

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR ANNABA UNIVERSITY

UNIVERSITE BADJI MOKHTAR ANNABA

جامعة باجي مختار – عنابة

Année universitaire: 2015-2016



FACULTE DES SCIENCES

DEPARTEMENT DE CHIMIE

THÈSE

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

Modèles QSAR pour la prédiction de la toxicité aquatique :

**-De dérivés benzéniques substitués vis-à-vis de Pemiphales promelas
-D'alcools et d'amines vis-a-vis de Tetrahymena pyriformis**

Domaine : Chimie

Spécialité : Chimie Analytique et Environnement

Présentée par : Ziani Nadia

Directeur de thèse : Mr. D. MESSADI

Pr.

Université de Annaba

Devant le jury

Présidente : M^{me} S. ALI MOKHNACHE

Pr.

Université de Annaba

Examinatrice: M^{me} L. DIB

Pr.

Université de Annaba

Examineur: Mr. A.H. GHEID

Pr.

Université DE Souk-Ahras

Examineur: Mr. R. MERDES

Pr.

Université de Guelma

Examineur: Mr. M. KADRI

Pr.

Université de Guelma

SOMMAIRE

REMERCIEMENTS	
DEDICACE	
RESUMES	
LISTE DES TABLEAUX	
LISTE DES FIGURES	
SYMBOLES ET ABREVIATIONS	
INTRODUCTION GENERALE	1
PARTIE I : Généralités	
I-Le benzène et ses dérivés	4
I-1- Introduction et aperçu historique	4
I-2 Utilisation du benzène et de ses dérivés	4
I-3-Influence du benzène et ses dérivés sur l'environnement	4
I-4 Dose létale	7
I-4-1-Définition	7
I-4-2- Forme de toxicité	8
I-4-2-a Toxicité aiguë	8
I-4-2-b Toxicité subaiguë	8
I-4-2-c Toxicité à long terme	9
I-4-3- Utilisation de la dose létale 50	9
I-4-4-Identification de la toxicité	9
I-4-5- Identification du pouvoir pathogène	10
II- Les alcools et les amines	10
II-1 Introduction	10
II-2 Les alcools	11
II-2-1 Utilisation	11
II-2-2 Propriétés	12
II-2-3 Toxicité	12
II-3 Les amines	12
II-3-1 Utilisation	13
II-3-2 Propriétés	13
II-3-3 Toxicité	13
II-4 Concentration d'inhibition de la croissance	14
II-4-1 Définition	14
II-4-2 Pourquoi 50%	14
Référencés bibliographiques	15
PARTIE II : Aspects théoriques de la modélisation moléculaire	
I-Optimisation de la géométrie de la molécule	17
I-1-Généralités	17
I-2-Méthodes semi- empiriques utilisées	19.
I-2-1 Le cadre Hartree Fock -Roothaan	19
I-2-2 Analyse de population de Mulliken	21
I-2-3 Les méthodes semi-empiriques.	23

I-3- Champ de force	28
I-3-1-Définition	28
I-3-2- Quelques exemples	29
I-3-3-Représentation simple d'un champ de force	30
I-3-4-Exemple de calcul	32
I-3-5 Champs de force MM2 et MM+	33
I-3-5-1 Champ de force MM2	33
I-3-5-2 Champ de force MM+	38
I-Modélisation et évaluation des modèles	40
II-1 Modélisation	40
II-1-1 La régression linéaire multiple	40
II-1-1 Modèle QSXR	41
II-1-3 Les réseaux de neurones artificiels	41
II-1-4 Régression machines à vecteur de support	43
II-2 Développement et évaluation de modèle	44
II-2-1 Sélection d'un sous- ensemble de descripteurs	44
II-2-2 Principe	44
II-2-3 Initiation aléatoire du modèle	45
II-2-4 Etape de croisement	45
II-2-5 Etape de mutation	45
II-2-6 Condition d'arrêt	46
II-3 Développement des modèles	46
II-3-1 Paramètre d'évaluation de la qualité de l'ajustement	46
II-3-2 Robustesse du modèle	47
II-3-3 Domaine d'application	47
II-3-4 Test de randomisation	48
II-3-5 Validation externe	48
Référencés bibliographiques	50
CHAPITRE III : Résultats et discussions	
I- Modélisation de la concentration d'inhibition 50 de la croissance	
I-1 INTRODUCTION	53
I-2 Coefficient de partage	54
I-2-1 Propriété de partage	54
I-2-2 Détermination expérimentale de logP	55
I-2-3 Méthode d'estimation de logP	56
I-2-3-1 ClogP	57
I-2-3-2 AlogP	58
I-2-3-3 MlogP	59
I-3 Collecte des données expérimentales	59
I-3-1 Le protozoaire (Tetrahymena pyriformis)	59
I-3-2 Test de toxicité (méthode et matériels)	60
I-3-3 Mécanisme de l'action toxique	61
I-3-4 Narcose apolaire	61
I-4 Résultats et discussion	62

I-5 Conclusion	67
II-Modélisation de la concentration létale 50	
II-1 Introduction	86
II-2 Matériels et méthode	70
II-2-1 Ensemble des données	70
II-2-2 Calcul des descripteurs	73
II-2-3 L'algorithme de Kennard et Stone	73
II-2-4 Sélection des descripteurs	74
II-3- Résultats et discussion	76
II-3-1 Résultats de la régression linéaire multiple	76
II-3-1-1 Analyse de contribution des descripteurs	78
II-3-1-3 Signification des descripteurs sélectionnés	79
1 - Polarisabilité	79
2 - Indice d'autocorrélation de Moran de distance topologique 1 pondéré par les masses atomiques m :MATS 1m	80
3- RDF020v	80
4- Descripteurs de charge	81
5- E1s	83
6- Clogp	84
II-3-1-4 Droite d'ajustement	84
II-3-1-5 Domaine d'application	84
II-3-1-6 Test de randomisation	85
II-3-2 Machine à support vecteur	86
II-3-3 Réseaux de neurones artificiels	86
II-4 Conclusion	88
Références bibliographiques	84
CONCLUSION GENERALE	97
ANNEXE I Présentation des données	99
ANNEXE II Publications	85

Remerciements

Ce travail de thèse a été réalisé au sein du Laboratoire de Sécurité Environnementale et Alimentaire de l'Université Badji Mokhtar Annaba sous la direction de M^r le Professeur MESSADI DJELLOUL.

Avant tout, je dois remercier Dieu le tout puissant qui m'a donné l'envie et la force

pour mener à terme ce travail;

Je tiens à remercier très sincèrement le Professeur MESSADI DJELLOUL de l'université BADJI MOKHTAR ANNABA, mon directeur de thèse. Ce fut un grand plaisir de travailler avec lui, durant la préparation du Magister puis du Doctorat. Professionnellement, j'ai beaucoup appris avec lui tout au long de ces années d'études, où à maintes reprises son expérience et ses conseils m'ont été d'une grande utilité et d'un apport inestimable pour ma formation post-graduée. Je lui suis reconnaissante de la confiance qu'il m'a témoignée.

Je tiens également à remercier les membres du jury de thèse, pour avoir pris le temps de lire ce manuscrit et de juger mon travail :

- 1- *M^{ME} ALI MOKHNACHE SALIMA*
- 2- *M^{ME} DIB LYNDA*
- 3- *M^R GHEID ABDELHAK*
- 4- *M^R MERDES RACHID*
- 5- *M^R KADRI MEKKI*

Je remercie toute l'équipe du laboratoire LASEA et en particulierement , M^{lle} AMIRAT Khadidja, M^r HADDAG Hamza, M^r KERTIOU Noureddine, et M^r DRIOUCH Youssef

Et enfin J'adresse tous mes chaleureux remerciements à tous ceux qui contribué de près ou de loin à la réalisation de ce travail.

ZIANI NADIA.....

Je dédie ce travail

À mes parents, mes frères et mes sœurs,

À mes nièces (Roïya, Nayrouz, et Rahaf).

*À ma grande et chère famille,
source intarissable de réconfort, de joie et d'émulation.*

À mon fiancé : Ziani Hacem.

À ma chère tante : Ziani Naffissa

À mes amis

Vous êtes toujours présents dans mon cœur et mon esprit.

Ziani Nadia.....

ملخص

عدة تقريبات تم استعمالها لربط عدة خصائص (التركيز المعيق للنمو - التركيز القاتل) لمجموعة مختلفة الأهمية لمركبات كيميائية (كحولات - أمينات - مشتقات البنزين).

أنجزت دراسة العلاقة بين الكمية البنوية و النشاط (QSAR) لتحقيق التسمم النسبي لمزيج مكون من 21 كحول (ذات سلاسل خطية و متفرعة) و 9 أمينات اليفاتية عادية للتعبير على تركيز المعيقات بنسبة لنمو 50% (IGC₅₀) (*Tetrahymena pyriformis*).

طريقة الانحدار الخطي البسيط اعتمدت على الموصفات الجزئية النظرية (هندسية), ثلاثية الأبعاد 3D, المتحصل عليها اعتمادا على برنامج DRAGON, والواصفات المختلفة log P المحسوبة. الصلابة و القدرة التنبؤية للنموذج تم التحقق منها اعتمادا على التصديق الداخلي (تصديق متقاطع) (Loo, LMO, bootstrap) كذلك التصديق الخارجي و Clog P ظهر بأنه أحسن واصف لنمذجة التسمية, و الذي يمكن تعويضه بالواصف الهندسي ADDD, بدون تغيير ملحوظ للمعايير الإحصائية.

كما انجزت دراسة اخرى علاقة كمية بنية - نشاط على مجموعة من مشتقات البنزين للتنبأ بالتسمم القاتل بنسبة 50 % ل pemiphales promelas.

التقنيات المستعملة في هذه الدراسة هي على التوالي طريقة الانحدار الخطي المتعدد - الشبكة العصبونية الاصطناعية - Support vector machine.

قسمت مجموعة المركبات الكيميائية الى مجموعتين بواسطة خوارزمية Kennard and Stone الاولى متكونة من 74 مركب من اجل تكوين النموذج و الثانية تتكون من 18 مركب من اجل اثبات التصديق الخارجي. تم اختيار النموذج عن طريق نظرية FIT بحيث تمت المصادقة عليه عن طريق التقنيات المذكورة سابقا اثبتت النتائج المتحصل عليها صلابة و استقرار النموذج المختار.

الكلمات الدالة

الكحولات و الامينات - مشتقات البنزين - التسمم المائي - الواصفات الجزئية - المركبات الكيميائية - العلاقة كمية بنية / نشاط - الانحدار الخطي المتعدد - الشبكة العصبونية الاصطناعية - Support vector machine.

Abstract

Different approaches were used to relate some compound properties to their chemical structure (Lethal concentration 50 (LC50), inhibitory growth concentration 50 (IGC50)) for more or less significative series of compound (alcohols, amines, benzenes derivatives).

A Quantitative Structure- Activity Relationship (QSAR) study was undertaken to evaluate the relative toxicity of a mixed series of 21 (linear and branched-chain) alcohols and 9 normal aliphatic amines in term of the 50% inhibitory growth concentration (IGC₅₀) of *Tetrahymena pyriformis*. The applied simple linear regression approach is based on theoretical 3D (geometrical) molecular descriptors from DRAGON package, and some calculated logP descriptors. The robustness and the predictive performance of the models were verified using both internal (cross-validation by LOO and LMO; bootstrap) and external statistical validations. ClogP turned out to be the best descriptor to model the considered endpoint. It may be interchanged with geometrical descriptor ADDD without relevant variations in the statistical parameters.

Another QSAR study was developed on a data set of benzene derivatives for the prediction of the lethal concentration 50 of *Pemiphales promelas*. The techniques used are: multiple linear regression, regression neural network and support vectors regression. Relations between the structure and the activity were examined quantitatively by using theoretical descriptors. The data set were devised in two disjoined sets of 74 observation and 18 of prediction using the algorithm of Kennard and Stone, the size of the model (six descriptors) was conditioned by the optimal value of the function of the FIT of Kubinyi. The model obtained was examined by the three approaches (MLR, ANN, SVM), the results obtained are very similar which confirms the stability and the robustness of our model.

Key words: *Study QSAR- Alcohols and Amines – Benzene derivatives - Aquatic toxicity – Molecular descriptors – chemical compounds - QSAR – Simple linear regression - Artificial neural networks –Support vector machine.*

Résumé

Différentes approches ont été utilisées pour relier certaines propriétés (CL50, CIC50) de séries de composés chimiques (alcools, amines, dérivés benzéniques) à leurs structures.

Une étude Relation Quantitative Structure- Activité (QSAR) a été réalisée pour évaluer la toxicité relative d'un mélange composé de 21 alcools (à chaînes linéaires et ramifiées) et 9 amines aliphatiques normales, en terme de concentration d'inhibition 50% de la croissance (CIC50) de *Tetrahymena pyriformis*. L'approche par régression linéaire simple est basée sur des descripteurs moléculaires théoriques (géométriques) 3D obtenus à l'aide du logiciel DRAGON et différents descripteurs logP calculés. La robustesse et la capacité prédictive des modèles ont été vérifiées à l'aide de statistiques de validations internes (validations croisées LOO et LMO ; bootstrap) et externe. Clogp s'est avéré le meilleur descripteur pour la modélisation de la grandeur d'intérêt considérée. Il peut être remplacé par le descripteur géométrique ADDD sans variations appréciables des paramètres statistiques.

Une autre étude QSAR a été développée sur un ensemble de dérivés benzéniques pour la prédiction de la concentration létale 50 de *Pemiphales promelas*. Les techniques utilisées sont : la régression linéaire multiple, la régression par réseaux de neurones et la regression par supports vecteurs. Les relations entre la structure et l'activité ont été examinées quantitativement en utilisant des descripteurs théoriques. Les données ont été éclatées en deux sous-ensembles disjoints de calibrage (74 observations) et de prédiction (18 observations) en utilisant l'algorithme de Kennard et Stone, la taille du modèle (six régresseurs) à été conditionnée par la valeur optimale de la fonction de FIT de Kubinyi. Les approches linéaires (RNA, SVM) testées conduisent à des résultats très proches ce qui confirme la stabilité et la robustesse du modèle basé sur les six descripteurs sélectionnés.

Mots clés : Etude QSAR - Composés chimiques- alcools et amines- dérivés benzéniques - Toxicité aquatique- Descripteurs moléculaires - Régression linéaire simple – Régression linéaire multiple- Réseau de neurones artificiel – Machine à support vecteur.



LISTE DES TABLEAUX

	Titre	Page
	Partie II	
Tableau I	Etude comparative des techniques ab initio, semi-empirique et mécanique moléculaire	39
	Partie III	
Tableau I	valeurs des toxicités relatives et des descripteurs moléculaires des alcools et amines aliphatiques sélectionnés	63
Tableau II	Coefficients des modèles calculés par les moindres carrés ordinaires.	65
Tableau III	Résumé des statistiques obtenues pour les modèles unidimensionnels calculés.	65
Tableau IV	Les composés et les résultats prédictif de l'activité biologique pCL50	70
Tableau V	Comparaison de la performance des modèles de différentes tailles	77
Tableau VI	Caractéristiques des descripteurs sélectionnés par MLR	79
Tableau VII	Matrices de corrélation	80
Tableau VIII	Structure optimale du réseau de neurones	87
Tableau IX	Les paramètres statistiques	87



LISTE DES FIGURES

	Titre	Page
	Partie II	
Figure 1	Représentation schématique des quatre contributions à un champ de force de MM : élongation de liaison, flexion angulaire	31
Figure 2	Un modèle de champ de force typique pour le propane contient 10 termes d'élongation de liaison, 18 termes de flexion angulaire, 18 termes de torsion et 27 interactions de non – liaison.	32
Figure 3	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.	34
Figure 4	Deux façons pour modéliser les contributions de la variation d'angle extraplanaire	35
Figure 5	Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces.	36
	Partie III	
Figure 1	Activités observées en fonction de celles obtenues par validation croisée pour l'ensemble de calibrage.	66
Figure 2	variation de R^2 et Q^2 en fonction de la taille du modèle	76
Figure 3	La contribution relative en fonction du nombre des descripteurs	78
Figure 4	Droite d'ajustement des pCL50 observés en fonction des pCL50 calculés	85
Figure 5	Diagramme de Williams pour tous les composés.	86
Figure 6	Test de randomisation	86



SYMBOLES ET ABREVIATIONS

ALOGP	Coefficient de partage de Ghose-Crippen.
ADDD	Indice du rapport moyen des distances.
AG	Algorithme génétique
AM1	Austin Model 1.
CAS	Chemical abstract service
CE	Conseil européen.
CIC50	Concentration d'inhibition 50% de la croissance.
Clogp	Coefficient de partage calculé par le logiciel Chemdraw
DMC	Droite des moindres carrés.
EPA	Agence des Etats-Unis pour la protection de l'environnement.
EQMC	Erreur quadratique moyenne sur l'ensemble de calibrage.
EQMP	Erreur quadratique moyenne sur l'ensemble de prédiction.
EQMP_{ext.}	Erreur quadratique moyenne sur l'ensemble de prédiction externe.
E_s	Energie stérique.
e_i	Résidu : différence entre les valeurs observée et estimée du composé i.
e_{i std}	Résidu de prédiction standardisé.
F	Statistique de Fisher.
FIT	Fonction de Kubinyi.
H	Matrice de projection, ou matrice chapeau.
h_{ii}	Eléments diagonaux de la matrice chapeau.
h*	Valeur critique des leviers.
K	Indice de corrélation multi variable.
LC50	Concentration létale 50%
LogP	Coefficient de partage octanol /eau.
LOO	Cross-validation by leave-one-out: Validation croisée par omission d'une observation.
Mlogp	Coefficient de partage de Moriguchi calculé par le logiciel DRAGON
N	Dimension de la population.
OCDE	Organisation de Coopération et de développement Economique.
OMS	Organisation mondiale de la santé.
n-p	Nombre de degrés de liberté de la somme des carrés des résidus.
p	Nombre de descripteurs en comptant la constante (nombre de paramètres).
Pc	Probalité de croisement
pCIC50	log 1/(CIC50).
pCL50	log 1/(CL50).
PM3	Parameterized model number 3.
pm	Pico mètre
P_M	Probalité de mutation
ppm	Partie par million.
PRESS	Somme des carrés des erreurs de prédiction.
QSAR	Quantitative Structure/ Activity Relationships.
QSPR	Quantitative Structure/ Property Relationships.
Q²_{loo}	Coefficient de prédiction.
Q²_{boot}	Coefficient de prédiction par la technique du bootstrap.

R^2	Coefficient de détermination.
R^2_{adj}	Coefficient de détermination ajusté.
RND	Randomisation.
rw	Rayon de Van der Waals.
S	Erreur standard.
SCE	Somme des carrés des écarts.
SCT	Somme des carrés totale.
Tr	Ensemble de calibrage (training).
VIF	Facteur d'inflation de la variance.
\bar{y}	Valeur moyenne des valeurs observées
Y	Variable expliquée.
y_i	Valeur observée.
\hat{y}_i	Valeur estimée.
X	Matrice des valeurs observées des variables explicatives.
X'	Matrice transposée de X.



INTRODUCTION GENERALE

A priori, l'intérêt pour les Relations Quantitatives Structure / Activité (QSAR : Quantitative Structure / Activity Relationships) semble d'origine récente. Cet intérêt est stimulé par la nécessaire disponibilité des données caractéristiques de la grande masse des composés organiques commercialisés, qui ne sont pas toujours sans risques pour la santé publique et l'environnement, et dopé par la disponibilité générale et le développement rapide des moyens de calculs. En fait, la recherche de corrélations systématiques entre la structure de composés organiques et certaines de leurs propriétés remonte au tout début de la chimie organique. L'augmentation régulière des points de fusion et d'ébullitions avec celle de la longueur des chaînes, ou des masses, au sein de séries homologues d'hydrocarbures en est un exemple pertinent. Pareillement, A.F.A. Cros a noté, en 1863, une corrélation inverse entre la solubilité dans l'eau d'alcools aliphatiques primaires et le nombre de groupes méthylènes présents dans leurs molécules. Même la notion de coefficient de partage entre des phases aqueuse et lipophile, qui joue un rôle important dans les études QSAR concernant l'environnement et la biologie, remonte à une centaine d'années quand, respectivement, Overton et Meyer mettent en évidence et décrivent, en 1899, une corrélation entre la puissance narcotique des composés organiques non électrolytes et leurs solubilité dans l'eau et les lipides. Plus important encore, la méthodologie QSAR a été souvent utilisée, ces dernières années, en agriculture et en pharmacie pour tester et prédire les activités pharmacologiques et biologiques ainsi que leurs cinétiques et leurs effets en vue de réduire le coût et les efforts nécessaires pour la conception et la synthèse de nombreux produits, dont les médicaments. En dépit de ces premiers développements, les données déduites de la méthodologie QSAR sont considérées avec suspicion, la préférence étant accordée aux données expérimentales. Cependant, de plus en plus semble s'enraciner la conviction que la gestion systématique et globale du risque encouru, à cause de la multitude des substances chimiques présentes sur le marché et dans l'environnement, ne peut reposer uniquement sur la seule disponibilité des données expérimentales. Les fichiers de données expérimentales, complets, homogènes et précis s'ils sont parfois disponibles, peuvent faire défaut même pour les composés du commerce les plus courants et les plus importants.

Par ailleurs, la détermination expérimentale systématique au laboratoire de toutes les données manquantes se traduirait 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. De plus, si la détermination expérimentale des données coûte chère elle n'en reste pas moins entachée de larges marges d'erreurs. Par exemple, on peut imaginer que l'écart supérieur à 7 observé dans l'estimation par QSAR des valeurs de logP puisse être dû davantage à une surestimation expérimentale de la solubilité dans l'eau, plutôt qu'à un biais dans la méthodologie QSAR. En fait on a pu affirmer que le logP pour certaines classes de composés (HAP, PCB) peut être calculé à l'avance avec une précision meilleure que la répétabilité des mesures inter-laboratoires. Dans toutes les techniques QSAR, les applications pratiques nécessitent au départ un ensemble de composés d'activités connues : les données d'estimation. Cet ensemble, dit de calibrage, est utilisé dans une première étape pour le calcul d'un modèle prédictif. Ce qui conduit à des équations mathématiques simples associant, aussi bien que possible, les descripteurs (régresseurs) et les propriétés mesurées (observations ou encore variables dépendantes ou à expliquer). Si l'ensemble de calibrage constitue un échantillon représentatif de la population, on admet alors que l'introduction de nouveaux éléments dont la valeur de la variable dépendante est inconnue (qu'on désigne encore par données de prédiction), n'affectera pas la stabilité du modèle, et que l'on peut ainsi estimer ces valeurs manquantes avec suffisamment de confiance.

Pour l'estimation du risque de toxicité que peuvent présenter les polluants organiques, particulièrement vis-à-vis des organismes aquatiques et des mammifères, les données de toxicité aiguë sont nécessaires.

En dépit des banques de données qui portent sur, ou qui incluent, des fichiers de toxicité comme dans le "Registry of Toxic Effects of Chemical Substances" (RTECS), l'"Aquatic Toxicity Information Retrieval" (AQUIRE), l'"Environmental Chemicals Data and Information Network" (ECDIN), il n'en reste pas moins vrai qu'une petite fraction, seulement, du grand nombre de composés polluants et/ou toxiques est couverte par des données expérimentales de toxicité fiables et confirmées. Ainsi, la dérivation et l'estimation des données de toxicité « aiguë » à l'aide de modèles QSAR validés, constituent de plus en plus, un outil complémentaire important.

L'Agence Américaine de Protection de L'Environnement (Duluth, Minnesota) a établi un programme pour la génération des données de haute qualité sur la toxicité des poissons,

qui ont été publiées dans une série de volumes. Une partie de ces données servira de base à l'étude QSAR présentée dans ce travail.

Dans ce travail de thèse, nous nous sommes intéressés à la toxicité des alcools aliphatiques et des amines caractérisée par la concentration d'inhibition 50 % de la croissance (**CIC50**) et la concentration létale 50 % (**CL50**) des dérivés benzéniques vis-à-vis de *Pemiphales promelas*, qui seront reliées à des descripteurs, calculés en utilisant uniquement la structure des composés chimiques après optimisation de leur géométrie.

Cette thèse comporte trois parties,

- La première présente des généralités sur les composés étudiés, ainsi que leurs toxicités et leurs influences sur l'environnement.
- Dans la deuxième partie nous avons présenté les différentes méthodes de la modélisation moléculaire utilisées lors de l'étude théorique.
- Dans la troisième partie nous avons présenté et discuté les principaux résultats fournis par la modélisation moléculaire.

PARTIE I

GENERALITES

- *Le benzène et ses dérivés*
- *Les alcools et les amines*

I- LE BENZENE ET SES DERIVES

I-1-Introduction et aperçu historique.

Le benzène, hydrocarbure aromatique monocyclique, est un liquide cancérigène. Avant les années 1920, le benzène était fréquemment utilisé comme solvant industriel, particulièrement pour dégraisser les métaux. Lorsque sa toxicité devint évidente, il est remplacé par d'autres solvants pour les applications nécessitant une exposition directe de l'utilisateur. Le benzène est utilisé en majeure partie comme intermédiaire dans la synthèse d'autres composés chimiques.

Un grand nombre de composés chimiques très importants dans l'industrie sont obtenus en remplaçant un ou plusieurs atomes d'hydrogène du benzène par d'autres groupements fonctionnels.

I-2 Utilisation du benzène et de ses dérivés.

Les dérivés du benzène, sont utilisés dans la fabrication des polymères et des plastiques ; le phénol, intervient dans celle des résines et des adhésifs. Des quantités moindres de benzène sont signalées dans la fabrication de pneus, de lubrifiants, de colorants, de détergents, de médicaments, d'explosifs ou de pesticides.

Le toluène sert à élever l'indice d'octane dans les carburants. Il sert également de solvant pour les peintures. On s'en sert comme produit de départ pour divers procédés industriels: la synthèse du caoutchouc, du phénol, du TNT, du diisocyanate de toluène nécessaire pour obtenir la mousse de polyuréthane. On s'en sert également dans l'imprimerie, les adhésifs, les laques, et le tannage du cuir. Dans les années 1980, l'éthylbenzène, intermédiaire dans la préparation du styrène, représentait le principal dérivé du benzène.

I-3 Influence du benzène et ses dérivés de substitution sur l'environnement.

Les propriétés cancérigènes du benzène proviennent de ce qu'il se comporte comme un agent intercalant (c'est-à-dire qu'il se glisse entre les bases nucléotidiques des acides nucléiques, dont l'ADN, provoquant des erreurs de lecture et/ou de réplication). Il existe d'autres

agents intercalant (comme le bromure d'éthidium, ou BET, utilisé en biologie expérimentale pour marquer l'ADN, notamment au cours des électrophorèses). Tous les composés plans ne sont toutefois pas cancérigènes. L'acide benzoïque, par exemple, très proche du benzène, et dont la base conjuguée est absolument plane, n'est pas cancérigène (il est utilisé comme conservateur dans divers types de sodas). De même, la phénylalanine, un acide aminé qui comporte un groupement phényle (un cycle benzénique), n'est pas cancérigène.

L'intoxication par le benzène seul porte le nom de benzénisme ; celle par le benzène ou ses dérivés (toluène, xylène...) porte le nom de benzolisme. L'inhalation d'un taux très élevé de benzène peut causer la mort, tandis que des taux élevés peuvent occasionner des somnolences, des vertiges, une accélération du rythme cardiaque, des maux de tête, des tremblements, la confusion ou la perte de connaissance. Une exposition de cinq à dix minutes à un taux de benzène dans l'air de 2 % environ suffit pour entraîner la mort. *La dose létale par ingestion est de 50 mg/kg.* L'ingestion de nourritures ou de boissons contenant des taux élevés de benzène peut occasionner des vomissements, une irritation de l'estomac, des vertiges, des somnolences, des convulsions, une accélération du rythme cardiaque, voire la mort. L'effet principal d'une exposition chronique au benzène est un endommagement de la moelle osseuse, qui peut occasionner une décroissance du taux de globules rouges dans le sang et une anémie. Il peut également occasionner des saignements et un affaiblissement du système immunitaire. L'effet du benzène sur la fertilité de l'homme ou le bon développement du fœtus n'est pas connu. Enfin, le benzène est reconnu comme étant une substance cancérigène.

Le xylène a un effet nocif sur le cerveau. Des niveaux d'expositions élevés pour des périodes même courtes peuvent entraîner des maux de tête, un défaut de coordination des muscles, des vertiges, la confusion et des pertes du sens de l'équilibre. Des expositions à des taux élevés pendant de courtes périodes de temps peuvent également occasionner une irritation de la peau, des yeux, du nez et de la gorge, des difficultés de respiration, des problèmes pulmonaires, une augmentation des temps de réaction, des pertes de mémoire, des irritations d'estomac et des altérations du fonctionnement du foie et des reins. Des taux d'exposition très élevés peuvent entraîner la perte de conscience voire la mort. Des études sur des animaux ont montré que des concentrations élevées de xylène augmentent le nombre d'animaux mort-nés, ainsi que des retards de croissance et de développement. Dans beaucoup de cas, ces mêmes concentrations ont également des effets négatifs sur la santé des mères. L'effet d'expositions

de la mère à de faibles concentrations de xylène sur le fœtus n'est pas connu à l'heure actuelle.

L'aniline est une substance très toxique qui doit être manipulée avec précaution. Une exposition à des concentrations élevées peut être mortelle. Elle peut être absorbée par inhalation, ingestion et contact avec la peau, *y compris sous forme vapeur*.

Les chlorobenzènes sont des substances toxiques qui doivent être manipulées avec précautions.

Les nitrobenzènes peuvent causer des empoisonnements graves par ingestion, inhalation ou contact avec la peau. Ils réagissent avec l'hémoglobine du sang et l'empêchent de réagir avec l'oxygène. Ils peuvent également entraîner des troubles du système nerveux central, causant un sentiment de faiblesse, des maux de tête et des vomissements. Un taux élevé de nitrobenzène peut entraîner la mort en moins d'une heure ; en outre, son effet toxique est exacerbé par la prise d'alcool.

Le phénol est rapidement absorbé par toutes les voies d'exposition. L'absorption est estimée à 70 - 80 % en 6 heures pour une exposition à des vapeurs de phénol à des concentrations comprises entre 6 et 20 mg/m³ (1,6 et 5,2 ppm). Le phénol est ensuite rapidement distribué dans tous les tissus. Les organes cibles sont le cerveau et les reins [1]. Le foie, les poumons et la muqueuse gastro-intestinale sont les principaux sites de métabolisation du phénol. Ceux-ci dépendent de la voie d'exposition. Le phénol se conjugue pour former des sulfo- et glucuro- conjugués. Le phénylsulfate est le principal métabolite, (2/3) sont excrétés dans les urines en 24 heures. Cette sulfatation se réalise dans de nombreux tissus. Seule une petite fraction de phénol est transformée en catéchol ou en hydroquinone [2-8]. La formation de métabolites réactifs comme le 4,4-biphénol ou le diphénoquinone est rapportée lors d'études réalisées in vitro avec des neutrophiles humaines actives ou des leucocytes [9]. Le phénol est essentiellement éliminé par voie urinaire [10-12]. On trouve du phénol normalement dans les urines des sujets sans exposition connue [13]. Cependant, il existe une corrélation entre les concentrations urinaires en phénol et l'exposition humaine. Les principaux métabolites urinaires sont le phényl glucuronide, le phényl sulfate, 1,4-dihydroxybenzene glucuronide et le 1,4-dihydroxybenzene sulfate [12,14,15].

I-4- Dose létale 50(DL₅₀).**I-4-1 Définition.**

La dose létale 50 ou DL50 (*LD50 en anglais pour Lethal Dose 50*) ou CL50 (concentration létale 50) est un indicateur quantitatif de la toxicité d'une substance.

Cet indicateur mesure la dose de substance causant la mort de 50 % d'une population animale donnée (souvent des souris ou des rats) dans des conditions d'expérimentation précises.

La dose minimale mortelle chez l'animal, ou dose létale, est toujours délicate à déterminer de façon précise. On préfère établir la DL 50 définie comme "l'estimation statistique d'une dose unique de produit supposée pour tuer 50 % des animaux" en expérimentation.

L'essai est pratiqué habituellement sur 5 ou 6 lots d'animaux, généralement le rat. Chaque animal d'un même lot reçoit une dose identique (dose unique) de la substance à tester, mais la dose administrée est différente d'un lot à l'autre, afin que le pourcentage en mortalité varie entre 0 et 100. La voie d'administration est celle qui sera utilisée en clinique s'il s'agit d'un médicament (voie orale, injection, etc.), ou celle par laquelle la substance pourra pénétrer dans l'organisme s'il s'agit d'un produit chimique (voie orale, inhalation, voie transcutanée, etc...).

Après l'administration les animaux sont observés pendant 14 jours au cours desquels les examens cliniques sont fréquents. L'instant et les circonstances de la mort sont soigneusement notés. Les animaux encore en vie à la fin de l'essai sont sacrifiés. Tous les animaux (morts en cours d'essai et sacrifiés en fin d'essai) font l'objet d'une autopsie.

On construit ensuite la courbe donnant le pourcentage de mortalité en fonction du logarithme de la dose. C'est une courbe en "S", dite courbe de Trévan, qui peut être linéarisée par des moyens appropriés. On en déduit la DL50 (exprimée en mg par kg de poids corporel), dont on calcule aussi l'écart type.

Pour une substance administrée par voie orale on considère que :

- si la DL 50 est < 5 mg/kg, le produit est extrêmement toxique ;
- si la DL 50 est comprise entre 5 et 50 mg/kg, le produit est très toxique ;

- si la DL 50 est comprise entre 50 et 500 mg/kg, le produit est toxique ;
- si la DL 50 est comprise entre 0,5 et 5 g/kg, le produit est peu toxique ;
- si la DL 50 est > 5 g/kg, le produit n'est pas toxique ou très peu.

I-4-2 Formes de toxicité.

Les effets toxiques des produits chimiques, et comment ils provoquent ces effets, peuvent être divisés de plusieurs façons. Pour les études de propriétés par QSAR, les effets toxiques ont été divisés en trois grandes catégories : toxicité récepteur négociée ; toxicité aiguë non récepteur négociée ; et les effets supposés sur la santé humaine [16].

I-4-2-a Toxicité aiguë.

L'effet d'une dose unique entraîne généralement la mort.

L'expérimentation animale permet de quantifier la dose qui provoque la mort de 50% des sujets soumis au produit, cette dose létale est utilisée pour classer des produits chimiques en vue de leur étiquetage et conditionne l'affichage des pictogrammes « très toxique », « toxique » et « nocif ».

Exemples de produits occasionnant des intoxications aiguës : le monoxyde de carbone (CO) issu d'une combustion incomplète (incendie), l'acide cyanhydrique (HCN), le dichlore (Cl_2) gaz issu, par exemple, de la réaction de l'eau de javel avec un acide...

I-4-2-b Toxicité subaiguë.

Due à l'administration répétée d'un produit, sur une période n'excédant pas trois mois. Exemple le tétrachlorure de carbone (CCl_4), solvant chloré, une intoxication grave se manifeste par hépatite, suivie d'une cirrhose pouvant évoluer vers un cancer.

I-4-2-c Toxicité à long terme (chronique).

S'évalue après exposition répétée à de faibles concentrations du produit, pendant toute la durée de vie de l'animal. Cette exposition est utilisée pour déterminer les effets à long terme d'une substance. Ces effets, fonction de la dose totale absorbée, permettent de déterminer des doses seuils ou valeurs limites d'exposition, souvent utilisées pour fixer les limites à ne pas dépasser avec les substances cancérogènes et tératogènes.

I-4-3 Utilisation de la dose létale 50.

On administre généralement le toxique à des animaux répartis en plusieurs groupes et ce, à des doses croissantes suffisantes pour obtenir un pourcentage de mortalité s'échelonnant entre 0 et 100 %. L'effet d'une substance est, globalement, inversement proportionnel à la masse de l'animal à qui elle est administrée, c'est pourquoi elle est mesurée en g/kg. En général, si la toxicité immédiate est semblable chez tous les types d'animaux, elle sera probablement semblable chez les humains. Lorsque les DL_{50} sont différentes chez diverses espèces animales, on doit faire des approximations et des hypothèses lors de l'estimation de la dose mortelle probable chez les humains.

I-4-4 Identification de la toxicité.

La DL_{50} sert à mesurer toute la toxicité d'une substance, mesure qui s'effectue via des études qualitatives (non mesurables) et quantitatives (mesurables dont la DL_{50}).

La DL_{50} sert souvent de départ aux études de toxicité car elle fournit un minimum de connaissances en identifiant les symptômes de l'intoxication et la dose toxique.

