

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR ANNABA UNIVERSITY

UNIVERSITE BADJI MOKHTAR ANNABA



جامعة باجي مختار – عنابة

Année universitaire: 2020-2021

FACULTE DES SCIENCES

DEPARTEMENT DE CHIMIE

THÈSE

Présentée en vue de l'obtention du diplôme de Doctorat en Sciences

Calculs de modèles hybrides algorithme génétique / réseaux de neurones pour la toxicité et les propriétés des polluants potentiels de l'environnement

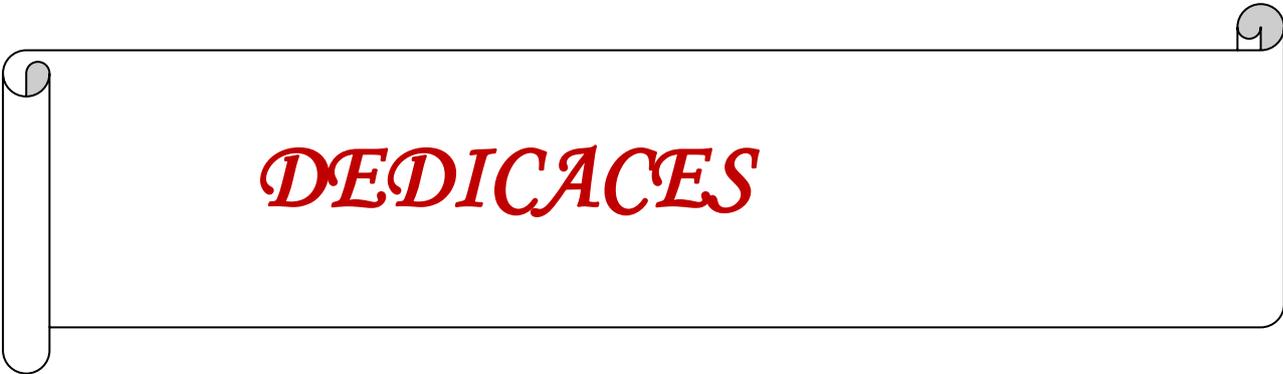
Domaine : Chimie

Spécialité : Chimie Analytique et Environnement

Présentée par : BOUAOUNE AHMED

Devant le jury

Président	M. LARKEM Ali	Pr	Université Badji Mokhtar Annaba
Directeur de thèse	M ^{me} . SALIMA Ali Mokhnache	Pr	Université Badji Mokhtar Annaba
Examineurs	M. BENLECHEB Tahar	MCA	Université Abbas el Ghrour Khenchela
	M. LAHMAR Hicham	MCA	Université de M ^{ed} .S .Ben Yahia Jijel
	M. DEMS Mohamed Abd Esselem	MRA	Centre de Recherche de Biotechnologie (CRBTs) Constantine



DEDICACES



Dédicaces

Je dédie ce Travail :

A tous ceux que j'ai de plus cher au monde :
Mes très chers parents : « ma chère mère,
Et à l'esprit de mon cher père, ma petite famille. »

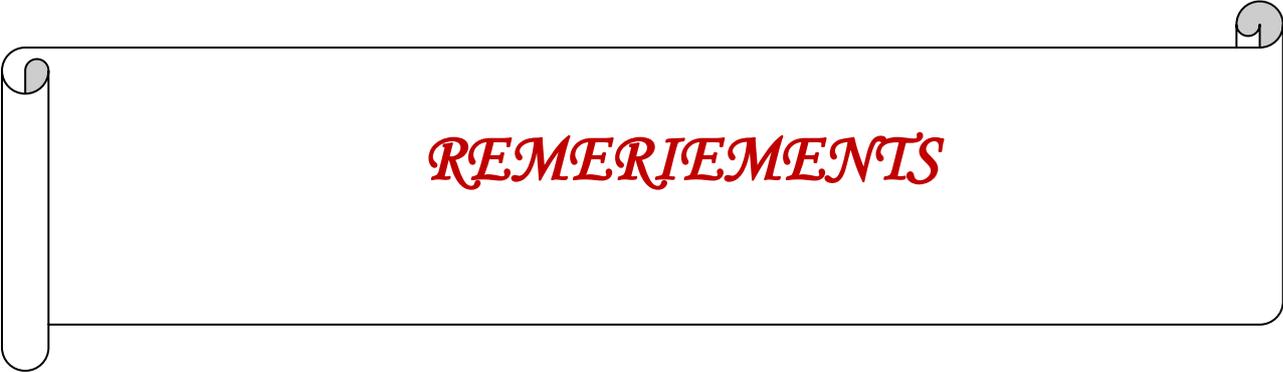
En gratitude de leur aide, soutien sans relâche, leur
amour et leur présence en continus tout au long de mes
Études et de ma vie.

- Que Dieu vous garde -
À mes chers frères .À mes chères sœurs .À Tous les
membres de ma famille. Et à tous mes professeurs aux
différentes étapes de mon éducation.

Egalement, je dédie ce modeste travail à toute l'équipe
du labo 34, département de Chimie
Et à tous mes amis.

A-B





REMERIEMENTS

Remerciements

Avant tout, je remercie Dieu le tout puissant de m'avoir donné la force et la foi et de m'avoir permis d'arriver à ce stade.

La présente étude a été réalisée au laboratoire de Sécurité Environnementale et Alimentaire de l'Université d'ANNABA sous la direction de madame. Salima Ali Mokhnache Professeur, Aussi, je me permets de lui exprimer ma profonde reconnaissance, pour le bienveillant intérêt qu'elle a accordée quant à la réalisation de cette étude.

J'exprime ma profonde et respectueuse gratitude à Monsieur LARKEM Ali, Professeur à l'Université d'Annaba, qui m'a fait l'honneur d'accepter de présider le jury de cette thèse.

Mes vifs remerciements vont également à messieurs BENLECHEHB Tahar Maître de conférences à l'université de Khenchela, AHMER Hichem Maître de conférences à l'université de Jijel, et DEMES Mohamed Abdelessalam Maître de recherche au centre CRBTs Constantine, pour l'honneur qu'ils m'ont fait en acceptant d'examiner ce travail.

Enfin, est avec beaucoup de gratitude que je remercie tous les membres de l'équipe LASEA pour leurs soutiens, leurs amitiés et leurs aides, surtout Haddag Hamza et Kertiou Noureddine. J'ai eu beaucoup de plaisir à partager des bons moments à leurs côtés.

Enfin, Je tiens à présenter ma reconnaissance et mes remerciements à ma famille, qui est ma source d'inspiration et mon plus grand soutien.

ملخص

تم تطوير اثنين من نماذج QSAR لتثبيط نمو الميكروبات بواسطة الأنيلات والفينولات. تم تقسيم مجموعات البيانات المتاحة بشكل عشوائي إلى مجموعات المعايرة والتحقق.

تعتبر نماذج QSAR المقترحة مستقرة وقوية مع أداء جيد للتنبؤ. وهي تنبؤيه للمواد الكيميائية المستخدمة في تطوير النماذج (التحقق الداخلي من المواد الكيميائية للمعايرة) وكذلك للمواد الكيميائية التي لا تستخدم في تطوير النماذج (التحقق الإحصائي الخارجي من المواد الكيميائية المصادق عليها). كما تم وصف مجالات التطبيق لنماذج QSAR. هذه النماذج هي أفضل من تلك المقترحة في العمل الأصلي الذي تم جمع البيانات منه. العوامل التي تحكم الأنشطة البيولوجية هي الحجم الجزيئي والشكل، وتفاعلات الجزيء مع البيئة.

الكلمات المفتاحية :

العوامل السامة، نمو الأنواع الميكروبية، نموذج QSAR الهجين، المصادقة الإحصائية الخارجية، مجال التطبيق، الشبكات العصبية الاصطناعية.

Résumé

Deux modèles QSAR de l'inhibition de la croissance microbienne par les anilines et les phénols ont été développés. Les ensembles de données disponibles ont été divisés au hasard en des ensembles de calibrage et de validation.

Les deux modèles QSAR proposés sont stables, robustes et avec de bonnes performances d'ajustement et de prédiction. Ils sont prédictifs pour les produits chimiques utilisés dans le développement du modèle (validation interne sur les produits chimiques de calibrage) et également pour les produits chimiques non utilisés dans le développement du modèle (validation statistique externe sur les composés de validation). Les domaines d'applications des modèles QSAR ont également été décrits. Ces modèles sont meilleurs que ceux proposés dans le travail original d'où ont été prélevés les données. Les facteurs régissant les activités biologiques sont la taille, la forme moléculaires et les interactions de la molécule avec son milieu environnant ou sa cible.

Mots clés :

Agents toxiques, croissance d'espèces microbiennes, modèle hybride QSAR, validation statistique externe, domaine d'applicabilité, réseaux de neurones artificiels.

Abstract

Two QSAR models of microbial growth inhibition by anilines and phenols have been developed. The available datasets were randomly divided into training and validation sets.

The proposed QSAR models are stable, robust, with good fit and prediction performances. They are predictive for chemicals used in model development (internal validation on training chemicals) and also for chemicals not used in model development (external statistical validation on test chemicals). The applicability domains of the QSAR models have also been described. These models are better than those proposed in the original work from which the dataset was collected. Factors governing biological activities are molecular size, shape and interactions of the molecule with its surrounding environment or its target.

Keywords:

Toxic agents, microbial species growth, hybrid QSAR model, external statistical validation, applicability domain, artificial neural networks.



SYMBOLES ET ABREVIATIONS

SYMBOLES ET ABREVIATIONS

- AM1:** Austin Model 1.
- CAS :** Chemical abstract service
- EQMC:** Ecart quadratique moyen calculé sur l'ensemble de calibrage.
- EQMP:** Ecart quadratique moyen de prédiction.
- EQMP_{ext.}:** Ecart quadratique moyen de prédiction calculé sur l'ensemble de validation externe.
- e_i :** Résidu : différence entre les valeurs observée (y_i) et estimée (\hat{y}_i).
- e_{istd}:** Résidu standardisé.
- F :** Statistique de Fisher.
- FIV :** Facteur d'inflation de la variance.
- GA:** Algorithme génétique (GeneticAlgorithm).
- H :** Matrice de projection, ou matrice chapeau.
- h_i:** Eléments diagonaux de la matrice chapeau.
- IGC₅₀ :** Concentration d'inhibition 50% de la croissance.
- LMO :** Cross-validation by leave-many-out: Validation croisée par omission d'un ensemble d'observations.
- LOO :** Cross-validation by leave-one-out: Validation croisée par omission d'une observation
- log_{k_{wo}} /logP :** Coefficient de partage octanol /eau.
- MSE :** Erreur quadratique moyenne.

- n :** Dimension de la population (échantillon).
- p :** Nombre de descripteurs
- PMC :** Réseaux multicouches Perception.
- PRESS :** Somme des carrés des erreurs de prédiction.
- Q^2_{LOO} :** Coefficient de prédiction.
- Q^2_{Ext} :** Coefficient de prédiction. Externe
- QSAR :** Quantitative Structure/ Activity Relationships.
(Relations Quantitatives Structure/ Activité).
- QSPR :** Quantitative Structure/ Propriety Relationships.
(Relations Quantitatives Structure/ Propriété).
- QSTR :** Quantitative Structure/Toxicité Relative
(Relations Quantitatives Toxicity /Rerrelationships)
- R^2 :** Coefficient de détermination.
- R^2_{adj} :** Coefficient de détermination ajusté.
- RLM (MLR) :** Régression linéaire multiple (Multiple Linear Regression).
- RMSE :** Racine de l'écart quadratique moyen (Root Mean Squared Error).
- RNA :** Réseaux de neurones artificiels.
- S :** Erreur standard.
- SCE :** Somme des carrés des écarts.
- SCT :** Somme des carrés totale.

$\tilde{\mathbf{X}}$: Matrice des valeurs observées des variables explicatives.

$\tilde{\mathbf{X}}'$: Matrice transposée de $\tilde{\mathbf{X}}$.

x_j : Variable explicative.

x_j : $j^{\text{ème}}$ valeur .

y_i : Valeur observée.

\hat{y}_i : Valeur estimée.

$\hat{y}_{(i)}$: Valeur prédite.

SOMMAIRE

Liste des tableaux	
Liste des figures	
symbols et abriviations	1
Introduction générale	2
PARTIE I : ETUDE BIBLIOGRAPHIQUE	
I-Toxicité des anilines et phénols	5
I.1 Notions et définitions	5
I.1.1 Définition de la toxicologie	5
I.1.2.Définition d'un toxique (poison)	6
I.1.3.Définition de la toxicité	6
I.1.4.Définition de la dose	6
I.1.5.Types de toxicité	6
I.1.6. Evaluation de la toxicité	6
I.1.7.Toxicité des produits chimiques organiques	7
I.2. Généralités sur les anilines	8
I.2.1. Historique	8
I.2.2. Propriétés physico-chimiques	8
I.2.3.Stabilité et réactivité	9
I.2.4 Fabrication	10
I.2.5 Utilisation	10
I.2.6 Dangers	11
I.2.7 Informations toxicologiques	11
I.3 Généralités sur les phénols	11
I.3.1 Les phénols dans la nature	11
I.3.2 Utilisation	12

I.3.3 Intérêt industriel	12
I.3.4 Dangers toxiques des phénols	12
I.3.5 Tableau clinique de l'intoxication	13
II- Modélisation QSAR/QSPR	14
II.1 Définition des QSAR/QSPR	15
II.2 Descripteurs moléculaires	15
II.2.1 Définition	16
II.2.2 Types de descripteurs moléculaires	16
II.3 Modèles QSAR de la toxicité (QSTR)	18
III. Bases de modélisation moléculaire	21
III-1 Généralités	21
III-2 Méthodes semi-empirique utilisées.	22
III-2.1 Cadre Hartree - Fock – Roothaan	22
III.2.2 Analyse de population de Mulliken	25
III.2.3 Méthodes semi-empiriques.	26
III.3.Champ de force.	31
III.3.1Définition	31
III.3.2 Quelques exemples	31
III.3.3 Représentation simple d'un champ de force	32
III.3.4 Champ de force MM+	38
IV.Développement et validation de modèles QSAR	40
IV-1 Modélisation	40
IV-2 Régression linéaire multiple (MLR)	40
IV-3 Réseaux de neurones artificiels	41
IV-2 Développement et évaluation de modelé.	42
IV-2.1 Sélection d'un sous-ensemble de descripteurs	42
IV-2.2 Principe de l'algorithme génétique (AG)	42

IV-2.3 Initialisation aléatoire du modèle	43
IV-2.4 Etape de croisement	43
IV-2.5 Etape de mutation	43
IV-2.6 Conditions d'arrêt	44
IV-3.1 Robustesse du modèle	45
IV-3.2 Domaine d'application	45
IV-3.3 Test de randomisation	46
IV-3.4 Validation externe	46

PARTIE II : APPLICATION

I- Toxicité des anilines : Approche QSAR	48
I-1 Collecte de données et méthodologie	48
I-2 Présentation et discussion du modèle QSAR	50
I-2-1 Qualités internes du modèle QSAR	50
I-2-2 Qualité externe du modèle QSAR	53
I-2-3 Qualité d'ajustement du modèle QSAR	54
I-2-4 Domaine d'application	54
I-2-5 Test de randomisation	55
I-3 Interprétation mécanistique du modèle	56
I-4 Comparaison avec le modèle original	57
II. Toxicité des phénols: Approche QSAR	62
II-1 Collecte de données et méthodologie	64
II-2 Présentation et discussion du modèle QSAR	64
II-2-1 Qualités internes du modèle QSAR	64
II-2-2 Qualité externe du modèle QSAR	69
II-2-3 Qualité d'ajustement du modèle QSAR	70
II-2-4 Domaine d'application	70
II-2-5 Test de randomisation	71

II-3 Interprétation mécanistique du modèle	72
II-4 Comparaison avec le modèle original	72
II-5 Modèle par réseaux de neurones artificiel	78
Conclusion générale	85
Références bibliographiques	88
ANNEXES	96



Listes des tableaux

LISTE DES TABLEAUX

Tableau 1. Données de toxicité pour les produits chimiques de haute production	7
Tableau 2. Propriétés physiques.	10
Tableau 3. Mécanismes possibles pour la toxicité.	19
Tableau 4. Etude comparative des techniques <i>ab initio</i>, semi –empirique et mécanique moléculaire.	39
Tableau 5. Milieu de croissance de <i>Tetrahymona pyriformis</i>.	51
Tableau 6. Matrice de corrélation des descripteurs du modèle.	53
Tableau 7. Valeurs de $R_{3v} + R_{Gyr}$ et de la concentration inhibitrice de la croissance pour l'ensemble des 48 anilines.	54
Tableau 8. Paramètres statistiques pour l'ensemble de calibrage.	54
Tableau 9. Valeurs de $pIGC_{50}$ expérimentales prédites et calculées, leviers et résidus standardisés de prédictions des 48 anilines.	55
Tableau 10. Paramètres statistiques pour l'ensemble de validation.	56
Tableau 11. Paramètres statistiques du modèle de l'approche Schultz <i>et al.</i>	61
Tableau 12. Valeurs des descripteurs selon Schultz <i>et al.</i> valeurs prédites et calculées, de valeurs des leviers et des résidus standardisés de prédiction.	62
Tableau 13. Valeurs de $\log K_{ow}$, $S - CH_3$ et de la concentration inhibitrice de la croissance pour l'ensemble des 95 phénols.	66
Tableau 14. Matrice de corrélation des descripteurs du modèle.	67
Tableau 15. Paramètres statistiques pour l'ensemble de calibrage (Phénols).	67
Tableau 16. Valeurs de $pIGC_{50}$ expérimentales prédites et calculées, leviers et résidus standardisés de prédictions des 95 phénols.	68
Tableau 17. Paramètres statistiques pour l'ensemble de validation pour les 33 phénols.	70
Tableau 18. Paramètres statistiques du modèle de l'approche Schultz <i>et al.</i> (Phénols).	74

Tableau 19. Valeurs des descripteurs selon Schultz <i>et al.</i> valeurs prédites et calculées de $pIGC_{50}$, valeurs des leviers et des résidus standardisés de prédictions (Phénols).	75
Tableau 20. Paramètres statistiques du modèle RNA.	80
Tableau 21. Valeurs expérimentales, calculées et prédites de $pIGC_{50}$ et résidus.	81



Listes des figures

LISTE DES FIGURES

Figure 1 : Milieu et différents éléments pouvant affecter l'écosystème et l'organisme vivant.	5
Figure. 2 : Structures de quelques phénols naturels.	12
Figure 3 : Différentes espèces aquatiques utilisées pour l'étude de la toxicité aquatique.	19
Figure 4 : Représentation schématique des quatre contributions à un champ de force de MM élongation de liaison, flexion angulaire.	34
Figure 5 : Sous un terme extra - planaire, l'atome d'oxygène de la cyclobutanone est prévu de se situer hors du plan du cycle / gauche plutôt que dans le plan .	35
Figure 6 : Deux façons pour modéliser les contributions de la variation d'angle. extra planaire.	36
Figure 7 : Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces.	37
Figure 8 : Représentation d'un neurone formel.	41
Figure 9 : Représentation d'un réseau de neurones.	42
Figure 10 : Valeurs expérimentales et prédites par LOO de l'ensemble de calibrage.	55
Figure 11 : Droite d'ajustement du modèle QSAR.	57
Figure 12 : Diagramme de Williams du modèles QSAR.	58
Figure 13 : Test de randomisation du modèle QSAR.	59
Figure 14 : Droite d'ajustement du modèle basé sur l'approche de Schultz <i>et al.</i>	63
Figure 15 : Diagramme de Williams du modèle QSAR basé sur l'approche de Schultz <i>et al.</i>	64
Figure 16 : Droite d'ajustement du modèle QSAR (Phénols).	71
Figure 17 : Diagramme de Williams du modèle QSAR (Phénols).	72
Figure 18 : Test de randomisation du modèle QSAR (Phénols).	72
Figure 19 : Droite d'ajustement du modèle basé sur l'approche de Schultz <i>et al.</i> (Phénols).	77
Figure 20 : Diagramme de Williams du modèle QSAR basé sur l'approche de Schultz <i>et al.</i> (Phénols).	78

Figure 21 : Evolution des erreurs EQMC (SDEC° et EQMP (SDEP) en fonction du nombre de neurones.	79
Figure 22 : Evolution des erreurs EQMC (SDEC) et EQMP (SDEP) en fonction du nombre d'itérations	80 83
Figure 23 : Droite d'ajustement du modèle RNA-QSAR.	84
Figure 24 : Valeurs des résidus en fonction des valeurs expérimentales de <i>pIGC₅₀</i>	



INTRODUCTION GÉNÉRALE

La détermination expérimentale systématique au laboratoire de toutes les données concernant les produits chimiques se traduit par une lourde charge, économiquement inacceptable, pour les industriels et les autorités de régulation, et dépasse de loin, de toutes les façons, les capacités de recherche disponibles. Étant donné que le risque pour la santé associé à l'exposition doit être évalué pour des centaines de substances chimiques, et vu la réduction des études. La expérimentales de toxicité est encouragée pour des raisons éthiques et économiques, le développement de méthodes alternatives sur ordinateur (*in silico*) est fortement encouragé.

Les techniques QSAR (Quantitative Structure Activity Relationship) pour s'appliquent avec succès dans les études de toxicité de certains composés vis-à-vis des animaux et de certaines espèces végétales. Les modèles QSAR proprement dit sont des modèles qui prédisent des activités biologiques, comme la toxicité. Les modèles qui prédisent des propriétés, physicochimiques ou biochimiques, comme des coefficients de partage ou des constantes métaboliques à partir de la structure moléculaire sont des modèles de relation quantitative structure-propriété (QSPR). Ces études sont d'un grand intérêt. Les relations structure / toxicité constituent de nos jours des outils (*in silico*) utiles qui aident à l'établissement d'expressions mathématiques simples utiles pour le calcul à l'avance des propriétés toxiques.

Les études QSAR courantes en toxicologie mettent en jeu des descripteurs 2D comme le coefficient de partage octanol / eau ($\log P$) ou d'autres paramètres, qui simulent les différentes interactions moléculaires.

Il est couramment admis que l'action des polluants sur les espèces animales s'exprime par une perturbation fonctionnelle des membranes cellulaires. Donc la toxicité éventuelle d'une molécule est profondément liée à sa tendance à s'accumuler dans les membranes cellulaires. Pour décrire les membranes cellulaires on peut utiliser le modèle simple constitué par un milieu apolaire, à savoir l'octanol. Il n'est donc pas surprenant de construire des modèles QSAR satisfaisants en faisant intervenir le coefficient de partage octanol / eau ($\log P$). Pour améliorer les capacités prédictives de ces modèles, on y incorpore généralement d'autres descripteurs 2D ou 3D.

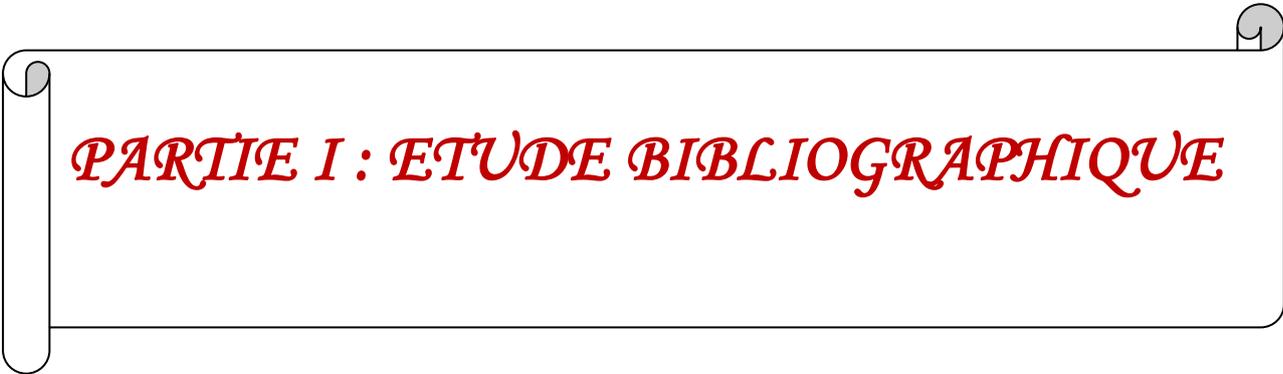
Nous nous sommes intéressés, dans ce travail à la toxicité des phénols et des anilines diversement substitués en utilisant des protozoaires ciliés comme système biologique. La

concentration d'inhibition à 50% de la croissance (IGC_{50}), d'une population de *Tetrahymena pyriformis* a été prise en compte : le logarithme de son inverse, $pIGC_{50} = \log (IGC_{50})^{-1}$, a servi d'indicateur de toxicité. L'ensemble de données relatives aux phénols a été tiré du travail de Schultz *et al.* [39] Comme pour les anilines et les modèles obtenues dans ce travail comparés à la source des données précédemment mentionnée.

Cette thèse comporte deux grandes parties :

- La première partie comporte des généralités sur les composés étudiés ainsi que leurs toxicités en plus les différentes méthodes de la modélisation moléculaire utilisées et les bases des modèles QSAR, complétée par les méthodes statistiques appliquées.
- La présentation et la discussion des principaux résultats fournis par la modélisation QSAR pour les deux ensembles relatifs aux anilines et phénols résume la seconde partie.

Nous avons réuni dans l'annexe en plus de l'article sanctionnant ce travail, les structures des 48 anilines et des 95 phénols, aussi que leurs numéros de CAS.



PARTIE I : ETUDE BIBLIOGRAPHIQUE

I. Toxicité des anilines et phénols

L'organisme humain ainsi que tous les autres organismes vivants sont en relation avec leur milieu par un ensemble d'échanges qui contribuent à maintenir un équilibre dynamique. Par exemple, la respiration permet d'absorber l'oxygène de l'air et d'y rejeter du dioxyde de carbone. Quoique nous fassions, le milieu nous influence et nous l'influons. Ce principe d'action-réaction signifie que toute action a des conséquences. Le milieu ne constitue cependant pas un tout homogène, mais plutôt un ensemble composé de nombreux éléments, comprenant les produits chimiques qui peuvent affecter la santé des organismes vivants et l'équilibre de l'écosystème. (*Figure 1*) [1].

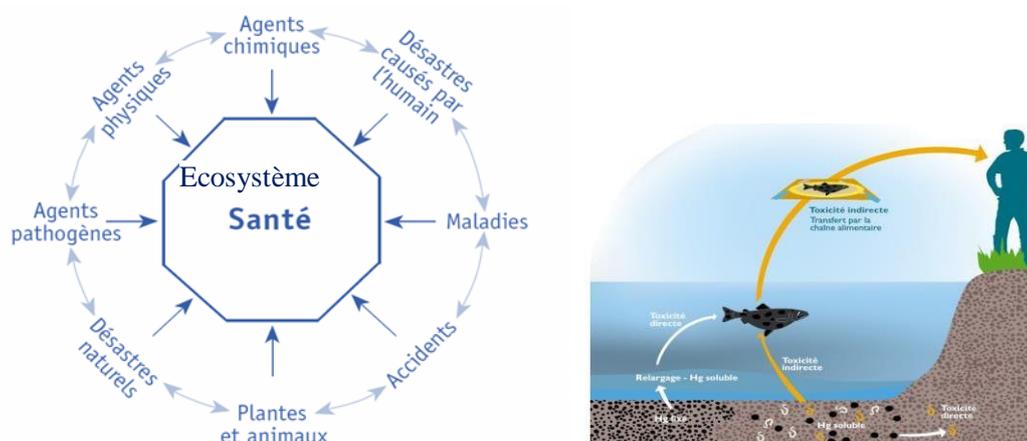


Figure 1: Milieu et différents éléments pouvant affecter l'écosystème et l'organisme vivant [1]

Chaque année, l'industrie met des centaines de nouveaux produits sur le marché, venant ainsi accroître le nombre de ceux qui existent déjà. Il est important de connaître l'innocuité (qualité de ce qui n'est pas nuisible) ou la nocivité (caractère de ce qui est nuisible) des produits chimiques pour bien en saisir les effets sur notre santé et sur notre environnement.

I.1 Notions et définitions [2]

I.1.1 Définition de la toxicologie

La toxicologie est depuis longtemps reconnue comme étant la science des poisons. Elle étudie les effets nocifs des substances chimiques sur les organismes vivants.

I.1.2 Définition d'un toxique (poison)

Un poison, ou toxique, est une substance capable de perturber le fonctionnement normal d'un organisme vivant. Il peut être de source naturelle (ex : poussières, pollen) ou artificielle (ex. : urée formaldéhyde), ou de nature chimique (ex. : acétone, benzène, anthrax...) ou biologique (ex. : aflatoxines, anthrax).

I.1.3 Définition de la toxicité :

Il s'agit de la capacité inhérente à une substance chimique de produire des effets nocifs chez un organisme vivant et qui en font une substance dangereuse. On parle alors d'un effet toxique, cet effet peut être local qui survient au point de contact, ou bien systémique qui survient à un endroit éloigné du point de contact initial. Les principales façons d'absorption au contact sont : l'inhalation (voie respiratoire), l'absorption par la peau (voie cutanée), l'ingestion (voie digestive).

I.1.4 Définition de la dose

La dose est la quantité d'une substance à laquelle un organisme vivant est exposé.

I.1.5 Types de toxicité

a. Toxicité aiguë: est définie comme celle qui résulte de l'exposition unique et massive (ou de doses ramassées dans le temps) à un produit chimique entraînant des dommages corporels pouvant conduire à la mort .Elle introduit la notion de dose « absorbée » (par ingestion, inhalation ou contact cutané) et se mesure par la DL 50 (dose létale, ou dose provoquant la mort de 50% des animaux exposés à une dose unique du produit incriminé), exprimée en mg/kg de l'animal d'expérience retenu.

b. Toxicité chronique : est le résultat de l'exposition prolongée à plus ou moins faible dose à un xénobiotique toxique dont les effets néfastes ne se feront sentir que quelques mois à quelques années voire des dizaines d'années plus tard.

c. Toxicité subaiguë : correspond à un stade d'exposition intermédiaire de l'ordre de trois mois.

I.1.6 Evaluation de la toxicité

On peut citer quatre catégories pour évaluer un effet toxique (toxicité)

- Les études épidémiologiques, qui comparent plusieurs groupes d'individus ou les études de cas.
- Les études expérimentales in vivo, qui utilisent des animaux (ex. : lapin, rat et

I. Toxicité des anilines et phénols

souris, *Tetrahymina-pyriiformis*, poissons, algues...) ou les bio-essais sont utilisés

- Les études in vitro, effectuées sur des cultures de tissus ou des cellules.
- Les études théoriques par modélisation en l'occurrence les méthodes QSAR.

I.1.7 Toxicité des produits chimiques organiques

Sur 100 000 produits chimiques libérés dans l'environnement, moins de 1 à 5% ont des données de toxicité disponibles. Même pour les produits chimiques de volume de production élevé ou HPVCs (High Production Volume Chemicals), (les substances chimiques produites en quantités >1000 tonnes par an dans l'UE ou > environ 442 tonnes par an aux Etats-Unis) il y'a un manque d'informations concernant leur toxicité, comme le montre le Tableau 1 [3].

Tableau 1. Données de toxicité pour les produits chimiques de haute production

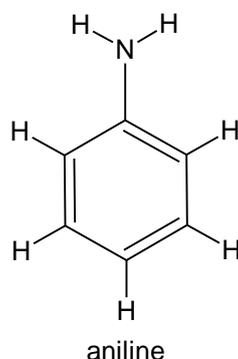
	Union européenne (UE)	Etats-Unis (USA)
Nombre	2465	2863
Données complètes de toxicité	3%	7%
Donnée spartielles de toxicité	43%	50%
Pas de données de toxicité	54%	43%

Avec une préoccupation croissante sur l'environnement et la santé humaine, les gouvernements et les organismes de réglementation à travers le monde cherchent à évaluer les risques éco-toxicologiques posés par la libération des substances chimiques. Ils ont proposé que 30000 produits chimiques existants soient testés sur les animaux pour une gamme des effets toxiques [4]. Ce serait évidemment une tâche très coûteuse, impliquant l'utilisation des milliers d'animaux, par conséquent le recours aux méthodes alternatives est devenu d'une grande importance.

I. Toxicité des anilines et phénols

I.2 Généralités sur les anilines :

L'aniline, connue également comme phénylamine ou aminobenzène, est un composé organique aromatique de formule chimique $C_6H_5NH_2$. C'est une amine primaire aromatique dérivée du benzène



I.2.1 Historique :

L'aniline a été isolée pour la première fois en 1826 par Otto Unverdorben par distillation de l'indigo. Celui-ci la baptisa cristalline. En 1834, Friedrich Runge parvint à isoler du goudron de houille une substance qui une fois traitée par du chlorure de chaux prend une couleur bleue. Il la baptisa kyanol ou cyanol. En 1841, C. J. Fritzsche obtint une substance huileuse en traitant de l'indigo avec de la potasse, qu'il baptisa aniline, d'après le nom d'une plante produisant de l'indigo, Indigo fera anil. Le mot anil est lui-même issu des termes sanskrits nīla, bleu profond, et nīlā plante d'indigo. A peu près en même temps, N.N. Zinin découvrit que la réduction du nitrobenzène permet d'obtenir une base qu'il baptisa benzidam. Par la suite, August Wilhelm von Hofmann étudia ces différentes substances et démontra en 1855 qu'elles sont identiques. La première utilisation d'aniline à l'échelle industrielle concerna la fabrication de la mauvéine, un colorant violet découvert en 1856 par William Henry Perkin.

I.2.2 Propriétés physico-chimiques

L'aniline, liquide à la température ambiante, est une substance huileuse incolore. A l'air, elle peut s'oxyder lentement et former une résine de couleur rouge-brune. L'aniline possède une odeur désagréable. Elle est aisément inflammable.

L'aniline est une base faible. En effet, les amines aromatiques sont généralement des bases nettement plus faibles que les amines aliphatiques. En effet, le doublet porté par l'atome d'azote est en partie délocalisé (mésomérie), ce qui n'est plus le cas sous la forme protonée

I. Toxicité des anilines et phénols

(forme acide) où le doublet est localisé sur la liaison N-H. La forme basique est donc plus stabilisée par mésomérie que la forme acide, d'où une constante d'acidité abaissée. Elle réagit avec les acides forts en formant des sels contenant l'ion anilinium ($C_6H_5-NH_3^+$). Elle réagit également avec les halogénures d'acyle (comme par exemple le chlorure d'éthanol CH_3COCl) en formant des amides. Les amides formées à partir de l'aniline sont parfois nommés anilides : $CH_3-CO-NH-C_6H_5$ est par exemple l'acétanilide.

L'aniline réagit avec les iodures d'alkyle en formant des amines secondaires ou tertiaires. L'oxydation de l'aniline a été très étudiée. En solution basique, elle réagit pour former de l'azobenzène. L'acide chromique permet de la transformer en quinone. Elle réagit avec les ions chlorates en présence de sels métalliques (notamment de vanadium) en formant du noir d'aniline. Elle réagit avec l'acide chlorhydrique et le chlorate de potassium en formant du chloranile. L'oxydation par le permanganate de potassium produit du nitrobenzène en milieu neutre, de l'azobenzène de l'ammoniaque et de l'acide oxalique en milieu basique, et du noir d'aniline en milieu acide. Elle réagit avec l'acide hypochloreux en formant du para-amino phénol et du para-amino diphénylamine.

Tout comme le benzène ou le phénol, l'aniline est réactive par substitution électrophile aromatique. Par exemple, elle peut subir une sulfonation pour former de l'acide sulfonique, qui peut être transformé en sulfonamides (médicaments très utilisés au début du XX^e siècle comme antiseptique).

L'aniline réagit avec l'acide nitreux en formant des sels de diazonium. Par leur intermédiaire, le groupement $-NH_2$ peut être transformé de manière simple en groupement $-OH$, $-CN$ ou halogénure.

I.2.3 Stabilité et réactivité

- Conditions à éviter : action de la lumière (décomposition) .
- Matières à éviter : oxydant (entre autres peroxydes, perchlorates, acide perchlorique, acide nitrique, oxygène), halogénures métalloïdes, métaux alcalins, métaux alcalino-terreux, acides, composés nitrés organiques, benzène/dérivés du benzène.
- Autres informations : Peut exploser avec l'air sous forme de vapeur/gaz.

Tableau 2. Propriétés physiques. [5]

Aspect : liquide	Couleur : incolore
Odeur : caractéristique	pH : 36 g/l eau (20°C) environ 8,8
Viscosité dynamique : (32 °C) 4,4 mPa*s	Masse moléculaire : 93,13
Température de fusion : -6 °C	Température d'ébullition : 184 °C
Température d'auto-inflammation : environ 530°C	Point d'éclair : 76°C
Limites d'explosivité dans l'air	
Inférieure : 1,2 Vol ; % Supérieure : 11 Vol%	
Pression de vapeur : (20 °C) 0,5 mbar	Masse volumique : (20 °C) 1,2 g/cm ³
Solubilité dans l'eau (20 °C) : 36 g/l	Solubilité dans l'éthanol (20°C) : facile
Decomposition thermique : environ 190°C	Coefficient de partage n-Octanol/eau : 0,91

I.2.4 Fabrication

L'aniline peut être fabriquée à partir du benzène en deux étapes. Au cours de la première étape, le benzène subit une nitration (substitution électrophile aromatique utilisant de l'acide nitrique) pour former du nitrobenzène. Au cours de la seconde étape, le nitrobenzène est réduit pour former l'aniline. Une grande variété de réactifs réducteurs peuvent être employés au cours de cette seconde étape, dont notamment du dihydrogène (en présence d'un catalyseur), du sulfure d'hydrogène, ou des métaux comme le fer, le zinc ou l'étain.

I.2.5 Utilisation

À l'origine, le grand intérêt commercial de l'aniline était la possibilité qu'elle créait de produire des teintures avec un bon rendement. La découverte de la mauvéine par William Henry Perkin en 1858 constitua ainsi le début de la découverte d'un grand nombre d'agents colorants qui se comptèrent bientôt par centaines. En sus des teintures, l'aniline était le produit de départ dans la synthèse d'un grand nombre de médicaments.

À l'heure actuelle, l'utilisation la plus importante de l'aniline concerne la fabrication du 4,4'-MDI (4,4'-Méthylènebis phényle isocyanate), qui utilise environ 85% de l'aniline produite. Parmi les autres utilisations, on peut citer la fabrication chimique de caoutchouc (9%), d'herbicides (2%) et de pigments ou agents colorants (2%).

I.2.6 Dangers

Nocif par inhalation, par contact avec la peau et par ingestion. Possibilité d'effets irréversibles.

Toxique : risque d'effets graves pour la santé en cas d'exposition prolongée par inhalation, par contact avec la peau et par ingestion.

I.2.7 Informations toxicologiques

Toxicité aiguë : DL50 (voie orale, rat) = 250 mg/kg.

Toxicité chronique/long-terme : Le soupçon d'effet cancérigène demande un supplément de recherche.

Autres informations toxicologiques:

- En cas d'inhalation de vapeurs : irritation des muqueuses.
- En cas de contact avec la substance : irritation sur : yeux, peau, muqueuse. Effet possible en cas de contact avec la substance : dermatite.
- En cas d'absorption de grandes quantités : méthémoglobinémie avec céphalées, nausées, troubles du rythme cardiaque, dyspnée.

I.3 Généralités sur les phénols

Les phénols sont des produits caustiques qui coagulent les albumines. Leur action est très irritante pour les muqueuses et la peau. Ils sont toxiques, on doit les manipuler avec précaution, mais dans le cas du phénol en solution diluée, il joue le rôle d'antiseptique en médecine vétérinaire.

I.3.1 Les phénols dans la nature :

Un grand nombre de phénols et d'éthers phénoliques sont produits dans la nature. L'acide « o-hydroxybenzoïque », connu sous le nom d'acide « salicylique », peut être tiré du sol (SALIXGENUS), tandis que son ester méthylique est le principe odoriférant de l'essence de « WINTERGREEN » et, entre dans la composition de nombreux aliments.

L'acide « acétylsalicylique » ou « aspirine » utilisé comme analgésique, n'est pas un produit naturel.

I. Toxicité des anilines et phénols

Le thymol qui donne son odeur au thym est utilisé comme antiseptique dans les dentifrices, alors que l'eugénol, qui donne son parfum aux clous de girofles, est l'antiseptique préféré des dentistes. Très proche de l'« eugénol », le safrole est le constituant principal de l'huile de sassafras et possède l'odeur caractéristique de la « rootbeer ». Les structures de quelques phénols sont représentées dans la figure 2.

Certains phénols ont des effets physiologiques très puissants, ainsi le principe irritant du « poison ivy », contient le système 1,2-dihydroxybenzène (catéchol), substitué en 3 par une longue chaîne alkylée d'insaturation variable.[6]

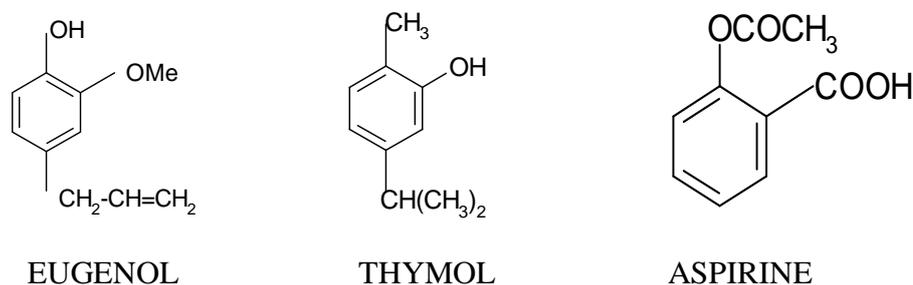


Figure 2 : Structures de quelques phénols naturels

I.3.2 Utilisation :

Le benzène et ses dérivés tels que les phénols sont utilisés dans l'industrie du caoutchouc, des peintures, vernis, matières plastiques et en métallurgie. Il reste encore un solvant couramment employé non seulement en industrie, mais aussi dans les laboratoires de recherches et d'analyse. Il est d'utilisation très large dans les pays du tiers monde (dégraissage à sec, « pressing » locaux par exemple).

