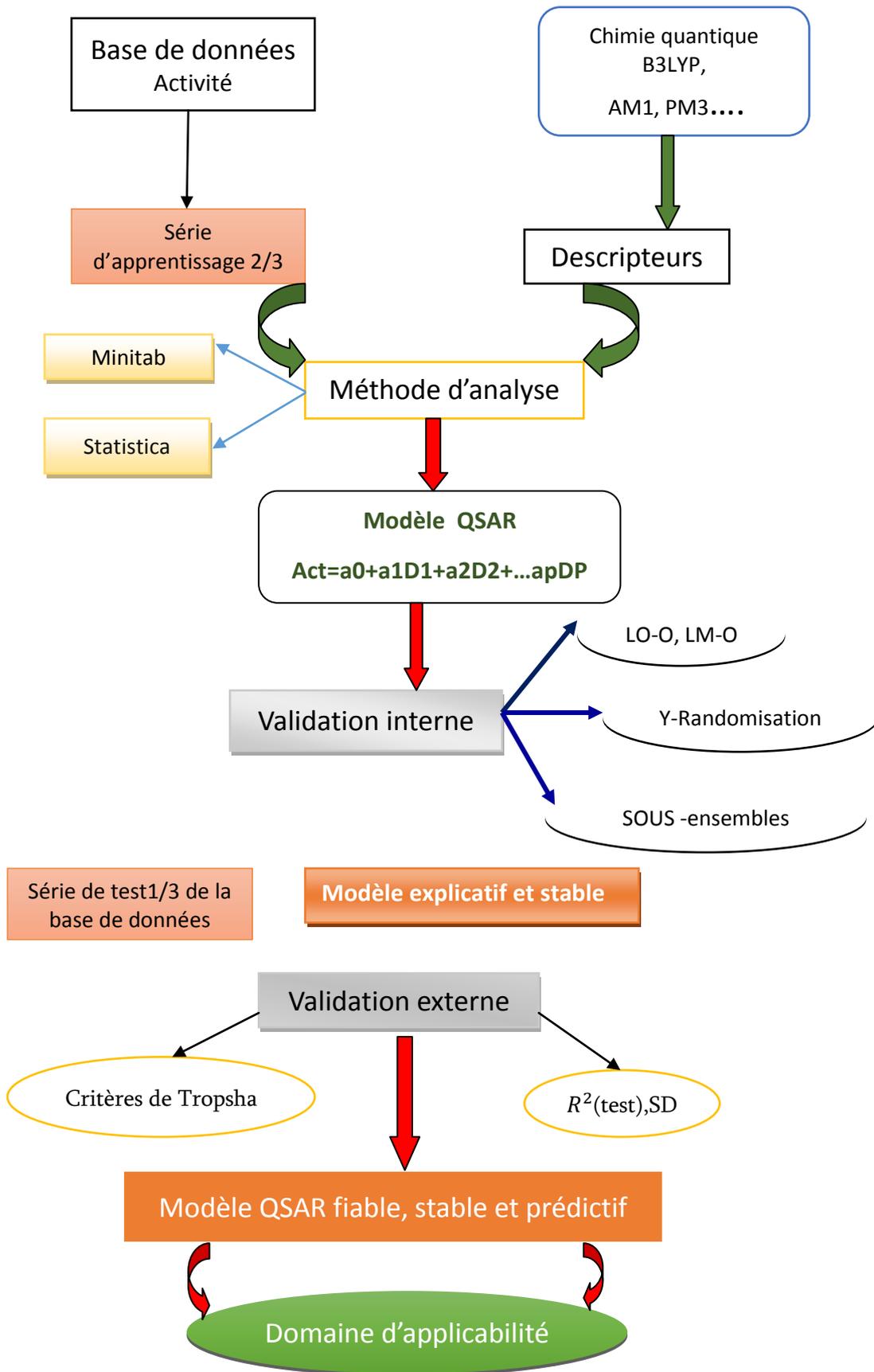


Figure.2.Méthodologie QSAR

Références bibliographiques

- [1] T.W. Schultz, M.T. Cronin, J.D. Walker, A.O. Aptula, Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective, *Journal of Molecular structure: THEOCHEM*, 622 (2003) 1-22.
- [2] R. Hmamouchi, M. Bouachrine, T. Lakhlifi, Tentative Pratique du Relation Quantitatives Structure-Activité/Propriété (QSAR/QSPR), *Revue Interdisciplinaire*, 1.
- [3] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, *Journal of Chemometrics: A Journal of the Chemometrics Society*, 15 (2001) 559-569.
- [4] K. BELLIFA, Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires.«Modélisation QSAR.
- [5] A. Fortuné, 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.
- [6] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. Gonzalez, A new search algorithm for QSPR/QSAR theories: Normal boiling points of some organic molecules, *Chemical Physics Letters*, 412 (2005) 376-380.
- [7] A.D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, *Physical review A*, 38 (1988) 3098.
- [8] K.L. Vine, J.M. Locke, M. Ranson, S.G. Pyne, J.B. Bremner, In vitro cytotoxicity evaluation of some substituted isatin derivatives, *Bioorganic & medicinal chemistry*, 15 (2007) 931-938.
- [9] B. Efron, Estimating the error rate of a prediction rule: improvement on cross-validation, *Journal of the American statistical association*, 78 (1983) 316-331.
- [10] M. Efronson, Multiple regression analysis, *Mathematical methods for digital computers*, (1960) 191-203.
- [11] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, *Journal of molecular biology*, 55 (1971) 379-IN374.
- [12] P. Rogers, Genetics and biochemistry of Clostridium relevant to development of fermentation processes, *Advances in applied microbiology*, Elsevier1986, pp. 1-60.
- [13] J.B. Ball, T.L. Nero, D. Iakovidis, L. Tung, G. Jackman, W.J. Louis, Computer-aided mapping of the. beta.-adrenoceptor I: Explanation for effect of para substitution on blocking activity at the. beta.-1-adrenoceptor, *Journal of medicinal chemistry*, 35 (1992) 4676-4682.
- [14] G. Crippen, A. Ghose, Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions, *J Chem Inf Comput Sci*, 27 (1987) 21-35.
- [15] S.K. Balasubramanian, J.A. Karrh, H. Patwardhan, Audience response to product placements: An integrative framework and future research agenda, *Journal of advertising*, 35 (2006) 115-141.
- [16] 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 & Combinatorial Science*, 22 (2003) 69-77.
- [17] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R.

Todeschini, D. Fourches, A. Varnek, Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection, *Journal of chemical information and modeling*, 48 (2008) 1733-1746.

- [18] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships, *ATLA*, 33 (2005) 155-173.

Chapitre II

Partie B-Méthodes Quantiques

Les méthodes dérivant de la mécanique quantique sont nombreuses et chacune présente avantages et inconvénients. Choisir la plus exacte n'est pas nécessairement une garantie de succès : il convient d'identifier celle qui se prêtera le mieux aux études envisagées.

Dans cette partie du chapitre II, nous décrivons très brièvement l'ensemble des principes de base des méthodes de la chimie quantique utilisées dans ce mémoire.

Les méthodes quantiques

1- Méthodes semi-empiriques:

Les méthodes semi-empiriques sont dérivées des méthodes ab-initio. Elles considèrent deux approximations dont la première consiste à prendre en considération que les électrons de valence. La deuxième approximation, nommée RDN (Recouvrement Différentiel Nul) ou NDO (Neglect of Differential Overlap) a été proposée en 1952 par Pann dans le but de réduire le nombre d'intégrales à calculer, en particulier les intégrales de Répulsion biélectroniques multicentrées [1-3]. On distingue plusieurs variantes :

-**CNDO** (Complete Neglect of Differential Overlap) [4] proposée par Pople, Segal et Santry en 1965. Elles négligent complètement les intégrales de répulsions entre les atomes Non chimiquement liés. Son défaut majeur réside dans son échec pour le traitement des Systèmes à couche ouverte.

-**INDO** (Intermediate Neglect of Differential Overlap) [4]: proposée par Pople, Beveridge et Dobosh en 1967. C'est une correction de la méthode CNDO. Elle permet de distinguer entre les états singulets et les états triplets d'un système en conservant les intégrales d'échange.

-**MINDO/3** (Modified Intermediate Neglect of Differential Overlap) [5, 6] proposée par Bingham, Dewar et Lo en 1975. La paramétrisation est effectuée en se référant aux résultats expérimentaux et non pas aux résultats ab-initio. Cependant, elle surestime le

calcul exact des chaleurs de formation, des potentiels d'ionisation et des distances interatomiques dans la plupart des molécules de dimension moyenne.

-**MNDO** (Modified Neglect of Diatomic Overlap): Proposée par Dewar et Thiel en 1977 [7]

Cette méthode est basée sur l'approximation NDDO (Neglect of diatomic Differential Overlap). Elle néglige le recouvrement différentiel entre orbitales atomiques sur des atomes différents. Elle a été élaborée de façon à reproduire les grandeurs physicochimiques caractérisant les états fondamentaux d'un grand nombre de composés organiques et minéraux.

Les enthalpies de formation, les chaleurs de réactions, les barrières de rotation, les potentiels d'ionisation, les moments dipolaires et les géométries moléculaires calculées par MNDO sont en bon accord avec les données expérimentales.

-**AM1** (Austin Model 1): Proposée par Dewar en 1985 [8] c'est une amélioration de la Méthode MNDO. Elle diffère de MNDO dans l'expression pour le calcul des répulsions cœur cœur (noyau-électrons internes).

-**PM3** (Parametric Method 3): Proposée par Stewart en 1989 [9]. PM3 est la troisième paramétrisation de MNDO [9]; AM1 étant considérée la seconde. Elles négligent généralement le calcul de certaines intégrales et les remplacent par des paramètres expérimentaux.

2- Théorie de la Fonctionnelle de Densité (DFT):

Actuellement, la DFT a un grand intérêt car elle peut traiter des systèmes de grande taille en incluant les effets de corrélation électronique.

La théorie de la fonctionnelle de la densité exprime l'énergie en fonction de la densité électronique. Les premiers à avoir exprimé l'énergie en fonction de la densité furent L.H. Thomas (1927) [10] E. Fermi (1927, 1928) [11] et P.A. Dirac (1930) [12] sur le modèle du gaz d'électrons non interagissant.

Le but des méthodes DFT est de déterminer des fonctionnelles qui permettent de donner une relation entre la densité électronique et l'énergie [13].

La DFT a véritablement débuté avec les théorèmes fondamentaux de Hohenberg et Kohn en 1964 [14], qui établissent une relation fonctionnelle entre l'énergie de l'état fondamental et sa densité électronique.

Les deux théorèmes montrent l'existence d'une fonctionnelle de la densité qui permet de calculer l'énergie de l'état fondamental d'un système.

2.1- Le Premier théorème de Hohenberg et Kohn (1964):

« L'énergie moléculaire, la fonction d'onde et toutes les autres propriétés électroniques de l'état fondamental sont déterminées à partir de la densité électronique de l'état fondamental $\rho_0(x, y, z)$ [15].

A partir de ce théorème, nous constatons que la densité électronique $\rho(r)$ peut être utilisée comme variable pour la résolution de l'équation de Schrodinger.

Le nombre d'électrons de système moléculaire peut être connu en intégrant sa densité électronique:

$$\int \rho(\mathbf{r}) d\mathbf{r} = n \quad (1)$$

Rappelons l'Hamiltonien d'un système moléculaire polyélectronique :

$$H = -\frac{1}{2} \sum_i \Delta_i + \sum_{i,j} \frac{1}{r_{ij}} + \sum_i v(r_i) \quad (2)$$

$$\text{avec : } v(\mathbf{r}_i) = -\sum_{\alpha} \frac{Z_{\alpha}}{r_{i\alpha}} \quad (3)$$

Où $V(\mathbf{r}_i)$ est le potentiel externe qui correspond à l'attraction de l'électron (i) avec tous les noyaux qui sont externes par rapport au système d'électrons.