Il faut malgré tout la considérer avec prudence car c'est souvent une étude préliminaire (première analyse) qui peut être influencée par plusieurs facteurs, tels l'espèce animale, le sexe, l'âge, le moment de la journée, etc.

Elle a cependant une valeur limitée, car elle ne concerne que la mortalité, d'où l'apparition de valeurs comme l' IC_{50} .

Il existe d'autres méthodes d'étude de la toxicité, par exemple les tests d'irritation de la peau et de corrosion des yeux, qui font généralement partie d'un programme d'évaluation toxicologique.

I-4-5 Identification du pouvoir pathogène.

La DL50 est une des deux données servant à mesurer le pouvoir pathogène d'un germe. La seconde donnée étant la Dose Minimale Infectante (DMI).

II- LES ALCOOLS ET LES AMINES.

II-1 Introduction.

Les alcools et les amines sont des composés organiques qui sont largement utilisés dans le milieu industriel et dans la vie quotidienne. Que ce soit professionnellement ou à la maison, ces deux familles de produits sont incontournables. Les alcools sont trouvés à l'état pur ou en mélange dans des préparations spécifiques. Ils sont utilisés comme diluants des encres d'imprimerie, des résines, des vernis, des peintures et des colles à moquette. Ils sont aussi largement utilisés comme excipients pour les produits pharmaceutiques ou cosmétiques ou comme milieu réactionnel dans l'industrie chimique [17]. Les amines sont utilisées comme intermédiaires chimiques pour la synthèse de produits pharmaceutiques, cosmétiques, détergents.... Elles sont aussi utilisées comme solvants et inhibiteurs de corrosion [18].

Les alcools ont des effets néfastes bien connus, entraînant notamment des incoordinations motrices ou une excitation intellectuelle. Les alcools liquides et leurs vapeurs sont irritants pour la peau, les yeux et les muqueuses en cas de contact prolongé ou répété. L'inhalation accidentelle d'une grande quantité de vapeurs d'alcool peut conduire à des syndromes ébrieux ou narcotiques avec nausées, malaises, vomissements et maux de tête [17]. Quant aux effets indésirables des amines, les problèmes de santé pouvant se développer chez les travailleurs surexposés sont divers, allant de l'irritation cutanée au cancer [19].

Le transport, la distribution, l'accumulation et l'absorption des xénobiotiques (médicaments ou toxiques), pour certains leur fixation sur les protéines plasmatiques, leur passage à travers

les membranes, leur entrée dans les cellules, les interactions enzyme-inhibiteur ou ligand-récepteur de nature hydrophobe, l'activité pharmacologique et pharmacocinétique des médicaments, la toxicité des médicaments ou des contaminants,..., les propriétés liées à la formulation telles que la solubilité, sont autant d'exemples dans lesquels la lipophilie des molécules constitue un descripteur physico-chimique de la première importance. La lipophilie d'une molécule est mesurée classiquement par son aptitude à se distribuer dans un système biphasique soit liquide-liquide soit solide-liquide. Le coefficient de partage dans le système *n*-octanol/eau est connu depuis longtemps comme étant un des paramètres physico-chimiques quantitatifs qui est le mieux corrélé à l'activité des molécules organiques [20].

II-2- Les alcools :

En chimie organique, un **alcool** est un composé organique dont l'un des carbones (celui ci étant tétraédrique) est lié à un groupement hydroxyle (-OH). L'éthanol (ou alcool éthylique) entrant dans la composition des boissons alcoolisées est un cas particulier d'alcool, mais tous les alcools ne sont pas propres à la consommation. En particulier, le méthanol est toxique et mortel à haute dose.

II-2-1- Utilisation.

Les alcools sont utilisés dans l'industrie chimique comme:

- Solvants : l'éthanol, peu toxique, est utilisé dans les parfums et les médicaments ;
- Combustibles : le méthanol et l'éthanol peuvent remplacer l'essence et le fioul car leur combustion ne produit pas de fumées toxiques ;
- Réactifs : les polyuréthanes, les esters ou les alcènes peuvent être synthétisés à partir des alcools ;
- Antigels : la basse température de solidification de certains alcools comme le méthanol et l'éthylène glycol en font de bons antigels.

II-2-2- Propriétés des alcools.

Les alcools sont des composés amphotères, c'est-à-dire qu'ils sont à la fois acide et base. En d'autres termes, ils peuvent être protonés par action d'un acide ou déprotonés par action d'une base. Dans le cas de cette déprotonation il sera possible de faire une *O*-alkylation, et donc d'obtenir un éther. Dans le cas de la protonation, on fait un ion oxonium qui conduira à un carbocation qui pourra subir un réarrangement de façon à donner le carbocation le plus stable.

II-2-3 Toxicité des alcools.

De tous les alcools, le plus toxique est le méthanol dans la mesure où il exerce une action sélective au niveau du nerf optique, pouvant provoquer la cécité ou la mort. Les effets néfastes de l'absorption d'éthanol sont aussi bien connus, l'alcoolémie entraînant notamment des incoordinations motrices ou une excitation intellectuelle. De manière générale, les manifestations d'une intoxication modérée se traduiront par des maux de tête, des troubles digestifs et un syndrome ébrieux.

Les alcools liquides et leurs vapeurs sont irritants pour la peau, les yeux et les muqueuses en cas de contact prolongé ou répété. L'alcool furfurylique, plus agressif que les autres alcools, peut provoquer des larmoiements à de très faibles expositions (15 ppm) et des irritations respiratoires.

L'inhalation accidentelle d'une grande quantité de vapeurs d'alcool peut conduire à des syndromes ébrieux ou narcotiques avec nausées, malaises, vomissements et maux de tête.

II-3-Les amines :

Les amines sont des composés azotés qui dérivent formellement de l'ammoniac NH_3 par remplacement d'un ou plusieurs atomes d'hydrogène par des groupes carbonés. Le nombre n des atomes d'hydrogène liés à l'azote, définit la classe de l'amine. Leur découverte est due au chimiste français Wurtz en 1849.

<i>N</i>	2	1	0
Classe	Primaire	Secondaire	Tertiaire

II-3-1 Utilisation.

Les amines sont utilisées en grande quantité dans des procédés industriels variés, comme un intermédiaire réactionnel, dans des mélanges ou comme solvant. On les emploie surtout dans l'industrie chimique, dans celle des polymères et du caoutchouc, en agriculture comme pesticides, de même que dans la composition de peintures, d'adhésifs et de textiles, ainsi que dans l'industrie pharmaceutique [18].

II-3-2 Propriétés des amines.

La présence de l'atome d'azote est la cause des propriétés des amines. Cet atome présente un doublet non liant, ce qui donne aux amines un caractère basique et nucléophile. Dans le cas d'amines primaires et secondaires, la liaison N-H peut se rompre, ce qui leur donne un (faible) caractère acide. Les amines sont volatiles, ont une odeur forte et sont hydrosolubles [19].

II-3-3- Toxicité des amines.

Les amines sont un produit irritant et corrosif pour la peau, les yeux, les voies respiratoires et digestives. La gravité des symptômes peut varier selon les conditions d'exposition (durée de contact, concentration du produit, etc...).

L'exposition aux brouillards peut causer une irritation des yeux, de la peau et des voies respiratoires. L'exposition à de fortes concentrations peut provoquer l'œdème pulmonaire. Les symptômes de l'œdème pulmonaire (principalement toux et difficultés respiratoires) se manifestent souvent après un délai pouvant aller jusqu'à 48 heures. L'effort physique peut aggraver ces symptômes. Le repos et la surveillance médicale sont par conséquent essentiels [21].

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.

II-4 Concentration d'inhibition de la croissance CIC50.

II-4-1 Définition :

La concentration inhibitrice médiane de la croissance (CIC50) est une mesure de l'efficacité d'un composé donné pour inhiber une fonction biologique ou biochimique spécifique. Souvent, le composé en question est un éventuel médicament. Cette mesure quantitative indique quelle quantité d'un médicament ou d'une autre substance (inhibiteur) est nécessaire pour inhiber à moitié un processus biologique donné (ou un élément d'un processus, par exemple une enzyme, un paramètre cellulaire, un récepteur cellulaire ou un microorganisme). En d'autres termes, c'est la demi (50 %) concentration inhibitrice (CI) d'une substance (CI à 50 %, ou CI50). Elle est couramment utilisée comme mesure de l'efficacité d'un médicament antagoniste en recherche pharmacologique. Selon la FDA (Food and Drug Administration), la CI50 représente la concentration d'un médicament qui est requise pour une inhibition à 50 % *in vitro*. Elle est comparable à la Concentration médiane de l'Efficacité (CE50) pour les médicaments agonistes. La CE50 représente aussi la concentration plasmatique nécessaire pour obtenir 50 % d'un effet maximal *in vivo* [17].

II-4-2 Pourquoi 50 % ?

C'est pour des raisons de représentativité statistique qu'on utilise la valeur 50 %, plutôt que 0 %, 5 %, 95 %, ou 100%. En effet, la courbe de Gauss est « plate » vers 50 %, ce qui fait qu'un échantillon est plus représentatif lorsqu'un seuil est franchi à 50 %.



REFERENCES BIBLIOGRAPHIQUES

- [1] Lo Dico C., Caplan Y., Levine B., Smyth D.F, and Samialek J. (1989) – Phenol: tissue distribution in a fatality. *J Forensic Science* Vol 34(4), pp 1013-1015.
- [2] Williams R.T.(1938) – Studies in detoxification. I. The influence of (a) dose and (b) o, m, p substitution on the sulfate detoxification of phenol in the rabbit. *J Biochemistry*, Vol 32, pp 878-887.
- [3]Williams R.T. (1959) – Detoxication mechanisms. *J Biochemistry*. New York, Chichester, Brisbane, Toronto, pp.237-295.
- [4]Garton G.A, and Williams R.T. (1949) –Studies in détoxication 26. The fate of phenol, phenyl sulfuric acids and phenylglucuronide in the rabbit in relation to the metabolism of benzene. *J Biochemistry*, Vol 45, pp 158-163.
- [5]Bray H.G., Humphris B.G., Thorpe W.V, and White K, and Wood P.B. (1952a) –Kinetic studies of the metabolism fore in organic compounds . 3. The conjugation of phenol with glucuronic acid. *J Biochemistry*, Vol 52, pp 416-419.
- [6]Bray H.G., Humphris B.G., Thrope W.V., White K and Wood p.b. (1952 b) – Kenetic studies of the metabolism fore in organic compounds. 4. The conjugation of phenol with sulfuric acid. *J Biochemistry*, Vol 52, pp 419-423.
- [7] Bray H.G., Humphris B.G., Thrope W.V., and White K . (1952 c) – Kinetic studies of the metabolism fore in organic compounds. 5. A mathematical model expressing the metabolic fate of phenols, benzoic acids and their precursors. *J Biochemistry*, Vol 52, pp 423-430.
- [8]Parke D.V, and Williams R.T. (1953) – Studies in detoxification. 54. The metabolism of benzene (a) formation of phenylglucuronide and phenylsulfuric acid from ^{14}C benzene. (b) The metabolism of ^{14}C phenol. *J Biochemistry*, Vol 55, pp 337-340.
- [9]Eastmond D.A., Smith M .T., Ruzo L.G, and Ross D. (1986)- Metaobolic activation of phenol by human myeloperoxidase and horseradish peroxidase. *J Molecular Pharmacology*, Vol 30(6), pp 674-679.
- [10]Deichmann W.B. (1944)- Phenols studies- V: the distribution detoxification and excretion of phenol in the mammalian body. *Archives Biochemistry* Vol 2, pp 345-355.
- [11]Capel I.D., French M.R., Millburn P., Smith R.L, and Williams R.T. (1972a)-Species variations in phenol metabolism. *J Biochemistry*, Vol 127, pp 25-26.

- [12] Capel I.D., French M.R., Millburn P., Smith R.L, and Williams R.T. (1972b)- The fate of ^{14}C phenol in various species. *J Xenobiotica*, Vol 2, pp 25-34.
- [13] Bruce R.M., Santodonato J, and Neal N.W. (1987) – Summary review of the health effects associated with phenol. *J Toxicology & Industrial Health*, Vol 3(4), pp 535-568.
- [14] Tremaine L.M., Diamond G.L, and Quebbeman A.J. (1984) – In vivo quantification of renal glucuronide and sulfate conjugation of 1- naphthol and p-nitrophenol in the rat. *J Biochemical Pharmacology* Vol 33, pp 419-427.
- [15] Wheldrake, J. F., Baudinette, R.V, and Hewitt, S. (1978) – The metabolism of phenol in a desert rodent *Notomys alexis*. *J Comparative Biochemistry and Physiology.*, C61, pp 103-107.
- [16] Creighton T. E., (1993) – *Proteins: Structures and Molecular Properties*, Second ed., Freeman, New York, p 335.
- [17] Boust C., (2004) – Institut national de recherche et sécurité .1^{ère} édition avril, pp 1-6, www.inrs.fr.
- [18] Lauwerys R. R., Haufroid V., Huet P, and Lison D. (2007) – *Toxicologie industrielle et intoxications professionnelles*, pp 643-644.
- [19] Les amines aliphatiques, *Encyclopædia Universalis France S.A* (2005 et 2009) – [http://www. Encyclopædia .com](http://www.Encyclopædia.com).
- [20] Carpy A., (1999) – Importance de la lipophilie en modélisation moléculaire, *J Analyses magazine*, Vol 1, pp 18 et 21.
- [21] National Institute for Occupational Safety and Health, *RTECS (Registry of toxic effects of chemical substances)*, Hamilton, Ont.: Canadian Centre for Occupational Health and Safety (CD-ROM) <http://ccinfoweb.ccohs.ca/rtecs/search.html> .

PARTIE II

*ASPECTS THEORIQUES DE LA MODELISATION
MOLECULAIRE*

*I-Optimisation de la géométrie de
la molécule*

*II- Modélisation et évaluation de la qualité
d'un modèle*



*OPTIMISATION DE LA
GEOMETRIE DE LA MOLECULE*

I-1 GENERALITES

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:

- 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 [1] 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^{-ax}) ou d'orbitales gaussiennes (e^{-ax^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'ensemble exact de fonctions de base utilisé est souvent spécifié par une abréviation du genre STO - 3G ou 6 - 311 ++ g **.

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 décades, 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 (MM), 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 [2-4]. 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 [5].

I-2 METHODES SEMI-EMPIRIQUES UTILISEES.

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.

I-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) [6, 7].

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), $\Phi_a = \phi_a \alpha$, $\bar{\Phi} = \phi_a \beta$.

On représentera ψ_0 par :

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

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) [8, 9].

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 [10]). Elles servent dans la description des états excités.

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

$$\mathbf{F} \mathbf{C} = \mathbf{S} \mathbf{C} \boldsymbol{\varepsilon} \quad (3)$$

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 :

$$\begin{aligned} F_{\mu\nu} &= H_{\mu\nu} + J_{\mu\nu} - \frac{1}{2} K_{\mu\nu} \\ F_{\mu\nu} &= H_{\mu\nu} + \sum_{\rho}^n \sum_{\sigma}^m P_{\rho\sigma} \left[\langle \mu\nu/\rho\sigma \rangle - \frac{1}{2} \langle \mu\sigma/\rho\nu \rangle \right] \end{aligned} \quad (4)$$

Avec :

$$H_{\mu\nu} = \int \chi_{\mu}^*(1) \hat{h} \chi_{\nu}(1) d\tau_1 \quad (5)$$

$$\langle \mu\nu/\rho\sigma \rangle = \iint \chi_{\mu}^*(1) \chi_{\nu}(1) \frac{1}{r_{12}} \chi_{\rho}^*(2) \chi_{\sigma}(2) d\tau_1 d\tau_2 \quad (6)$$

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 monoélectroniques 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 bi-électroniques 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 :

$$E_{el} = 2 \sum_a^m \epsilon_a - \frac{1}{2} \sum_{\mu\nu}^m P_{\mu\nu} \left(J_{\mu\nu} - \frac{1}{2} K_{\mu\nu} \right) \quad (8)$$

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) [11].

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) [12, 13].

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

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

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

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 électronique 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.

I-2- 2-Analyse de population de Mulliken [14, 15].

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 :

$$d\rho = \psi^* \psi d\tau \quad (10)$$

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

$$dQ = e dP \quad (11)$$

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 :

$$\int_E d\rho = N \quad (\text{N électrons}) \quad (12)$$

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

$$\int_E \psi^* \psi d\tau = N = \sum_{i=1}^m 2 \int \psi^*(\nu) \psi_i(\nu) d\tau_\nu \quad (13)$$

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

En décomposant sur la base atomique, on obtient :

$$\sum_{i=1}^m 2 \int \left(\sum_{l=1}^n c_{li}^* \varphi_l(\nu) \right) \left(\sum_{m=1}^n c_{mi} \varphi_m(\nu) \right) d\tau_\nu = 2 \sum_{i=1}^m \sum_{l=1}^n \sum_{m=1}^n c_{li}^* c_{mi} \int \varphi_l(\nu) \varphi_m(\nu) d\tau_\nu \quad (14)$$

$$= \sum_{l=1}^n \sum_{m=1}^n P_{lm} S_{lm} = \sum_{l=1}^n \left[\sum_{m=1}^n P_{lm} \right] = N$$

Avec :

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

Et :

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

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 $l^{\text{ème}}$ 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 $q_{l(L)}$ ($l \in L$) :

$$Q_L = \sum_{l(L)} q_{l(L)} \quad (15)$$

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

$$\sum_l q_l = N \quad (16)$$

$$\sum_L Q_L = N$$

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 :

$$C_L = Z_L - Q_L \quad (17)$$

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

I-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) [16], 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 [17] 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.

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 :

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

où :

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

$n(\text{val})$ 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) $H_{\mu_A \mu_B}^c = \langle \mu_A(1) | \widehat{H}_{(1)}^c | \mu_B(1) \rangle$, 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 :

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

Des considérations de la théorie des groupes [18] permettent d'annuler $\langle \mu_A | -\frac{1}{2} \nabla_1^2 + V_A | \nu_A \rangle$, de telle sorte que :

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

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

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

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

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

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

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

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

$U_{\mu_A \mu_A}^c = \langle \mu_A | -\frac{1}{2} \nabla^2 + V_A | \nu_A \rangle$ est évalué à partir de paramètres tirés de spectres atomiques (les paramètres utilisés pour l'atome de carbone : U_{ss} et U_{pp}). Donc :

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

L'évaluation de $\langle \mu_A \nu_A | s_B s_B \rangle$ 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 $g_{\mu \nu} \langle \mu_A \mu_A | \nu_A \nu_A \rangle$, soit des intégrales d'échange

$$h_{\mu \nu} \langle \mu_A \nu_A | \mu_A \nu_A \rangle.$$

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 [19].

- 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.

I-3 Champ de force.

I-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.

I-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/mm2 mm3 chtml](http://enropa.chem.uga.edu/allinger/mm2%20mm3.html)
introduit par Allinger et al [20-23], largement utilisé pour le traitement de petites molécules.

- **AMBER**: [http:// amber. Scripps.edu](http://amber.Scripps.edu)
(Assisted Method Building and Energy Refinement) introduit par Cornell et al [24], 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 [22] 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 [25-27], 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.

I-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 :

$$\begin{aligned} V(\mathbf{r}^N) = & \sum_{\text{liaisons}} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} (1 - \cos(n\omega - \gamma)) \\ & + \sum_{i=1}^N \sum_{j=i+1}^N \left(4 \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \end{aligned} \quad (36)$$

$V(\mathbf{r}^N)$ 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 suivante :

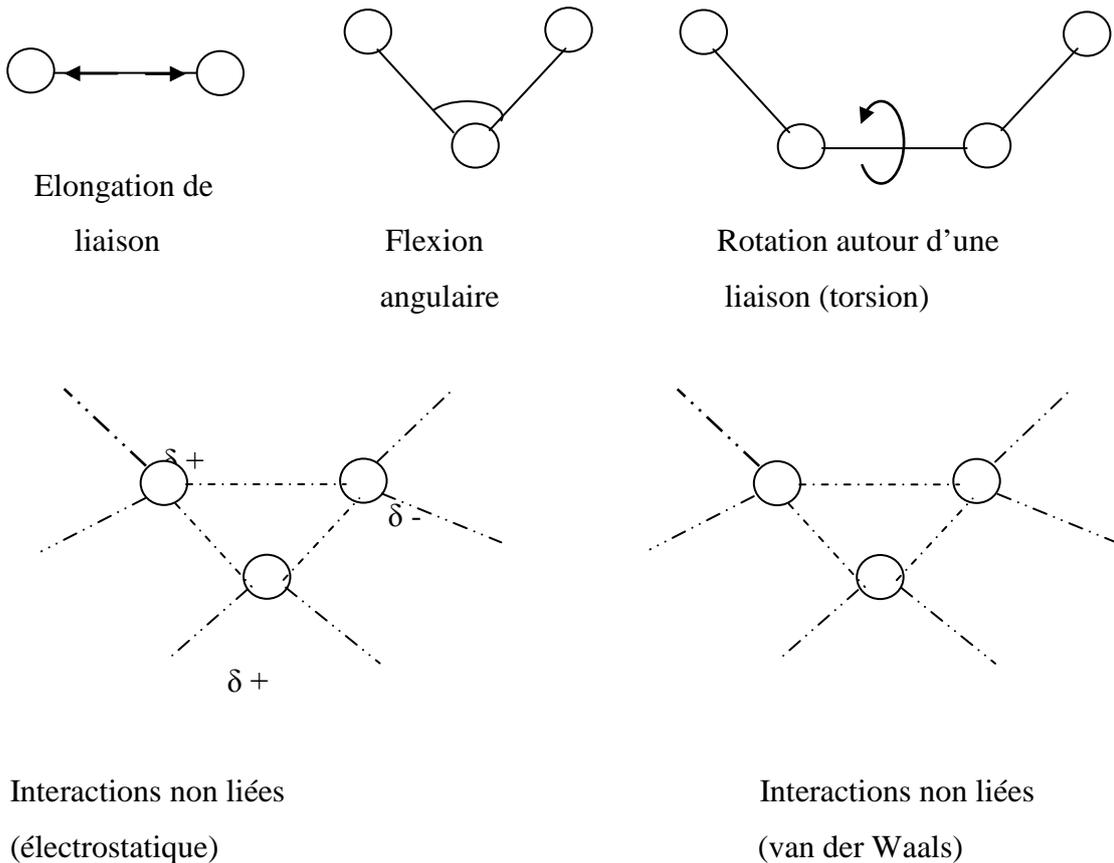


Fig.-1 : Représentation schématique des quatre contributions à un champ de force de MM : élongation de liaison, flexion angulaire

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.

I-3 - 4 Exemple de calcul : énergie d'une conformation du propane

A titre d'illustration nous montrons comment la relation (36) peut être utilisée pour calculer l'énergie de conformation du propane (Fig. 2).

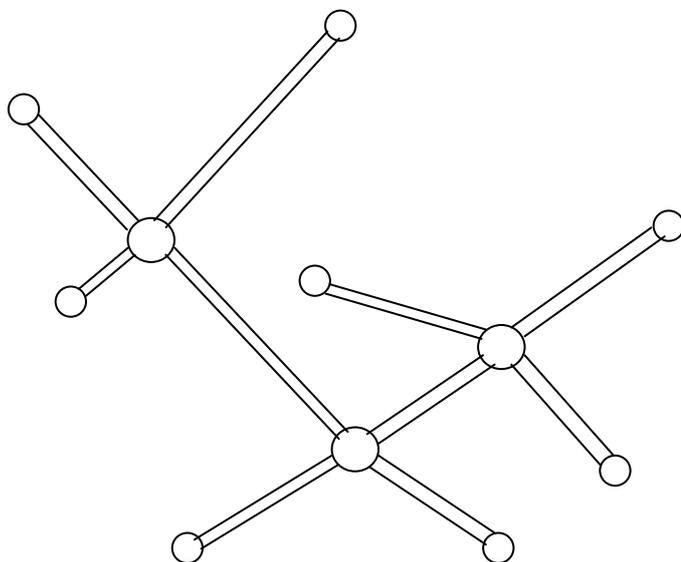


Fig.2: Un modèle de champ de force typique pour le propane contient 10 termes d'élongation de liaison, 18 termes de flexion angulaire, 18 termes de torsion et 27 interactions de non – liaison.

Le propane possède 10 liaisons : 2 liaisons C-C et 8 liaisons C-H. Les liaisons C-C sont symétriques et équivalentes, mais les liaisons C-H appartiennent à 2 classes, un groupe comprend les 2H liés au carbone central du méthylène (CH_2) et un groupe correspondant aux 6 hydrogènes liés aux carbones des groupements méthyls.

Dans certains champs de force compliqués des paramètres différents seront utilisés pour ces 2 types de liaison C-H, mais dans la plupart des champs de force les paramètres de liaison (k_i et $l_{i,0}$) seront utilisés pour les 8 liaisons C-H. Il y a 18 angles de valence différents pour le propane, comprenant un angle C-C-C, 10 angles C-C-H et 7 angles H-C-H. Il est à noter que tous les angles sont pris en compte dans le modèle de champ de force quoique certains d'entre eux peuvent ne pas être indépendants des autres. Il y a 18 termes de torsion : 12 de type H-C-C-H et 6 du type H-C-C-C. Chacun d'eux est modélisé par un développement en série de cosinus présentant des minima pour les conformations trans et gauche. Finalement, Il y a 27 termes de non-liaison à calculer, impliquant 21 interactions H-H et 6 interactions H-C. La contribution électrostatique sera obtenue en appliquant la loi de Coulomb aux charges atomiques partielles et la contribution de van der Waals en utilisant un potentiel de Lennard – Jones avec des paramètres ϵ_0 et σ appropriés. Un assez grand nombre de termes sont ainsi inclus dans le modèle de champ de force, même pour une molécule aussi simple que le propane. Même ainsi, le nombre de termes (73) est beaucoup moindre que le nombre d'intégrales qui seraient impliquées dans un calcul quanto-mécanique équivalent.

I-3 – 5 – Champs de force MM2 et MM+.

I-3 – 5 – 1 - Champ de force MM2.

* 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 :

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

* 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.

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

* 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 (Fig.3).

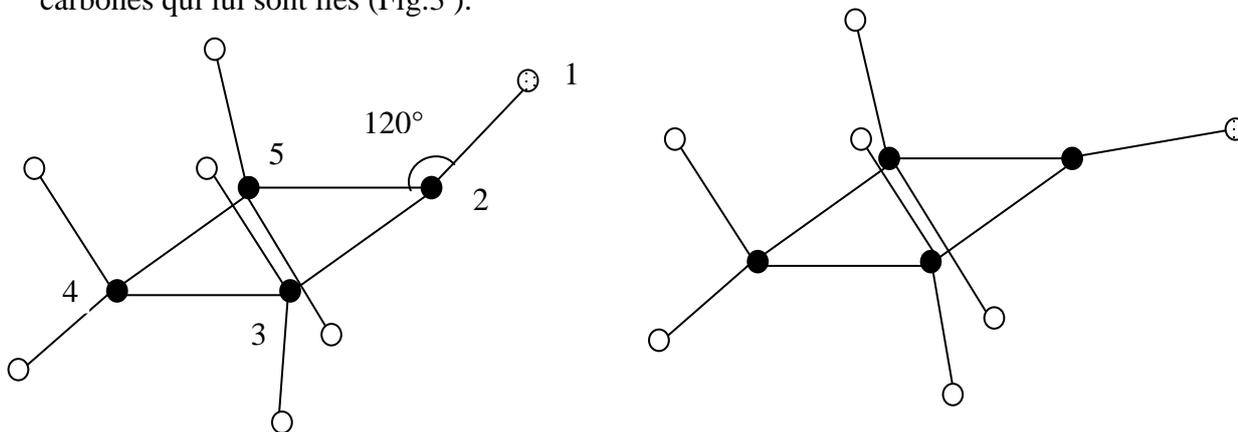


Fig.3 : Sous un terme extra - planaire, l'atome d'oxygène de la cyclobutanone est prévu de se situer hors du plan du cycle / gauche plutôt que dans le plan.

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 maximisée dans l'arrangement coplanaire, sera plus réduite si l'atome d'oxygène est dévié 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 (Fig. II-3) 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 .

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 (Fig. II-4). 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 (Fig. II- 4). 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 :

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

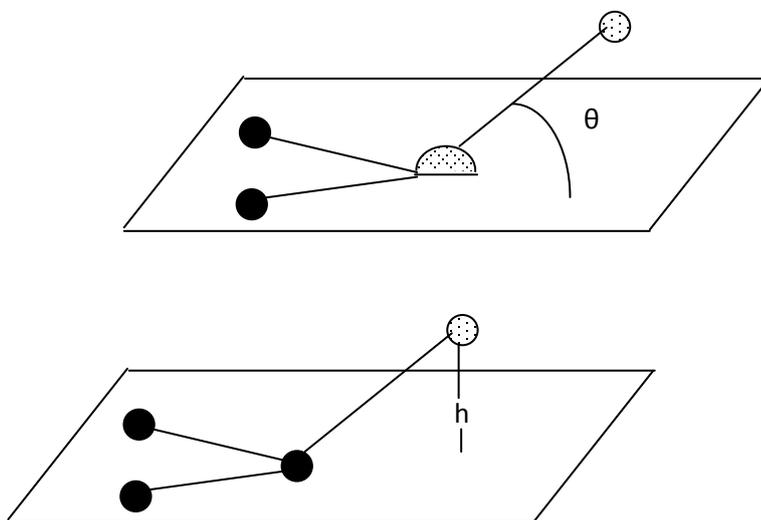


Fig.4 : Deux façons pour modéliser les contributions de la variation d'angle extra – planaire.

* 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 « élancement – 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 $\Delta \theta = \theta - \theta_0$

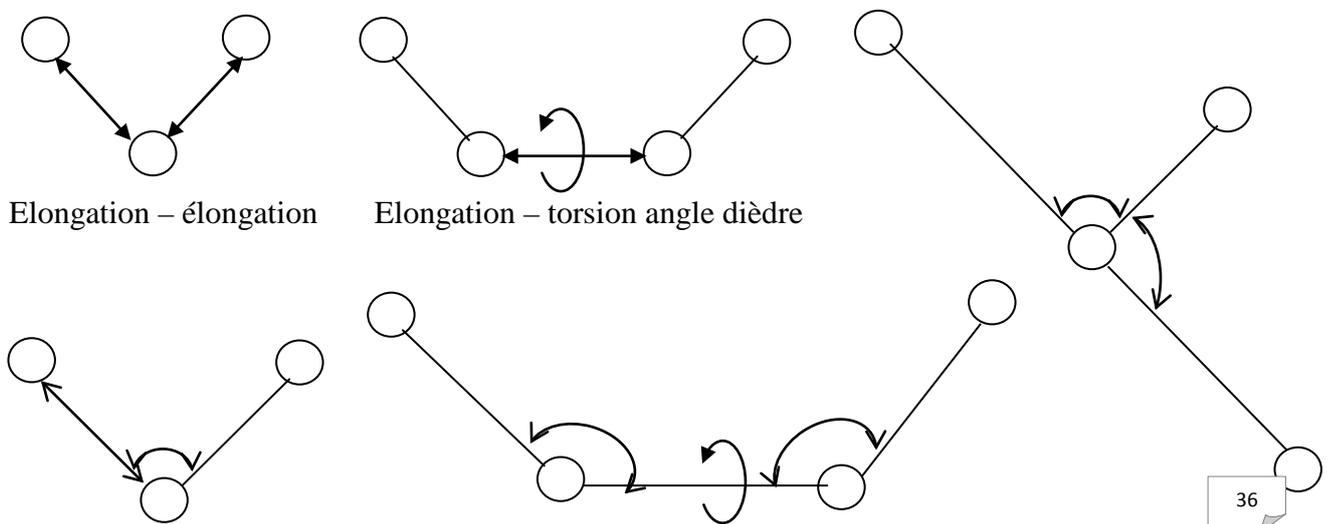
l_{10} , l_{20} et θ_0 représentent les valeurs de référence pour l_1 , l_2 et θ respectivement.

Les termes de croisement les plus utilisés sont (Fig. II- 5) :

* élancement – élancement et élancement – variation angulaire, pour deux liaisons à un même atome ;

* élancement – 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 élancement – variation angulaire.



ensuite utilisés pour affecter des longueurs de liaisons et des constantes de force de référence pour les liaisons conjuguées.

I-3 – 5 – 2 - 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.

Le tableau suivant [28] compare les trois techniques computationnelles majeures évoquées.

Tableau I : Etude comparative des techniques ab initio, semi-empirique et mécanique moléculaire (d'après [28]).

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. <ul style="list-style-type: none"> - Peut être appliquée à des composés inorganiques, organiques, 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. <ul style="list-style-type: none"> - Applicable uniquement à l'état fondamental.

II-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 $\{ y(x^i), i = 1 \dots N \}$. La détermination des valeurs de ces paramètres nécessite la mise en œuvre de méthodes d'optimisation qui diffèrent selon le type de modèle choisi.

Nous allons présenter les trois types de modèles exploités dans cette thèse.

II-1-1- Régression linéaire multiple (MLR) [29-31]

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 $x = \{ x_k, k = 1 \dots q \}$. Ces entrées peuvent être des fonctions non paramétrées, ou à paramètres fixés, de ces variables. L'équation linéaire recherché est de la forme :

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

où $\theta = \{ \theta_k, k=1 \dots q \}$ est le vecteur des paramètres; X, matrice des observations de taille (N, q), 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 :

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

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 :

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

La fonction $J(\theta)$ est une fonction positive quadratique en θ : 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 grandeur étudiée.

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

II-1-2 MODELE QSXR (X = activité, propriété, toxicité, rétention) non linéaires: Réseaux de neurones artificiels (RNA) ; Régression par machines à vecteurs de support (SVR).

Les modèles QSXR intrinsèquement non linéaires peuvent être développés en utilisant les réseaux de neurones artificiels, ou la régression par machines à vecteurs de support.

II-1-2-1-Réseaux de neurones artificiels [32-35]

Un réseau de neurones artificiels (RNA) est un programme informatique conçu pour apprendre à partir de données qui lui sont présentées, en s'inspirant du schéma d'apprentissage effectué par le cerveau humain où le neurone est l'unité fonctionnelle de base

du système nerveux. Les RNA représentent un moyen puissant pour le développement des relations non linéaires entre variables, ce qui en fait d'excellents outils de prédiction dans différents domaines.

Les réseaux multicouches, qui utilisent l'apprentissage supervisé [36], sont des puissants réseaux de neurones ; ils comportent en plus de l'entrée et de la couche de sortie, de une à plusieurs couches cachées (figure 6). Chaque neurone est uniquement relié à tous les neurones de la couche

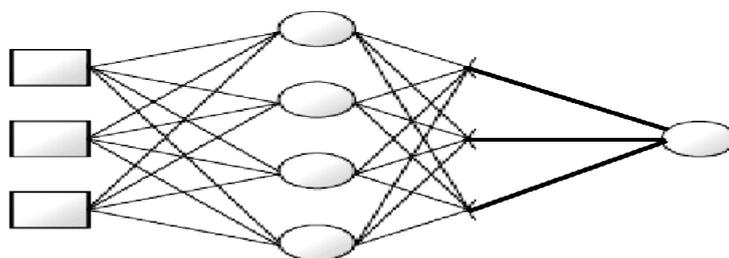


Figure.6: Réseau de neurones avec 3 entrées, une couche cachée avec 4 neurones et une couche de sortie avec 1 neurone.

Chaque connexion étant caractérisée par un poids. La façon dont chaque neurone transforme son entrée dépend du poids et du biais qui lui sont associés, ces 2 paramètres étant modifiables. L'apprentissage d'un RNA peut être réalisé à l'aide de l'algorithme de rétropropagation. Dans ce but, les valeurs d'entrée sont présentées, après une transformation éventuelle, au réseau qui les propage jusqu'à la couche de sortie et donne ainsi la réponse au réseau. Dans une deuxième étape les bonnes sorties correspondantes sont représentées aux neurones de la couche de sortie qui calculent l'écart, modifient leurs poids et leurs biais, et rétropropagent l'erreur jusqu'aux entrées pour permettre aux neurones cachés de modifier leurs poids et leurs biais, de la même façon. Ce processus itératif est stoppé en utilisant comme critère « l'arrêt précoce », c'est-à-dire dès que l'indice de performance (erreur quadratique moyenne : EQM) calculé sur les données de test cesse de s'améliorer.

Dans la plupart des applications des RNA à la chimie l'utilisation d'une seule couche cachée semble suffire [36]. 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 [37] 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.