I.3.3 Intérêt industriel :

Les phénols ont un intérêt industriel en tant que matière première très importante de l'industrie organique telle que : fabrication des matières plastiques, parfumerie, matières colorantes, produits pharmaceutiques et explosifs, développement photo.....

Pendant longtemps, la presque totalité du benzène était extraite des goudrons de houille, obtenue par distillation de ces goudrons entre 150⁰C et 210⁰C par la soude, et les phénols donnent des phénates solubles qui se séparent du reste de l'huile.

I.3.4 Dangers toxiques des phénols :

I. Toxicité des anilines et phénols

Les dangers toxiques des phénols [7] se manifestent lors de l'utilisation et selon le type de l'exposition (chronique ou aiguë).

I.3.5 Tableau clinique de l'intoxication :

a. Toxicité aiguë (accidentelle) :

- **Inhalation** : vomissement, somnolence pouvant aller jusqu'au coma, convulsion si forte exposition.
- **Ingestion** : troubles digestifs, troubles neurologiques, pneumopathie peut être noté.

b. Effets hématopoïtiques:

Anémie, polyglobulie. Ces anomalies régressent à l'arrêt de l'exposition. Mais l'intoxication se prolonge lors d'une réexposition.

c. Effet leucémique :

Le pouvoir leucémogène se manifesterait pour des expositions supérieures à 10 ppm. Le benzène et ses dérivés sont classés cancérigènes catégorie 1 selon la classification de l'IARC (International Agency for Research on Cancer).

L'utilisation des outils informatiques chez les chimistes est devenue obligatoire afin de bien manipuler les informations moléculaires qui ont été, au cours des dernières années, stockées numériquement sur les ordinateurs dans des bases de données en très grandes quantités. De plus, la multiplication des données exploitables par les chimistes a donné lieu à une obligation de la numérisation, afin d'être capable de stocker, visualiser et traiter ces mêmes données aisément.

La discipline décrivant l'utilisation des outils informatiques pour traiter et résoudre des problèmes à la fois dans le domaine chimique et biologique est désignée par la Chémoinformatique [8]. Ses utilisations sont très variées et vont de la création et l'utilisation de base de données de petites molécules à la manipulation de fichiers en passant par les études statistiques. Cependant, son application la plus communément admise est dans le domaine de la recherche de nouveaux médicaments (Drug eDiscovery), domaine dans lequel elle joue un rôle central dans l'analyse et l'interprétation des données de structures et de propriétés collectées au cours des criblages à haut débit (technique se fait par les biologistes et visant à identifier des molécules nouvelles et potentiellement actives dans des bases de données de composés).

L'émergence de cette discipline peut être mise en parallèle avec la multiplication des données chimiques stockées numériquement. En effet, les quantités de données générées par les nouvelles approches de Drug design n'ont eu de cesse d'augmenter et il s'est avéré nécessaire, pour traiter les résultats de criblage à haut débit ou encore de la chimie combinatoire, de développer et d'utiliser des techniques informatiques [9].

Les avancées technologiques de la dernière décennie ont rendu possibles de nombreuses découvertes et applications inaccessibles auparavant. Par exemple, le nombre de composés disponibles dans les études de criblage a augmenté de manière exponentielle. En parallèle, les développements techniques dans le domaine de l'informatique et des technologies de communication ont permis la création de bases de données de composés comportant des millions d'entrées. L'exemple le plus pertinente pour illustrer ces avancées est la base de données «PubChem» développée par le NIH (National Institute of Health) [10]. Avec un contenu de plus de 31 millions de composés reliés à leurs activités biologiques, ce genre de bases de données nécessite le développement et l'utilisation d'outils mathématiques et statistiques afin de pouvoir accéder à de nouvelles découvertes en termes de développement

de nouveaux médicaments et à la compréhension «des relations entre structure et activité (QSAR)» ou bien «des relations entre structure et propriétés (QSPR)».

Les premiers essais de modélisation d'activités de molécules datent de la fin du 19^{ème} siècle, lorsque Crum-Brown et Frazer [11] postulèrent que l'activité biologique d'une molécule est une fonction de sa constitution chimique. Mais ce n'est qu'en 1964 que furent développés les modèles de «contribution de groupes», qui constituent les réels débuts de la modélisation QSAR. Depuis, l'essor de nouvelles techniques de modélisation par apprentissage, linéaires d'abord, puis non linéaires, ont permis la mise en place de nombreuses méthodes ; elles reposent pour la plupart sur «la recherche d'une relation entre un ensemble de nombres réels, descripteurs de la molécule, et la propriété ou l'activité que l'on souhaite prédire».

II.1 Définition des QSAR/QSPR

Lors d'une étude de «QSAR (Qualitative Structure-Activity Relationships)» ou de «QSPR (Qualitative Structure-Property Relationships)», on étudie les relations entre la structure et l'activité (propriété) d'un composé ou molécule, par exemple les effets d'une variation chimique locale sur une molécule à l'activité connue. En effet, certains changements chimiques sur certaines parties d'une molécule peuvent entraîner des variations de son activité biologique en agissant sur l'interaction avec la cible [12].

Ainsi, la méthodologie QSAR/QSPR permet de trouver un modèle mathématique qui met en corrélation l'activité (propriété) et la structure au sein d'une famille de composés. De nombreuses méthodes conceptuellement différentes peuvent être utilisées pour mettre en place les modèles mathématiques permettant de détecter des relations de type QSAR. Ainsi, les études QSAR/QSPR sont basées sur des méthodes informatiques, celles de modélisations, déjà exploitées dans différents domaines.

Les grandes phases de la mise en place d'un modèle QSAR/QSPR peuvent être décrites comme suit : Extraire les descripteurs à partir de la structure moléculaire, choisir les descripteurs adaptés à l'étude par rapport à l'activité (propriété) analysée, utiliser les descripteurs comme variables explicatives pour définir une relation qui les corrèle à l'activité en question, et enfin chaque modèle doit être validé sur des jeux de données de test [13-14].

II.2 Les descripteurs moléculaires

De nombreuses recherches ont été menées, au cours des dernières décennies, pour trouver la

meilleure façon de représenter l'information contenue dans la structure des molécules, et ces structures elles-mêmes, en un ensemble de nombres réels appelés descripteurs ; une fois que ces nombres sont disponibles, il est possible d'établir une relation entre ceux-ci et une propriété ou activité moléculaire, à l'aide d'outils de modélisation classiques.

II.2.1 Définition

Les descripteurs numériques réalisent un codage de l'information chimique en un vecteur de réels. Tout simplement, un descripteur moléculaire est une représentation mathématique d'une molécule, qui contient à la fois des informations sur la structure, et donc, implicitement ou explicitement, sur ses propriétés physico-chimiques. Ces informations peuvent être encodées par des valeurs scalaires, des vecteurs ou des chaînes de bits [13 - 15].

II.2.2 Types de descripteurs moléculaires [8]

On dénombre aujourd'hui plus de 10000 descripteurs moléculaires, qui quantifient des caractéristiques physico-chimiques ou structurelles de molécules. Ils peuvent être obtenus de manière empirique ou non-empirique, mais les descripteurs calculés, et non mesurés, sont à privilégier : ils permettent en effet d'effectuer des prédictions sans avoir à synthétiser les molécules, ce qui est un des objectifs de la modélisation. Il existe cependant quelques descripteurs mesurés : il s'agit généralement de données expérimentales plus faciles à mesurer que la propriété ou l'activité à prédire (coefficient de partage eau-octanol [16], polarisabilité, ou potentiel d'ionisation). Les descripteurs moléculaires sont fréquemment classés par rapport à la dimensionnalité de la représentation moléculaire sur laquelle ils sont calculés : On parlera alors de descripteurs 1D, 2D ou 3D [12].

a) Les descripteurs 1D :

Sont appelés «descripteurs constitutionnels» et sont faciles et rapides à calculer. Ils sont accessibles à partir de la formule brute de la molécule (par exemple C_6H_6O pour le phénol), et décrivent des propriétés globales du composé. Il s'agit par exemple de sa composition, c'est-à-dire les atomes qui le constituent, ou de sa masse molaire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères [13-14].

b) Les descripteurs 2D :

Les descripteurs moléculaires utilisent la représentation des molécules comme des

graphes sont dits «descripteurs 2D» et contiennent des informations à propos de la connectivité ou à propos de certains fragments moléculaires, mais aussi des estimations des propriétés physico-chimiques. C'est à partir de ce niveau que l'on peut espérer la capture d'informations chimiques pertinentes pour la prédiction de la majorité des propriétés moléculaires. On trouvera dans cette catégorie les descripteurs suivants :

Les indices topologiques, qui considère la structure du composé comme un graphe, les atomes étant les sommets et les liaisons sont les arêtes. De nombreux indices quantifiant la connectivité moléculaire ont été développés en se basant sur cette approche, comme par exemple l'indice de Wiener [17], qui compte le nombre total de liaisons dans les chemins les plus courts entre toutes les paires d'atomes (en excluant les hydrogènes), et qui sera également l'axe central de cette thèse.

Les indices constitutionnels, qui se basent sur des motifs sous-structuraux. Par exemple, les empreintes BCI (Barnard Chemical Information Ltd) [18] sont des ensembles de bits indiquant la présence ou l'absence de certains fragments dans une molécule. Les fragments prennent en compte les atomes et leurs plus proches voisins, les paires d'atomes et les séquences ou encore les fragments basés sur des cycles. L'approche des clés MDL est une approche similaire comprenant la recherche des 166 fragments MDL [13, 14, 19].

Ces descripteurs 2D reflètent bien les propriétés physiques dans la plupart des cas, mais sont insuffisants pour expliquer de façon satisfaisante certaines propriétés ou activités, telles que les activités biologiques.

c) Les descripteurs 3D :

Les descripteurs 3D d'une molécule sont évalués à partir des positions relatives de ses 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 s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent de la structure 3D. On distingue plusieurs familles importantes de descripteurs 3D :

Les descripteurs géométriques : parmi ceux qui sont les plus importants sont le volume moléculaire, la surface accessible au solvant et le moment principal d'inertie.

Les descripteurs électroniques : ils permettent de quantifier différents types d'interactions inter et intramoléculaires, de grande influence sur l'activité biologique de molécules. Le

calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie stérique est minimale, et fait souvent appel à la chimie quantique. Par exemple, les énergies de la plus haute orbitale moléculaire occupée et de la plus basse vacante sont des descripteurs fréquemment sélectionnés. Le moment dipolaire, le potentiel d'ionisation, et différentes énergies relatives à la molécule sont d'autres paramètres importants.

Les descripteurs spectroscopiques : les molécules peuvent être caractérisées par des mesures spectroscopiques, par exemples par leurs fonctions d'onde vibrationnelles. En effet, les vibrations d'une molécule dépendent de la masse des atomes et des forces d'interaction entre ceux-ci ; ces vibrations fournissent donc des informations sur la structure de la molécule et sur sa conformation. Les spectres infrarouges peuvent être obtenus soit de manière expérimentale, soit par calcul théorique, après recherche de la géométrie optimale de la molécule. Ces spectres sont alors codés en vecteurs de descripteurs de taille fixe. Les descripteurs de type MORSE [20] (Molecule Representation of Structures based on Electron diffraction) font appel au calcul des intensités théoriques de diffraction d'électrons.

II.3 Modèles QSAR de la toxicité (QSTR) :

Les QSAR pour la toxicité remonte au 19^{ème} siècle. En 1863, A.F.A. Crois à l'Université de Strasbourg a observé que la toxicité des alcools à des mammifères augmente lorsque la solubilité des alcools dans l'eau diminue. Dans les années 1890, Hans Horst Meyer de l'Université de Marburg et Charles Ernest Overton de l'Université de Zurich, ont noté que la toxicité des composés organiques était tributaire de la lipophilie [21]. Au début des années 1960 Corwin Hansch a proposé un modèle mathématique pour corréliser l'activité biologique et la structure chimique [22], cette date est considérée comme étant la naissance des méthodes QSAR. Depuis, l'utilisation des QSAR en toxicologie n'a pas cessé d'évoluer.

Des modèles QSAR sont maintenant mis au point en utilisant une variété d'approches, de méthodes d'analyse de données et de paramètres [23].

Un grand nombre d'études QSAR de toxicité et, en particulier la toxicité aiguë ont été publiées dans la littérature. La plupart des données de toxicité pour l'environnement ont été obtenues en utilisant des espèces aquatiques (représentées dans la figure 3) en l'occurrence les poissons, les daphnies, les protozoaires « *Tetrahymena Pyriformis* », *Vibrio fischeri*, les algues ...

Cronin et Dearden [24- 25] ont examiné la littérature concernant la modélisation QSAR de la toxicité aquatique. Plusieurs modes d'action ont été identifiés chez les espèces aquatiques, à savoir narcose non polaire, narcose polaire, découplage de la phosphorylation

II. Modélisation QSAR/QSPR

oxydative, irritation de la membrane respiratoire, inhibition d'acétylcholinestérase, saisie de système nerveux central, inhibition de photosynthèse, et l'alkylation. Cependant, ceux-ci sont généralement plus largement regroupés comme : narcose non polaire, narcose polaire, réactifs sélectifs (électrophiles), et les molécules à mécanismes d'action spécifiques.

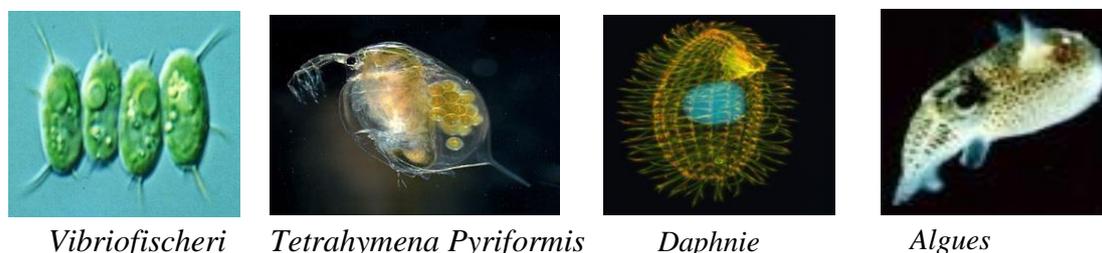


Figure 3 : Différentes espèces aquatiques utilisées pour l'étude de la toxicité aquatique [1].

Verhaar *et al.* [26] ont développé un système sur la base de la présence de groupes fonctionnels pour classer les produits chimiques dans ces quatre groupes **Tableau 3** [1].

Tableau 3. Mécanismes possibles pour la toxicité [26]

Mécanisme (Mode d'action)	Structures déterminantes
Narcose non-polaire	Par exemple les alcanes saturés avec halogène et /ou substituants alcoxy (alcools aliphatiques, les cétones, les éthers, des amines)...
Narcose polaire	Les phénols, phénols et anilines avec trois ou moins d'atomes d'halogène, et/ou substituants alkyle...
Formation des radicaux libres	Phénols et anilines avec quatre ou plusieurs atomes d'halogène, ou plus d'un groupe nitro, ou seul un groupe nitro, et plus d'un groupe d'halogène...
Electrophiles /pro-électrophiles	Certains nitrobenzènes ; des noyaux benzéniques sans aniline ou de phénol qui ont deux groupes nitro sur un noyau ; phénols avec un seul groupe nitro et un halogène ; composés aromatiques ayant deux ou plusieurs groupes hydroxy dans la position ortho ou para, quinines ; aldéhydes ; composés aromatiques avec deshalogènes; cétènes; époxydes...

La dépendance de la toxicité des narcoses avec le coefficient de partage, en particulier le coefficient de partage octanol-eau, a été montrée par de nombreux auteurs [27-35]. Le QSAR représentant une toxicité de référence a été dérivé pour un groupe de narcotiques non polaires (alcools saturés, cétones, nitriles, esters et des composés

contenant du soufre). Ils ont conclu que le coefficient de partage octanol-eau est suffisant pour expliquer la toxicité des narcoses apolaires tandis que pour les narcoses polaires il faut la présence d'un descripteur supplémentaire qui explique le caractère électronique des molécules.

En outre, des modèles de combinaison des deux groupes de composés (narcoses apolaires et polaires) et les deux types de descripteurs ont été développés. Freidig et Hermens [28] ont conclu qu'en utilisant des modèles QSAR distincts pour les composés agissant par des mécanismes différents, y compris un descripteur qui caractérise le mécanisme de toxicité particulière, donne de meilleurs résultats que l'utilisation d'un modèle unique qui combine tous les composés avec les mêmes descripteurs.

Certains auteurs ont utilisé l'approche surface de réponse sur la base de l'hydrophobie et l'électrophilie des composés. Dans cette approche, les QSAR comprennent un descripteur qui caractérise la bio-absorption et la distribution (généralement partage octanol-eau ou coefficients de distribution ($\log P$ ou $\log D$) et un descripteur de réactivité électrophile (habituellement LUMO ou la super delocalisabilité maximale (A_{\max})). Cette approche a été appliquée à des différentes espèces aquatiques, y compris la bactérie *Vibriofischeri* [36], les protozoaires *Tetrahymena pyriformis* [37-39], les algues de *Scenedesmus* [40] et *Chlorellavulgaris* [41]. L'avantage de l'approche surface de réponse est qu'elle est simple et a une interprétation mécanistique. Alors que certains auteurs (par exemple, Cronin et Schultz [37], Cronin *et al.* [36], Cronin *et al.* [39]) ont utilisé LUMO comme descripteur de réactivité électrophile entraînant des interactions covalentes dans les systèmes biologiques. Dimitrov *et al.* [42-43] ont suggéré que LUMO peut être également utilisé pour décrire l'interaction électrophile non covalente des produits chimiques narcotiques avec le site d'action. Certains auteurs ont prolongé la démarche réponse-surface en ajoutant un indicateur supplémentaire et d'autres paramètres afin d'améliorer l'ajustement statistique qui est difficile des modèles [44-47]. Toutefois, selon Schultz *et al.* [48] la modélisation QSAR des composés électrophiles en raison de la limitation des données et des descripteurs par rapport à la modélisation QSAR des composés agissant par d'autres mécanismes toxiques. xxx. Cependant, des QSAR basés sur des indices topologiques pour l'étude de la toxicité ont fait l'objet de nombreux travaux [32, 49-52].

III-1 Généralités

Les techniques de calcul qui peuvent fournir la valeur de l'énergie d'une géométrie, aussi particulière que l'état fondamental, appartiennent à plusieurs catégories [53]:

- méthodes *ab initio*,
- méthodes semi-empiriques,
- méthodes empiriques,
- mécanique moléculaire.

Concernant les deux premières méthodes, elles sont fondées sur l'évaluation des interactions électroniques complètes ou partiellement négligées. Le terme *ab initio* est réservé aux calculs déduits directement des principes théoriques, sans faire intervenir de données expérimentales. Deux méthodes fondamentales sont proposées pour la résolution de l'équation de Schrödinger à partir des principes de base. La théorie des orbitales moléculaires (OM) tend à établir une expression pour la fonction d'onde ψ , alors que dans la théorie de la fonctionnelle de la densité (DFT), la distribution de la densité électronique (ρ) joue ce rôle. Le fondement de la DFT est associé à un théorème dû à Hohenberg et Kohn [54] qui ont démontré que toutes les propriétés d'un système dans un état fondamental non dégénéré sont complètement déterminées par sa densité électronique.

Le type le plus courant de calcul *ab initio*, ou calcul Hartree-Fock (HF), repose sur l'approximation principale du champ central. Le calcul variationnel mis en œuvre conduit à des énergies supérieures aux énergies réelles (Théorème de Eckart) et tendent vers une valeur limite appelée limite de Hartree Fock. La seconde approximation dans les calculs HF consiste à décrire la fonction d'onde par une « fonction utile » qui est connue exactement pour quelques systèmes mono-électroniques. Les fonctions les plus souvent utilisées sont des combinaisons linéaires d'orbitales de type Slater ($e^{-a x}$) ou d'orbitales gaussiennes ($e^{-a x^2}$), dont les abréviations sont, respectivement, STO (pour Slater Type Orbitals) et GTO (pour Gaussian Type Orbitals). La fonction d'onde est obtenue à partir de combinaisons linéaires d'orbitales, ou plus souvent à partir de combinaisons de fonctions d'un ensemble de base. A cause de cette approximation, la plupart des calculs HF conduisent à des énergies supérieures à la limite HF.

L'utilisation de bases de fonctions gaussiennes permet de calculer toutes les intégrales de la méthode sans autres approximations que celles inhérentes à la méthode elle-même.

Réservées initialement au traitement de petites molécules (une dizaine d'atomes), les méthodes *ab initio* ont été étendues, ces dernières décennies, à des systèmes de quelques centaines d'atomes, comme conséquence de l'augmentation de la puissance des ordinateurs (hardware et software).

Une approximation sur l'hamiltonien est considérée comme une méthode semi-empirique.

Les méthodes semi-empiriques sont moins contraignantes en moyens de calculs. De plus, l'incorporation de paramètres déduits des données expérimentales dans certaines de ces méthodes permet de prédire quelques propriétés avec une meilleure précision que celle obtenue avec les méthodes *ab initio* les plus élaborées.

Les méthodes de champ de force ne demandent pas de temps excessifs de calcul pour donner des informations sur l'énergie de la molécule étudiée. La mécanique moléculaire (M M), appelée parfois : calcul par champ de force empirique, (empirical Force Field, EFF, en anglais), permet le calcul de la structure et de l'énergie d'entités moléculaires [55-57]. D'une part, les distributions électroniques ne sont pas explicitement détaillées (à quelques exceptions près), d'autre part, la recherche de l'énergie minimale par optimisation de la géométrie joue un rôle primordial.

L'énergie de la molécule est exprimée sous la forme d'une somme de contributions associées aux écarts de la structure par rapport à des paramètres structuraux de référence. Les variables de calcul sont alors les coordonnées internes du système : longueur de liaison, angles de valence, angles dièdres et distances entre les atomes non liés. Un calcul de MM aboutit à une disposition des noyaux telle que la somme de toutes les contributions énergétiques est minimisée ; ses résultats concernent surtout la géométrie et l'énergie de système [58].

III-2 Méthodes semi-empiriques utilisées

Les méthodes AM1 et PM3 utilisées étant des re-paramétrisations de la méthode MNDO, nous présenterons ces trois méthodes, en rappelant au préalable le cadre des équations (*ab initio*) HFR (Hartree-Fock-Roothaan) sur lequel elles sont basées et les approximations supplémentaires auxquelles il est fait recours.

III-2.1 Le cadre Hartree - Fock – Roothaan

Les méthodes *ab initio* utilisent l'équation de Schrödinger électronique obtenue après séparation des mouvements électroniques et nucléaires (approximation de Born-Oppenheimer) [59, 60].

Dans la méthode Hartree – Fock la fonction d'onde ψ d'un système à N électrons est représentée par un déterminant de Slater ψ_0 de spin orbitales ϕ unique. Les spin orbitales consistent en des produits d'orbitales moléculaires (OM) ϕ et de fonctions de spin (α ou β), $\phi_a = \phi_a \alpha$, $\bar{\phi}_a = \phi_a \beta$.

On représentera ψ_0 par :

$$\Psi_0 = |\Phi_1 \bar{\Phi}_1 \Phi_2 \bar{\Phi}_2 \dots \Phi_M \bar{\Phi}_M\rangle \quad (1)$$

pour un système à couches complètes comportant N électrons (auquel cas $M = \frac{N}{2}$).

Chaque OM est développée sous forme d'une combinaison linéaire de fonctions de base, appelées conventionnellement orbitales atomiques (OM-CLOA, combinaison linéaire d'orbitales atomiques), quoiqu'elles ne soient pas généralement, solutions du problème HF atomique.

$$\phi_a = \sum_{\mu}^m c_{\mu a} \chi_{\mu} \quad (2)$$

En tenant compte de (1), on obtient après multiplication à gauche par une fonction spécifique, intégration et application du principe variationnel, un système d'équations linéaires, ou équations de Roothaan – Hall (pour un système à couches complètes) [61,62].

Signalons que la résolution des équations de Roothaan – Hall fournit un total de m (= nombre de fonctions de base) orbitales moléculaires (OM) dont n sont occupées et (m - n) libres ou virtuelles. Celles-ci sont orthogonales à toutes les orbitales occupées, mais n'ont pas d'interprétation physique directe exceptée comme affinité électronique (via le théorème de Koopmans [63]). Elles servent dans la description des états excités.

L'équation (3) condense, sous forme matricielle, les équations de Roothaan – Hall.

$$\mathbf{F} \mathbf{C} = \mathbf{S} \mathbf{C} \boldsymbol{\varepsilon} \quad (1)$$

où:

- la matrice \mathbf{F} de Fock est l'opérateur hamiltonien effectif,
- \mathbf{C} est la matrice des coefficients des OM, $c_{\mu a}$,
- \mathbf{S} est la matrice de recouvrement,
- et $\boldsymbol{\varepsilon}$ une matrice diagonale comportant les énergies orbitales.

La matrice de Fock, \mathbf{F} , comporte toutes les informations relatives au système quantomécanique, c'est – à – dire toutes les interactions prises en compte dans les calculs. Sa formulation *ab initio* est la suivante :

$$\mathbf{F}_{\mu\nu} = \mathbf{H}_{\mu\nu} + \mathbf{J}_{\mu\nu} - \frac{1}{2} \mathbf{K}_{\mu\nu}$$

$$\mathbf{F}_{\mu\nu} = \mathbf{H}_{\mu\nu} + \sum_{\rho}^n \sum_{\sigma}^m \mathbf{P}_{\rho\sigma} \left[\langle \mu\nu/\rho\sigma \rangle - \frac{1}{2} \langle \mu\sigma/\rho\nu \rangle \right] \quad (2)$$

Avec :

$$\mathbf{H}_{\mu\nu} = \int \chi_{\mu}^*(1) \hat{h} \chi_{\nu}(1) d\tau_1 \quad (3)$$

$$\langle \mu\nu/\rho\sigma \rangle = \iint \chi_{\mu}^*(1) \chi_{\nu}(1) \frac{1}{r_{12}} \chi_{\rho}^*(2) \chi_{\sigma}(2) d\tau_1 d\tau_2 \quad (4)$$

III. Bases de modélisation moléculaires

$$\text{et } P_{\rho\nu} = 2 \sum_a^m C_{\rho a}^* C_{\nu a} \quad (7)$$

où μ, ν, ρ et σ désignent des orbitales atomiques, et $H_{\mu\nu}$ des intégrales mono-électroniques représentant les valeurs moyennes de l'opérateur associé à l'énergie cinétique et l'opérateur énergie potentielle d'interaction noyau – électron (\hat{V}_{en}). Les $\langle \mu\nu/\rho\sigma \rangle$ sont des intégrales de répulsion bioélectroniques représentant \hat{V}_{ee} (opérateur d'interaction entre les électrons eux - mêmes), et les $P_{\rho\nu}$ sont les éléments de la matrice densité \mathbf{P} .

$\mathbf{J}_{\mu\nu}$ et $\mathbf{K}_{\mu\nu}$ sont les représentations matricielles des opérateurs coulombien \hat{J} et d'échange \hat{K} respectivement.

L'énergie électronique (E_{el}) peut être exprimée au moyen des valeurs propres ϵ_a :

$$E_{el} = 2 \sum_a^m \epsilon_a - \frac{1}{2} \sum_{\mu\nu}^m P_{\mu\nu} \left(\mathbf{J}_{\mu\nu} - \frac{1}{2} \mathbf{K}_{\mu\nu} \right) \quad (5)$$

Comme la matrice de Fock dépend des coefficients des orbitales, les équations de Roothaan doivent être résolues de façon itérative en utilisant la procédure du champ auto-cohérent ou SCF (pour : Self Consistent Field) [64].

Une étape importante de la procédure SCF est la conversion de l'équation générale aux valeurs propres (3) en une équation ordinaire par une transformation orthogonale (méthode d'orthogonalisation de Löwdin) [65,66].

$$\mathbf{F}^\lambda \mathbf{C}^\lambda = \mathbf{S}^{-1/2} \mathbf{F}$$

$$\text{Avec } \mathbf{F}^\lambda = \mathbf{S}^{-1/2} \mathbf{F} \mathbf{S}^{-1/2} \quad (9)$$

$$\text{et : } \mathbf{C}^\lambda = \mathbf{S}^{1/2} \mathbf{C}$$

Notons que $\mathbf{S}^{-1/2}$ qui est obtenue à partir de la matrice de recouvrement \mathbf{S} qui n'est jamais singulière, n'est jamais singulière non plus.

Signalons que les approximations électroniques et CLOA (utilisation d'un nombre limité d'orbitales atomiques) et un problème de corrélation limitent la méthode HFR. On dépasse ces limitations par l'utilisation de fonctions corrélées ou en faisant intervenir l'interaction de configuration.

III.2.2 Analyse de population de Mulliken [67, 68].

L'analyse de population répond à la question de savoir comment, lors de la formation d'une molécule à N électrons, répartir équitablement ces électrons entre les atomes, de telle sorte que la notion d'atome dans la molécule ne disparaisse pas complètement, sans enfreindre toutefois les principes de la mécanique ondulatoire et en tenant compte de la géométrie particulière de la molécule.

Soit $d\tau (= d\tau_1 d\tau_2 \dots d\tau_i)$ un élément de volume, ψ la fonction d'onde déterminantale, la probabilité :

$$dp = \psi^* \psi d\tau \tag{6}$$

représente, à un facteur près, la quantité d'électricité dans $d\tau$ à un moment donné, puisque :

$$dQ = edP \tag{7}$$

En prenant pour unité de charge, la charge de l'électron on n'introduit plus e avec dP.

Ainsi, si on intègre dP sur l'espace de configuration E, il vient :

$$\int_E dp = N \quad (\text{N électrons}) \tag{8}$$

Le résultat serait une somme sur les orbitales occupées (les ψ du déterminant de Slater), c'est -à- dire :

$$\int_E \psi^* \psi d\tau = N = \sum_{i=1}^m 2 \int \psi^*(\nu) \psi_i(\nu) d\tau_\nu \tag{9}$$

Les orbitales étant normées : $\langle \psi_i(\nu) / \psi_i(\nu) \rangle = 1$

En décomposant sur la base atomique, on obtient :

$$\begin{aligned} \sum_{i=1}^m 2 \int \left(\sum_{l=1}^n c_{li}^* \varphi_l(\nu) \right) \left(\sum_{m=1}^n c_{mi} \varphi_m(\nu) \right) d\tau_\nu &= 2 \sum_{i=1}^m \sum_{l=1}^n \sum_{m=1}^n c_{li}^* c_{mi} \int \varphi_l(\nu) \varphi_m(\nu) d\tau_\nu \\ &= \sum_{l=1}^n \sum_{m=1}^n p_{lm} s_{lm} = \sum_{l=1}^n \left[\sum_{m=1}^n p_{lm} \right] = N \end{aligned} \tag{10}$$

Avec :

$$s_{lm} = \int \varphi_l^*(\nu) \varphi_m(\nu) d\tau_\nu$$

et :

$$P_{Lm} = \sum_{i=1}^m 2 C_{li}^* C_m$$

III. Bases de modélisation moléculaires

En posant : $q_l = \sum_{m=1}^n P_{lm} S_{lm}$, on se donne un moyen de répartir le nombre d'électrons de la molécule : q_l est la quantité d'électricité qui peut être attribuée à la 1^{ère} orbitale atomique de base.

Remarque : la relation $\int_E \psi^* \psi d\tau = N$ met en exergue les principes de la mécanique ondulatoire, alors que la géométrie particulière de la molécule ressort dans p_{lm} et S_{lm} . La quantité d'électricité attribuée à l'atome L est la somme des $q_{l(L)}$ ($l \in L$) :

$$Q_L = \sum_{l(L)} q_{l(L)} \quad (11)$$

La relation d'identité initiale $\sum_l \sum_m p_{lm} S_{lm} = N$ peut être écrite sous la forme :

$$\sum_l q_l = N \quad (12)$$

$$\sum_L Q_L = N$$

q_l est la densité électronique de l'orbitale, et Q_L celle de l'atome L.

La charge, C_L , de l'atome L dans la molécule est :

$$C_L = Z_L - Q_L \quad (13)$$

Z_L étant le nombre d'électrons de l'atome isolé, et Q_L la quantité d'électricité qu'il possède dans l'atome.

III.2.3 Les méthodes semi-empiriques.

Dans les méthodes semi-empiriques on simplifie l'approche Hartree-Fock - Roothaan.

1) Dans la construction de Ψ_0 : seuls les électrons de valence sont traités de façon explicite en utilisant un ensemble de base minimal. Ce qui signifie que les atomes H sont décrits par une fonction 1s, les éléments Li à F par un ensemble {2s, 2p}, les éléments Na à Cl par un ensemble {3s, 3p}, Ca, K, et Zn à Br avec un ensemble {4s, 4p}, Sc – Cu avec un ensemble de base {4s, 4p, 3d} ; etc...

On tient compte des électrons de cœur soit en corrigeant la charge nucléaire, soit en introduisant des fonctions pour modéliser les répulsions simultanées entre noyaux d'une part et entre électrons de cœur d'autre part.

2) Dans la construction de F^λ on néglige une grande part des interactions, en particulier dans la partie bi-électronique $\langle \mu\nu/\rho \sigma \rangle$. Toutes les intégrales mettant en jeu des orbitales atomiques centrées sur plus de 2 noyaux sont négligées. Certaines classes d'intégrales sont remplacées par des paramètres. C'est le cas, en particulier, des intégrales mono-électroniques bi-centres $H_{\mu\nu}$ qui sont, pour une large part, responsables de la liaison chimique.

La façon d'introduire ces simplifications dans le modèle permet de distinguer entre les différentes méthodes.

Une autre façon de réduire les intégrales bi-électroniques est l'approximation du recouvrement différentiel nul (RDN) dans laquelle on néglige tous les produits des fonctions de base dépendant des coordonnées d'un même électron localisé sur des atomes différents. Cela signifie que tous les produits des fonctions orbitales atomiques $\chi_\mu \chi_\nu$ sont posés égaux à zéro et l'intégrale de recouvrement se réduit à $S_{\mu\nu} = \delta_{\mu\nu}$ ($\delta_{\mu\nu}$ est le symbole de Kronecker ; $\delta_{\mu\nu} = 0$ si $\mu \neq \nu$ et $\delta_{\mu\nu} = 1$ si $\mu = \nu$).

Dans l'approximation RDN, toutes les intégrales tri et tétra-centres s'annulent ce qui transforme la matrice de recouvrement en une matrice unité. Les intégrales mono-électroniques tri-centres sont égalées à zéro. Toutes les intégrales bi-électroniques tri et tétra-centres sont négligées.

Les paramètres sont imposés pour compenser les approximations. Ainsi toutes les intégrales restantes sont remplacées par des paramètres convenables ajustés sur des grandeurs fournies par l'expérience.

Toutes les méthodes semi-empiriques modernes sont basées sur l'approche MNDO (Modified Neglect of Differential Overlap) [69], dans laquelle des paramètres sont assignés aux différents types d'atomes puis ajustés de telle sorte à reproduire certaines propriétés comme les chaleurs de formation, les variables géométriques, les moments dipolaires et les énergies de première ionisation.

Les paramètres sont conçus séparément pour des classes de composés tels que les hydrocarbures, les systèmes CHO, les systèmes CHN, etc...

Les méthodes AM1 et PM3 [70] appartiennent aux dernières versions de la méthode MNDO.

Dans la méthode MNDO les paramètres associés aux intégrales bi-électroniques mono-centres sont basés sur des données spectroscopiques relatives aux atomes isolés et l'évaluation des autres intégrales bi-électroniques repose sur les interactions multipole-multipole de l'électrostatique classique. Dans cette méthode, des composés contenant H, Li, Be, B, C, N, O, F, Al, Si, Ge, Sn, Pb, P, S, Cl, Br, I, Zn, et Hg ont été paramétrés.

III. Bases de modélisation moléculaires

L'hamiltonien associé aux électrons de valence est donné par :

$$\hat{H}_{\text{Val}} = \sum_{i=1}^{n(\text{val})} \left[-\frac{1}{2} \nabla_i^2 + V(i) \right] + \sum_{i=1}^{n(\text{val})} \sum_{j \neq i} \frac{1}{r_{ij}} \quad (19)$$

qui se simplifie en :

$$\hat{H}_{\text{Val}} = \sum_{i=1}^{n(\text{val})} \hat{H}_{\text{val}}^c(i) + \sum_{i=1}^{n(\text{val})} \sum_{j \neq i} \frac{1}{r_{ij}} \quad (20)$$

$$\text{où :} \quad \hat{H}_{\text{val}}^c(i) = \left[-\frac{1}{2} \nabla_i^2 + V(i) \right] \quad (21)$$

(val) désigne le nombre d'électrons de valence du système, $V(i)$ est l'énergie potentielle de l'électron i dans le champ des noyaux et des électrons de cœur, $\hat{H}_{\text{val}}^c(i)$ est la contribution mono-électronique à \hat{H}_{Val} .

Les éléments de la matrice de Fock sont calculés à l'aide de l'équation :

$$F_{\text{val},rs} = H_{\text{val},rs}^c + \sum_{t=1}^b \sum_{u=1}^b P_{tu} \left[(rs|tu) - \frac{1}{2} (ru|ts) \right] \quad (22)$$

Dans la méthode MNDO les éléments de la matrice de Fock peuvent être calculés comme suit.

Les éléments de la matrice de cœur (intégrale de résonance de cœur) $H_{\mu_A \mu_B}^c = \langle \mu_A(1) | \hat{H}_{(1)}^c | \mu_B(1) \rangle$, avec des orbitales atomiques centrées sur les atomes A et B sont donnés par :

$$H_{\mu_A \mu_B}^c = \frac{1}{2} \left(\beta_{\mu_A} + \beta_{\nu_B} \right) S_{\mu_A \nu_B} \quad ; \quad A \neq B \quad (23)$$

où les β sont les paramètres de chaque orbitale. Par exemple, le carbone avec les orbitales atomiques de valence 2s 2p, centrées sur le même atome de carbone, aura les paramètres β_{C2s} et β_{C2p} .

Les éléments de la matrice de cœur à partir d'orbitales atomiques différentes centrées sur le même atome sont fournis par l'équation (24) : $H^c(1) = -\frac{1}{2} \nabla_1^2 + V(1)$, où $V(1)$ est l'énergie potentielle de l'électron de valence 1 dans le champ du cœur. Décomposant $V(1)$ en contributions individuelles de cœurs atomiques, il vient :

$$H^c(1) = -\frac{1}{2} \nabla_1^2 + V_A(1) + \sum_{B \neq A} V_B(1) \quad (24)$$

Ainsi :

$$H_{\mu_A \nu_B}^c = \left\langle \mu_A \left| -\frac{1}{2} \nabla_1^2 + V_A \right| \nu_A \right\rangle + \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle \quad (25)$$

Des considérations de la théorie des groupes [71] permettent d'annuler $\left\langle \mu_A \left| -\frac{1}{2} \nabla_1^2 + V_A \right| \nu_A \right\rangle$, de telle sorte que :

$$H_{\mu_A \nu_B}^c = \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle \quad (26)$$

Si l'on considère que l'électron 1 interagit avec un point du cœur de charge C_B , alors :

$$V_B = -\frac{C_B}{r_{1B}} \quad (27)$$

$$\left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle = -C_B \left\langle \mu_A \left| \frac{1}{r_{1B}} \right| \nu_A \right\rangle \quad (28)$$

Dans la méthode MNDO, $\left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle = -C_B \left\langle \mu_A \nu_A \left| s_B s_B \right\rangle$, où s_B est l'orbitale de valence s centrée sur l'atome B :

$$H_{\mu_A \nu_B}^c = \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle = -\sum_{B \neq A} C_B \left\langle \mu_A \nu_A \left| s_B s_B \right\rangle; \mu_A \neq \nu_A \quad (29)$$

Les éléments de la matrice de cœur : $H_{\mu_A \mu_A}^c = \left\langle \mu_A(1) \left| \hat{H}^c \right| \mu_A(1) \right\rangle$ sont calculés en utilisant la relation :

$$H_{\mu_A \mu_A}^c = \left\langle \mu_A \left| -\frac{1}{2} \nabla^2 + V_A \right| \nu_A \right\rangle + \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle = U_{\mu_A \mu_A}^c + \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle \quad (30)$$

$U_{\mu_A \mu_A}^c$ est évalué à partir de paramètres tirés de spectres atomiques (les paramètres utilisés pour l'atome de carbone : U_{ss} et U_{pp}). Donc :

$$H_{\mu_A \nu_A}^c = U_{\mu_A \mu_A} + \sum_{B \neq A} C_B \left\langle \mu_A \nu_A \left| s_B s_B \right\rangle \quad (31)$$

L'évaluation de $\left\langle \mu_A \nu_A \left| s_B s_B \right\rangle$ est réalisée comme suit :

- 1) Toutes les intégrales tri et tétra – centres sont annulées dans la méthode RDN.
- 2) Les intégrales de répulsion électroniques mono-centres sont soit des intégrales coulombiennes $g_{\mu \nu} \left\langle \mu_A \mu_A \left| \nu_A \nu_A \right\rangle$, soit des intégrales d'échange $h_{\mu \nu} \left\langle \mu_A \nu_A \left| \mu_A \nu_A \right\rangle$.

Pour l'atome de carbone, par exemple, les intégrales sont g_{ss} , g_{sp} , g_{pp} , $g_{pp'}$, h_{sp} et $h_{pp'}$, p et p' étant portées par des axes différents.