L'énergie totale peut s'écrire comme une somme de trois fonctionnelles:

$$E[\rho] = T[\rho] + V_{ee}[\rho] + V_{ne}[\rho] \quad (4)$$

Cette fonctionnelle contient $T[\rho]$, l'énergie cinétique électronique et $V_{ee}[\rho]$, l'énergie potentielle due à l'interaction entre électrons et $V_{ne}[\rho]$ qui exprime l'attraction électron noyau

2.2- Deuxième théorème de Hohenberg et Kohn:

Pour une densité d'essai $\rho(r)$ telle que $\rho(r) \geq 0$ et $\int \rho(r) dr = n$, l'inégalité est vérifiée:

$$E_0[\rho] \leq E[\rho] \quad (5)$$

D'après la définition:

$$\delta E = \int \frac{\delta E}{\delta \rho} \delta \rho dr = 0 \quad (6)$$

Cette relation est donc vérifiée si:

$$\frac{\delta E}{\delta \rho} = 0 \quad (7)$$

2.3- Méthode de Kohn et Sham:

Les deux théorèmes de Hohenberg et Kohn cités précédemment ne donnent pas une procédure pour calculer l'énergie E_0 à partir de la densité électronique ρ ni comment déterminer cette dernière (ρ) sans déterminer la fonction d'onde. Pour résoudre ce problème, Kohn et Sham (1955) [16] ont élaboré une méthode pratique pour déterminer E_0 à partir de ρ_0 . Ils ont considéré un système fictif de référence S constitué par les n électrons non interagissant.

Kohn et Sham ont reformulé le problème de manière à ce que le système de référence ait la même densité que l'état fondamental. Dans ce système de référence, étant donné que les électrons ne réagissent pas entre eux l'Hamiltonien s'écrit:

$$\hat{H}_s = \sum_{i=1}^n \left[-1/2 \nabla_i^2 + v_s(r_i) \right] = \sum_{i=1}^n h_i^{KS} \quad (8)$$

Avec :

$$h_i^{KS} = -1/2 \nabla_i^2 + v_s(r_i) \quad (9)$$

2.4- Terme d'échange-corrélation:

L'expression de l'énergie d'échange-corrélation est la suivante:

$$\mathbf{E}_X[\rho] = \mathbf{E}_x[\rho] + \mathbf{E}_c[\rho] \quad (10)$$

Le premier terme correspond à l'énergie d'échange qui exprime les interactions électroniques de même spin. Le second est l'énergie de corrélation qui exprime les interactions électroniques de spins différents.

La fonctionnelle d'énergie d'échange-corrélation est définie comme suit:

$$\mathbf{E}_X[\rho] = \mathbf{T}[\rho] + \mathbf{V}_{ee}[\rho] \quad (11)$$

2.5- L'approximation locale LDA

Dans l'approximation LDA (Local Density Approximation), il a été supposé que la densité électronique varie uniformément.

L'énergie d'échange-corrélation s'écrit :

$$\mathbf{E}_{XC}^{LD}[\rho] = \int \rho(\mathbf{r}) \epsilon_{XC}(\rho) d\mathbf{r} = \mathbf{E}_X^{LDA}[\rho] + \mathbf{E}_C^{LDA}[\rho] \quad (12)$$

L'énergie d'échange-corrélation par électron s'écrit :

$$\epsilon_X(\rho) = \epsilon_x[\rho] + \epsilon_c[\rho] \quad (13)$$

2.6- L'approximation des gradients généralisés GGA

Cette approximation consiste donc à exprimer la fonctionnelle d'énergie d'échange-corrélation en fonction de la densité électronique et de son gradient. Cette technique est appelée « approximation de l'expansion du gradient » (GEA). Elle se révèle

efficace pour les systèmes dont la densité électronique ne varie que lentement. Pour les systèmes chimiques, il s'avère qu'elle donne des résultats moins bons que LDA.

La solution consiste alors à réécrire l'expression d'échange-corrélation

$$E_{xc}^{GGA} = \int \epsilon_{xc}^{GGA}(\rho, \nabla \rho) d\mathbf{r} \quad (14)$$

De nombreuses fonctionnelles ont été développées depuis, tant pour l'échange que pour la corrélation.

Parmi les plus connues et les plus utilisées on peut citer les fonctionnelles d'échange de Becke (B88) [17] et de Perdew et Wang (PW91) [18]. Pour la corrélation, on dispose, entre autres, des fonctionnelles de Perdew (P86) [19], de Lee, Yang et Parr (LYP) [20] et de Perdew et Wang (PW91) [19].

Toutes ces fonctionnelles permettent une amélioration de l'estimation des énergies de liaison dans les molécules, ainsi que des barrières d'énergie par rapport à l'approximation locale LDA. Il faut encore citer les fonctionnelles dites « hybrides », basées sur le formalisme de la connection adiabatique [21].

2.7-Fonctionnelle hybride B3LYP

C'est une fonctionnelle à 3 paramètres combinant les fonctionnelles d'échanges suivantes : fonctionnelle locale, de Becke(88) et celle de Hartree-Fock avec les fonctionnelles de corrélation de Vosko, wilk, Nusair (LDA) et de gradient GGA de Lee, Wang et Parr. L'énergie de cette fonctionnelle s'écrit :

$$E_{xc}^{B3LYP} = (1 - a_0 - a_x) E_x^{LDA} + a_x E_x^{B88} + a_c E_c^{LYP} (1 - a_c) + E_c^{VWN} + a_0 E_x^{HF} \quad (15)$$

Où $a_0 = 0.20$, $a_x = 0.72$ et $a_c = 0.81$. Ces paramètres sont des quantités semi-empiriques déterminées par un lissage des chaleurs de formation d'un ensemble standard de molécules. Cette fonctionnelle donne des résultats remarquablement précis pour un grand nombre de systèmes [22].

Références bibliographiques

- [1] J. Pople, D. Beveridge, P. Dobosh, Approximate self-consistent molecular-orbital theory. V. Intermediate neglect of differential overlap, *The Journal of Chemical Physics*, 47 (1967) 2026-2033.
- [2] J.A. Pople, G.A. Segal, Approximate self-consistent molecular orbital theory. III. CNDO results for AB₂ and AB₃ systems, *The Journal of Chemical Physics*, 44 (1966) 3289-3296.
- [3] L. Ordronneau, V. Aubert, R. Métivier, E. Ishow, J. Boixel, K. Nakatani, F. Ibersiene, D. Hammoutène, A. Boucekkine, H. Le Bozec, Tunable double photochromism of a family of bis-DTE bipyridine ligands and their dipolar Zn complexes, *Physical Chemistry Chemical Physics*, 14 (2012) 2599-2605.
- [4] J.A. Pople, D.P. Santry, G.A. Segal, Approximate self-consistent molecular orbital theory. I. Invariant procedures, *The Journal of Chemical Physics*, 43 (1965) S129-S135.
- [5] R.C. Bingham, M.J. Dewar, D.H. Lo, Ground states of molecules. XXV. MINDO/3. Improved version of the MINDO semiempirical SCF-MO method, *Journal of the American Chemical Society*, 97 (1975) 1285-1293.
- [6] N.C. Baird, M.J. Dewar, Ground States of σ -Bonded Molecules. IV. The MINDO Method and Its Application to Hydrocarbons, *The Journal of Chemical Physics*, 50 (1969) 1262-1274.
- [7] M.J. Dewar, W. Thiel, Ground states of molecules. 38. The MNDO method. Approximations and parameters, *Journal of the American Chemical Society*, 99 (1977) 4899-4907.
- [8] M.J. Dewar, E.F. Healy, A.J. Holder, Y.C. Yuan, Comments on a comparison of AM1 with the recently developed PM3 method, *Journal of computational chemistry*, 11 (1990) 541-542.
- [9] J.J. Stewart, Optimization of parameters for semiempirical methods II. Applications, *Journal of computational chemistry*, 10 (1989) 221-264.
- [10] L.H. Thomas, The calculation of atomic fields, *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, 1927, pp. 542-548.
- [11] E. Fermi, Un metodo statistico per la determinazione di alcune priorieta dell'atome, *Rend. Accad. Naz. Lincei*, 6 (1927) 32.
- [12] P.A. Dirac, Note on exchange phenomena in the Thomas atom, *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, 1930, pp. 376-385.
- [13] A.J. Cohen, P. Mori-Sánchez, W. Yang, Fractional spins and static correlation error in density functional theory, *AIP*, 2008.
- [14] P.W. Ayers, Atoms in molecules, an axiomatic approach. I. Maximum transferability, *The Journal of Chemical Physics*, 113 (2000) 10886-10898.
- [15] S. Segura, J.A.R. Martínez, M.G. Moreno-Armenta, First-principles study of the (0001)-MgB₂ surface finished in Mg and B, *Journal of Physics: Conference Series*, IOP Publishing, 2014, pp. 012010.
- [16] J. Sun, A. Ruzsinszky, J.P. Perdew, Strongly constrained and appropriately normed semilocal density functional, *Physical review letters*, 115 (2015) 036402.

- [17] A.D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, *Physical review A*, 38 (1988) 3098.
- [18] P. Ziesche, H. Eschrig, *Electronic Structure of Solids' 91: Proceedings of the 75. WE-Heraeus-Seminar and 21st Annual International Symposium on Electronic Structure of Solids Held in Gaussig (Germany), March 11-15, 1991*, Akademie Verlag 1991.
- [19] J.P. Perdew, Density-functional approximation for the correlation energy of the inhomogeneous electron gas, *Physical Review B*, 33 (1986) 8822.
- [20] B. Miehlich, A. Savin, H. Stoll, H. Preuss, Results obtained with the correlation energy density functionals of Becke and Lee, Yang and Parr, *Chemical Physics Letters*, 157 (1989) 200-206.
- [21] J.B. Staunton, The electronic structure of magnetic transition metallic materials, *Reports on Progress in Physics*, 57 (1994) 1289.
- [22] V. Barone, Inclusion of Hartree-Fock exchange in the density functional approach. Benchmark computations for diatomic molecules containing H, B, C, N, O, and F atoms, *Chemical physics letters*, 226 (1994) 392-398.

Chapitre II

Partie C-Méthodes Statistiques

II.3- Méthodes d'analyse statistique

L'élaboration d'un modèle QSAR nécessite l'utilisation d'une méthode statistique d'analyse des données.

Dans cette partie du deuxième chapitre nous donnerons les principes de base des trois méthodes statistiques utilisées dans ce mémoire.

II.3.1. Méthodes de régression linéaire (SLR, MLR)

L'objectif de la régression linéaire est de trouver une **relation linéaire** entre une variable Y à expliquer (dépendante) appelée aussi réponse et une ou plusieurs variables X_i dites explicatives (indépendantes, descripteurs).

II.3.1.1. Régression linéaire simple (RLS)

S'il s'agit d'un seul descripteur, la relation linéaire entre les deux variables, explicative (X) et à expliquer (Y) est de la forme suivante :

$$Y = a_0 + a_1 X$$

Puisque la relation qu'on étudie est une simplification de la réalité, que l'on a forcément oublié des facteurs explicatifs de Y, on introduit un terme aléatoire (perturbation, appelé encore résidu) noté ε à cette relation déterministe. Ainsi le modèle que l'on étudie est le suivant :

$$Y = \beta_0 + \beta_1 X_i + \varepsilon \quad (1)$$

Avec β_0 (appelé *Intercept*) et β_1 sont les paramètres inconnus, appelés coefficients de régression, que l'on cherche à déterminer. ε , le terme d'erreur.