II-1-2-2-Régression par machines à vecteurs de support [38, 39]

La régression par machines à vecteurs de support (SVR) consiste à trouver la fonction $f(x)$ qui a au plus une déviation ε par rapport aux exemples d'apprentissage (x_i, y_i) , pour $i=1, \dots, N$, et qui est la plus plate possible. Cela revient à ne pas considérer les erreurs inférieures à ε et à interdire celles supérieures à ε [40]. Maximiser la platitude de la fonction permet de minimiser la complexité du modèle qui influe sur ses performances en généralisation. En effet, la théorie de l'apprentissage [38] permet de borner l'erreur de généralisation par une somme de deux termes : l'un dépendant de la complexité du modèle et l'autre dépendant de l'erreur sur les données d'apprentissage [41].

Les méthodes SVM sont basées sur le contrôle de la complexité du modèle lors de l'apprentissage.

Dans la méthode SVM, différents hyperparamètres apparaissent : C , qui représente le compromis entre la complexité du modèle et l'erreur sur les données d'apprentissage ; ε , qui correspond à la largeur du tube d'insensibilité ; les éventuels paramètres de la fonction noyau k (σ, γ, \dots). Ces hyperparamètres sont en général réglés en fonction et d'une estimation de l'erreur de généralisation qui peut être évaluée sur un jeu indépendant de données de validation croisée. Cela implique de réaliser l'apprentissage pour différentes valeurs et estimer leurs performances. Dans le cas d'une estimation de l'erreur de généralisation par validation croisée, cette procédure peut se révéler très coûteuse en temps de calcul.

Une idée permettant d'éviter les temps de calcul trop importants dans la recherche des hyperparamètres optimaux est donnée dans [40].

II-2- DEVELOPPEMENT ET EVALUATION DE MODELE.

II-2-1 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 (GA) que nous avons appliqués dans le présent travail, et que nous rappelons succinctement.

II-2-2 Principe.

Dans la terminologie des algorithmes génétiques, le vecteur binaire \tilde{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.

II-2-3 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) ;

II-2-4 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).

II-2-5 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.

II-2-6 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 [42], 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.

II-3 Développement des modèles

Les techniques les plus courantes pour établir des modèles QSAR utilisent l'analyse de régression (régression linéaire multiple : RLM ; projection des structures latentes par les moindres carrés partiels : PLS), les réseaux neuronaux (RNA), et machine à support de vecteur (SVM).

II-3-1 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) :

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

II-3-2 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) [43]. 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 Residual Sum of Squares) :

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 \quad (54)$$

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

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

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 $Q_{\text{LoO}}^2 > 0,5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [44].

II-3-3 Domaine d'application.

Le domaine d'application a été discuté à l'aide du diagramme de Williams (traité en détail dans [44 ,45], 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)

$$h_i = x_i (X^T X)^{-1} x_i^T \quad (56)$$

Où (x_i) est le vecteur ligne des descripteurs du composé i et 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 $h_{ii} < h^*$, 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 $h_{ii} > h^*$ 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.

II-3-4 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) [47].

II-3-5 Validation externe.

En plus du test de randomisation, il est intéressant [48], pour juger de la qualité du modèle, de considérer le coefficient de prédiction externe calculé comme suit :

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

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 EQMP_{ext}).

Ces valeurs RMSE sont mieux adaptées, pour juger de la qualité d'un modèle que les valeurs de R^2 et Q^2 seules, qui constituent de bons tests uniquement pour des données réparties régulièrement.

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

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

II-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 $\{ y(x^i), i = 1 \dots N \}$. La détermination des valeurs de ces paramètres nécessite la mise en œuvre de méthodes d'optimisation qui diffèrent selon le type de modèle choisi.

Nous allons présenter les trois types de modèles exploités dans cette thèse.

II-1-1- Régression linéaire multiple (MLR) [28-30]

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 $x = \{ x_k, k = 1 \dots q \}$. Ces entrées peuvent être des fonctions non paramétrées, ou à paramètres fixés, de ces variables. L'équation linéaire recherché est de la forme :

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

où $\theta = \{ \theta_k, k=1 \dots q \}$ est le vecteur des paramètres; X, matrice des observations de taille (N, q), 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 :

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

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 :

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

La fonction $J(\theta)$ est une fonction positive quadratique en θ : 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 grandeur étudiée.

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

II-1-2 MODELE QSXR (X = activité, propriété, toxicité, rétention) non linéaires: Réseaux de neurones artificiels (RNA) ; Régression par machines à vecteurs de support (SVR).

Les modèles QSXR intrinsèquement non linéaires peuvent être développés en utilisant les réseaux de neurones artificiels, ou la régression par machines à vecteurs de support.

II-1-2-1-Réseaux de neurones artificiels [31-34]

Un réseau de neurones artificiels (RNA) est un programme informatique conçu pour apprendre à partir de données qui lui sont présentées, en s'inspirant du schéma d'apprentissage effectué par le cerveau humain où le neurone est l'unité fonctionnelle de base

du système nerveux. Les RNA représentent un moyen puissant pour le développement des relations non linéaires entre variables, ce qui en fait d'excellents outils de prédiction dans différents domaines.

Les réseaux multicouches, qui utilisent l'apprentissage supervisé [35], sont des puissants réseaux de neurones ; ils comportent en plus de l'entrée et de la couche de sortie, de une à plusieurs couches cachées (figure 6). Chaque neurone est uniquement relié à tous les neurones de la couche

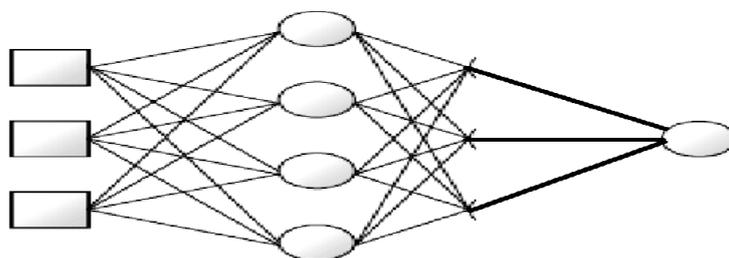


Figure.6: Réseau de neurones avec 3 entrées, une couche cachée avec 4 neurones et une couche de sortie avec 1 neurone.

Chaque connexion étant caractérisée par un poids. La façon dont chaque neurone transforme son entrée dépend du poids et du biais qui lui sont associés, ces 2 paramètres étant modifiables. L'apprentissage d'un RNA peut être réalisé à l'aide de l'algorithme de rétropropagation. Dans ce but, les valeurs d'entrée sont présentées, après une transformation éventuelle, au réseau qui les propage jusqu'à la couche de sortie et donne ainsi la réponse au réseau. Dans une deuxième étape les bonnes sorties correspondantes sont représentées aux neurones de la couche de sortie qui calculent l'écart, modifient leurs poids et leurs biais, et rétropropagent l'erreur jusqu'aux entrées pour permettre aux neurones cachés de modifier leurs poids et leurs biais, de la même façon. Ce processus itératif est stoppé en utilisant comme critère « l'arrêt précoce », c'est-à-dire dès que l'indice de performance (erreur quadratique moyenne : EQM) calculé sur les données de test cesse de s'améliorer.

Dans la plupart des applications des RNA à la chimie l'utilisation d'une seule couche cachée semble suffire [35]. 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 [36] 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.

II-1-2-2-Régression par machines à vecteurs de support [37,38]

La régression par machines à vecteurs de support (SVR) consiste à trouver la fonction $f(x)$ qui a au plus une déviation ε par rapport aux exemples d'apprentissage (x_i, y_i) , pour $i=1, \dots, N$, et qui est la plus plate possible. Cela revient à ne pas considérer les erreurs inférieures à ε et à interdire celles supérieures à ε [39]. Maximiser la platitude de la fonction permet de minimiser la complexité du modèle qui influe sur ses performances en généralisation. En effet, la théorie de l'apprentissage [37] permet de borner l'erreur de généralisation par une somme de deux termes : l'un dépendant de la complexité du modèle et l'autre dépendant de l'erreur sur les données d'apprentissage [40].

Les méthodes SVM sont basées sur le contrôle de la complexité du modèle lors de l'apprentissage.

Dans la méthode SVM, différents hyperparamètres apparaissent : C , qui représente le compromis entre la complexité du modèle et l'erreur sur les données d'apprentissage ; ε , qui correspond à la largeur du tube d'insensibilité ; les éventuels paramètres de la fonction noyau k (σ, γ, \dots). Ces hyperparamètres sont en général réglés en fonction et d'une estimation de l'erreur de généralisation qui peut être évaluée sur un jeu indépendant de données de validation croisée. Cela implique de réaliser l'apprentissage pour différentes valeurs et estimer leurs performances. Dans le cas d'une estimation de l'erreur de généralisation par validation croisée, cette procédure peut se révéler très coûteuse en temps de calcul.

Une idée permettant d'éviter les temps de calcul trop importants dans la recherche des hyperparamètres optimaux est donnée dans [39].

II-2- DEVELOPPEMENT ET EVALUATION DE MODELE.

II-2-1 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 (GA) que nous avons appliqués dans le présent travail, et que nous rappelons succinctement.

II-2-2 Principe.

Dans la terminologie des algorithmes génétiques, le vecteur binaire \tilde{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.

II-2-3 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) ;

II-2-4 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).

II-2-5 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.

II-2-6 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 [41], 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.

II-3 Développement des modèles

Les techniques les plus courantes pour établir des modèles QSAR utilisent l'analyse de régression (régression linéaire multiple : RLM ; projection des structures latentes par les moindres carrés partiels : PLS), les réseaux neuronaux (RNA), et machine à support de vecteur (SVM).

II-3-1 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) :

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

II-3-2 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) [42]. 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 Residual Sum of Squares) :

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 \quad (54)$$

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

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

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 $Q_{\text{LoO}}^2 > 0,5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [43].

II-3-3 Domaine d'application.

Le domaine d'application a été discuté à l'aide du diagramme de Williams (traité en détail dans [44 ,45], 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)

$$h_i = x_i (X^T X)^{-1} x_i^T \quad (56)$$

Où (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 $h_{ii} < h^*$, 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 $h_{ii} > h^*$ 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.

II-3-4 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) [46].

II-3-5 Validation externe.

En plus du test de randomisation, il est intéressant [47], pour juger de la qualité du modèle, de considérer le coefficient de prédiction externe calculé comme suit :

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

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 EQMP_{ext}).

Ces valeurs RMSE sont mieux adaptées, pour juger de la qualité d'un modèle que les valeurs de R^2 et Q^2 seules, qui constituent de bons tests uniquement pour des données réparties régulièrement.

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

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



REFERENCES BIBLIOGRAPHIQUES

- [1] Hohenberg , P., Kohn,W. (1964)- Inhomogeneous Electron Gas, j,Physical Review; vol. 136 No. 3B, pp.B 864 –B871.
- [2] Allinger, N. L. (1976)- Calculation of Molecular Structure and Energy by Force-Field Methods, j Advances in Physical Organic Chemistry, Vol. 13,pp. 1-82.
- [3] Niketic S. R., Rasmussen, K. (1977)- The Consistent Force Field: A Documentation, Springer, Berlin.
- [4] Burbert U., Allinger, N. L. (1982)- Molecular Mecanics, American Chemical Society.,Washington.
- [5] Lomas, J. S. (1986)- La mécanique moléculaire, une méthode non quantique pour le calcul de la structure et de l'énergie d'entités moléculaire, j L'actualité chimique, Vol. 22 No. 3,pp. 7 – 20.
- [6] Kolos, W., Wolniewicz, L. (1964)- Accurate adiabatic treatment of the ground state of the hydrogen molecule, j of chemical physics., Vol. 41 No. 12,pp. 3663.
- [7] Sutcliffe , B. T. (1997)- The Nuclear Motion Problem in Molecular Physics, j Advances in Quantum Chemistry, vol. 28 ,pp. 65.
- [8] Roothan, C.C.J. (1951)- New Developments in Molecular Orbital, j Theory,Reviews Of Modern Physics, vol. 23 No. 2,pp.69.
- [9] Hall, G. G. (1951)- The molecular orbital theory of chemical valency. VIII A method of calculating ionization potentials, j Proceedings of the Royal Society of London A, vol 205 No 1083, p.p.541
-
- [10] Koopmans, T. A. (1933)- The distribution of wave function and characteristic value among the individual electrons of an atom, j Physica, vol. 1 , pp.104-113.
- [11] Blinder S. M. (1965)- Basic Concepts of Self-Consistent-Field Theory, j American Journal of Physics,vol.33,pp.431.
- [12] Löwdin, P. O. (1950)- «On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals, J of Chemical Physics, vol. 18,pp.365.
- [13] Löwdin P. O., (1970)- On the Orthogonality problem. J Advances in Quantum Chemistry, Vol 5, pp 185-199
- [14] Mulliken (1955)- Electronic Population Analysis on LCAO-MO Molecular Wave Functions.(I). J Chemical Physics, Vol 23, pp 1833- 1840
- [15] Mulliken (1955)- Electronic Population Analysis on LCAO-MO Molecular Wave Functions.(II). Populations, Bond Orders, and Covalent Bond Energies. J Chemical

Physics., Vol 23, pp 1841- 1846.

[16] Dewar, M. J. S. and Thiel, W. (1977)- Ground States of Molecules.38. The MNDO Method. Approximations and Parameters , J of the American Chemical Society , vol .99 No 15, pp.4899-4907.

[17] Dewar, M. J .S., Zoebisch, E. G., Healy, E. F., Stewart ,J. J. P. (1985)- The development and use of quantum mechanical molecular models. 76. AMI: a new general purpose quantum mechanical molecular model, J of the American Chemical Society, Vol .107, pp.3902-3909.

[18] Stewart J. J. P. (1989)- Optimization of parameters for semiempirical methods I. Method", J of Computational Chemistry, Vol. 10 No. 2, pp. 209–220.

[19] Dewar, M. J. S. and Thiel, W. (1977)- A Semiempirical Model for the Two-Center Repulsion Integrals in the NDDO Formalism, J Theoretica chimica acta, vol .46,pp. 89-104.

[20] Allinger N.L., (1977) - J American Chemical Society., Vol 99, pp 8127-8134

[21] Burkert U., Allinger N.L., (1982, 1986) - Molecular Mechanics, ACS Monograph No177, American Chemical Society, Washington, DC.,

[22] Allinger N. l., Yuh, Y.H, and Li J-H., (1989)- J of the American Chemical Society. Vol 111, pp 8551-8565.

[23] Allinger N.l., Chem k., Lii J-H., J. (1996) -Computational Chemistry. Vol 17, pp 642-668

[24]. Mc kerell A.D, Jr., Bashford D., Bellott M., Dunbrack R. L., Jr., Evanseck J.D., Field M.J., Fischer S., Gao J., Guo H., Ha S., Joseph D. Carthy Mc, Kuchnir L., Ruczera, K. . Lau F.T.K, Mattos C., Michnick, S. Ngo T., Nguyen D.T., Prodhom B., Reiher W.E. Roux B ., Schlenkrich D. Smith M., Stote J.C., Stramb R., Watanabe J., Wiokiewicz- Kuczera, M., Yin J, and Karplus M., (1998) - J Chemical Physics, Vol 102, pp 3586-3616.

[25] Halgren T.A., (1996)- J Computational Chemistry Vol 17, pp 490-519.

[26] Halgren T.A, (1996)- J Computational Chemistry Vol 17, pp 520-552, 553-586, 6116-6641.

[27] Halgren T.A., Nacbar R.B., (1996)- J Computational Chemistry. Vol 17, pp 587-615.

[28] Ramachandran K.I , Deepa, G , Namboori, K. Computational Chemistry and Molecular Modeling. Principles and Applications 2008. DOI 10. 1007/978-3-540-77304-7.

[29] Dodge Y., Rousson. (2004)- Analyse de regression appliqué Dunold, Paris.

[30]Montgomery,D.C , Peck, E.A. (1992)-Introduction to Linear Regression Analysis. Snd Edition. John Wiley & sons. Inc.

[31] Draper N.R., Smith H., (1998)- Applied Regression Analysis. Third Edition. John Wiley& Sons, Inc..

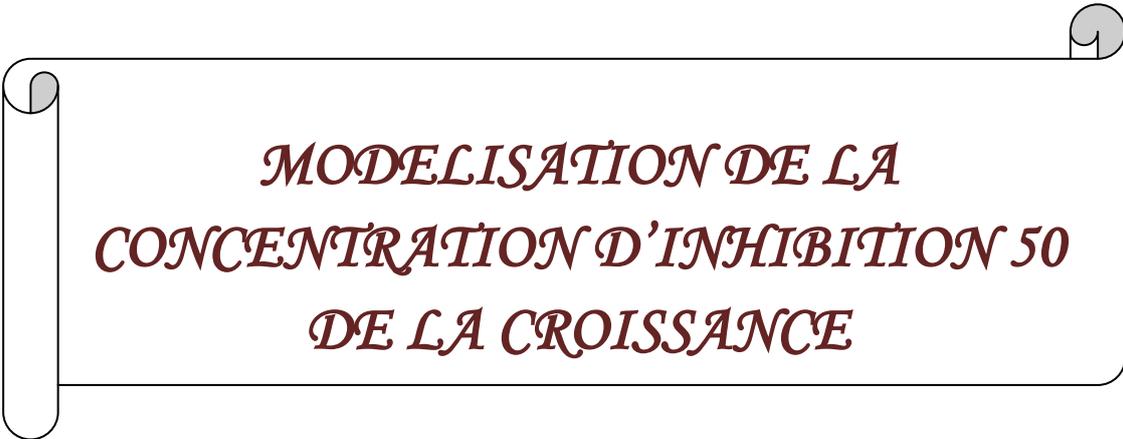
- [32] Jodouin J.F., (1994)- Les réseaux de neurones, Editions Hermes.
- [33] Davalo E & Naim, P. (1998)- les réseaux de neurones, Edition Eyrolle.
- [34] Zupan, J. Gasteiger J. (1999)- Neural Networks in Chemistry and Drug Desgn. Second Edition. Wiley VCH, New York.
- [35] Badran F, Thria S., Hérault L. (2004)- Réseaux de neurons: Méthodologie et applications, Editions EYROLLES.
- [36] Hecht-Nilson R., (1990)- Neurocomputing. Addition-Wesly Publishing Company.
- [37] MATLAB. (2004)-Version 7.0.0. 1992 0 (Release 14). The Language of Technical Computing. The Math Works, Inc. May 6.
- [38] Vapnik Y.N, (2000)- The nature of statistical learning theory. Springer-Verlag, New York, USA.
- [39] Nianyi Chem, Wencong Lu, Jie Yang, Guozheng hi. (2004)- Support Vector Machine in Chemistry. World Scientific, New Jersey.
- [40] Smola, A.J. Schölkopf, B. (2004)- A tutorial on support vector regression. Statistics and Computing, 14(3), pp 199-222.
- [41] Cristianini N., Shawe-Taylor J. (2000)- An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.
- [42] Kubinyi H, (1994)- Quantitative Structure-activity Relationships., Vol 13 , pp 285.
- [43] Draper N.R, and Smith H., (1998)- Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New York.
- [44] Eriksson L., Jaworska J, and Worth A.P., (2003)- Perspective M.T.D., Vol 111(10), pp 1361-1375.
- [45] Erikson L., Jaworska J., Worth A. P., Cronin M. T. D., Mc Dowell R. M, and Gramatica P., (2003) - Methods for Reliability and uncertainty Assessment and for Applicability Evaluations of Classification-and Regression - Based QSARs. J Environmental Health Perspectives. Vol 111 (10). pp 1361-1375.
- [46] Tropsha A, Gramatica P, and Grombar V.K. (2003)- The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR & J Combinatorial Science., Vol 22. pp 69-76.
- [47] Sharma B.K, Singh P, Pilania P, Sarbhai K, Yenamandra S, and Prabhakar. CP-(2011) - MLR/PLS directed QSAR study on apical sodium-codependent bile acid transporter inhibition activity of benzothiepinines J Mol Divers Vol 15. pp 135–147.
- [48] Todeschini R., Consonni V., Mannhold R., Kubinyi H, and Timmerman H., eds., Wiley (2000)- Handbook of Molecular Descriptors, VCH, Verlag Gmbh, Weinheim

PARTIE III

RESULTATS ET DISCUSSION

*I- Modélisation de la concentration
d'inhibition 50 de la croissance*

*II- Modélisation de la concentration
 létale 50*



*MODELISATION DE LA
CONCENTRATION D'INHIBITION 50
DE LA CROISSANCE*

I- Modélisation

I-1 INTRODUCTION.

L'impact du danger potentiel des produits chimiques non testés, qui est un défi auquel sont confrontés les organismes de réglementation nationaux et internationaux [1-4], peut être mesuré par des études expérimentales, mais cette approche peut s'avérer fort coûteuse en temps et argent [5]. Une alternative est de s'appuyer sur des modèles QSAR (Quantitative Structure/ Activity Relationship) qui décrivent une relation mathématique entre les caractéristiques structurales d'un ensemble de produits et l'activité particulière qui leur est associée [6,7].

Plusieurs modèles QSAR concernant la prédiction de la toxicité chimique aiguë pour l'environnement aquatique ont été publiés [8-12]. Ils sont principalement basés sur le coefficient de partage n-octanol / eau. Quoique le nombre de composés pour lesquels on dispose de la valeur mesurée de cette caractéristique soit estimé à 30 000 [13], ce qui semble élevé à première vue, ce nombre est en fait négligeable par rapport au nombre croissant de composés pour lesquels les valeurs du coefficient de partage n-octanol/ eau sont souhaitées mais font défaut.

En outre, la détermination expérimentale de cette caractéristique est fastidieuse, prend du temps et exige une grande pureté du soluté [14] ; aucune de ces conditions préalables n'étant compatible avec des techniques à haut débit il y a donc un intérêt permanent pour les méthodes de prédiction des valeurs du coefficient de partage n-octanol / eau.

Au cours des décennies récentes diverses approches (méthodes fragmentales, à base d'atomes, conformations dépendantes) [15-18] ont été développées, mises en application pour la plupart et disponibles comme programmes informatiques. Cependant, même dans ces calculs il n'est pas rare d'avoir des différences de plusieurs ordres de grandeur [19-20].

Pour ces raisons, le coefficient de partage n-octanol/ eau ne peut être considéré comme un descripteur non équivoque, ce qui a amené différents auteurs [19-24] à proposer des modèles de toxicité basés exclusivement sur d'autres descripteurs moléculaires structurels théoriques.

Dans cette partie nous proposerons des modèles QSAR prédictifs, obtenus par régression linéaire simple, pour évaluer la toxicité relative de composés chimiques organiques en termes (du logarithme de l'inverse) de la concentration d'inhibition 50% de la croissance : CIC50 (50% Inhibitory Growth Concentration) de *Tetrahymena pyriformis*. Des modèles basés sur différents types de coefficients de partage n-octanol/ eau seront comparés au modèle optimal construit en utilisant un seul descripteur (géométrique) 3D calculé à partir de la structure chimique.

I-2 COEFFICIENT DE PARTAGE

I-2-1 Propriétés de partage [25].

Le partage d'une molécule entre une phase aqueuse et une phase lipidique conditionne en partie ses propriétés biologiques telles que le transport, le passage à travers les membranes, la biodisponibilité (distribution et accumulation), l'affinité pour un récepteur et la fixation par une protéine, l'activité pharmacologique ou encore la toxicité. S'agissant de contaminants, ce même partage conditionne leur devenir dans notre environnement en particulier leur accumulation dans les organismes aquatiques.

Depuis les travaux de Collander à la fin des années 1950 [26], puis ceux du groupe de Hansch quelques années plus tard [27,28], le coefficient de partage d'une molécule dans un système biphasique constitué de deux solvants non miscibles (le plus souvent le système n-octanol/eau), est reconnu pour sa faculté à mimer le passage de cette molécule à travers les membranes biologiques. Pour des solutions diluées, ce coefficient de partage n-octanol/ eau est le rapport de la concentration d'une molécule de soluté dans le n-octanol sur sa concentration dans l'eau lorsque le système biphasique est en équilibre. En fait, comme les valeurs mesurées s'étendent sur au moins douze unités de grandeur (10^{-4} - 10^8), on utilise de préférence les logarithmes décimaux. Le partage est donc une propriété physico-chimique importante qui peut être utilisée pour représenter la nature lipophile ou hydrophile d'une molécule.

* Remarque : dans notre mémoire on parlera de lipophilie ou d'hydrophilie (Grand Larousse Universel). Certains auteurs utilisent les termes lipophilité et hydrophilité.

Selon les auteurs, on parle de coefficient ou de constante de partage ou de distribution ou encore de rapport de distribution. Outre P , plusieurs autres symboles sont utilisés tels que K_D , K_p , K_{OW} ou encore D . On parle également de P' ou de D' , le coefficient de partage apparent, signifiant qu'il n'est valable que sous certaines conditions expérimentales utilisées lors de sa détermination. En effet il arrive que le soluté soit présent dans chaque phase sous plusieurs formes dont une seule intervient dans le partage. C'est le cas par exemple pour les acides faibles AH . Dans la phase organique l'acide faible existe sous sa forme non-dissociée AH alors que dans l'eau, il se trouve à la fois sous forme AH et sous forme ionisée A^- . Le coefficient de partage apparent dépend alors du pH . La connaissance du pK_a permet le calcul du coefficient du partage vrai.

Le $\log P$ constitue un paramètre unique qui regroupe plusieurs effets : tous les types d'interactions non covalentes, la solvatation ainsi qu'une composante d'entropie. Il est très largement utilisé dans des études de relations quantitatives structure – activité (QSARs) dans les sciences pharmaceutiques, biochimiques, toxicologiques et dans les sciences de l'environnement. La lipophilie intéresse donc tout autant la communauté qui étudie les problèmes de santé humaine que celle qui est impliquée dans les problèmes de l'environnement.

I-2-2 Détermination expérimentale des $\log P$.

Chaque fois que cela sera possible, le $\log P$ d'une molécule fera l'objet d'une détermination expérimentale. La méthode dite des flacons agités ou « shake – flask » [29], est la plus ancienne. La molécule étudiée, interagit avec les deux phases en équilibre d'une manière qui mime la façon dont par exemple, un ligand se fixe sur le site actif d'un récepteur. Malgré ses imperfections (problèmes d'adsorption par le verre, formations d'émulsion lors de l'agitation, domaine de mesure étroit de l'ordre de -3 à +4, durée de l'ordre de 30 minutes par échantillon.....), elle reste cependant la méthode de choix pour des molécules organiques originales (méthode la plus précise, pour la gamme la plus large de solutés neutres comme chargés; la structure chimique n'a pas à être connue avant de commencer la procédure) et de ce fait, elle est préconisée comme procédure standard de caractérisation par l'OCDE [30].

Toutefois, elle tend à être supplantée par les méthodes chromatographiques. En particulier la chromatographie liquide haute performance à polarité inversée (CLHP-PI) [31], adaptée aux études de criblage, est elle aussi préconisée par l'OCDE [30]. Dans ce cas on

utilise comme indice de lipophilie, une valeur déduite de la mesure des temps de rétention, $\log k_w$.

I-2-3 Méthodes d'estimation de $\log P$.

La plupart des méthodes expérimentales de détermination de $\log P$ souffrent du même inconvénient, à savoir que leur domaine d'application est relativement étroit. D'autre part, du fait de la nature intrinsèque de certaines molécules, leurs $\log P$ sont inaccessibles à l'expérience. C'est le cas en particulier des surfactants qui ont tendance à s'accumuler à l'interface du système biphasique au lieu de se disperser dans les deux phases. Enfin, dans le domaine de conception assistée par ordinateur ou dans le domaine de la chimie combinatoire, les chercheurs travaillent sur des modèles moléculaires avant même que les molécules aient été synthétisées ceci explique le succès des nombreuses méthodes d'estimation de $\log P$ qui ont été décrites dans la littérature depuis plus de quarante ans.

La première méthode de calcul de $\log P$ a été développée par Hansch et Fujita (système π) [32]. Les imperfections relevées ont conduit Rekker à développer la première approche de contribution fragmentale [33-36]. Puisque la définition de fragment peut être ambiguë, Broto et al. [37], suivis par d'autres, ont développé des systèmes de calculs basés sur des contributions atomiques. Toutes les méthodes qui divisent les molécules en sous-structures sont appelées approches par sous-structure. L'utilisation exclusive des sous-structures atomiques caractérisent des méthodes de contribution d'atomes, tandis que l'utilisation additionnelle de plus grands groupes est typique des méthodes fragmentales. L'addition des contributions des sous-structures conduit finalement à la valeur de $\log P$.

En revanche, les approches basées sur la molécule entière emploient les « potentiels de lipophilie moléculaire », les indices topologiques, ou les propriétés moléculaires pour quantifier $\log P$.

Le « potentiel de lipophilie moléculaire » MLP (sigle anglo-saxon pour « Molecular Lipophilicity Potential » traduit l'influence des contributions fragmentales de lipophilie d'une molécule sur son environnement. Cette grandeur dépend de la distance à laquelle elle est calculée :

$$\text{MLP} = \sum_i f_i \cdot g(d_i) \quad (1)$$

Où f_i est la constante lipophile du fragment i telle qu'elle a été définie dans le calcul de $\log P$ et $g(d_i)$ est une fonction de la distance d_i entre le fragment i et un point de l'espace environnant de la molécule.

Ce « potentiel » peut être considéré comme une extension naturelle du « moment dipolaire hydrophobe » μ introduit par Eisenberg & McLachlan par analogie avec le moment dipolaire électrostatique :

$$\mu = \sum_i f_i \vec{r}_i \quad (2)$$

Où \vec{r}_i est le vecteur joignant l'origine à un atome du fragment i .

Plusieurs méthodes de calcul qui diffèrent soit par la base de constantes fragmentales, soit par l'expression de la fonction $g(d_i)$, sont actuellement utilisées.

I-2-3-1 Clog P (calculated logP) [38].

La version du logiciel du système développé par Hansch et Leo [39], utilise le système additif de Rekker [40], et est connue sous l'appellation ClogP (ou logP calculé).

Dans ce système la lipophilie d'un composé est estimée par la somme de la lipophilie de ses fragments et de termes de corrections :

$$C \log P = \sum_{i=1}^n a_i * f_i + \sum_{j=1}^m b_j * \vec{r}_j \quad (3)$$

Où : f_i = constante de lipophilie du fragment f_i

a_i = nombre de fragments f_i

F_j = facteur de correction F_j

b_j = nombre de facteurs de correction F_j

Les constantes fragmentales ont été tirées de solutés où le fragment intervient isolément. De plus les liaisons environnantes, du fragment considéré, sont prises en compte (alkyle, benzyle, vinyle, styryle, et voisins aromatiques) conduisant à cinq valeurs par fragment. Si un fragment, en combinaison avec les liaisons environnantes est absent, mais au moins deux

valeurs pour le même fragment avec différents voisins ont pu être trouvées, on tentera une interpolation pour obtenir les données manquantes.

Les facteurs de correction ont été calculés à partir des corrections nécessaires pour les interactions spécifiques modélisées. Par exemple, l'interaction de deux groupes hydroxyles dans le diéthylène glycol augmente la valeurs logP de 0,85 par rapport à deux groupes hydroxyles qui n'interagissent pas. Cette valeur est alors considérée comme le terme de correction pour deux groupes hydroxyles voisins [41].

La décomposition de la structure moléculaire en fragments (atomes ou groupes polyatomiques) est effectuée en utilisant ainsi une solution unique.

I-2-3-2 Approche de Ghose et Crippen basée sur un système d'incrément atomiques purs [42 ,43].

Le système d'incrément atomique, AlogP, développé par Ghose et Crippen est le plus fréquemment utilisé des méthodes de contribution d'atomes. Celles-ci simplifient à la fois la reconnaissance des fragments et les calculs pour les corrections de stucture ne sont pas appliqués (voir éq 4). Deux problèmes restent cependant posés : le premier concerne le grand nombre d'atomes nécessaires pour décrire un ensemble de molécules, le second a trait aux structures isomères ces méthodes ne prenant pas en charge la flexibilité de conformation.

Pour chaque composé, la valeur de logP est estimée par :

$$A \log P = \sum_i n_i a_i \quad (4)$$

Où n_i est l'occurrence du type d'atome i et a_i la constante d'hydrophobie correspondante.

Le modèle AlogP mis en œuvre dans le logiciel DRAGON [44] a été évalué sur un ensemble de 2648 composés dont les valeurs expérimentales de logP ont été extraites de la base de données ouverte NCI. Le coefficient de corrélation obtenu est $r=0,915$.

I-2-3-3 Modèle de Moriguchi basé sur des paramètres structuraux : MlogP

Les approches de « molécule entière » utilisent des descriptions de la molécule entière pour calculer logP. Ces modèles essaient d'éviter les imperfections des approches fragmentales, telles que la simplification des effets stériques, et l'impossibilité de calculer les logP des structures pour lesquelles les valeurs des fragments n'existent pas, ou le problème soulevé par le cas des isomères.

Il s'agit d'un modèle décrit par une équation de régression sur la base de treize paramètres [45, 46] y compris la somme des hydrophobies atomiques, les effets de proximité, les liaisons insaturées, les propriétés amphotères et les fonctions spéciales telles que la présence d'un azote quaternaire, le nombre de groupes nitro, ou un simulacre pour la présence de la p-lactame.

Les coefficients de régression ont été évalués sur un ensemble de 1230 molécules organiques comprenant , les composés aliphatiques, aromatiques et hétérocycliques contenant les atomes suivants : C, H, N, O, S, P, F, Cl, Br, I [44]. Les paramètres statistiques du modèle sont $r=0,952$; $SE=0,422$; $F_0(13 ; 1216) = 900,4$.

I-3 COLLECTE DES DONNEES EXPERIMENTALES.

I-3-1 Le protozoaire *Tetrahymena pyriformis*.

Les protozoaires sont des organismes unicellulaires microscopiques qui, à cause de leur membrane nue, ne se retrouvent que dans les habitats humides ou aquatiques comme les océans, les lacs ou le sol. Chaque protozoaire est une cellule très spécialisée capable de remplir toutes les fonction vitales.

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 [47].

Tetrahymena pyriformis, est un protozoaire cilié omniprésent d'eau douce de taille variable, mais dont on considère qu'il mesure, couramment, 50 μm de long sur 30 μm de large. Les chercheurs en biochimie utilisent *Tetrahymena pyriformis* comme organisme modèle, comme agent de transformation d'un polluant organique, et comme biotest permettant d'évaluer la toxicité de molécules chimiques et d'échantillons naturels.

I-3-2 Test de toxicité: méthode et matériels.

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 50% de la croissance (CIC50), dont le logarithme de l'inverse, soit $pCIC50 = \log (CIC50)^{-1}$, servira d'indicateur de toxicité.

Les essais ont été menés, à $27 \pm 1^{\circ}C$, dans des erlenmeyers de 250ml contenant 50 ml d'un milieu dont la composition est la suivante :

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

Ce milieu est inoculé avec 0,25mL d'une culture contenant approximativement 36000 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.

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.

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% (350ml par fiole), quantité qui n'altère pas la reproduction de *Tetrahymena pyriformis* [48, 49].

I-3-3 Mécanismes de l'action toxique.

Il est évident que la séparation des composés chimiques industriels en groupes toxicologiques significatifs ne doit pas reproduire les catégories de la chimie organique. Toutefois les effets toxiques des composés organiques industriels doivent relever de trois catégories de cause et d'effet générales. Celles-ci comprennent :

*1-Les toxicités non spécifiques ou physiquement réversibles dues à une perturbation membranaire.

*2-Les toxicités non spécifiques non réversibles ou toxicités chimiques dues à l'action directe du toxique sur un groupe fonctionnel ou une macromolécule, et

*3- Les réponses toxiques spécifiques.

La première catégorie comprend les composés chimiques qui sont à l'origine d'effets physiologiques réversibles, effets physiques par nature et par conséquent généralement indépendants de la structure moléculaire. C'est la principale catégorie de toxicité pour les composés organiques industriels. Cette catégorie consiste en une série de mécanismes de l'action toxique, ayant en commun les faits que la réponse relative à l'équilibre soit reliée à la concentration externe et qu'elle soit fonction du rapport hydrophobie/ hydrophilie.

Le niveau de l'étape limitative pour ces mécanismes d'action est lié à la capacité du composé toxique d'atteindre le site d'action. Par conséquent chaque mode d'action dans cette catégorie peut être estimé par un modèle QSAR à un seul descripteur log Kow.

I-3-4 Narcose apolaire.

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. La narcose apolaire est le mode d'action toxique des composés chimiques à la fois non ioniques et non réactifs.

Les alcools aliphatiques sont, classiquement, les narcotiques apolaires industriels les plus étudiés.

Dans cette étude deux toxiques ont été étudiés : un ensemble de 21 alcools à chaînes linéaires et ramifiées et 9 amines aliphatiques normales, choisis pour leurs diversités dans les longueurs et les ramifications des chaînes.

L'ensemble des données expérimentales a été prélevé dans [50].

I-4 RESULTATS ET DISCUSSION.