- 3) Les intégrales de répulsion bi-centres sont calculées à partir des valeurs d'une intégrale mono-centre et la distance inter - nucléaire en utilisant une procédure d'expansion multipole[72].

4) Le terme de répulsion cœur – cœur est donné par :

$$V_{CC} = \sum_{B \neq A} \sum_A [C_A C_B (s_A s_B / s_B s_B) + f_{AB}] \quad (32)$$

où :

$$f_{AB} = f_{AB}^{MNDO} = \left[C_A C_B (s_A s_B / s_B s_B) \left(e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} \right) \right] \quad (33)$$

α_A et α_B sont les paramètres des atomes A et B. Pour les paires O-H et N-H, par exemple, on aura :

$$f_{AH}^{MNDO} = \left[(C_A C_H (s_A s_H) s_H s_H) \left(R_{AH} e^{-\alpha_A R_{AH}} + e^{-\alpha_H R_{AH}} \right) \right] \alpha_A \alpha_H \quad (34)$$

où A désigne soit N soit O.

Dans la méthode MNDO, les paramètres suivants doivent être optimisés :

- 1) Les intégrales mono-électroniques mono-centres U_{ss} et U_{pp} .
- 2) L'exposant ξ de la STO. Pour la MNDO $\xi_s = \xi_p$.
- 3) β_s et β_p . La méthode MNDO suppose que $\beta_s = \beta_p$.

Dans la méthode AM1, $\xi_s \neq \xi_p$.

Des composés comportant différents atomes (H, B, Al, C, Si, Ge, Sn, N, P, O, S, F, Cl, Br, I, Zn et Hg) ont été paramétrés dans AM1.

On a :

$$f_{AB}^{AM1} = f_{AB}^{MNDO} + \frac{C_A C_B}{R_{AB}} \left[\sum_k a_{kA} \exp \left[-b_{kA} (R_{AB} - C_{BA})^2 \right] \right] + \frac{C_A C_B}{R_{AB}} \left[\sum_k a_{kB} \exp \left[-b_{kB} (R_{AB} - C_{KB})^2 \right] \right] \quad (35)$$

Stewart a re-paramétré les valeurs pour générer la série PM. Celle qui dérive de AM1 est connue sous l'appellation PM3 (Parametric Method 3).

Dans la méthode PM3, les intégrales de répulsion mono-centres sont paramétrées par optimisation. La fonction de répulsion de cœur contient seulement deux fonctions gaussiennes par atome. Des composés comportant des atomes parmi : H, C, Si, Ge, Sn, Pb, N, P, As, Sb, Bi, O, S, Se, Te, F, Cl, Br, I, Al, Ga, In, Te, Be, Mg, Zn, Cd et Hg ont été paramétrés dans PM3.

III.3. Champ de force.

III.3.1 Définition :

La mécanique moléculaire est une méthode d'analyse conformationnelle basée sur l'utilisation de champs de forces empiriques et la minimisation d'énergie.

Dans un sens général, la mécanique moléculaire traite les atomes (ou les noyaux) d'une molécule comme des masses ou des sphères reliées par des ressorts de différentes forces représentant les liaisons.

Les interactions entre particules (de type atomique) sont traitées à l'aide de fonctions de potentiel tirées de la mécanique classique : fonctions de potentiel individuelles pour décrire les différents types d'interactions.

Les fonctions d'énergie potentielle comportent des paramètres empiriques décrivant des interactions entre des ensembles d'atomes. La paramétrisation est faite à partir de données expérimentales (RMN, RX, calculs *ab initio*) sur le plus grand ensemble possible de molécules. Le choix des données expérimentales est important et le modèle obtenu en dépend étroitement. Les constantes sont ajustées pour rendre l'expression de l'énergie potentielle, E , la plus générale possible.

Les fonctions de potentiel et les paramètres exploités pour l'évaluation des interactions sont désignés par 'champ de force'.

Une hypothèse importante est que le champ de force déterminé à partir d'un ensemble de molécules est transférable à d'autres molécules.

Notons qu'un ensemble de paramètres développés et testés sur un nombre relativement petit de cas est encore applicable à une plus large gamme de problèmes. En outre, les paramètres développés à partir de données relatives à de petites molécules peuvent être utilisés pour étudier des molécules plus grandes tels que les polymères.

III.3.2 Quelques exemples :

Parmi les champs disponibles nous citerons les plus répandus largement utilisés pour le traitement de petites molécules.

- **MM2, MM3 et MM4**, <http://enropa.chem.uga.edu/allinger/mm2mm3.html> introduit par Allinger *et al* [73-76], largement utilisé pour le traitement de petites molécules.

- **AMBER**: <http://amber.scripps.edu>

(Assisted Method Building and Energy Refinement) introduit par Cornell *et al* [77], très largement utilisé dans le traitement des protéines et des acides nucléiques.

- **CHARMM:** <http://yuri.harvard.edu>

(Chemistry at Harvard Macromolecular Mechanics) développé par Mackerall, Karplus *et al* [75] qui est largement utilisé pour la simulation de petites molécules et jusqu'aux complexes solvatés de grandes macromolécules biologiques.

- **MMFF :** (Merck Molecular Force Field) développé par Halgren [78-80], il est similaire à MM3 dans la forme, mais en diffère par son application focalisée sur les processus de condensation de phases (en dynamique moléculaire). Il reproduit l'exactitude MM3 pour les petites molécules, et est applicable aux protéines et autres systèmes d'importance biologique.

Les différents champs de force se distinguent par trois aspects principaux :

- 1) La forme de la fonction de chaque terme énergétique.
- 2) Le nombre de termes croisés (qui reflètent le couplage entre coordonnées internes) inclus.
- 3) Le type d'information utilisé pour ajuster les paramètres.

III.3.3 Représentation simple d'un champ de force :

Beaucoup de champs de force utilisés actuellement pour la modélisation moléculaire peuvent s'interpréter en termes d'une représentation relativement simple à quatre composantes des forces intra et intermoléculaires internes au système considéré. Les pénalités énergétiques sont associées aux écarts des liaisons et des angles par rapport à leurs valeurs de 'référence' ou 'd'équilibre', il y a une fonction qui décrit la façon dont l'énergie change lors de la rotation des liaisons, et finalement le champ de force contient des termes décrivant l'interaction entre des parties non liées du système. Des champs de forces plus sophistiqués peuvent comporter des termes additionnels, mais ils présentent invariablement ces quatre composantes. Un fait intéressant lié à cette représentation est qu'on peut imputer les variations à des coordonnées internes spécifiques telles que les longueurs et les angles de liaisons, la rotation des liaisons ou les mouvements des atomes relativement les uns aux autres. Ce qui permet de comprendre facilement comment les changements des paramètres du champ de force affectent ses performances, et aident également dans le processus de paramétrisation.

Pour des molécules seules ou des ensembles d'atomes et / ou de molécules, un tel champ de force a la forme fonctionnelle suivante :

$$\begin{aligned}
 \mathbf{V}(\mathbf{r}^N) = & \sum_{\text{liaisons}} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} (1 - \cos(n\omega - \gamma)) \\
 & + \sum_{i=1}^N \sum_{j=i+1}^N \left(4 \varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \quad (36)
 \end{aligned}$$

$\mathbf{V}(\mathbf{r}^N)$ représente l'énergie potentielle qui est fonction des positions (\mathbf{r}) des N particules (habituellement les atomes)

Les diverses contributions sont représentées schématiquement sur la figure 4.

Le premier terme de l'équation (36) modèle l'interaction entre paires d'atomes liés, représentée dans ce cas par un potentiel harmonique qui fournit la variation de l'énergie lorsque la longueur de liaison l_i dévie de sa valeur de référence (à l'équilibre) $l_{i,0}$. Le second terme est une sommation sur tous les angles de valence de la molécule, encore modélisé par un potentiel harmonique (un angle de valence est l'angle formé par 3 atomes A-B-C où A et C sont tous deux liés à B). Le troisième terme dans l'équation (36) renseigne sur la variation d'énergie lorsqu'une liaison tourne. La quatrième contribution est le terme de non liaison, qui est calculé entre toutes les paires d'atomes (i et j) localisés sur différentes molécules ou appartenant à la même molécule mais séparés par au moins 3 liaisons (c'est – à – dire avec une relation 1, n où $n \geq 4$). Dans un champ de force simple le terme de non liaison comprend un terme de potentiel coulombien pour les interactions électrostatiques et le potentiel de Lennard – Jones pour les interactions de van der Waals.

- Elongation des liaisons : MM2 a été paramétré pour ajuster les distances moyennes calculées à partir du mouvement vibrationnel à température ambiante à celles obtenues par diffraction électronique.

Pour mieux reproduire la courbe de Morse, MM2 fait intervenir un terme quadratique et un autre cubique :

$$v(l) = \frac{k}{2} (l - l_0)^2 \left[1 - k'(l - l_0) \right] \quad (37)$$

- Variation des angles : Les déviations des angles de leurs valeurs de référence sont souvent écrites en utilisant la loi de Hooke ou potentiel harmonique :

$$v(\theta) = \frac{k}{2} (\theta - \theta_0)^2 \quad (38)$$

Comme pour les termes d'élongation des liaisons, la précision du champ de force peut être augmentée par l'incorporation de termes d'ordres supérieurs. MM2 contient un terme d'ordre 4 en plus du terme quadratique.

$$v(\theta) = \frac{k}{2} (\theta - \theta_0)^2 \left[1 - k'(\theta - \theta_0)^2 \right] \quad (39)$$

III. Bases de modélisation moléculaires

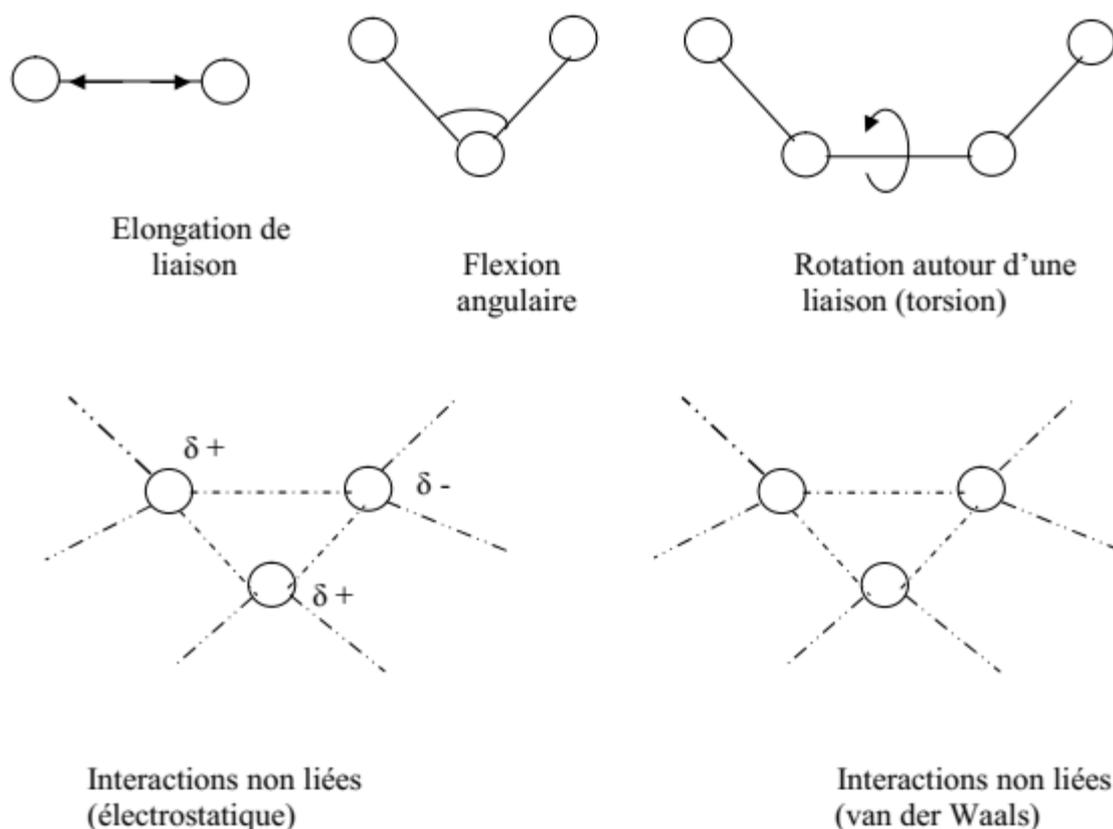


Figure 4 : Représentation schématique des quatre contributions à un champ de force de MM : élongation de liaison, flexion angulaire [53].

- Torsion des angles dièdres : correspond à la rotation d'une liaison selon l'angle dièdre ω formé par 4 atomes. Le champ de force MM2 utilise 3 termes d'une série de Fourier :

$$v(\omega) = \frac{V_1}{2} (1 + \cos \omega) + \frac{V_2}{2} (1 - \cos 2\omega) + \frac{V_3}{3} (1 + \cos 3\omega) \quad (40)$$

Une interprétation physique a été attribuée à chacun des 3 termes à partir de l'analyse de calculs *ab initio* effectués sur des hydrocarbures fluorés simples.

- Angle dièdre impropre ou déviation extra - planaire :

Examinons comment la cyclobutanone serait modélisée en utilisant uniquement un champ de force comportant des termes standards pour l'élongation des liaisons et les variations angulaires du type de ceux apparaissant dans l'équation (40). La structure d'équilibre obtenue avec un tel champ de force sera caractérisée par la localisation de l'atome d'oxygène hors du plan formé par l'atome de carbone contigu (à l'oxygène) et les 2 carbones qui lui sont liés (figure 5).

Dans cette configuration, les angles vers l'oxygène adoptent des valeurs proches de la valeur de référence 120° . Expérimentalement, il est établi que l'atome d'oxygène demeure dans le plan du cyclobutane bien que les angles C-C=O soient grands (133°). Ceci parce que l'énergie de liaison

π , qui est maximalisée dans l'arrangement coplanaire, sera plus réduite si l'atome d'oxygène est dévié

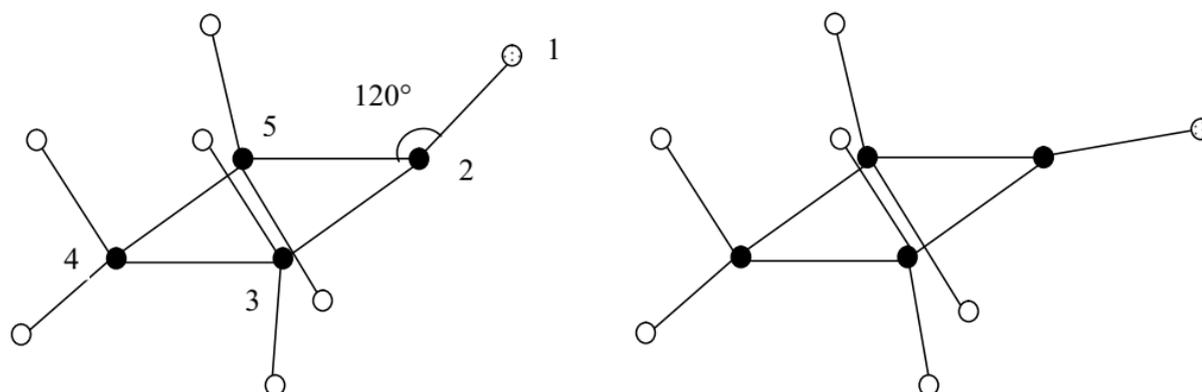


Figure 5: Sous un terme extra - planaire, l'atome d'oxygène de la cyclobutanone est prévu de se situer hors du plan du cycle / gauche plutôt que dans le plan [53].

hors du plan. Pour réaliser la géométrie désirée il est nécessaire d'ajouter un (ou des) terme(s) additionnel(s) dans le champ de force qui maintienne(nt) le carbone sp^2 et les 3 atomes qui lui sont liés dans le même plan. La plus simple façon de le réaliser est d'utiliser un terme de variation angulaire extra – planaire.

Il y a plusieurs façons d'incorporer des termes de variation extra – planaire dans un champ de force. Une approche possible est de traiter les 4 atomes comme un angle de torsion impropre pour lequel les 4 atomes (figure 6) ne sont pas liés dans la séquence 1 – 2 – 3 – 4. Une façon de définir, dans ce cas, une torsion impropre consiste à impliquer les atomes 1 – 5 – 3 – 2 de la figure 6.

Un potentiel de torsion de la forme suivante :

$$v(\omega) = k (1 - \cos 2\omega) \quad (41)$$

peut être utilisé pour maintenir l'angle de rotation impropre à 0° ou 180° .

Il existe diverses autres façons pour inclure la contribution de la variation d'angle extra – planaire. Par exemple, une définition qui est la plus proche de la notion de « variation extra – planaire » implique le calcul de l'angle entre une liaison à partir de l'atome central et le plan défini par l'atome central et les 2 autres atomes (figure 6). La valeur 0° correspond aux 4 atomes coplanaires. Une troisième approche consiste à calculer la hauteur de l'atome central au – dessus du plan défini par les 3 autres atomes (figure 6). Avec ces 2 définitions la déviation de la coordonnée extra – planaire (angle ou distance) peut être modélisée à l'aide d'un potentiel harmonique de la forme :

$$v(\theta) = \frac{k}{2} \theta^2 \quad ; \quad v(h) = \frac{k}{2} h^2 \quad (42)$$

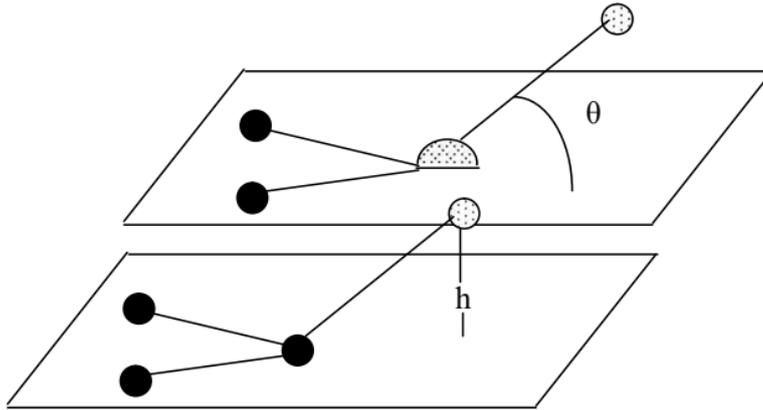


Figure 6: Deux façons pour modéliser les contributions de la variation d'angle extra – planaire [53].

- Termes de croisement : Les termes de croisement permettent de tenir compte des interactions entre l'élongation des liaisons, la variation des angles et la torsion des angles dièdres. Par exemple, si les liaisons C-O et O-H de l'angle C-O-H sont étirées, alors la distance entre les atomes terminaux (C et H) est augmentée, ce qui rend plus facile la diminution de l'angle COH. Pareillement, la diminution de l'angle COH tend à être accompagnée d'une augmentation des longueurs des liaisons O-H et C-O. Pour tenir compte de cette interaction, on peut ajouter un terme de croisement « élongation – variation angulaire ». (stretch - bend) de la forme :

$$v_{\Delta \theta} = \frac{1}{2} k_{12} (\Delta l_1 + \Delta l_2) \Delta \theta \quad (43)$$

avec $\Delta l_1 = l_1 - l_{10}$; $\Delta l_2 = l_2 - l_{20}$ et $\Delta \theta = \theta - \theta_0$

l_{10} , l_{20} et θ_0 représentent les valeurs de référence pour l_1 , l_2 et θ respectivement.

Les termes de croisement les plus utilisés sont (figure 7) :

- élongation – élongation et élongation – variation angulaire, pour deux liaisons à un même atome ;
- élongation – torsion angle dièdre, variation angulaire - torsion angle dièdre et variation angulaire - variation angulaire pour 2 angles avec un atome central commun.

Le champ de force MM2 fait intervenir uniquement un terme de croisement élongation – variation angulaire.

III. Bases de modélisation moléculaires

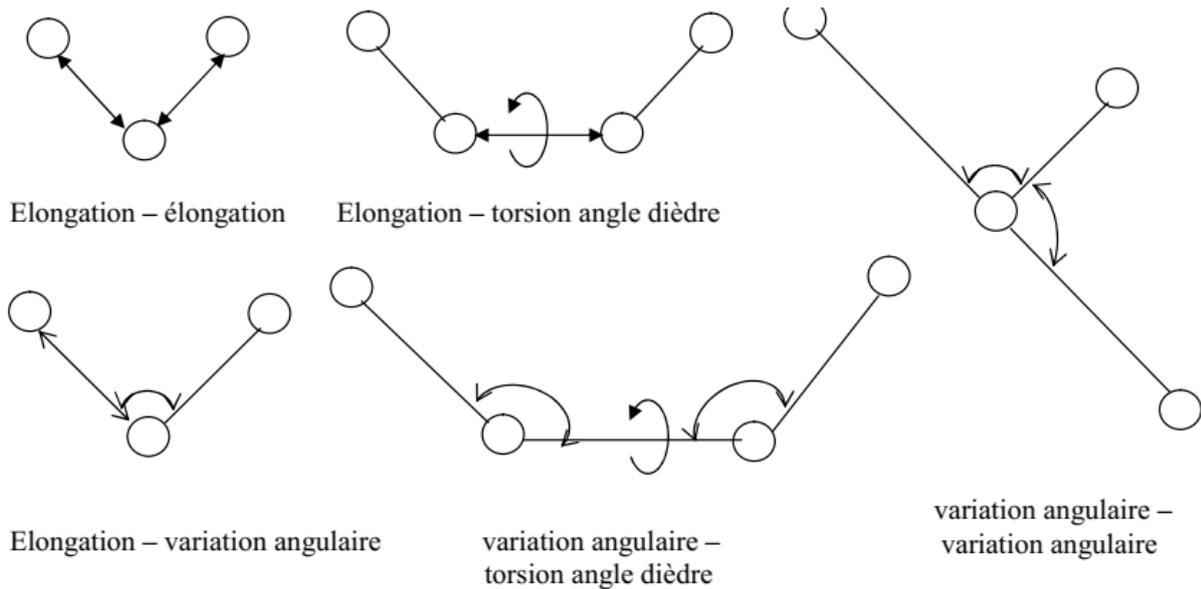


Figure 7: Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces [53].

- Interactions électrostatiques : Le terme électrostatique v_{es} est généralement défini par la somme des interactions électrostatiques entre toutes les paires d'atomes exceptées les paires 1, 2 et 1, 3 :

$$v_{es} = \sum_{1, \geq 4} v_{es, ij}, \text{ où les atomes } i, j \text{ vérifient la relation } (1, \geq 4v_{es, ij}).$$

V_{es} est généralement calculé en affectant des charges atomiques partielles à chaque atome et en appliquant un potentiel coulombien.

MM2 n'utilise pas cette procédure mais associe un moment dipolaire avec chaque liaison puis calcule V_{es} comme somme des énergies potentielles d'interactions entre moments de liaisons. Chaque moment dipolaire étant localisé au centre, et dirigé le long de cette liaison. Les valeurs des moments de liaisons de certains types de liaisons ont été choisies pour ajuster les moments dipolaires expérimentaux de petites molécules.

L'équation (44) décrit un modèle de charge, basé essentiellement sur les dipôles centrés, utilisé dans MM2 [73].

$$v_{es} = \frac{\mu_i \mu_j}{k r^3} (\cos \chi - 3 \cos \alpha_i \cos \alpha_j) \quad (44)$$

χ et α_i, α_j désignent respectivement les angles entre les dipôles et les angles entre chaque dipôle et le vecteur de connexion.

- Interactions de van der Waals : La plupart des champs de force utilisent le potentiel 12 – 6 de Lennard Jones.

MM2 utilise le potentiel de Buckingham, avec un terme attractif proportionnel à r^{-6} et un terme répulsif proportionnel à $e^{-\alpha r}$ où α est un paramètre :

$$V_{vdw} = A e^{-\alpha r} - \frac{B}{r^6} \quad (45)$$

- Liaisons conjuguées : MM2 utilise une procédure générale qui consiste à effectuer des calculs semi-empiriques sur les électrons π pour en tirer les ordres de liaisons, qui sont ensuite utilisés pour affecter des longueurs de liaisons et des constantes de force de référence pour les liaisons conjuguées.

III.3.4 Champ de force MM+.

C'est une extension du Champ de force MM2, avec l'ajout de quelques paramètres additionnels. MM+ est un champ de force robuste, il a l'aptitude de prendre en considération les paramètres négligés dans d'autres champs de force et peut donc s'appliquer pour des molécules plus complexes tels que les composés inorganiques.

III. Bases de modélisation moléculaires

Le tableau suivant [81] compare les trois techniques computationnelles majeures évoquées.

Tableau 4: Etude comparative des techniques ab initio, semi –empirique et mécanique moléculaire.

Ab initio	Semi -empirique	Mécanique moléculaire
<ul style="list-style-type: none"> • Prise en compte de tous les électrons. • Limitée à quelques dizaines d'atomes. • Nécessite un super ordinateur. • Peut être appliquée à des composés inorganiques, organique, organométalliques ,et aux fragments moléculaires (composants catalytiques d'enzymes). • Vide, solvation implicite. • Applicable à l'état fondamental, et aux états de transition et excité. 	<ul style="list-style-type: none"> • Ignore certains électrons (simplification). • Limitée à quelques centaines d'atomes. • Peut être appliquée à des composés inorganiques, organiques, organométalliques et de petits oligomères (peptides, nucléotides, saccharides). • Vide, solvation implicite. • Applicable à l'état fondamental, et aux états de transition et excité. 	<ul style="list-style-type: none"> • Ignore tous les électrons ,seuls les noyaux sont considérés. • Molécules contenant des milliers d'atomes. • Peut être appliquée aux composés inorganiques, organiques, oligo–nucléotides, peptides, saccharides,métallo-organiques et inorganiques. • Vide, solvation implicite ou explicite. • Applicable uniquement à l'état fondamental.

IV-1 Modélisation

La modélisation par apprentissage consiste à trouver le jeu de paramètres qui conduit à la meilleure approximation possible de la fonction de régression, à partir des couples entrées/sortie constituant l'ensemble d'apprentissage (ou de calibrage) ; le plus souvent, ces couples sont constitués d'un ensemble de vecteurs de variables (descripteurs dans le cas de molécules) $\{ x^i, i = 1 \dots n \}$, et un ensemble de mesures de la grandeur à modéliser $\{ y(x^i), i = 1 \dots n \}$ [53]. La détermination des valeurs de ces paramètres nécessite la mise en œuvre de méthodes d'optimisation qui diffèrent selon le type de modèle choisi.

Dans cette thèse deux types de méthodes ont été exploitées.

IV-2 Régression linéaire multiple (MLR) [82-84]

La régression linéaire multiple est la méthode la plus simple de modélisation, elle consiste à rechercher une équation linéaire par rapport à ses paramètres reliant la variable à modéliser au vecteur d'entrées $\mathbf{x} = \{x_k, k = 1 \dots p\}$. Ces entrées peuvent être des fonctions non paramétrées, ou à paramètres fixés, de ces variables. L'équation linéaire recherchée est de la forme :

$$g(\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^q \theta_k x_k = \mathbf{X} \boldsymbol{\theta} \quad (47)$$

Où $\boldsymbol{\theta} = \{\theta_k, k=1 \dots p\}$ est le vecteur des paramètres; \mathbf{X} , matrice des observations de taille (n, p) , est définie comme la matrice dont les éléments de la colonne k prennent pour valeurs les n mesures de la variable k . Pour chaque élément i de la base d'apprentissage, le résidu est défini comme la différence entre la valeur de la grandeur à modéliser pour cet élément i et l'estimation du modèle :

$$R_i = y^i - g(\mathbf{x}^i, \boldsymbol{\theta}) \quad (48)$$

L'apprentissage est réalisé par minimisation de la fonction de coût des moindres carrés, qui mesure l'ajustement du modèle g aux données d'apprentissage :

$$J(\boldsymbol{\theta}) = \sum_{i=1}^N (R_i)^2 = \sum_{i=1}^N [y^i - g(\mathbf{x}^i, \boldsymbol{\theta})]^2 = \|\mathbf{y} - \mathbf{X} \boldsymbol{\theta}\|^2 \quad (49)$$

La fonction $J(\boldsymbol{\theta})$ est une fonction positive quadratique en $\boldsymbol{\theta}$: son minimum est unique. Il est donné par :

$$\theta_{mc} = (X^T X)^{-1} X^T y \quad (50)$$

Les paramètres θ_k sont appelés coefficients de régression partielle ; chacun d'eux mesure l'effet de la variable explicative x_k concernée sur la propriété modélisée lorsque les autres variables explicatives sont maintenues constantes.

La régression linéaire est facile à mettre en œuvre, et les coefficients θ_k obtenus peuvent être interprétés : ils mesurent l'influence de chacune des variables sur les grandeurs étudiées.

Cependant, il est souvent nécessaire d'avoir recours à des modèles de plus grande complexité.

IV-3 Réseaux de neurones artificiels [85-88]

Les réseaux de neurones formels [85] étaient, à l'origine, une tentative de modélisation mathématique des systèmes nerveux, initiée dès 1943 par McCulloch et Pitts [86].

Un *neurone formel* est une fonction non linéaire paramétrée, à valeurs bornées, de variables réelles. Le plus souvent, les neurones formels réalisent une combinaison linéaire des entrées reçues, puis appliquent à cette valeur une « fonction d'activation » f , généralement non linéaire. La valeur obtenue y est la sortie du neurone. Un neurone formel est ainsi représenté sur la Figure 8.

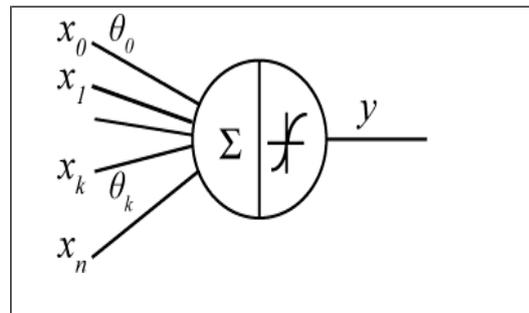


Figure 8 : Représentation d'un neurone formel

Les $\{x_k\}_{k=1,\dots,n}$ sont les variables, ou *entrées* du neurone, et les $\{\theta_k\}_{k=0,\dots,n}$ sont les *paramètres*, également appelés synapses ou poids. Le paramètre θ_0 est le paramètre associé à une entrée fixée à 1, appelée biais. L'équation du neurone est donc :

$$y = f \left(\theta_0 + \sum_{k=1}^n \theta_k x_k \right) \quad (46)$$

Les fonctions d'activation les plus couramment utilisées sont la fonction tangente

hyperbolique, la fonction sigmoïde et la fonction identité.

Les neurones seuls réalisent des fonctions assez simples, et c'est leurs compositions qui permettent de construire des fonctions aux propriétés particulièrement intéressantes. On appelle ainsi *réseau de neurones* une composition de fonctions « neurones » définies ci-dessus.

La Figure 9 représente un réseau de neurones non bouclé, organisé en couches (perceptron multicouche), qui comporte N_e variables, une couche de N_c neurones cachés, et N_s neurones de sortie

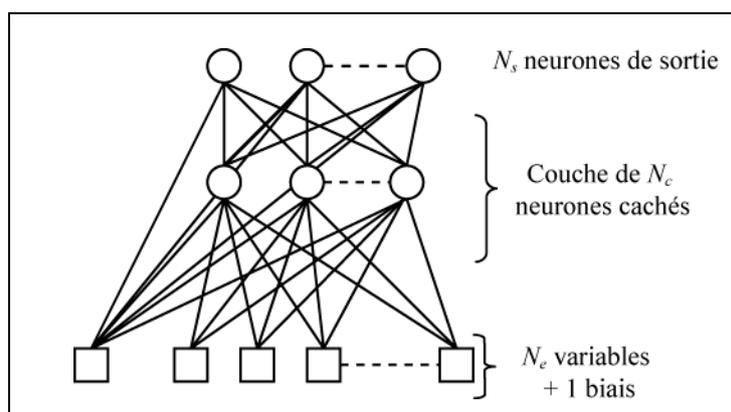


Figure 9 : Représentation d'un réseau de neurones

À chaque connexion est associé un paramètre. Les sorties du réseau sont donc des fonctions non-linéaires de ses variables et de ses paramètres. Le nombre de degrés de liberté, c'est-à-dire de paramètres ajustables, dépend du nombre de neurones de la couche cachée ; il est donc possible de faire varier la complexité du réseau en augmentant ou en diminuant le nombre de neurones cachés.

IV-3.1 Propriétés des réseaux de neurones

Les réseaux de neurones ont pour but de modéliser des processus, à partir d'exemples de couples entrées / sorties. Ils ont la propriété d'*approximation universelle* : un réseau de neurones comportant un nombre fini de neurones cachés, de même fonction d'activation, et un neurone de sortie linéaire, est capable d'approcher uniformément, avec une précision arbitraire, toute fonction bornée suffisamment régulière, sur un domaine fini de l'espace de ses variables. De plus, il s'agit d'*approximateurs parcimonieux* : une approximation par un réseau de neurones nécessite en général moins de paramètres que les approximateurs usuels. Le nombre de paramètres nécessaires pour obtenir une précision donnée augmente en effet linéairement avec le nombre de variables pour un réseau de neurones, alors qu'il croît exponentiellement pour un modèle linéaire par rapport aux paramètres. Cette propriété est très importante, car les réseaux de neurones demandent de ce fait moins d'exemples que d'autres approximateurs pour l'apprentissage.

IV-3.2 Apprentissage des réseaux des neurones

Considérons un ensemble d'apprentissage, constitué de n couples entrées / sorties, c'est-à-dire d'un ensemble des variables $\{y(x^i), i=1\dots n\}$ et d'un ensemble de mesures de la grandeur à modéliser $\{y(x^i), i=1\dots n\}$. Pour une complexité donnée, l'apprentissage s'effectue par minimisation de la fonction de coût des moindres carrés, définie par :

$$J(\theta) = \frac{1}{2} \sum_{j=1}^N [y(x^i) - g(x^i, \theta)]^2 \quad (47)$$

La minimisation de cette fonction s'effectue par une descente de gradient. Cet algorithme a pour but de converger, de manière itérative, vers un minimum de la fonction de coût, à partir de valeurs initiales des poids aléatoires. À chaque étape, le gradient de la fonction est calculé, à l'aide de l'algorithme de *rétro-propagation*. Puis les paramètres sont modifiés en fonction de ce gradient, dans la direction de la plus forte pente, vers un minimum local de J . Cette descente peut être effectuée suivant plusieurs méthodes : gradient simple ou méthodes du second ordre, dérivées de la méthode de Newton. Les méthodes du second ordre, généralement plus efficaces, sont les plus utilisées. La procédure de minimisation est arrêtée lorsqu'un critère est satisfait : le nombre maximal d'itérations est atteint, la variation du module du vecteur des paramètres ou du gradient de la fonction de coût est trop faible...

IV-3.3 Des réseaux de neurones particuliers : réseaux de fonctions radiales de Base

Les réseaux de neurones de fonctions radiales de base (souvent notés RBF), sont des réseaux dont la couche cachée est composée de neurones à fonction d'activation gaussienne radiale. La fonction d'activation d'un neurone j est par exemple :

$$f(x, \sigma_j, \theta_j) = \exp\left(-\frac{\|x - \theta_j\|^2}{2\sigma_j^2}\right) \quad (48)$$

où σ_j est l'écart-type de la gaussienne et θ_j est le vecteur des coordonnées de son centre. Les neurones de sortie sont à fonction d'activation linéaire, et sont reliés aux neurones de la couche cachés par des poids β_j ajustables. Une sortie y d'un réseau à N_c neurones cachés est ainsi déterminée par :

$$y = \sum_{j=1}^{N_c} \beta_j f(x, \sigma_j, \theta_j) \quad (49)$$

La sortie est une fonction linéaire des poids β_j et non-linéaire des paramètres des gaussiennes. Si les paramètres des gaussiennes sont fixés, la sortie devient une fonction linéaire des poids β_j , et les réseaux perdent leur propriété de parcimonie. De plus, la

réussite de l'apprentissage dépend fortement de l'initialisation. Enfin, la configuration d'un réseau de neurones RBF optimal est difficile. Lors du processus d'apprentissage du réseau, deux stratégies sont possibles. La première consiste à modifier simultanément tous les paramètres du réseau (les coordonnées des centres des fonctions radiales, leur écart-type et les poids β_j), par descente de gradient. Cependant, les dynamiques de convergence des fonctions radiales et des poids β_j sont différentes, et les poids convergent plus rapidement que les autres paramètres. L'apprentissage conduit très souvent à un minimum local. Une autre méthode consiste à optimiser séparément les paramètres de la couche cachée, par apprentissage non-supervisé, et les poids entre la couche cachée et la couche de sortie, par descente dégradée.

Dans la plupart des applications des RNA à la chimie l'utilisation d'une seule couche cachée semble suffire [89]. Nous avons donc utilisé dans ce travail un réseau standard à 3 couches comprenant l'entrée, la sortie et une couche cachée. L'algorithme de Levenberg-Marquardt conçu pour faciliter certains problèmes de convergence est l'un des plus utilisés pour l'apprentissage des réseaux, d'autant plus qu'il s'adapte très bien avec le choix de l'erreur quadratique moyenne comme indice de performance.

Nous avons donc utilisé l'algorithme Levenberg-Marquardt de rétropropagation (fonction TRAINLM de la boîte à outils du logiciel MATLAB 7.0 [90] pour l'apprentissage du réseau. Les fonctions de transfert sigmoïde (tangente hyperbolique) et linéaire ont été adoptées comme fonctions d'activation, respectivement pour les couches cachée et de sortie.

IV-4 Développement et évaluation de modèle

IV-4.1 Sélection d'un sous-ensemble de descripteurs

Des logiciels spécialisés permettent le calcul de plus de 6000 descripteurs moléculaires appartenant à différentes classes. Plutôt que de chercher à expliquer la variable dépendante (grandeur d'intérêt) par tous les régresseurs (descripteurs moléculaires), on peut chercher seulement un ensemble réduit de régresseurs qui donne une reconstitution aussi satisfaisante de la variable à expliquer. Parmi les stratégies mises en œuvre pour la sélection d'un ensemble réduit de variables explicatives, on peut citer : les méthodes de pas-à-pas (méthode descendante ; méthode ascendante et méthode dite stepwise), ainsi que les algorithmes génétiques.

En général, la comparaison se fait à l'avantage des algorithmes génétiques (AG) que nous avons appliqués dans le présent travail, et que nous rappelons succinctement.

IV-4.2 Principe de l'algorithme génétique

Dans la terminologie des algorithmes génétiques, le vecteur binaire I , appelé "chromosome", est un vecteur de dimension p où chaque position (un "gène") correspond à une valeur (1 si elle figure dans le modèle, 0 sinon). Chaque chromosome représente un modèle basé sur un ensemble de variables explicatives.

On commence par définir le paramètre statistique à optimiser (par exemple maximiser Q^2 en utilisant la validation croisée par "leave-one-out" ; (cf. infra), avec la taille P de la population du modèle (par exemple, $P = 100$), et le nombre maximum de variables L permises pour le modèle (par exemple, $L = 10$) ; le minimum de variables permises est généralement supposé égal à 1. De plus, une probabilité de croisement p_c (habituellement élevée $p_c = 0,9$), et une probabilité de mutation p_M (habituellement faible, $p_M = 0,1$) doivent être également définies.

Après définition des principaux paramètres, la mise en œuvre de l'algorithme génétique est démarrée, son évolution comprend trois étapes principales.

IV-4.3 Initialisation aléatoire du modèle.

La population est constituée au départ de modèles aléatoires avec des variables comprises entre 1 et L , puis les modèles sont ordonnés eu égard au paramètre statistique sélectionné – la qualité du modèle – (le meilleur modèle est en première position, le plus mauvais en position P) ;

IV-4.4 Etape de croisement.

A partir de la population, on sélectionne des paires de modèles (aléatoirement, ou avec une probabilité proportionnelle à leur qualité). Puis, pour chaque paire de modèles on conserve les caractéristiques communes, c'est-à-dire les variables exclues dans les 2 modèles restent exclues, et les variables intégrées dans les 2 modèles sont conservées. Pour les variables sélectionnées dans un modèle et éliminées dans l'autre, on en essaye un certain nombre au hasard que l'on compare à la probabilité de croisement p_c : si le nombre aléatoire est inférieur à la probabilité de croisement, la variable exclue est intégrée au modèle et vice versa. Finalement, le paramètre statistique du nouveau modèle est calculé : si la valeur de ce paramètre est meilleure que la plus mauvaise pour la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans la cas contraire, il n'est pas pris en compte. Cette procédure est répétée pour de nombreuses paires (100 fois par exemple).

IV-4.5 Etape de mutation.

Pour chaque modèle de la population (c'est-à-dire pour chaque chromosome) p nombres aléatoires sont éprouvés, et, un à la fois, chacun est comparé à la probabilité de mutation, p_M , définie : chaque gène demeure inchangé si le nombre aléatoire correspondant excède la probabilité de mutation, dans le cas contraire, on le change de 0 à 1 ou vice versa. Les faibles valeurs de p_M permettent uniquement peu de mutations, conduisant à de nouveaux chromosomes peu différents des chromosomes générateurs.

Après obtention du modèle transformé, on en calcule le paramètre statistique : si cette valeur est meilleure que la plus mauvaise de la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans le cas contraire il n'est pas pris en compte.

IV-4.6 Conditions d'arrêt.

Les étapes de croisement et de mutation sont répétées jusqu'à la rencontre d'une condition d'arrêt (par exemple un nombre maximum d'itérations défini par l'utilisateur), ou qu'il est mis fin arbitrairement au processus.

Une caractéristique importante de la sélection d'un ensemble réduit de variables par algorithme génétique est qu'on n'obtient pas nécessairement un modèle unique, mais le résultat consiste habituellement en une population de modèles acceptables ; cette caractéristique, parfois considérée comme un désavantage, fournit une opportunité pour procéder à une évaluation des relations avec la réponse à différents points de vue.