On note que les paramètres β_0 et β_1 sont les mêmes pour toutes les observations, autrement dit que **l'influence de X sur Y est la même pour toutes les observations.**

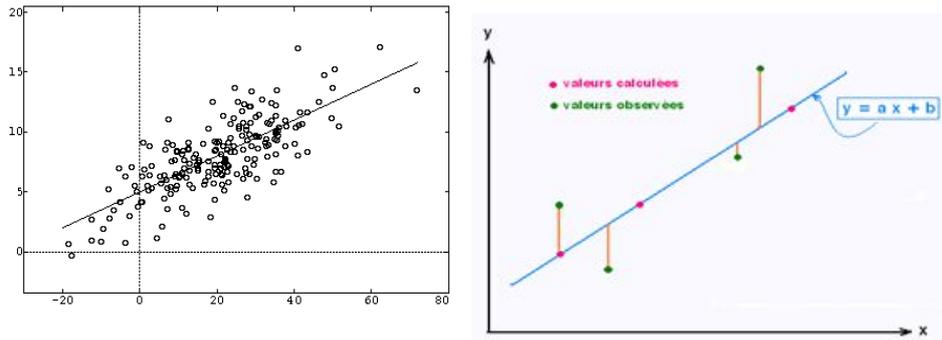


Figure.1. Représentation graphique de la régression

Pour trouver la droite qui passe « au plus près » de tous les points il faut se donner un critère d'ajustement. Le critère retenu pour déterminer la droite passant au plus près de tous les points sera tel que : « La somme des carrés des écarts (**SCE**) des points observés M_i à la droite solution soit minimum »

La droite solution sera appelée droite de **régression** de Y sur X. Le critère d'ajustement est le «critère des Moindres Carrés Ordinaires » (Ordinary Least Squares).

1- Méthode d'estimation des paramètres β_0 et β_1

Critère des Moindres Carrés Ordinaires » (MCO, Ordinary Least Squares),

La Somme des Carrés des Ecarts (SCE) est donnée par :

$$S = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \quad (2)$$

En mathématique cette fonction S est minimum lorsque les dérivées de S par rapport à β_0 et β_1 s'annulent.

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \quad (3)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \quad (4)$$

Chapitre II : Bases théoriques-Partie C-Méthodes statistiques

$\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs de β_0 et β_1

Ces dérivées s'annulent pour deux valeurs $\hat{\beta}_0$ et $\hat{\beta}_1$ solutions des deux équations à deux inconnues :

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} \quad (5)$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad (6)$$

$$\hat{\beta}_0 = \bar{Y} - \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \bar{X}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (7)$$

\hat{Y} est l'estimation de Y obtenue à partir de l'équation de régression.

L'écart entre cette prédiction \hat{Y} et Y est appelé résidu

$$Y_i - \hat{Y}_i = \varepsilon_i \quad (8)$$

la variation résiduelle du coefficient de régression « a » appelée σ^2 est donnée par :

$$S^2 = \frac{1}{n-2} \sum \varepsilon_i^2 \quad (9)$$

2- Analyse de la variance

$$Y = \hat{Y} + \varepsilon$$

Et la somme des carrés est donnée comme suit :

Chapitre II : Bases théoriques-Partie C-Méthodes statistiques

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2 \quad (10)$$

Ces diverses sommes des carrés sont définies comme suit :

SST (Sum of Squares Total) représente la somme des carrés des écarts total.

$$SST = \sum (Y_i - \bar{Y})^2 \quad (11)$$

Cette variation totale à expliquer, peut se décomposer en :

-Somme des carrés expliquée par le model (Sum of Squares Regression)

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 \quad (12)$$

-Somme des carrés des résidus (Sum of Squares Error qui représente la partie que le modèle n'explique pas)

$$SSE = \sum \varepsilon_i^2 = \sum (Y_i - \hat{Y}_i)^2 \quad (13)$$

On a alors l'équation d'analyse de la variance suivante :

$$SST = SSR + SSE \quad (14)$$

La représentation géométrique de tous ces écarts est donnée sur la **figure.2**

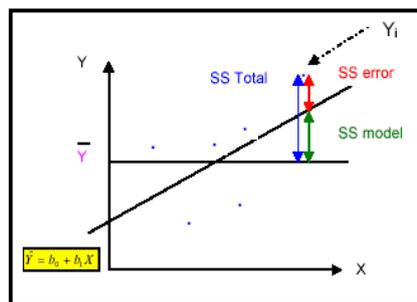


Figure.2. Décomposition des différents écarts

II.3.1.2. Régression linéaire Multiple (MLR)

Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante (à expliquer) Y (ici, l'activité) et une série de p variables indépendantes (explicatives) X_i (ici, les descripteurs).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (15)$$

Cette équation est linéaire par rapport aux paramètres (coefficients de régression) $\beta_0, \beta_1, \dots, \beta_p$.

La détermination de l'équation (15) se fait alors à partir d'une base de données de n échantillons (observations) pour laquelle à la fois les variables indépendantes et la variable dépendante sont connues.

Il s'agit donc de considérer un système d'équations qui peut être donné sous la forme matricielle suivante :

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & & & & \\ 1 & & & & \\ \dots & & & & \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

❖ Estimation des paramètres statistiques du modèle

Dans le cas d'un modèle à p variables, le critère des moindres carrés s'écrit :

$$S(\beta_0, \dots, \beta_p) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1i} - \dots - \beta_p X_{pi})^2$$

Les valeurs des β qui minimisent ce critère seront les solutions $\beta_0, \beta_1, \dots, \beta_p$ du système linéaire de $(p+1)$ équations à $(p+1)$ inconnues obtenues comme suit :

$$\hat{Y} = X \hat{\beta} = XB$$

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

(16)

Avec

X' : la matrice transposée de la matrice des variables explicatives X

$(X'X)^{-1}$: la matrice inverse de la matrice $(X'X)$

Y : vecteur des valeurs de la variable à expliquer

❖ **Écarts type des coefficients de régression (erreurs sur les coefficients de régression)**

On peut calculer les écart-type des estimateurs (les coefficients de régression) en utilisant la relation $\sigma^2 (X'X)^{-1}$ avec σ^2 la moyenne des carrés des erreurs ($SSE/n-1-p$)

La diagonale de cette matrice nous donnera les écarts type des estimateurs (erreurs sur les coefficients de régression).

❖ **Confiance à accorder aux résultats**

On cherche à tester la validité d'hypothèses concernant les (vrais) paramètres du modèle c'est-à-dire confirmer l'existence du modèle et la contribution de chaque descripteur dans l'explication de la réponse Y .

- **Test de la signification globale de la régression :**

✓ **Test de Fisher-Snedecor**

Ce test a surtout un intérêt dans le cadre de la régression multiple, c'est-à-dire avec p régresseurs X_1, X_2, \dots, X_p :

Ce test permet de connaître l'apport global de l'ensemble des variables X_1, \dots, X_p à la détermination de Y .

On teste l'hypothèse nulle : $H_0: \beta_1 = \dots = \beta_p = 0$ contre H_a : il existe au moins un β_j parmi β_1, \dots, β_p non égal à 0.

$$F = \frac{MS_{Model}}{MS_{Error}} \quad (17)$$

avec

$$MS_{Model} = \frac{SS_{Model}}{p} \quad (18)$$

et

$$MS_{Error} = \frac{SS_{Error}}{n-p-1} \quad (19)$$

Si H_0 est vraie et sous réserve des suppositions suivantes :

Ce rapport F est une valeur observée d'une variable à p et $n-p-1$ degrés de liberté (p : nombre de variables explicatives, n : nombre d'observations).

La statistique F suit une loi de Fisher-Snedecor

$$F = \frac{MS_{Model}}{MS_{Error}} \approx F(p, n-p-1)$$

Si la quantité F observée dépasse seuil, on rejette l'hypothèse H_0 (au niveau $\alpha=5\%$) et dans le cas contraire, on conserve H_0 .

$$\text{Si } F_{\text{observé}} \geq F_{1-\alpha}(p, n-p-1)$$

alors $H_0 : \beta_1 = \dots = \beta_p = 0$ doit être rejetée au niveau α . Cependant, la plupart des logiciels ne demandent pas de fixer a priori un niveau α pour effectuer le test : ils donnent le niveau à partir duquel notre décision aurait changé : c'est la notion de **p-value**.

Si p-value $\leq \alpha$ alors on rejette l'hypothèse nulle

✓ **Test de signification de chaque paramètre (chaque descripteur) t-Student**

Pour voir la contribution de chaque paramètre dans l'explication de la variable dépendante Y on utilise la statistique « t » définie auparavant en régression simple.

Chapitre II : Bases théoriques-Partie C-Méthodes statistiques

On calcule t-test pour chaque paramètre β_i en utilisant la formule ci-dessous

$$t\text{-observé} = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \quad (20)$$

Avec $s(\hat{\beta}_i)$ est l'erreur sur le paramètre β_i

En pratique si on choisit le risque $\alpha = 5\%$ et si n est grand ($n > 30$) pour approcher la loi de Student par la loi normale la valeur critique de t-test oscille le plus souvent autour de 2 (1,96).

Indicateurs de qualité d'une régression linéaire

❖ Coefficient de détermination R^2

Pour évaluer la précision avec laquelle la dépendance trouvée décrit la variance de la variable dépendante, le coefficient de détermination (R^2) est utilisé.

Un modèle, que l'on qualifie de bon, possède des estimations **proches** des **vraies valeurs** de Y . Les deux quantités SCE totale (SST) et SCE modèle (SSR) sont des sommes des carrés donc toujours positives ou nulles. Le rapport de ces deux quantités nous donne le coefficient de détermination R^2 :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (21)$$

Il est compris entre 0 et 1. Plus il est proche de 1 et plus la régression permet d'expliquer une grande partie de la variance totale de la variable à expliquer.

❖ Déviation standard (SD)

La fiabilité de la prédiction de la variable dépendante peut être évaluée également par la valeur de l'**erreur type d'estimation** ou déviation standard (SD).

$$s^2 = MSE = \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-p-1} = \frac{SSE}{n-p-1} \quad (22)$$

Root Mean Square Error= SD

$$\text{Root MSE} = \text{SD} = s \quad (23)$$

L'estimation de l'erreur-type appelé aussi déviation standard SD est une mesure de la dispersion des valeurs observées de la variable dépendante sur la droite de régression (de surface).

Les petites valeurs de SD signifient un bon ajustement statistique du modèle et une forte fiabilité de la prédiction.

▪ Sélection de variables et choix du modèle dans le cas d'une régression multiple

Parmi l'ensemble des p variables disponibles, toutes n'ont pas nécessairement un intérêt dans la modélisation de Y , et il peut alors être néfaste de les utiliser. Nous sommes alors en présence de différents modèles possibles parmi lesquels il faut faire un choix.

❖ Critères de comparaison de modèle

✓ Coefficient de détermination R^2

Il n'est pas recommandé d'utiliser R^2 pour comparer des modèles avec un nombre différent de descripteurs, le coefficient R^2 nous dira toujours de choisir le modèle avec le plus grand nombre de descripteurs car son R^2 sera plus important (on projette sur un espace plus grand), même si les variables sont sans effets sur la réponse Y .

Par conséquent, un autre paramètre statistique peut être utilisé en régression multiple, appelé R^2 ajusté (R^2 ajusté).