Le meilleur modèle non $-\log P$ à une dimension a été obtenu avec l'indice du rapport moyen des distances, noté *ADDD* (pour average distance degree). Il renseigne sur les repliements moléculaires [51, 52], liés à la facilité de diffusion au travers des barrières biologiques telles que les membranes.

Les matrices distance, distance, notées D/D , sont définies par le rapport entre les matrices des distances géométriques, r_{ij} , et topologiques d_{ij} :

$$[G/D]_{ij} = \begin{cases} \frac{r_{ij}}{d_{ij}} & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases} \quad (5)$$

Les sommes des éléments des lignes de ces matrices renferment des informations sur le repliement moléculaire ; en effet, dans des structures fortement repliées, elles ont tendance à être relativement petites du fait que les distances inter-atomiques sont petites, alors que les distances topologiques augmentent à mesure que la taille de la structure augmente.

Par conséquent la somme moyenne des éléments d'une ligne est un invariant moléculaire appelé rapport moyen de distance -à- distance, défini par :

$$ADDD = \frac{1}{A} \sum_{i=1}^A \sum_{j=1}^A \frac{r_{ij}}{d_{ij}} \quad i \neq j \quad (6)$$

A étant le nombre d'atomes de la molécule.

Dans le tableau I nous avons réuni le numéro d'enregistrement du service des composés chimiques (numéro de CAS), ainsi que les valeurs $-\log \text{CIC50}$, AlogP , MlogP , ClogP et ADDD des alcools et des amines aliphatiques sélectionnés.

Tableau I - Valeurs des toxicités relatives et des descripteurs moléculaires des alcools et amines aliphatiques sélectionnés

Composés	Numéro CAS	$-\log \text{CIC50}$	AlogP	MlogP	ClogP	ADDD
Méthanol	67-56-1	-2,77	-0,358	-0,814	-0,764	4,964
Ethanol	64-17-5	-2,41	0,009	-0,172	-0,235	8,108
1-propanol	71-23-8	-1,84	0,515	0,347	0,294	11,098
1-pentanol	71-41-0	-1,12	1,427	1,209	1,352	17,194
1-hexanol	111-27-3	-0,47	1,883	1,587	1,881	20,305
1-heptanol	111-70-6	0,02	2,339	1,940	2,410	23,451
1-nonanol	143-08-8	0,77	3,252	2,591	3,468	29,859
1-décanol	112-30-1	1,1	3,708	2,894	3,997	33,118
1-dodecanol	112-53-8	2,07	4,620	3,467	5,055	39,716
1-tridecanol	112-70-9	2,28	5,077	3,739	5,584	43,053
2-propanol	67-63-0	-1,99	0,368	0,347	0,074	10,987
2-méthyl-1-butanol	137-32-6	-1,13	1,290	1,209	1,222	16,810
3-méthyl-1-butanol	123-51-6	-1,13	1,223	1,209	1,222	16,833
3-méthyl-2-butanol	598-75-4	-1,08	1,211	1,209	1,002	16,622
(tert)pentanol	75-85-4	-1,27	1,097	1,209	1,002	16,640

Composés	Numéro CAS	-logCIC50	AlogP	MlogP	ClogP	ADDD
1-propylamine	107-10-8	-0,85	0,225	0,347	0,394	11,968
1-hexylamine	11-26-2	-0,34	1,594	1,587	1,981	21,226
1-heptylamine	111-68-2	0,1	2,050	1,940	2,510	24,393
1-octylamine	111-86-4	0,51	2,506	2,274	3,039	27,602
1-undecylamine	7307-55-3	2,26	3,875	3,186	4,626	37,408
1-butanol*	71-36-3	-1,52	0,971	0,800	0,823	14,138
1-octanol*	111-87-5	0,5	2,796	2,274	2,939	26,640
1-undecanol*	112-42-5	1,87	4,164	3,186	4,526	36,401
2-pentanol*	6032-29-7	-1,25	1,348	1,209	1,132	16,888
3-pentanol*	584-02-1	-1,33	1,416	1,209	1,132	16,908
(neo) pentanol*	75-84-3	-0,96	1,108	1,209	1,092	16,710
1-butylamine*	109-73-9	-0,7	0,681	0,800	0,923	15,015
1-amylamine*	110-58-7	-0,61	1,137	1,209	1,452	18,101
1-nonylamine*	112-20-9	1,59	2,962	2,591	3,568	30,837
1-decylamine*	2016-57-0	1,95	3,418	2,894	4,097	34,110

(*) Élément de l'ensemble de validation.

Les ordonnées à l'origine (β_0) et les pentes (β_1) de chacun des modèles simples (unidimensionnel) calculés sont présentées dans le tableau II

Tableau II Coefficients des modèles calculés par les moindres carrés ordinaires.

X	AlogP	MlogP	ClogP	ADDD
β_0	-2,144 ($\pm 0,139$)	-2,230 ($\pm 0,134$)	-1,998 ($\pm 0,088$)	-3,333 ($\pm 0,134$)
β_1	0,939 ($\pm 0,057$)	1,191 ($\pm 0,068$)	0,815 ($\pm 0,033$)	0,138 ($\pm 0,006$)

Les paramètres statistiques rapportés ci après (tableau III) montrent clairement la différence dans l'ajustement et les performances de prédiction pour les descripteurs logP sélectionnés, ClogP fournissant incontestablement les meilleurs résultats.

Un autre fait remarquable observé est que ClogP et le descripteur moléculaire théorique ADDD peuvent être échangés sans variations pertinentes dans les résultats statistiques.

Tableau III Résumé des statistiques obtenues pour les modèles unidimensionnels calculés.

X	R ²	Q ²	Q ² _{L(5)O}	Q ² _{boot}	Q ² _{ext}	SDEC	SDEP	SDEP _{ext}	F _(p=0.000)	SE
Alog P	0,9372	0,9188	0,9165	0,9063	0,8584	0,365	0,415	0,554	268,4	0,385
Mlog P	0,9444	0,9291	0,9270	0,9160	0,8998	0,344	0,388	0,461	305,52	0,362
Clog P	0,9713	0,9630	0,9621	0,9580	0,9352	0,247	0,280	0,371	610,02	0,260
ADDD	0,9711	0,9624	0,9616	0,9566	0,9326	0,248	0,283	0,378	603,82	0,261

Carbo- Dorca et al. [53] ont rapporté une étude QSAR où les mêmes données ont été examinées. Ces auteurs ont construit un modèle prédictif en utilisant un descripteur moléculaire quantique associé à l'énergie moyenne de répulsion électronique. Les paramètres trouvés ($R^2 = 0,9240$; $Q^2 = 0,9090$, et $SE = 0,415$) sont inférieurs à ceux de la présente approche.

La valeur de R^2 (tableau III) montre le bon ajustement du modèle. En général, plus la valeur de F est grande plus le modèle prédit le mieux les valeurs des propriétés des molécules testées. La grande valeur de $F = 603,82$ indique que le modèle fait un excellent travail de prédiction des valeurs de pCIC50. Le modèle est robuste, la différence entre R^2 et Q^2 est faible ($< 1\%$). La figure (1) reproduit, pour l'ensemble de calibrage, les valeurs observées de pCIC50 en fonction de celles obtenues par validation croisée.

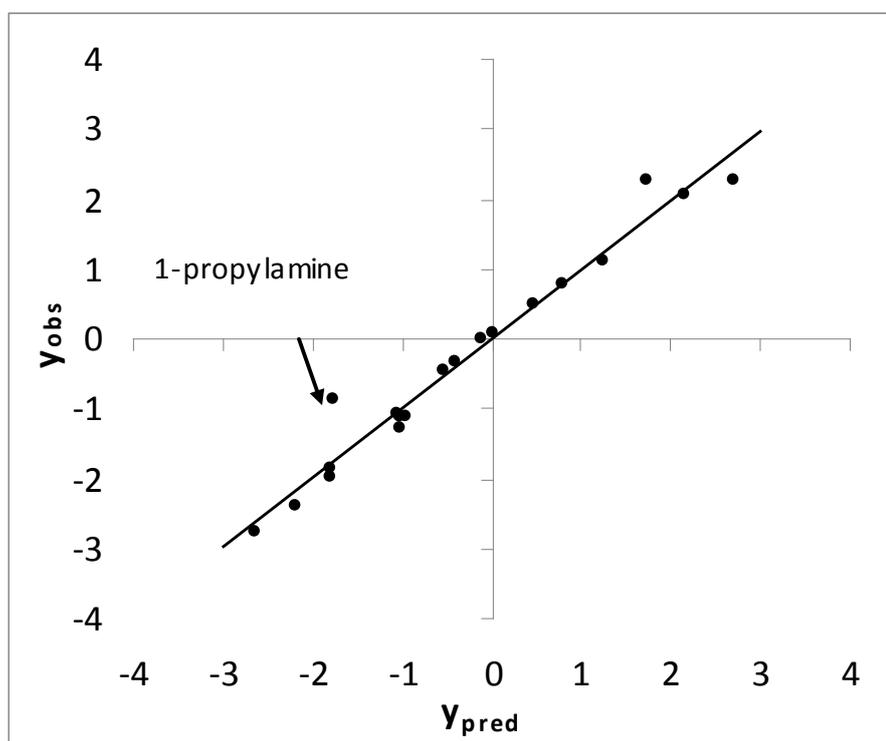


Figure.1 : Activités observées en fonction de celles obtenues par validation croisée pour l'ensemble de calibrage.

La dispersion des points est faible, bien qu'il y ait un point un peu éloigné du reste (1-Propylamine). Ce composé chimique dont les résidus standardisés (e_{std}) sont supérieurs à 3 unités d'écart-type (en valeur absolue) est un point aberrant sérieux dans tous les modèles envisagés. La similarité de EQMP et EQMC indique que le modèle a une capacité de prédiction interne pas trop dissemblable de son pouvoir d'ajustement.

	f(Alogp)	f(Mlogp)	f(Clogp)	f(ADDD)
e_{std}	-3,35	-3,13	-3,68	-3,7

Le modèle démontre une très bonne stabilité dans la validation interne (la différence entre Q^2 , et $Q^2_{L(5)O}$ atteint 0,33%), alors que la technique du bootstrap confirme la capacité de prédiction interne et la stabilité du modèle.

Quoique de petite taille, l'ensemble des données a été soumis à la validation statistique en le séparant au préalable, de façon aléatoire, en deux sous-ensembles disjoints : l'un de calibrage (20 composés chimiques) pour le calcul du modèle et l'autre de test (10 composés chimiques) pour sa validation externe. La petite taille de l'ensemble des données expérimentales publiées [50] ne permettait pas une séparation plus radicale. L'information fournie par Q^2_{ext} est quelque peu optimiste. En effet, avec les petits ensembles de données (20-30) une prédiction externe complète de nouveaux composés chimiques est seulement réalisable a posteriori, cas par cas.

I-5 CONCLUSION.

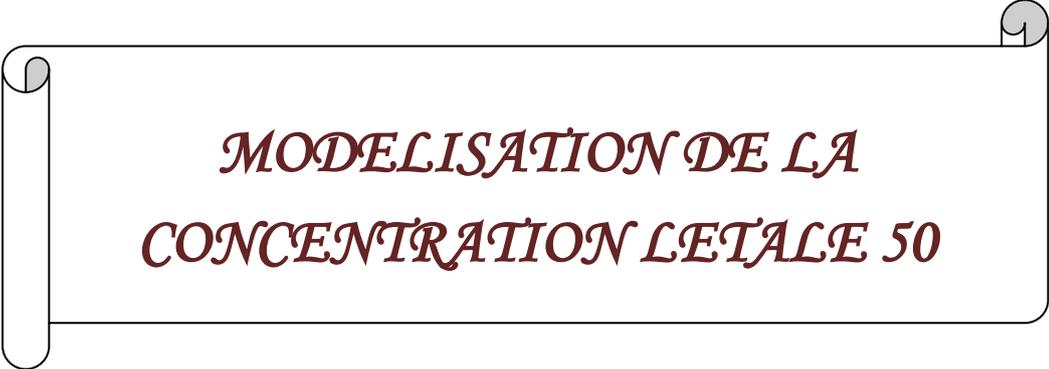
Des résultats et discussion qui ont précédé, nous concluons que :

1-Parmi les descripteurs logP choisis pour modéliser l'inhibition de la croissance de *Tetrahymena pyriformis* par les alcools et les amines aliphatiques, ClogP est le meilleur.

2-Le descripteur géométrique ADDD (sigle anglo-saxon pour « Average Distance/ Distance Degree ») qui est un descripteur moléculaire théorique, et ClogP peuvent être échangés sans variations pertinentes dans les résultats statistiques.

3-Le modèle non - logP obtenu dans ce travail a de très bonnes performances d'ajustement, il est robuste et possède un pouvoir prédictif acceptable. Les paramètres de validation interne (Q^2_{LOO} et $Q^2_{L(5)O}$) sont similaires aux paramètres d'ajustement.

Il est à noter que le composé chimique 16 (1-propylamine) caractérisé par des résidus standardisés supérieurs en valeurs absolues à 3unités d'écart-type est un point aberrant sérieux dans tous les modèles envisagés.



*MODELISATION DE LA
CONCENTRATION LETALE 50*

II- Modélisation

II-1 Introduction

L'environnement est en permanence exposé aux substances chimiques, comme le benzène et ses dérivés, utilisées dans des procédés industriels. La connaissance de la toxicité aquatique est nécessaire pour l'évaluation du risque et/ ou du danger que peuvent constituer les substances chimiques vis-à-vis des organismes aquatiques aussi bien marins que d'eau douce [54]. La mesure des effets toxiques demande du temps et de l'argent, aussi le recours à la modélisation à partir de la structure des composés est une alternative intéressante qui peut être exploitée dans la lutte contre la pollution environnementale, la conception de médicaments avec effet minimal de toxicité et l'interprétation des mécanismes de toxicité [55, 56].

Les modèles structure - toxicité sont à la frontière de la biologie, de la chimie et des statistiques. Le rapprochement de ces trois domaines a permis de développer les relations structure - activité comme sous-discipline reconnue de la toxicologie. On estime que les décennies à venir verront la grande utilisation des relations quantitatives structure - activité (QSARs) pour la prévision de la toxicité de produits chimiques futurs ou déjà existants. La plus grande focalisation portera sur leur application pour réduire, voire remplacer, l'utilisation des animaux dans l'évaluation toxicologique afin de déterminer la régulation des produits chimiques (par exemple dans la législation REACH) [57]. En 1962, Hansch et al [58] publièrent un article mettant en corrélation l'activité biologique et le coefficient de partage octanol-eau [59], aussi considère-t-on cette année comme la date de naissance officielle des QSARs.

Les relations quantitatives structure - activité sont des modèles mathématiques en termes de descripteurs moléculaires. Le modèle QSAR est utile pour la compréhension des facteurs régissant et la conception de composés efficaces [58]. Les principaux problèmes rencontrés dans ce type de recherche portent toujours sur la description de la structure moléculaire à l'aide de descripteurs moléculaires appropriés et la sélection de méthodes de modélisation convenables. Actuellement, plusieurs types de descripteurs moléculaires tels que les indices topologiques et les paramètres quato-chimiques ont été proposés pour la description des caractéristiques structurales des molécules [60-62]. Plusieurs méthodes

chimométriques différentes, tels que la régression linéaire multiple (RLM), la régression des moindres carrés partiels (PLS), différents types de réseaux de neurones (RNA), les algorithmes génétiques (AG) et les machines à vecteur support (SVM) peuvent être utilisés pour le calcul de modèles mettant en corrélation les structures des molécules et leurs propriétés [63].

Le succès de tout modèle de QSAR dépend de l'exactitude des données d'entrée, de la sélection des descripteurs pertinents qui représentent quantitativement les variations des propriétés structurales des molécules, les outils statistiques et la validation du modèle développé [64-68]. Les stratégies de validation vérifient la fiabilité des modèles développés quant à leur éventuelle application à un nouvel ensemble de données et de juger du niveau de confiance lié à la prédiction. Habituellement quatre stratégies sont adoptées pour la validation des modèles QSAR [69]: a/ validation interne ou validation croisée; b/ validation en éclatant au préalable les données en deux ensembles de calibrage et de test pour la validation externe; c/ validation externe véritable par application du modèle sur de nouvelles données externes et d/ randomisation des données ou Y-scrambling. On aboutit ainsi à une relation mathématique simple, de la forme:

$$\text{Propriété} = f(\text{descripteurs structuraux})$$

Les techniques de relation quantitatives structure - activité englobent les mesures chimiques et les analyses biologiques de départ jusqu'aux techniques statistiques et l'analyse des résultats [59, 70, 71].

Dans cette partie une étude QSAR à été réalisée en vue de développer un modèle qui relie la concentration létale de 92 dérivés benzéniques substitués à leurs structures représentées par des descripteurs théoriques. Les algorithmes génétiques ont été imposés dans le choix des descripteurs les plus instructifs parmi ceux calculés à l'aide du logiciel Dragon (version 5.3) [44]. Les descripteurs sélectionnés ont été exploités pour le développement, selon différentes approches (régression linéaire multiple; réseaux de neurones artificiels ; machine à vecteur support), d'un modèle visant à prédire le logarithme (décimal) de l'inverse de la concentration létale 50 % (1 / CL50) soit $pCL50 = \log(1/CL50)$. Les modèles ont été validés en divisant au préalable, à l'aide de l'algorithme CADEX de Kennard et Stone, les données expérimentales disponibles en deux sous – ensembles disjoints :

- l'un de calibrage, comprenant 74 composés (environ 80% des données) pour le calcul du modèle,

- et l'autre, constitué des 18 composés restants (environ de 20%), dit de test, pour la validation externe.

Différentes techniques statistiques ont été employées pour faire ressortir les conditions structurales à même de développer un modèle de toxicité idéal.

Trois objectifs ont été assignés à ce travail : (1) exploration des relations structure/activité de la toxicité aquatique de composés divers, (2) choix parmi tous les modèles chimiométriques comparables, du meilleur modèle prédictif de la toxicité aquatique, (3) vérification de la performance et de la stabilité du modèle obtenu par les trois approches : régression linéaire multiple ; réseaux de neurones artificiels ; machine à vecteur support.

II-2 MATERIELS ET METHODES

II-2-1 Ensemble des données

Les valeurs expérimentales pCL50 relatives à 92 dérivés benzéniques substitués (tableau IV), ont été extraites de la littérature [72, 73].

Tableau IV : Les Composés et résultats prédictifs de l'activité biologique pCL50

Composés	LC50	SVM	MLR-LOO	ANN (6-4-1)
Chlorobenzene	-3.77	-3.83035	-3.74546	-3.74151
1.2-dichlorobenzene	-4.4	-4.48338	-4.38271	-4.43937
1.4-dichlorobenzene	-4.56	-4.56954	-4.46907	-4.51028
1.2.3-trichlorobenzene	-4.89	-4.98056	-4.88053	-4.98078
1.2.4-trichlorobenzene	-4.83	-5.14178	-5.05285	-5.0448
1.3.5-trichlorobenzene	-4.74	-5.10292	-5.00723	-5.076
1.2.3.4-tetrachlorobenzene	-5.35	-5.57135	-5.49524	-5.41645
1.2.4.5-tetrachlorobenzene	-5.85	-5.49538	-5.42123	-5.33095
1-chloro-3-methyl-benzene	-3.84	-4.09821	-4.038	-4.03969
1-chloro-4-methyl-benzene	-4.33	-4.22253	-4.11184	-4.15287
1.2.4-trichloro-5-methyl-benzene	-5.06	-5.38214	-5.29733	-5.24486
1.2-dichloro-4-methyl-benzene	-4.6	-4.84391	-4.74415	-4.79413
1.2.3.4.5-pentachloro-6-methyl-benzene	-6.15	-6.2801	-6.20402	-5.83564
Benzene	-3.09	-2.93762	-2.86614	-3.0203

Composés	LC50	SVM	MLR-LOO	ANN (6-4-1)
Toluene	-3.13	-3.31579	-3.28472	-3.26926
1.2-xylene	-3.48	-3.62715	-3.59907	-3.55371
1.4-xylene	-3.48	-3.58264	-3.55134	-3.50956
nitrobenzene	-2.97	-3.33114	-3.28405	-3.31017
1-methyl-2-nitro-benzene	-3.59	-3.68566	-3.58487	-3.60532
1-methyl-3-nitro-benzene	-3.65	-3.81668	-3.71205	-3.73011
1.2-dimethyl-3-nitro-benzene	-4.39	-4.11069	-3.97564	-4.02403
1.2-dimethyl-4-nitro-benzene	-4.21	-4.2327	-4.10444	-4.13594
1-chloro-2-nitro-benzene	-3.72	-3.93677	-3.92089	-3.88837
1-chloro-4-nitro-benzene	-4.42	-4.64256	-4.49677	-4.51232
1.2-dichloro-3-nitro-benzene	-4.66	-4.50215	-4.50354	-4.51735
2.4-dichloro-1-nitro-benzene	-4.46	-4.60542	-4.61065	-4.59398
1.3-dichloro-5-nitro-benzene	-4.58	-4.64973	-4.64906	-4.65265
1-chloro-2-methyl-3-nitro-benzene	-4.52	-4.4445	-4.38609	-4.38701
4-chloro-1-methyl-2-nitro-benzene	-4.44	-4.36645	-4.31572	-4.32848
Phenol	-3.45	-3.22425	-3.31629	-3.25839
2-methylphenol	-3.77	-3.53301	-3.57994	-3.49417
4-methylphenol	-3.74	-3.58894	-3.64372	-3.55421
2.6-dimethylphenol	-3.75	-3.84254	-3.87107	-3.78263
2.3.6-trimethylphenol	-4.21	-4.15463	-4.16903	-4.14101
4-ethylphenol	-4.07	-3.90861	-3.91137	-3.85323
4-propylphenol	-4.09	-4.25004	-4.24035	-4.20591
4-butylphenol	-4.47	-4.60121	-4.57657	-4.55226
4-tert-butylphenol	-4.46	-4.39709	-4.57808	-4.51038
4-methyl-2-tert-butyl-phenol	-4.9	-4.58099	-4.70879	-4.74544
4-pentylphenol	-5.12	-4.92698	-4.85182	-4.84189
4-(2-methylbutan-2-yl)phenol	-4.81	-4.68858	-4.83424	-4.77966
2-prop-2-enylphenol	-3.96	-3.90149	-3.88034	-3.81388
2-phenylphenol	-4.76	-4.45127	-4.57126	-4.58508
naphth-1-ol	-4.5	-4.26329	-4.35766	-4.34269
4-chlorophenol	-4.18	-3.93639	-3.9572	-3.91333
4-chloro-3-methyl-phenol	-4.33	-4.16685	-4.17845	-4.15503
4-chloro-3.5-dimethyl-phenol	-4.66	-4.46813	-4.47471	-4.47649
4-methoxyphenol	-3.05	-3.3507	-3.3751	-3.31629
4-phenoxyphenol	-4.58	-4.56774	-4.71578	-4.68027
(2S)-2-amino-3-(3H-imidazol-4yl)propanoic	-3.63	-3.6473	-3.67662	-3.59185
Aniline	-2.91	-3.11501	-3.1868	-3.13259
2-methylaniline	-3.12	-3.47612	-3.51939	-3.39356
4-methylaniline	-3.72	-3.48476	-3.46109	-3.39861
N,N-dimethylaniline	-3.33	-3.46437	-3.40551	-3.38112
2-ethylaniline	-3.21	-3.57518	-3.48017	-3.42124

Composés	LC50	SVM	MLR-LOO	ANN (6-4-1)
4-ethylaniline	-3.52	-3.58244	-3.45392	-3.43907
4-butylianiline	-4.16	-4.16551	-4.01168	-4.01078
2,6-dipropan-2-ylaniline	-4.06	-4.10066	-4.2282	-4.09578
2-chloroaniline	-4.31	-3.83238	-3.72177	-3.70656
4-chloroaniline	-3.67	-3.84311	-3.80996	-3.74746
2,5-dichloroaniline	-4.99	-4.56596	-4.44545	-4.47459
3,4-dichloroaniline	-4.39	-4.43828	-4.38394	-4.37531
2,3,6-trichloroaniline	-4.73	-5.13898	-5.06185	-4.99604
2,4,5-trichloroaniline	-4.92	-5.09255	-5.0133	-4.98045
4-bromoaniline	-3.56	-3.84621	-3.68878	-3.66952
4-fluro-3-(trifluoromethyl) aniline	-3.77	-3.66578	-3.68848	-3.66669
4-fluro-2-(trifluoromethyl) aniline	-3.78	-4.14417	-4.09831	-4.0351
2,3,4,5,6-pentafluroaniline	-3.69	-3.77702	-4.06334	-3.82059
3-phenylmethoxyaniline	-4.34	-4.59314	-4.58536	-4.5044
4-hexoxyaniline	-4.78	-4.80503	-4.62292	-4.63588
2-nitroaniline	-4.15	-3.63788	-3.43008	-3.51284
4-nitroaniline	-3.23	-3.44366	-3.42475	-3.40801
2-chloro-4-nitro-aniline	-3.93	-4.06466	-4.03416	-3.98182
4-ethoxy-2-nitro-aniline	-3.85	-4.18033	-4.01758	-4.00761
1,3-dichlorobenzene	-4.28	-4.43198	-4.17	-4.17
1,2,3,5-tetrachlorobenzene	-5.43	-5.6432	-5.67913	-5.67913
2,4-dichloro-1-methyl-benzene	-4.54	-4.69806	-4.60076	-4.60076
1-methyl-4-nitro-benzene	-3.67	-3.73465	-3.45165	-3.45165
1-chloro-3-nitrobenzene	-4.01	-4.08952	-3.92861	-3.92861
1,4-dichloro-2-nitro-benzene	-4.59	-4.59178	-4.51358	-4.51358
3-methylphenol	-3.48	-3.48421	-3.4927	-3.4927
2,4-dimethylphenol	-3.86	-3.91718	-3.9888	-3.9888
3,4-dimethylphenol	-3.92	-3.87816	-3.91363	-3.91363
3-methoxyphenol	-3.22	-3.06706	-3.07564	-3.07564
3-methylaniline	-3.47	-3.40947	-3.55747	-3.55747
3-ethylaniline	-3.65	-3.57373	-3.43421	-3.43421
3-chloroaniline	-3.98	-3.81601	-3.73967	-3.73967
2,4-dichloroaniline	-4.41	-4.47374	-4.50667	-4.50667
3,5-dichloroaniline	-4.62	-4.42679	-4.39361	-4.39361
2, 3,4-trichloroaniline	-5.15	-4.96576	-4.99111	-4.99111
3-nitroaniline	-3.24	-3.57797	-3.75482	-3.75482

II-2-2 Calcul des descripteurs

Rappelons que la représentation numérique de la structure chimique (descripteurs moléculaires) est une étape importante de l'investigation QSAR. Les performances du modèle élaboré et la précision des résultats sont étroitement liées au mode de détermination de ces descripteurs.

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 [74]. Les géométries finales d'énergie conformationnelle minimale ont été obtenues par la méthode semi-empirique PM3, 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 [44] pour le calcul de plus de 1664 descripteurs (en tenant compte de ceux calculés à l'aide du logiciel HyperChem) appartenant à 20 classes différentes.

En utilisant les options correspondantes du logiciel DRAGON, nous avons d'abord éliminé les descripteurs à valeurs constantes (écarts types inférieurs à 0,001) qui n'apportent aucune information, ensuite ceux qui sont hautement corrélés ($R > 0,95$) et qui véhiculent une information redondante. Pour chaque paire de descripteurs corrélés, est éliminé automatiquement celui qui présente les plus hautes corrélations croisées avec les autres descripteurs.

II-2-3 l'algorithme de Kennard et Stone (CADEX)[75]

Il est important de définir rationnellement un ensemble de calibrage pour la construction du modèle et un ensemble de test externe sur lequel évaluer le pouvoir de prédiction de ce modèle. L'objectif de cette sélection vise la génération de deux ensembles présentant une diversité moléculaire similaire, de telle sorte à être réciproquement représentatifs et couvrir les principales caractéristiques structurales et physiologiques de l'ensemble global des données.

De nombreuses procédures peuvent être adoptées pour la sélection des ensembles de calibrage et de test, ce dernier doit comprendre de 15 à 40% des composés de l'ensemble complet des données.

L'algorithme CADEX de Kennard et Stone sélectionné est une technique séquentielle qui maximise les distances euclidiennes entre les nouveaux échantillons sélectionnés et ceux qui le sont déjà. Elle commence par situer les deux échantillons les plus éloignés l'un de l'autre, qui sont retirés de la base de données initiales et affectés à l'ensemble de calibrage.

Pour chaque échantillon non sélectionné (ech.i), l'algorithme calcule la distance vers chaque échantillon déjà sélectionné, et attribue à (ech.i) la plus petite des distances.

L'échantillon (ech.i) associé à la plus grande distance est donc le plus éloigné de tous les échantillons déjà sélectionnés ; c'est donc celui qui est sélectionné. La procédure est répétée jusqu'à l'obtention du nombre d'échantillons désirés pour l'ensemble de calibrage.

Le fait de sélectionner les échantillons les plus éloignés les uns des autres introduit une grande diversité dans l'ensemble de calibrage ; l'obtention d'une répartition uniforme est un autre avantage de cette technique. L'algorithme CADEX de Kennard et Stone est considéré comme l'un des meilleurs moyens pour la construction des ensembles de calibrage et de validation (test)[67, 76]

II-2-4. Sélection des descripteurs

L'analyse par régression linéaire multiple et la sélection des variables explicatives ont été réalisées avec le logiciel MOBY DIGS [77] 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 selection ou AG-VSS) [78].

Dans le logiciel Moby Digs 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.

Tout d'abord des modèles à 1 ou 2 variables explicatives ont été développés par la procédure « all-subset-method » afin d'explorer toutes les combinaisons de basses dimensions.

Le nombre de descripteurs est par la suite augmenté d'une unité à chaque fois, pour construire des modèles de plus hauts rangs. L'algorithme génétique est arrêté lorsque l'augmentation de la taille du modèle n'entraîne plus un accroissement significatif des valeurs de Q^2 . Une attention particulière a été réservée à la colinéarité des descripteurs moléculaires sélectionnés, par application de la règle QUIK (Q Under Influence of K) [79], une condition nécessaire pour la validité du modèle. Un modèle n'est acceptable que lorsque la corrélation globale du bloc $[x+y]$ (k_{xy}) est supérieure à la corrélation du bloc de la variable x (k_{xx}) soit : $(k_{xy} - k_{xx}) \geq 0$, x désignant les descripteurs moléculaires et y la variable réponse.

La colinéarité dans l'ensemble originale des descripteurs moléculaires se traduit par de nombreux modèles similaires qui, plus ou moins, ont la même capacité prédictive (dans le logiciel Moby Digs 100 modèles de différentes dimensions). Parmi les modèles de même performance ceux caractérisés par les plus grands Dk ($k_{xy} - k_{xx}$) sont sélectionnés pour être soumis à un contrôle plus poussé.

Pour chaque ordre (dimension du modèle) les meilleurs modèles sont sélectionnés, et le modèle final choisi parmi ceux-ci.

Le modèle retenu doit présenter une corrélation suffisante et, en même temps prévenir contre toute sur-paramétrisation, ce qui conduirait à une perte du pouvoir de prédiction pour les échantillons externes à l'ensemble de calibrage.

D'un point de vue statistique le rapport entre le nombre d'échantillons (n) et le nombre de descripteurs (m) ne doit pas être trop faible. Habituellement la valeur $n/m \geq 5$ est recommandée [80].

II-3 RESULTATS ET DISCUSSION.

II-3-1 Résultat de la régression linéaire multiple.

La figure 2 reproduit les valeurs de R^2 et Q^2 obtenues au cours de l'analyse AG/ RLM.

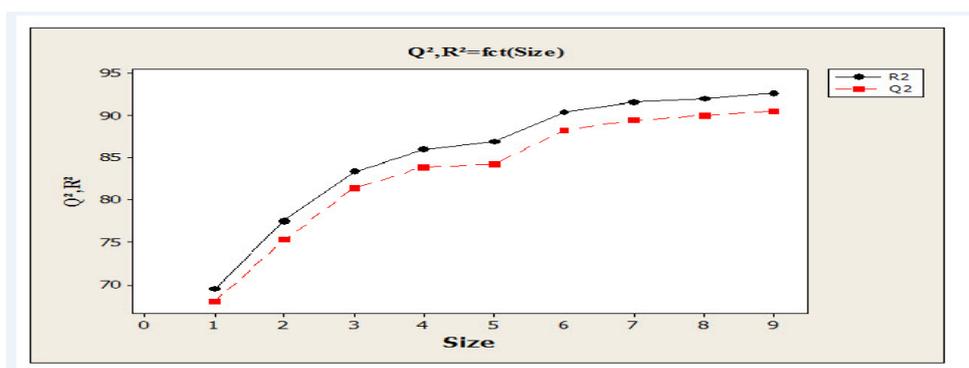


Figure. 2 variation de R^2 et Q^2 en fonction de la taille du modèle

Manifestement $pCL50$ maximum n'est pas, séparément, en corrélation linéaire avec quelque descripteur moléculaire que ce soit, les valeurs de Q^2 , dans chaque cas, étant faible. La valeur de Q^2 augmente graduellement avec le nombre de descripteurs lorsque l'ajout d'un descripteur (augmentation de la dimension du modèle d'une unité) n'améliore pas de manière significative les statistiques d'un modèle, on a déterminé la taille optimale du sous-ensemble.

Tableau V: Comparaison de la performance des modèles de différentes tailles.

N	R2	R2-R1	Q2	Q2-Q1	Q2 _{ext}
1	0.6961		0.6804		0.6992
2	0.7751	0.079	0.7532	0.0728	0.9378
3	0.8337	0.0586	0.8147	0.0615	0.9299
4	0.86	0.0263	0.8393	0.0246	0.917
5	0.8688	0.088	0.8427	0.0034	0.927
6	0.9039	0.0351	0.8829	0.0402	0.9538
7	0.9152	0.0113	0.8946	0.0117	0.9319
8	0.9203	0.0051	0.9005	0.0059	0.9003
9	0.9257	0.0054	0.9055	0.005	0.8574

Plusieurs bons modèles, basés sur différents ensembles de descripteurs moléculaires, ont pu être calculés.

Le modèle optimal (construit sur l'ensemble de calibrage) choisi, est de dimension 6. Il a pour équation :

$$pLC50 = - 1.36 - 0.164 Po + 6.24 MATS1m + 0.550 RDF020v - 0.433 E1s + 0.0625 PCWTe - 0.261 C \log p \quad (7)$$

$$n_{tr} = 74 ; R^2 = 0.904 ; Q_{LOO} = 0.8829 ; S = 0.21 ; F = 105,04 ; P = 0,000$$

Où : Po désigne la polarisabilité; MATS1m, l'indice d'autocorrelation de Moran de distance topologique 1 pondéré par les masses atomiques m ; RDF020v, la fonction de distribution radiale -2,0, pondérée par les volumes de Van der Waals ; E1s, représente les éléments de la matrice d'adjacence des arrêtes; PCWT^E, est l'indice électronique topologique pondéré par la charge partielle; ClogP, le logarithme du coefficient de partage octanol/eau calculé.

De plus amples informations concernant ces descripteurs peuvent être trouvées dans le guide de l'utilisateur « du logiciel Dragon et dans les références indiquées.

La valeur de R² indique que 0,904 de variation totale est expliquée par le modèle (bon ajustement), alors que la valeur très élevée du rapport de la variance expliquée par le modèle à la variance résiduelle (F = 105,04 ; p = 0,000) montre que l'équation (7) permet une très bonne prédiction des n (=74) valeurs de pCL50 de l'ensemble de calibrage, (erreur standard s= 0.21), la valeur très élevée de Q²_{LOO}, qui diffère très peu de celle de R², renseigne sur la robustesse du modèle.

II-3-1-1 Analyse de contribution des descripteurs.

Basées sur une procédure précédemment décrite [81, 82], les contributions relatives de chacun des six descripteurs au modèle ont été déterminées et sont représentées dans la figure (3).