Notons que la taille du modèle est fixée par la valeur optimale de la fonction FIT de Kubinyi [91], calculée selon :

$$FIT = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{(n+p^2)} \quad (51)$$

p : désignant le nombre de variables du modèle et R^2 le coefficient de détermination.

Ce critère permet de comparer entre modèles construits sur un même nombre n de données, mais avec un nombre de variables p différent.

IV-5. Paramètres d'évaluation de la qualité de l'ajustement.

Deux paramètres sont couramment utilisés :

Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \quad (52)$$

où \hat{y}_i est la valeur estimée du paramètre physique, et \bar{y} la moyenne des valeurs observées.

La racine de l'erreur quadratique moyenne de prédiction (désignée également par EQMP) :

$$EQMP = \sqrt{\frac{1}{n} \sum_1^n (y_i - \hat{y}_{(i)})^2} \quad (53)$$

IV-5.1 Robustesse du modèle.

La stabilité du modèle a été explorée en utilisant la "validation croisée par omission d'une observation" (LOO : cross-validation by leave-one-out) [84]. Elle consiste à recalculer le modèle sur $(n - 1)$ composés de calibrage, le modèle obtenu servant alors à estimer la valeur de la propriété du composé éliminé noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des n composés de l'ensemble de calibrage.

"La somme des carrés des erreurs de prédiction", désignée par l'acronyme PRESS (pour : Predictive ResidualSum of Squares) :

$$PRESS = \sum_1^n (y_i - \hat{y}_{(i)})^2 \quad (54)$$

est une mesure de la dispersion de ces estimations. On l'utilise pour définir le coefficient de prédiction :

$$Q_{LOO}^2 = \frac{SCT - PRESS}{SCT} \quad (55)$$

Contrairement à R^2 , qui augmente avec le nombre de paramètres de la régression, le facteur Q_{LOO}^2 affiche une courbe avec maximum (ou avec palier), obtenu pour un certain nombre de variables explicatives, puis décroît par la suite de façon monotone. Ce fait confère une grande importance au coefficient Q_{LOO}^2 . Une valeur de $Q_{LOO}^2 > 0,5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [92].

IV-5.2 Domaine d'application.

Le domaine d'application a été discuté à l'aide du diagramme de Williams (traité en détail dans [93] représentant les résidus de prédiction standardisés en fonction des valeurs des

leviers h_i . L'équation (56) définit le levier d'un composé dans l'espace original des variables indépendantes (x_i)

$$h_i = \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \quad (56)$$

Où (\mathbf{x}_i) est le vecteur ligne des descripteurs du composé i et \mathbf{X} ($n \times p$) la matrice du modèle déduite des valeurs des descripteurs de l'ensemble de calibrage ; l'indice T désigne le vecteur (ou la matrice) transposé (e).

La valeur critique du levier (h^*) est fixée à $3(p+1)/n$. Si $h_i < h^*$, la probabilité d'accord entre les valeurs mesurée et prédite du composé i est aussi élevée que celle des composés de calibrage. Les composés avec $h_i > h^*$ renforcent le modèle quand ils appartiennent à l'ensemble de calibrage, mais auront, sinon, des valeurs prédites douteuses sans pour autant être forcément aberrantes, les résidus pouvant être bas.

IV-5.3 Test de randomisation

Ce test permet de mettre en évidence des corrélations dues au hasard. Il consiste à générer un vecteur « propriété considérée » par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle QSAR, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas) [94].

IV-5.4 Validation externe.

En plus du test de randomisation, il est intéressant [95,96], pour juger de la qualité du modèle, de considérer le coefficient de prédiction externe calculé comme suit :

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{next} (\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{ntr} (y_i - \bar{y}_{tr})^2 / n_{tr}} \quad (57)$$

La racine de l'écart quadratique moyen (RMSE pour Root Mean Squared Error), calculée sur différents ensembles:

- Ensemble d'estimation (appelée EQMC)
- Ensemble de prédiction externe (désignée par EQMPext).

Ces valeurs RMSE sont mieux adaptées, pour juger de la qualité d'un modèle que les valeurs de R^2 et Q^2 seules.

IV. Développement et validation de modèles QSAR

où n_{tr} et n_{ext} sont respectivement le nombre d'observations dans les sous ensemble de calibrage et validation et \bar{y}_{tr} étant la valeur des y pour l'ensemble de calibrage

$$EQMC = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (58)$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{n_{ext}}} \quad (59)$$

Une validation externe supplémentaire selon (Golbraikh et Tropsha, 2002) est appliquée uniquement à l'ensemble de test. Selon les critères recommandés de Tropsha *et al*, un modèle QSAR /QSPR prédictif, doit satisfaire aux conditions suivantes :

$$1) Q_{EXT}^2 > 0.5 \quad (56-a)$$

$$2) R^2 > 0.6 \quad (56-b)$$

$$3) (R^2 - R_0^2)/R^2 < 0.1 \quad \text{et} \quad 0.85 < k < 1.15 \quad (56-c)$$

$$(R^2 - R_0'^2)/R^2 < 0.1 \quad \text{et} \quad 0.85 < k' < 1.15 \quad (56-d)$$

où

$$R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (57-a)$$

$$R_0^2 = 1 - \frac{\sum (y_i - y_i^{f_0})^2}{\sum (y_i - \bar{y})^2} \quad (57-b)$$

$$R_0'^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{f_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \quad (57-c)$$

$$k = \frac{\sum (y_i \tilde{y}_i)}{\sum (\tilde{y}_i)^2} \quad (57-d)$$

$$k' = \frac{\sum (y_i \tilde{y}_i)}{\sum (y_i)^2} \quad (57-e)$$

où R est le coefficient de corrélation entre les valeurs calculées et expérimentales dans l'ensemble de test; R_0^2 (valeurs calculées par rapport aux observées) et $R_0'^2$ (valeurs observées par rapport aux calculées) sont les coefficients de détermination; k et k' sont les pentes des droites de régressions passant par l'origine pour les valeurs calculées par rapport aux valeurs observées et observées par rapport aux calculées, respectivement; $y_i^{f_0}$ et $\tilde{y}_i^{f_0}$ sont définis

IV. Développement et validation de modèles QSAR

respectivement par : $y_i^{r_0} = k \tilde{y}$ et, $\tilde{y}_i^{r_0} = k' y$; et les sommations sont sur tous les échantillons dans l'ensemble de test.

La raison d'utiliser et d'exiger des valeurs de k qui sont proches de 1 est que lorsque sont comparés les propriétés réelles par rapport aux prédites, un ajustement précis est nécessaire, non seulement une corrélation.



PARTIE II : APPLICATIONS

I-1 Collecte de données et méthodologie :

Il est généralement admis que la toxicité de nombreuses substances, particulièrement les produits chimiques organiques industriels, est la conséquence de leur solubilité dans les lipides, alors que leurs caractéristiques moléculaires spécifiques ont peu ou pas d'influence. Leur mode d'action consisterait en la destruction des processus physiologiques associés aux membranes cellulaires.

Les protozoaires sont souvent utilisés pour l'évaluation de la toxicité. Les méthodes mises en œuvre sont basées sur des critères morphologiques, ultra-structuraux, éthologiques et métaboliques [97, 98].

L'inhibition de la croissance d'une population est un indicateur très en vogue, parce qu'il peut être déterminé directement ou indirectement à l'aide d'un équipement électronique. Ce qui permet l'acquisition rapide des observations nécessaires pour les analyses de régression. Nous considérerons la concentration d'inhibition à 50% de la croissance (IGC_{50}), dont le logarithme de l'inverse soit $pIGC_{50} = \log(IGC_{50})^{-1}$, servira d'indicateur de toxicité.

Les tests de toxicité ont été réalisés par Schultz *et al* [99] en examinant la croissance d'une population de *Tetrahymona pyriformis*. Les essais ont été menés dans des erlenmeyers de 250 ml, contenant 50 ml d'un milieu dont la composition est précisée ci-après :

Tableau 5. Milieu de croissance de *Tetrahymona pyriformis*

Eau distillée	1000 ml
Proteose peptone	20 g
D-glucose	5 g
extrait de levure	1 g
FeEDTA	1 mL d'une solution à 3 % (masse/v)
pH	7,35

La température a été fixée à $27 \pm 1^\circ\text{C}$.

Ce milieu est inoculé avec 0,25 ml d'une culture contenant approximativement 36 000 cellules par ml. La croissance des ciliés est suivie par spectrophotométrie, en mesurant la densité optique (absorbance) à 540 nm après 48 heures d'incubation. (On pourra trouver dans [100, 101] des indications plus complètes).

Plusieurs critères ont guidé au choix des composés toxiques examinés. Tous sont disponibles dans le commerce avec une pureté suffisante (95 % et plus), ce qui ne nécessite pas une re-purification préalablement au test. Des précautions ont été observées afin d'assurer une diversité concernant, à la fois, les propriétés physico-chimiques et la position des substituants.

Les solutions stocks des divers composés toxiques, ont été préparées dans le diméthylsulfoxyde (DMSO) à des concentrations de 5, 10, 25 et 50 grammes par litre. Dans chaque cas, le volume de solution stock ajouté à chaque fiole est limité par la concentration finale de DMSO qui ne doit pas excéder 0,75 % (350 ml par fiole), quantité qui n'altère pas la reproduction de *Tétrahymena* [100, 101].

L'ensemble de données relatives aux anilines (tiré de Schultz *et al* [99]) a été divisé au hasard en un ensemble d'apprentissage (31 objets), utilisé pour développer le modèle QSAR, et un ensemble de validation (17 objets), utilisé uniquement pour les statistiques de validation externe.

Les structures de toutes les molécules ont été pré-optimisées à l'aide du champ de force MM⁺ de la mécanique moléculaire (algorithme Polak-Ribiere) en utilisant le programme HyperChem 6.03 [102]. Les géométries finales d'énergie conformationnelle minimale ont été obtenues par la méthode semi-empirique AM1, dans le cadre du formalisme RHF sans interaction de configuration. En appliquant pour norme limite, une racine du carré moyen du gradient égale à 0,001 kcal/mol. Les géométries ainsi optimisées ont été transférées dans le logiciel DRAGON [103] pour calculer 1664 descripteurs dont 271 du type Géométrique et GETAWAY (GEometry, Topology and Atoms Weighted Assembly). Les descripteurs avec des valeurs constantes ou quasi constantes dans chaque groupe ont été rejetés. Pour chaque paire de descripteurs corrélés (avec un coefficient de corrélation $r \geq 0,95$), celui présentant la plus forte corrélation des paires avec les autres descripteurs a été exclu.

L'algorithme génétique (AG) [104] a été considéré comme supérieur aux autres méthodes de sélection de variables. Ainsi, la sélection des variables a été effectuée sur l'ensemble d'apprentissage, en utilisant l'AG dans la version de MobyDigs de Todeschini [105] en maximisant la variance expliquée par validation croisée par omission d'une observation Q_{LoO}^2 . en utilisant la régression par les moindres carrés ordinaires et la sélection de sous-ensembles de variables explicatives par algorithme génétique (Genetic Algorithm-Variable Subset Sélection ou GA-VSS) [106].

Dans le logiciel MobyDigs les processus de croisement et de mutation de l'algorithme génétique sont contrôlés par un paramètre T variant de 0 à 1. Les paramètres de l'algorithme génétique ont été fixés comme suit : population des modèles Pop = 100 ; valeur de T fixée à 0,5 pour équilibrer les rôles des deux processus de croisement et de mutation.

I-2 Présentation et discussion du modèle QSAR

I-2-1 Qualités internes du modèle QSAR

L'application de la méthode GA-VSS a conduit à meilleurs modèles pour la prédiction de $pIGC_{50}$ basés sur différents ensembles de descripteurs moléculaires. Le meilleur modèle bidimensionnel a été construit en utilisant le rayon de giration ($RGyr$) [Le symbole ($RGyr$) correspond au rayon de giration (masse pesée). C'est parmi les descripteurs], et l'auto-corrélation maximale du décalage 3 pondérée par le volume atomique de van der Waals ($R3v+$), [Le symbole ($R3v+$) correspond à R autocorrélation maximale du décalage 3 /pondérée par les volumes atomiques de van der Waals. Il fait partie des descripteurs GETAWAY]. La matrice de corrélation (tableau 6) montre que les descripteurs choisis sont corrélés avec la variable à expliquer ($pIGC_{50}$) et ne sont pas corrélés entre eux. Toutes les données concernant les valeurs de $RGyr$, $R3v+$ et de l'activité biologique sont résumées dans le tableau 7.

Tableau 6. Matrice de corrélation des descripteurs du modèle

	$pIGC_{50}$	$RGyr$
$RGyr$	0,844	
$R3v+$	0,437	-0,060

L'équation du modèle optimal peut s'écrire:

$$pIGC_{50} = -3,602(\pm 0,174) + 1,439(\pm 0,069) RGyr + 16,342 (\pm 1,416) R3v+ \quad (62)$$

Tous les paramètres statistiques pertinents sont rapportés dans le tableau 8.

Les valeurs R^2 et R_{adj}^2 attestent des bonnes performances d'ajustement du modèle qui, de plus, est très fortement significatif (grande valeur du paramètre de Fisher, F).

Le modèle est robuste, la différence entre R^2 et Q^2 est faible (1%). La figure 10 montre un tracé des valeurs de $pIGC_{50}$ expérimentale et prédites par validation croisée LOO. La dispersion des points est faible bien qu'il y ait deux points un peu éloignés.

I. Toxicité des anilines : Approche QSAR

Le modèle montre une très bonne stabilité dans la validation interne (la différence entre Q_{LOO}^2 et $Q_{LMO/50}^2$ est d'environ 1%), tandis que le bootstrap confirme la prédictivité interne et la stabilité du modèle.

Pour calculer les valeurs des descripteurs $R3v+$, $RGyr$ avec $pIGC_{50}$, voir le tableau 7, ainsi que les valeurs statistiques pour l'ensemble de calibrage, voir le tableau 8.

Tableau 7. Valeurs de $R3v+$, $RGyr$ et de la concentration inhibitrice de la croissance $pIGC_{50}$ pour l'ensemble des 48 anilines. Les 17 premiers composés sont l'ensemble de validation

Composé	$pIGC_{50}$	$RGyr$	$R3v+$	Composé	$pIGC_{50}$	$RGyr$	$R3v+$
4-hexylaniline	2,04	3,434	0,023	3,4-dimethylaniline	-0,29	2,132	0,029
2,3-chloro - - -	1,02	2,169	0,087	3-ethyl - - -	-0,12	2,204	0,021
4-methyl - - -	-0,02	2,008	0,021	2-chloro - - -	-0,09	1,982	0,041
2,4,6-trichloro - - -	1,01	2,588	0,039	2,4-dimethyl - - -	-0,30	2,133	0,022
2-bromo - - -	0,46	2,013	0,056	2-ethyl - - -	-0,25	2,107	0,023
4-butyl - - -	1,05	2,998	0,022	3-fluoro - - -	0,04	1,932	0,025
2-chloro-6-methyl - - -	0,12	2,156	0,037	2-propyl - - -	0,06	2,414	0,023
2-phenyl - - -	0,86	2,680	0,019	3-chloro - - -	0,09	2,111	0,042
3-iodo - - -	0,61	2,158	0,071	2-isopropyl - - -	0,10	2,190	0,024
3,4,5-trichloro - - -	1,51	2,451	0,088	4-isopropyl - - -	0,21	2,403	0,024
4-ethyl - - -	0,04	2,286	0,021	2-chloro-5-methyl - - -	0,20	2,245	0,037
3-chloro-4-methyl - - -	0,45	2,208	0,049	4-octyl - - -	2,34	3,914	0,022
5-chloro-2-methyl - - -	0,51	2,300	0,03	2-iodo - - -	0,35	1,981	0,070
2,6-dichloro - - -	0,33	2,291	0,04	4-chloro-2-methyl - - -	0,35	2,298	0,033
3-phenyl - - -	0,78	2,815	0,021	3-chloro-2-methyl - - -	0,45	2,153	0,044
2,5-dichloro - - -	0,58	2,448	0,039	2,4-dichloro - - -	0,56	2,390	0,040
3,5-dichloro - - -	0,71	2,423	0,038	4-propyl - - -	0,49	2,611	0,024
2-methyl - - -	-0,55	1,892	0,024	3-bromo - - -	0,52	2,179	0,058
3-methyl - - -	-0,43	1,968	0,026	2,6-dichloro-3-methyl - - -	0,69	2,407	0,041
2,6-dimethyl - - -	-0,43	2,047	0,024	4-phenyl - - -	0,95	2,904	0,022
3,5-dimethyl - - -	-0,37	2,158	0,017	3,4-dichloro - - -	1,14	2,281	0,089
2,5-dimethyl - - -	-0,35	2,134	0,023	2,4,5-trichloro - - -	1,30	2,577	0,086
2-fluoro - - -	-0,31	1,846	0,025	2,3,4-trichloro - - -	1,35	2,405	0,087
2-chloro-4-methyl - - -	0,24	2,211	0,039	4-pentyl - - -	1,67	3,246	0,022

Tableau 8. Paramètres statistiques pour l'ensemble de calibrage.

n_{tr}	Q_{LOO}^2	R^2	$Q_{LMO/50}^2$	Q_{BOOT}^2	R_{adj}^2	$SDEC$	$SDEP$	S	F
31	93,85	94,99	92,34	92,48	94,64	0,151	0,168	0,159	265,64

I. Toxicité des anilines : Approche QSAR

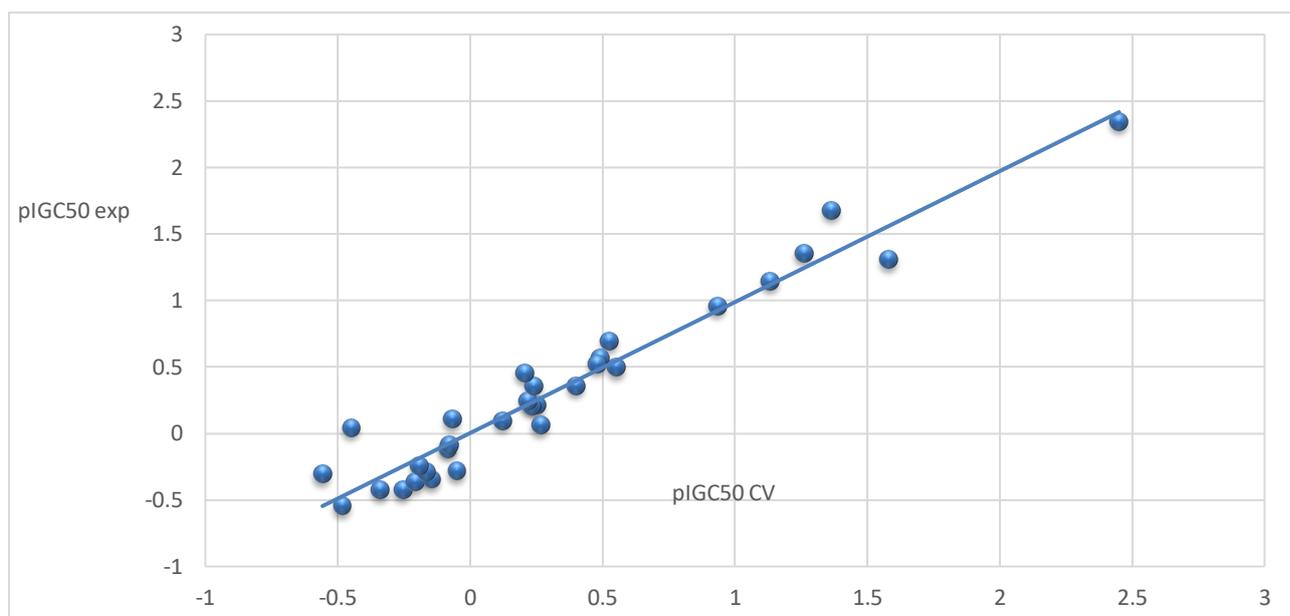


Figure 10 : Valeurs expérimentales et prédites par LOO de l'ensemble de calibrage

Les valeurs prédites expérimentales et calculées des deux ensembles (calibrage et validation) sont dans le tableau 9, en plus des valeurs des leviers et des résidus standardisés de prédictions.

Tableau 9. Valeurs de $pIGC_{50}$ expérimentales prédites et calculées, leviers et résidus standardisés de prédictions des 48 anilines

N°	Composés	$pIGC_{50exp}$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_i	e_{istd}
1	4-hexylaniline	2,04	2,04	-	1,7172	0,276
2	2,3-chloro aniline	1,02	1,02	-	0,9418	0,228
3	4-methyl aniline	-0,02	-0,02	-	-0,3685	0,069
4	2,4,6-trichloro aniline	1,01	1,01	-	0,7607	0,048
5	2-bromo aniline	0,46	0,46	-	0,2106	0,075
6	4-butyl aniline	1,05	1,05	-	1,0732	0,133
7	2-chloro-6-methyl aniline	0,12	0,12	-	0,1060	0,036
8	2-phenyl aniline	0,86	0,86	-	0,5663	0,078
9	3-iodo aniline	0,61	0,61	-	0,6645	0,126
10	3,4,5-trichloro aniline	1,51	1,51	-	1,3641	0,243
11	4-ethyl aniline	0,04	0,04	-	0,0317	0,050
12	3-chloro-4-methyl aniline	0,45	0,45	-	0,3770	0,046
13	5-chloro-2-methyl aniline	0,51	0,51	-	0,1990	0,035
14	2,6-dichloro aniline	0,33	0,33	-	0,3494	0,033
15	3-phenyl aniline	0,78	0,78	-	0,7934	0,095
16	2,5-dichloro aniline	0,58	0,58	-	0,5591	0,037
17	3,5-dichloro aniline	0,71	0,71	-	0,5068	0,035
18	2-methyl aniline	-0,55	-0,55	-0,4865	-0,481	0,079
19	3-methyl aniline	-0,43	-0,43	-0,3444	-0,3385	0,064
20	2,6-dimethyl aniline	-0,43	-0,43	-0,2633	-0,253	0,059

Tableau 9. Suite et fin

N°	Composés	$pIGC_{50exp}$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_i	e_{istd}
21	3,5-dimethyl aniline	-0,37	-0,37	-0,2179	-0,2069	0,067
22	2,5-dimethyl aniline	-0,35	-0,35	-0,1544	-0,1435	0,053
23	2-fluoro aniline	-0,31	-0,31	-0,5364	-0,5575	0,085
24	2-chloro-4-methyl aniline	0,24	0,24	0,2179	0,2171	0,034
25	3,4-dimethylaniline	-0,29	-0,29	-0,0593	-0,0489	0,043
26	3-ethyl aniline	-0,12	-0,12	-0,0863	-0,0844	0,054
27	2-chloro aniline	-0,09	-0,09	-0,0791	-0,0785	0,053
28	2,4-dimethyl aniline	-0,30	-0,30	-0,1722	-0,1647	0,055
29	2-ethyl aniline	-0,25	-0,25	-0,1933	-0,1900	0,055
30	3-fluoro aniline	0,04	0,04	-0,4126	-0,4472	0,071
31	2-propyl aniline	0,06	0,06	0,2487	0,2582	0,048
32	3-chloro aniline	0,09	0,09	0,1229	0,1244	0,041
33	2-isopropyl aniline	0,10	0,10	-0,0575	-0,0653	0,048
34	4-isopropyl aniline	0,21	0,21	0,2492	0,2511	0,046
35	2-chloro-5-methyl aniline	0,20	0,20	0,2342	0,2353	0,033
36	4-octyl aniline	2,34	2,34	2,3920	2,4519	0,535
37	2-iodo aniline	0,35	0,35	0,3933	0,4001	0,137
38	4-chloro-2-methyl aniline	0,35	0,35	0,2451	0,2415	0,033
39	3-chloro-2-methyl aniline	0,45	0,45	0,2161	0,2061	0,041
40	2,4-dichloro aniline	0,56	0,56	0,4919	0,4895	0,035
41	4-propyl aniline	0,49	0,49	0,5487	0,5525	0,061
42	3-bromo aniline	0,52	0,52	0,4823	0,4794	0,071
43	2,6-dichloro-3-methyl aniline	0,69	0,69	0,5328	0,5269	0,036
44	4-phenyl aniline	0,95	0,95	0,9378	0,9363	0,113
45	3,4-dichloro aniline	1,14	1,14	1,1357	1,1342	0,250
46	2,4,5-trichloro aniline	1,30	1,30	1,5128	1,5827	0,247
47	2,3,4-trichloro aniline	1,35	1,35	1,2815	1,2600	0,239
48	4-pentyl aniline	1,67	1,67	1,4302	1,3656	0,212

I-2-2 Qualité externe du modèle QSAR:

$SDEP_{ext}$ est un peu différent de $SDEP$ (Tableau 10) ; le modèle fonctionne un peu moins bien dans la prédiction externe que dans la prédiction interne.

Tableau 10. Paramètres statistiques pour l'ensemble de validation

n_{ext}	Q_{ext}^2	$SDEP_{ext}$
17	92,13	0,184

La validation externe supplémentaire selon (Golbraikh et Tropsha 2002) décrite précédemment confirme la validité du modèle qu'on propose. Les résultats sont comme suit:

- 1) $Q_{EXT}^2 = 0,9141 > 0,5$
- 2) $R^2 = 0,9264 > 0,6$
- 3) $(R^2 - R_0^2)/R^2 = -0,0396 < 0,1$ et $0,85 < k = 1,1249 < 1,15$
 $(R^2 - R_0'^2)/R^2 = -0,0151 < 0,1$ et $0,85 < k' = 0,8576 < 1,15$

I-2-3 Qualité d'ajustement du modèle QSAR:

Le graphique représentant les valeurs prédites, calculées en fonction des expérimentales (figure 11) est caractérisé par une faible dispersion autour de la droite ce qui indique un bon ajustement.

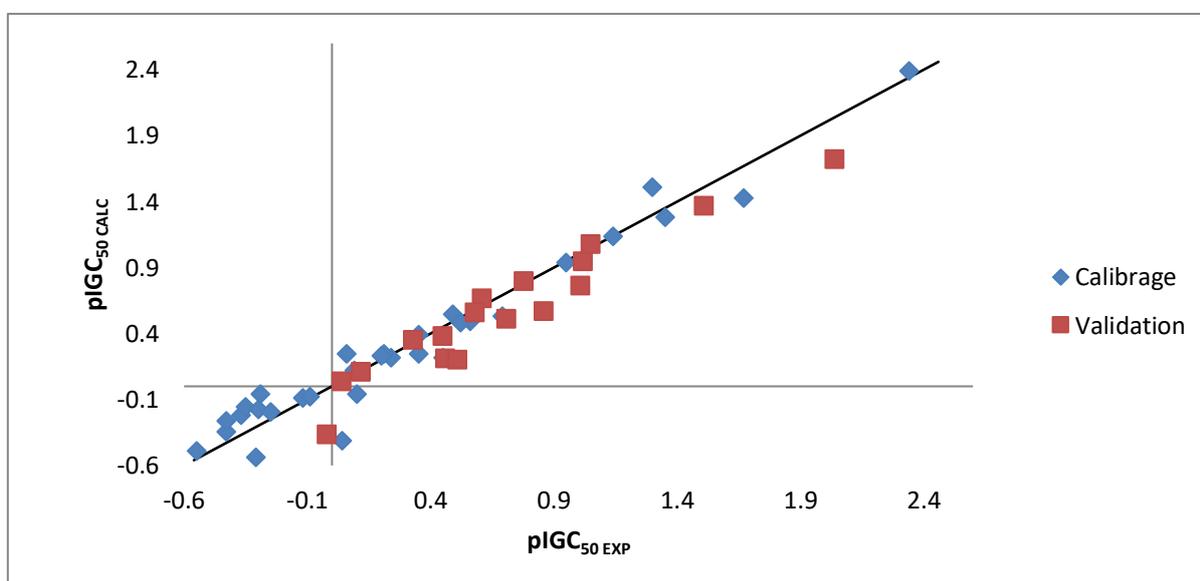


Figure 11: Droite d'ajustement du modèle QSAR.

I-2-4 Domaine d'application:

Comme le montre le diagramme de Williams (figure 12), le seul composé chimique avec un effet de levier élevé ($h_i > h^* = 0,29$) de l'ensemble de calibrage (4-octylaniline) est parfaitement prédit, comme cela se produit normalement pour les produits chimiques influents de l'ensemble de calibrage. Une seule valeur aberrante est observée (3-fluoroaniline) qui est caractérisé par son résidu normalisé supérieur à trois unités d'écart-type (3σ).

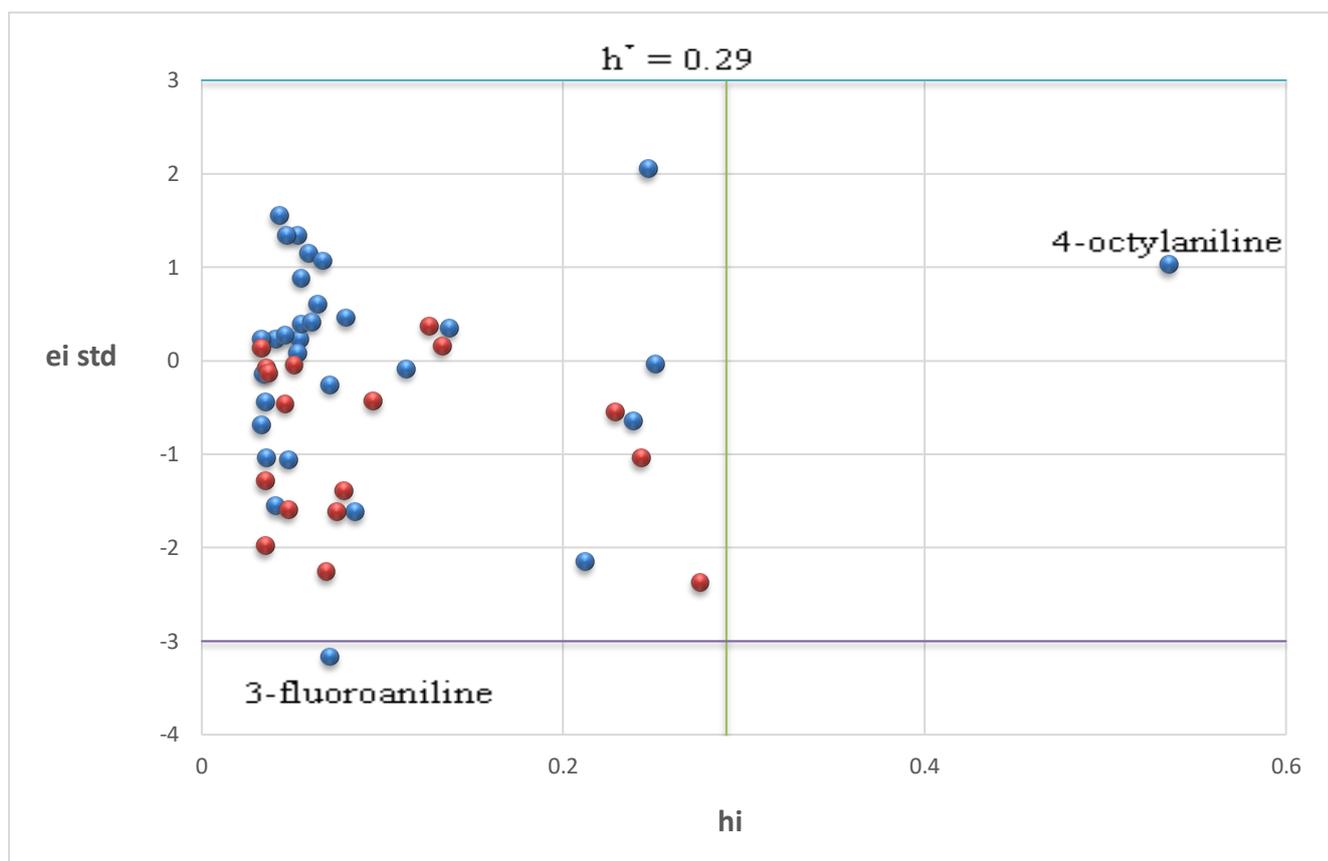


Figure 12 : Diagramme de Williams du modèle QSAR

I-2-5 Test de randomisation :

Afin de mettre en évidence l'existence de corrélations fortuites, le test de randomisation (permutations de Y) [107] a été adopté. Ce test consiste à construire un vecteur de propriétés dont les composantes sont les composantes du vecteur de la propriété étudiée (dans ce cas $pIGC_{50}$), mais permutées aléatoirement dans leurs positions. Ce nouveau vecteur d'activité est utilisé comme s'il s'agissait d'un vecteur expérimental, et un modèle QSAR est calculé de la manière habituelle. Ce processus a été répété 300 fois, afin de tester la relations structure / activité établie.

Comme on peut l'observer (figure 13), les réponses permutées donnent des modèles prédictifs médiocres, tous ayant $Q^2 < 20$. D'autre part, les paramètres statistiques des $pIGC_{50}$ correctement ordonnés donnent de bons paramètres statistiques, et par conséquent, le modèle est isolé en haut du graphique.

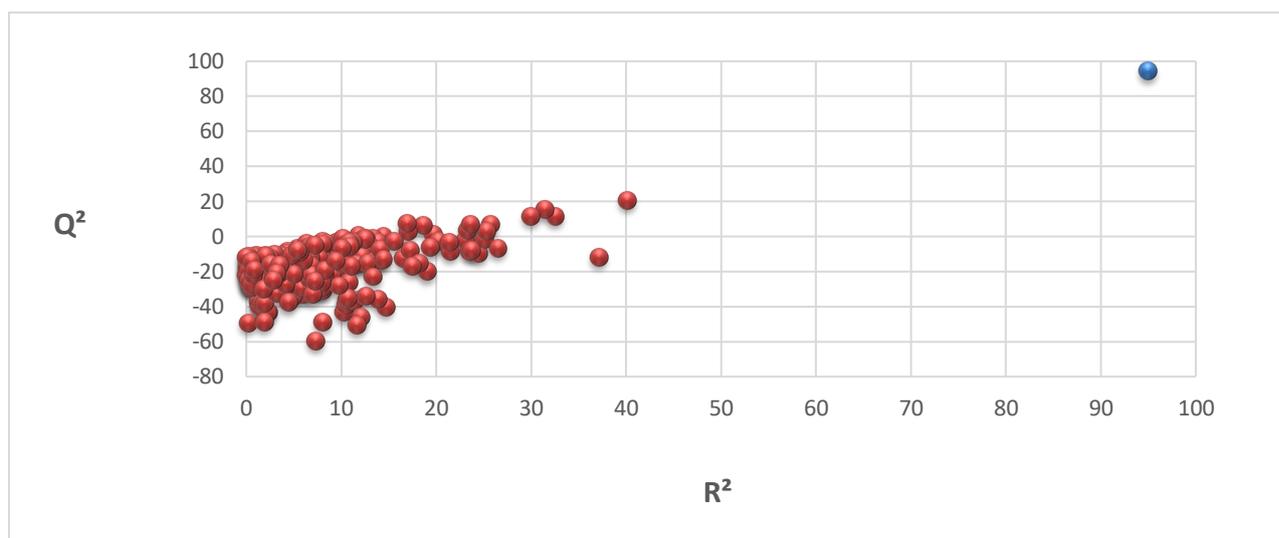


Figure 13: Teste de randomisation du modèle QSAR. (Les croix représentent les activités ordonnées au hasard et le carré correspond aux activités réelles).

I-3 Interprétation mécanistique du modèle:

En interprétant les descripteurs dans le modèle proposé, il est possible d'avoir un aperçu des facteurs qui sont probablement liés à l'inhibition de la croissance microbienne. Sur les deux descripteurs, un est un descripteur GETAWAY ($R3v+$) et l'autre est géométrique ($RGyr$).

Les descripteurs GETAWAY R qui sont représentés par Rkw ont été calculés comme suit. La matrice d'influence moléculaire dénotée H ressemble à la matrice chapeau (ou d'influence) définie dans les diagnostics de régression [108].

La valeur de H a été calculée à partir de la matrice moléculaire M (M a A lignes correspondant aux coordonnées cartésiennes x, y, z de chaque atome dans la structure moléculaire optimisée) comme suit :

$$H = M(M^T M)^{-1} M^T \quad (63)$$

où l'exposant T se réfère à la matrice transposée. La contribution maximale à l'autocorrélation à chaque retard représenté par $Rkw+$ est défini comme :

$$Rkw+ = \max_{ij} \left[\frac{\sqrt{h_{ii} h_{jj}}}{r_{ij}} w_i w_j \delta(k, d_{ij}) \right] \quad i \neq j \text{ et } k=1,2,\dots,8 \quad (64)$$

Où $Rkw +$ est l'indice R maximal pondéré de kième ordre, r_{ij} sont les distances géométriques entre chaque paire d'atomes i et j, d_{ij} est le diamètre topologique, h_{ii} et h_{jj} sont les termes diagonaux de la matrice H et δ est une fonction delta de Dirac définie comme :

$$\delta(k, d_{ij}) = \begin{cases} 1 & \text{si } d_{ij} = k \\ 0 & \text{si } d_{ij} \neq k \end{cases} \quad (65)$$

$R3v +$ décrit la taille et la forme des molécules. Comme nous le savons la taille, la forme et la symétrie des molécules jouent un rôle clé dans le processus de distribution de la molécule entre deux phases liquides immiscibles. En même temps, ce descripteur indique le rôle du volume (v) dans la variation de l'activité. La taille moyenne est une propriété simple et très significative d'une molécule [109]. Facilement obtenue à partir d'expériences de diffusion de la lumière, une mesure commune de la taille moyenne telle que le rayon de giration ($RGyr$) fournit des informations précieuses sur l'interaction de la molécule avec son milieu environnant ou sa cible.

I-4 Comparaison avec le modèle original:

Schultz *et al.* [99] ont évalué la toxicité relative de 48 anilines sélectionnées en utilisant les caractéristiques de croissance de la population de *Tetrahymena pyriformis* (concentration causant 50% d'inhibition de la croissance) comme critère. Ces auteurs ont montré qu'une simple corrélation de $pIGC_{50}$ ($= \log IGC_{50}^{-1}$) par rapport à au coefficient de partage octanol-eau ($\log P$) peut modéliser la toxicité environnementale. La prédictivité de ce modèle QSAR dépendant du $\log P$ peut être améliorée en ajoutant $\sum \sigma$ (la somme des paramètres électroniques de substitution σ), en tant que second descripteur orthogonal.

Pour 3 composés (4-hexylaniline, 4-octylaniline et 4-pentylaniline) les valeurs de $\log P$ ont été tirées du livre de Hansch et Leo [110] d'où ont été puisées les valeurs des $\sum \sigma$ et pour le 2-phenylaniline le logiciel CLOGP version 3.34 [111] a été utilisé. Toutes les autres valeurs de $\log P$ ont été calculées en utilisant des valeurs tabulées dans le travail de Norrington *et al.* [112].

Les paramètres statistiques rapportés par les auteurs sont uniquement liés aux performances d'ajustement et sont les suivant:

I. Toxicité des anilines : Approche QSAR

$$r^2 = 0,887, \quad s = 0,226, \quad F=177,41$$

Pour l'équation de régression paramétrée pour les 48 anilines:

$$pIGC_{50} = -1,373 + 0,251 \sum \sigma + 0,339 \log K_{ow} \quad (66)$$

Si on l'établi sur les 48 anilines donne les résultats suivant:

$$r^2 = 0,941, \quad s = 0,164, \quad F=356,88$$

En utilisant le même ensemble de calibrage que précédemment, nous avons calculé un modèle sur les descripteurs moléculaires sélectionnés par Schultz et al. Il a pour équation:

$$pIGC_{50} = -1,404 (\pm 0,113) - 0,4848 (\pm 0,123) \sum \sigma + 0,727 (\pm 0,047) \log K_{ow} \quad (67)$$

Les paramètres d'ajustement et de prédiction correspondants indiqués ci-dessous dans le tableau 11 montrent que notre modèle est légèrement meilleur que celui basé sur l'approche de Schultz *et al.*

Tableau 11. Paramètres statistiques du modèle de l'approche Schultz *et al.*

n_{tr}	n_{ext}	Q_{LOO}^2	R^2	$Q_{LMO/50}^2$	Q_{BOOT}^2	R_{adj}^2
31	17	89,24	91,62	85,38	86,06	91,03
Q_{ext}^2	<i>SDEC</i>	<i>SDEP</i>	<i>SDEP</i> _{ext}	s	F	
81,28	0,196	0,222	0,293	0,206	153,1523	

Les valeurs prédites et calculées de $pIGC_{50}$ des deux ensembles (calibrage et validation), en plus des valeurs des leviers et des résidus standardisés de prédictions sont dans le tableau 12 pour le modèle de l'approche de Schultz *et al.* Les valeurs de $\sum \sigma$ et $\log K_{ow}$ y sont aussi réunis. La numérotation correspond à celle du tableau 9 de notre modèle.

Comme dans le cas de notre modèle ; la validation externe selon Tropsha *et al.* a été appliquée et confirme la validité du modèle basé sur les descripteurs choisi par Schultz *et al.* Les résultats sont comme suit:

- 1) $Q_{EXT}^2 = 0,7948 > 0,5$
- 2) $R^2 = 0,7852 > 0,6$
- 3) $(R^2 - R_0^2)/R^2 = -0,2696 < 0,1$ et $0,85 < k = 0,9610 < 1,15$
 $(R^2 - R_0'^2)/R^2 = -0,2525 < 0,1$ et $0,85 < k' = 0,9248 < 1,15$

Si l'on établissait notre modèle (Basé sur $R3v$ + et $RGyr$) en utilisant les 48 anilines (c.-à-d. sans validation externe) les statistiques seraient.