✓ Coefficient de détermination ajusté $R^2_{\text{ajusté}}$:

Ce coefficient est utilisé en régression multiple par ce qu'il tient compte du nombre de paramètres du modèle.

$$R_{\text{adj}}^2 = 1 - \frac{(n - \text{intercept})(1 - R^2)}{n - p} \quad (24)$$

$$\text{Ou bien } R_{\text{adj}}^2 = \frac{R^2(n-1) - p}{n - p - 1}$$

Cette formule indique notamment que $R^2\text{-adj}$ est toujours inférieur à R^2 , et ceci d'autant plus que le modèle contient un grand nombre de prédicteurs (descripteurs).

II.3.2.Méthode des moindres carrés partiels PLS

1-Définition

La régression PLS est une technique récente qui généralise et combine les caractéristiques de l'analyse sur composantes principales et de la régression multiple.

Elle permet de relier un ensemble de variables dépendantes Y à un ensemble de variables indépendantes X_i lorsque le nombre de variables indépendantes est élevé et/ou elles **sont fortement corrélées**.

Il s'agit d'étudier les relations entre les variables dépendantes Y et les variables indépendantes X_j en **prenant en compte les relations internes entre les variables** de chaque groupe.

2- Principe de la méthode PLS

Le principe de cette méthode consiste aux recherches successives de composantes h ($h=1, \dots, A$) combinaisons linéaires des variables de départ : $t_h = Xw_h$, liées avec Y , et ensuite les utiliser pour modéliser Y linéairement.

La première étape de la PLS consiste à :

- **Maximiser la corrélation** entre la première composante t_1 et la variable à expliquer Y (régression ordinaire).

Chapitre II : Bases théoriques-Partie C-Méthodes statistiques

- **Maximiser la variance** de la première composante t_1 afin que la composante représente au mieux toutes les variables explicatives.

3- Construction de la première composante $h=1$

La première composante a pour formule

$$t_1 = XW = W_{11}X_1 + W_{12}X_2 + W_{13}X_3 + \dots + W_{1p}X_p \quad (25)$$

avec w c'est le poids de chaque variable explicative X dans la composante t_1 , il peut être obtenu comme suit :

$$W_{1j} = \frac{\text{Cov}(X_j, Y)}{\sqrt{\sum_{k=1}^p \text{Cov}^2(X_k, Y)}} \quad (26)$$

Et l'équation de la régression simple de Y sur t_1 est donnée comme suit

$$Y = c_1 t_1 + Y^{[1]} \quad (27)$$

avec c_1 c'est le coefficient de régression de Y sur la première composante t_1 et $Y^{[1]}$ c'est le terme d'erreur de Y qui n'a pas été expliqué par la première composante t_1 .

En faisant p régressions simples des variables $X_1 \dots X_p$ sur t_1 , nous obtenons les termes d'erreurs des X , ces derniers vont servir comme données pour construire une deuxième composante t_2 .

4. Construction des composantes suivantes

$h = 2$ (deuxième composante)

De la même manière que la première la deuxième composante sera construite en utilisant les résidus des X et de Y non expliqués par la première composante.

$$w_{2j} = \frac{\text{Cov}(X_j^{[1]}, Y^{[1]})}{\sqrt{\sum_{k=1}^p \text{Cov}^2(X_k^{[1]}, Y^{[1]})}} \quad (28)$$

$$Y = c_1 t_1 + c_2 t_2 + Y^{[2]}$$

(29)

Avec $Y^{[2]}$ c'est le résidu de Y non expliqué par la deuxième composante, c_1 et c_2 sont les coefficients de régression de Y sur t_1 et sur t_2 respectivement.

On peut recommencer la procédure jusqu'à obtenir $h = A$ composantes = ?

4- Choix de la dimension du modèle (A) (nombre de composantes)

Le problème qui se pose après avoir construit un modèle est de connaître son aptitude à prédire les réponses de nouvelles variables d'entrée $[X]$. Pour le cas du PLS, ceci consiste surtout à évaluer le nombre optimal de composantes à inclure dans le modèle.

En effet, si **trop peu** de composantes sont incluses, il se peut que des phénomènes importants se produisant au sein des données étudiées ne soient pas modélisés. Si on prend par contre **trop de composantes**, on a le phénomène de surévaluation (**overfitting**) : on modélise les phénomènes sans importance comme les bruits et ceci peut fausser les prédictions.

5- Critère de choix de nombre de composantes:

➤ Critère de « validation croisée »

La validation croisée consiste à séparer les données $\{[X],[Y]\}$ en deux groupes :

- Un groupe d'étalonnage qui va servir à élaborer le modèle
- Un groupe-test sur lequel le modèle sera appliqué pour évaluer sa capacité de prédiction.

- La validation croisée « Leave One Out » (LOO-CV) :

Le LOO-CV est une variante de la validation croisée dans laquelle le groupe d'étalonnage

Chapitre II : Bases théoriques-Partie C-Méthodes statistiques

est composé des $n-1$ échantillons et le groupe test est donc formé par l'échantillon restant.

Le modèle sera tout d'abord ajusté sur les $n-1$ échantillons. Les prévisions des variables $[Y]$ seront calculées sur le groupe test. La somme des carrés des erreurs (différence entre les $[Y]$ calculés ou prédits et les valeurs connues des $[Y]$) sera ensuite évaluée. Cette procédure est répétée n fois jusqu'à ce que chacune des échantillons serve une fois de test.

La somme de tous les carrés des erreurs obtenues forme ce qu'on appelle le **PRESS**

$$PRESS_h = \sum_{i=1}^n (y - \widehat{y}_i^h)^2 \quad (30)$$

où \widehat{Y}_{ih}^k est la prévision de Y^k en utilisant le i -ème échantillon comme groupe-test et avec un modèle PLS à h composantes.

Le **PRESS** est très utile pour la validation du modèle car c'est à partir de cette grandeur qu'on peut considérer qu'une composante PLS est significative ou non. En effet, cette quantité décroît en fonction du nombre de composantes pour atteindre une valeur minimale et se stabiliser par la suite. C'est ce minimum qui détermine le nombre de composantes à retenir pour le modèle.

Références bibliographiques

- [1] P.-A. Cornillon, É. Matzner-Løber, Régression: théorie et applications, Springer2007.
- [2] K. BELLIFA, Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires."Modélisation QSAR".
- [3] W.C. Parr, M.A. Smith, Developing case-based business statistics courses, The American Statistician, 52 (1998) 330-337.
- [4] R.D. Cook, S. Weisberg, An introduction to regression graphics, John Wiley & Sons2009.
- [5] J. JACQUES, Modélisation statistique, Poly Tech Lille.
- [6] P. Besse, Apprentissage statistique & data mining, Toulouse: INSA, (2009).
- [7] G. McLachlan, Discriminant analysis and statistical pattern recognition, John Wiley & Sons2004.
- [8] J.-P. Nakache, J. Confais, Statistique explicative appliquée: analyse discriminante, modèle logistique, segmentation par arbre, Editions Technip2003.
- [9] G. Saporta, Probabilités, analyse des données et statistique, Editions Technip2006.

Chapitre III
Résultats et Discussion

Introduction

Le phénol est un produit chimique aromatique couramment rencontré dans les effluents domestiques et industriels ce qui constitue une préoccupation mondiale en toxicologie. Lorsqu'il atteint les milieux aquatiques, des dommages importants sont observés chez toutes les espèces aquatiques, poissons, daphnies et algues.

L'impact du danger potentiel de ces produits, un défi qui conforte les organismes de réglementation internationaux [1], peut être mesuré par voie expérimentale, mais cette approche est très coûteuse en argent et en temps [2]. Pour cette raison l'utilisation des méthodes de modélisation QSAR pour l'étude de la toxicité est devenue d'une grande importance.

L'objectif de ce travail est d'étudier et de prédire la toxicité d'une série constituée de 30 composés phénoliques vis-à-vis deux algues marines *Chlorellavulgaris* et *Dunaliella tertiolecta*, en utilisant des descripteurs ciblés et en prenant en considération le mode d'action de chaque sous série de la base de données. Cette étude QSAR a été réalisée en utilisant un nombre réduit de descripteurs pertinent et en respectant toute la méthodologie QSAR et tous les critères d'OECD.

Chapitre III : Résultats et discussion

Tableau.1. Base de données [3]

N°	Nom du composé	Symbol	Mode d'action (MOA)	pT-48h	pT-72h	pT-96h	pT-96-1
1	Phénol	P	Polar narcotic	-0,24	-0,28	-0,30	-0,60
2	2-chlorophenol	2-CP	Polar narcotic	0,44	0,34	0,31	0,17
3	3-chlorophenol*	3-CP*	Polar narcotic	0,91	0,62	0,49	0,36
4	4-chlorophenol	4-CP	Polar narcotic	0,81	0,67	0,63	0,46
5	2,3-dichlorophenol	2,3-DCP	Polar narcotic	0,98	0,92	0,88	1,09
6	2,4-dichlorophenol	2,4-DCP	Polar narcotic	1,23	1,20	1,14	1,24
7	2,5- dichlorophenol	2,5-DCP	Polar narcotic	0,98	0,97	0,95	1,12
8	2,6-dichlorophenol	2,6-DCP	Polar narcotic	0,43	0,30	0,25	0,88
9	3,4-dichlorophenol*	3,4-DCP*	Polar narcotic	1,43	1,35	1,35	1,47
10	3,5-dichlorophenol	3,5-DCP	Polar narcotic	1,71	1,46	1,39	1,67
11	2,3,4-trichlorophenol	2,3,4-TCP	Polar narcotic	1,88	1,68	1,65	1,64
12	2,3,5-trichlorophenol	2,3,5-TCP	Polar narcotic	1,78	1,78	1,69	1,86
13	2,3,6-trichlorophenol	2,3,6-TCP	Polar narcotic	0,80	0,60	0,58	1,51
14	2,4,5-trichlorophenol*	2,4,5-TCP*	Polar narcotic	1,99	1,90	1,73	1,67
15	2,4,6- trichlorophenol	2,4,6-TCP	Polar narcotic	0,75	0,68	0,67	1,53
16	3,4,5-trichlorophenol	3,4,5-TCP	Polar narcotic	2,22	2,15	2,15	2,18
17	2,3,4,5-tetrachlorophenol	2,3,4,5-TeCP	Uncoupler	2,59	2,46	2,41	2,34
18	2,3,4,6-tetrachlorophenol	2,3,4,6-TeCP	Uncoupler	2,29	2,04	1,85	1,45
19	2,3,5,6-tetrachlorophenol	2,3,5,6-TeCP	Uncoupler	2,32	2,11	1,90	1,43
20	Pentachlorophenol	PCP	Uncoupler	3,03	2,96	2,91	1,45
21	1,2,3-trihydroxybenzene	1,2,3-THB	Pro-electrophile	0,59	0,57	0,53	1,28
22	Hydroquinone	HQ	Pro-electrophile	2,56	2,52	2,51	0,02
23	Chlorohydroquinone*	CLHQ*	Pro-electrophile	2,59	2,59	2,58	1,13
24	tetrachlorohydroquinone	TetraCLHQ	Pro-redox cyler	1,68	1,66	1,64	1,46
25	Catéchol*	Cat*	Pro-electrophile	0,73	0,67	0,60	0,27
26	4-chlorocatechols	4-CLCat	Pro-electrophile	0,86	0,74	0,70	1,13
27	3,5dichloro catéchols	3,5-DiCLCat	Pro-electrophile	1,19	1,03	0,94	1,95
28	Résorcinol	Res	Polar narcotic	-0,42	-0,42	-0,43	-0,49
29	4-chloro résorcinol*	4-CLRes*	Polar narcotic	0,16	0,04	0,00	0,27
30	4,6-dichloro résorcinol	4,6-DiCLRes	Polar narcotic	0,51	0,49	0,49	1,02

*(Série de test), (pT-48,pT-72,pT-96 :Algue 1,pT-96-1 :Algue 2)

Calcul des descripteurs

Les géométries d'équilibre des composés phénoliques ont été obtenues en utilisant le programme GAUSSIAN G09 [4] avec différentes méthodes en l'occurrence, la méthode semi empirique AM1 et la DFT B3LYP/6-31G(d, p), ce qui nous a permis de générer plusieurs descripteurs.