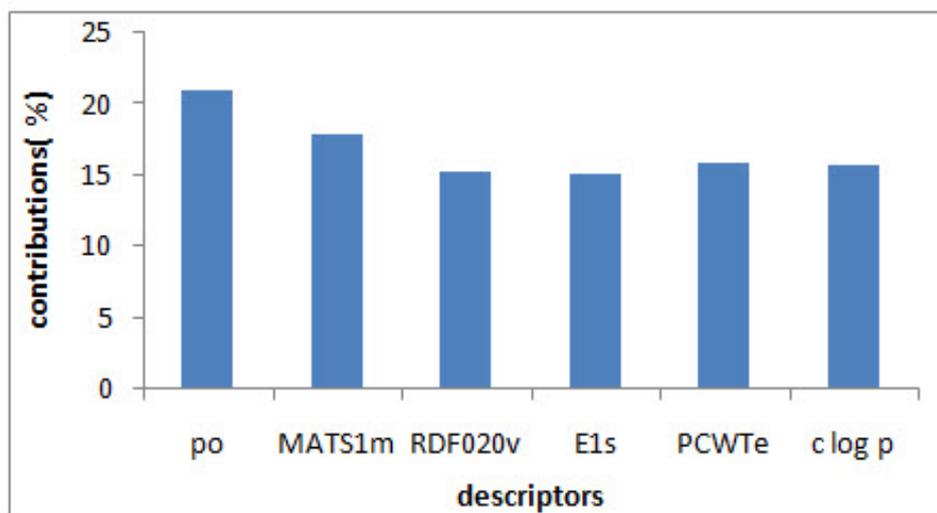


Figure. 3 La contribution relative en fonction du nombre de descripteurs

D'après cet histogramme on constate que les six descripteurs sont nécessaires pour la construction du modèle QSAR.

Le tableau VI résume les caractéristiques des descripteurs sélectionnés

Tableau VI. Caractéristiques des descripteurs sélectionnés par MLR

Predictor	Coef	SE Coef	T	P	VIF
Constant	-1.3635	0.1683	-8.10	0.000	-
Po	-0.16427	0.01795	-9.15	0.000	4.326
MATS1m	6.2420	0.9533	6.55	0.000	1.171
RDF020v	0.5503	0.1764	3.12	0.003	3.603
E1s	-0.4333	0.1070	-4.05	0.000	1.319
PCWTe	0.06254	0.01146	5.46	0.000	2.016
Clog p	0.26124	0.05402	-4.84	0.000	3.932

La valeur de t pour un descripteur est liée à sa signification statistique. Les valeurs absolues élevées des t rapportées indiquent que les coefficients de régression sont significativement plus grands que l'écart type. La probabilité de t (p) pour un descripteur donne sa signification statistique lorsqu'il est impliqué dans un modèle QSRR global ; elle renseigne sur les interactions entre descripteurs. Les descripteurs auxquels correspondent des probabilités de t inférieures à 0.05 sont considérés comme statistiquement significatifs pour un modèle donné, c'est-à-dire que leur influence sur la variable dépendante n'est pas due au hasard [83]. Les valeurs des probabilités de t pour les six descripteurs sont très petites, ce qui indique qu'ils sont hautement significatifs. Les valeurs des facteurs d'inflation de la variance et la matrice de corrélation reproduite dans le tableau (VII) suggèrent que ces descripteurs sont faiblement corrélés entre eux [84].

Table VII: Matrice de corrélation

	pCL50	Po	MATS1m	RDF020v	E1s	PCWTe
Po	-0.653					
MATS1m	-0.065	0.181				
RDF020v	0.269	0.381	-0.170			
E1s	-0.332	-0.076	0.240	-0.378		
PCWTe	0.098	0.482	-0.145	0.607	-0.417	
$c \log p$	-0.834	0.607	0.220	-0.332	0.187	0.051

II -3-1-2- Signification des descripteurs sélectionnés

1- Polarisabilité

Toute molécule, polaire ou non polaire, est polarisable, c'est-à-dire que ses électrons peuvent être déplacés sous l'effet d'un champ électrique \vec{E} ce qui induit une polarisation P dans la molécule. La polarisation est proportionnelle à l'intensité du champ électrique. Les plus simples relations entre P et \vec{E} , données sous forme scalaire, sont :

$$P = \chi E = \frac{\epsilon - 1}{4\pi} E \quad (8)$$

Où : π est la susceptibilité diélectrique et ϵ la constante diélectrique ou permittivité.

La polarisation peut être séparée en deux contributions principales : polarisation induite P_x , conséquence des effets de translation, et polarisation dipolaire P_μ , conséquence de l'orientation des moments permanents. De plus, on peut considérer la polarisation induite comme conséquence d'une contribution d'une polarisation électronique P_E et d'une polarisation atomique P_A :

$$\mathbf{P} = \mathbf{P}_\alpha + \mathbf{P}_\mu = \mathbf{P}_E + \mathbf{P}_A + \mathbf{P}_\mu \quad (9)$$

Notons le rôle de la polarisabilité dans :

- Les interactions dipôle-dipôle induit et,
- Les interactions dipôle induit-dipôle instantané.

2- Indice d'autocorrélation de Moran de distance topologique 1 pondéré par les masses atomiques m : MATS 1m :

L'indice d'autocorrélation de Moran MATS kw , où w est la propriété atomique utilisée pour pondérer le graphe moléculaire et k le décalage (distance topologique), est calculé en appliquant le coefficient de Moran [85] au graphe moléculaire :

$$\text{MATS}kw = \frac{\frac{1}{\Lambda} \sum_{i=1}^{nsk} \sum_{j=1}^{nsk} \delta_{ij} (w_i - \bar{w})(w_j - \bar{w})}{nsk \sum_{i=1}^{nsk} (w_i - \bar{w})^2}}{\frac{1}{nSK} \sum_{i=1}^{nSK} (w_i - \bar{w})^2} \quad (10)$$

Où : w est n'importe quelle propriété atomique, \bar{w} sa moyenne pour la molécule, $n s k$ est le nombre d'atomes autres que l'hydrogène, δ_{ij} est le delta de Kronecker ($\delta_{ij}=1$ si d_{ij} sinon zéro, d_{ij} est la distance topologique entre les 2 atomes considérés) est la somme des deltas de Kronecker, c'est-à-dire le nombre de paires d'atomes distants de k .

3-RDF020v

Est une fonction de Distribution Radiale-2.0/pondérée par les volumes de van der Waals – la fonction de distribution radiale (RDF pour Radial distribution function) du logiciel DRAGON [44] encode des caractéristiques atomiques et géométriques des structures chimiques dans l'espace 3D selon les propriétés des paires atomiques A_i et A_j qui dans ce travail correspondent au volume de van der Waals. L'équation (10) utilise un terme gaussien comme fonction de distance :

$$RDF(r) = f \sum_{i=1}^{N-1} \sum_{j=i+1}^N A_i A_j e^{-B(r-r_{ij})^2} \quad (11)$$

Où f est un facteur d'échelle, N le nombre d'atomes i et j et r_{ij} les distances interatomiques. Un paramètre de lissage, B , définit la distribution de probabilité des distances particulières des atomes qui peuvent être considérées comme leurs vibrations dans la molécule.

Selon [86] RDF peut être interprété comme la densité de probabilité de trouver un atome dans un volume sphérique de rayon r . Sur la base de cette définition il est possible de supposer que RDF020v un code une contribution partielle dans les régions où les atomes sont distants de $r = 2 \text{ \AA}$ du « centre géométrique » de chaque molécule. Le programme calcule trente descripteurs pondérés par les volumes atomiques avec des pas de $r = 0.5 \text{ \AA}$ et uniquement RDF020v a un rapport avec le modèle QSAR.

4- Descripteurs de charge

Ce sont des descripteurs électroniques définis en termes de charges atomiques utilisées pour décrire les aspects électroniques aussi bien de la molécule entière que de régions particulières, comme les atomes, les liaisons, les fragments moléculaires et les orbitales. Les charges renseignent sur l'importance de la localisation de la densité électronique dans une molécule : des valeurs q_i négatives signifient la localisation d'un excès de charge électronique au centre i , alors que des valeurs positives indiquent que le centre i est électroniquement déficitaire.

Dans une molécule les charges électriques constituent les forces motrice des interactions électrostatiques, et il est bien connu que les densités des charges électroniques jouent un rôle fondamental dans les propriétés physicochimiques et l'affinité de liaison.

Les descripteurs de charge sont calculés par les méthodes de la chimie computationnelle et font partie des descripteurs quato-chimiques. Dans le cadre de la chimie quantique, l'analyse de population est l'outil de base utilisé pour le calcul des charges atomiques.

Dans l'analyse de population de Mulliken on a vu [cf : (partie II) I-2-2] que le schéma de partage est basé sur l'utilisation de la matrice densité, \mathbf{P} et la matrice de recouvrement, \mathbf{S} . La charge atomique nette q_a du $a^{\text{ème}}$ atome est définie par :

$$q_a = Z_a - \sum_{\mu=1}^{nOA} \sum_{\nu=1}^{nOA} P_{\mu\nu} \cdot S_{\mu\nu} \quad \mu \in a \quad (12)$$

n OA désigne le nombre d'orbitales (les fonctions de base), Z_a est la charge nucléaire effective du $a^{\text{ème}}$, $P_{\mu\nu}$ l'élément de la matrice densité qui correspond à l'orbitale atomique μ centrée sur le $a^{\text{ème}}$ et $S_{\mu\nu}$ l'élément correspondant aux orbitales μ et ν de la matrice de recouvrement. La somme porte sur toutes les orbitales atomiques du $a^{\text{ème}}$ atome .

Comme pour tous les autres schémas de répartition de densité électronique dans les molécules, l'analyse de population de Mulliken est arbitraire et dépend fortement de la base particulière choisie. Cependant, la comparaison des analyses de population pour une série de molécules est utile pour la description quantitative des interactions intra-moléculaires, la réactivité chimique et l'information structurale.

Dans une autre approche, l'analyse de population de Lowdin [partie II (I-2-1)] , les orbitales atomiques sont d'abord transformées en un ensemble orthogonalisé, la même transformation s'appliquant aux coefficients de l'orbitale moléculaire.

De plus, des charges atomiques nettes peuvent être également calculées en faisant la distinction entre la densité électronique σ , $q_{a,\sigma}$, et la densité électronique π , $q_{a,\pi}$.

Les descripteurs électroniques topographiques sont calculés à partir des charges atomiques partielles q comme suit [87-89] :

$$T^E = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{|q_i - q_j|}{r_{ij}^2} ; \quad C_{T^E} = \sum_{l=1}^B \left(\frac{q_i - q_j}{r_{ij}^2} \right)_b \quad (13)$$

Dans le premier indice toutes les paires d'atomes (liés ou non) sont prises en compte, alors que le second indice ne fait intervenir que les paires d'atomes i-j liés ; les r_{ij} sont les distances interatomiques. Ces descripteurs sont calculés de façon à refléter, jusqu'à un certain point, les différences de dimensions, de formes et de constitution, ces grandeurs affectant les distributions de charges électroniques et les distances interatomiques des molécules.

L'indice électronique topologique pondéré par la charge partielle : $PCWT^E$ (pour partial charge weighted topological electronic index) est un descripteur électronique moléculaire défini [90] par:

$$PCWT^E = \frac{1}{Q_{max}^-} \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{|q_i - q_j|}{r_{ij}^2} = \frac{T^E}{Q_{max}^-} \quad (14)$$

Q_{max}^- est la charge négative maximale:

$$Q_{max}^- = \max_i (q_i^-)$$

Où les q_i^- sont les charges atomiques négatives nettes.

5- E1s

La matrice d'adjacence des arêtes, déduite du graphe moléculaire, chiffre la connectivité entre les arêtes du graphe. C'est une matrice systématique de dimension $B*B$, où B est le nombre de liaisons, déduite habituellement d'un graphe moléculaire dépourvu des atomes d'hydrogène [91].

Les éléments $[E]_{ij}$ de la matrice sont égaux à 1 si les arêtes e_i et e_j sont adjacentes et à 0 sinon

$$[{}^E A]_{ij} \equiv [E]_{ij} = \begin{cases} 1 & \text{si } (i,j) \\ 0 & \text{sinon} \end{cases} \quad (15)$$

Les moments spectraux de la matrice d'adjacence des arêtes sont définis [89] par:

$$\mu_k = t_r(E^K) \quad (16)$$

où k est la puissance de la matrice d'adjacence et t_r la trace de cette matrice, c'est-à-dire la somme de ses éléments diagonaux.

Le moment spectral d'ordre k peut être exprimé par une combinaison linéaire des fréquences N_k des différents fragments structuraux du graphe. En pratique les N_k sont des descripteurs de comptage comme, par exemple, le nombre de types d'atomes, ou le nombre de fragments à deux atomes etc....

6- Clogp [conf (partie III) I-2-3-1]

II-3-1-3 Droite d'ajustement

D'après la figure suivante il est clair que les valeurs calculées de l'activité calculées sont très proches des valeurs expérimentales.

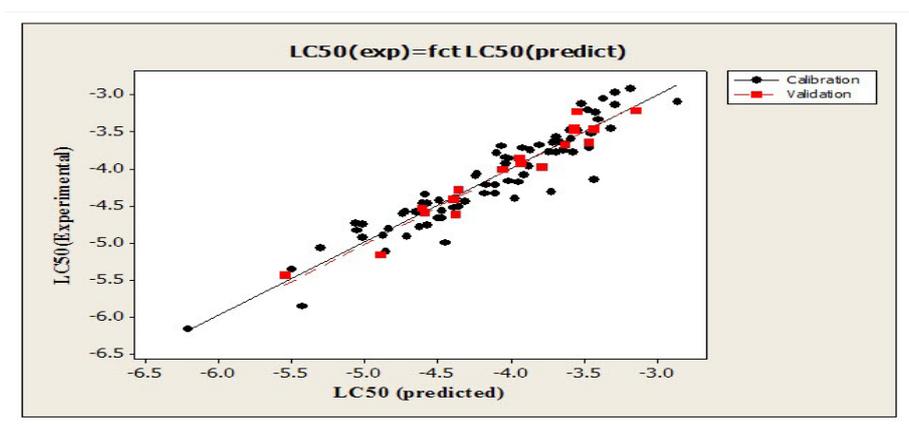


Figure. 4 Droite d'ajustement des pCL50 observés en fonction des pCL50 calculés.

II-3-1-4 Domaine d'application

Tous les résidus standardisés $e_{i \text{ std}}$ sont inférieurs à 3unités d'écart type (3s) à l'exception du composé 71(2,6-dipropan-2-ylaniline)

Les valeurs de h_{ii} , $i^{\text{ème}}$ terme diagonal de la matrice de projection : $\underline{H} = \underline{X}(\underline{X}'\underline{X})^{-1} \underline{X}'$ où \underline{X} est la matrice des valeurs observées des variables explicatives et \underline{X}' sa transposée. La valeur critique pour déterminer les points leviers correspond à $h^* = \frac{3p}{n} = 3*7/74 = 0,28$. On constate que tous les h_i sont inférieurs à cette valeur critique 0,28 à l'exception du composé 68 (3-ethylaniline).

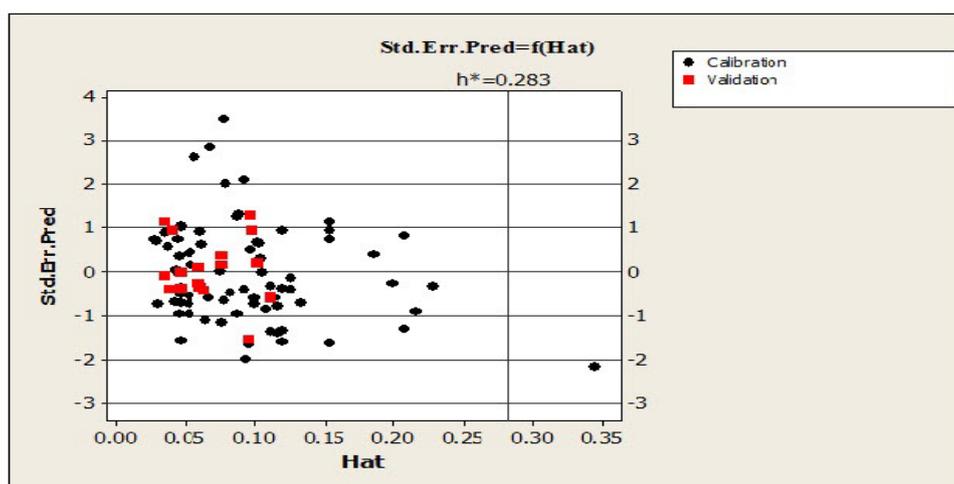


Figure. 5 Diagramme de Williams pour tous les composés.

II-3-1-5 Test de randomisation

Dans le but d'établir que le modèle obtenu n'est pas dû au hasard ou à une sur-spécification, nous avons appliqué le test de randomisation.

Ainsi 100 nouveaux vecteurs de **la concentration létale 50** ont été générés par permutation des positions des composantes du vecteur réel:

$$y = (y_1, y_2, \dots, y_{27}) \xrightarrow{RND} y_{RND} = (y_8, y_5, \dots, y_2)$$

et utilisés comme sources d'observations pour des modèles QSAR dans les conditions optimales établies .

La figure (6) qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés au modèle réel de départ.

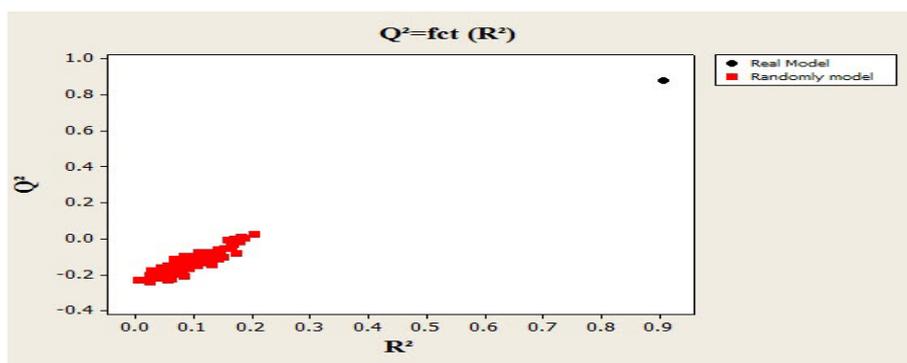


Figure. 6 Test de randomisation

Il est clair que les statistiques obtenues pour les vecteurs modifiés de la concentration létale **50** sont plus petites que celles du modèle QSPR réel, ce qui permet d'affirmer que le modèle proposé n'est pas aléatoire.

II-3-2 Modèle de régression par machines à supports vecteurs (SVM)

Une méthode complémentaire basée sur les méthodes dites SVM (Support Vector Machine) formulées pour la régression, a été appliquée en exploitant le logiciel Molégro Data Modeler (MDM) [92].

La procédure de réglage des paramètres C, Sigma (σ) et Epsilon (ϵ) se fait en inspectant toutes les possibilités résultant de la variation de chacun des paramètres pour un intervalle donné.

Pour chaque jeu de paramètres un modèle est créé et évalué sur un jeu de calibrage (RMSE cal) et un second de validation (RMSE val). Les modèles seront classés par ordre croissant du RMSE cal.

Les hyper paramètres C, ϵ et σ ont donc été déterminés empiriquement en cherchant à minimiser le taux d'erreur sur la base de validation. Les valeurs ($C = 88888,9$ et $\sigma = 1e^{-05}$) ont permis d'obtenir des : ($Q^2 = 0,8882$; RMSE= 0,222) pour l'ensemble de calibrage, et ($Q^2 = 0,947$; RMSE= 0,146) pour l'ensemble de validation.

II-3-3 Réseaux de neurones artificiels

Les six (6) descripteurs du modèle RLM ont été utilisés pour l'optimisation neuronale. Après avoir cherché le nombre de neurones dans la couche cachée, l'algorithme d'apprentissage se présente comme suit :

Tableau VIII : Structure optimale du réseau de neurones

Nombre d'entrées	06 (les descripteurs)
Nombre de sortie	01 (p CL50)
Nombre de couches cachées	01
Nombre de neurones dans la couche cachée	04
Algorithme d'apprentissage	Rétropropagation du gradient de l'erreur
Fonction d'apprentissage	Tangente hyperbolique

Le tableau IX permet de comparer les paramètres statistiques obtenus avec chacune des trois (3) approches .

Tableau IX : Les paramètres statistiques :

Approche	Q^2	RMSE	Q^2_{ext}	RMSE _{ext}
MLR	0,8829	0,225	0,9538	0,141
RNA	0,898	0,214	0,9564	0,135
SVM	0,8882	0,222	0,947	0,146

Ce qui montre que le modèle basé sur les six (6) descripteurs sélectionnés est stable, robuste et significatif.

II-4 CONCLUSION

Le développement de relations quantitatives structure-propriété/ activité (QSPR/QSAR) par des descripteurs théoriques est un outil puissant non seulement pour la prédiction des propriétés chimiques, physiques et biologiques des composés, mais également

pour une compréhension profonde des mécanismes détaillés de la toxicité aquatique dans les dérivés benzéniques qui prédéterminent cette activité.

Dans cette étude nous avons développé une équation utile de QSAR qui relie des descripteurs chimiques théoriques aux propriétés toxicité aquatiques de 92 dérivés benzéniques. Pour chaque composé 1664 descripteurs (qui appartiennent à 20 classes) calculés par le logiciel Dragon. L'ensemble des données a été divisé en deux ensembles de calibrage et de prédiction, en utilisant l'algorithme de Kennard et Stone. Puis les meilleurs descripteurs ont été sélectionnés par « algorithme génétique » de Moby Dygs. Le modèle obtenu a une qualité statistique élevée et de faibles erreurs de prédiction. En général, On peut conclure que, pour cet ensemble de données, les combinaisons des techniques de modélisations ont comme conséquence une amélioration des modèles linéaires. Les résultats indiquent que les six descripteurs choisis jouent un rôle important dans la toxicité aquatique des dérivés benzéniques.

La méthode QSAR a été appliquée pour la prédiction de la concentration létale des dérivés du benzène. Un modèle linéaire à six descripteurs a été développé par MLR, avec Q^2 de 0,8829 et de RMSE du 0,225, NNA avec Q^2 de 0,898, RMSE de 0,214, SVM avec Q^2 de 0,8882, RMSE de 0,222 pour l'ensemble de calibrage. Plusieurs techniques de validation, y compris laissent-un-dehors la contre-vérification, test de randomisation, et la validation externe, a illustré la fiabilité du modèle proposé. Tous les descripteurs impliqués peuvent être directement calculés à partir de la structure moléculaire des composés, ainsi le modèle proposé est prédictif et pourrait être employé pour estimer la concentration létale des dérivés benzéniques appartenant au même domaine chimique.



REFERENCES BIBLIOGRAPHIQUES

- [1] Zeeman M., Aver C.M., Clements R.G., Nabholtz J.V, and Boethling R.S., (1995)- U.S. EPA Regulatory Perspectives on the use of QSAR for new and existing chemical evaluations SAR QSAR, J Environmental. Research, Vol. 3(3), pp 179-201.
- [2] Walker J.D., (2003)-Applications of QSARs in toxicology: a US Government perspective, Journal of Molecular Structure – J Theochem, Vol. 622(1-2), pp 167-184.
- [3] Bradbury S.P., Russon C.L., Ankley G.T., Schultz T.W, and Walker J.D., (2003)- Overview of data and conceptual approaches for derivation of Quantitative Structure – Activity Relationships, for ecotoxicological effects of organic chemicals, J Environmental Toxicology & Chemistry, Vol. 22 (8), pp 1789-1798.
- [4] European Commission. White Paper on a strategy for a future Community Policy for Chemicals., (2001).[http:// europa .eu.int / comm / enterprise / reach /](http://europa.eu.int/comm/enterprise/reach/).
- [5] Toussaint M.W., Shedd T.R, Van der Schalie W.H, and Leather G.R., (1995)- A comparison of standard acute toxicity tests with rapid screening toxicity tests. J Environmental Toxicology & Chemistry, Vol. 14(5), pp 907-915.
- [6] Kubinyi H., (2002)- From Narcosis to Hyperspace: The History of QSAR, Quantitative Structure.-Activity Relationships., Vol. 21(4), pp 348-356.
- [7] [http:// e c b .j r c .i t / QSAR /](http://ecb.jrc.it/QSAR/).
- [8] Schultz T.W., Cronin M.T.D., Walker J.D, and Aptula A.O., (2003)-Quantitative structure –activity relationships (QSAR_s) in toxicology: a historical perspective, J Molecular Structure –Theochem, Vol.622(1-2), pp 1-22.
- [9] Posthumus R, and Slooff W., (2001)- Implementation of QSARs in ecotoxicological risk assessments RIVM report 601516003.
- [10] Dearden J.C., (2002)- Prediction of Environmental Toxicity and Fate Using Quantitative Structure –Activity Relationships (QSARs), J Brazilian Chemical Society, Vol . 13 (6), pp 754-762.
- [11] Schultz T.W., Cronin M.T.D, and Netzeva T.I., (2003)- The Present Status of QSAR In Toxicology, J Molecular Structure -Theochem, Vol. 622(1-2), pp 23-38.
- [12] Cronin M.T.D. and Dearden J.C., (1995)- QSAR in toxicology .1. Prediction of Aquatic Toxicity, Quantutative Structure.-Activity. Relationships., Vol.14(1), pp 1-7.
- [13] Mannhold R, and Van de Waterbeemd H., (2001)- Substructure and whole molecule approaches for calculating logP, J Computer- Aided Molecular Design, Vol. 15(4), pp 337-354.

- [14] Mannhold R, and Rekker R.F., (2000)- The hydrophobic fragmental constant approach for calculating log P in octanol/water and aliphatic hydrocarbon/water systems. J Perspectives in Drug Discovery & Design, Vol.18(1), pp 1-18.
- [15] Benfenati E., Gini G., Piclin N., Roncaglioni A, and Vari M.R ., (2003)-Predicting log P of pesticides using different software, J Chemosphere, Vol.53(9), pp 1155-1164.
- [16] Manhold R., and Petrauskas A., (2003)- Substructure versus Whole-molecule Approaches for Calculating Log P, J QSAR & Combinatorial Science, Vol .22(4), pp 466-475.1
- [17] <http://clogP.pomona.edu/medchem/chem/clogP/index.html>.
- [18] Klopman G., Li J.K., Wang S, and Dimayuga M., (1994)-Computer Automated log P calculations based on an extended group contribution approach, J Chemical. Information. Computer Sciences, Vol.34(4), pp 752-781.
- [19] Kaiser K.L.E., (2003)- The use of neural networks in QSARs for acute aquatic toxicological endpoints, J Molecular Structure Theochem, Vol .622(1-2), pp 85-95.
- [20] Papa E., Villa F, and Gramatica P., (2005)- Statistically Validated QSARs Based on Theoretical Descriptors , for Modeling Aquatic Toxicity of Organic Chemicals in *Pemiphales promelas* (Fathead Minnow), J Chemical Information & Modeling, Vol.45(5), pp 1256-1266.
- [21] Roy K, and Ghosh G., (2009)- QSTR with extended topochemical atom (ETA) indices. 12. QSAR for the toxicity of diverse aromatic compounds to *Tetrahymena pyriformis* using chemometric tools. J Chemosphere, Vol.77(7), pp 999-1009.
- [22] Zhao Y.H., Zhang X.J., WEN Y., Sun F.T., Guo Z., Qin W.C., Qin H.W., Xu J.L., Sheng L.X, and Abraham M.H., (2010)- Toxicity of organic chemicals to *Tetrahymena pyriformis* : Effect of polarity and ionization on toxicity. J Chemosphere, Vol. 79(1), pp 72-77.
- [23] Roy K, and Das R.N., (2010)-QSTR with extended topochemical atom (ETA) indices.14. QSAR modeling of toxicity of aromatic aldehydes to *Tetrahymena pyriformis*. J Hazardous Materials, Vol. 183(1-3), pp 913-922.
- [24] Bouaoune A., Lourici L., Haddag H, and Messadi D., (2012)-Inhibition of Microbial Growth by anilines: A QSAR study, J Environmental Science and Engineering., A1, Vol. 1(5A), pp 663-671.
- [25] Carpy A., (1999)- Importance de la lipophile en modélisation moléculaire. Analyses, Vol. 27(1), pp 3-6.
- [26] Callander R. (1957)- Annual Review of Plant Physiology. Vol. 8, pp 335-348.
- [27] Hansch C., Quinlan J.E, and Lawrence G.L.(1968)- . J. Organic Chemistry. Vol.33, pp 347-350.

- [28] Leo A.J., Hansch C , and Elkins D. (1971)- Partition coefficients and their uses. Chemical Reviews. Vol.71, pp 525-616.
- [29] <http://www.Organic-Chemistry.Org/prop/peo>.
- [30] OECD Guidelines for testing of Chemicals. (1992)- No.107 OECD, Paris,.
- [31] OECD Guidelines for Testing of Chemicals (1992)- No. 117, OECD, Paris,.
- [32] Fujita T., Iwasa J, and Hansch C., (1964)- A New Substituent Constant, n , Derived from Partition Coefficients. J. American Chemical society, Vol 86, pp 5175-5180.
- [33] Nys C. G, and Rekker R. F., (1973)- Statistical Analysis of a series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. The introduction of Hydrophobic Fragmental Constants (f values), J Chimica Therapeutica, Vol 8, pp 521-535.
- [34] Rekker R. F, and Dekort M., Eur. (1979)- Medicinal Chemistry. Vol 14, pp 479-488.
- [35] Nus G. G., Rekker R. F., (1974)- J Chimica Therapeutica, Vol 9, pp 361-375
- [36] Rekker R. F., (1977)- The hydrophobic fragmental constant. Its derivation and application. A means of characterizing membrane systems. Pharmacochem. Library Vol.1, Elsevier, Amsterdam,
- [37] Broto P, Moreau G, Vanduycke, Eur C. (1984)- Journal of Medicinal Chemistry, Vol 19, pp 71-78.
- [38] <http://clogP.pomona.edu/medchem/chem/clogP/index.html>.
- [39]- Hansch C, and Leo A., (1979)- Substituent Constants for Correlation Analysis in Chemistry and Biology. Wiley & Sons Inc. New York (NY), p 339.
- [40] Nys C.G, and Rekker R.F., (1973)- Statistical Analysis of a series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. The introduction of Hydrophobic Fragmental Constants (f values), J Chimica Therapeutica Vol.8, pp 521-535.
- [41] Kleinöder T., Yan A, and Spycher S., (2008)- Estimation of Octanol/Water Partition coefficient ($\log P_{ow}$). In Chemoinformatics J. Gasteiger, T. Engel (Eds), WILEY-VCH GmbH & Co. KGaA, Weinheim, pp 492-494.

- [42]- Ghose A. K, and Crippen M., (1986)- Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed-Quantitative Structure –activity relationships .I-Partition Coefficients as a Measure of Hydrophobicity, J Computational Chemistry, Vol 7, pp 565-577.
- [43] Viswanadhaan V. N, and Ghose A.K., (1989)- Information and Computer Sciences, Vol 29, pp 163-172.
- [44] Todeschini R., Consonni V., Mauri A, and Pavan M. (2005)- DRAGON Software for the Calculation of Molecular Descriptors–version 5.4 for Windows, Talete s.r.l., Milano, Italy.
- [45] Moriguchi I, Hirono S., Liu O., Nakagome I, and Matsushita Y., (1992)- Simple Method of Calculating Octanol/Water Partition Coefficient, J Chemical & Pharmaceutical Bulletin, Vol.40 (1), pp 127-130.
- [46] Moriguchi I., Hirono S., Nakagome I, and Hirano H, (1994)- Comparison of Reliability of LogP Values for Drugs Calculated by Several Methods, J Chemical & Pharmaceutical Bulletin, Vol. 42(4), pp 976-978.
- [47] Persoone G, and Dive D., (1978)- Ecotoxicology and environmental safety, Vol 2, pp 105-144.
- [48] Schultz T.W., Cajina-Quezada M., Dumont J.N., (1980). Archives Environmental Contamination and Toxicology, Vol 9, pp 591-598.
- [49] Schultz T.W., Cajina-Quezada M., (1982). Archives of Environmental Contamination and Toxicology. Vol 11, pp 353-361.
- [50] Schultz T.W., Lin D.T., Wilke T.S, and Arnold L.M., (1990)- Quantitative structure-activity relationships for the *Tetrahymena pyriformis* population growth endpoint :a mechanism of action approach. In Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology W. Karcher,J. Devillers (Eds) Kluwer Academic Publishers. Dordrecht, Vol.1., 241-262.
- [51] Randic M., Kleiner A.F., and De Alba L.M., (1994)- Distance Matrices, J Chemical. Information. Computer Sciences, Vol. 34(2), 277-286.

- [52] Randic M, and Krilov G., (1999)-On a characterization of the Folding of Proteins, International Journal of Quantum Chemistry, Vol. 75(6), 1017-1026.
- [53] Carbö-Dorca R., Robert D., Amat Li., Gironés X, and Besalu E., (2000)- Molecular Quantum Similarity in QSAR and Drug Design. J Springer-Verlag Berlin Heidelberg. Vol. 42, p 123.
- [54] Netzeva TI, Pavan M, and Worth AP (2008) - Review of (quantitative) structure–activity relationships for acute aquatic toxicity. J QSAR Combinatorial Science, Vol 27, pp 77–90.
- [55] Jalali-Heravi M, and Kyani A (2008)- Comparative structure–toxicity relationship study of substituted benzenes to *Tetrahymena pyriformis* using shuffling-adaptive neuro fuzzy inference system and artificial neural networks. Chemosphere; Vol 72, pp 733–740.
- [56] Zarei K, Atabati M, and Kor K, (2014)- Bee Algorithm and Adaptive Neuro-Fuzzy Inference System as Tools for QSAR Study Toxicity of Substituted Benzenes to *Tetrahymena pyriformis*. J Bull Environ Contam Toxicol; Vol 92, pp 642-649.
- [57] Worth, A.P., Bassan, A., De Bruijn, J., Gallegos Saliner, A., Netzeva, T., Pavan, M., Patlewicz, G., Tsakovska, I. and Eisenreich, S. (2007)- The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods, J SAR & QSAR in Environmental Research, Vol. 18, pp 111-125.
- [58] Hansch C, Maloney P.P, Fujita T. and Muir R.M., (1962)- Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients, J Nature, Vol 194, pp 178-180.
- [59] Bordbar M, Ghasemi J, Fall AY, and Fazaeli R. (2013)- Chemometric Modeling to Predict Aquatic Toxicity of Benzene Derivatives Using Stepwise-Multi Linear Regression and Partial Least Square. Asian Journal of Chemistry, Vol 25(1) pp 331-342.
- [60] Karelson M, (2000)-Molecular Descriptors in QSAR/QSPR. John Wiley & Sons: New York.
- [61] Devillers J, and Balaban, A. T., (1999)- Topological Indices and Related Descriptors in QSAR and QSPR. Eds, Gordon and Breach: Amsterdam, Netherlands;
- [62] Todeschini R, and Consonni, V., (2000)- Handbook of Molecular Descriptors, Wiley-VCH: Weinheim, Germany.

- [63] Darang R, Minaoui B, and Fakir M. (2012)- QSAR Models for Prediction Study of HIV protease inhibitors Using Support Vector Machines Neural Networks and Multiple Linear j.arabjc.10.021
- [64] Tong,W., Hong, H., Xie, Q., Shi, L., Fang, H. and Perkins., R. (2005) - Assessing QSAR limitations –a regulatory perspective, j Current Computer-Aided Drug Design, Vol. 1, pp. 195-205.
- [65] He L. and Jurs, P.J. (2005)- Assessing the reliability of a QSAR model's predictions, J Molecular Graphics and Modelling, Vol. 23, pp. 503-523.
- [66] Ghafourian T. and Cronin, M. (2005)-The impact of variable selection on the modelling of oestrogenicity, J SAR QSAR Environmental Research, Vol. 16, pp. 171-190.
- [67] Tropsha, A., Gramatica, P. and Gombar, V.K. (2003)- The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, J QSAR & Combinatorial Science, Vol. 22, pp. 69-77.
- [68] Golbraikh, A. and Tropsha, A. (2002)- Beware of q²!, J Molecular Graphics and Modelling, Vol. 20, pp. 269-276.
- [69] Wold, S. and Eriksson, L. (1995)- Chemometric Methods in Molecular Design, VCH Publisher, Weinheim.
- [70] Brown, S.D., Sum, S.T., Despagne, F. and Lavine, B.K. (1996)- Chemometrics, J Analytical Chemistry, Vol. 68 No. 1080, pp. 21-61.
- [71] Roy, K. and Leonard, J.T. (2005)- QSAR analyses of 3-(4-benzylpiperidin-1-yl)-Nphenylpropylamine derivatives as potent CCR5 antagonists, J Chemical Information and Modeling, Vol. 45, pp. 1352-1368.
- [72] Kier, L.B. and Hall, L.H. (1986)- Molecular Connectivity in Structure-Activity Analysis, John Wiley & Sons Inc., NewYork, NY.
- [73] Kier, L.B. and Hall, L.H. (1975)- Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, NY.
- [74] HyperChemTM (2002)- Release 6.03 for Windows, Molecular Modeling System.

[75] Keunard R.W, and Stone L.A. (1969) Computer Aided Design of Experiments. Vol. 11(1) pp. 137-148.