I. Toxicité des anilines : Approche QSAR

On remarque que les valeurs de Q_{EXT}^2 et R^2 sont très inférieures à celle trouvées pour notre modèle 0,9141 et 0,9264 respectivement

Tableau 12. Valeurs des descripteurs selon Schultz *et al.* [99], valeurs prédites et calculées de $pIGC_{50}$, valeurs des leviers et des résidus standardisés de prédictions

N°	$\log K_{ow}$	$\Sigma \sigma$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_i	e_{istd}
1	4,210	-0,150	2,040	-	1,6226	0,2780
2	2,630	0,600	1,020	-	0,7740	0,1260
3	1,380	-0,170	-0,020	-	-0,4907	0,0810
4	3,210	0,690	1,010	-	1,2445	0,1670
5	1,790	0,230	0,460	-	-0,0117	0,0650
6	2,880	-0,160	1,050	-	0,6292	0,0780
7	2,080	0,060	0,120	-	0,1300	0,0350
8	2,840	-0,010	0,860	-	0,6647	0,0520
9	2,370	0,350	0,610	-	0,4719	0,0600
10	3,910	0,970	1,510	-	1,8870	0,3380
11	1,370	0,340	0,040	-	-0,2761	0,1310
12	2,430	0,200	0,450	-	0,4512	0,0380
13	2,430	0,200	0,510	-	0,4512	0,0380
14	2,290	0,460	0,330	-	0,4603	0,0890
15	2,670	0,060	0,780	-	0,5688	0,0390
16	2,630	0,600	0,580	-	0,7740	0,1260
17	2,980	0,740	0,710	-	1,0953	0,1810
18	1,390	-0,170	-0,550	-0,4832	-0,4773	0,0810
19	1,400	-0,070	-0,430	-0,4323	-0,4325	0,0750
20	1,880	-0,340	-0,430	-0,1928	-0,1700	0,0880
21	1,900	-0,140	-0,370	-0,0909	-0,0762	0,0500
22	1,890	-0,240	-0,350	-0,1419	-0,1273	0,0650
23	1,150	0,060	-0,310	-0,5616	-0,5921	0,1080
24	2,070	0,060	0,240	0,1226	0,1182	0,0360
25	1,880	-0,240	-0,290	-0,1493	-0,1395	0,0660
26	1,840	-0,070	-0,120	-0,1051	-0,1044	0,0460
27	1,590	0,230	-0,090	-0,1604	-0,1668	0,0830
28	1,960	-0,340	-0,300	-0,1333	-0,1174	0,0870
29	1,890	-0,150	-0,250	-0,1027	-0,0947	0,0510
30	1,880	-0,150	0,040	-0,1101	-0,1183	0,0520
31	2,430	-0,130	0,060	0,3076	0,3203	0,0490
32	1,940	0,370	0,090	0,1608	0,1674	0,0850
33	2,270	-0,150	0,100	0,1799	0,1840	0,0490
34	2,260	-0,150	0,210	0,1725	0,1705	0,0480
35	2,090	0,140	0,200	0,1722	0,1711	0,0390
36	5,270	-0,150	2,340	2,4109	2,5143	0,5930
37	2,090	0,180	0,350	0,1897	0,1826	0,0420
38	2,320	0,060	0,350	0,3085	0,3071	0,0320
39	2,420	0,200	0,450	0,4438	0,4435	0,0380

Tableau 12. Suite et fin

40	2,520	0,460	0,560	0,6313	0,6379	0,0850
41	2,380	-0,130	0,490	0,2704	0,2594	0,0480
42	2,070	0,390	0,520	0,2662	0,2435	0,0820
43	2,780	0,390	0,690	0,7942	0,8020	0,0700
44	2,640	-0,010	0,950	0,5160	0,4968	0,0420
45	2,890	0,600	1,140	0,9674	0,9418	0,1290
46	3,560	0,830	1,300	1,5658	1,6543	0,2500
47	3,560	0,830	1,350	1,5658	1,6377	0,2500
48	3,680	-0,150	1,670	1,2285	1,1320	0,1790

La qualité de l'ajustement du modèle basé sur l'approche de Shultz *et al.* représenté dans la figure 14 est assez bonne mais inférieure à celle de notre modèle (figure 11). On remarque une augmentation de la dispersion des points autour de la droite.

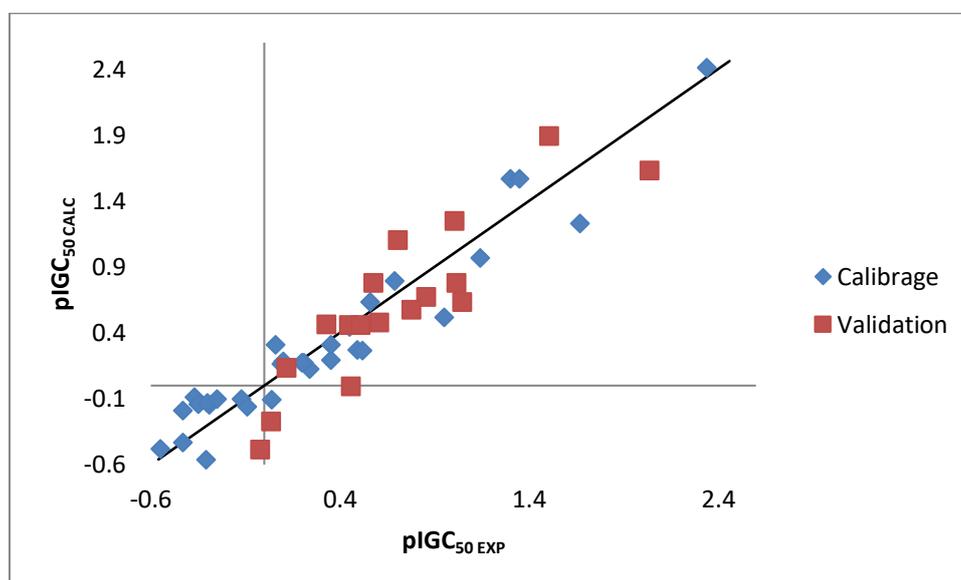


Figure 14 : Droite d'ajustement du modèle basé sur l'approche de Schultz *et al.*

Concernant le domaine d'application du modèle basé sur les descripteurs de Schultz *et al.* On remarque que le composé de l'ensemble de calibrage (N°36 : 4-octylaniline) signalé influant précédemment pour notre modèle (figure 12) l'est aussi pour l'approche Schultz *et al.* mais dans son influence est plus grande (le h_i passant de 0,535 à 0,593). Aucun composé aberrant n'est à remarquer mais le 4-pentylaniline est à la limite des 3σ ($e_{istd} = -2,9692$).

Le 3, 4,5-trichloroaniline (N° 10) est un composé influant de l'ensemble de validation ce qui rend sa prédiction douteuse car en dehors du domaine d'application.

I. Toxicité des anilines : Approche QSAR

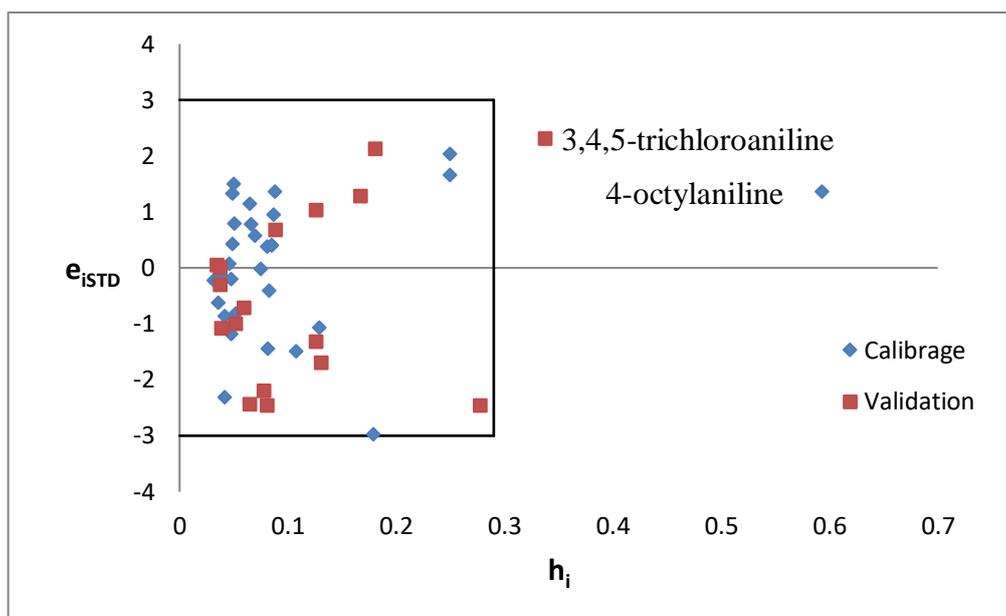


Figure 15: Diagramme de Williams du modèle QSAR basé sur l'approche de Schultz *et al.*

Le modèle obtenu et discuté dans cette partie est meilleur que celui de Schultz *et al.* car ce dernier présente les défauts suivants :

- Le modèle de Schultz *et al.* établie sur les 48 anilines est dépourvu de validation statistique externe et les statistiques d'ajustement rapportés ne suffisent pas à juger de sa qualité.
- Les valeurs de K_{ow} (pour la plupart) ne sont pas des valeurs expérimentales et comme déjà remarqué, sont de trois sources différentes ce qui est à éviter lors de la construction d'un modèles QSAR.
- Une fois reparamétré sur les mêmes ensembles de validation et de calibrage les statistiques ne sont pas très bonnes. R^2 et Q_{EXT}^2 de la validation selon Tropsha *et al.* en sont des exemples pertinents.
- Le 3,4,5-trichloroaniline de l'ensemble de validation, en dehors du domaine d'application (figure 15) est une limitation pour l'utilisation du modèle.
- Les trois points précédemment cités remettent en cause l'utilisation de $\log K_{ow}$ et $\sum \sigma$ comme meilleur choix des descripteurs pour la prédiction de la toxicité vis-à-vis de *Tetrahymena pyriformis* pour les anilines.

II-1 Collecte de données et méthodologie :

L'ensemble de données relatives aux phénols (tiré de Schultz *et al* [99] comme pour les anilines précédemment traités) a été divisé au hasard en un ensemble d'apprentissage (62 objets), utilisé pour développer le modèle QSAR, et un ensemble de validation (33 objets), utilisé uniquement pour la validation statistique externe (tableau 13).

Les structures de toutes les molécules ont été pré-optimisées à l'aide du champ de force MM⁺ de la mécanique moléculaire (algorithme Polak-Ribiere) en utilisant le programme HyperChem 6.03 [102]. Les géométries finales d'énergie conformationnelle minimale ont été obtenues par la méthode semi-empirique AM1. Les géométries ainsi optimisées ont été transférées dans le logiciel DRAGON [103] pour calculer 1664 descripteurs. Les descripteurs avec des valeurs constantes ou quasi constantes dans chaque groupe ont été rejetés. Pour chaque paire de descripteurs corrélés (avec un coefficient de corrélation $r \geq 0,95$), celui présentant la plus forte corrélation des paires avec les autres descripteurs a été exclu.

La structure chimique de chaque composé a été introduite sur un PC en utilisant le programme E-CALC [113] pour calculer les indices d'état électrotopologique (E-state indices) de Kier et Hall [114, 115].

La sélection des variables a été effectuée sur l'ensemble d'apprentissage, en utilisant l'algorithme génétique (AG) dans la version de MobyDigs de Todeschini [105] en maximisant la variance expliquée par validation croisée par omission d'une observation Q_{LOO}^2 . en utilisant la régression par les moindres carrés ordinaires pour la sélection de sous-ensembles. Dans le logiciel MobyDigs les processus de croisement et de mutation de l'algorithme génétique sont contrôlés par un paramètre T variant de 0 à 1. Les paramètres de l'algorithme génétique ont été fixés comme suit : population des modèles Pop = 100 ; valeur de T fixée à 0,5 pour équilibrer les rôles des deux processus de croisement et de mutation.

II-2 Présentation et discussion du modèle QSAR :

II-2-1 Qualités internes du modèle QSAR :

Parmi plus de 100 modèles simples avec deux variables explicatives, nous avons choisi le modèle avec la meilleure valeur du paramètre de prédiction Q_{LOO}^2 . Ce dernier a été construit en utilisant : $\log_{k_{ow}}$ et s-CH3. Les valeurs de ces descripteurs sont aussi résumées dans le tableau 13.

Tableau 13. Valeurs de $\log K_{ow}$, $S - CH_3$ et de la concentration inhibitrice de la croissance $pIGC_{50}$ pour l'ensemble des 95 phénols. Les 33 derniers composés sont l'ensemble de validation.

Composé	$pIGC_{50}$	$\log K_{ow}$	$S - CH_3$	Composé	$pIGC_{50}$	$\log K_{ow}$	$S - CH_3$
3-hydroxybenzylalcohol	-1,04	0,44	0	4bromo-2,6-diméthylphénol	1,28	3,93	3,7641
4-hydroxyphenethyl alcool	-0,83	0,67	0	4-cyclopentylphénol	1,29	3,69	0
4-acetoamidophenol	-0,82	0,49	1,4374	3,5-dichlorophénol	1,56	3,35	0
3-hydroxyphénol	-0,65	0,81	0	4-chloro-2-isopropylphénol	1,86	4,71	4,0385
3-aminophénol	-0,54	0,25	0	2, 4,6-tribromophénol	2,05	4,02	0
Phénol	-0,43	1,48	0	4-(tert)octylphénol	2,09	5,31	6,8137
4-hydroxybenzylcyanide	-0,38	0,9	0	2, 4,5-trichlorophénol	2,1	3,85	0
2-hydroxybenzaldoxime	-0,25	1,1	0	2,6-diphénylphénol	2,11	5,25	0
4-méthoxyphénol	-0,14	1,57	1,5897	4-hydroxybenzamide	-0,78	0,33	0
4-acétylphénol	-0,09	1,45	1,4924	2-méthylphénol	-0,27	2,12	1,8704
3-méthylphénol	-0,06	2,12	1,9442	3,4-diméthylphénol	0,12	2,77	4,0038
2,5-diméthylphénol	0,01	2,77	3,8412	2-hydroxybenzaldehyde	0,48	2,07	0
4-hydroxypropiophenone	0,06	1,98	1,8147	2-chloro-5-méthylphénol	0,64	2,85	1,8989
2-acétylphénol	0,08	2,08	1,4256	3-(tert)butylphénol	0,73	3,45	6,3803
methyl-4-hydroxybenzoate	0,08	1,98	1,3147	2-isopropylphénol	0,8	3,05	4,1291
3,5-diméthylphenol	0,11	2,77	3,93	3, 4,5-triméthylphénol	0,93	3,42	6,0637
2-éthylphénol	0,18	2,65	2,0229	4-(tert)pentylphénol	1,23	3,98	6,6246
4-éthylphénol	0,21	2,65	2,0897	6-(tert) butyl-2,4-diméthylphénol	1,25	4,75	10,3354
2-chlorophénol	0,28	2,2	0	4-bromo-6-chloro-2-méthylphénol	1,28	3,87	1,7976
2,3,5-triméthylphénol	0,36	3,33	5,9011	2-(ter) butyl-4-méthylphénol	1,3	4,1	8,3325
2,6-difluorophénol	0,4	1,65	0	4-hexyloxyphénol	1,65	4,22	2,1953
4-isopropylphénol	0,47	3,05	4,2627	2,6-di (tert) butyl-4-méthylphénol	1,79	6,08	14,8844
2-bromophénol	0,5	2,35	0	4-heptyloxyphénol	2,03	4,75	2,2152
3-nitrophenol	0,51	1,85	0	2-hydroxybenzamide	-0,24	0,96	0
ethyl-4-hydroxybenzoate	0,57	2,51	1,7483	3-méthylhydroxybenzoate	-0,05	1,95	1,298
$\alpha\alpha\alpha$ -trifluoro-4-crésol	0,62	2,88	0	4-fluorophénol	0,02	1,91	0
4-propylphénol	0,64	3,18	2,1457	2,3-diméthylphénol	0,12	2,77	3,8886
4-bromophénol	0,68	2,63	0	4-hydroxybenzaldehyde	0,27	1,44	0
4-chloro-2-méthylphénol	0,7	3,13	1,8068	2-fluorophénol	0,28	1,63	0
2-bromo-4-méthylphénol	0,79	2,91	1,9764	3-éthylhydroxybenzoate	0,48	2,51	1,7368
3-chloro-4-fluorophénol	0,84	2,78	0	4-cyanophénol	0,52	1,6	0
4-(tert)butylphénol	0,91	3,45	6,4565	2-nitrophenol	0,67	1,85	0
3-chlorophénol	0,96	2,48	0	4-chloro-3-méthylphénol	0,8	3,13	1,8493
4-(sec)butylphénol	0,98	3,58	4,3511	3-iodophénol	1,12	2,89	0
4-benzyloxyphénol	1,04	3,14	0	3-phénylphénol	1,35	3,36	0
4-chloro-3,5-diméthylphénol	1,2	3,78	3,7402	2,4-bromophénol	1,4	3,37	0
4-hydroxyphenylmethane	1,2	3,69	1,9856	2, 4,6-trichlorophénol	1,7	3,69	0
2-(tert)butylphénol	1,24	3,45	6,2561	4-bromo-2,6-dichlorophénol	1,78	3,84	0
2,3-dichlorophénol	1,27	3,07	0	2-hydroxybenzyl alcool	-0,95	0,44	0

Tableau 13 suite et fin

Composé	$pIGC_{50}$	$\log K_{ow}$	$S - CH_3$	Composé	$pIGC_{50}$	$\log K_{ow}$	$S - CH_3$
3-acétylphénol	-0,38	1,45	1,467	4-chlorophénol	0,55	2,48	0
3-méthoxyphénol	-0,14	1,57	1,5643	3-isopropylphénol	0,61	3,05	4,2119
3-cyanophénol	-0,07	1,6	0	4-butoxyphénol	0,7	3,16	2,1252
4-éthoxyphénol	0,01	2,1	1,9235	4-iodophénol	0,85	2,89	0
2-cyanophénol	0,03	1,6	0	4-hydroxybenzophenone	1,02	3,08	0
3-hydroxybenzaldéhyde	0,08	1,44	0	2-phénylphénol	1,09	3,36	0
3-éthylphénol	0,23	2,65	2,0643	2,5-dichlorophénol	1,13	3,07	0
2-allylphénol	0,35	2,64	0	4-phénylphénol	1,38	3,36	0
3-fluorophénol	0,47	1,91	0				

L'équation du modèle optimal peut s'écrire comme suit :

$$pIGC_{50} = -1,136(\pm 0,078) + 0,701(\pm 0,031) \log K_{ow} - 0,089(\pm 0,013) S - CH_3 \quad (68)$$

Ici : $\log K_{ow}$ est tiré de la littérature [99] et $s-CH_3$ est calculé avec le logiciel E-CALC.

Le tableau 14 résume le teste de corrélation concernant les variables du modèle. Aucune anomalie n'est à signalée.

Tableau 14 Matrice de corrélation des descripteurs du modèle

	$pIGC_{50}$	$\log K_{ow}$
$\log K_{ow}$	0,919	
$S - CH_3$	0,345	0,602

Les paramètres statistiques pertinents sont rapportés dans le tableau ci-dessous :

Tableau 15 Paramètres statistiques pour l'ensemble de calibrage (Phénols)

n_{tr}	R^2	Q_{LOO}^2	$Q_{LMO/50}^2$	Q_{BOOT}^2	R_{adj}^2	$SDEC$	$SDEP$	S	F
62	91,15	90,16	89,57	89,43	90,85	0,238	0,251	0,244	303,81

Les valeurs R^2 et R_{adj}^2 attestent des bonnes performances d'ajustement du modèle qui, de plus, est très fortement significatif (grande valeur du paramètre de Fisher F).

Le modèle est robuste, la différence entre R^2 et Q^2 est faible. Le modèle montre une très bonne stabilité dans la validation interne (la différence entre Q_{LOO}^2 et $Q_{LMO/50}^2$ est d'environ 1%), tandis que le bootstrap confirme la prédictive interne et la stabilité du modèle.

II. Toxicité des phénols : Approche QSAR

Les valeurs prédites expérimentales et calculées des deux ensembles (calibrage et validation) sont dans le tableau 16, en plus des valeurs des leviers et des résidus standardisés de prédictions.

Tableau 16. Valeurs de $pIGC_{50}$ expérimentales prédites et calculées, leviers et résidus standardisés de prédictions des 95 phénols

N°	Composés	$pIGC_{50exp}$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_i	e_{istd}
1	3-hydroxybenzylalcool	-1,04	-0,8101	-0,8274	0,075	0,9796
2	4-hydroxyphenethyl alcool	-0,83	-0,655	-0,6662	0,063	0,7408
3	4-acetoamidophenol	-0,82	-0,9294	-0,92	0,086	-0,4688
4	3-hydroxyphénol	-0,65	-0,563	-0,568	0,057	0,3671
5	3-aminophénol	-0,54	-1,0004	-0,9607	0,086	-1,9736
6	Phénol	-0,43	-0,0861	-0,0982	0,035	1,4344
7	4-hydroxybenzylcyanide	-0,38	-0,5119	-0,5049	0,053	-0,5555
8	2-hydroxybenzaldoxime	-0,25	-0,3701	-0,3646	0,046	-0,5039
9	4-méthoxyphénol	-0,14	-0,1775	-0,1762	0,034	-0,1565
10	4-acétylphénol	-0,09	-0,2581	-0,2517	0,038	-0,7023
11	3-méthylphénol	-0,06	0,1833	0,1779	0,022	1,008
12	2,5-diméthylphénol	0,01	0,476	0,4653	0,023	1,9316
13	4-hydroxypropiophenone	0,06	0,0921	0,0913	0,024	0,133
14	2-acétylphénol	0,08	0,1985	0,196	0,022	0,491
15	methyl-4-hydroxybenzoate	0,08	0,137	0,1357	0,023	0,2364
16	3,5-diméthylphenol	0,11	0,4658	0,4574	0,024	1,4757
17	2-éthylphénol	0,18	0,5487	0,5426	0,016	1,5233
18	4-éthylphénol	0,21	0,5421	0,5367	0,016	1,3723
19	2-chlorophénol	0,28	0,4102	0,4067	0,027	0,5409
20	2,3,5-triméthylphénol	0,36	0,6887	0,675	0,042	1,376
21	2,6-difluorophénol	0,40	0,0087	0,021	0,032	-1,6296
22	4-isopropylphénol	0,47	0,6279	0,6242	0,024	0,6549
23	2-bromophénol	0,50	0,5122	0,5119	0,027	0,0507
24	3-nitrophenol	0,51	0,1509	0,1613	0,029	-1,4931
25	ethyl-4-hydroxybenzoate	0,57	0,467	0,4688	0,017	-0,4256
26	aaa-trifluoro-4-crésol	0,62	0,8928	0,8835	0,034	1,1377
27	4-propylphénol	0,64	0,9084	0,9033	0,019	1,1106
28	4-bromophénol	0,68	0,7091	0,7082	0,03	0,121
29	4-chloro-2-méthylphénol	0,70	0,9024	0,8984	0,02	0,838
30	2-bromo-4-méthylphénol	0,79	0,728	0,729	0,017	-0,2563
31	3-chloro-4-fluorophénol	0,84	0,8125	0,8134	0,032	-0,1145
32	4-(tert) butylphénol	0,91	0,6993	0,7099	0,05	-0,8856
33	3-chlorophénol	0,96	0,5927	0,603	0,028	-1,5268

II. Toxicité des phénols : Approche QSAR

Tableau 16. Suite

N°	Composés	$pIGC_{50exp}$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_i	e_{istd}
34	4-(sec) butylphénol	0,98	0,9882	0,988	0,024	0,0339
35	4-benzyloxyphénol	1,04	1,0669	1,0658	0,041	0,1127
36	4-chloro-3,5-diméthylphénol	1,20	1,182	1,1825	0,025	-0,0747
37	4-hydroxyphenylmethane	1,20	1,2776	1,2752	0,031	0,323
38	2-(tert) butylphénol	1,24	0,7028	0,7276	0,046	-2,254
39	2,3-dichlorophénol	1,27	1,0065	1,0167	0,039	-1,1012
40	4bromo-2,6-diméthylphénol	1,28	1,2857	1,2855	0,028	0,0236
41	4-cyclopentylphénol	1,29	1,4621	1,4515	0,062	0,7281
42	3,5-dichlorophénol	1,56	1,1957	1,2131	0,048	-1,5296
43	4-chloro-2-isopropylphénol	1,86	1,8051	1,8081	0,054	-0,2313
44	2, 4,6-tribromophénol	2,05	1,6515	1,6829	0,079	-1,7014
45	4-(tert) octylphénol	2,09	1,973	1,9824	0,08	-0,5001
46	2, 4,5-trichlorophénol	2,1	1,5236	1,5637	0,07	-2,4487
47	2,6-diphénylphénol	2,11	2,6362	2,5454	0,173	2,3701
48	4-hydroxybenzamide	-0,78	-0,9156	-0,9046	0,082	-0,5799
49	2-méthylphénol	-0,27	0,1946	0,1845	0,022	1,9247
50	3,4-diméthylphénol	0,12	0,4591	0,4508	0,024	1,4069
51	2-hydroxybenzaldehyde	0,48	0,3109	0,3155	0,027	-0,7024
52	2-chloro-5-méthylphénol	0,64	0,6948	0,6939	0,017	0,2264
53	3-(tert) butylphénol	0,73	0,7159	0,7166	0,048	-0,0591
54	2-isopropylphénol	0,8	0,6323	0,636	0,023	-0,6952
55	3, 4,5-triméthylphénol	0,93	0,7144	0,7237	0,043	-0,9034
56	4-(tert) penthylphénol	1,23	1,058	1,0666	0,05	-0,7228
57	6-(tert) butyl-2,4-diméthylphénol	1,25	1,2812	1,277	0,134	0,1372
58	4-bromo-6-chloro-2-méthylphénol	1,28	1,4237	1,4181	0,039	0,601
59	2-(ter) butyl-4-méthylphénol	1,3	0,9717	0,999	0,083	-1,4049
60	4-hexyloxyphénol	1,65	1,6271	1,6282	0,049	-0,0964
61	2,6-di (tert) butyl-4-méthylphénol	1,79	1,8126	1,8056	0,308	0,1111
62	4-heptyloxyphénol	2,03	1,9954	1,9981	0,077	-0,1476
63	2-hydroxybenzamide	-0,24	-0,4628	-	0,05	-0,9369
64	3-méthylhydroxybenzoate	-0,05	0,1161	-	0,024	0,6889
65	4-fluorophénol	0,02	0,2033	-	0,028	0,7621
66	2,3-diméthylphénol	0,12	0,4611	-	0,023	1,4141
67	4-hydroxybenzaldehyde	0,27	-0,1262	-	0,036	-1,6533
68	2-fluorophénol	0,28	0,007	-	0,032	-1,1368
69	3-éthylhydroxybenzoate	0,48	0,4698	-	0,017	-0,042
70	4-cyanophénol	0,52	-0,014	-	0,032	-2,2245
71	2-nitrophenol	0,67	0,1613	-	0,029	-2,1151
72	4-chloro-3-méthylphénol	0,8	0,8946	-	0,02	0,3915
73	3-iodophénol	1,12	0,8905	-	0,034	-0,9568
74	3-phénylphénol	1,35	1,2201	-	0,048	-0,5454

Tableau 16. Suite et fin

N°	Composés	$pIGC_{50exp}$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_{ii}	e_{istd}
75	2,4-bromophénol	1,4	1,2271	-	0,048	-0,726
76	2, 4,6-trichlorophénol	1,7	1,4515	-	0,061	-1,0508
77	4-bromo-2,6-dichlorophénol	1,78	1,5567	-	0,068	-0,948
78	2-hydroxybenzyl alcool	-0,95	-0,8274	-	0,074	0,522
79	3-acétylphénol	-0,38	-0,2495	-	0,037	0,5451
80	3-méthoxyphénol	-0,14	-0,174	-	0,034	-0,1417
81	3-cyanophénol	-0,07	-0,014	-	0,032	0,2331
82	4-éthoxyphénol	0,01	0,1658	-	0,022	0,6454
83	2-cyanophénol	0,03	-0,014	-	0,032	-0,1834
84	3-hydroxybenzaldéhyde	0,08	-0,1262	-	0,036	-0,8605
85	3-éthylphénol	0,23	0,5389	-	0,016	1,2764
86	2-allylphénol	0,35	0,7152	-	0,03	1,5194
87	3-fluorophénol	0,47	0,2033	-	0,028	-1,1083
88	4-chlorophénol	0,55	0,603	-	0,028	0,2204
89	3-isopropylphénol	0,61	0,6287	-	0,023	0,0775
90	4-butoxyphénol	0,7	0,8911	-	0,019	0,7907
91	4-iodophénol	0,85	0,8905	-	0,034	0,169
92	4-hydroxybenzophenone	1,02	1,0238	-	0,039	0,0157
93	2-phénylphénol	1,09	1,2201	-	0,048	0,5462
94	2,5-dichlorophénol	1,13	1,0167	-	0,038	-0,4733
95	4-phénylphénol	1,38	1,2201	-	0,048	-0,6714

II-2-2 Qualité externe du modèle QSAR :

L'application du modèle au 33 composés de validation conduit aux statistiques du tableau 17 où on constate que $SDEP_{ext}$ est un peu différent de $SDEP$ (tableau 15) ; le modèle fonctionne un peu moins bien dans la prédiction interne que dans la prédiction externe.

Tableau 17. Paramètres statistiques pour l'ensemble de validation pour les 33 phénols

n_{ext}	Q_{ext}^2	$SDEP_{ext}$
33	91,84	0,229

Les résultats de la validation externe selon (Golbraikh et Tropsha 2002) confirment la validité du modèle et sont comme suit :

- 1) $Q_{EXT}^2 = 0,8689 > 0,5$
- 2) $R^2 = 0,8696 > 0,6$

$$3) (R^2 - R_0^2)/R^2 = -0,1486 < 0,1 \quad \text{et} \quad 0,85 < k = 1,0271 < 1,15$$

$$(R^2 - R_0'^2)/R^2 = -0,1292 < 0,1 \quad \text{et} \quad 0,85 < k' = 0,8968 < 1,15$$

II-2-3 Qualité d'ajustement du modèle QSAR :

Le graphique représentant les valeurs prédites et calculées en fonction des expérimentales (figure16) est caractérisé par une faible dispersion autour de la droite ce qui indique un bon ajustement.

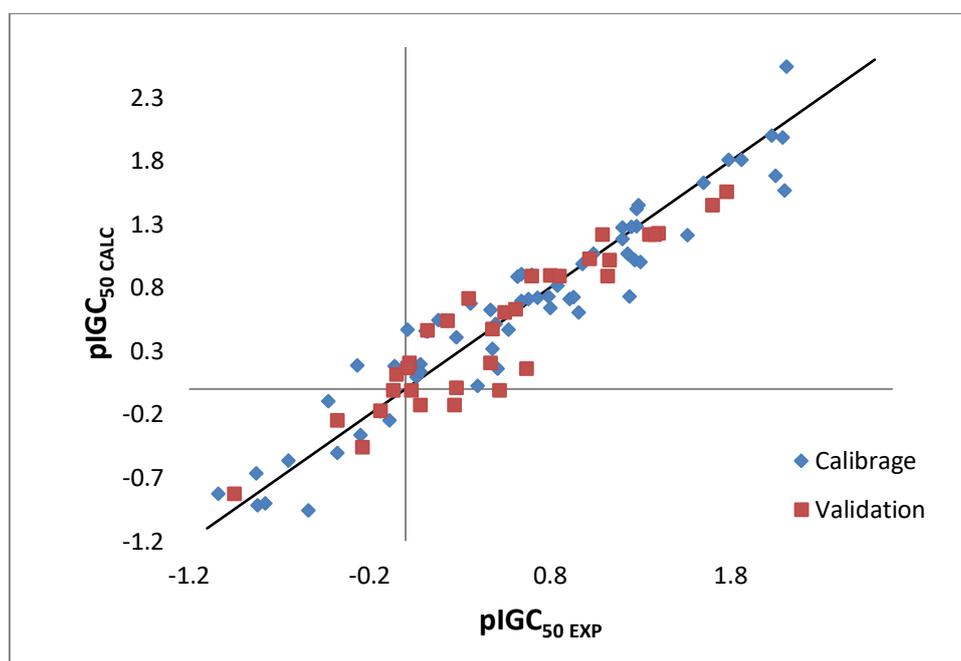


Figure 16 : Droite d'ajustement du modèle QSAR (Phénols).

II-2-4 Domaine d'application :

Comme le montre le diagramme de Williams (figure 17), deux phénols de l'ensemble de calibrage ont un effet de levier élevé ($h_i > h^* = 0,145$). Le 2,6-di(tert)butyl-4-méthylphénol est parfaitement prédits mais le 2,6-diphénylphénol l'est un peu moins bien. Aucune valeur aberrante n'est observée, de plus toutes les valeurs des résidus standardisés de prédictions sont comprises entre +2,5 et -2,5.

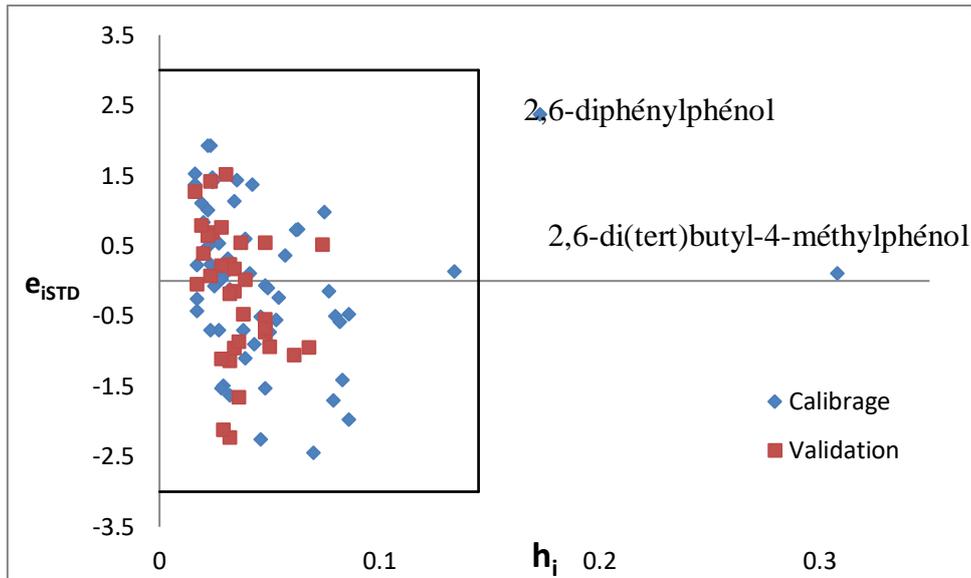


Figure 17 : Diagramme de Williams du modèle QSAR (Phénols)

II-2-5 Teste de randomisation :

Afin de mettre en évidence l'existence de corrélations fortuites, le test de randomisation a été adopté. Comme on peut l'observer (figure 18), les réponses permutées donnent de mauvais modèles, tous ayant $Q^2 < 20$. D'autre part, les paramètres statistiques des $pIGC_{50}$ correctement ordonnés donnent de bons paramètres statistiques. La figure 18 montre une séparation nette entre les statistiques des réponses randomisées et celle du modèle initial. Cela suggère qu'une relation structure-toxicité a été établie et qu'elle n'est pas due au hasard.

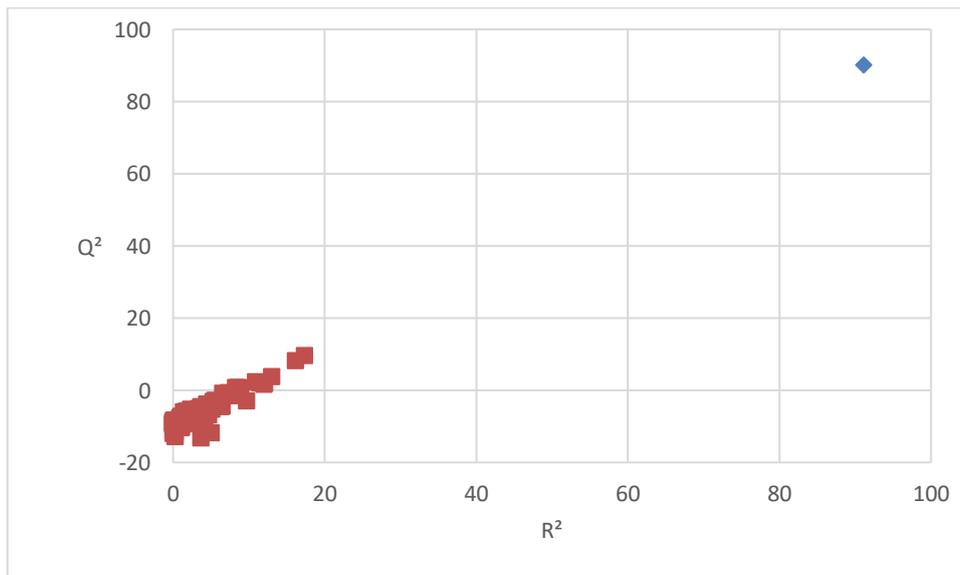


Figure 18 : Teste de randomisation du modèle QSAR (Phénols). (Les croix représentent les activités ordonnées au hasard et le carré correspond aux activités réelles.)

II-3 Interprétation mécanistique du modèle :

Le premier descripteur est le coefficient de partage n-octanol / eau qui est le rapport entre la concentration à l'équilibre d'une substance chimique dans l'octanol et la concentration de la même substance dans l'eau. Il peut être considéré comme une membrane qui sépare l'intérieur et l'extérieur de l'organisme vivant (*Tetrahymena pyriformis*). C'est un facteur d'hydrophobicité, la sélection du $\log K_{ow}$ comme le descripteur le plus utilisé dans les modèles QSTR (avec un nombre différent de descripteurs) souligne l'importance de ce paramètre, donc ce descripteur est le descripteur le plus important.

Le second descripteur (S-CH3) est un indices d'état électrotopologique (E- state indices). Les indices déterminent les sommes de l'état électrotopologique (état E) et / ou les comptes de chaque type d'atome.

II-4 Comparaison avec le modèle original :

Schultz *et al.* [99] ont évalué la toxicité relative de 95 phénols et ont montré qu'une simple corrélation de $pIGC_{50}$ par rapport à au coefficient de partage n-octanol-eau ($\log K_{ow}$) peut modéliser la toxicité environnementale. La prédictivité de ce modèle QSAR dépendant du $\log K_{ow}$ peut être améliorée en ajoutant $\sum \sigma$ (la somme des paramètres électroniques de substitution σ), en tant que second descripteur orthogonal.

Les valeurs de $\sum \sigma$ ont été tirées du livre de Hansch *et Leo* [110] et celles des $\log K_{ow}$ ont été calculées avec le logiciel CLOGP version 3.34 [111].

Comme pour les anilines de la partie I, les paramètres statistiques rapportés par les auteurs sont uniquement liés aux performances d'ajustement et sont les suivant :

$$r^2 = 0,904, s = 0,234, F=435,47$$

Avec l'équation de régression :

$$pIGC_{50} = -1,206 + 0,686 \sum \sigma + 0,640 \log K_{ow} \quad (69)$$

En comparaison notre modèles établi sur les 95 phénols donne les résultats suivant:

$$r^2 = 0,902, s = 0,237, F=423,52$$

En adoptant les mêmes sous-ensembles de calibrage et de validation combinée aux descripteurs choisis par Schultz *et al.*, nous avons calculé un modèle qui a pour équation :

$$pIGC_{50} = -1,202 (\pm 0,083) - 0,646 (\pm 0,102) \sum \sigma + 0,633 (\pm 0,026) \log K_{ow} \quad (70)$$

II. Toxicité des phénols : Approche QSAR

Les paramètres d'ajustement indiqués ci-dessous dans le tableau 18 montrent que notre modèle est légèrement meilleur que celui basé sur l'approche de Schultz *et al.* Ce dernier est meilleur en ce qui concerne la validation externe $Q_{ext}^2 = 93,08 > 91,84$.

Tableau 18. Paramètres statistiques du modèle de l'approche Schultz *et al.* (Phénols)

n_{tr}	n_{ext}	Q_{LOO}^2	R^2	$Q_{LMO/50}^2$	Q_{BOOT}^2	R_{adj}^2
62	33	89,72	90,86	89,12	88,92	90,55
Q_{ext}^2	<i>SDEC</i>	<i>SDEP</i>	<i>SDEP</i> _{ext}	s	F	
93,08	0,242	0,257	0,210	0,248	293,2538	

Les résultats de la validation externe selon Tropsha *et al.* sont comme suit :

- 1) $Q_{EXT}^2 = 0,8889 > 0,5$
- 2) $R^2 = 0,8981 > 0,6$
- 3) $(R^2 - R_0^2)/R^2 = -0,0965 < 0,1$ et $0,85 < k = 1,0944 < 1,15$
 $(R^2 - R_0'^2)/R^2 = -0,0755 < 0,1$ et $0,85 < k' = 0,8586 < 1,15$

Ces statistiques confirment la validité du modèle basé sur les descripteurs choisis par Schultz *et al.* On remarque que ces valeurs sont légèrement supérieures à celles trouvées pour notre modèle.

Le tableau 19 renferme les valeurs prédites et calculées de $pIGC_{50}$ des deux ensembles (calibrage et validation), en plus des valeurs des leviers et des résidus standardisés de prédictions pour le modèle de l'approche de Schultz *et al.* en plus des valeurs de $\sum \sigma$ et $\log K_{ow}$. La numérotation correspond à celle du tableau 16 de notre modèle.