La valeur de l'hydrophobie (log P) pour chaque produit chimique a été calculée à l'aide du logiciel ACD / Labs [5].

Chapitre III : Résultats et discussion

Tableau.2. Signification des descripteurs

Descripteur	Signification
E_{HOMO}	L'énergie de l'orbitale la plus haute occupée
E_{LUMO}	L'énergie de l'orbitale la plus basse vide
LA DURETE η	$\eta = E_{LUMO} - E_{HOMO}$
LE POTENTIEL	$\mu = (E_{HOMO} + E_{LUMO})/2$
MD	Moment dipolaire
ΔH	L'enthalpie de formation
ΔG	L'énergie libre
ELECTROPHILIE	$\omega = \mu^2 / (2 * \eta)$
Log P	Le coefficient de partage

Les tableaux ci-dessous regroupent les différents descripteurs calculés pour toutes les molécules de notre base de données.

Chapitre III : Résultats et discussion

Tableau. 3. Descripteurs calculés par la méthode DFT-B3LYP/6-31G (d,p)

Numéro	E _{Homo} (ev)	E _{Lumo} (ev)	Dureté η (ev)	Potentiel (ev)	Polarisabilité	MD	charge C/OH	ΔH (Kcal/mol)	ΔG (Kcal/mol)	Electrophilie(ev)	log p
1	-5,965232	0,01496	5,980192	-2,975136	13,12550887	1,3355	-0,089593	-307,367213	-307,402648	0,74006271	1,48
2	-6,253824	-0,370464	5,88336	-3,312144	11,36074257	0,9121	-0,074866	-766,97173	-767,010346	0,93231571	2,04
4	-6,091712	-0,419696	5,672016	-3,255704	14,71350887	2,4886	0,327445	-766,969457	-767,008342	0,93437752	2,43
5	-6,488288	-0,676192	5,812096	-3,58224	11,6719908	1,3292	-0,095338	-1226,569265	-1226,611108	1,10394283	2,83
6	-6,350384	-0,766768	5,583616	-3,558576	12,61919163	1,0861	0,352047	-1226,572391	-1226,614437	1,13398407	2,99
7	-6,512224	-0,73576	5,776464	-3,623992	13,12254403	1,4474	-0,073626	-1226,57318	-1226,615205	1,13679563	2,88
8	-6,483664	-0,696592	5,787072	-3,590128	11,83511387	2,0519	0,379474	-1226,569988	-1226,611906	1,11360452	2,61
10	-6,631088	-0,736032	5,895056	-3,68356	16,4685542	2,3842	-0,103225	-1226,571168	-1226,61342	1,15084694	3,33
11	-6,504064	-0,986544	5,51752	-3,745304	12,2800916	1,8777	0,362692	-1686,164964	-1686,210061	1,27116005	3,66
12	-6,754848	-1,005584	5,749264	-3,880216	13,2059051	1,0312	-0,095249	-1686,169408	-1686,214662	1,30939162	3,69
13	-6,6232	-0,978656	5,644544	-3,800928	12,74587737	0,6225	-0,097793	-1686,166644	-1686,211809	1,27973612	3,33
15	-6,564992	-1,059168	5,505824	-3,81208	13,27890413	1,413	0,381382	-1686,169295	-1686,214643	1,31968929	3,58
16	-6,565536	-0,988448	5,577088	-3,776992	16,7708714	3,3527	-0,095622	-1686,162856	-1686,208232	1,27895315	4,02
17	-6,704528	-1,241136	5,463392	-3,972832	13,19992427	2,0772	-0,088226	-2145,760071	-2145,808399	1,44446839	4,39
18	-6,672976	-1,260992	5,411984	-3,966984	12,98368893	1,0059	0,39099	-2145,761221	-2145,809638	1,45389954	4,17
19	-6,847872	-1,223456	5,624416	-4,035664	12,940663	1,3872	0,395562	-2145,762068	-2145,81046	1,44784667	4,06
20	-6,811968	-1,442144	5,369824	-4,127056	12,85442837	1,9031	0,400318	-2605,351458	-2605,402884	1,58595433	4,78
21	-5,593952	0,501296	6,095248	-2,546328	71,61545793	2,9347	-0,136819	-457,798105	-457,839164	0,53187223	0,29
22	-5,420144	-0,074528	5,345616	-2,747336	29,1215033	2,641	-0,105062	-382,579151	-382,617446	0,70598553	0,64
24	-6,325088	-1,242224	5,082864	-3,783656	23,66189957	2,6484	-0,136918	-2220,971808	-2221,022671	1,40826636	3,79
26	-5,829232	-0,20944	5,619792	-3,019336	44,58045743	2,4448	-0,111711	-842,184697	-842,226405	0,81109673	2,15
27	-6,188544	-0,526864	5,66168	-3,357704	28,68204697	2,4209	-0,13934	-1301,787554	-1301,832239	0,99565643	3,04
28	-5,778912	0,182512	5,961424	-2,7982	24,2666844	1,3357	0,326116	-382,582617	-382,620658	0,65671585	0,76
30	-6,196976	-0,604656	5,59232	-3,400816	21,1327233	0,5855	0,35397	-1301,788472	-1301,832934	1,03405648	2,58
3	-6,29	-0,384064	5,905936	-3,337032	15,180466	1,1301	-0,100617	-766,970086	-767,008946	0,94276187	2,4
9	-6,330528	-0,71536	5,615168	-3,522944	15,71193403	2,6215	-0,095752	-1226,566979	-1226,609128	1,10514364	3,22
14	-6,550848	-1,037136	5,513712	-3,793992	13,35588577	2,0434	0,357204	-1686,1688	-1686,214099	1,30532528	3,71
23	-5,728592	-0,444176	5,284416	-3,086384	27,0387252	1,4863	-0,120425	-842,183326	-842,224768	0,90130737	1,52
25	-5,62632	0,204	5,83032	-2,71116	42,8334191	2,4859	-0,128538	-382,582594	-382,620841	0,63035893	0,88
29	-5,972304	-0,219504	5,7528	-3,095904	22,3030967	1,03	0,350546	-842,186483	-842,227721	0,8330397	1,67

Chapitre III : Résultats et discussion

Tableau. 4. Descripteurs calculés par la méthode semi-empirique AM1

Numéro	E _{HOMO} (ev)	E _{LUMO} (ev)	Dureté η (ev)	Potentiel(ev)	polarisabilité	MD	charge C/OH	ΔH (kcal/mol)	ΔG (kcal/mol)	Electrophilie(ev)
1	-9,110912	0,397664	9,508576	-4,356624	16,44546557	1,2326	-0,097003	48,8315732	26,4840596	0,99805548
2	-9,255616	0,029648	9,285264	-4,612984	17,32236773	0,937	-0,084983	37,2289133	12,8771327	1,14588133
4	-9,122608	0,094928	9,217536	-4,51384	17,5185571	1,4763	-0,090067	36,6716844	12,268448	1,10521681
5	-9,425888	-0,261936	9,163952	-4,843912	18,17954163	0,8353	-0,021178	26,8166399	0,49008531	1,2802055
6	-9,267584	-0,244528	9,023056	-4,756056	18,11915427	0,4017	-0,080141	25,663904	-0,73293168	1,2534594
7	-9,404128	-0,324768	9,07936	-4,864448	18,75946667	1,144	-0,07665	25,7078297	-0,69528108	1,30311246
8	-9,185168	0,065824	9,250992	-4,559672	20,02193787	2,3983	-0,093975	39,3135015	14,90148	1,12369618
10	-9,533056	-0,284512	9,248544	-4,908784	18,7666823	1,4075	-0,021313	25,4957313	-0,93749994	1,30270021
11	-9,376112	-0,472192	8,90392	-4,924152	18,36693457	1,0161	-0,020695	16,7927951	-11,4984932	1,36160662
12	-9,58936	-0,577728	9,011632	-5,083544	19,19166163	0,4243	-0,013887	15,8364699	-12,5320022	1,43383682
13	-9,452	-0,55216	8,89984	-5,00208	21,30589243	0,8099	-0,021865	18,2781113	-10,1010285	1,40568844
15	-9,386448	-0,500752	8,885696	-4,9436	21,22140153	1,0726	-0,087792	17,3920672	-11,0598638	1,3751979
16	-9,418	-0,466752	8,951248	-4,942376	18,76448933	2,0559	-0,020578	16,5072781	-11,8317011	1,36445111
17	-9,504224	-0,750992	8,753232	-5,127608	19,0890392	1,2208	-0,014473	7,31990415	-22,933608	1,50186604
18	-9,449552	-0,751264	8,698288	-5,100408	21,3389493	0,4747	-0,021895	8,73619422	-21,6039143	1,49536103
19	-9,625264	-0,816	8,809264	-5,220632	21,9580457	1,0655	-0,023136	8,7663147	-21,5763038	1,54695094
20	-9,569504	-0,976752	8,592752	-5,273128	21,7642008	1,2416	-0,023994	0,753012	-31,4583313	1,61798449
21	-8,95968	0,269552	9,229232	-4,345064	112,3249207	2,4309	-0,209447	-32,5533363	-58,704188	1,02281431
22	-8,721952	0,220048	8,942	-4,250952	38,84346393	0,0002	-0,178102	8,89495425	-15,313754	1,01043351
24	-9,238752	-0,933232	8,30552	-5,085992	52,97100637	0,0036	-0,115697	-26,7494963	-58,9533095	1,55723631
26	-8,967568	0,001088	8,968656	-4,48324	62,7865014	1,4348	-0,166571	-3,64395057	-29,9755252	1,12053807
27	-9,21536	-0,353872	8,861488	-4,784616	66,7903672	1,567	-0,099156	-14,246987	-42,5878487	1,29168771
28	-9,045904	0,320688	9,366592	-4,362608	27,37522297	1,2078	0,116852	7,89282078	-16,0893564	1,01596977
30	-9,219712	-0,262208	8,957504	-4,74096	25,54249097	1,022	0,130511	-14,1503505	-42,1485917	1,25462973
3	-9,296416	0,039168	9,335584	-4,628624	15,5587548	0,2357	-0,022719	36,7607908	12,3694771	1,14744617

Chapitre III : Résultats et discussion

9	-9,256432	-0,199376	9,057056	-4,727904	15,08334753	1,3564	-0,021226	26,2198778	-0,15248493	1,23401447
14	-9,383728	-0,55488	8,828848	-4,969304	16,1481932	1,3034	-0,073366	15,6833574	-12,6976649	1,39848269
23	-8,902832	-0,110432	8,7924	-4,506632	34,97248313	1,2718	-0,174184	-2,36634021	-28,55296	1,15495951
25	-8,881344	0,297024	9,178368	-4,29216	64,1994255	2,1281	-0,169375	8,27936694	-16,0366456	1,00359004
29	-9,094592	-0,007344	9,087248	-4,550968	24,65572383	0,3598	0,131878	-3,60253491	-29,5789389	1,13958097

Chapitre III : Résultats et discussion

Analyse statistique

Des modèles QSAR ont été élaborés en utilisant les méthodes de régression SLR et PLS (partiel least square) sur le programme MINITAB VERSION 17 [6]

Les quatre toxicités sont prises comme des variables dépendantes alors que les descripteurs calculés sont des variables indépendantes.