[76] Wu, W., Walczak, B., Massart, D.L., Heuerding, S., Erni, F., Last, I.R. and Prebble, K.A. (1996)- Artificial neural networks in classification of NIR spectral data: design of the training set, J Chemometrics and Intelligent Laboratory Systems, Vol. 33, pp. 35-46.

[77] Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. and Pavan, M. (2009)- Moby Digs Software for multilinear regression analysis and variable subset selection by genetic algorithm, Release 1.1 for Windows, Milano.

[78] Leardi, R., Boggia, R. and Terrile, M. (1992)- Genetic algorithms as a strategy for feature selection, J Chemometrics, Vol. 6, pp. 267-281.

[79] Todeschini, R. (1997)- Data correlation, number of significant principal components and shape of molecules. The K correlation index. *Anal. Chim. Acta*, 348: 419-430.

[80] Todeschini, R., Consonni, V., Maiocchi, A. (1999). The K correlation index: theory development and its applications in chemometrics. *Chemom. Intell. Lab. Syst.*, 46 : 13-29.

[81] Zheng, F., Bayram, E., Sumithran, S.P., Ayers, J.T., Zhan, C.G., Schmitt, J.D., Dwoskin, L.P. and Crooks, P.A. (2006)- QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release, *Bioorganic & Medicinal Chemistry*, Vol. 14, pp 3017-3037.

[82] Guha, R. and Jurs, P.C. (2005)- Interpreting computational neural network QSAR models: a measure of descriptor importance, *J Chemical Information and Modeling*, Vol. 45 No. 3, pp. 800-806.

[83] Ramsey, L.F. and Schafer, W.D. (1997)- *The Statistical Sleuth*, Wadsworth Publishing Company.

[84] Holder, A.J., Yourtee, D.M., White, D.A., Glaros, A.G. and Smith, R.J. (2003)- Chain melting temperature estimation for phosphatidyl cholines by quantum mechanically derived quantitative structure property relationships, *J Computer-Aided Molecular Design*, Vol. 17, pp. 223-230.

- [85] P.A.P.Moran. (1950)-Notes on Continuous Stochastic Phenomena. *J. Biometrika*, Vol 37, pp.17-23.
- [86] Hemmer M.C., Steinhauer V., Gasteiger J, (1999)- the Prediction of the 3D Structure of Organic Molecules from Their Infrared Spectra', *Vibrat. Spectroscopy, Handbook Vibrational Spectroscopy.*,Vol 19, pp 151-164.
- [87] Osmialowsky R., Halkiewicz J., Radecki A, and Kliszan R. (1985)- Quantum Chemical Parameters in Correlation Analysis of Gas-Liquid Chromatographic Retention Indices of Amines. *J chromatography.*,Vol 346, pp 53-60.
- [88] Osmaliowsky K., alkiewicz J. H, and Kliszan R. (1986)- Quantum Chemical Parameters in Correlation Analysis of Gas-Liquid Chromatographic Retention Indices of Amines. II. Topological Electronic Index. *J chromatography.*, Vol 361, pp 63-69.
- [89] Katritzky A. R, and Gordeeva E.V.. (1993) - Traditional Topological Indices vs Electronic, Geométrical, and Combined Molecular Descriptors in QSAR /QSPR Research. *J Chemical Information and Computer Sciences*, Vol 33, pp 835-857
- [90] Zefirov N.S., Kirpichenok M.A., Izmailov F.F, and Trofimov M.I. (1987)- Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson's Principle. *Dokl. Akad. Nauk. SSSR*, Vol 296, pp 883-887
- [91] Bonchev D. (1983)- Information Theoretic Indices for Characterization of Chemical Structure. *Research Studies Press, Chichester (éTK)*
- [92] Molegro (2009) Data Modeller (MDM) V.2.1.0 copyright Molegro.



CONCLUSION GENERALE

CONCLUSION GENERALE

Le transport, la distribution, l'accumulation et l'absorption des polluants, pour certains leur fixation sur certaines cibles, leur passage à travers les membranes, leur entrée dans les cellules.....sont autant d'exemples dans lesquels la « modélisation moléculaire » joue un rôle de première importance, par suite du développement prodigieux de l'informatique, et de l'existence sur le marché de logiciels professionnels adaptés.

En partant du principe que toutes les propriétés moléculaires sont non seulement codées par la structure de ces molécules mais résulteraient de celle-ci, on essaie d'établir, à travers un modèle mathématique, une liaison entre la structure moléculaire et n'importe quelle propriété.

Les molécules à analyser sont caractérisées par des descripteurs déduits de la structure (constitution, parfois configuration, et même conformation moléculaires) ou propriétés (physiques, chimiques, biologiques) des molécules. Idéalement, ces descripteurs devraient être rapidement calculables et facilement interprétables.

Le succès de l'approche QSAR/QSPR dépend de façon critique de la définition précise et de l'utilisation appropriée des descripteurs moléculaires.

La recherche de modèles linéaires statistiques prédictifs basés sur des descripteurs théoriques n'ayant pas conduit à des résultats probants d'autres voies sont en cours d'investigation (relation non linéaires entre la grandeur d'intérêt et les descripteurs théoriques).

Les études QSAR ont été réalisées pour évaluer la toxicité aquatique des polluant organique, en utilisant plusieurs approches.

L'approche par régression linéaire simple (RLS) à été réalisée pour construire plusieurs modèles simples c'est-c-à-d un modèle avec un seul descripteur, basés sur les différents coefficients de partages calculées (Alogp, Mlogp, Clogp) et un autre descripteur géométrique 3D (ADDD) obtenu à l'aide du logiciel dragon pour la prédiction de la concentration d' inhibition de la croissance 50 de *tetrahymena pyriformis* par une série de 30 composés de 21alcools et 9 amines.

CONCLUSION GENERALE

Les résultats statistiques obtenus indiquent que le coefficient de partage influe sur la concentration inhibitrice 50, Clogp est le meilleur descripteur choisi pour l'explication de cette grandeur. Il peut le remplacer par le descripteur géométrique ADDD sans variation des valeurs des paramètres statistiques.

Les trois approches par la régression linéaire multiple, le réseau de neurone artificielle, les machines à vecteur de support (MLR, RNA, SVM) ont été appliqués pour relier la toxicité (concentration létale) vis-à-vis du vairon (*Pimephales promelas*) d'une série de composés organiques polluants potentiels de l'environnement aquatique.

Les 92 données de base ont été éclatés en deux ensemble par l'algorithme de Kennard et Stone, un ensemble de 74 composés utilisées pour la construction de modèle et un autre ensemble de 18 composés pour la prédiction externe. La taille du modèle (6 descripteurs moléculaires) à été fixée en maximisant la fonction FIT de Kubinyi, la sélection des variables explicatives à été réalisée par algorithme génétique, dans la version Mobidygs de Todeschini, en maximisant Q^2_{LOO} .

Le modèle QSAR choisi à été développé en utilisant l'analyse de régression multilinéaire, le réseau de neurone artificielle à trois couches (les entrées, une couche cachée, et une couche de sortie) et la machine à vecteur de support avec les valeurs des paramètres optimaux (C, γ et ϵ) (88888,9, $1e^{-05}$ et 0,22222) respectivement. Les résultats obtenus par les trois approches (MLR , ANN , SVM) montre que le modèle choisi à six descripteurs est stable, robuste et très significatives.

Le test de randomisation associé au modèles obtenus permet d'assurer qu'une relation structure/activité réelle a été établie.

Les résultats de la validation externe suggèrent tout à la fois, une bonne capacité prédictive (faible valeur de RMSE calculées sur les différents ensembles), et une possibilité d'extension suffisante (valeurs proches ou similaires) du modèle.

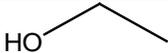
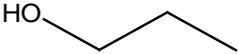
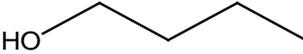
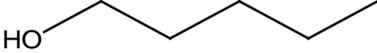
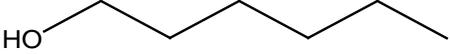
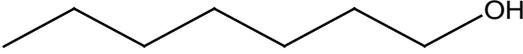
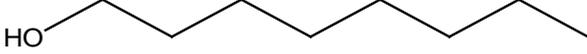
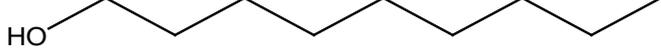
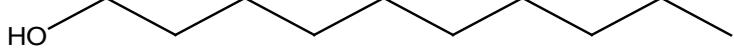
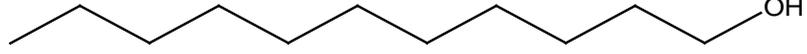
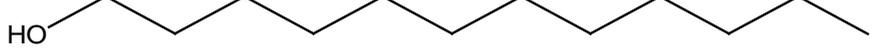
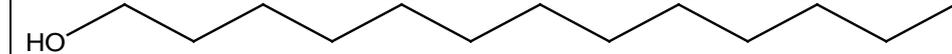
ANNEXE

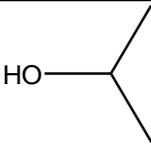
Présentation des données.

Tableau I – Toxicité (pCIC50) pour les alcools aliphatiques et les amines.

Tableau II: Toxicité (pCL50) vis-à-vis du vairoon de dérivés benzéniques.

Tableau I – Toxicité (CIC50) pour les alcools aliphatiques et les amines

N°	Composés (IUPAC)	Structure	N° CAS	pCIC50
1	Méthanol		67-56-1	-2.77
2	Ethanol		64-17-5	-2.41
3	propan-1-ol		71-23-8	-1.84
4	butan-1-ol		71-36-3	-1.52
5	pentan-1-ol		71-41-0	-1.12
6	hexan-1-ol		111-27-3	-0.47
7	heptan-1-ol		111-70-6	0.02
8	octan-1-ol		111-87-5	0.5
9	nonan-1-ol		143-08-8	0.77
10	décan-1-ol		112-30-1	1.1
11	undécan-1-ol		112-42-5	1.87
12	dodecan-1-ol		112-53-8	2.07
13	tridecan-1-ol		112-70-9	2.28

14	propan-2-ol		67-63-0	-1.99
15	pentan-2-ol	OH	6032-29-7	-1.25
16	pentan-3-ol	OH	584-02-1	-1.33
17	2-méthyl-1-butanol	OH	137-32-6	-1.13
18	3-méthyl-1-butanol	HO	123-51-6	-1.13
19	3-méthyl-2-butanol	OH	598-75-4	-1.08
20	(tert) pentanol	HO	75-85-4	-1.27
21	(neo) pentanol	HO	75-84-3	-0.96
22	1-propylamine	H ₂ N	107-10-8	-0.85
23	1-butylamine	H ₂ N	109-73-9	-0.7

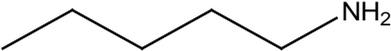
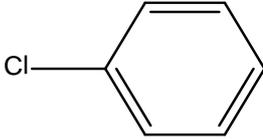
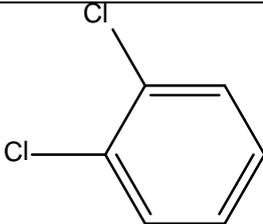
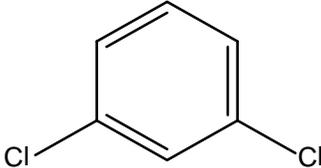
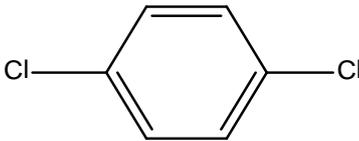
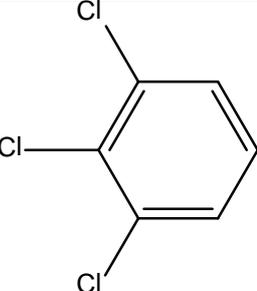
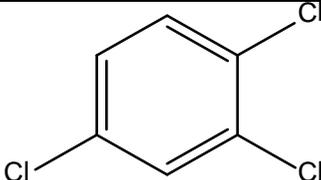
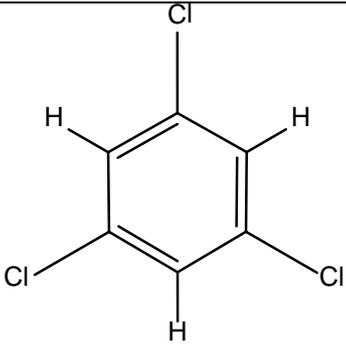
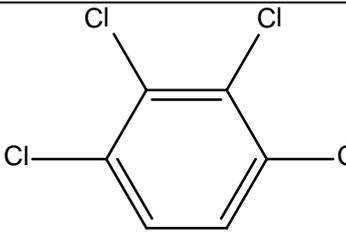
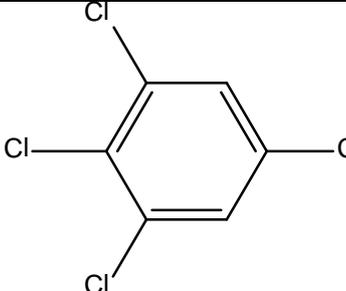
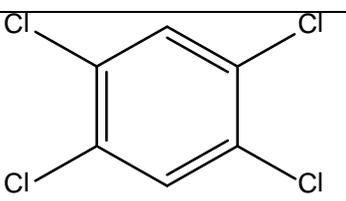
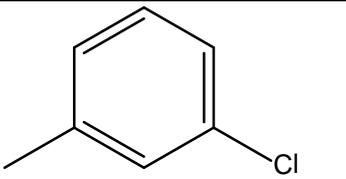
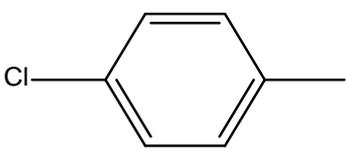
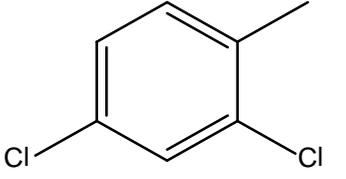
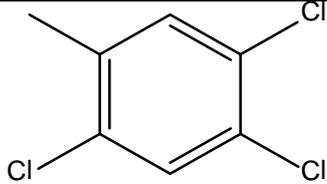
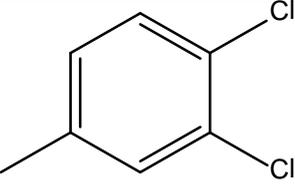
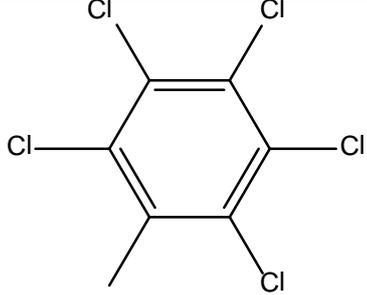
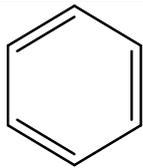
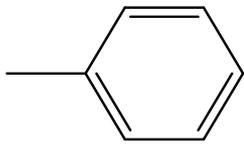
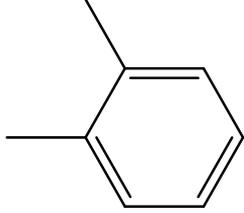
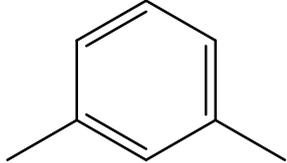
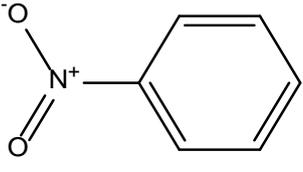
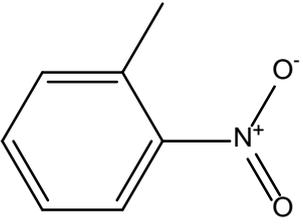
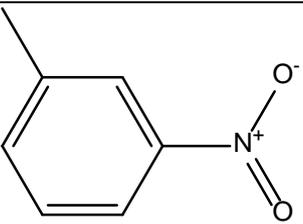
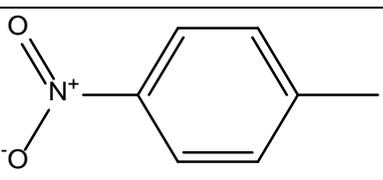
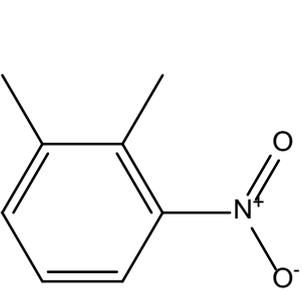
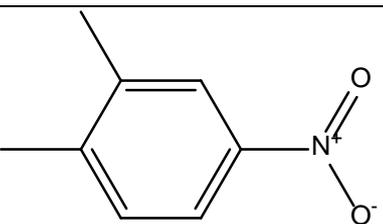
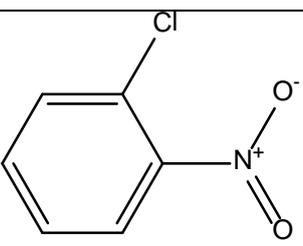
24	1-amylamine		110-58-7	-0.61
25	1-hexylamine	H ₂ N	111-26-2	-0.34
26	1-heptylamine	NH ₂	111-68-2	0.1
27	1-octylamine	H ₂ N	111-86-4	0.51
28	1-nonylamine	H ₂ N	112-20-9	1.59
29	1-décylamine	H ₂ N	2016-57-0	1.95
30	1-undécylamine	NH ₂	7307-55-3	2.26

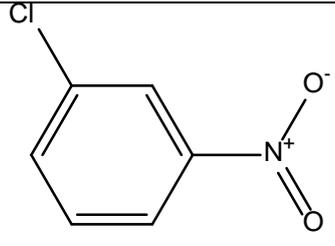
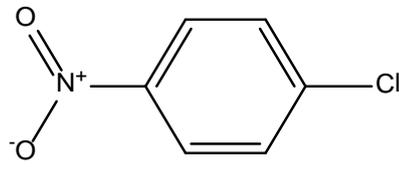
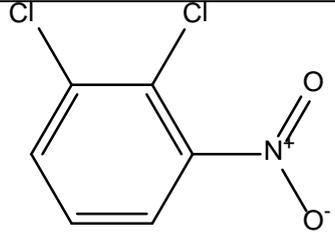
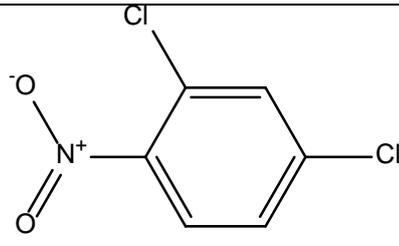
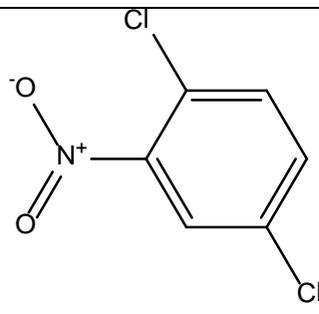
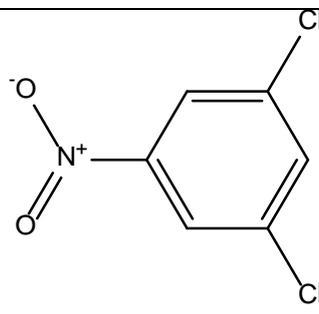
Tableau II: Toxicité (pCL50) vis-à-vis du vairon de dérivés benzéniques

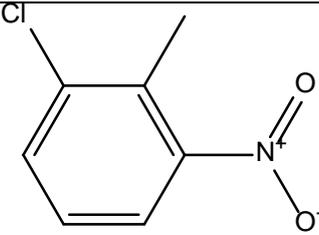
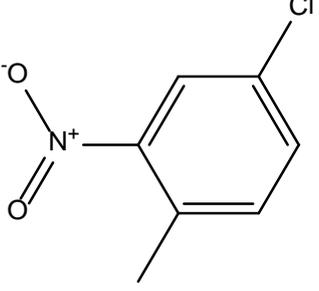
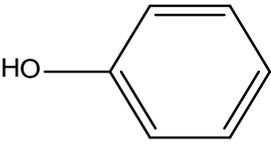
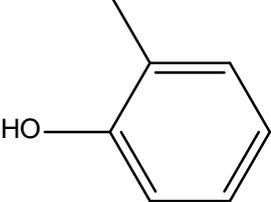
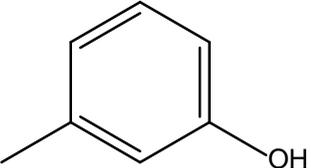
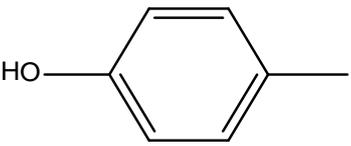
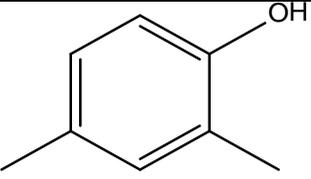
N°	Nom IUPAC	Formules	N° CAS	pCL50
1	Chlorobenzene		50717-45-8	-3.77
2	1.2-dichlorobenzene		95-50-1	-4.4
3	1.3-dichlorobenzene		541-73-1	-4.28
4	1.4-dichlorobenzene		106-46-7	-4.56
5	1.2.3-trichlorobenzene		87-61-6	-4.89
6	1.2.4-trichlorobenzene		120-82-1	-4.83

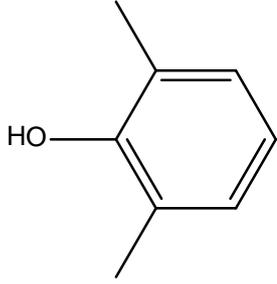
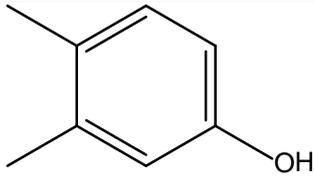
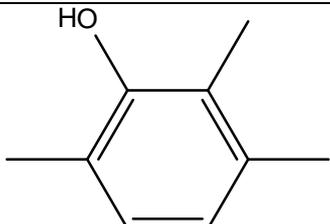
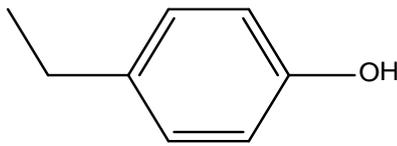
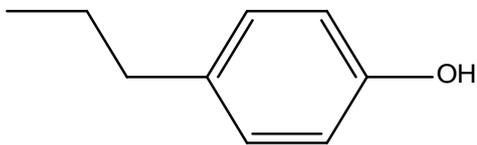
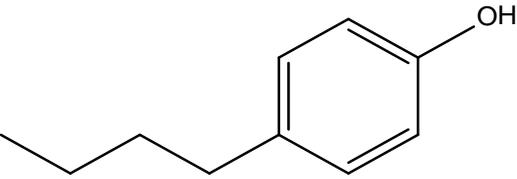
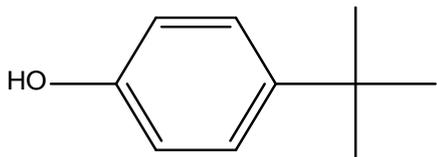
7	1.3.5-trichlorobenzene		108-70-3	-4.74
8	1.2.3.4-tetrachlorobenzene		39905-57-2	-5.35
9	1.2.3.5-tetrachlorobenzene		634-90-2	-5.43
10	1.2.4.5-tetrachlorobenzene		95-94-3	-5.85
11	1-chloro-3-methyl-benzene		108-41-8	-3.84
12	1-chloro-4-methyl-benzene		106-43-4	-4.33
13	2,4-dichloro-1-methyl-benzene		95-73-8	-4.54

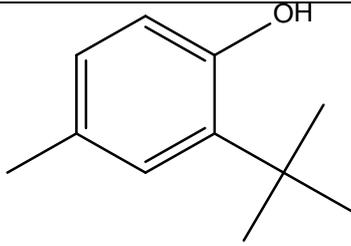
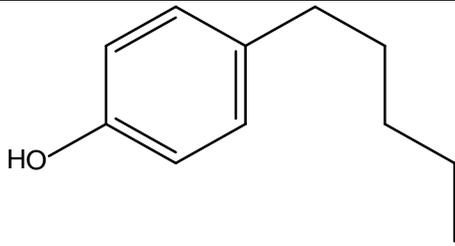
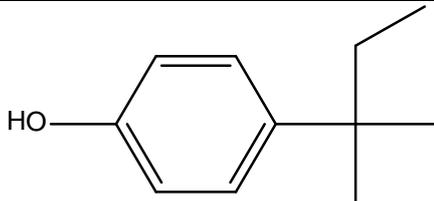
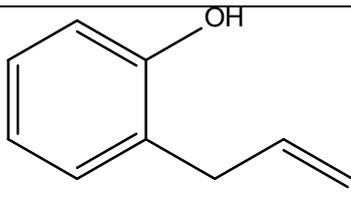
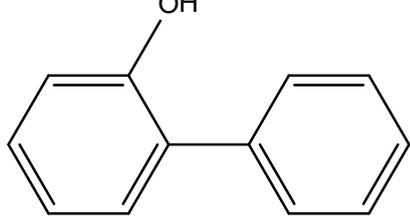
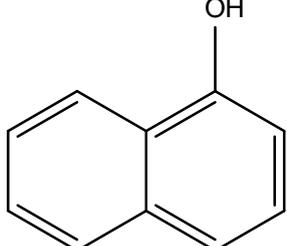
14	1.2.4-trichloro-5-methyl-benzene		6639-30-1	-5.06
15	1.2-dichloro-4-methyl-benzene		95-75-0	-4.6
16	1.2.3.4.5-pentachloro-6-methyl-benzene		69911-61-1	-6.15
17	Benzene		71-43-2	-3.09
18	Toluene		108-88-3	-3.13
19	1.2-xylene		95-47-6	-3.48
20	1.3-xylene		108-38-3	-3.45
21	1.4-xylene		106-42-3	-3.48

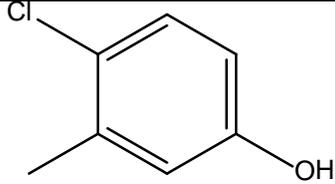
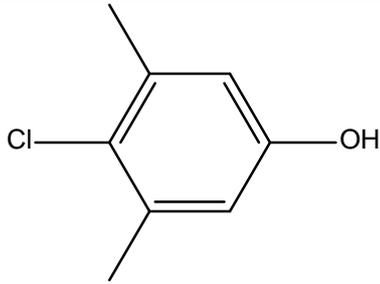
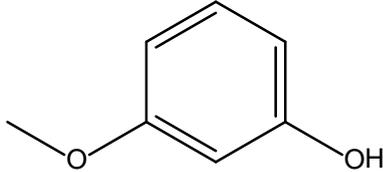
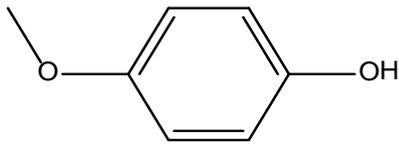
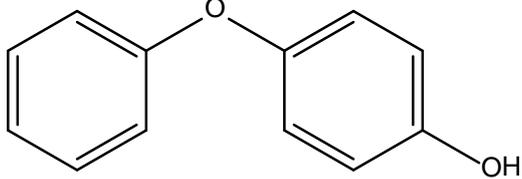
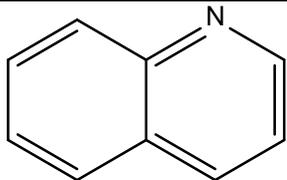
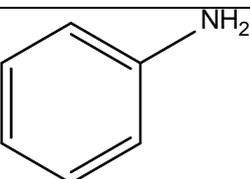
22	nitrobenzene		98-95-3	-2.97
23	1-methyl-2-nitro-benzene		88-72-2	-3.59
24	1-methyl-3-nitro-benzene		99-08-1	-3.65
25	1-methyl-4-nitro-benzene		99-99-0	-3.67
26	1,2-dimethyl-3-nitro-benzene		83-41-0	-4.39
27	1,2-dimethyl-4-nitro-benzene		99-51-4	-4.21
28	1-chloro-2-nitro-benzene		88-73-3	-3.72

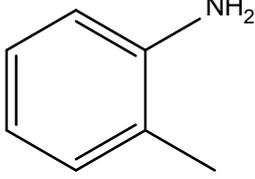
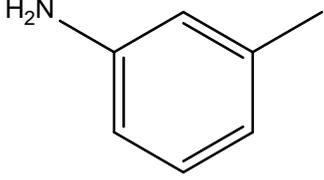
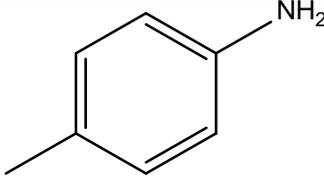
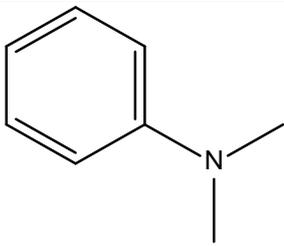
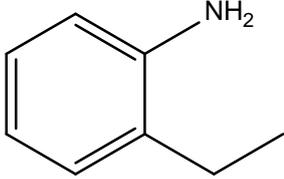
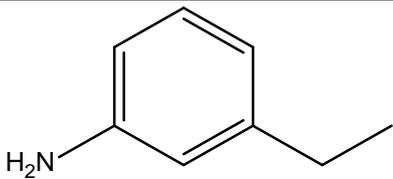
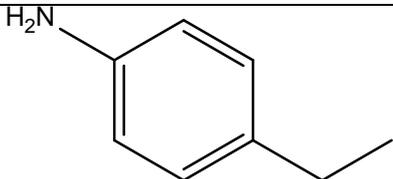
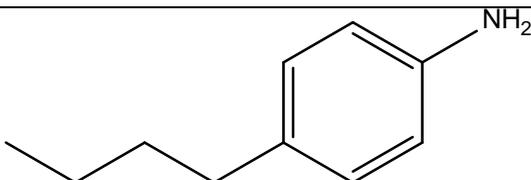
29	1-chloro-3-nitrobenzene		121-73-3	-4.01
30	1-chloro-4-nitro-benzene		100-00-5	-4.42
31	1,2-dichloro-3-nitro-benzene		3209-22-1	-4.66
32	2,4-dichloro-1-nitro-benzene		611-06-3	-4.46
33	1,4-dichloro-2-nitro-benzene		624-19-1	-4.59
34	1,3-dichloro-5-nitro-benzene		618-62-2	-4.58

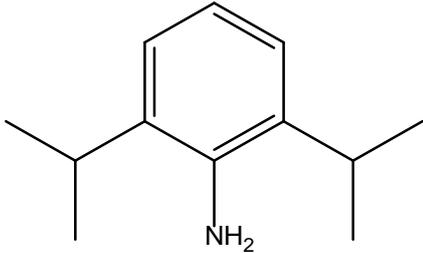
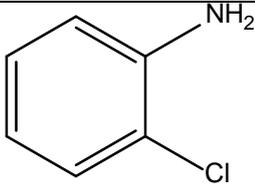
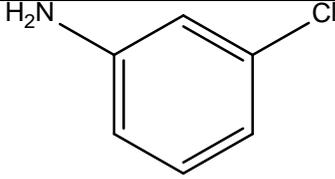
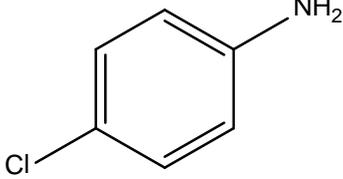
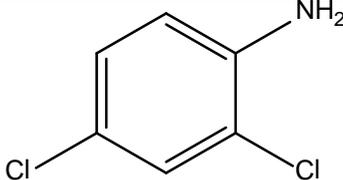
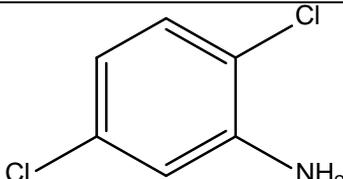
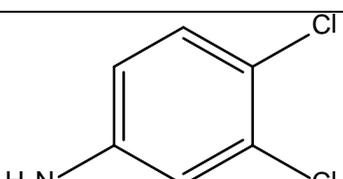
35	1-chloro-2-methyl-3-nitro-benzene		83-42-1	-4.52
36	4-chloro-1-methyl-2-nitro-benzene		89-59-8	-4.44
37	Phenol		8002-07-1	-3.45
38	2-methylphenol		95-48-7	-3.77
39	3-methylphenol		3019-89-4	-3.48
40	4-methylphenol		72269-62-9	-3.74
41	2,4-dimethylphenol		105-67-9	-3.86

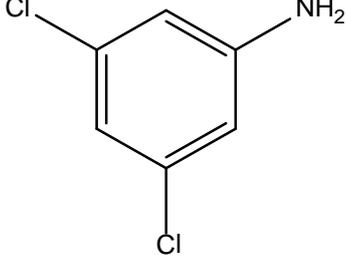
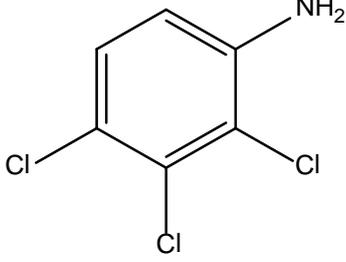
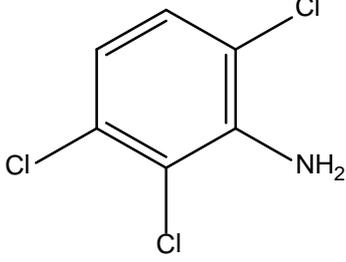
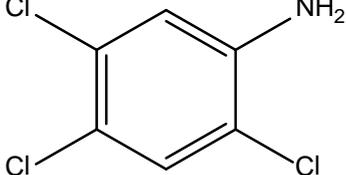
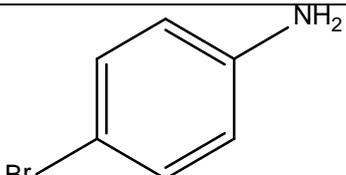
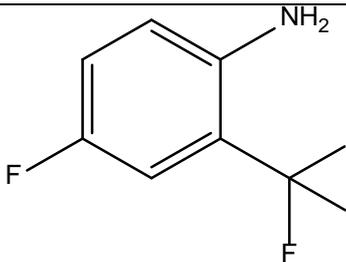
42	2.6-dimethylphenol		576-26-1	-3.75
43	3.4-dimethylphenol		95-65-8	-3.92
44	2.3.6-trimethylphenol		50356-13-3	-4.21
45	4-ethylphenol		19277-91-9	-4.07
46	4-propylphenol		645-56-7	-4.09
47	4-butylphenol		1638-22-8	-4.47
48	4-tert-butylphenol		98-54-4	-4.46

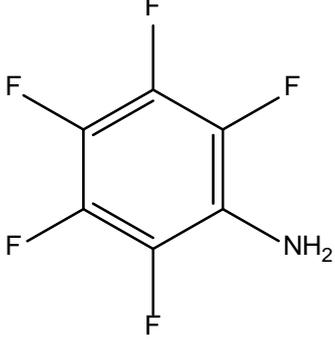
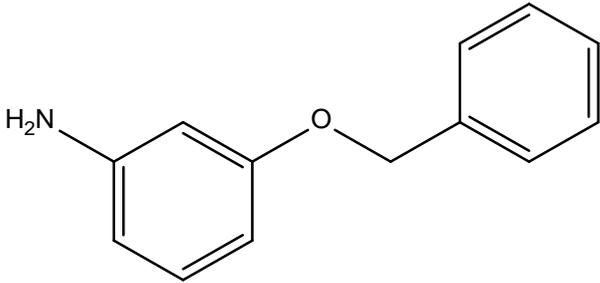
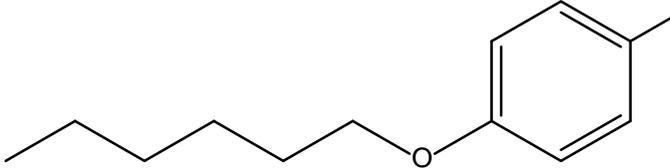
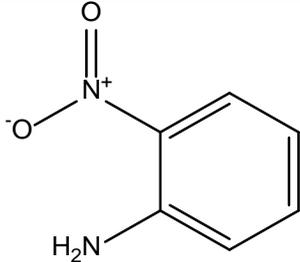
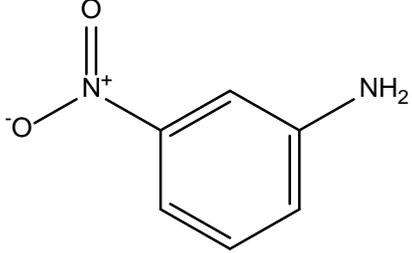
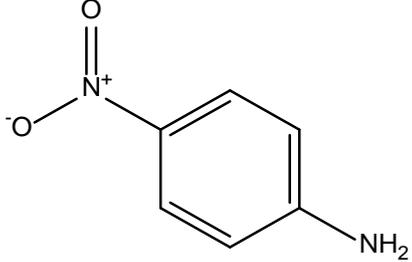
49	4-methyl-2-tert-butyl-phenol		29959-28-2	-4.9
50	4-pentylphenol		65916-15-6	-5.12
51	4-(2-methylbutan-2-yl)phenol		80-46-6	-4.81
52	2-prop-2-enylphenol		3383-08-2	-3.96
53	2-phenylphenol		90-43-7	-4.76
54	naphthaen-1-ol		90-15-3	-4.5
55	4-chlorophenol		106-48-9	-4.18