Tableau 19. Valeurs des descripteurs selon Schultz *et al.* [99], valeurs prédites et calculées de $pIGC_{50}$, valeurs des leviers et des résidus standardisés de prédictions (Phénols)

N°	$\log K_{ow}$	$\Sigma \sigma$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_i	e_{istd}
1	0,44	0	-0,91	-0,92	0,085	0,54
2	0,67	0	-0,77	-0,78	0,073	0,23
3	0,49	0	-0,9	-0,89	0,082	-0,33
4	0,81	0,12	-0,61	-0,61	0,059	0,17
5	0,25	0,16	-0,98	-0,94	0,085	-1,85
6	1,48	0	-0,26	-0,27	0,04	0,71
7	0,9	0,01	-0,64	-0,63	0,061	-1,09
8	1,1	0,1	-0,45	-0,44	0,048	-0,83
9	1,57	-0,27	-0,4	-0,38	0,064	-1,08
10	1,45	0,5	0,05	0,04	0,053	0,56
11	2,12	-0,07	0,1	0,09	0,027	0,65
12	2,77	-0,14	0,47	0,46	0,023	1,88
13	1,98	0,5	0,39	0,37	0,046	1,36
14	2,08	0,5	0,45	0,44	0,045	1,55
15	1,98	0,45	0,35	0,34	0,04	1,12
16	2,77	-0,14	0,47	0,46	0,023	1,47
17	2,65	-0,15	0,38	0,38	0,025	0,83
18	2,65	-0,15	0,38	0,38	0,025	0,71
19	2,2	0,23	0,34	0,34	0,022	0,25
20	3,33	-0,21	0,78	0,77	0,028	1,73
21	1,65	0,12	-0,1	-0,08	0,03	-2,03
22	3,05	-0,13	0,65	0,64	0,022	0,73
23	2,35	0,23	0,43	0,43	0,021	-0,27
24	1,85	0,71	0,42	0,43	0,079	-0,38
25	2,51	0	0,38	0,39	0,019	-0,76
26	2,88	0,54	0,99	0,97	0,055	1,53
27	3,18	-0,15	0,72	0,71	0,023	0,31
28	2,63	0,23	0,61	0,61	0,02	-0,29
29	3,13	0,6	1,2	1,17	0,07	2,1
30	2,91	0,06	0,68	0,68	0,016	-0,46
31	2,78	0,43	0,84	0,84	0,038	-0,02
32	3,45	-0,2	0,85	0,85	0,028	-0,24
33	2,48	0,37	0,6	0,61	0,03	-1,49
34	3,58	-0,12	0,99	0,99	0,024	0,03
35	3,14	-0,03	0,76	0,77	0,018	-1,13
36	3,69	-0,09	1,07	1,08	0,025	-0,52
37	3,78	0,09	1,25	1,25	0,027	0,21
38	3,45	-0,2	0,84	0,85	0,028	-1,63
39	3,07	0,6	1,12	1,13	0,069	-0,63
40	3,93	-0,11	1,21	1,21	0,029	-0,28
41	3,69	-0,22	0,98	0,99	0,031	-1,26
42	3,35	0,74	1,38	1,4	0,105	-0,78
43	4,71	0	1,77	1,78	0,053	-0,35

Tableau 19. Suite

N°	$\log K_{ow}$	$\Sigma\sigma$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_i	e_{istd}
44	4,02	0,69	1,75	1,79	0,118	-1,27
45	5,31	-0,2	2,03	2,03	0,078	-0,27
46	3,85	0,83	1,71	1,77	0,148	-1,69
47	5,25	-0,02	2,11	2,11	0,077	-0,01
48	0,33	0,36	-0,76	-0,76	0,08	0,09
49	2,12	-0,17	0,04	0,03	0,036	1,28
50	2,77	-0,24	0,41	0,4	0,032	1,17
51	2,51	0,37	0,63	0,63	0,03	0,62
52	2,85	0,16	0,71	0,71	0,018	0,27
53	3,45	-0,1	0,92	0,92	0,022	0,78
54	3,05	-0,13	0,64	0,64	0,022	-0,65
55	3,42	-0,41	0,69	0,7	0,05	-1,01
56	3,98	-0,24	1,16	1,16	0,037	-0,29
57	4,75	-0,54	1,48	1,46	0,086	0,95
58	3,87	0,29	1,44	1,43	0,043	0,67
59	4,1	-0,37	1,15	1,15	0,05	-0,63
60	4,22	-0,34	1,23	1,25	0,049	-1,74
61	6,08	-0,57	2,36	2,28	0,144	2,49
62	4,75	-0,34	1,56	1,59	0,063	-1,98
63	0,96	0,36	-0,36	-	0,054	-0,51
64	1,95	0,37	0,27	-	0,033	1,32
65	1,91	0,06	0,05	-	0,026	0,11
66	2,77	-0,24	0,4	-	0,032	1,13
67	1,44	0,45	0	-	0,047	-1,12
68	1,63	0,06	-0,13	-	0,032	-1,69
69	2,07	0,42	0,38	-	0,036	-0,41
70	1,6	0,66	0,24	-	0,072	-1,18
71	1,85	0,78	0,47	-	0,092	-0,84
72	3,13	0,16	0,88	-	0,019	0,34
73	2,89	0,35	0,85	-	0,03	-1,09
74	3,36	0,06	0,96	-	0,019	-1,57
75	3,37	0,46	1,23	-	0,051	-0,71
76	3,69	0,69	1,58	-	0,104	-0,51
77	3,84	0,69	1,67	-	0,109	-0,45
78	0,44	0	-0,92	-	0,084	0,11
79	1,45	0,38	-0,04	-	0,042	1,41
80	1,57	0,12	-0,13	-	0,032	0,04
81	1,6	0,56	0,17	-	0,057	1,01
82	2,1	-0,24	-0,03	-	0,043	-0,16
83	1,6	0,66	0,24	-	0,072	0,87
84	1,44	0,35	-0,06	-	0,04	-0,59
85	2,65	-0,07	0,43	-	0,02	0,82
86	2,64	-0,15	0,37	-	0,025	0,09
87	1,91	0,34	0,23	-	0,031	-1

Tableau 19. Suite et fin

N°	$\log K_{ow}$	$\Sigma \sigma$	$pIGC_{50pred}$	$pIGC_{50calc}$	h_i	e_{istd}
88	2,48	0,23	0,52	-	0,02	-0,14
89	3,05	-0,07	0,68	-	0,019	0,3
90	3,16	-0,32	0,59	-	0,038	-0,45
91	2,89	0,18	0,74	-	0,019	-0,43
92	3,08	0,43	1,03	-	0,041	0,02
93	3,36	-0,01	0,92	-	0,019	-0,7
94	3,07	0,6	1,13	-	0,068	-0,01
95	3,36	-0,01	0,92	-	0,019	-1,88

La qualité de l'ajustement du modèle basé sur l'approche de Shultz *et al.* Des phénols, représenté dans la figure 19 est légèrement supérieur à celle de notre modèle (figure 16).

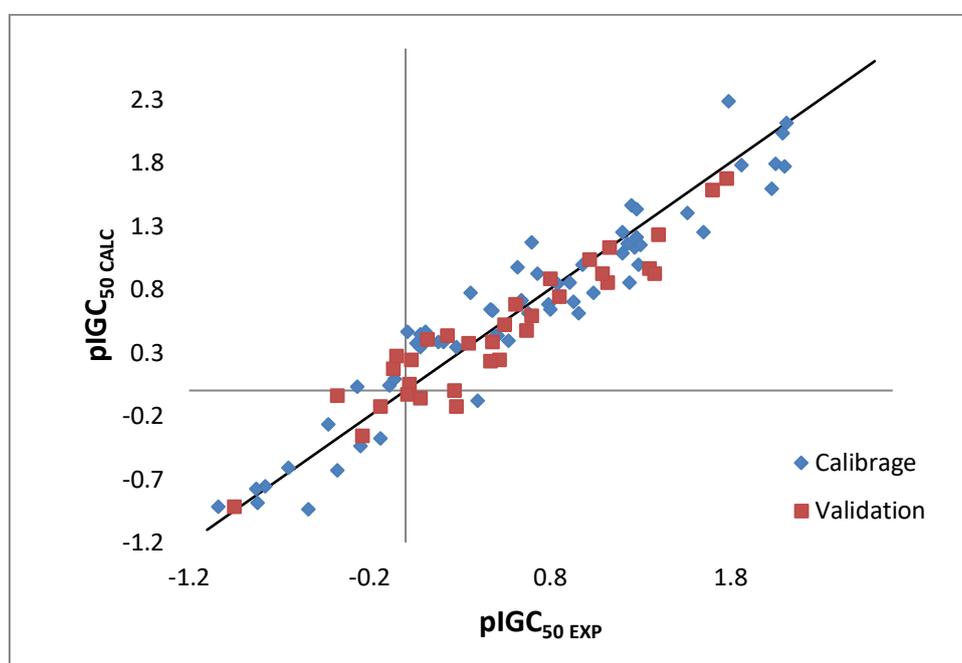


Figure 19 : Droite d'ajustement du modèle basé sur l'approche de Schultz *et al.* (Phénols)

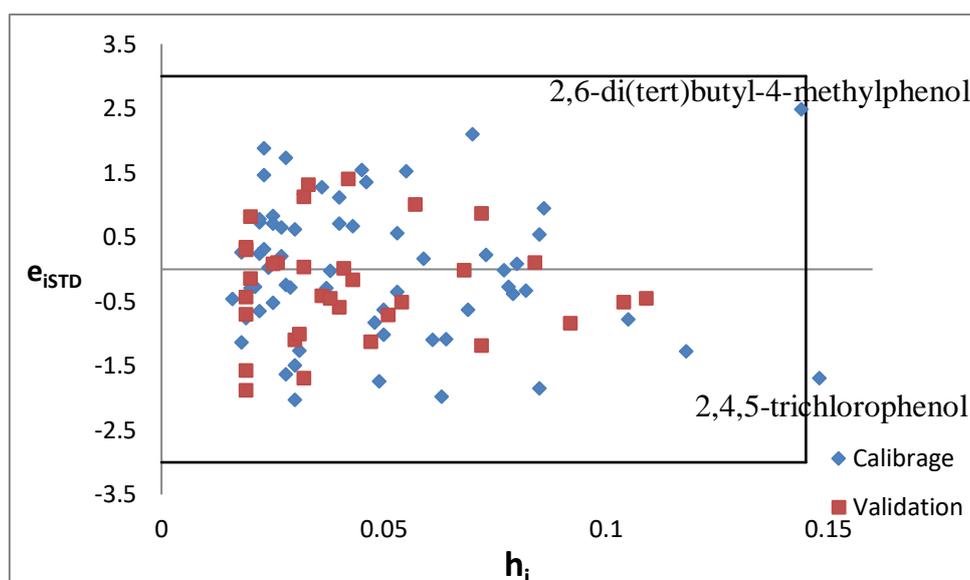


Figure 20 : Diagramme de Williams du modèle QSAR basé sur l'approche de Schultz *et al.*
(Phénols)

Aucune valeur aberrante n'est observée dans le domaine d'application du modèle basé sur les descripteurs de Schultz *et al.* défini par le diagramme de Williams (figure 20) de plus toutes les valeurs des résidus standardisés de prédictions sont comprises entre +2,5 et -2,5. Contrairement à notre modèle, celui de Schultz *et al.* n'est caractérisé qu'avec un seul composé influant qui est le 2,4,5-trichlorophénol. Le 2,6-di(tert)butyl-4-méthylphénol signalé précédemment comme influant (figure 18) est à la limite de h^* avec un $h_i = 0,144$.

En conclusion de cette comparaison nous notons que :

- Comme pour les anilines, le modèle de Schultz *et al.* établie sur les 95 phénols est dépourvu de validation statistique externe.
- Notre modèle est meilleur que celui de Schultz *et al.* en tenant compte des performances d'ajustement et de prédiction interne.
- Concernant les performances de prédiction externe, les deux modèles sont valides selon la procédure de Tropsha *et al.* mais le modèle basé sur $\sum \sigma$ est meilleur ($Q_{ext}^2 = 93,08$).
- Comme le montre la figure 17 le 2,6-di(tert)butyl-4-méthylphénol est problématique car il a un $h_i = 0,308 > h^*$.

Ces remarques conduisent à dire que le modèle basé sur $\sum \sigma$ et $\log K_{ow}$ est globalement meilleur que celui que nous présentant. L'amélioration que nous apportant à l'approche de Schultz *et al.* concerne la validation statistique externe et interne et l'approche des leviers pour définir le domaine d'application.

II-5 Modèle par réseaux de neurones artificiel

Afin d'améliorer la qualité du modèle QSAR pour la prédiction de la concentration inhibitrice de la croissance vis à vis de *Tetrahymena pyriformis* nous avons exploité un réseau de neurones artificiel (RNA) à trois couches contenant une couche d'entrée recevant les descripteurs calculés, une couche cachée contenant un nombre de neurones à choisir et une couche de sortie fournissant la réponse du RNA (c-à-d $pIGC_{50_{pred}}$).

La couche d'entrée est formée des deux descripteurs précédemment choisis dans notre modèle linéaire $\log K_{ow}$ et S-CH3.

Pour choisir le nombre de neurones dans la couche cachée n nous avons procédé sur les 62 phénols de calibrages dont 33 choisis aléatoirement pour tester l'apprentissage du réseau par rétro-propagation de l'erreur. Pour chaque nombre de neurone (de 1 à 7) les SDEC (pour l'ensemble de calibrage) et SDEP ; (pour l'ensemble de teste) ont été calculés. La représentation graphique des résultats est la figure 21.

Le nombre de neurones optimal est fixé à 4 et il est caractérisé par les plus petites valeurs des SDEC et SDEP, 0,179 et 0,175 respectivement.

L'étape finale du choix du réseau est de trouver le nombre d'itération optimal en procédant sur les mêmes sous-ensembles précédemment définis en faisant varier les itérations de 10 à 100. Pour 4 neurones dans la couche cachée.

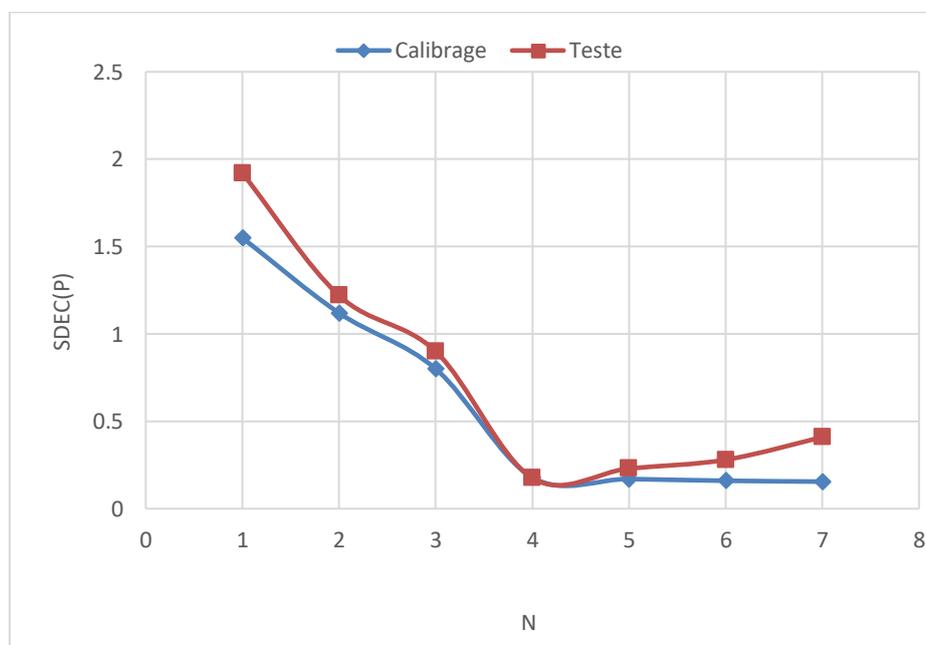


Figure 21 : Evolution des erreurs SDEC et SDEP en fonction du nombre de neurones.

II. Toxicité des phénols : Approche QSAR

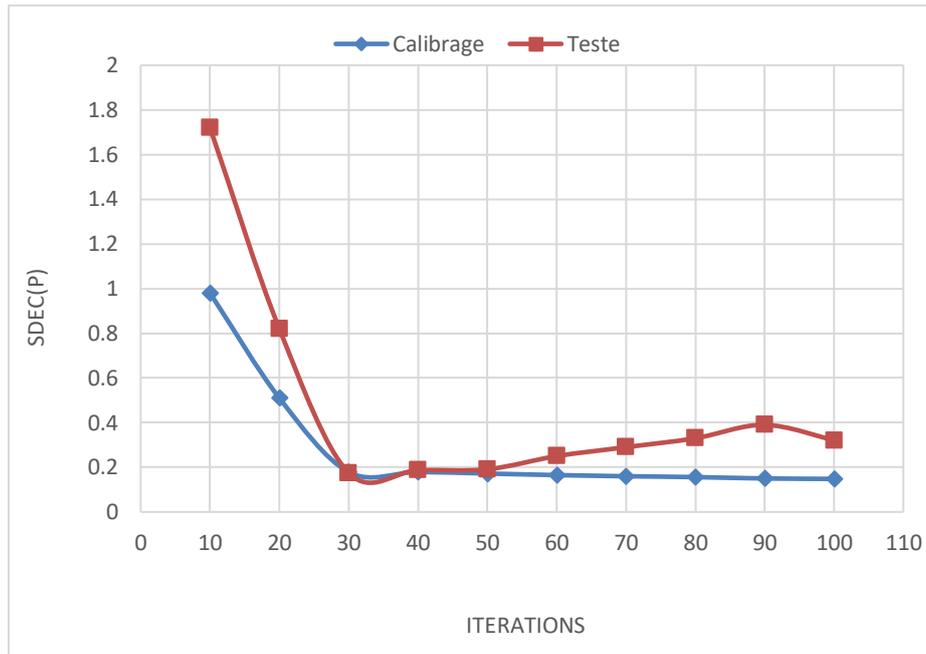


Figure 22 : Evolution des erreurs SDEC et SDEP en fonction du nombre d'itérations.

La figure 22 représente les résultats de cette procédure où l'on remarque que 30 itérations pour lesquels SDEC et SDEP sont très petits suffisent à l'apprentissage du réseau.

Le meilleur modèle RNA a été trouvé pour les paramètres suivants : 4 neurones dans la couche cachée et un nombre d'itérations = 30. Dans ces conditions, les résultats obtenus sont présentés dans le tableau 20.

Tableau 20. Paramètres statistiques du modèle RNA.

n_{tr}	n_{ext}	R^2	Q_{ext}^2	$SDEC$	$SDEP_{ext}$
62	33	95,01	94,91	0,1787	0,1806

L'approche non-linéaire améliore notablement les statistiques du modèle. En comparaison aux tableaux 15 et 17, toutes les statistiques sont meilleures : R^2 passent de 91,15 à 95,01, Q_{ext}^2 de 91,84 à 94,91 et les $SDEC$ et $SDEP_{ext}$ sont plus faibles.

Les valeurs expérimentales, calculées et prédites de la concentration inhibitrice de la croissance ainsi que les résidus sont dans le tableau 21. La numérotation correspond à celle du tableau 16.

II. Toxicité des phénols : Approche QSAR

Tableau 21. Valeurs expérimentales, calculées et prédites de $pIGC_{50}$ et résidus.

N°	$pIGC_{50 \text{ exp}}$	$pIGC_{50 \text{ RNA}}$	e_i
1	-1,04	-0,7529	-0,2871
2	-0,83	-0,6882	-0,1418
3	-0,82	-0,7316	-0,0884
4	-0,65	-0,6256	-0,0244
5	-0,54	-0,7824	0,2424
6	-0,43	0,0476	-0,4776
7	-0,38	-0,573	0,193
8	-0,25	-0,4133	0,1633
9	-0,14	-0,2088	0,0688
10	-0,09	-0,293	0,203
11	-0,06	-0,0679	0,0079
12	0,01	0,1298	-0,1198
13	0,06	-0,0575	0,1175
14	0,08	0,0694	0,0106
15	0,08	0,0798	0,0002
16	0,11	0,1114	-0,0014
17	0,18	0,3459	-0,1659
18	0,21	0,3312	-0,1212
19	0,28	0,4929	-0,2129
20	0,36	0,756	-0,396
21	0,4	0,2596	0,1404
22	0,47	0,5178	-0,0478
23	0,5	0,5392	-0,0392
24	0,51	0,4211	0,0889
25	0,57	0,2465	0,3235
26	0,62	0,8843	-0,2643
27	0,64	0,8722	-0,2322
28	0,68	0,7073	-0,0273
29	0,7	0,8679	-0,1679
30	0,79	0,6541	0,1359
31	0,84	0,8138	0,0262
32	0,91	0,8579	0,0521
33	0,96	0,6077	0,3523
34	0,98	1,0044	-0,0244
35	1,04	1,0701	-0,0301
36	1,2	1,1401	0,0599
37	1,2	1,2601	-0,0601
38	1,24	0,8735	0,3665
39	1,27	1,0184	0,2516
40	1,28	1,2387	0,0413
41	1,29	1,604	-0,314
42	1,56	1,2457	0,3143
43	1,86	1,9148	-0,0548
44	2,05	1,9077	0,1423
45	2,09	2,1107	-0,0207
46	2,1	1,7687	0,3313
47	2,11	2,1312	-0,0212
48	-0,78	-0,7719	-0,0081
49	-0,27	-0,0474	-0,2226

Tableau 21. Suite et fin

N°	$pIGC_{50 \text{ exp}}$	$pIGC_{50 \text{ RNA}}$	e_i
50	0,12	0,0961	0,0239
51	0,48	0,4755	0,0045
52	0,64	0,603	0,037
53	0,73	0,8642	-0,1342
54	0,8	0,5331	0,2669
55	0,93	0,8529	0,0771
56	1,23	1,2261	0,0039
57	1,25	1,3952	-0,1452
58	1,28	1,4672	-0,1872
59	1,3	1,2438	0,0562
60	1,65	1,7624	-0,1124
61	1,79	1,7087	0,0813
62	2,03	2,0637	-0,0337
63	-0,24	-0,5316	0,2916
64	-0,05	0,077	-0,127
65	0,02	0,4457	-0,4257
66	0,12	0,12	0
67	0,27	-0,0062	0,2762
68	0,28	0,2372	0,0428
69	0,48	0,2493	0,2307
70	0,52	0,202	0,318
71	0,67	0,4211	0,2489
72	0,8	0,8636	-0,0636
73	1,12	0,8914	0,2286
74	1,35	1,255	0,095
75	1,4	1,2645	0,1355
76	1,7	1,604	0,096
77	1,78	1,7592	0,0208
78	-0,95	-0,7529	-0,1971
79	-0,38	-0,289	-0,091
80	-0,14	-0,2041	0,0641
81	-0,07	0,202	-0,272
82	0,01	-0,0664	0,0764
83	0,03	0,202	-0,172
84	0,08	-0,0062	0,0862
85	0,23	0,3368	-0,1068
86	0,35	0,7143	-0,3643
87	0,47	0,4457	0,0243
88	0,55	0,6077	-0,0577
89	0,61	0,5236	0,0864
90	0,7	0,8589	-0,1589
91	0,85	0,8914	-0,0414
92	1,02	1,0257	-0,0057
93	1,09	1,255	-0,165
94	1,13	1,0184	0,1116
95	1,38	1,255	0,125

En utilisant les valeurs expérimentales et prédites de l'ensemble de validation les statistiques de tropsha *et al.* prouvent la validité du modèle RNA et sont les suivants :

- 1) $Q_{EXT}^2 = 0,9184 > 0,5$
- 2) $R^2 = 0,9115 > 0,6$
- 3) $(R^2 - R_0^2)/R^2 = -0,0918 < 0,1$ et $0,85 < k = 1,0092 < 1,15$
 $(R^2 - R_0'^2)/R^2 = -0,0857 < 0,1$ et $0,85 < k' = 0,9418 < 1,15$

La qualité d'ajustement du modèle RNA-QSAR peut être vérifiée par la figure 23 où l'on remarque que la dispersion des points est moindre que celle du modèle RLM apprécié dans la figure 16 ce qui confirme la nette amélioration apportée par la modélisation non linéaire.

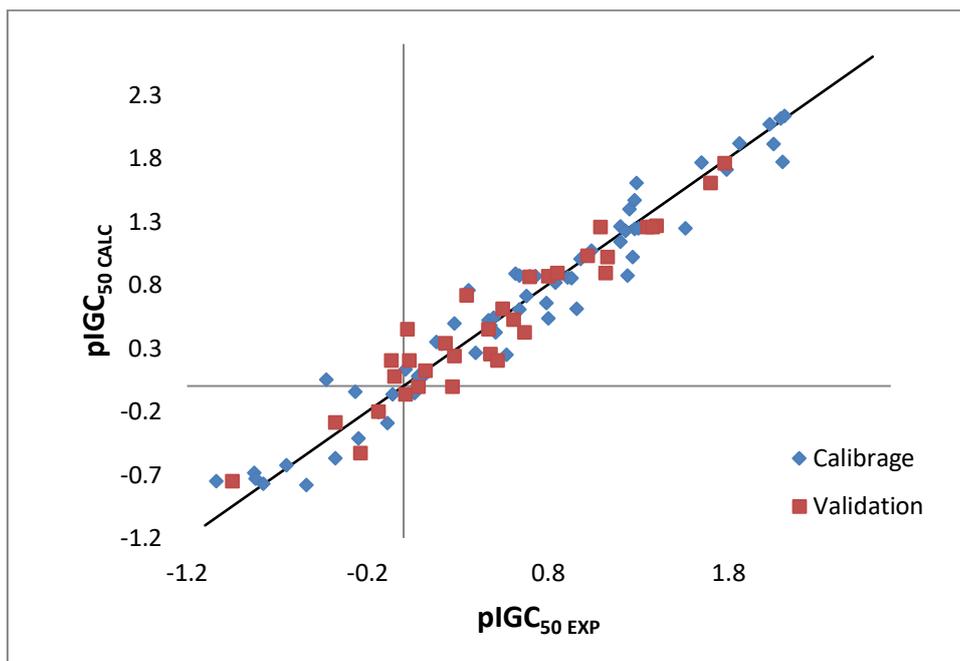


Figure 23 : Droite d'ajustement du modèle RNA-QSAR.

En représentant les valeurs des résidus en fonction des valeurs expérimentales de $pIGC_{50}$ (figure 24) on remarque que toutes les valeurs sont comprises entre $\pm SDEC = \pm 0,5361$ pour l'ensemble de calibrage et $\pm SDEP_{ext} = \pm 0,5418$ pour l'ensemble de validation ce qui prouve que les erreurs de calculs ou de prédiction sont acceptables.

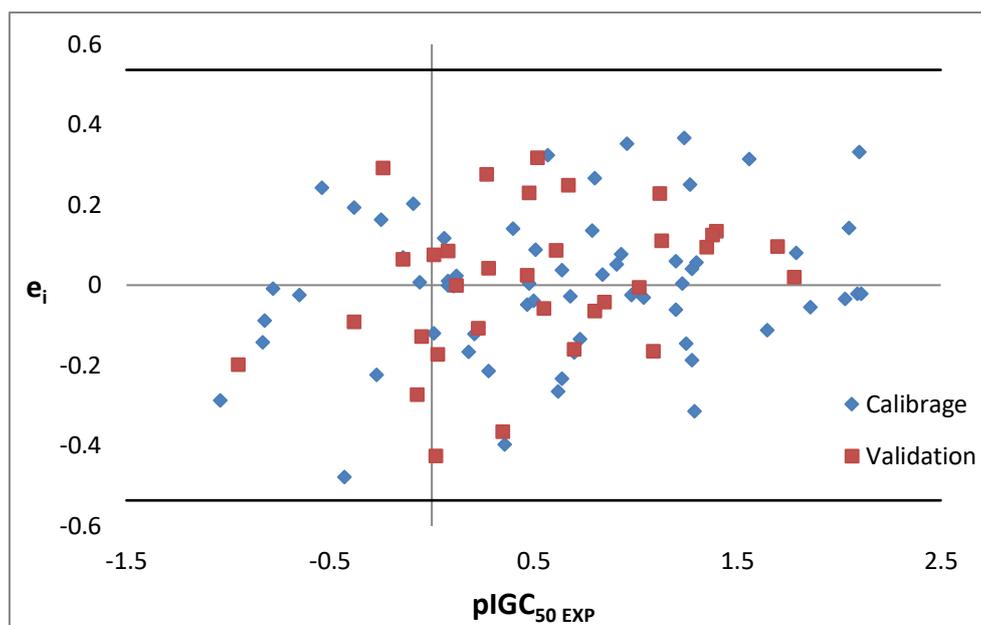


Figure 24 : Valeurs des résidus en fonction des valeurs expérimentales de $pIGC_{50}$



CONCLUSION GÉNÉRALE

Conclusion générale

La toxicité relative à 48 anilines et 95 phénols utilisant les caractéristiques de croissance de la population de *Tetrahymena pyriformis* (concentration causant 50% d'inhibition de croissance), disponible dans la littérature, a été étudiée. Au début, l'ensemble des données a été divisé au hasard en un ensemble de calibrage (31 produits chimiques pour les anilines et 62 pour les phénols) utilisé pour établir le modèle QSAR, et un ensemble de validation (17 produits chimiques pour les anilines et 33 pour les phénols) pour la validation statistique externe.

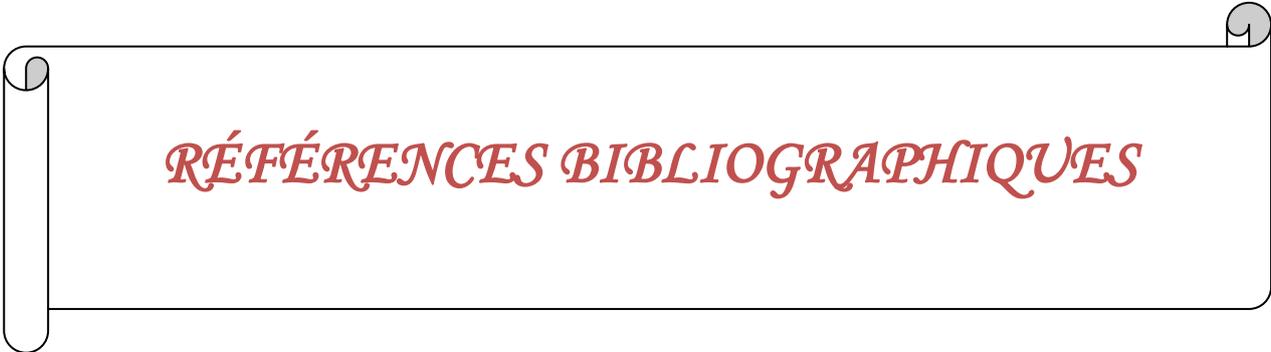
Concernant les anilines, un modèle biparamétrique a été développé en utilisant, comme variables indépendantes, des descripteurs théoriques 3D dérivés du logiciel DRAGON qui sont $R3v+$ et $RGyr$. Le modèle de moindres carrés ordinaires (MCO), en sélectionnant les sous-ensembles de variables par algorithmes génétiques (GA-VSS), a été examiné pour la robustesse (validation croisée Q_{LOO}^2 , Randomisation de Y), et pour la capacité prédictive par les méthodes de validation interne (validation croisée Q_{LMO}^2 , bootstrap) et externe (Q_{ext}^2). Les descripteurs inclus dans le modèle QSAR indiquent que la toxicité est liée, à la taille et à la forme moléculaire, et à l'interaction de la molécule avec son milieu environnant ou sa cible. De plus, le domaine d'applicabilité du modèle a été discuté. Le modèle QSAR proposé dans ce travail est stable, robuste, avec de bonnes performances d'ajustement, et de prédiction. Le modèle obtenu et discuté est meilleur et que celui de Schultz *et al.* basé sur le $\log P$ et $\sum \sigma$ statistiquement parlant, et qui est en plus entaché de plusieurs anomalies qui limitent son utilisation. L'absence de validation externe indispensable dans les études QSAR, et les différents points influents et aberrants qui apparaissent en adoptant une approche QSAR standard en utilisant les descripteurs originaux sont parmi les plus importantes anomalies remarquées.

Pour les phénols, un modèle avec deux variables électrotopologiques explicatives a été construit. Ce modèle donne de bons résultats tant en calibrage qu'en validation ($R^2 = 91,15\%$, $Q_{EXT}^2 = 91,84\%$) et, la comparaison avec le modèle original basé comme pour les anilines sur le $\log P$ et $\sum \sigma$ est en faveur de ce dernier seulement après utilisation d'une approche QSAR usuelle et seulement en ce qui concerne la validation statistique externe. L'amélioration que nous apportons à l'approche de Schultz *et al.* concerne la validation statistique externe et interne et l'approche des leviers pour définir le domaine d'application.

Conclusion générale

Afin d'améliorer la qualité du modèle QSAR pour la prédiction de la concentration inhibitrice de la croissance vis-à-vis de *Tetrahymena pyriformis* des phénols, nous avons exploité un réseau de neurones artificiel (RNA) à trois couches contenant une couche d'entrée recevant les descripteurs calculés, une couche cachée contenant un nombre de neurones à choisir et une couche de sortie fournissant la réponse du RNA qui a 4 neurones dans la couche cachée et un nombre d'itérations = 30 comme paramètres. Toutes les statistiques sont meilleures : R^2 passent à 95,01 Q_{ext}^2 , à 94,91 et les $SDEC$ et $SDEP_{ext}$ sont plus faibles. Le modèle RNA-QSAR est plus performant que le MLR-QSAR et peut servir à prédire la toxicité des phénols.

Il serait plus intéressant, (à l'avenir) d'adopter l'approche QSAR en utilisant une base de données qui nous est propre en collectant les données n'ayant pas encore fait l'objet de ce genre d'approche. De plus une interprétation mécanistique plus poussée du modèle QSAR, malgré qu'elle ne soit pas obligatoire, serait un plus pour cette modélisation mais nécessiterait des notions solide, en biologie et toxicologie.



RÉFÉRENCES BIBLIOGRAPHIQUES

- [1]. Mme. ERRAHOUI née BELLIFA KHADIDJA. Etude des relations quantitatives structure–toxicité des composés chimiques à l’aide des descripteurs moléculaires. « Modélisation QSAR » Thèse de Doctorat. Université Abou BekrBelkaïd de Tlemcen. **2015**.
- [2]. Notions de Toxicologie www.csst.qc.ca
- [3]. J. C. Dearden, Prediction of Environmental Toxicity and Fate Using Quantitative Structure-Activity Relationships (QSARs). *Journal of the Brazilian Chemical Society*, **2002**, 13,754-762.
- [4]. Commission of the European Communities; White Paper on a Strategy for a Future Chemicals Policy, Brussels, Belgium, **2001**.
<http://europa.eu.int/comm/environment/chemicals/whitepaper.htm>.
- [5]. http://www.csst.qc.ca/prevention/reptox/Pages/fiche-complete.aspx?no_produit=422
(Accédé le 10/02/2017)
- [6]. M. Atanasova et F.Ribarova. Phénols et flavonoïdes totaux dans les extraits secs des feuilles des bouleaux argentés bulgares (*Betula pendula*). *Revue de génie industriel*, **2009**, 4, 21-25.
- [7]. Phénol Fiche toxicologique n° 15.
http://www.inrs.fr/publications/bdd/fichetox/fiche.html?refINRS=FICHETOX_15
(Accédé le 12/02/2017)
- [8]. Mohamed ESSALIH. L’étude des indices topologiques, leurs applications en QSAR/QSPR et leurs corrélations aux représentations moléculaires «Plerograph» et «Kenograph». Thèse de Doctorat. Université Mohammed V-Agdal Faculté des Sciences de Rabat. **2013**
- [9]. E.Russo. Chemistry plans a structural overhaul. *Nature*, **2002**, 419, 4–7.
- [10]. E.Bolton, Y.Wang, P.A.Thiessen et S.H.Bryant. Chapter 12 pubchem : Integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry*, **2008**, 4, 217–241
- [11]. A.Crum-Brown et T.Frazer. On the connection between chemical constitution and physiological action. *Journal of Anatomy and Physiology*, **1868**, 2;224–242.
- [12]. A.Z.Dudek, T.Arodz, and J.Gálvez. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Combinatorial Chemistry & High Throughput Screening*, **2006**, 9, 213–228.
- [13]. F.Bonachera. Les triplets pharmacophoriques flous : développement et applications. Thèse de Doctorat. Université Lille1 sciences et technologies. **2011**.

- [14]. A.Goulon-Sigwalt-Abram. Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments. Thèse de Doctorat. Université Pierre et Marie Curie (Paris 6), **2008**.
- [15]. M.V.Diudea, I.Gutman et L.Jantschi. Molecular Topology. Science publishers, **1999**.
- [16]. C.Hansch, A.Leoet D.Hoekmann. Exploring QSAR : hydrophobic, electronic and steric constants. **1995**.
- [17]. H.Wiener. Structural determination of paraffin boiling points. *Journal of Chemical Information and Computer Sciences*, **1947**, 69, 17–20
- [18]. J.M.Barnard et G.M.Downs. Chemical fragment generation and clustering software. *Journal of Chemical Information and Computer Sciences*, **1997**, 89, 141–142.
- [19]. J.L.Durant, B.A.Leland, D.R.Henry et J.G.Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, **2002**, 42, 1273– 1280
- [20]. J.H.Schuur, P.Selzer et J.Gasteiger. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *Journal of Chemical Information and Computer Sciences*, **1996**; 36, 334–344.
- [21]. R. L. Lipnick. Charles Ernest Overton: narcosis studies and a contribution to general pharmacology.*Trends in Pharmacological Sciences*, **1986**, 7, 161–164.
- [22]. C. Hansch, P.P. Maloney, T. FujitaetR.M. Muir. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients.*Nature*,**1962**, 194, 178-180.
- [23]. I. Lessigiarska, A.P. Worthet T.I. Netzeva, Comparative review of QSARs for acute toxicity. EUR report No. 21559 EN. EC Joint Research Centre, Ispra, Italy (**2005b**).
- [24]. M.T.D.Cronin et J.C. Dearden, *Quantitative Structure-Activity Relationships*, **1995**, 14, 1-7.
- [25]. M.T.D. Cronin, T. I. Netzeva, J. C. Dearden, R. Edwards et A. D.P. Worgan. Assessment and Modeling of the Toxicity of Organic Chemicals to *Chlorella vulgaris*: Development of a Novel Database. *Chemical Research in Toxicology*, **2004**, 17, 545-554.
- [26]. H.J.M.Verhaar, C. J. Van Leeuwen et J.L.M. Hermens. Classifying environmental pollutants 1. Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere*, **1992**, 25, 471-491
- [27]. J.Dearden, ECVAM Workshop on The Use of Computer Models as Alternatives to

Animal Experiments in Chemical Risk Assessment, October 3-4, 2002, Praha, Czech Republic.

- [28]. A.P. Freidiget J.L.M. Hermens. Narcosis and chemical reactivity QSARs for acute fish toxicity. *Quantitative Structure-Activity Relationships*, **2000**, 19, 547-553.
- [29]. S. Kapur, A. Shusterman, R.P. Verma, C. Hanschet C.D. Selassie, Toxicology of benzylalcohols: a QSAR analysis. *Chemosphere*, **2000**, 41, 1643-1649
- [30]. T.F. Parkerton et W.J. Konkel. Application of quantitative structure-activity relationships for assessing the aquatic toxicity of phthalate esters. *Ecotoxicology and Environmental Safety*, **2000**, 45, 61-78.
- [31]. J.G. Bundy, A.W.J. Morriss, D.G. Durham, C.D. Campbell et G.I. Paton. Development of QSARs to investigate the bacterial toxicity and biotransformation Potential of aromatic heterocyclic compounds. *Chemosphere*, **2001**, 42, 885-892.
- [32]. P. Gramatica, M. Vighi, F. Consolaro, R. Todeschini, A. Finizio et M. Faust. QSAR approach for the selection of congeneric compounds with a similar toxicological mode of action. *Chemosphere*, **2001**, 42, 873-883.
- [33]. S. Renet P.D. Frymier. Estimating the toxicities of organic chemicals to bioluminescent bacteria and activated sludge. *Water Research*, **2002**, 36, 4406-4414.
- [34]. L.E. Sverdrup, T. Nielsen et P.H. Krogh. Soil ecotoxicity of polycyclic aromatic hydrocarbons in relation to soil sorption, lipophilicity, and water solubility. *Environmental Science & Technology*, **2002**, 36, 2429-2435
- [35]. A.D.P. Worgan, J.C. Dearden, R. Edwards, T.I. Netzeva et M.T.D. Cronin. Evaluation of an over short-term algal toxicity assay by the development of QSARs and inter-species relationships for narcotic chemicals. *QSAR & Combinatorial Science*, **2003**, 22, 204-209.
- [36]. M.T.D. Cronin, G.S. Bowers, G. D. Sinks et T.W. Schultz, Structure-toxicity relationships for aliphatic compounds encompassing a variety of mechanisms of toxic action to *V. fischeri*. *SAR and QSAR in Environmental Research*, **2000**, 11, 301-312.
- [37]. M.T.D. Cronin et T.W. Schultz. Development of quantitative structure-activity relationships for the toxicity of aromatic compounds to *T. pyriformis*: comparative assessment of the methodologies. *Chemical Research in Toxicology*, **2001**, 14, 1284-1295
- [38]. M.T.D. Cronin, A.O. Aptula, J.C. Duffy, T.I. Netzeva, P.H. Rowe, I.V. Valkova et T.W. Schultz. Comparative assessment of methods to develop QSARs for the

- prediction of the toxicity of phenols to *Tetrahymena pyriformis*. *Chemosphere*, **2002**, 49, 1201–1221.
- [39]. T.W. Schultz, D.T. Lin, T.S. WilkeetL.M. Arnold, Quantitative structure- activity relationships for the *Tetrahymena pyriformis* population growth endpoint: a mechanism of action approach. In: Karcher,W and Devillers, J.(Eds), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*. Kluwer Academic Publishers, **1990**, 61-82.
- [40]. A.D.P. Worgan. The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. *SAR and QSAR in Environmental Research*. **2002**, 13, 167-176
- [41]. M.T.D. Cronin, J.C. Dearden, J.C. Duffy, R.Edwards, N.Manga, A.P.Worth et A.D.P. Worgan. The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. *SAR and QSAR in Environmental Research*,**2002**, 13, 167-176.
- [42]. S.D. Dimitrov, O. G. Mekenyan et T. W. Schultz. Interspecies modeling of narcotic toxicity to aquatic animals. *Bulletin of Environmental Contamination and Toxicology*.**2000**, 65, 399-406.
- [43]. S.D. Dimitrov, O.G. Mekenyan, G. D. Sinks et T. W. Schultz. Global modeling of narcotic chemicals: ciliate and fish toxicity. *Journal of Molecular Structure: THEOCHEM*, **2003**, 622, 63-70.
- [44]. H. Schmitt, R. Altenburger, B. Jastorff et G.Schüürmann, Quantitative structure-activity analysis of the algae toxicity of nitroaromatic compounds. *Chemical Research in Toxicology*,**2000**, 13,441-450.
- [45]. H. Huang, X. Wang, W. Ou, J. Zhao, Y. Shao et L. Wang. Acute toxicity of benzene derivatives to the tadpoles (*Rana japonica*) and QSAR analyses. *Chemosphere*, **2003**, 53, 963-970.
- [46]. M.T.D. Cronin, T.I. Netzeva, J.C. Dearden, R.EdwardsetA.D.P. Worgan. Assessment and Modeling of the Toxicity of Organic Chemicals to *Chlorella vulgaris*: Development of a Novel Database. *Chemical Research in Toxicology*,**2004**, 17, 545-554.
- [47]. T.I. Netzeva, J.C. Dearden, R. Edwards, A.D.P. WorganetM.T.D. Cronin. QSAR analysis of the toxicity of aromatic compounds to *Chlorella vulgaris* in a novel short-term assay. *Journal of Chemical Information and Computer Sciences*, **2004**, 44, 258-

265.