Résultats et discussion

La base de données contient 30 composés phénoliques de différents modes d'action, est divisée en deux sous-séries, une série d'apprentissage constituée de 24 composés et les 6 molécules restantes forment la série de test.

En effectuant des régressions simples et multiples, il a été remarqué que deux composés se présentent comme des outliers avec un résiduel qui dépasse 2,5, il s'agit de THB (N°21), et QH (N°22). L'élimination de ces deux composés a permis d'améliorer d'une manière très significative la qualité des modèles.

1- Méthode de régression linéaire simple SLR

Des modèles avec un seul descripteur sont élaborés pour les deux méthodes B3LYP et AM1, les modèles retenus ($R^2 > 30\%$) avec leurs paramètres statistiques sont présentés dans les tableaux 5 et 6.

Tableau.5 : Modèles QSAR avec un seul descripteur (B3LYP)

N°	Modèle	R^2	R^2_{adj}	SD
1	pT48 = -13,13 - 61,0 E _{Homo}	58,04%	55,94%	0,597822
	pT72 = -12,42 - 57,4 E _{Homo}	54,64%	52,38%	0,602399
	pT96 = -12,21 - 56,5 E _{Homo}	55,47%	53,24%	0,583219
	pT96-1 = -9,75 - 46,5 E _{Homo}	47,82%	45,21%	0,559218
2	PT48 = -0,127 - 49,07 E _{Lumo}	73,43%	72,10%	0,475729
	pT72 = -0,193 - 46,14 E _{Lumo}	69,11%	67,57%	0,497124
	pT96 = -0,206 - 46,37 E _{Lumo}	73,13%	71,79%	0,452989

Chapitre III : Résultats et discussion

	$pT96-1 = 0,144 - 37,37 E_{Lumo}$	60,51%	58,54%	0,486478
3	$pT48 = 16,50 - 73,4 \eta$	40,02%	37,02%	0,714723
	$pT72 = 15,44 - 69,0 \eta$	37,65%	34,53%	0,706299
	$pT96 = 16,02 - 71,9 \eta$	42,76%	39,90%	0,661209
	$pT96-1 = 12,81 - 55,9 \eta$	32,99%	29,64%	0,633724
4	$pT48 = -6,37 - 57,76 \mu$	70,69%	69,23%	0,499616
	$pT72 = -6,06 - 54,32 \mu$	66,55%	64,87%	0,517364
	$pT96 = -6,05 - 54,15 \mu$	69,29%	67,76%	0,484276
	$PT96-1 = -4,61 - 43,99 \mu$	58,26%	56,17%	0,500195
5	$PT48 = -2,380 + 85,4 \omega$	74,36%	73,08%	0,467319
	$pT72 = -2,326 + 80,6 \omega$	70,59%	69,12%	0,485085
	$pT96 = -2,332 + 80,6 \omega$	73,91%	72,60%	0,446418
	$PT96-1 = -1,450 + 62,2 \omega$	56,02%	53,82%	0,513405
6	$PT48 = -1,411 + 0,8582 \log p$	86,86%	86,20%	0,334541
	$pT72 = -1,390 + 0,8035 \log p$	81,05%	80,10%	0,389426
	$pT96 = -1,412 + 0,8086 \log p$	86,00%	85,30%	0,327052
	$PT96-1 = -0,845 + 0,6573 \log p$	72,41%	71,03%	0,406649

Tableau.6 : Modèles QSAR avec un seul descripteur (AM1)

<i>N°</i>	<i>Modèle</i>	<i>R²</i>	<i>R² adj</i>	<i>SD</i>
1	$PT48 = -32,77 - 99,2 E_{Homo}$	55,18%	52,94%	0,617844
	$pT72 = -31,19 - 94,1 E_{Homo}$	52,91%	50,55%	0,613813
	$pT96 = -30,53 - 92,2 E_{Homo}$	53,16%	50,82%	0,598133
	$PT96-1 = -23,26 - 71,3 E_{Homo}$	40,51%	37,54%	0,597104

Chapitre III : Résultats et discussion

2	PT48 = 0,565 - 54,92 E _{Lumo}	74,06%	72,77%	0,470012
	pT72 = 0,446 - 52,54 E _{Lumo}	72,15%	70,76%	0,472049
	pT96 = 0,445 - 52,15 E _{Lumo}	74,45%	73,17%	0,441774
	PT96-1 = 0,672 - 41,81 E _{Lumo}	61,00%	59,05%	0,483458
3	PT48 = 22,27 - 63,6 η	50,35%	47,87%	0,650271
	pT72 = 21,32 - 61,1 η	49,60%	47,08%	0,635032
	pT96 = 21,35 - 61,2 η	52,12%	49,73%	0,604726
	PT96-1 = 17,82 - 50,3 η	44,79%	42,03%	0,575232
4	PT48 = -12,90 - 79,6 μ	75,81%	74,60%	0,453906
	pT72 = -12,40 - 76,0 μ	73,51%	72,19%	0,460364
	pT96 = -12,26 - 75,11 μ	75,27%	74,03%	0,434602
	PT96-1 = -9,40 - 59,6 μ	60,42%	58,44%	0,487052
5	PT48 = -4,577 + 121,1 ω	75,51%	74,29%	0,456703
	pT72 = -4,488 + 116,2 ω	73,96%	72,66%	0,456453
	pT96 = -4,429 + 114,8 ω	75,68%	74,46%	0,430986
	PT96-1 = -3,076 + 88,8 ω	57,60%	55,48%	0,504116
6	PT48 = -1,411 + 0,8582 log p	86,86%	86,20%	0,334541
	pT72 = -1,390 + 0,8035 log p	81,05%	80,10%	0,389426
	pT96 = -1,412 + 0,8086 log p	86,00%	85,30%	0,327052
	PT96-1 = -0,845 + 0,6573 log p	72,41%	71,03%	71,03%

D'après les deux tableaux il s'avère que deux méthodes donnent des modèles avec des performances très proches.

Chapitre III : Résultats et discussion

Les quatre meilleurs modèles obtenus avec un seul descripteur pour les deux méthodes AM1 et B3LYP sont ceux obtenus avec $\log p$, l'électrophilie, E_{lumo} , et le potentiel respectivement.

Dans le but d'obtenir des modèles plus explicatifs, des modèles avec plusieurs descripteurs doivent être élaborés.

- L'analyse de la matrice de corrélation est indispensable pour élaborer des modèles avec plus d'un descripteur.

Matrice de corrélation (B3LYP)

	E_{Homo}	E_{lumo}	dureté η	potentiel	électrophilie
Elumo	0.887				
dureté η	0.348	0.742			
Potentiel	0.961	0.980	0.594		
électrophilie	-0.903	-0.996	-0.710	-0.985	
$\log p$	-0.867	-0.970	-0.709	-0.954	0.966

Matrice de corrélation (AM1)

	E_{Homo}	E_{lumo}	dureté η	potentiel	électrophilie
Elumo	0.754				
dureté η	0.388	0.898			
Potentiel	0.882	0.974	0.776		
électrophilie	-0.798	-0.996	-0.863	-0.987	
$\log p$	-0.801	-0.935	-0.775	-0.944	0.938

D'après les matrices de corrélation, les descripteurs présents dans les modèles avec un seul descripteur sont fortement corrélés entre eux, ceci dit qu'on ne peut pas les combiner pour obtenir un modèle MLR. Pour contourner ce problème nous avons utilisé la méthode des moindres carrés partiels (PLS).

Les meilleurs modèles QSAR obtenus avec la méthode d'analyse PLS contiennent deux descripteurs, $\log P$ et le potentiel chimique.

2- Méthode des moindres carrés partiels PLS

Chapitre III : Résultats et discussion

Les modèles avec leurs paramètres statistiques sont donnés dans les tableaux ci-dessous :

Tableau.7 : Modèles QSAR avec deux descripteurs (**Méthode B3LYP**)

<i>Algue</i>	<i>Modèle</i>	R^2	R^2_{cv}	F
Algue 1	pT48 = 1,9751 + 1,3275 log p + 36,7043 μ (1)	89,43%	85,89%	80,41
	pT72 = 1,5359+ 1,2089log p +31,7097 μ (2)	83,09%	77,96%	46,68
	pT96 = 2,0556+ 1,2892log p + 37,5874 μ (3)	89,00%	84,66%	76,91
Algue 2	pT96-1 = 2,0086+ 1,0528log p + 30,9346 μ (4)	75,00%	63,56%	28,51

Tableau 8 : Modèles QSAR avec deux descripteurs (**Méthode AM1**)

<i>Algue</i>	<i>Modèle</i>	R^2	R^2_{cv}	F
Algue 1	pT48 = 1,9751 + 1,3275 log p + 36,7043 μ	86,92%	81,93%	63,18
	pT72 = 1,5359+ 1,2089log p +31,7097 μ	79,59%	75,76%	77,89
	pT96 = 2,0556+ 1,2892log p + 37,5874 μ	86,04%	80,08%	58,58
Algue 2	pT96-1 = 2,0086+ 1,0528log p + 30,9346 μ	73,00%	62,00%	25,69

D'après les deux tableaux il s'avère que

- ✓ Les deux méthodes B3LYP/DFT et AM1/semi empirique donnent des modèles avec la même tendance
- ✓ La méthode B3lyp a montré des paramètres statistiques un peu meilleurs que ceux obtenus avec la méthode AM1.

Le modèle obtenu avec l'algue 1 avec une durée d'exposition de 96h est utilisé pour prédire la toxicité des phénols et de la corrélérer avec la toxicité vis-à-vis l'algue 2 (pT96-1), pour la même durée d'exposition 96h, le graphe ci-dessous représente la corrélation obtenue entre la toxicité prédite et la toxicité expérimentale.

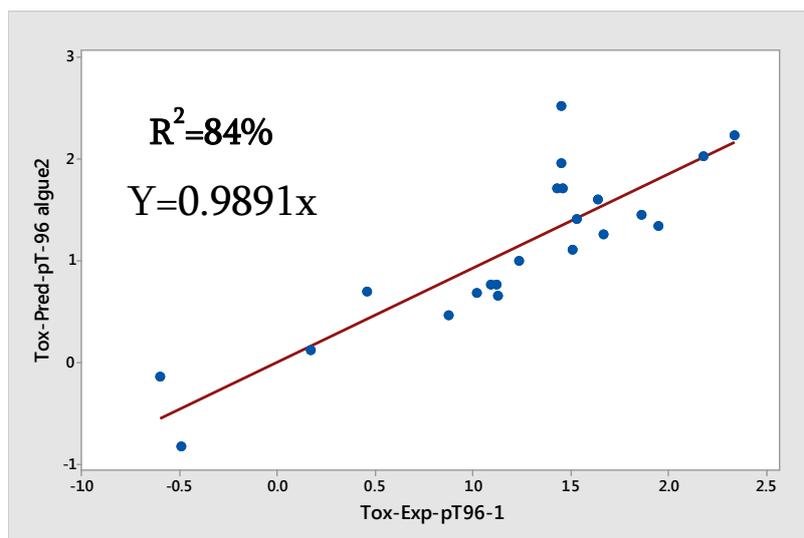


Figure.1. Corrélation entre la toxicité prédite (Eq.3) et la toxicité observée

Une bonne corrélation est obtenue ce qui confirme que modèle élaboré peut être utilisé pour expliquer et prédire les toxicités des deux algues marines (algue 1: *Dunaliella tertiolecta* et algue 2: *Chlorella vulgaris*)

3- Validation des meilleurs modèles QSAR

1) *Validation interne*

Pour s'assurer de la stabilité des modèles élaborés, une validation croisée et une y-randomisation sont effectuées.