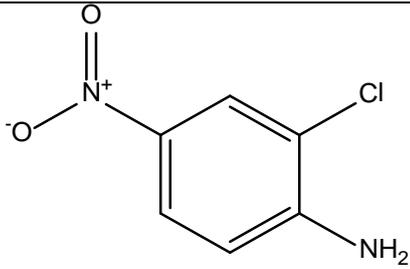
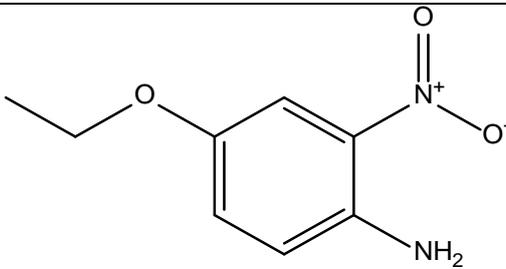
56	4-chloro-3-methyl-phenol		59-50-7	-4.33
57	4-chloro-3,5-dimethyl-phenol		88-04-0	-4.66
58	3-methoxyphenol		150-19-6	-3.22
59	4-methoxyphenol		150-76-5	-3.05
60	4-phenoxyphenol		634-67-3	-4.58
61	(2S)-2-amino-3-(3H-imidazol-4yl)propanoic		6027-02-7	-3.63
62	aniline		62-53-3	-2.91

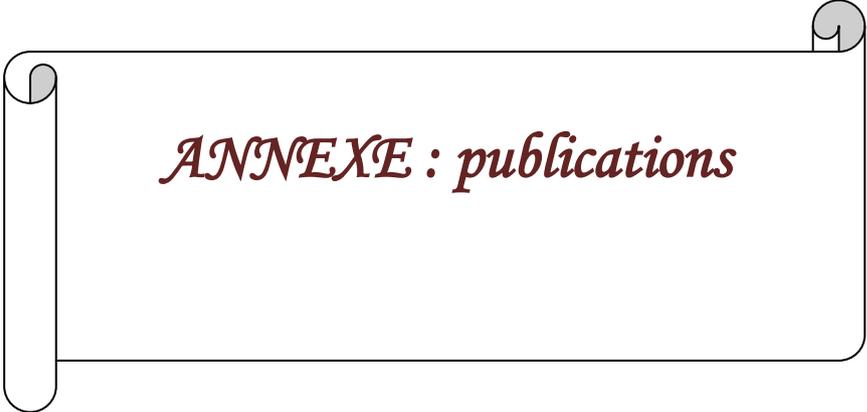
63	2-methylaniline		626-43-7	-3.12
64	3-methylaniline		95-53-4	-3.47
65	4-methylaniline		106-49-0	-3.72
66	N,N-dimethylaniline		58888-49-6	-3.33
67	2-ethylaniline		578-54-1	-3.21
68	3-ethylaniline		587-02-0	-3.65
69	4-ethylaniline		589-16-2	-3.52
70	4-butylaniline		104-13-2	-4.16

71	2,6-dipropan-2-ylaniline		24544-04-5	-4.06
72	2-chloroaniline		95-51-2	-4.31
73	3-chloroaniline		141-85-5	-3.98
74	4-chloroaniline		4084-48-4	-3.67
75	2,4-dichloroaniline		554-00-7	-4.41
76	2,5-dichloroaniline		95-82-9	-4.99
77	3,4-dichloroaniline		616-86-4	-4.39

78	3,5-dichloroaniline		95-53-4	-4.62
79	2,3,4-trichloroaniline		831-82-3	-5.15
80	2,3,6-trichloroaniline		88963-39-7	-4.73
81	2,4,5-trichloroaniline		636-30-6	-4.92
82	4-bromoaniline		89-61-2	-3.56
83	4-fluoro-3-(trifluoromethyl)aniline		2357-47-3	-3.77
84	4-fluoro-2-(trifluoromethyl)aniline		393-39-5	-3.78

85	2,3,4,5,6-pentafluoroaniline		771-60-8	-3.69
86	3-phenylmethoxyaniline		1484-26-0	-4.34
87	4-hexoxyaniline		634-66-2	-4.78
88	2-nitroaniline		88-74-4	-4.15
89	3-nitroaniline		99-09-2	-3.24
90	4-nitroaniline		66827-74-5	-3.23

91	2-chloro-4-nitro-aniline		121-87-9	-3.93
92	4-ethoxy-2-nitro-aniline		95-76-1	-3.85



ANNEXE : publications

Inhibition of *Tetrahymena pyriformis* growth by Aliphatic Alcohols and Amines: a QSAR Study

Nadia Ziani, Khadidja Amirat & Djelloul Messadi

Laboratoire de Sécurité Environnementale et Alimentaire (LASEA),
Université Badji Mokhtar – Annaba, 23 000, Annaba, Algérie

Soumis le : 07.06.2013

Révisé le : 30.04.2014

Accepté le : 04.06.2014

ملخص

أنجزت دراسة العلاقة بين الكمية البنيوية والنشاط (QSAR) لتحقيق التسمم النسبي لمزيج مكون من 21 كحول (ذات سلاسل خطية و متفرعة) و 9 أمينات اليقاتية عادية للتعبير على تركيز المعيقات بنسبة لنمو 50% (IGC_{50}) (*Tetrahymena pyriformis*). طريقة الانحدار الخطي البسيط اعتمدت على الموصفات الجزئية النظرية (هندسية), ثلاثية الأبعاد 3D, المتحصل عليها اعتمادا على برنامج DRAGON, والواصفات المختلفة log P المحسوبة. الصلابة و القدرة التنبؤية للنموذج تم التحقق منها اعتمادا على التصديق الداخلي (تصديق متقاطع) (Loo, LMO, bootstrap) و كذلك التصديق الخارجي Clog P ظهر بأنه أحسن واصف لنمذجة التسممية, و الذي يمكن تعويضه بالواصف الهندسي ADDD, بدون تغير ملحوظ للمعايير الإحصائية.

الكلمات المفتاحية: كحول وأمينات-تسمم مائي QSAR - الوصفات الهندسية و خصائص مائية - الانحدار الخطي البسيط.

Résumé

Une étude Relation Quantitative Structure- Activité (QSAR) a été réalisée pour évaluer la toxicité relative d'un mélange composé de 21 alcools (à chaînes linéaires et ramifiées) et 9 amines aliphatiques normales, en terme de concentration d'inhibition 50% de la croissance (IGC_{50}) de *Tetrahymena pyriformis*. L'approche par régression linéaire simple est basée sur des descripteurs moléculaires théoriques (géométriques) 3D obtenus à l'aide du logiciel DRAGON et différents descripteurs logP calculés. La robustesse et la capacité prédictive des modèles ont été vérifiées à l'aide de statistiques de validations internes (validations croisées LOO et LMO ; bootstrap) et externe. Clog s'est avéré le meilleur descripteur pour la modélisation de la grandeur d'intérêt considérée. Il peut être remplacé par le descripteur géométrique ADDD sans variations appréciables des paramètres statistiques.

Mots clés: Alcools et Amines – Toxicité aquatique – QSAR – Descripteurs géométriques et caractère hydrophobe – Régression linéaire simple.

Abstract

A Quantitative Structure- Activity Relationship (QSAR) study was undertaken to evaluate the relative toxicity of a mixed series of 21 (linear and branched-chain) alcohols and 9 normal aliphatic amines in term of the 50% inhibitory growth concentration (IGC_{50}) of *Tetrahymena pyriformis*. The applied simple linear regression approach is based on theoretical 3D (geometrical) molecular descriptors from DRAGON package, and some calculated logP descriptors. The robustness and the predictive performance of the models were verified using both internal (cross-validation by LOO and LMO; bootstrap) and external statistical validations. ClogP turned out to be the best descriptor to model the considered endpoint. It may be interchanged with geometrical descriptor ADDD without relevant variations in the statistical parameters.

Keywords: Alcohols and Amines –Aquatic toxicity – QSAR – Geometrical descriptors and hydrophobic character – Simple linear regression.

Auteur correspondant : d_messadi@yahoo.fr

1. INTRODUCTION

The impact of the potential hazard of untested chemicals, a challenge confronting national and international regulatory agencies [1-4] can be measured by experimental investigations, but this approach is both quite expensive and time consuming [5]. An alternative is to rely on QSAR (Quantitative Structure-Activity Relationships) models that describe a mathematical relationship between the structural features of a set of chemicals and the particular activity associated with them [6,7].

Several QSAR models predicting acute chemical toxicity for aquatic environment have been published [8-12]. They are based mainly on the logarithm of the octanol-water coefficient (logP, also referred to as logKow) as this hydrophobicity term reproduces the ability of a substance to enter cells through the lipid membranes and indicates both the toxicant uptake and baseline toxicity.

Albeit the number of compounds with a measured value for the logP was estimated to be 30 000 [13], which seems at a first glance to be high, this is negligible compared to the rapidly increasing number of compounds for which logP values are desired but missing. Furthermore, the experimental determination is tedious, time-consuming and demands a high purity of the solute [14]; none of these preconditions are compatible with high-throughput techniques, there is, therefore, an ongoing interest in methods for the prediction of logP values.

Over recent decades various approaches (fragmental, atom-based, conformation – dependent methods) [15-18] have been developed that are mostly implemented and available as computer programs. However, even in these calculations it is not uncommon to have differences of several order of magnitude [19, 20].

For these reasons logP cannot be considered a univocal descriptor, which brought different authors [19-24] to propose toxicity models based exclusively on other structural theoretical molecular descriptors.

The present paper proposes predictive simple linear regression QSAR models to evaluate the relative toxicity of organic chemicals, in terms (of the logarithm of the inverse) of the 50% inhibitory growth concentration (IGC50) of

Tetrahymena pyriformis. Models based on different kinds of logP (calculated values for AlogP, MlogP and ClogP), are compared to the optimal model constructed using a single 3D (geometrical) descriptor calculated from the chemical structure alone.

2. METHODS

2.1 Experimental Data

Two different toxicants were studied: a set of 21 (linear and branched-chain) alcohols and 9 normal aliphatic amines, selected to reflect diversity in chain length and branching.

These toxicants, which are both nonionic and nonreactive, inhibit the growth concentration of *Tetrahymena pyriformis* the most tested common freshwater hymenostome ciliate, which approximately measures 50 μm in length and 30 μm in width [25].

The ciliates were grown in axenic culture with population density being measured spectrophotometrically as optical density (absorbance) at 540 nm following 48h of incubation.

The set of experimental data was taken from Schultz [26].

2.2 Estimation of octanol /water partition coefficient

ClogP (\equiv *calculated logP*) [17]

The software version of the system developed by Hansch and Leo [27], using the Rekker's additive scheme [28], is known as ClogP (or calculated logP). It is based on different fragmental constants and correction terms. Fragment constants were derived from solutes where the fragment occurs in isolation. Furthermore, the bonding environment was taken into account (alkyl, benzyl, vinyl, styryl, and aromatic neighbors) resulting in five values per fragment. If a fragment in combination with the bonding environment was missing but at least two values for the same fragment with different neighbors could be found an interpolation was attempted to derive the missing data. The correction factors have been calculated from the corrections required for the specific interactions being modelled. For instance, the interaction of the two hydroxyl groups in diethylene glycol increases the logP value by 0.85 compared with two hydroxyl

groups that do not interact. This value is then taken as the correction term for a two-neighbored hydroxyl group [29].

The decomposition of the molecular structure into fragments is performed by using a unique and simple set of rules, thus obtaining a unique solution; the fragments are either atoms or polyatomic groups.

AlogP (\equiv Ghose and Crippen model based on atomic increment system) [30]

Several models have been published where the fragments are defined on a purely atomic level. This simplifies both the recognition of fragments and the calculation, as correction substructures are not applied (see Eq.(1)).

The most frequently used atomic increment system, AlogP, was developed by Ghose and Crippen [31]. Atoms are classified by their neighboring environment and carbon atoms additionally by their hybridization.

Estimated logP for any compound is given by:

$$\text{AlogP} = \sum_i n_i a_i \quad (1)$$

where n_i is the occurrence of the i th atom type and a_i is the corresponding hydrophobicity constant.

The AlogP model implemented in DRAGON has been evaluated on a set of 2648 compounds with known experimental logP taken from the NCI open Data Base. The resulted correlation coefficient r is 0.915.

MlogP (\equiv Moriguchi model based on structural parameters)

This is a model described by a regression equation based on 13 structural parameters [32, 33].

The regression coefficients have been evaluated by a training set of 1230 organic molecules including general aliphatic aromatic and heterocyclic compounds containing the following atoms: C, H, N, O, S, P, F, Cl, Br, I [30]. The statistical parameters of the model are $r = 0.952$; $SE = 0.422$; $F_0(13;1216) = 900.4$

2.3 Geometrical Descriptors Generation

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program [34] and pre optimized using MM⁺ molecular mechanics method (Polack-Ribiere

algorithm). The final geometries of the minimum energy conformation were obtained by the semi empirical PM3 method at a restricted Hartree-Fock level with no configuration interaction applying a gradient norm limit of $0.01 \text{ kcal. \AA}^{-1} \cdot \text{mol}^{-1}$ as a stopping criterion.

The resulted geometries were used as input for the generation of (74) 3D- geometrical descriptors using the DRAGON software (version 5.3) [30].

Geometrical descriptors being defined from the three dimensional structure of the molecule, which involves the knowledge of the relative positions of the atoms in 3D space provide information and discrimination power also for similar molecular structures and molecule conformations.

2.4 Chemometric Methods

Models with one variable were performed by the software MOBYDYGS [35] using the Ordinary Least Square regression (OLS) method.

Population of 74 regression models corresponding to each of the 74 Geometry descriptors were ordered according to their decreasing internal predictive performance, verified by Q^2 , optimal non-logP model was then selected and compared to the three logP – based models.

The goodness of fit of the calculated models were assessed by means of the multiple determination coefficients, R^2 , and the standard deviation error in calculation (SDEC).

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

Cross validation techniques allow the assessment of internal predictivity (Q^2_{LMO} cross validation; bootstrap) in addition to the robustness of model (Q^2_{LOO} cross validation).

Cross validation methods consist in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out. This procedure is repeated for all compounds of the training set, obtaining a prediction for everyone. If each compound is taken away one at a time the cross validation procedure is called leave-one-out technique (LOO technique), otherwise

leave-more-out technique (LMO technique). An LOO or LMO correlation coefficient, generally indicated with Q^2 , is computed by evaluating the accuracy of these “test” compounds prediction.

$$Q^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS} \quad (3)$$

The “hat” of the variable y , as is the usual statistical notation, indicates that it is a predicted value of the studied property, and the sub index “i/i” indicates that the predicted values come from models built without the predicted compound.

TSS is the total sum of squares.

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

$$SDEP = \sqrt{PRESS/n} \quad (4)$$

A value $Q^2 > 0.5$ is generally regarded as a good result and $Q^2 > 0.9$ as excellent [36, 37].

However, studies [38, 39] have indicated that while Q^2 is a necessary condition for high predictive power a model, is not sufficient.

To avoid overestimating the predictive power of the model LMO procedure (repeated 5000 times, with 5 objects left out at each step) was also performed ($Q^2_{L(5)O}$).

In bootstrap validation technique K n-dimensional groups are generated by a randomly repeated selection of n -objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times for each validated model.

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^2_{ext} , which is defined as

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2 / n_{tr}} = 1 - \frac{PRESS / n_{ext}}{TSS / n_{tr}} \quad (5)$$

Here 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.

The data set randomly was divided into a training set (20 objects) used to develop the QSAR models and a validation set (10 objects), used only for statistical external validation.

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:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} (y_i - \bar{y})^2} \quad (6)$$

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

3. RESULTS AND DISCUSSION

The best one dimensional non-logP model was obtained using the average distance-distance degree (ADDD) index. It encodes information on the molecular folding [40, 41] information about molecular diffusion easiness through biological barriers like membranes. Distance/distance matrices, denoted as D/D , were defined as quotient matrices in terms of geometric r_{ij} distances and topological distance d_{ij} :

$$[G/D]_{ij} = \begin{cases} \frac{r_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i=j \end{cases} \quad (7)$$

The row sums of these matrices contain information on the molecular folding; in effect, in highly folded structures, they tend to be relatively small as the inter-atomic distances are small while the topological distances increase as the size of the structure increases.

Therefore, the average row sum is a molecular invariant called average distance-distance degree, that is:

$$ADDD = \frac{1}{A} \sum_{i=1}^A \sum_{j=1, j \neq i}^A \frac{r_{ij}}{d_{ij}} \quad (8)$$

A being the number of molecule atoms. Table 1 lists the CAS number, $-\log IGC_{50}$, $AlogP$, $MlogP$, $ClogP$ and ADDD values of the selected aliphatic alcohols and amines.

Table 1. Relative toxicity and molecular descriptors data for the selected aliphatic alcohols and amines

Compound	CAS number ^(a)	-logIGC ₅₀	AlogP	MlogP	ClogP	ADDD
Methanol	67-56-1	-2.77	-0.358	-0.814	-0.764	4.964
Ethanol	64-17-5	-2.41	0.009	-0.172	-0.235	8.108
1-propanol	71-23-8	-1.84	0.515	0.347	0.294	11.098
1-pentanol	71-41-0	-1.12	1.427	1.209	1.352	17.194
1-hexanol	111-27-3	-0.47	1.883	1.587	1.881	20.305
1-heptanol	111-70-6	0.02	2.339	1.940	2.410	23.451
1-nonanol	143-08-8	0.77	3.252	2.591	3.468	29.859
1-decanol	112-30-1	1.1	3.708	2.894	3.997	33.118
1-dodecanol	112-53-8	2.07	4.620	3.467	5.055	39.716
1-tridecanol	112-70-9	2.28	5.077	3.739	5.584	43.053
2-propanol	67-63-0	-1.99	0.368	0.347	0.074	10.987
2-methyl-1-butanol	137-32-6	-1.13	1.290	1.209	1.222	16.810
3-methyl-1-butanol	123-51-6	-1.13	1.223	1.209	1.222	16.833
3-methyl-2-butanol	598-75-4	-1.08	1.211	1.209	1.002	16.622
(tert)pentanol	75-85-4	-1.27	1.097	1.209	1.002	16.640
1-propylamine	107-10-8	-0.85	0.225	0.347	0.394	11.968
1-hexylamine	11-26-2	-0.34	1.594	1.587	1.981	21.226
1-heptylamine	111-68-2	0.1	2.050	1.940	2.510	24.393
1-octylamine	111-86-4	0.51	2.506	2.274	3.039	27.602
1-undecylamine	7307-55-3	2.26	3.875	3.186	4.626	37.408
1-butanol*	71-36-3	-1.52	0.971	0.800	0.823	14.138
1-otanol*	111-87-5	0.5	2.796	2.274	2.939	26.640
1-undecanol*	112-42-5	1.87	4.164	3.186	4.526	36.401
2-pentanol*	6032-29-7	-1.25	1.348	1.209	1.132	16.888
3-pentanol*	584-02-1	-1.33	1.416	1.209	1.132	16.908
(neo) pentanol*	75-84-3	-0.96	1.108	1.209	1.092	16.710
1-butylamine*	109-73-9	-0.7	0.681	0.800	0.923	15.015
1-amylamine*	110-58-7	-0.61	1.137	1.209	1.452	18.101
1-nonylamine*	112-20-9	1.59	2.962	2.591	3.568	30.837
1-decylamine*	2016-57-0	1.95	3.418	2.894	4.097	34.110

(a): Chemical Abstract Services registry number ; (*): validation set compound.

Intercepts (β_0) and slopes (β_1) of the calculated one dimensional models are shown in table 2

Table 2. Coefficients for the ordinary least squares calculated models.

X	AlogP	MlogP	ClogP	ADDD
β_0	-2.144 (± 0.139)	-2.230 (± 0.134)	-1.998 (± 0.088)	-3.333 (± 0.134)
β_1	0.939 (± 0.057)	1.191 (± 0.068)	0.815 (± 0.033)	0.138 (± 0.006)

Relevant statistical parameters reported in table 3 below clearly show the difference in fitting and prediction performances for the selected logP descriptors, ClogP appear as the best ones.

Another remarkable fact observed is that ClogP and ADDD theoretical molecular descriptor can be interchanged without relevant variations in the statistical results.

Table 3. Summary statistics for the one dimensional calculated models.

X	R ²	Q ²	Q ² _{L(5)O}	Q ² _{boot}	Q ² _{ext}	SDEC	SDEP	SDEP _{ext}	F _(p=0.000)	SE
Alog P	0.9372	0.9188	0.9165	0.9063	0.8584	0.365	0.415	0.554	268.4	0.385
Mlog P	0.9444	0.9291	0.9270	0.9160	0.8998	0.344	0.388	0.461	305.52	0.362
Clog P	0.9713	0.9630	0.9621	0.9580	0.9352	0.247	0.280	0.371	610.02	0.260
ADDD	0.9711	0.9624	0.9616	0.9566	0.9326	0.248	0.283	0.378	603.82	0.261

Carbö –Dorca *et al.* [42] reported a QSAR study where the same data was examined, these authors constructed a predictive model using, as a molecular descriptor, the expectation value of the inter electronic repulsion energy operator presented as a kind of quantum self-similarity measure (QS-SM). The correlation results reached R² = 0.9240, Q² = 0.9090 and SE = 0.415, which are inferior than the present approach.

The value of R² attests the good fitting performances of the model. In general, the larger the magnitude of the F ratio, the better the model predicts the property values in the training set. The large F ratio of 603.82 indicates that the model does an excellent job of predicting the -log IGC50 values. The model is robust, the difference between R² and Q² is small (<1%). Figure 1 shows a plot contrasting experimental and cross-validated -logIGC₅₀. The point dispersion is small, although there is one point a little bit far away from the rest (1-propylamine). SDEP is similar to SDEC, so this model has internal predictivity not so dissimilar from fitting power.

The model demonstrates a very good stability in internal validation (difference between Q² and Q²_{L(5)O} is 0.33%), while bootstrapping confirms the internal predictivity and stability of the model.

Though small sized the data set underwent statistical validation by preliminary random splitting of the chemicals into training (20 chemicals) and validation (10 chemicals) sets. The small size of the published experimental

data set [26] did not allow a more drastic splitting. The information obtained by Q²_{ext} is

somewhat optimistic. In fact with small data sets (20-30 chemicals), completely new chemicals external predictivity can only be verified *a posteriori*, case -by-case.

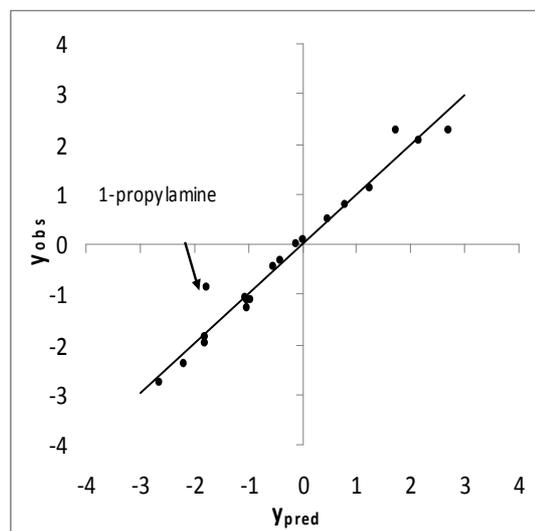


Figure-1 Experimental versus cross-validation activity for the training set objects.

4. CONCLUSION

From the results and discussion above we conclude that:

1. Among the logP descriptors selected to model the inhibition of *Tetrahymena pyriformis*

growth by aliphatic alcohols and amines, Clog P is the best.

2. Geometrical descriptor ADDD (Average Distance/Distance Degree), a theoretical molecular descriptor, and Clog P can be interchanged without relevant variations in the statistical results.

3. The non- logP model obtained in this study has very good fitting performances, is robust and with acceptable predictive power. The internal validation parameters (Q^2_{LOO} , $Q^2_{L(5)O}$ and bootstrap) are similar to the fitting parameters.

Notice that chemical 19 (1-propyl amine) with cross validated standardized residuals (not reported here) greater than 3 standard deviation units, is an heavy outlier in all the models considered in this paper.

REFERENCES

- [1] Zeeman M., Aver C.M., Clements R.G., Nabholz J.V. & Boethling R.S., 1995. U.S. EPA Regulatory Perspectives on the use of QSAR for new and existing chemical evaluations SAR QSAR, *Environmental Research*, Vol. 3(3),179-201.
- [2] Walker J.D., 2003. Applications of QSARs in toxicology: a US Government perspective, *Journal of Molecular Structure - Theochem*, Vol. 622(1-2), 167-184.
- [3] Bradbury S.P., Russon C.L., Ankley G.T., Schultz T.W. & Walker J.D., 2003. Overview of data and conceptual approaches for derivation of Quantitative Structure –Activity Relationships, for ecotoxicological effects of organic chemicals, *Environmental Toxicology & Chemistry*, Vol. 22 (8), 1789-1798.
- [4] European Commission. White Paper on a strategy for a future Community Policy for Chemicals., 2001.[http:// europa .eu.int / comm / enterprise / reach /](http://europa.eu.int/comm/enterprise/reach/).
- [5] Toussaint M.W., Shedd T.R., Van der Schalie W.H. & Leather G.R., 1995. A comparison of standard acute toxicity tests with rapid screening toxicity tests. *Environmental Toxicology & Chemistry*, Vol. 14(5), 907-915.
- [6] Kubinyi H., 2002. From Narcosis to Hyperspace: The History of QSAR, *Quantitative Structure.-Activity Relationships.*, Vol. 21(4), 348-356.
- [7] [http:// e c b . j r c . i t / Q S A R /](http://ecb.jrc.it/QSAR/).
- [8] Schultz T.W., Cronin M.T.D., Walker J.D. & Aptula A.O., 2003.Quantitative structure –activity relationships (QSARs) in toxicology: a historical perspective, *Journal of Molecular Structure –Theochem*, Vol.622(1-2), 1-22.
- [9] Posthumus R. & Slooff W., 2001. Implementation of QSARs in ecotoxicological risk assessments RIVM report 601516003.
- [10] Dearden J.C., 2002. Prediction of Environmental Toxicity and Fate Using Quantitative Structure –Activity Relationships (QSARs), *Journal of Brazilian Chemical Society*, Vol . 13 (6), 754-762.
- [11] Schultz T.W., Cronin M.T.D. & Netzeva T.I., 2003. The present status of QSAR in toxicology, *Journal of Molecular Structure -Theochem*, Vol. 622(1-2), 23-38.
- [12] Cronin M.T.D. & Dearden J.C., 1995. QSAR in toxicology .1. Prediction of Aquatic Toxicity, *Quantitative Structure.-Activity. Relationships.*, Vol.14(1), 1-7.
- [13] Mannhold R. & van de Waterbeemd H., 2001. Substructure and whole molecule approaches for calculating logP, *Journal of Computer- Aided Molecular Design*, Vol. 15(4), 337-354.
- [14] Mannhold R. & Rekker R.F., 2000. The hydrophobic fragmental constant approach for calculating log P in octanol/water and aliphatic hydrocarbon/water systems. *Perspectives in. Drug Discovery & Design*, Vol.18(1), 1-18.
- [15] Benfenati E., Gini G., Piclin N., Roncaglioni A. & Vari M.R ., 2003.Predicting log P of pesticides using different software, *Chemosphere*, Vol.53(9), 1155-1164.
- [16] Manhold R. & Petrauskas A., 2003. Substructure versus Whole-molecule Approaches for Calculating Log P, *QSAR & Combinatorial Science*, Vol. 22(4), 466-475.1
- [17] <http://clogP.pomona.edu/medchem/chem/clogP/index.html>.
- [18] Klopman G., Li J.K., Wang S. & Dimayuga M., 1994.Computer Automated log P calculations based on an extended group contribution approach, *Journal of Chemical. Information. Computer Sciences*, Vol.34(4), 752-781.
- [19] Kaiser K.L.E., 2003. The use of neural networks in QSARs for acute aquatic toxicological endpoints, *Journal of Molecular Structure. Theochem*, Vol .622(1-2), 85-95.
- [20] Papa E., Villa F. & Gramatica P., 2005. Statistically Validated QSARs Based on Theoretical Descriptors , for Modeling Aquatic Toxicity of Organic Chemicals in *Pemphales promelas* (Fathead Minnow), *Journal of Chemical Information & Modeling*, Vol.45(5), 1256-1266.
- [21] Roy K. & Ghosh G., 2009.QSTR with extended topochemical atom (ETA) indices. 12. QSAR for the toxicity of diverse aromatic compounds to *Tetrahymena pyriformis* using chemometric tools. *Chemosphere*, Vol. 77(7), 999-1009.
- [22] Zhao Y.H., Zhang X.J.,WEN Y., Sun F.T., Guo Z., Qin W.C., Qin H.W.,Xu J.L., Sheng L.X. & Abraham M.H., 2010. Toxicity of organic chemicals to *Tetrahymena pyriformis* : Effect of polarity and ionization on toxicity. *Chemosphere*, Vol. 79(1), 72-77.
- [23] Roy K. & Das R.N., 2010.QSTR with extended topochemical atom (ETA) indices.14. QSAR modeling of toxicity of aromatic aldehydes to *Tetrahymena pyriformis*. *Journal of Hazardous Materials*, Vol. 183(1-3), 913-922.
- [24] Bouaoune A., Lourici L., Haddag H. & Messadi D., 2012. Inhibition of Microbial Growth by anilines: A QSAR study, *Journal of Environmental Science and Engineering.*, A1, Vol. 1(5A), 663-671.
- [25] Hill D.L., 1972. The biochemistry and physiology of *Tetrahymena*. Academic Press, New York & London, 230p
- [26] Schultz T.W., Lin D.T., Wilke T.S. & Arnold L.M., 1990. Quantitative structure-activity relationships for the

- Tetrahymena pyriformis* population growth endpoint : a mechanism of action approach. In Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology W. Karcher, J. Devillers (Eds) Kluwer Academic Publishers. Dordrecht, Vol.1., 241-262.
- [27] Hansch C. & Leo A., 1979. Substituent Constants for Correlation Analysis in Chemistry and Biology. Wiley & Sons Inc. New York (NY), 339p.
- [28] Nys C.G. & Rekker R.F., 1973. Statistical Analysis of a series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. The introduction of Hydrophobic Fragmental Constants (f values), *Chimica. Therapeutica* Vol.8, 521-535.
- [29] Kleinöder T., Yan A & Spycher S., 2008. Estimation of Octanol/Water Partition coefficient ($\log P_{ow}$). In *Cheminformatics J. Gasteiger, T. Engel (Eds)*, WILEY-VCH GmbH & Co. kgA, Weinheim, 492-494.
- [30] Todeschini R., Consonni V., Mauri A. & Pavan M. 2005 DRAGON Software for the Calculation of Molecular Descriptors—version 5.3 for Windows, Talete s.r.l., Milano, Italy.
- [31] Ghose A.R. & Crippen G.M., 1986. Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed-Quantitative Structure -activity relationships .I-Partition Coefficients as a Measure of Hydrophobicity, *Journal of Computational Chemistry*, Vol. 7(4), 565-577.
- [32] Moriguchi I., Hirono S., Liu O., Nakagome I. & Matsushita Y., 1992. Simple Method of Calculating Octanol/Water Partition Coefficient, *Chemical & Pharmaceutical Bulletin*, Vol.40(1), 127-130.
- [33] Moriguchi I., Hirono S., Nakagome I. & Hirano H., 1994. Comparison of Reliability of LogP Values for Drugs Calculated by Several Methods, *Chemical & Pharmaceutical Bulletin*, Vol. 42(4), 976-978.
- [34] HYPERCHEM/CHEMPLUS ver.7.03 for Windows Autodesk Inc., Sausalito, CA, U.S.A. 2002.
- [35] MOBYDIGS – Models BY Descriptors In Genetic Selection – ver. 1.1 for Windows, Talete S.r.l., Milano, Italy.
- [36] Eriksson L., Jaworska J., Worth A., Cronin M Mc., Dowell R.M. & Gramatica P., 2003. Methods for Reliability, uncertainty assessment , and applicability evaluations of regression based and classification QSARs, *Environmental Health Perspectives*, Vol. 111(10),1361-1375.
- [37] Tropsha A., Gramatica P. & Grombar V.K., 2003.The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR & Combinatorial Science*, Vol. 22(1), 69-77.
- [38] Kubinyi H., Hamprecht F.A. & Mietzner T., 1998. Three- dimensional quantitative similarity-activity relationships (3D QSAR) from SEAL similarity matrices, *Journal of Medicinal Chemistry*, Vol.41(14), 2553-2564.
- [39] Golbraikh A. & Tropsha A., 2002.Beware of q^2 , *Journal of Molecular.Graphics*, Vol.20(4), 269-276.
- [40] Randic M., Kleiner A.F. & De Alba L.M., 1994. Distance Matrices, *Journal of Chemical. Information. Computer Sciences*, Vol. 34(2), 277-286.
- [41] Randic M. & Krilov G.,1999.On a characterization of the Folding of Proteins, *International Journal of Quantum Chemistry*, Vol. 75(6), 1017-1026.
- [42] Carbö-Dorca R., Robert D., Amat Li., Gironés X. & Besalu E., 2000. Molecular Quantum Similarity in QSAR and Drug Design. Springer-Verlag Berlin Heidelberg. Vol. 42, 123p.



Management of Environmental Quality: An International

Chemometric modeling to predict aquatic toxicity of benzene derivatives in

Pimephales Promelas

Nadia Ziani Khadidja Amirat Djelloul Messadi

Article information:

To cite this document:

Nadia Ziani Khadidja Amirat Djelloul Messadi , (2016), "Chemometric modeling to predict aquatic toxicity of benzene derivatives in Pimephales Promelas", Management of Environmental Quality: An International Journal, Vol. 27 Iss 3 pp. 299 - 312

Permanent link to this document:

<http://dx.doi.org/10.1108/MEQ-05-2015-0082>

Downloaded on: 05 April 2016, At: 08:30 (PT)

References: this document contains references to 38 other documents.

To copy this document: permissions@emeraldinsight.com

The fulltext of this document has been downloaded 9 times since 2016*

Users who downloaded this article also downloaded:

(2016), "Chemometric modeling to predict retention times for a large set of pesticides or toxicants using hybrid genetic algorithm/multiple linear regression approach", Management of Environmental Quality: An International Journal, Vol. 27 Iss 3 pp. 313-325 <http://dx.doi.org/10.1108/MEQ-05-2015-0080>

(2016), "Assessment and management of water resources in the watershed of the middle Seybouse (Northeast Algeria)", Management of Environmental Quality: An International Journal, Vol. 27 Iss 3 pp. 326-337 <http://dx.doi.org/10.1108/MEQ-04-2015-0053>



Access to this document was granted through an Emerald subscription provided by emerald-srm:393177 []

For Authors

If you would like to write for this, or any other Emerald publication, then please use our Emerald for Authors service information about how to choose which publication to write for and submission guidelines are available for all. Please visit www.emeraldinsight.com/authors for more information.

About Emerald www.emeraldinsight.com

Emerald is a global publisher linking research and practice to the benefit of society. The company manages a portfolio of more than 290 journals and over 2,350 books and book series volumes, as well as providing an extensive range of online products and additional customer resources and services.

Emerald is both COUNTER 4 and TRANSFER compliant. The organization is a partner of the Committee on Publication Ethics (COPE) and also works with Portico and the LOCKSS initiative for digital archive preservation.

*Related content and download information correct at time of download.

Chemometric modeling to predict aquatic toxicity of benzene derivatives in *Pimephales Promelas*

Predict aquatic
toxicity of
benzene
derivatives

299

Nadia Ziani, Khadidja Amirat and Djelloul Messadi
*Environmental and Food Safety Laboratory, Faculty of Science,
Badji Mokhtar University Annaba, Annaba, Algeria*

Received 11 May 2015
Revised 11 May 2015
Accepted 16 June 2015

Abstract

Purpose – The purpose of this paper is to predict the aquatic toxicity (LC50) of 92 substituted benzenes derivatives in *Pimephales promelas*.

Design/methodology/approach – Quantitative structure-activity relationship analysis was performed on a series of 92 substituted benzenes derivatives using multiple linear regression (MLR), artificial neural network (ANN) and support vector machines (SVM) methods, which correlate aquatic toxicity (LC50) values of these chemicals to their structural descriptors. At first, the entire data set was split according to Kennard and Stone algorithm into a training set (74 chemicals) and a test set (18 chemical) for statistical external validation.