- [48]. T.W. Schultz, G.D. Sinkset A.P. Bearden. QSAR in aquatixotoxicology: a mechanism of action approach comparing toxic potency to *Pimephalespromelas*, *T. pyriformis*, and *V. fischeri*. In: Devillers, J. (Ed.), *Comparative QSAR*. Taylor & Francis New York, **1998**, 51-109.
- [49]. F.R. Burden. Quantitative structure-activity relationship studies using Gaussian processes. *Journal of Chemical Information and Computer Sciences*, **2001**, 41, 830-835.
- [50]. J. A. Grodnitzky et J. R. Coats. QSAR evaluation of monoterpenoids' insecticidal activity. *Journal of Agricultural and Food Chemistry*, **2002**, 50, 4576-4580.
- [51]. J. Huuskonen. QSAR modeling with the electrotopological state indices: predicting the toxicity of organic chemicals. *Chemosphere*, **2003**, 20, 949-953.
- [52]. K. Rose et L. H. Hall. E-state modelling of fish toxicity independent of 3D structure information. *SAR and QSAR in Environmental Research*, **2003**, 14, 113-129.
- [53]. Ziani Nadia. Modèles QSAR pour la prédiction de la toxicité aquatique : D'alcools et d'amines vis-a-vis de *Tetrahyemapyriformis*. De dérivés benzéniques substitués vis-à-vis de *Pemiphalespromelas*. Thèse de Doctorat. Université Badji Mokhtar Annaba. **2016**.
- [54]. P. Hohenberg et W. Kohn. Inhomogeneous Electron Gas, *physical review*, **1964**, 136, B864 –B871.
- [55]. N. L. Allinger. Calculation of Molecular Structure and Energy by Force-Field Methods, *Advances in physical organic chemistry*, **1976**, 13, 1-82.
- [56]. R. Niketic et S. K. Rasmussen. *The Consistent Force Field: A Documentation*, **1977**, Springer, Berlin.
- [57]. U. Burbert et N. L. Allinger. *Molecular Mechanics*. **1982**, American Chemical Society, Washington.
- [58]. J. S. Lomas. La Mécanique Moléculaire, une Méthode non Quantique pour le Calcul de la Structure et de l'énergie d'entités Moléculaire, *L'actualité chimique*, **1986**, 22, 7 – 20.
- [59]. W. Kolos et L. Wolniewicz .Accurate Adiabatic Treatment of the Ground State of the Hydrogen Molecule, *The Journal of Chemical Physics*, **1964**, 41, 3663.
- [60]. B. T. Sutcliffe. The Nuclear Motion Problem in Molecular Physics, *Advances in Quantum Chemistry*, **1997**, 28, 65 - 80.
- [61]. C. C. J. Roothan. New Developments in Molecular Orbital Theory, *Reviews of*

Modern Physics, **1951**, 23, 69 - 89

- [62]. G. G. Hall. The Molecular Orbital Theory of Chemical Valency VIII: A Method of Calculating Ionization Potentials, *Proceedings of the Royal Society of London A*, **1951**, 205, 541- 552.
- [63]. T. A. Koopmans. The Distribution of Wave Function and Characteristic Value Among the Individual Electrons of an Atom, *Physica*, **1933**, 1, 104-113.
- [64]. S. M. Blinder. Basic Concepts of Self-Consistent-Field Theory, *American Journal of Physics*, **1965**, 33, 431 - 443.
- [65]. P. O. Löwdin. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals, *The Journal of Chemical Physics*, **1950**, 18, 365 - 375.
- [66]. P. O. Löwdin. On the Orthogonality Problem. *Advances in Quantum Chemistry*, **1970**, 5, 185-199
- [67]. R. S. Mulliken. Electronic Population Analysis on LCAO-MO Molecular Wave Functions.(I), *The Journal of Chemical Physics*, **1955**, 23, 1833- 1840.
- [68]. R. S. Mulliken. Electronic Population Analysis on LCAO-MO Molecular Wave Functions.(II). Populations, Bond Orders, and Covalent Bond Energies, *The Journal of Chemical Physics*, **1955**, 23, 1841-1846.
- [69]. M. J. S. Dewar et W. Thiel. Ground States of Molecules. 38. The MNDO Method. Approximations and Parameter, *Journal of the American Chemical Society*, **1977**, 99, 4899-4907.
- [70]. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy et J. J. P. Stewart. The Development and Use of Quantum Mechanical Molecular Models. 76. AMI: a New General Purpose Quantum Mechanical Molecular Model, *Journal of the American Chemical Society*, **1985**, 3902-3909.
- [71]. J. J. P. Stewart. Optimization of Parameters for Semiempirical Methods I. Method, *Journal of computational chemistry*, **1989**, 10, 209-220.
- [72]. M. J. S. Dewar et W. Thiel. A Semiempirical model for the two-center repulsion integrals in the NDDO approximation, *Theoretica Chimica Acta*, **1977**, 46, 89-104.
- [73]. N.L. Allinger. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms, *Journal of the American Chemical Society*, **1977**, 99, 8127-8134.
- [74]. U. Burkert et N.L. Allinger. Molecular Mechanics, ACS Monograph No177, American Chemical Society, **1982**, **1986**, Washington, DC.

- [75]. N. L. Allinger, Y. H. Yu et J. H. Lii. Molecular mechanics. The MM3 force field for hydrocarbons. 1, *Journal of the American Chemical Society*, **1989**, 111, 8551-8566.
- [76]. N. L. Allinger, K. J. H. Chen et J. H. Lii. An improved force field (MM4) for saturated hydrocarbons, *Journal of computational chemistry*, **1996**, 17, 642-668.
- [77]. A. D. Jr Mc Kerell, D. Bashford, M. Bellott, R. L. Jr ; J. D. Dunbrack, M.J. Evanseck Field, S. Fischer, J. Gao, H. Guo, S. Ha, J. D. Mc Carthy, L. Kuchnir; K. Ruczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W. E. Reiher , B. Roux, D. Schlenkrich, M. Smith, J.C. Stote, R. Stramb, J. Watanabe, M. Wiokiewicz- Kuczera, J.Yin et M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins, *The Journal of Physical Chemistry B*, **1998**, 102, 3586-3616.
- [78]. T. A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, *Journal of computational chemistry*, **1996**, 17, 490-519.
- [79]. T. A. Halgren. *Journal of computational chemistry*, **1996**, 17, 520-552, 553-586.
- [80]. T. A. Halgren et R. B. Nachbar. Merck molecular force field. IV. conformational energies and geometries for MMFF94, *Journal of computational chemistry*, **1996**, 17, 587-615.
- [81]. K. I. Ramachandran, G. Deepa et K. Namboori. Computational Chemistry and Molecular Modeling. Principles and Applications. **2008**. Springer-Verlag, Heidelberg.
- [82]. Y. Dodge et V. Rousson. Analyse de régression appliquée. **2004**. Dunold, Paris.
- [83]. D.C. Montgomery et E. A. Peck. Introduction to Linear Regression Analysis. **1992**. John Wiley & Sons. Inc.
- [84]. N. R. Draper et H. Smith. Applied Regression Analysis. 1998. John Wiley & Sons, Inc.
- [85]. G. Dreyfus, et al. Réseaux de neurones, méthodologie et applications. Paris: Eyrolles, 2ème édition, 2004.
- [86]. W.S. McCulloch, et W.A. Pitts, logical calculus of ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 1943, 5, 115-133
- [87]. Aurélie GOULON-SIGWALT-ABRAM. Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments. Thèse de doctorat université Pierre et Marie Curie Paris 6 2008
- [88]. F. Badran, S. Thria et L. Héroult. Réseaux de neurones: Méthodologie et

- applications. **2004**. Editions Eyrolles.
- [89]. R. Hecht-Nielsen. Neurocomputing. **1990**. Addison-Wesley Publishing Company.
- [90]. MATLAB. Version 7.0.0. (Release 14). **2004**. The Language of Technical Computing. The Math Works, Inc.
- [91]. H. Kubinyi. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm, *Quantitative Structure-activity Relationships*, **1994**, 13, 285–294.
- [92]. L. Erikson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell et P. Gramatica. Methods for Reliability and uncertainty Assessment and for Applicability Evaluations of Classification and Regression Based QSARs. *Environmental and Health Perspectives*. **2003**, 111, 1361-1375.
- [93]. A. Tropsha, P. Gramatica et V. K. Grombar. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, **2003**, 22, 69-76.
- [94]. B. K. Sharma, P. Singh, P. Pilania, K. Sarbhai, S. Yenamandra et C. P. Prabhakar. CP-MLR/PLS directed QSAR study on apical sodium-codependent bile acid transporter inhibition activity of benzothiepinines, *Molecular Diversity*, **2011**, 15, 135–147.
- [95]. A. Golbraikh et A. Tropsha. Beware of q^2 , *Journal of Molecular Graphics and Modelling*, **2002**, 20, 269-276.
- [96]. A. Bouakkadia. Modélisation de quelques propriétés (cteH, S, Pv, Koc(w)) contrôlant l'évolution dans l'environnement d'une série d'herbicides. Thèse de Doctorat. Université Badji Mokhtar Annaba. **2016**.
- [97]. G. Persoone, D. Dive, 1978. *Ecotox. Environ. Saf.*, 2, 105-144.
- [98]. Hakim Hamada. Relation quantitative structure/activité d'une série de phénol. Mémoire de Magister. Université Badji Mokhtar Annaba. **2007**.
- [99]. T. W. Schultz, D. T. Lin, T. S. Wicke et L. M. Arnold. Quantitative Structure – Activity Relationships for the *Tetrahymena Pyriformis* Population Growth Endpoint : a Mechanism of Action Approach, in : W. Karcher, , Practical Applications of Quantitative Structure – Activity Relationships (QSAR) In Environmental Chemistry and Toxicology, J. Devillers (eds), Kluwer Academic Publishers, Dordrecht, **1990**, 241 – 262.
- [100]. T.W. Schultz, M. Cajina-Quezada, J.N. Dumont, 1980. *Arch. Environ. Contam. Toxicol.* 9, 591-598.
- [101]. T.W. Schultz, M. Cajina-Quezada, 1982. *Arch. Environ. Contam. Toxicol.* 11,

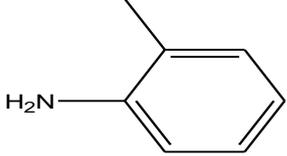
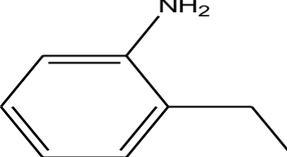
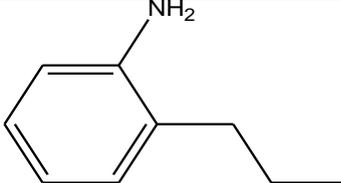
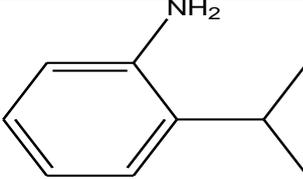
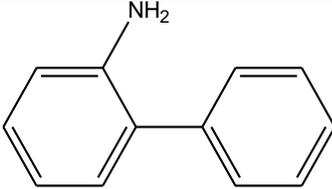
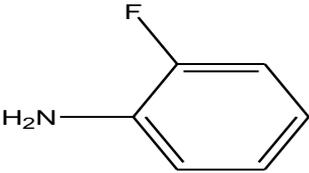
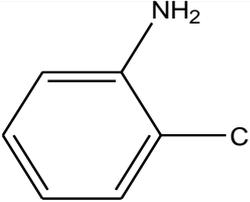
353-361.

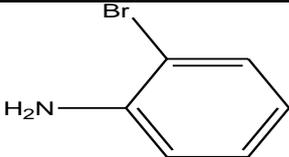
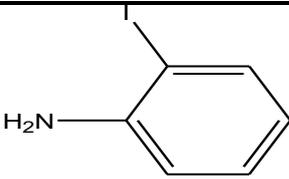
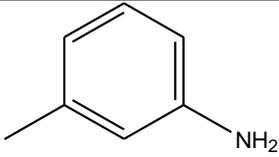
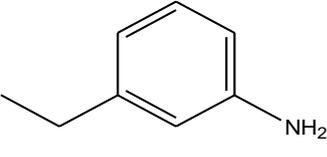
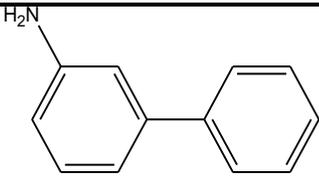
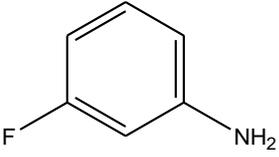
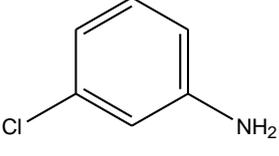
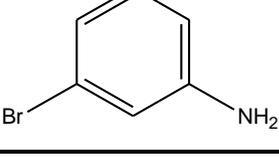
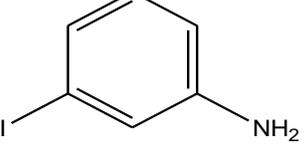
- [102]. Hyperchem™ Release 7, Hypercube for windows, Molecular Modeling system, **2000**.
- [103]. R. Todeschini, V. Consonni, M. Pavan, DRAGON Software for the Calculation of Molecular Descriptors, Release 5.4 for windows, Milano, **2006**.
- [104]. R. Leardi, R. Boggia et M. Terrile. Genetic Algorithms as a Strategy for Feature Selection, *Journal of Chemometrics*, **1992**, 6, 267 – 281.
- [105]. R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan, MOBYDIGS, version 1.1, Copyright TALETE srl, **2009**.
- [106]. M. Pavan, A. Mauri et R. Todeschini. Total Ranking Models by the Genetic Algorithm Variable Subset Selection (GA–VSS) Approach for Environmental Priority Settings, *Analytical and Bioanalytical Chemistry*, **2004**, 380, 430 – 444.
- [107]. S. Wold et L. Eriksson, Statistical Validation of QSAR Results In Chemometrics Methods in Molecular Design, H. Van de Waterbeemd., VCH Publishers, New York, **1995**, pp. 309 – 318.
- [108]. V. Consonni, R. Todeschini et M. Pavan, Structure / Response Correlations and Similarity / Diversity Analysis by GETAWAY Descriptors. 1– Theory of the Novel 3D Molecular Descriptors, *Journal of Chemical Information and Modeling*, **2002**, 380, 682 – 692.
- [109]. G. A. Arteca, Analysis of Shape Transitions Using Molecular Size Descriptors Associated with Inner and Outer Regions of a Polymer Chain, *Journal of Molecular Structure THEOCHEM*, **2003**, 630, 113 – 123.
- [110]. C. Hansch et A. Leo. Substituent Constants for Correlations in Chemistry and Biology. **1979**, Wiley Interscience, New York.
- [111]. CLOGP Pomona College and BioByte, Inc. of Claremont, CA. (1988-2018)
- [112]. F. E. Norrington et R. M. Hyde, S. G. Williams et R. Wootton. Physicochemical-Activity Relations in Practice. 1. Rational and Self-Consistent Data Bank, *Journal of Medicinal Chemistry*, **1975**, 18, 604-607.
- [113]. Kier–Hall, E–calc version 1.1, **1999**.
- [114]. L. B. Kier, L. H. Hall, Molecular Structure Description: The Electrotopological State, Academic Press, San Diego, **1999**.
- [115]. L. B. Kier, L. H. Hall, B. K. Mohny, The electro–topological state: An atom index for QSAR, *Quant. Struct.-Act. Relat.* **1991**, 10, 43–51.

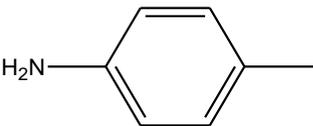
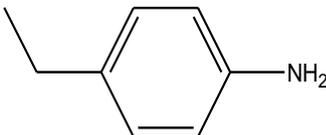
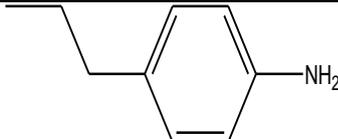
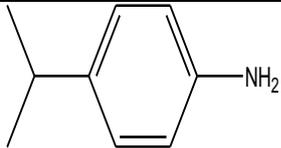
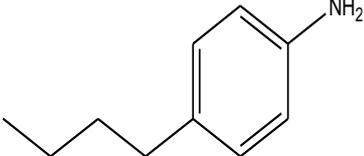
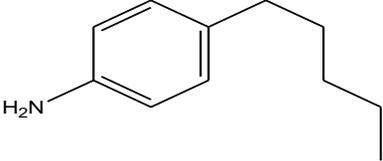
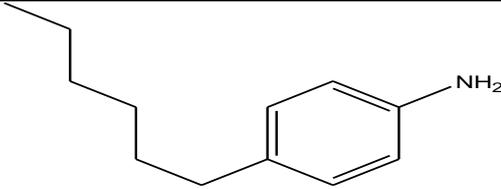
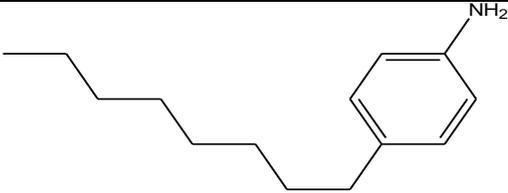
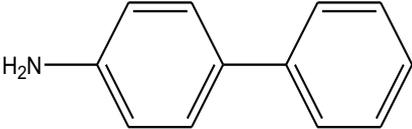


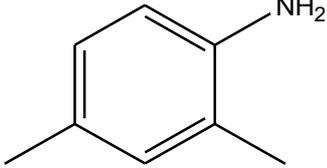
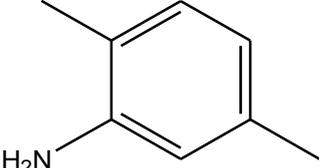
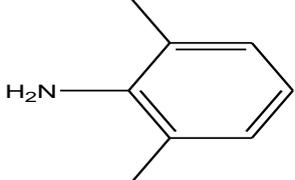
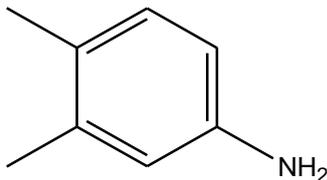
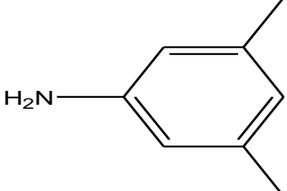
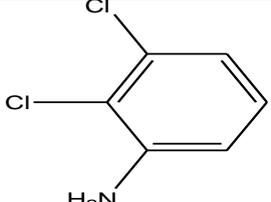
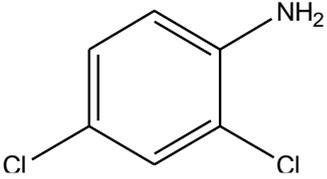
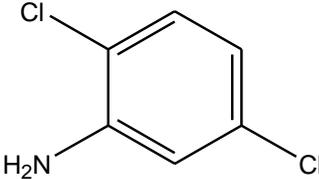
ANNEXES

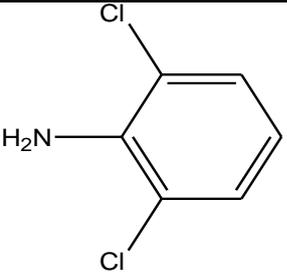
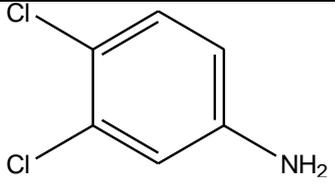
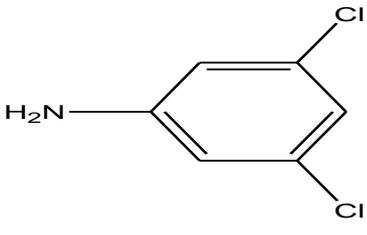
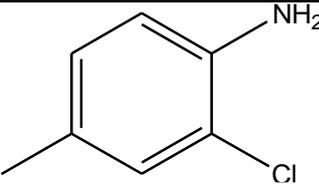
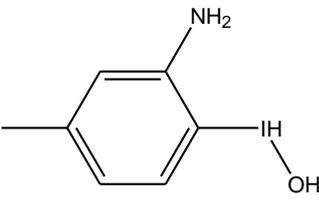
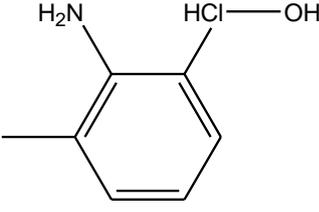
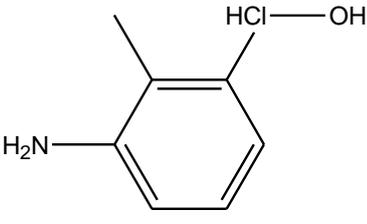
Tableau A-I : Structures des composés anilines

N°	Composés et numéro de CAS	structure
1	2-Methylaniline 95-53-4	
2	2-Ethylaniline 578-54-1	
3	2-Propylaniline 1821-39-2	
4	2-Isopropylaniline 643-28-7	
5	2-Phénylaniline 90-41-5	
6	2-Fluoroaniline 348-54-9	
7	2-Chloroaniline 95-51-2	

8	2-Bromoaniline 615-36-1	
9	2-Iodoaniline 615-43-0	
10	3-Methylaniline 108-44-1	
11	3-Ethylaniline 587-02-0	
12	3-Phénylaniline 2243-47-2	
13	3-Fluoroaniline 372-19-0	
14	3-Chloroaniline 108-42-9	
15	3-Bromoaniline 591-19-5	
16	3-Iodoaniline 626-01-7	

17	4-Méthylaniline 106-49-0	
18	4-Ethylaniline 589-16-2	
19	4-Propylaniline 2696-84-6	
20	4-Isopropylaniline 99-88-7	
21	4-Butylaniline 104-13-2	
22	4-Pentylaniline 33228-44-3	
23	4-Hexylaniline 33228-45-4	
24	4-Octylaniline 16245-79-7	
25	4-Phénylaniline 92-67-1	

26	2,4-diméthylaniline 95-68-1	
27	2-5-diméthylaniline 95-78-3	
28	2,6-diméthylaniline 87-62-7	
29	3,4-diméthylaniline 95-64-7	
30	3,5-diméthylaniline 108-69-0	
31	2,6-dichloroaniline 608-31-1	
32	2,4-dichloroaniline 554-00-7	
33	2,5-dichloroaniline 95-82-9	

34	2,6-dichloroaniline 608-31-1	
35	3,4-dichloroaniline 89059-40-5	
36	3,5-dichloroaniline 626-43-7	
37	2-chloro-4-méthylaniline 615-65-6	
38	2-iodo-5-méthylaniline 13194-69-9	
39	2-chloro-6-méthylaniline 87-63-8	
40	3-chloro-2-méthylaniline 87-60-5	

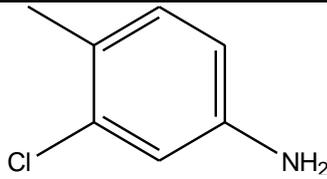
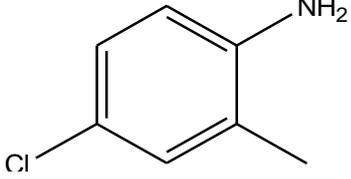
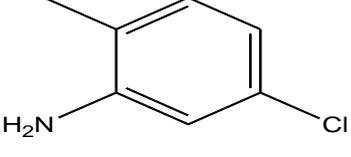
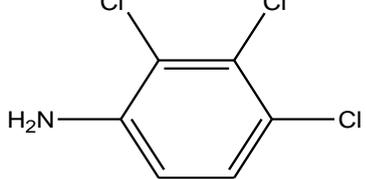
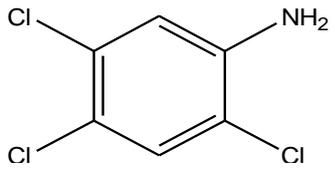
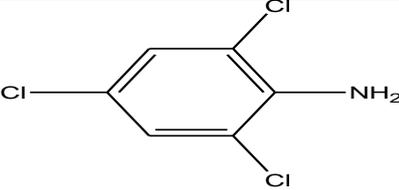
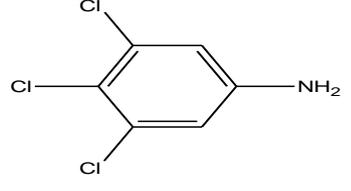
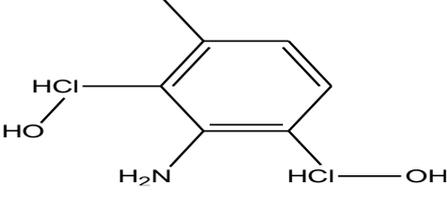
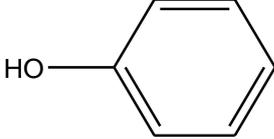
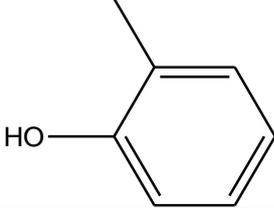
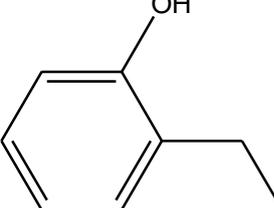
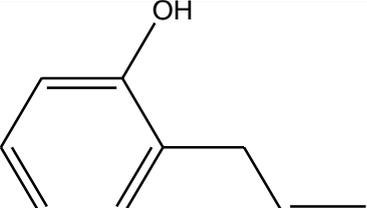
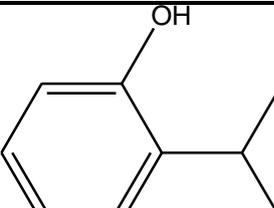
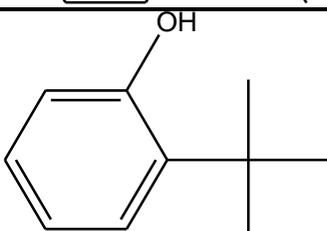
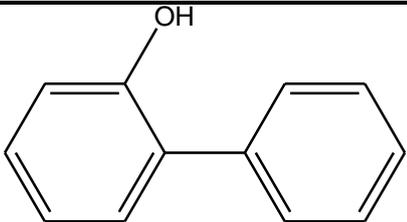
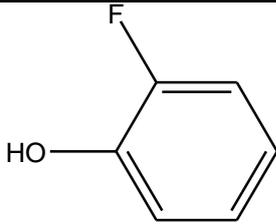
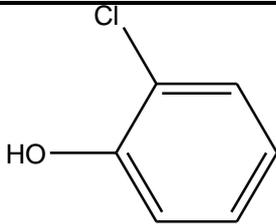
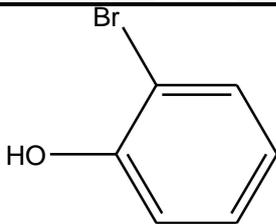
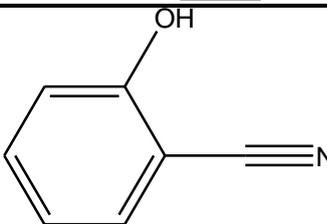
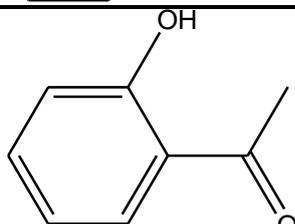
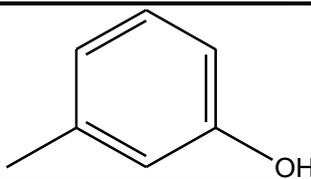
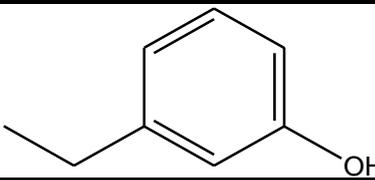
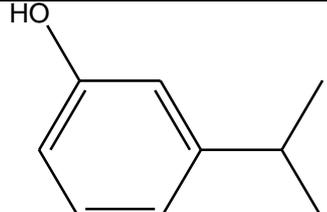
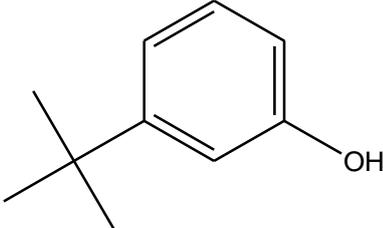
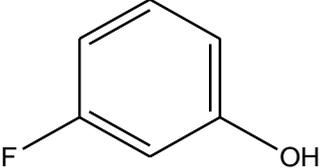
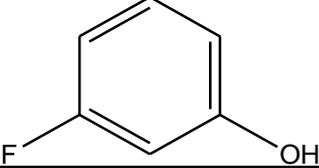
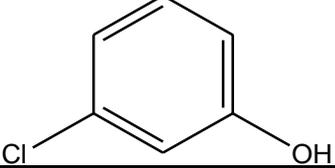
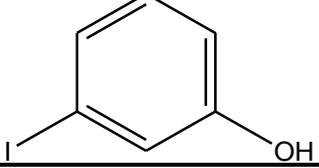
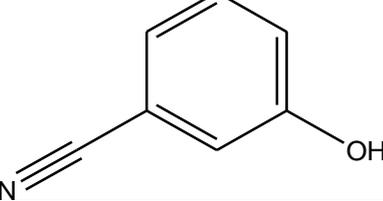
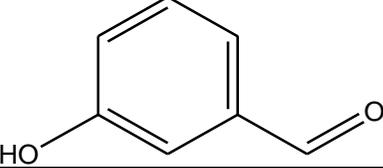
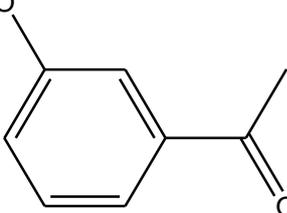
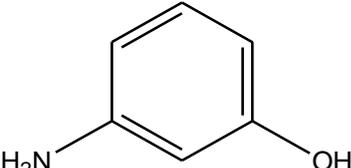
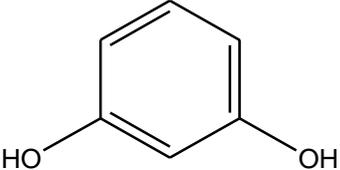
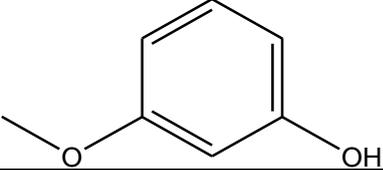
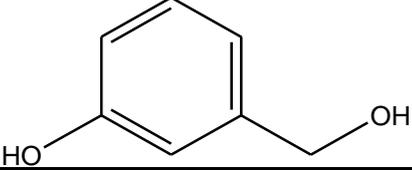
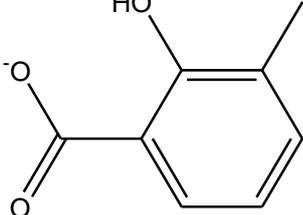
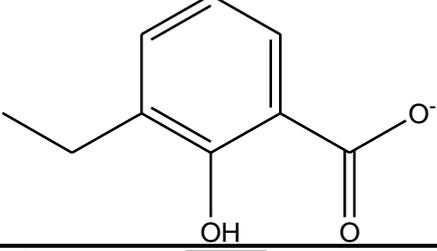
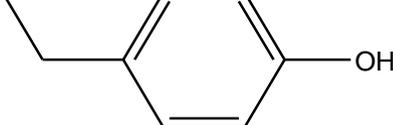
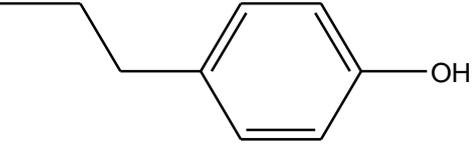
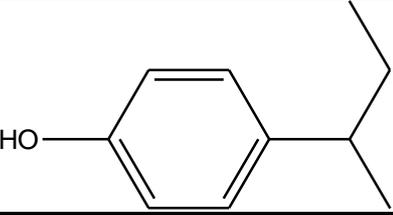
41	3-chloro-4-méthylaniline 33240-95-8	
42	4-chloro-2-méthylaniline 95-69-2	
43	5-chloro-2-méthylaniline 95-79-4	
44	2,3,4-trichloroaniline 634-67-3	
45	2,4,5-trichloroaniline 636-30-6	
46	2,4,6-trichloroaniline 634-93-5	
47	3,4,5-trichloroaniline 634-91-3	
48	2,6-dichlorol-3-méthylaniline 64063-37-2	

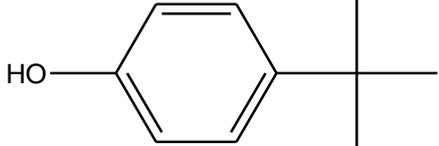
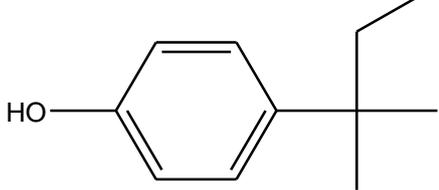
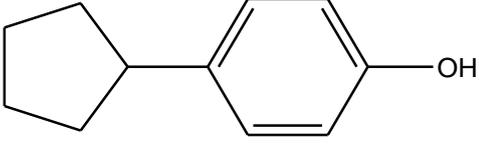
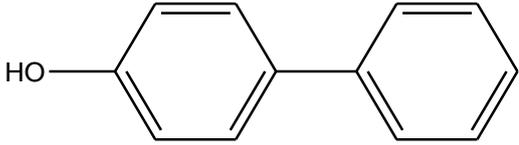
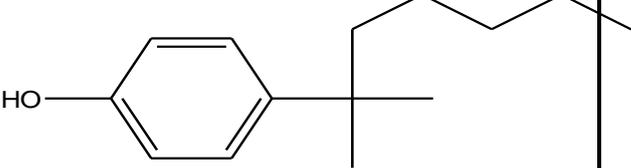
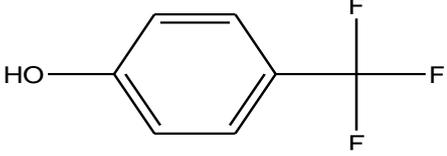
Tableau A-II : Structures des composés phénols

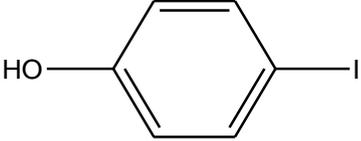
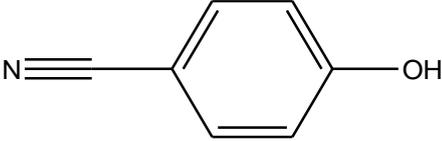
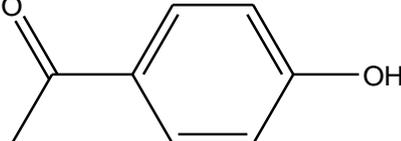
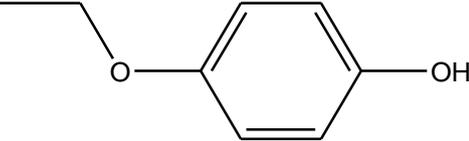
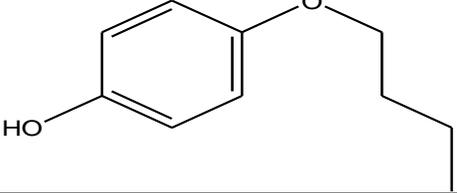
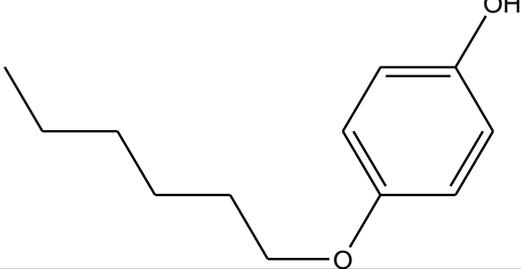
N°	Composés et numéro de CAS	structure
1	Phénol 108-95-2	
2	2-méthylphénol 95-48-7	
3	2-éthylphénol 90-00-6	
4	2-allylphénol 1745-81-9	
5	2-isopropylphénol 88-69-7	
6	2-(tert)butylphénol 88-18-6	
7	2-phénylphénol 287389-48-4	

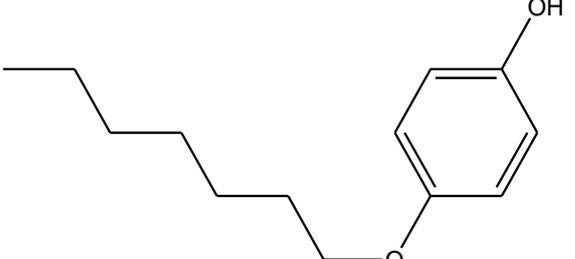
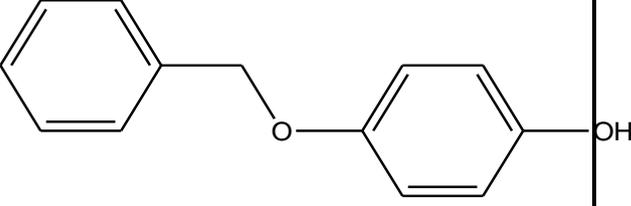
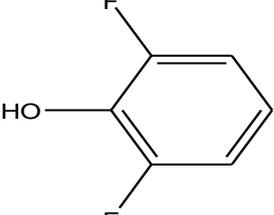
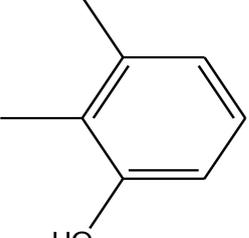
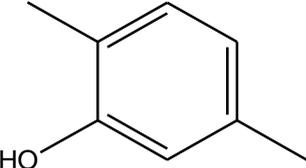
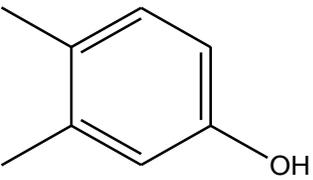
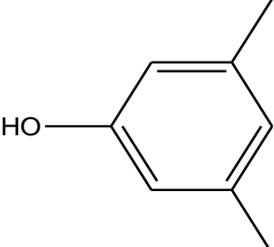
8	2-fluorophénol 367-12-4	
9	2-chlorophénol 95-57-8	
10	2-bromophénol 253429-15-1	
11	2-cyanophénol 73289-85-7	
12	2-acétylphénol 118-93-4	
13	3-méthylphénol 108-39-4	
14	3-éthylphénol 620-17-7	
15	3-isopropylphénol 618-45-1	

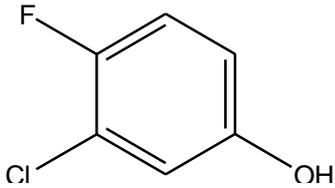
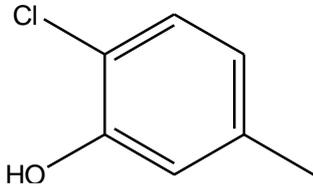
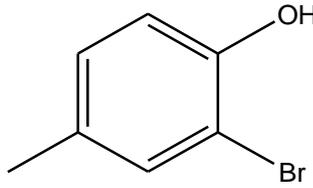
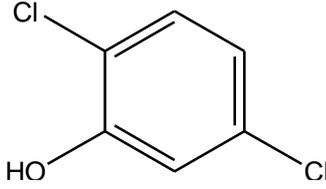
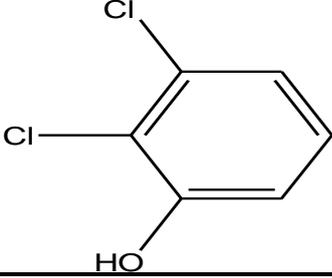
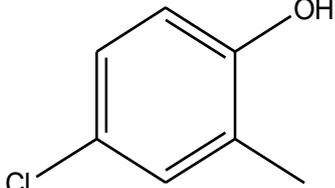
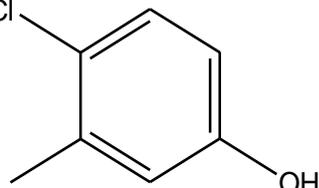
16	3-(tert)butylphénol 585-34-2	
17	3-phénylphénol 580-51-8	
18	3-fluorophénol 372-20-3	
19	3-chlorophénol 108-43-0	
20	3-iodophénol 626-02-8	
21	3-cyanophénol 873-62-1	
22	3-hydroxybenzaldéhyde 100-83-4	
23	3-acétylphénol 121-71-1	
24	3-aminophénol 591-27-5	