-**LOO leave one out** : les valeurs de R^2_{cv} pour les modèles Eq.1, Eq.2 et Eq.3 dépassent 0,5. En plus ces derniers sont proches de R^2 . Ces résultats montrent bien que notre trois meilleurs modèles QSAR ne sont pas sensibles à cette opération de mettre à part une molécule et de la remettre dans la série d'apprentissage. C'est une première information sur la stabilité de nos modèles QSAR.

Chapitre III : Résultats et discussion

- Y-Randomisation :

Tableau .9. Les dix premières itérations d'Y-randomisation (Algue 1)

Itération	$R^2_r(\text{eq},1)$	$R^2_{cv}(\text{eq},1)$	$R^2_r(\text{eq},2)$	$R^2_{cv}(\text{eq},2)$	$R^2_r(\text{eq},3)$	$R^2_{cv}(\text{eq},3)$
1	3,18%	0,00%	1,09%	0,00%	10,14%	0,00%
2	0,04%	0,00%	13,57%	0,00%	26,83%	1,22%
3	7,96%	0,00%	4,08%	0,00%	2,25%	0,00%
4	8,61%	0,00%	16,41%	0,00%	24,30%	7,19%
5	3,55%	0,00%	6,13%	0,00%	2,51%	0,00%
6	12,13%	0,00%	5,79%	0,00%	17,00%	0,00%
7	10,43%	0,00%	1,75%	0,00%	0,50%	0,00%
8	13,89%	0,00%	3,17%	0,00%	1,79%	0,00%
9	5,08%	0,00%	2,27%	0,00%	14,16%	0,00%
10	7,60%	0,00%	25,72%	0,00%	12,47%	0,00%

Tableau.10. Les dix premières itérations d'Y-randomisation (Algue 2)

Itération	$R^2_r(\text{eq},4)$	$R^2_{CvR}(\text{eq},4)$
1	12,17%	0,00%
2	13,26%	0,00%
3	8,23%	0,00%
4	32,83%	0,00%
5	2,99%	0,00%
6	39,97%	19,55%

Chapitre III : Résultats et discussion

7	9,38%	0,00%
8	12,18%	0,00%
9	9,36%	0,00%
10	6,92%	0,00%

Les résultats des 10 premières itérations de l'Y-randomisation sont donnés respectivement dans les tableaux 8 et 9 pour les deux types d'algue, il s'avère que les modèles élaborés avec les valeurs randomisées ont des R^2 inférieur à 40%, ces résultats indiquent que nos meilleurs modèles n'ont pas dû à une chance et ils sont stables.

D'après les résultats de la validation interne on peut conclure que ces modèles sont stables et possèdent **un grand pouvoir explicatif**.

2- Validation externe

Pour vérifier le pouvoir prédictif pour les modèles, une série de test composé de 6 molécules est utilisée.

L'analyse des résidus a montré composés possèdent un résidu standardisé $>2,5$ ce qui indique que ces derniers représentent des outliers (observations aberrantes).

L'élimination de ces deux molécules de la série de test a permis d'obtenir les résultats ci-dessous :

Tableau. 11 Paramètres statistiques de la série de test

	Modèle (log p+ potentiel)			
	<i>N</i>	<i>R</i> ²	<i>R</i> ² _{cv}	<i>SD</i>
pT48	4	96.94%	90.95%	0.171023
pT72	4	98.26%	90.79%	0.119871
pT96	4	99.10%	95.23%	0.0715411
pT96-1	4	92.17%	68.99%	0.211237

Tableau. 12. Critères de Tropsha

Toxicité	Critères de Tropsha			
	R^2_o	k	$R^2 - R^2_o/R^2$	$ R^2 - R^2_o $
pT48	94,15%	1,0931	0,0287	0,027
pT72	97,85%	1,1444	0,0041	0,041
pT96	92,62%	0,9879	0,0653	0,06
pT96-1	90,34%	0,9959	0,0195	0,018

D'après les deux tableaux ci-dessus il est clair que tous les paramètres statistiques (R^2 et R^2_{cv} de la série de test sont élevés et tous les critères de Tropsha sont vérifiés.

Les résultats obtenus montrent que nos modèles possèdent de grands pouvoirs prédictifs.

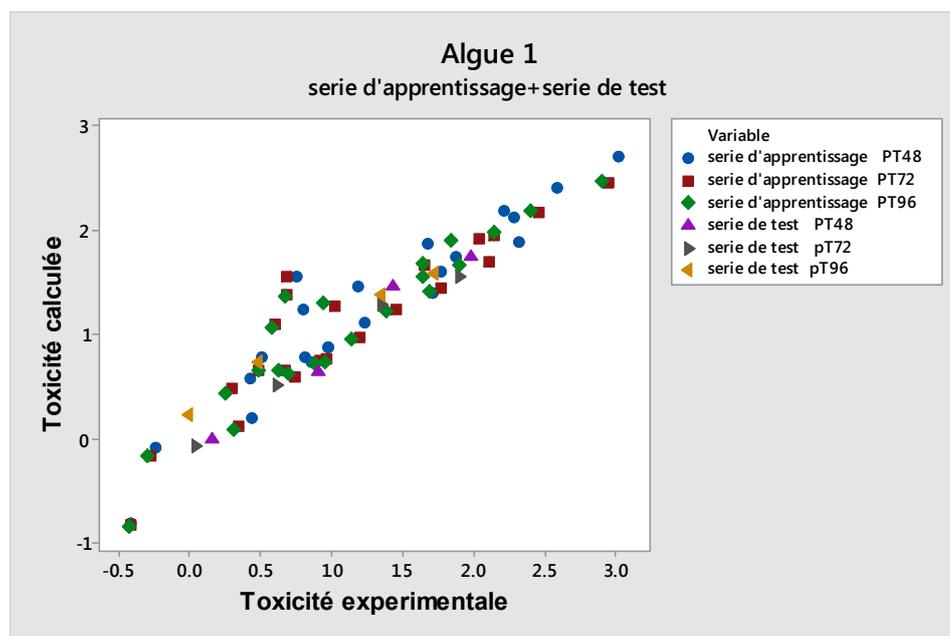


Figure.2. Variation de la toxicité calculée en fonction de la toxicité expérimentale pour l'algue 1

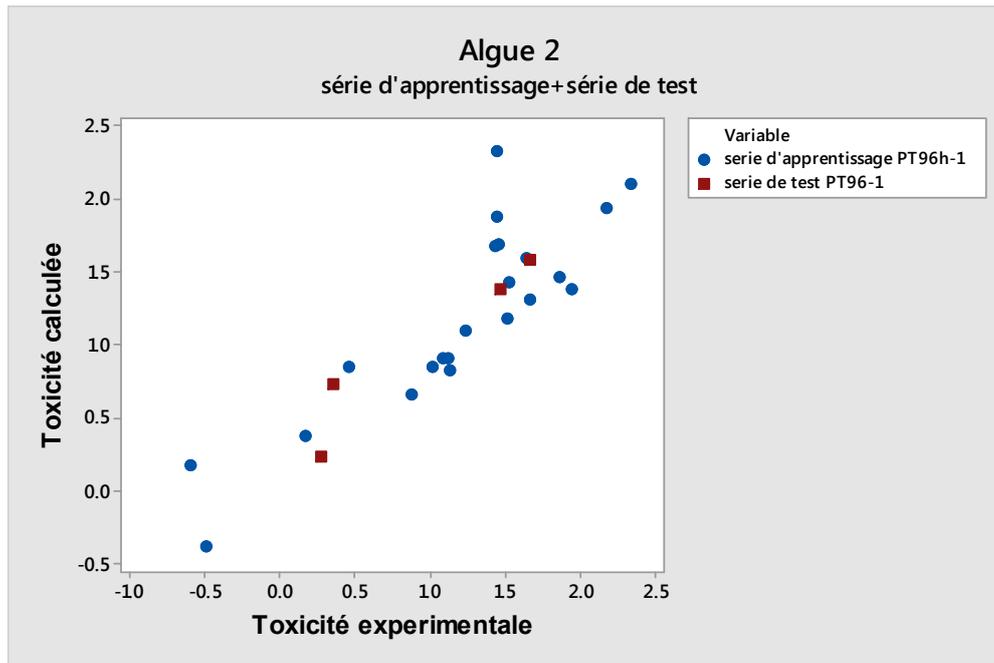


Figure.3. Variation de la toxicité calculée en fonction de la toxicité expérimentale pour l'algue 2

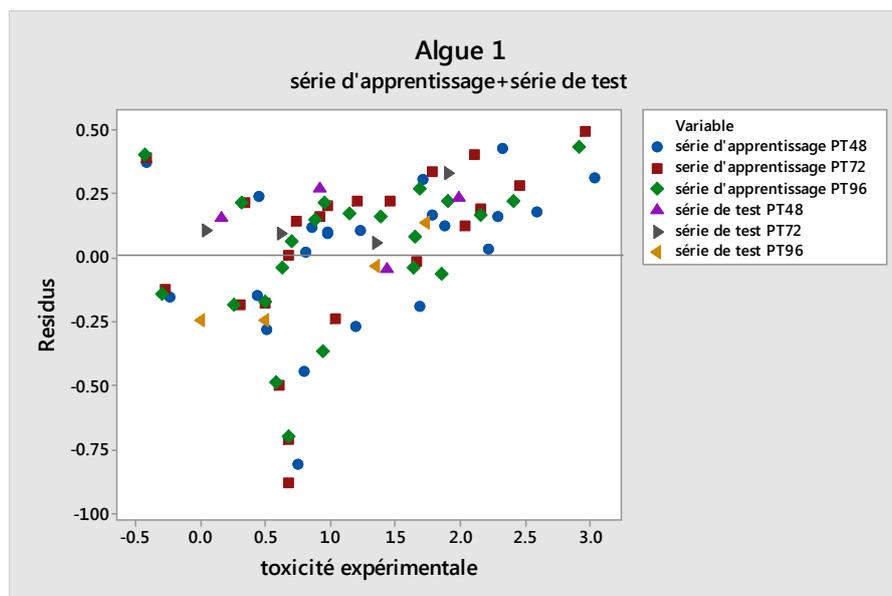


Figure.4. Variation des résidus en fonction de la toxicité expérimentale pour l'algue 1

D'après la figure 4, il s'avère que les résidus sont distribués autour du zéro pour les quatre toxicités expérimentales.

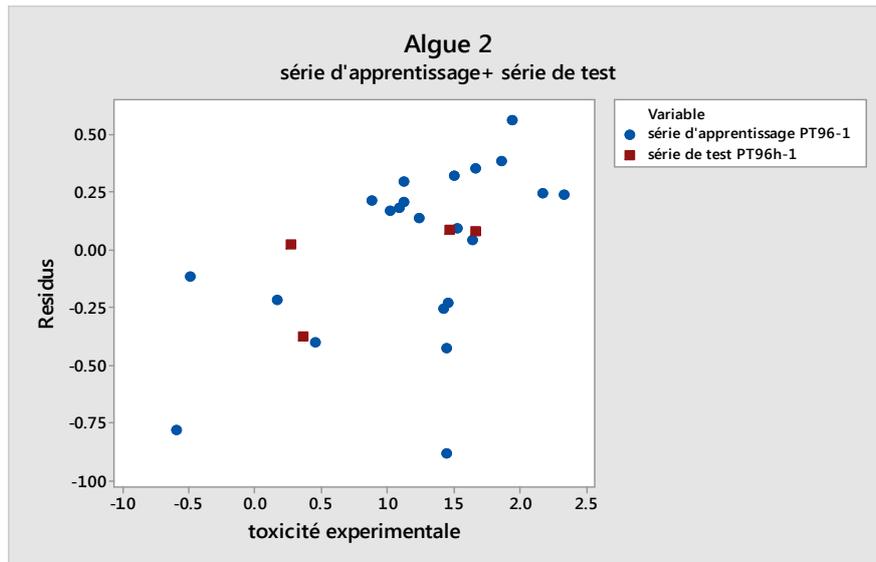


Figure.5. Variation des résidus en fonction de la toxicité expérimentale pour l'algue 2

La figure 5 montre qu'il y a une distribution plus au moins homogène des résidus autour du zéro

Domaine d'applicabilité

C'est une représentation des valeurs des résidus standardisés des composés de la série d'apprentissage en fonction de « *leverage* », ce domaine nous donne des informations sur les valeurs aberrantes (outliers). Si un des composés a un leverage supérieur à la valeur seuil $h^* = 3(p'+1)/n$ (avec p' le nombre de descripteur et n le nombre d'observations) ce composé est considéré comme un outlier hors du domaine d'applicabilité du modèle élaboré et la prédiction de sa toxicité est impossible.

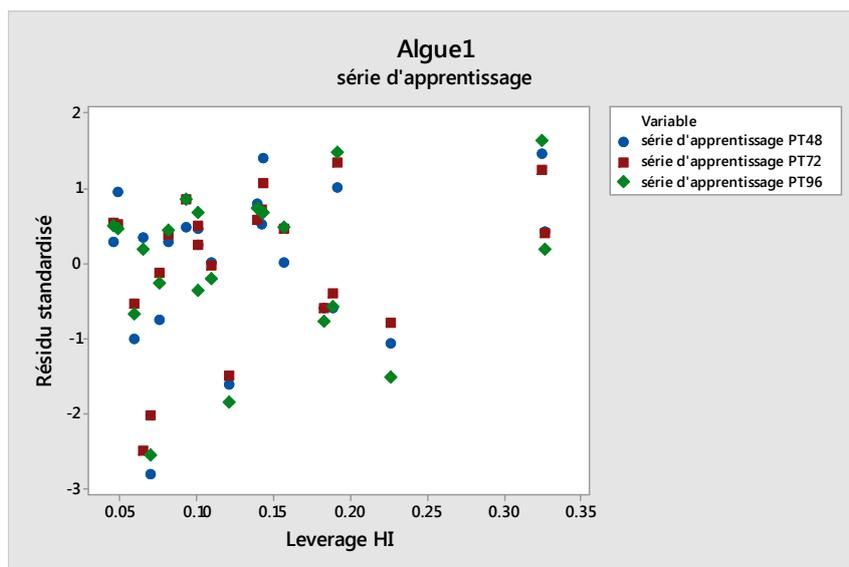


Figure.5. Domaine d'applicabilité des meilleurs modèles pour l'Algue 1

La figure 5 montre bien qu'il n'y a pas des composés qui ont des h supérieurs à la valeur seuil $h^*=0.40$.

Interprétation mécanistique des modèles QSAR

Dans les meilleurs modèles obtenus pour les différentes durées d'exposition pour les deux algues, les principaux facteurs qui peuvent expliquer la toxicité observée de cette série de phénols, sont le paramètre de lipophilie, présenté par $\log P$, et le potentiel chimique μ calculé selon la méthode B3lyp. Les deux descripteurs ont des signes positifs, ce qui signifie que le composé le plus toxique est caractérisé par une forte lipophilie et un grand potentiel chimique ce qui confirme encore une fois le mécanisme surface-réponse caractérisant la toxicité des phénols vis-à-vis de l'espèce *Tetrahyména pyriformis* [7-9].

Conclusion

La toxicité des 30 composés phénoliques a été étudiée en respectant tous les critères d'OECD et toute la méthodologie QSAR

Dans cette étude le descripteur électronique qui est le potentiel chimique qui exprime l'interaction entre la molécule toxique et l'organisme (espèce étudiée) a été utilisé en combinaison avec le coefficient de partage $\log p$ qui exprime le transfert entre deux phases

Chapitre III : Résultats et discussion

lipidique et aqueuse sont utiles pour étudier et prédire la toxicité de deux types d'algues marines, *Dunaliella tertiolecta* et *Chlorella vulgaris*.

Ces paramètres sont calculés avec deux méthodes la DFT/B3LYP6-31G (d, p) et une méthode semi-empirique 'AM1' en utilisant la méthode statistique PLS

La combinaison de ces deux descripteurs a augmenté le pouvoir explicatif des modèles QSAR représentés par les équations 1, 2,3, et 4.

Les modèles développés sont simples, clairs et ils ont une bonne stabilité et un grand pouvoir prédictif (tous les critères de Tropsha sont vérifiés). Ces modèles peuvent être utilisés pour la prédiction de la toxicité d'autres phénols ou composés narcoses polaires vis-à-vis d'autres algues.

Références bibliographiques

- [1] E. Papa, F. Villa, P. Gramatica, Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow), *Journal of chemical information and modeling*, 45 (2005) 1256-1266.
- [2] M.W. Toussaint, T.R. Shedd, W.H. van der Schalie, G.R. Leather, A comparison of standard acute toxicity tests with rapid-screening toxicity tests, *Environmental toxicology and chemistry*, 14 (1995) 907-915.
- [3] M.D. Ertürk, M.T. Saçan, M. Novic, N. Minovski, Quantitative structure-activity relationships (QSARs) using the novel marine algal toxicity data of phenols, *Journal of Molecular Graphics and Modelling*, 38 (2012) 90-100.
- [4] D. calculations were performed using Gaussian, 09, Revision D. 01, MJ Frisch, et al, Gaussian, Inc., Wallingford CT, (2009).
- [5] K. Mansouri, C. Grulke, A. Richard, R. Judson, A. Williams, An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling, *SAR and QSAR in Environmental Research*, 27 (2016) 911-937.
- [6] L.M. Arterburn, E.B. Hall, H. Oken, Distribution, interconversion, and dose response of n-3 fatty acids in humans, *The American journal of clinical nutrition*, 83 (2006) 1467S-1476S.
- [7] S. Enoch, M. Cronin, T.W. Schultz, J. Madden, An evaluation of global QSAR models for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*, *Chemosphere*, 71 (2008) 1225-1232.
- [8] M.T. Cronin, A.O. Aptula, J.C. Duffy, T.I. Netzeva, P.H. Rowe, I.V. Valkova, T.W. Schultz, Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*, *Chemosphere*, 49 (2002) 1201-1221.
- [9] A.O. Aptula, T.I. Netzeva, I.V. Valkova, M.T. Cronin, T.W. Schultz, R. Kühne, G. Schüürmann, Multivariate discrimination between modes of toxic action of phenols, *Quantitative Structure-Activity Relationships*, 21 (2002) 12-22.

Conclusion Générale

Conclusion générale

Ce présent travail a pour but d'étudier la toxicité des composés phénoliques en utilisant la méthode de modélisation QSAR

L'objectif est de développer des modèles QSAR fiables et clairs pour la prédiction de la toxicité de 30 composés phénoliques vis-à-vis deux algues marines '*Chlorella vulgaris*' et '*Dunaliella tertiolecta*'.

Deux méthodes de calcul, semi-empirique AM1 et la théorie de fonctionnelle de densité B3LYP/6-31G(d,p) sont utilisées pour le calcul des descripteurs.

Les meilleurs modèles sont obtenus avec les descripteurs calculés par la méthode DFT/B3lyp, en utilisant la méthode d'analyse PLS, ont montré que la combinaison des deux descripteurs, le potentiel chimique électronique avec le coefficient de partition 'log P' peut expliquer 89% de la toxicité des composés phénoliques étudiés.

On a effectué la validation interne d'une série d'apprentissage composée de 22 molécules avec deux méthodes (la validation croisée et l'Y-randomisation) pour montrer que notre modèle n'est pas due à une chance, on a utilisé également la validation externe pour une série de test constituée de 4 molécules. Les résultats de cette validation ont montré que les 5 critères de Tropsha sont vérifiés, ce qui confirme la stabilité, la fiabilité et le grand pouvoir prédictif des modèles QSAR obtenus.

D'après les modèles QSAR élaborés, les principaux paramètres qui peuvent expliquer la toxicité de cette série de phénols sont le potentiel chimique électronique et le coefficient de partage log p, ces deux derniers ont des coefficients positifs ce qu'indique qu'il y a une relation positive entre la toxicité et les deux descripteurs.

Le facteur d'hydrophobie, tel qu'il est exprimé par log P avec un coefficient positif, est utile pour décrire le transport vers le site d'action, la contribution des effets électroniques, quantifiée par le potentiel chimique électronique explique l'interaction avec le site d'action.

Les modèles élaborés pour ces les deux algues marines peuvent être utilisés pour la prédiction de nouvelles bases de données dont le mode d'action est narcose-polaire comme

Conclusion générale

ils peuvent être utilisés pour étudier la toxicité d'autres espèces aquatiques tels que les protozoaires, les daphnies et les poissons.

ملخص

في الآونة الأخيرة تزايد الاهتمام بالطحالب بسبب اهتماماتها الصيدلانية والغذائية والاقتصادية في هذه الدراسة، تتم دراسة سمية المركبات الفينولية في طحلبين بحريتين باستخدام طريقتين كميتين، $DFT/B3LYP-6-31G^*$ و AM1، لحساب الواصفات، نماذج QSAR التي تم الحصول عليها هي: وضعت مع طريقة الانحدار PLS الاستقرار والقوة التنبؤية لأفضل طرازات QSAR التي تم الحصول عليها عن طريق التحقق من الصحة الداخلية والعشوائية Y والتحقق الخارجي، يتم التحقق من مجال قابليتها للتطبيق من خلال طريقة "leverage" تؤكد النتائج أن أساليب QSAR موثوقة، وتتمتع باستقرار جيد وقدرة تنبؤية

Résumé

Ces derniers temps une attention croissante portée aux algues à cause de leurs intérêts industriels pharmaceutiques, alimentaires et économiques.

Dans la présente étude, la toxicité de 30 composés phénoliques vis-à-vis deux algues marines est étudiée à l'aide de deux méthodes quantiques la $DFT/B3LYP-6-31G^*$ et AM1, pour le calcul des descripteurs, les modèles QSAR obtenus sont élaborés avec la méthode de régression PLS

La stabilité, le pouvoir prédictif des meilleurs modèles QSAR obtenus ont été vérifiés par la validation interne, la Y randomisation et la validation externe, leurs domaine d'applicabilité est vérifié par la méthode de 'leverage'.

Les résultats confirment que les modèles QSAR élaborés sont fiables, ont une bonne stabilité et un grand pouvoir prédictif

ABSTRACT

Recently, there has been increasing attention to algae due to their pharmaceutical, food and economic industrial interests.

In this study the toxicity of 30 phenolic compounds to two marine algae is studied using two quantum methods $DFT/B3LYP-6-31G^*$ and AM1, for the calculation of the descriptors, the obtained QSAR models are developed with the PLS method.

The stability, predictive power of the best obtained QSAR models have been achieved through internal validation, Y randomization and external validation, their applicability domain is verified by the leverage method.

The results confirm that QSAR methods are reliable, stable and predictive.