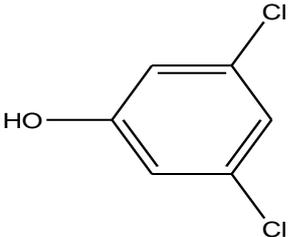
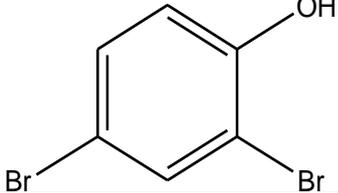
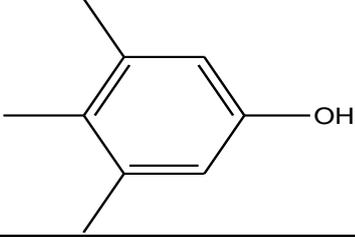
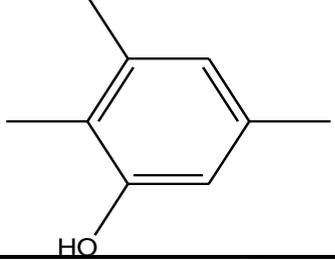
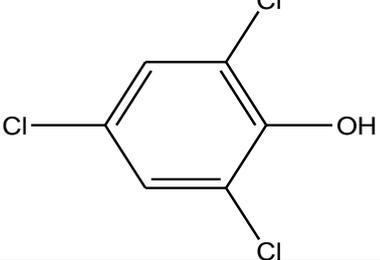
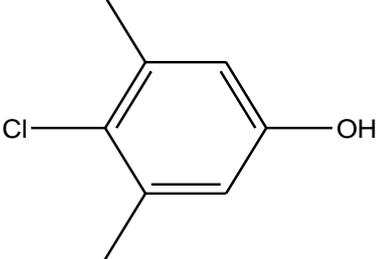
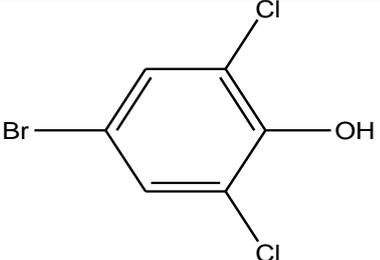
25	3-hydroxyphénol 108-46-3	
26	3-méthoxyphénol 150-19-6	
27	3-hydroxybenzylalcool 620-24-6	
28	3-méthylhydroxybenzoate 19438 -10 - 9	
29	3-éthylhydroxybenzoate 7781 -98 - 8	
30	4-éthylphénol 123-07-9	
31	4-propylphénol 645- 56 - 7	
32	4-isopropylphénol 99 -89- 8	
33	4-(sec)butylphénol 99-71-8	

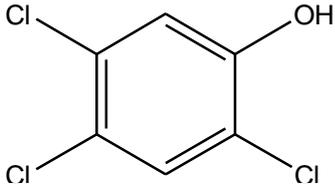
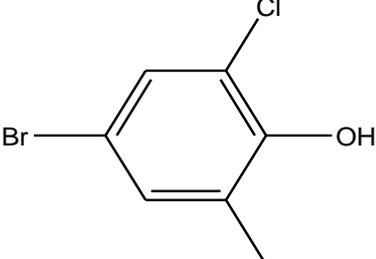
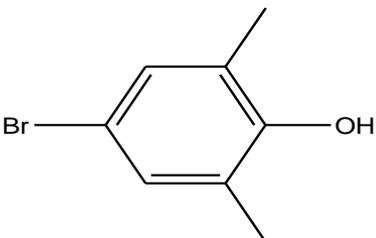
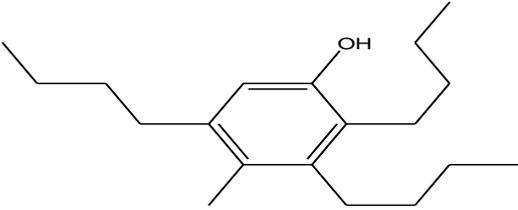
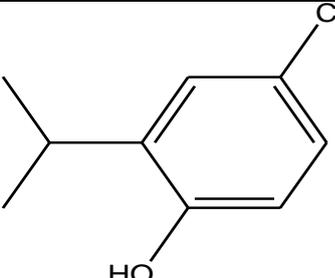
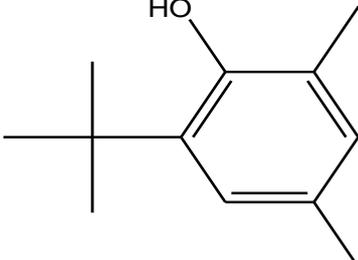
34	4-(tert)butylphénol 98 -54 -4	
35	4-(tert)pentylphénol 80 -46 -6	
36	4-cyclopentylphénol 1518-83 -8	
37	4-phénylphénol 92- 69- 3	
38	4-(tert)octylphénol 140-66-9	
39	$\alpha\alpha\alpha$ -trifluoro-4-crésol 402-45-9	
40	4-fluorophénol 371-41-5	
41	4-chlorophénol 106 -48-9	

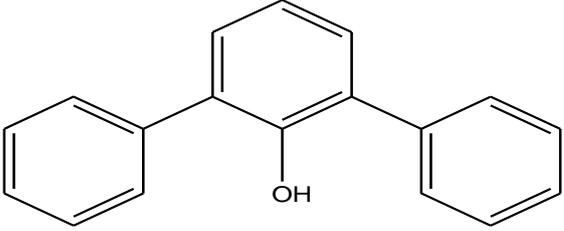
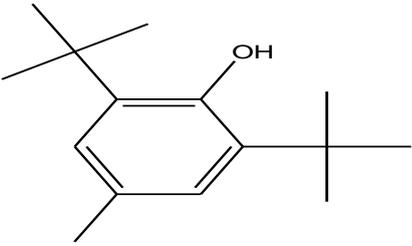
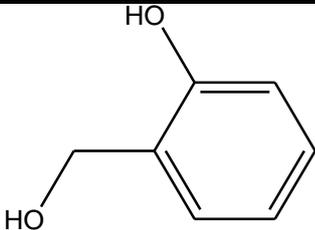
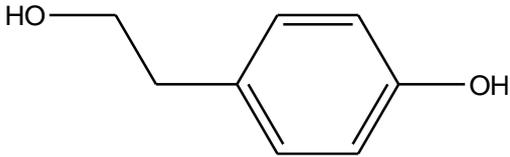
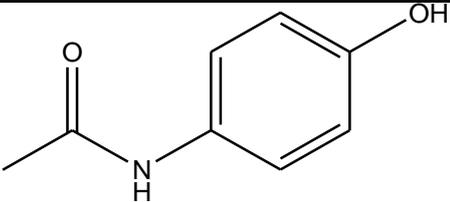
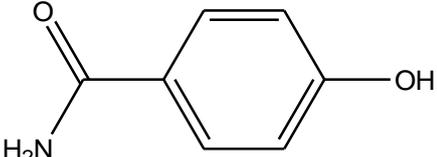
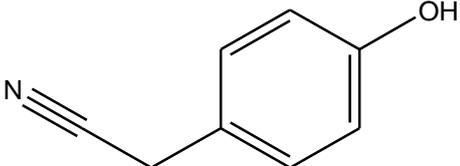
42	4-bromophénol 106 -41 - 2	
43	4-iodophénol 540-38-5	
44	4-cyanophénol 767 -00- 0	
45	4-acétylphénol 99 - 93 - 4	
46	4-méthoxyphénol 150 - 76 -5	
47	4-éthoxyphénol 622 -62 -8	
48	4-butoxyphénol 122 -94 -1	
49	4-hexyloxyphénol 18979 -55 -0	

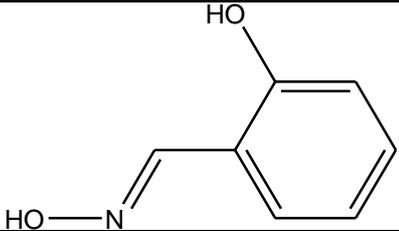
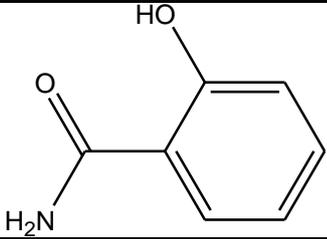
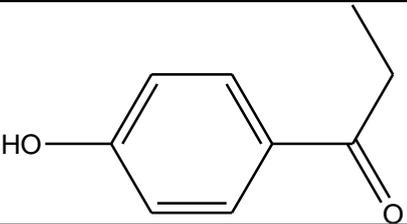
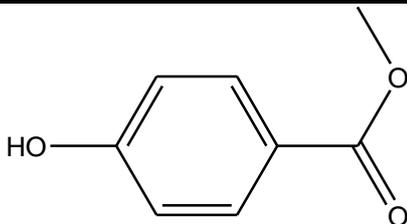
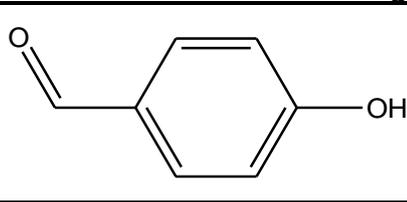
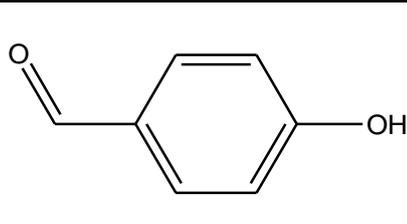
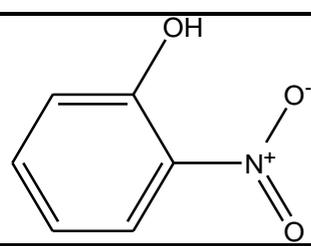
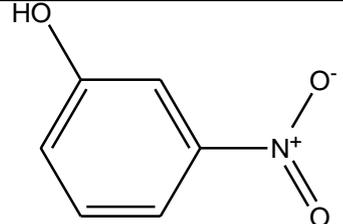
50	4-heptyloxyphénol 13037 - 86 -0	
51	4-benzyloxyphénol 103-16-2	
52	2,6-difluorophénol 6418 -38-8	
53	2,3-diméthylphénol 526 - 75 - 0	
54	2,5-diméthylphénol 95 - 87 -4	
55	3,4-diméthylphénol 95 -65 - 8	
56	3,5-diméthylphénol 108 - 68 -9	

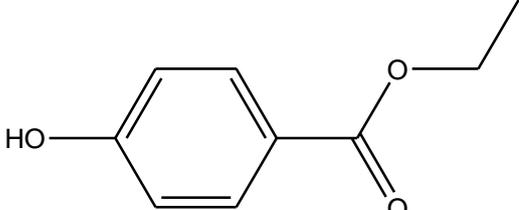
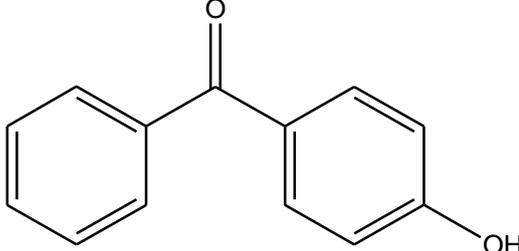
57	3-chloro-4-fluorophénol 2613 - 23-2	
58	2-chloro-5-méthylphénol 615 - 74 -7	
59	2-bromo-4-méthylphénol 6627 -55 -0	
60	2,5-dichlorophénol 583 -78 -8	
61	2,3-dichlorophénol 576 - 24 -	
62	4-chloro-2-méthylphénol 1570 - 64 -5	
63	4-chloro-3-méthylphénol 59- 50-7	

64	3,5-dichlorophénol 591 -35 -5	
65	2,4-dibromophénol 615 - 58 -7	
66	3,4,5-triméthylphénol 527 - 54- 8	
67	2,3,5-triméthylphénol 697- 82-5	
68	2,4,6-trichlorophénol 88 -06 - 2	
69	4-chloro-3,5-diméthylphénol 88 - 04 -0	
70	4-bromo-2,6-dichlorophénol 697 - 86 -9	

71	2,4,5-trichlorophénol 95-95 - 4	
72	4-bromo-6-chloro-2-méthylphénol 7530 -27 -0	
73	4bromo-2,6-diméthylphénol 697 - 86 - 9	
75	2-(ter) butyl-4-méthylphénol 2409 -55 -4	
76	4-chloro-2-isopropylphénol 89 - 68 - 9	
77	6-(tert) butyl-2,4-diméthylphénol 1879-09-0	

78	2,6-diphénylphénol 2432-11-3	
79	2,6-di(tert)butyl-4-méthylphénol 128-37-0	
80	2-hydroxybenzylalcohol 90-01-7	
81	4-hydroxyphénethylalcohol 501-94-0	
82	4-acetoamidophénol 103-90-2	
83	4-hydroxybenzamide 619-57-8	
84	4-hydroxybenzylcyanide 14191-95-8	

85	2-hydroxybenzaloxime 94-67-7	
86	2-hydroxybenzamide 65-45-2	
87	4'-hydroxypropiophenone 70-70-2	
88	methyl-4-hydroxybenzoate 99-76-3	
89	4-hydroxybenzaldehyde 123-08-0	
90	2-hydroxybenzaldehyde 90-02-8	
91	2-nitrophénol 88-75-5	
92	3-nitrophénol 554-84-7	

93	ethyl-4-hydroxybenzoate 120-47-8	
94	4-hydroxybenzophenone 1137-42-4	
95	4-hydroxyphénylmethane 88170-17-6	

Inhibition of Microbial Growth by Anilines: A QSAR Study

Ahmed Bouaoune, Leila Lourici, Hamza Haddag and Djelloul Messadi
Environmental and Food Safety Laboratory, Badji Mokhtar University, Annaba 23000, Algeria

Received: November 3, 2011 / Accepted: January 9, 2012 / Published: May 20, 2012.

Abstract: The relative toxicity of 48 anilines using the *Tetrahymena pyriformis* population growth characteristics IGC_{50} (concentration causing 50% growth inhibition), available in the literature, was studied. At first, the entire data set was randomly split into a training set (31 chemicals) used to establish the QSAR model, and a test set (17 chemicals) for statistical external validation. A biparametric model was developed using, as independent variables, 3D theoretical descriptors derived from DRAGON software. The GA-MLR (genetic algorithm variable subset selection) procedure was performed on the training set by the software mobydigs using the OLS (ordinary least squares) regression method, and GA (genetic algorithm)-VSS (variable subset selection) by maximising the cross-validated explained variance (Q_{LOO}^2). The obtained model was examined for robustness (Q_{LOO}^2 cross-validation, Y-scrambling) and predictive ability through both internal (Q_{LMO}^2 , bootstrap) and external validation (Q_{ext}^2) methods. Descriptors included in the QSAR model indicated that $\log IGC_{50}^{-1}$ value was related to molecular size and shape, and interaction of molecule with its surrounding medium or its target. Moreover, the applicability domain of the model was discussed.

Key words: Toxic agents, growth of microbial species, QSAR hybrid model, statistical external validation, applicability domain.

1. Introduction

Recently computational methods have been used to solve complex problems in many aspects of science. One particularly useful method—the development of QSARs (quantitative structure-activity relationships) has found application in environmental chemistry and ecotoxicology [1-5].

QSAR approach systematization which has to be associated to the work of Hansch and Fujita in 1964 [6] is based on the assumption that the structure of a molecule must contain the features responsible for its physical, chemical and biological properties and on the possibility of representing a molecule by numerical descriptors.

The underlying hypothesis for QSAR models is that all molecules interact with the receptor in same or similar mode of action [7].

The descriptors most used in the early QSAR analyses are the octanol/water partition coefficient ($\log P$), the Hammett δ constant [8, 9] acting as an electronic effect descriptor and the lipophilicity parameter π , which is defined by analogy to the electronic descriptor. Together with these empirical descriptors, the classical models employ other physical-chemical properties as parameters; some of them derived from quantum chemical calculations, namely: partial charges, HOMO/LUMO energies, etc..

An important topic in environmental chemistry and ecotoxicology consists of the effect of toxic agents on the growth of microbial species. The population growth of protozoa, in particular of ciliates, in varied concentrations of toxic substances has been assessed by comparing a number of specific experimental values, including population density [10], growth rates [11], growth curves [12] and number of generations [13].

Corresponding author: Leila Lourici, Ph.D., main research field: environmental chemistry. E-mail: leilalourici@yahoo.fr.

In all cases, the most tested species has been *Tetrahymena pyriformis*, a common freshwater hymenostome ciliate, which approximatively measures 50 μm in length and 30 μm in width [14]. Modern electronic equipment allows the easy determination of the population growth inhibition, providing a large collection of data for toxicological research.

Schultz et al. [15] evaluated the relative toxicity of 48 selected anilines using the *Tetrahymena pyriformis* population growth characteristics IGC_{50} (concentration causing 50% growth inhibition) as an endpoint. The authors shown that simple $pIGC_{50}^{-1}$ ($= \log IGC_{50}^{-1}$) versus $\log P$ correlation can model environmental toxicity. The predictability of this $\log P$ dependent QSAR can be improved with the addition of $\sum \delta$ (the summation of the substituent electronic parameter δ), as a second and orthogonal descriptor. The statistical parameters reported by the mentioned authors are only related to the fitting performances.

In 1988, QSAR techniques suffered a great transformation due to the introduction of the so-called three dimensional (3D) molecular parameters, which accounted for the influence of different conformers, stereoisomers or enantiomers.

Several principles for assessing the validity of QSARs were proposed in 2002, as the "Setubal Principles" [16]; these were then modified in 2004 as the OECD (Organisation for Economic Co-operation and Development) Principles for QSAR validation [17]. To facilitate the consideration of a QSAR model for regulatory purpose, it should be associated with the following information: (a) a defined endpoint; (b) an unambiguous algorithm; (c) a defined applicability domain (AD); (d) appropriate measures of goodness of fit, robustness and predictivity; and (e) a mechanistic interpretation, if possible. Thus, further QSAR development on anilines should follow these guidelines.

In this study a biparametric model for the toxicity

of aniline derivatives was developed using, as independent variables, 3D theoretical descriptors calculated from the chemical structure alone (Geometrical and GETAWAY descriptors). The available data set (taken from Schultz et al. [15]) was randomly split into training set (31 objects), used to develop the QSAR model, and a validation set (17 objects), used only for statistical external validation.

The model was examined for robustness and predictive ability through both internal and external validation methods. Finally, the QSAR applicability domain was discussed by the Williams plot of standardized residuals versus leverage values [18, 19].

2. Methodology

2.1 Descriptors Generation

The structures of the molecules were drawn using Hyperchem 6.03 software [20]. The final geometries were obtained with the semi empirical method AM1. All calculation were carried out at the RHF (restricted Hartree-Fock) level with non configuration interaction. The molecular structures were optimized using the algorithm Polak-Ribiere and a gradient norm limit of 0.001 kcal/Å. The resulted geometry was transferred into the software Dragon version 5.3 [21] to calculate 271 descriptors of the type Geometrical and GETAWAY (Geometry, Topology and Atoms Weighted Assembly). Descriptors with constant or near constant values inside each group were discarded. For each pair of correlated descriptors (with correlation coefficient $r \geq 0.95$), the one showing the highest pair correlation with the other descriptors was excluded.

The GA (Genetic Algorithm) [22] has been considered superior to other methods of variable selection techniques. So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini [23] by maximizing the cross-validated explained variance Q_{LOO}^2 .

2.2 Model Development and Validation

Multiple linear regression analysis and variable selection were performed by package MobyDigs for windows/PC [23], using OLS (ordinary least squares regression) method and, as previously indicated, GA-VSS (GA for variable subset selection).

The goodness of fit of the calculated model was assessed by means of the multiple determination coefficient, R^2 and the $SDEC$ (standard deviation error in calculation) defined as :

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

Cross validation techniques allow the assessment of internal predictivity (Q_{LMO}^2 cross-validation, bootstrap) in addition to the robustness of the model (Q_{LOO}^2 cross-validation, Y-scrambling).

Cross validation by the LOO (leave-one-out) procedure employs n training sets of $n-1$ objects in and predicting each excluded object in the test set. The cross validated explained Q_{LOO}^2 is defined as:

$$Q_{LOO}^2 = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

where y_i and \bar{y} are, respectively, the measured, and averaged (over the entire data set) values of the dependent variable; $\hat{y}_{i/i}$ denotes the response of the i -th object estimated by using a model obtained without using the i -th object; the summations run over all compounds in the training set.

The $PRESS$ (predictive residual sum of squares) measures the dispersion of the predicted values. It is used to define Q^2 , and the $SDEP$ (standard deviation error in prediction).

$$SDEP = \sqrt{\frac{PRESS}{n}} \quad (3)$$

A value $Q^2 > 0.5$ is generally regarded as a good result and $Q^2 > 0.9$ as excellent [18, 19].

However, studies have indicated that while Q^2 is a necessary condition for high predictive power in a model, its alone is not sufficient.

To avoid overestimating the predictive power of the model the leave-more out (LMO up to 50% of perturbation: LMO/50) procedure (repeated 8000 times in this study) was also performed. In a typical LMO validation, n objects of the data set are divided in G cancellation groups of equal size, $m_i (= n/G)$. Based on the value of n , G is generally selected between 2 and 10. A large number of models are developed with each of the $n - m_i$ objects in the training set and m_i objects in the validation set. For each corresponding model m_i objects are predicted and Q_{LMO}^2 computed (as average value of the number of validation runs).

In order to evidence the existence of fortuitous correlations, the randomization test (Y-scrambling) [24] was adopted. This test consists of building a property vector whose components are the components of the actual property vector, but randomly permuted in their positions. This new activity vector is used as if it was really an experimental one, and a QSAR model is computed in the usual way. This process was repeated 300 times, in order to test the capacity factor of the model to extract actual structure/activity relationships.

By bootstrap validation technique, the original size of the data set (n) is preserved for the training set, by the selection of n objects with repetition; in this way the training set usually consists of repeated objects and the evaluation set of the object left out [25]. The model is calculated on the training set and responses are predicted on the evaluation set. All the squared differences between the true response and the predictive response of the objects of evaluation set are collected in $PRESS$. This procedure of building training sets and evaluation sets is repeated 5,000 times in this study, $PRESS$ are summed and the average predictive power is calculated.

By using the selected model the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of Q_{ext}^2 , which is defined as:

$$Q_{ext}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y})^2 / n_{tr}} = 1 - \frac{PRESS/n_{ext}}{TSS/n_{tr}} \quad (4)$$

where n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap), and the number of training set objects, respectively.

Other useful parameters are R^2 , calculated for the validation chemicals by applying the model developed on the training set, and external standard deviation error of prediction ($SDEP_{ext}$), defined as:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2} \quad (5)$$

where the sum runs over the test set objects (n_{ext}).

2.3 QSAR AD (Applicability Domain)

The AD was discussed by the Williams plot [18, 19] of jackknifed residuals versus leverages (hat diagonal) values (h_i). The jackknifed residuals (or Studentized residuals) are the standardized cross-validated residuals. Each residuals is divided by its standard deviation, which is calculated without the i -th observation. The leverage (h_i) value of a chemical in the original variable space is defined as :

$$h_i = \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \quad (i = 1, \dots, n) \quad (6)$$

where \mathbf{x}_i is the descriptor row-vector of the query compound, and \mathbf{X} is the $n(p+1)$ matrix of p model parameter values for n training set compounds. The superscript T refers to the transpose of the matrix/vector.

The warning leverage value (h^*) is defined as $3(p+1)/n$. When h value of a compound is lower than h^* , the probability of accordance between

predicted and actual values is as high as that for the compounds in the training set. A chemical with $h_i > h^*$ will reinforce the model if the chemical is in the training set. But such a chemical in the validation set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residual may be low. Thus the leverage and the jackknifed residual should be combined for the characterization of the AD.

3. Results and Discussion

3.1 Development and Validation of QSAR Models

Application of the GA-VSS led to several good models for the prediction of $pIGC_{50}^{-1}$ based on different sets of molecular descriptors. The best two dimensional model was constructed using the radius of gyration (RG_{yr}) and R maximal autocorrelation of lag 3 weighted by van der Waals atomic volume v ($R3v+$). All data concerning value of RG_{yr} , $R3v+$ and biological activity are summarized in Table 1.

The equation of the optimal model can be written as:

$$pIGC_{50}^{-1} = -3.602 (\pm 0.174) + 1.439 (\pm 0.069) RG_{yr} + 16.342 (\pm 1.416) R3v+ \quad (7)$$

All relevant statistical parameters are reported in Table 2.

Values of R^2 and R_{adj}^2 attest the good fitting performances of the model which, moreover, is very highly significant (great value of the Fisher parameter F).

The model is robust, the difference between R^2 and Q^2 is small (1%). Fig. 1 shows a plot contrasting experimental and cross-validated $pIGC_{50}^{-1}$. The point dispersion is small, although in this case there are two points a little bit far away from the rest.

The model demonstrates a very good stability in internal validation (difference between Q_{LOO}^2 and

Table 1 Values of $RGyr$, $R3v+$ and inhibition of growth concentration $pIGC_{50}^{-1}$ for a set of 48 anilines. The first 17 chemicals are the test set.

Chemical	$pIGC_{50}^{-1}$	$RGyr$	$R3v+$	Chemical	$pIGC_{50}^{-1}$	$RGyr$	$R3v+$
4-hexylaniline	2.04	3.434	0.023	3,4-dimethylaniline	-0.29	2.132	0.029
2,3-chloro	1.02	2.169	0.087	3-ethyl	-0.12	2.204	0.021
4-methyl	-0.02	2.008	0.021	2-chloro	-0.09	1.982	0.041
2,4,6-trichloro	1.01	2.588	0.039	2,4-dimethyl	-0.30	2.133	0.022
2-bromo	0.46	2.013	0.056	2-ethyl	-0.25	2.107	0.023
4-butyl	1.05	2.998	0.022	3-fluoro	0.04	1.932	0.025
2-chloro-6-methyl	0.12	2.156	0.037	2-propyl	0.06	2.414	0.023
2-phenyl	0.86	2.680	0.019	3-chloro	0.09	2.111	0.042
3-iodo	0.61	2.158	0.071	2-isopropyl	0.10	2.190	0.024
3,4,5-trichloro	1.51	2.451	0.088	4-isopropyl	0.21	2.403	0.024
4-ethyl	0.04	2.286	0.021	2-chloro-5-methyl	0.20	2.245	0.037
3-chloro-4-methyl	0.45	2.208	0.049	4-octyl	2.34	3.914	0.022
5-chloro-2-methyl	0.51	2.300	0.03	2-iodo	0.35	1.981	0.070
2,6-dichloro	0.33	2.291	0.04	4-chloro-2-methyl	0.35	2.298	0.033
3-phenyl	0.78	2.815	0.021	3-chloro-2-methyl	0.45	2.153	0.044
2,5-dichloro	0.58	2.448	0.039	2,4-dichloro	0.56	2.390	0.040
3,5-dichloro	0.71	2.423	0.038	4-propyl	0.49	2.611	0.024
2-methyl	-0.55	1.892	0.024	3-bromo	0.52	2.179	0.058
3-methyl	-0.43	1.968	0.026	2,6-dichloro-3-methyl	0.69	2.407	0.041
2,6-dimethyl	-0.43	2.047	0.024	4-phenyl	0.95	2.904	0.022
3,5-dimethyl	-0.37	2.158	0.017	3,4-dichloro	1.14	2.281	0.089
2,5-dimethyl	-0.35	2.134	0.023	2,4,5-trichloro	1.30	2.577	0.086
2-fluoro	-0.31	1.846	0.025	2,3,4-trichloro	1.35	2.405	0.087
2-chloro-4-methyl	0.24	2.211	0.039	4-pentyl	1.67	3.246	0.022

Table 2 Statistical parameters of the developed model.

n_{tr}	n_{ext}	Q_{LOO}^2	R^2	$Q_{LMO/50}^2$	Q_{BOOT}^2	R_{adj}^2	Q_{ext}^2
31	17	93.85	94.99	92.34	92.48	94.64	92.13
$SDEC$		$SDEP$		$SDEP_{ext}$		s	
0.151		0.168		0.184		0.159	
						F	
						265.64	

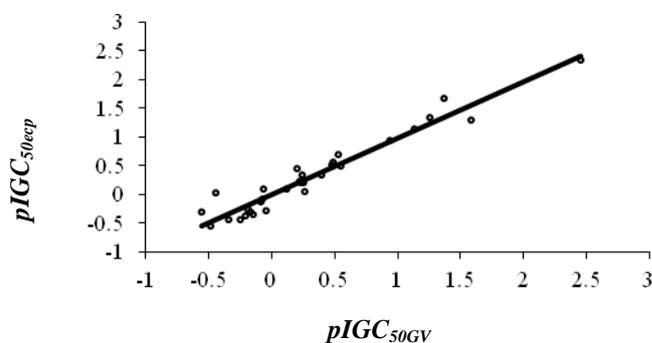


Fig. 1 Experimental ($pIGC_{50exp}$) versus cross-validated ($pIGC_{50cv}$) activity for the training set objects.

$Q_{LMO/50}^2$ is about 1%), while bootstrapping confirms the internal predictivity and stability of the model. $SDEP_{ext}$ is a little bit different from $SDEP$; the model works slightly worse in external prediction than in internal prediction. The model was also verified by Y-scrambling. Fig. 2 clearly ensures the existence of a linear relationship between $pIGC_{50}^{-1}$ and the descriptors $RGyr$ and $R3v+$. As can be observed the permuted responses yield poor predictive models, all having $Q^2 < 0.2$. On the other hand, the correctly ordered $pIGC_{50}^{-1}$ yield good statistical parameters, and therefore it is located isolated in the plot.

Using the same training set as before, a model was calculated by us on the molecular descriptors selected by Schultz et al. [15]. It follows the expression :

$$pIGC_{50}^{-1} = -1.404(\pm 0.113) - 0.4848(\pm 0.123) \sum \delta + 0.727(\pm 0.047) \log K_{ow} \quad (8)$$

The corresponding fitting and prediction parameters reported in Table 3 show that the model presented in this paper is slightly better than the one based on the

Schultz et al. [15] approach.

3.2 Mechanistic Interpretation

By interpreting the descriptors in the proposed model, it is possible to gain some insight into factors that are likely related to inhibition of microbial growth. Of the two descriptors, one is GETAWAY ($R3v+$) and one is Geometrical ($RGyr$).

R-GETAWAY descriptors which are represented by $Rk(w)$ were calculated as follows. The molecular influence matrix was denoted by H and resembled the leverage (or influence) matrix defined in regression diagnostics [26].

The value of H was calculated from the molecular matrix M (M has A rows corresponding to the Cartesian coordinates x, y, z of each atom in optimized molecular structure) as follows:

$$H = M (M^T M)^{-1} M^T \quad (9)$$

where the superscript T refers to the transposed matrix. The maximal contributed to the autocorrelation at each lag represented by $Rk(w)+$ can be defined as:

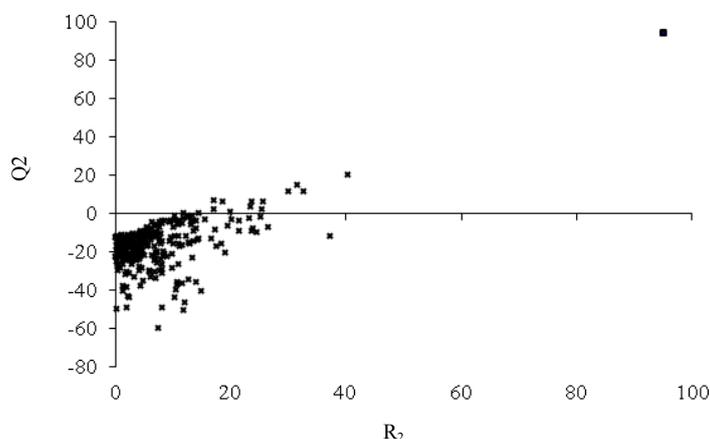


Fig. 2 Randomization test associated to the previous QSAR model. Crosses represent the randomly ordered activities, and the square corresponds to the real activities.

Table 3 Statistical parameters of the Schultz et al. [15] approach model.

n_{tr}	n_{ext}	Q_{LOO}^2	R^2	$Q_{LMO/50}^2$	Q_{BOOT}^2	R_{adj}^2	Q_{ext}^2
31	17	89.24	91.62	85.38	86.06	91.03	81.28
$SDEC$		$SDEP$		$SDEP_{ext}$		s	F
0.196		0.222		0.293		0.206	153.1523

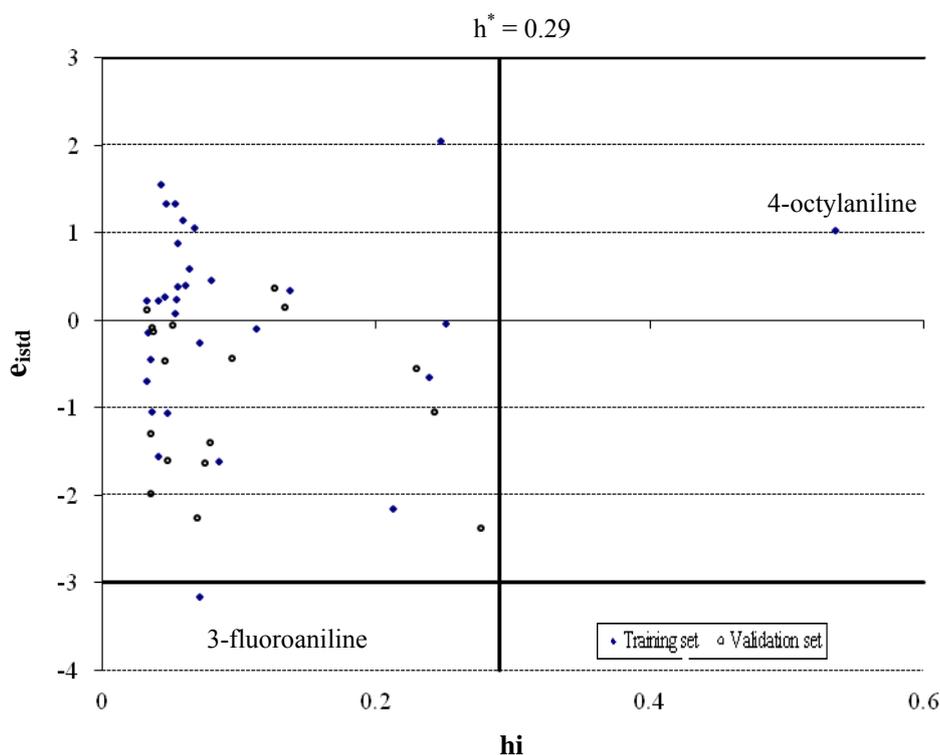


Fig. 3 Williams plot of the current QSAR model.

$$Rk_{w+} = \max_{ij} \left[\frac{\sqrt{h_{ii}h_{jj}}}{r_{ij}} w_i w_j \delta(k, d_{ij}) \right] \quad i \neq j \text{ and} \\ k = 1, 2, \dots, 8 \quad (10)$$

where $Rk(w)_+$ is the w -weighted k th order maximal R index, r_{ij} is the 3D geometric distances between each pair of atoms i and j , d_{ij} is the topological diameter, h_{ii} and h_{jj} are diagonal terms of the H matrix and δ is a Dirac delta function defined as:

$$\delta(k, d_{ij}) = \begin{cases} 1 & \text{if } d_{ij} = k \\ 0 & \text{if } d_{ij} \neq k \end{cases} \quad (11)$$

$R3v_+$ describes the size and the shape of the molecules. It is known that size, shape and symmetry of molecules play a key role in the process of distribution of molecule between two immiscible liquid phases. At the same time this descriptor indicates the role of the volume (v) in deciding the activity.

Mean size is a simple and very significant property of a molecule [27]. Easily obtained from light scattering experiments, a common measure of mean size such as the radius of gyration (RG_{yr}) provides valuable information on interaction of molecule with its surrounding medium or its target.

3.3 Applicability Domain

As shown in the Williams plot (Fig. 3), the only high leverage chemical ($h_i > h^* = 0.29$) of the training set (4-octylaniline) is perfectly predicted, as normally happens for chemicals influential in training sets. Only one outlier is observed (3-fluoroaniline) which can be judged by its standardized residual greater than three standard deviation units (3σ).

4. Conclusion

A QSAR model on inhibition of microbial growth by anilines was developed using the OECD guidelines.

The available data set was randomly split into training and validation sets.

The QSAR model proposed in this paper is stable, robust, with good fitting and predictive performance. It is predictive for the chemicals used in the model development (internal validation on training chemicals) and also for chemicals not used in the model development (statistical external validation on validation set chemicals). The AD of the QSAR model was also described. The factors governing biological activities are the molecular size and shape, and interactions of molecule with its surrounding medium or its target.

References

- [1] A.D. Deweese, T.W. Schultz, Structure-activity relationships for aquatic toxicity to *Tetrahymena*: Halogensubstituted aliphatic esters, *Environmental Toxicology* 16 (2001) 54-60.
- [2] J.D. Leblond, B.M. Applegate, F.M. Menn, T.W. Schultz, G.S. Sayler, Structure-toxicity assessment of metabolites of the aerobic bacterial transformation of substituted naphthalenes, *Environmental Toxicology and Chemistry* 19 (2000) 1235-1246.
- [3] A. Cotescu, M.V. Diudea, QSTR study on aquatic toxicity against *poecilia reticulata* and *tetrahymena pyriformis* using topological indices, *Internet Electronic Journal of Molecular Design* 5 (2006) 116-134.
- [4] F. Li, J. Chen, Z. Wang, J. Li, X. Qia, Determination and prediction of xenoestrogens by recombinant yeast-based assay and QSAR, *Chemosphere* 74 (2009) 1152-1157.
- [5] G.H. Lu, C. Wang, X.L. Guo, Prediction of toxicity of phenols and anilines to algae by quantitative structure-activity relationship, *Biomedical and Environmental Sciences* 21 (2008) 193-196.
- [6] C. Hansch, T. Fujita, ρ - σ - π analysis: A method for the correlation of biological activity and chemical structure, *Journal of the American Chemical Society* 86 (1964) 1616-1626.
- [7] M. Nendza, A. Wenzel, Discriminating toxicant classes by mode of action-1.(Eco) toxicity profiles, *Environmental Science and Pollution Research* 13 (2006) 192-203.
- [8] L.P. Hammett, The effect of structures upon the reactions of organic compounds, Benzene derivatives, *Journal of the American Chemical Society* 59 (1937) 96-103.
- [9] L.P. Hammett, *Physical Organic Chemistry*, Mc Graw Hill, New York, 1940.
- [10] J.S. Roth, Certain effects of 2-aminofluorene and α - and β -naphthylamines on *Tetrahymena pyriformis*, *Cancer Research* 2 (1954) 346-350.
- [11] N.R. Cooley, J.M. Keltner Jr, J. Forester, Polychlorinated biphenyls, aroclors 1248 and 1260: Effect on and accumulation by *tetrahymena pyriformis*, *The Journal of Protozoology* 20 (1975) 443-445.
- [12] S. Apostol, A bioassay of toxicity using protozoa in the study of aquatic environment pollution and its prevention, *Environmental Research* 6 (1973) 365-372.
- [13] D. Dive, H. Leclerc, Standardized test method using protozoa for measuring water pollutant toxicity, *Progress in Water Technology* 7 (1975) 67-72.
- [14] D.L. Hill, *The Biochemistry and Physiology of Tetrahymena*, Academic Press, New York, 1972.
- [15] T.W. Schultz, D.T. Lin, T.S. Wicke, L.M. Arnold, Quantitative structure-activity relationships for the *Tetrahymena pyriformis* population growth endpoint: A mechanism of action approach, in: W. Karcher, J. Devillers (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Kluwer Academic Publishers, Dordrecht, 1990, pp. 241-262
- [16] J.S. Jaworska, M. Comber, C. Auer, C.J. Van Leeuwen, Summary of a workshop on regulatory acceptance of (Q)SAR for human health and environmental endpoints, *Environmental Health Perspectives* 111 (2003) 1358-1360.
- [17] Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models [Online], Series on Testing and Assessment 69, OCDE's Environment Directorate, OECD Environment Health and Safety Publications, Mar. 30, 2007, <http://www.oecd.org/data/33/37/37849783.pdf>.
- [18] L. Eriksson, J. Jaworska, A. Worth, M. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs, *Environmental Health Perspectives* 111 (2003) 1361-1375.
- [19] A. Tropsha, P. Gramatica, V.K. Grombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR and Combinatorial Science* 22 (2003) 69-76.
- [20] Hyperchem™ Release 7, Hypercube for Windows, Molecular Modeling System, 2000.
- [21] R. Todeschini, V. Consonni, M. Pavan, DRAGON Software for the Calculation of Molecular Descriptors, Release 5.4 for Windows, Milano, 2006.
- [22] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, *Journal of Chemometrics*

- 6 (1992) 267-281.
- [23] R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan, MOBYDIGS, version 1.1, Copyright TALETE srl, 2009.
- [24] S. Wold, L. Eriksson, Statistical Validation of QSAR Results, Validation Tools in Chemometrics Methods in Molecular Design, VCH Publishers, New York, 1995, pp. 309-318.
- [25] B. Efron, The Jackknife, the Bootstrap and Other Resampling Planes, Society for Industrial and Applied Mathematics, Philadelphia, 1994.
- [26] V. Consonni, R. Todeschini, M. Pavan, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors, 1—Theory of the novel 3D molecular descriptors, Journal of Chemical Information and Modeling 42 (2002) 682-692.
- [27] G.A. Arteca, Analysis of shape transitions using molecular size descriptors associated with inner and outer regions of a polymer chain, Journal of Molecular Structure: THEOCHEM 630 (2003) 113-123.