Findings – Models with six descriptors were developed using as independent variables theoretical descriptors derived from Dragon software when applying genetic algorithm – variable subset selection procedure.

Originality/value – The values of Q2 and RMSE in internal validation for MLR, SVM, and ANN model were: (0.8829; 0.225), (0.8882; 0.222); (0.8980; 0.214), respectively and also for external validation were: (0.9538; 0.141); (0.947; 0.146); (0.9564; 0.146). The statistical parameters obtained for the three approaches are very similar, which confirm that our six parameters model is stable, robust and significant.

Keywords Support vector machine, Artificial neural networks, Quantitative structure-activity relationship, Benzene derivatives, Aquatic toxicity, Multiple linear regression

Paper type Research paper

1. Introduction

The environment is regularly exposed to chemical substances such as substituted benzenes through their use in industrial processes. Information on aquatic toxicity is required in order to assess hazard and risk of chemical substances to marine and freshwater organisms living in the water (Netzeva *et al.*, 2008). The measurement of toxic effects is an expensive and time consuming process. Therefore, prediction of toxicity from the structure of compounds can help in control of environmental pollutions and also in designing the new beneficial compounds such as pharmaceutical substances with the minimum effects of toxicity. These studies may be also useful in interpreting the mechanisms of toxicity (Jalali-Heravi and Kyani, 2008; Zarei *et al.*, 2014)

Structure-toxicity models exist at the intersection of biology, chemistry and statistics. The connection of these three subjects has permitted the development of structure-activity relationships as an accepted sub-discipline in toxicology. The next decade will see an increased using of quantitative structure-activity relationships (QSARs) to predict toxicity for new and existing chemicals. Much of the focus will be on their application to reduce or replace animal use in toxicological testing for the regulation of existing chemicals (e.g. in the REACH legislation) (Worth *et al.*, 2007). The official birth date of QSAR is considered to



be 1962, when Hansch *et al.* (1962) published a paper which showed a correlation between biological activity and octanol-water partition coefficient (Bordbar *et al.*, 2013).

QSAR is mathematical models of activity in terms of structural descriptors. The QSAR model is useful for understanding the factors controlling activity and for designing new potent compounds (Hansch *et al.*, 1962). The main problems encountered in this kind of research are still the description of the molecular structure using appropriate molecular descriptors and selection of suitable modeling methods. At present, many types of molecular descriptors such as topological indices and quantum chemical parameters have been proposed to describe the structural features of molecules (Karelson, 2000; Devillers and Balaban, 1999; Todeschini and Consonni, 2000). Many different chemometrics methods, such as multiple linear regression (MLR), Partial least squares regression, different types of neural networks (NN), genetic algorithms (GAs), and support vector machine (SVM) can be employed to derive correlation models between the molecular structures and properties (Darnag *et al.*, 2014).

The success of any QSAR model depends on the accuracy of input data, selection of appropriate descriptors that represent variations in structural property of molecules quantitatively and statistical tools and validation of the developed model (Tong *et al.*, 2005; He and Jurs, 2005; Ghafourian and Cronin, 2005; Tropsha *et al.*, 2003; Golbraikh and Tropsha, 2002). The validation strategies check the reliability of the developed models for their possible application on a new set of data and confidence of prediction can thus be judged. For validation of QSAR models usually four strategies are adopted (Wold and Eriksson, 1995): first, internal validation or cross-validation; second, validation by dividing the data set into training and test compounds; third, true external validation by application of model on external data; and fourth, data randomization or *Y*-scrambling. As a result, a simple mathematical relationship is established:

$$\text{Property} = f(\text{structural descriptors})$$

QSAR techniques include from chemical measurements and biological assays to the statistical techniques and interpretation of results (Bordbar *et al.*, 2013; Brown *et al.*, 1996; Roy and Leonard, 2005).

In this work a QSAR study is performed, to develop model that relate the structures of 92 substituted benzenes to theoretical descriptors. The GA was used to select the most informative descriptors from the calculated descriptors by Dragon (Version 5.4) software (Todeschini *et al.*, 2005). The selected descriptors were used to develop a model by different approach (MLR; artificial neural networks (ANN); SVM) for predicting the log (1/LC50) (decimal logarithm of the inverse 50 percent lethal concentration) values. We have validated the models by dividing the data set into training (74 compounds) and test set (18 compounds) by Kennard and Stones algorithm. Different statistical techniques were used to develop the model to highlight the structural requirements for an ideal aquatic lethal toxicity. The three objectives of the present paper have been: first, to explore the structure-activity relationships of aquatic toxicity of diverse compounds; second, to select the best predictive model from among all comparable chemometric models for the aquatic toxicity, and third, verification of the performance and stability of the obtained model by three approach (MLR; ANN; SVM).

2. Materials and methods

2.1 Data set

The known experimental pLC50 values (Table IV) for a diverse data set consisted of 92 substituted benzene compounds was taken from literature (Kier and Hall, 1986; Kier and Hall, 1975).

2.2 Calculation of descriptors

The chemical structure of each compound was sketched on a PC using the HYPERCHEM Program (HYPERCHEM Software, 2002) and pre-optimized using MM+ molecular mechanics method (Polak-Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi-empirical PM3 method at a restricted Hartree-Fock level with no configuration interaction, applying a gradient norm limit of 0.001 kcal Å⁻¹ mol⁻¹ as a stopping criterion. Then the geometries were used as input for the generation of 1,664 descriptors using the Dragon software (Version 5.4) (Todeschini *et al.*, 2005).

2.3 Kennard and Stones algorithm

Kennard and Stones algorithm has been widely used for splitting data sets into two subsets. This algorithm starts by finding two samples, based on the input variables that are the farthest apart from each other. These two samples are removed from the original data set and put into the calibration set. This procedure is repeated until the desired number of samples has been selected in the calibration set. The advantages of this algorithm are that the calibration samples always map the measured region of the input variable space completely with respect to the induced metric and that the no validation samples fall outside the measured region. Kennard and Stones algorithm has been considered as one of the best ways to build training and test sets (Tropsha *et al.*, 2003; Wu *et al.*, 1996).

2.4 Descriptor selection

MLR analysis and variable selection were performed by the software MOBY DIGS of Todeschini *et al.* (2009) using the ordinary least square regression method and GA-variable subset selection (Leardi *et al.*, 1992). The outcome of the application of the GAs is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by Q^2 . The models with lower Q^2 are those with fewer descriptors.

2.5 Chemometric methods

2.5.1 MLRs. MLR is a statistical tool that regresses independent variables against a dependent variable. The objective of MLR is to find a linear model of the property of interest, which takes the form:

$$y = a_0 + \sum_{i=1}^n a_i x_i \quad (1)$$

where y is the property, which is, the dependent variable; x_i the molecular descriptors; and a_i the coefficients of those descriptors and a_0 is the intercept of the equation.

2.5.2 ANN. ANN are artificial systems simulating the function of the human brain. Three components constitute a NN: the processing elements or nodes, the topology of the connections between the nodes, and the learning rule by which new information is encoded in the network. While there are a number of different ANN models, the most frequently used type of ANN in QSAR is the three-layered feed-forward network (Zupan and Gasteiger, 1993). In this type of networks, the neurons are arranged in layers (an input layer, one hidden layer and an output layer). Each neuron in any layer is fully connected with the neurons of a succeeding layer and no connections are between neurons belonging to the same layer.

2.5.3 SVM. SVM is gaining popularity due to many attractive features and promising empirical performance. It originated from early concepts developed by Cortes and Vapnik (1995). This method has proven to be very effective for addressing general purpose classification and regression problems (Darnag *et al.*, 2014).

Similar with other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters. They are capacity parameter C , of ϵ insensitive loss function, the kernel type K and its corresponding parameters. C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small then insufficient stress will be placed on fitting the training data. If C is too large then the algorithm will over fit the training data (Li *et al.*, 2007).

2.6 Model development and validation

2.6.1 Statistical parameters. The original data set (92 compounds) was split into a training set (74 compounds), used for establishing the QSAR models and selecting the parameters of the methods used, and a test set (18 compounds) for external validation.

To assess the predictivity of the developed QSAR models, several diagnostic statistical (Wehrens *et al.*, 2000) tools are used.

Root mean square error:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{m=1}^N (Y_m - \hat{Y}_m)^2} \quad (2)$$

Correlation coefficient:

$$Q^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

External correlation coefficient:

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

External root mean square error:

$$\text{RMSE}_{ext} = \sqrt{\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} (y_i - \bar{y})^2} \quad (5)$$

Variance inflation factor:

$$\text{VIF} = 1 / (1 - R_j^2) \quad (6)$$

In these equations: y_m is the desired output; \hat{y} the predicted value by model; \bar{y} the mean of dependent variable; n the number of the molecules in data set; n_{tr} the number of the molecules in training set; n_{ext} the number of the molecules in test set; n_{tr} the number of

the molecules in training set; and R_j^2 the squared correlation coefficient between the j th coefficients regressed against all the other descriptors in the model.

2.6.2 Applicability domain (AD) analysis. The AD of a QSAR model (Tropsha *et al.*, 2003; Shen *et al.*, 2004) must be defined if the model is to be used for screening new compounds. The AD is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSAR should make reliable predictions. This region is defined by the nature of the compounds in the training set, and can be characterized in various ways. In this work, the structural AD was verified by the leverage approach. The leverage h_i (Xu *et al.*, 2011) is defined as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (7)$$

where x_i is the descriptor row-vector the i th compound, x_i^T is the transpose of x_i , X is the descriptor matrix, X^T is the transpose of X . The warning leverage h^* is, generally, fixed at $3(m+1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation. In fact, leverage can be used as a quantitative measure of the model AD suitable for evaluating the degree of extrapolation. It represents a sort of compound distance from the model experimental space. The Williams plot, the plot of leverage values vs standardized residuals, was used to give a graphical detection of both the response outliers (Y outliers) and the structurally influential compounds (X outliers). In this plot, the two horizontal lines indicate the limit of normal values for Y outliers (i.e. samples with standardized residuals greater than 3 standard deviation units, $\pm 3s$); the vertical straight lines indicate the limits of normal values for X outliers (i.e. samples with leverage values greater than the threshold value, $h > h^*$). For a sample in the external test set whose leverage value is greater than h^* , its prediction is considered unreliable, because the prediction is the result of a substantial extrapolation of the model. Conversely, when the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and experimental values is as high as that for the compounds in the training set. It is noteworthy that the response outliers can be highlighted only for compounds with known responses and the possibility of a compound to be out of the structural AD of a model can be verified for every new compound, the only knowledge needed being the molecular structure information represented by the molecular descriptors selected in the model.

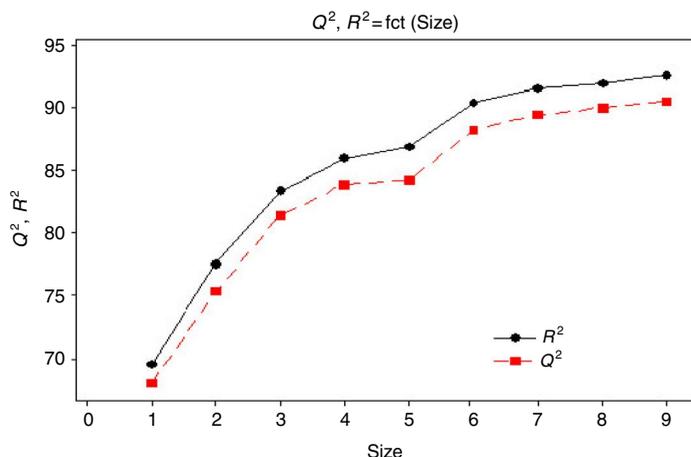
2.6.3 Y-randomization. Chance of correlations, if any, associated with the MLR models were recognized in randomization test by repeated scrambling of the biological response. The data sets with scrambled response vector have been reassessed by MLR. The resulting regression equations, if any, with correlation coefficients better than or equal to the one corresponding to the unscrambled response data were counted. Every model has been subjected to 100 such simulation runs. This has been used as a measure to express the percent chance correlation of the model under scrutiny (Sharma *et al.*, 2011).

3. Results and discussion

3.1 Results of the MLR model

The Q^2 , R^2 results during the GA MLR analysis are shown in Figure 1. Obviously, pLC50 max is not linearly correlated with any of the molecular descriptors since univariant correlations between pLC50 max and the different descriptors have poor Q^2 values. The Q^2 increases gradually with the increased number of descriptors. When

Figure 1.
 Q^2 and R^2 vs
number of latent
descriptors in the
best MLR equation



adding another descriptor did not significantly improve the statistics of a model, it was determined that the optimum subset size had been achieved.

The graph of the Figure 1, reproduces the variation of Q^2 and R^2 according to the variable number of the model, it is obvious that the optimum model comprises six descriptors (variables).

According to the Table I, it was clear that the size of models higher than six descriptors encounter problems in the external prediction Q^2_{ext} .

The linear function constructed from the six molecular descriptors and the training set has the following form:

$$\begin{aligned} \text{pLC50} = & -1.36 - 0.164 \text{ po} + 6.24 \text{ MATS1m} + 0.550 \text{ RDF020v} - 0.433 \text{ E1s} \\ & + 0.0625 \text{ PCWTe} - 0.261 \text{ c log } p \end{aligned}$$

$$Q^2 = 0.8829, \text{ RMSE} = 0.225, Q^2_{ext} = 0.9538, \text{ RMSE}_{ext} = 0.141.$$

3.1.1 Descriptor contribution analysis. Based on a previously described procedure (Zheng *et al.*, 2006; Guha and Jurs, 2005), the relative contributions of the six descriptors to the model were determined and are plotted in Figure 2.

Six descriptors were needed in the QSAR model, although it is not against the rule of thumb for building a linear model. It should be noted that the difference in the descriptor

No.	R^2	$R^2 - R^1$	Q^2	$Q^2 - Q^1$	Q^2_{ext}
1	0.6961		0.6804		0.6992
2	0.7751	0.079	0.7532	0.0728	0.9378
3	0.8337	0.0586	0.8147	0.0615	0.9299
4	0.86	0.0263	0.8393	0.0246	0.917
5	0.8688	0.088	0.8427	0.0034	0.927
6	0.9039	0.0351	0.8829	0.0402	0.9538
7	0.9152	0.0113	0.8946	0.0117	0.9319
8	0.9203	0.0051	0.9005	0.0059	0.9003
9	0.9257	0.0054	0.9055	0.005	0.8574

Table I.
Comparison the
performance of
different size models

contribution between any two descriptors used in the model is not significant, indicating that all of the descriptors are indispensable in generating the predictive.

3.1.2 Mechanism of toxicity. The selected descriptors by biological activity show different aspects of the mechanism of toxicity of the substituted benzenes. It has been established that chemical toxicity is a combination of uptake into or through biological membranes and the inter-action of toxicant with the site of action. Uptake of most organic chemicals to the site of action is made by passive diffusion. Hydrophobicity is the most important factor in diffusion and most often quantified by the octanol-water partition coefficient (log KOW). Selection of the log KOW as the most repetitive descriptor in all bee models (with different number of descriptors) supports the importance of this parameter, so this descriptor as the most important descriptor among six selected descriptors (Zarei *et al.*, 2014).

The high absolute *t*-values shown in Table II express that the regression coefficients of the descriptors involved in the MLR model are significantly larger than the standard deviation. The *t*-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (i.e. descriptors' interactions). Descriptors with *t*-probability values below 0.05 (95 percent confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance (Ramsey and Schafer, 1997). The smaller *t*-probability suggests the more significant descriptor. The *t*-probability values of the six descriptors are very small, indicating that all of them are highly significant descriptors. Models would not be accepted if they contain descriptors with VIFs above a value of five (Holder *et al.*, 2003). Correlation matrix as shown

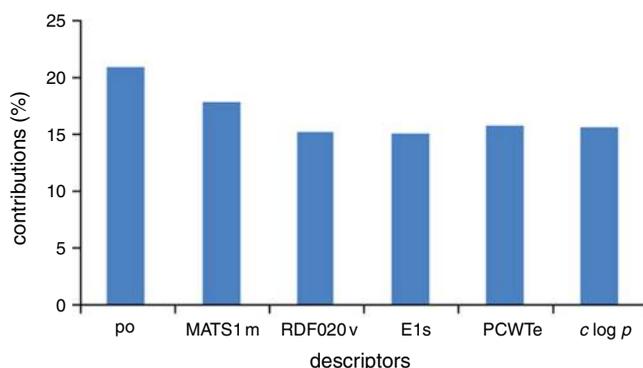


Figure 2. Relative contributions of the selected descriptors to the MLR model

Predictor	Coef	SE coef	<i>t</i>	<i>p</i>	VIF
Constant	-1.3635	0.1683	-8.10	0.000	-
po	-0.16427	0.01795	-9.15	0.000	4.326
MATS1m	6.2420	0.9533	6.55	0.000	1.171
RDF020v	0.5503	0.1764	3.12	0.003	3.603
E1s	-0.4333	0.1070	-4.05	0.000	1.319
PCWTe	0.06254	0.01146	5.46	0.000	2.016
<i>c log p</i>	0.26124	0.05402	-4.84	0.000	3.932

Table II. Characteristics of the selected descriptors in MLR model

in Table III suggests that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

3.1.3 Regression line of model MLR. According to the Figure 3 it was clear that the activity values calculated were very similar to the experimental values.

3.1.4 Domain of applicability in MLR. It needs to be pointed out that no matter how robust, significant and validated a QSAR model may be, it cannot be expected to reliably predict the modeled property for the entire universe of compounds. Therefore, before a QSAR model is put into use for screening compounds, its AD must be defined and predictions for only those compounds that fall in this domain can be considered as reliable. The AD of the MLR model was analyzed in the Williams plot (shown in Figure 4). There are one *X* outlier with leverage higher than the warning limit of 0.28 (Compound 68) and one *Y* outlier with residual higher than $\pm 3s$ (Compound 71) in the training set. Removing these two outliers could improve Q^2 between the experimental pLC50 values and the selected descriptors.

3.1.5 Y-randomization test. Y-randomization is an attempt to observe the action of chance in fitting given data. In other words it is applied to exclude the possibility of chance correlation. This technique ensures the robustness of a QSAR model (Tropsha *et al.*, 2003; Tropsha and Golbraikh, 2007) (Figure 5).

The low values of R^2 and Q^2 for the order model indicate that the selected model is not due to chance.

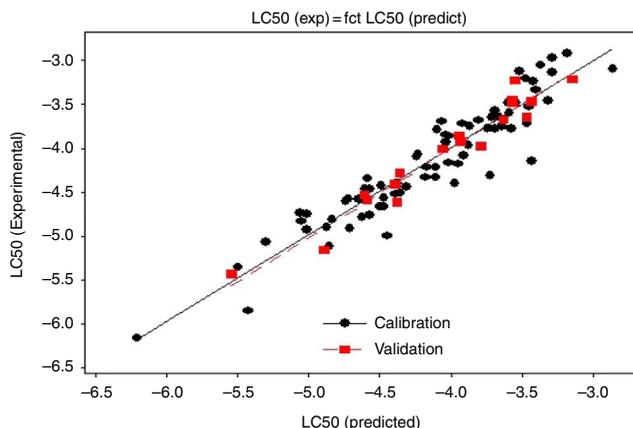
3.2 SVM

The training of the SVM model included the selection of capacity parameter C , ϵ of ϵ -insensitive loss function and the corresponding parameters of the kernel function. First, the kernel function should be decided, which determines the sample distribution in the

Table III.
Correlation matrix
between the selected
descriptors and
pLC50

	pLC50	po	MATS1m	RDF020v	E1s	PCWTe
po	-0.653					
MATS1m	-0.065	0.181				
RDF020v	0.269	0.381	-0.170			
E1s	-0.332	-0.076	0.240	-0.378		
PCWTe	0.098	0.482	-0.145	0.607	-0.417	
$c \log p$	-0.834	0.607	0.220	-0.332	0.187	0.051

Figure 3.
log (1/LC) observed
experimentally vs
log (1/LC) predicted
by MLR



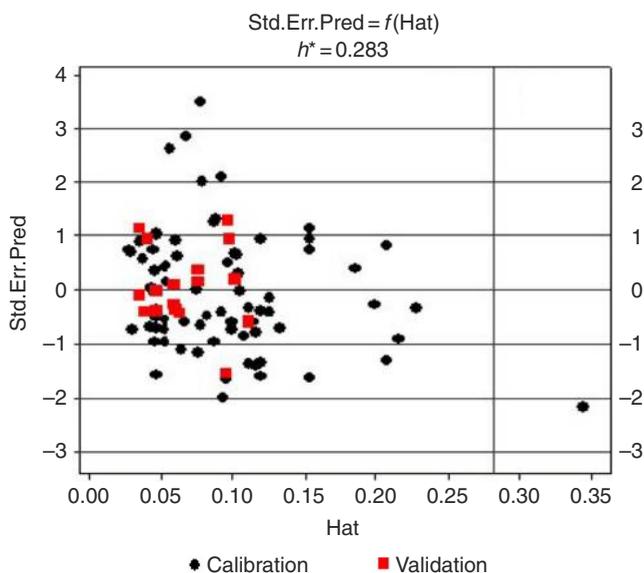


Figure 4.
Williams plot of the
current QSAR model

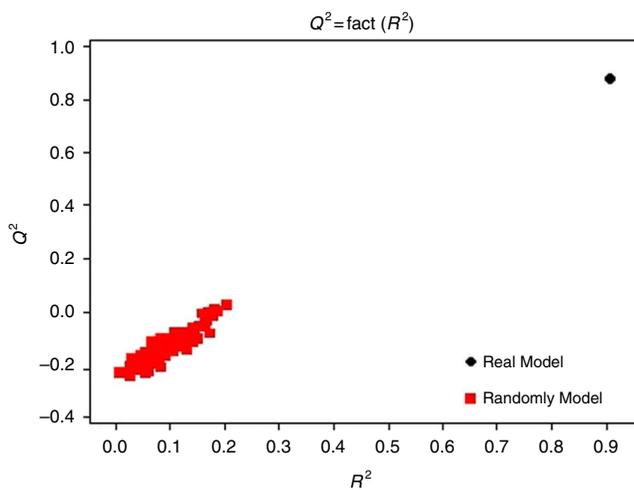


Figure 5.
Randomization test
associated to the
QSAR model

mapping space. Generally, using RBF kernel function will yield better prediction performance (Nianyi *et al.*, 2004). The optimal values of C , γ and ε are identified to be 88888.9, $1e^{-05}$ and 0.22222, respectively. The values of Q^2 and RMSE are 0.8882 and 0.222, respectively. The external values of Q^2 and RMSE are 0.947 and 0.146, respectively.

3.3 ANN

Among all architectures of ANN, the best one is 7-4-1 ($Q^2 = 0.898$ and $\text{RMSE} = 0.214$), ($Q_{ext}^2 = 0.9564$ and $\text{RMSE}_{ext} = 0.135$)

According to the Tables IV and V the statistics obtained by the three techniques (MLR, NNA, SVM) showed that our model was stable, robust and significant.

Compounds	LC50	SVM	MLR-LOO	ANN (7-3-4)
Chlorobenzene	-3.77	-3.83035	-3.74546	-3.74151
1,2-dichlorobenzene	-4.4	-4.48338	-4.38271	-4.43937
1,4-dichlorobenzene	-4.56	-4.56954	-4.46907	-4.51028
1,2,3-trichlorobenzene	-4.89	-4.98056	-4.88053	-4.98078
1,2,4-trichlorobenzene	-4.83	-5.14178	-5.05285	-5.0448
1,3,5-trichlorobenzene	-4.74	-5.10292	-5.00723	-5.076
1,2,3,4-tetrachlorobenzene	-5.35	-5.57135	-5.49524	-5.41645
1,2,4,5-tetrachlorobenzene	-5.85	-5.49538	-5.42123	-5.33095
1-chloro-3-methyl-benzene	-3.84	-4.09821	-4.038	-4.03969
1-chloro-4-methyl-benzene	-4.33	-4.22253	-4.11184	-4.15287
1,2,4-trichloro-5-methyl-benzene	-5.06	-5.38214	-5.29733	-5.24486
1,2-dichloro-4-methyl-benzene	-4.6	-4.84391	-4.74415	-4.79413
1,2,3,4,5-pentachloro-6-methyl-benzene	-6.15	-6.2801	-6.20402	-5.83564
Benzene	-3.09	-2.93762	-2.86614	-3.0203
Toluene	-3.13	-3.31579	-3.28472	-3.26926
1,2-xylene	-3.48	-3.62715	-3.59907	-3.55371
1,4-xylene	-3.48	-3.58264	-3.55134	-3.50956
Nitrobenzene	-2.97	-3.33114	-3.28405	-3.31017
1-methyl-2-nitro-benzene	-3.59	-3.68566	-3.58487	-3.60532
1-methyl-3-nitro-benzene	-3.65	-3.81668	-3.71205	-3.73011
1,2-dimethyl-3-nitro-benzene	-4.39	-4.11069	-3.97564	-4.02403
1,2-dimethyl-4-nitro-benzene	-4.21	-4.2327	-4.10444	-4.13594
1-chloro-2-nitro-benzene	-3.72	-3.93677	-3.92089	-3.88837
1-chloro-4-nitro-benzene	-4.42	-4.64256	-4.49677	-4.51232
1,2-dichloro-3-nitro-benzene	-4.66	-4.50215	-4.50354	-4.51735
2,4-dichloro-1-nitro-benzene	-4.46	-4.60542	-4.61065	-4.59398
1,3-dichloro-5-nitro-benzene	-4.58	-4.64973	-4.64906	-4.65265
1-chloro-2-methyl-3-nitro-benzene	-4.52	-4.4445	-4.38609	-4.38701
4-chloro-1-methyl-2-nitro-benzene	-4.44	-4.36645	-4.31572	-4.32848
Phenol	-3.45	-3.22425	-3.31629	-3.25839
2-methylphenol	-3.77	-3.53301	-3.57994	-3.49417
4-methylphenol	-3.74	-3.58894	-3.64372	-3.55421
2,6-dimethylphenol	-3.75	-3.84254	-3.87107	-3.78263
2,3,6-trimethylphenol	-4.21	-4.15463	-4.16903	-4.14101
4-ethylphenol	-4.07	-3.90861	-3.91137	-3.85323
4-propylphenol	-4.09	-4.25004	-4.24035	-4.20591
4-butylphenol	-4.47	-4.60121	-4.57657	-4.55226
4-tert-butylphenol	-4.46	-4.39709	-4.57808	-4.51038
4-methyl-2-tert-butyl-phenol	-4.9	-4.58099	-4.70879	-4.74544
4-pentylphenol	-5.12	-4.92698	-4.85182	-4.84189
4-(2-methylbutan-2-yl)phenol	-4.81	-4.68858	-4.83424	-4.77966
2-prop-2-enylphenol	-3.96	-3.90149	-3.88034	-3.81388
2-phenylphenol	-4.76	-4.45127	-4.57126	-4.58508
Naphth-1-ol	-4.5	-4.26329	-4.35766	-4.34269
4-chlorophenol	-4.18	-3.93639	-3.9572	-3.91333
4-chloro-3-methyl-phenol	-4.33	-4.16685	-4.17845	-4.15503
4-chloro-3,5-dimethylphenol	-4.66	-4.46813	-4.47471	-4.47649
4-methoxyphenol	-3.05	-3.3507	-3.3751	-3.31629
4-phenoxyphenol	-4.58	-4.56774	-4.71578	-4.68027
(2S)-2-amino-3-(3H-imidazol-4yl)propanoic	-3.63	-3.6473	-3.67662	-3.59185

Table IV.
Compounds and the
predicted results of
the biological
activity pLC50

(continued)

Compounds	LC50	SVM	MLR-LOO	ANN (7-3-4)
Aniline	-2.91	-3.11501	-3.1868	-3.13259
2-methylaniline	-3.12	-3.47612	-3.51939	-3.39356
4-methylaniline	-3.72	-3.48476	-3.46109	-3.39861
N,N-dimethylaniline	-3.33	-3.46437	-3.40551	-3.38112
2-ethylaniline	-3.21	-3.57518	-3.48017	-3.42124
4-ethylaniline	-3.52	-3.58244	-3.45392	-3.43907
4-butylaniline	-4.16	-4.16551	-4.01168	-4.01078
2,6-dipropan-2-ylaniline	-4.06	-4.10066	-4.2282	-4.09578
2-chloroaniline	-4.31	-3.83238	-3.72177	-3.70656
4-chloroaniline	-3.67	-3.84311	-3.80996	-3.74746
2,5-dichloroaniline	-4.99	-4.56596	-4.44545	-4.47459
3,4-dichloroaniline	-4.39	-4.43828	-4.38394	-4.37531
2,3,6-trichloroaniline	-4.73	-5.13898	-5.06185	-4.99604
2,4,5-trichloroaniline	-4.92	-5.09255	-5.0133	-4.98045
4-bromoaniline	-3.56	-3.84621	-3.68878	-3.66952
4-fluro-3-(trifluoromethyl) aniline	-3.77	-3.66578	-3.68848	-3.66669
4-fluro-2-(trifluoromethyl) aniline	-3.78	-4.14417	-4.09831	-4.0351
2,3,4,5,6-pentafluoroaniline	-3.69	-3.77702	-4.06334	-3.82059
3-phenylmethoxyaniline	-4.34	-4.59314	-4.58536	-4.5044
4-hexoxyaniline	-4.78	-4.80503	-4.62292	-4.63588
2-nitroaniline	-4.15	-3.63788	-3.43008	-3.51284
4-nitroaniline	-3.23	-3.44366	-3.42475	-3.40801
2-chloro-4-nitro-aniline	-3.93	-4.06466	-4.03416	-3.98182
4-ethoxy-2-nitro-aniline	-3.85	-4.18033	-4.01758	-4.00761
1,3-dichlorobenzene	-4.28	-4.43198	-4.17	-4.17
1,2,3,5-tetrachlorobenzene	-5.43	-5.6432	-5.67913	-5.67913
2,4-dichloro-1-methyl-benzene	-4.54	-4.69806	-4.60076	-4.60076
1,3-xylene	-3.45	-3.61481	-3.60976	-3.60976
1-methyl-4-nitro-benzene	-3.67	-3.73465	-3.45165	-3.45165
1-chloro-3-nitrobenzene	-4.01	-4.08952	-3.92861	-3.92861
1,4-dichloro-2-nitro-benzene	-4.59	-4.59178	-4.51358	-4.51358
3-methylphenol	-3.48	-3.48421	-3.4927	-3.4927
2,4-dimethylphenol	-3.86	-3.91718	-3.9888	-3.9888
3,4-dimethylphenol	-3.92	-3.87816	-3.91363	-3.91363
3-methoxyphenol	-3.22	-3.06706	-3.07564	-3.07564
3-methylaniline	-3.47	-3.40947	-3.55747	-3.55747
3-ethylaniline	-3.65	-3.57373	-3.43421	-3.43421
3-chloroaniline	-3.98	-3.81601	-3.73967	-3.73967
2,4-dichloroaniline	-4.41	-4.47374	-4.50667	-4.50667
3,5-dichloroaniline	-4.62	-4.42679	-4.39361	-4.39361
2,3,4-trichloroaniline	-5.15	-4.96576	-4.99111	-4.99111
3-nitroaniline	-3.24	-3.57797	-3.75482	-3.75482

Table IV.

Approaches	Q^2	RMSE	Q^2_{ext}	RMSE _{ext}
MLR	0.8829	0.225	0.9538	0.141
NNA	0.898	0.214	0.9564	0.135
SVM	0.8882	0.222	0.947	0.146

Table V.
Statistical parameters and predictive ability of training and testing data set

4. Conclusion

Development of quantitative structure-property/activity relationships (QSPR/QSAR) on theoretical descriptors is a powerful tool not only for prediction of the chemical, physical and biological properties/activities of compounds, but also for deeper understanding of the detailed mechanisms of aquatic toxicity in benzene derivatives that predetermine these activity.

In this paper we have developed a useful QSAR equation derived from theoretical chemical descriptors associated with aquatic toxicity properties of 92 benzene derivatives. For each compound 1,664 descriptors, 20 classes of Dragon descriptors are calculated. The data set was carefully split into training and test sets, guaranteeing enough molecular diversity in each subset, by using Kennard and Stones algorithm analysis. Then the best set of calculated descriptors was selected by GA of MOBY DIGS. This model with high statistical quality and low prediction errors was obtained. In general, it can be concluded that, for this data set, the combinations of linear modeling techniques result in an improvement of the linear models. The results indicate that six descriptors were selected and play an important role on the aquatic toxicity of benzene derivatives structure.

The QSAR method was applied to the prediction of the lethal concentration of benzene derivatives. A six-parameter linear model was developed by MLR, with Q^2 of 0.8829 and RMSE of the 0.225, NNA with Q^2 of 0.898, RMSE of 0.214, SVM with Q^2 of 0.8882, RMSE of 0.222 for training set. Several validation techniques, including leave-one-out cross-validation, randomization tests, and validation through the test set, illustrated the reliability of the proposed model. All of the descriptors involved can be directly calculated from the molecular structure of the compounds, thus the proposed model is predictive and could be used to estimate the lethal concentration of benzene derivatives.

References

- Bordbar, M., Ghasemi, J., Fall, A.Y. and Fazaeli, R. (2013), "Chemometric modeling to predict aquatic toxicity of benzene derivatives using stepwise-multi linear regression and partial least square", *Asian Journal of Chemistry*, Vol. 25 No. 1, pp. 331-342.
- Brown, S.D., Sum, S.T., Despagne, F. and Lavine, B.K. (1996), "Chemometrics", *Analytical Chemistry*, Vol. 68 No. 1080, pp. 21-61.
- Cortes, C. and Vapnik, V. (1995), "Support- vector networks", *Machine Learning*, Vol. 20 No. 3, pp. 273-297.
- Darnag, R., Minaoui, B. and Fakir, M. (2014), "QSAR models for prediction study of HIV protease inhibitors using support vector machines neural networks and multiple linear", *Arabian Journal of Chemistry* (in press), available at: <http://dx.doi.org/10.1016/j.arabjc.2012.10.021>
- Devillers, J. and Balaban, A.T. (1999), *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, Amsterdam.
- Ghafourian, T. and Cronin, M. (2005), "The impact of variable selection on the modelling of oestrogenicity", *SAR QSAR Environmental Research*, Vol. 16 Nos 1-2, pp. 171-190.
- Golbraikh, A. and Tropsha, A. (2002), "Beware of q^2 !", *Molecular Graphics and Modelling*, Vol. 20 No. 4, pp. 269-276.
- Guha, R. and Jurs, P.C. (2005), "Interpreting computational neural network QSAR models: a measure of descriptor importance", *Chemical Information and Modeling*, Vol. 45 No. 3, pp. 800-806.
- Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M. (1962), "Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients", *Nature*, Vol. 194, pp. 80-178.

- He, L. and Jurs, P.J. (2005), "Assessing the reliability of a QSAR model's predictions", *Molecular Graphics and Modelling*, Vol. 23 No. 6, pp. 503-523.
- Holder, A.J., Yourtee, D.M., White, D.A., Glaros, A.G. and Smith, R.J. (2003), "Chain melting temperature estimation for phosphatidyl cholines by quantum mechanically derived quantitative structure property relationships", *Computer-Aided Molecular Design*, Vol. 17 No. 2, pp. 223-230.
- HYPERCHEM Software (2002), Release 6.03 for Windows, Molecular Modeling System.
- Jalali-Heravi, M. and Kyani, A. (2008), "Comparative structure – toxicity relationship study of substituted benzenes to *Tetrahymena pyriformis* using shuffling-adaptive neuro fuzzy inference system and artificial neural networks", *Chemosphere*, Vol. 72 No. 5, pp. 733-740.
- Karelson, M. (2000), *Molecular Descriptors in QSAR/QSPR*, John Wiley & Sons, New York, NY.
- Kier, L.B. and Hall, L.H. (1975), *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, NY.
- Kier, L.B. and Hall, L.H. (1986), *Molecular Connectivity in Structure-Activity Analysis*, John Wiley & Sons Inc., New York, NY.
- Leardi, R., Boggia, R. and Terrile, M. (1992), "Genetic algorithms as a strategy for feature selection", *Chemometrics*, Vol. 6 No. 5, pp. 267-281.
- Li, X., Luan, F., Si, H., Hu, Z. and Liu, M. (2007), "Prediction of retention times for a large set of pesticides or toxicants based on support vector machine and the heuristic method", *Toxicology Letters*, Vol. 175 Nos 1-3, pp. 136-144.
- Netzeva, T.I., Pavan, M. and Worth, A.P. (2008), "Quantitative structure – activity relationships for acute aquatic toxicity", *QSAR & Combinatorial Science*, Vol. 27 No. 1, pp. 77-90.
- Nianyi, C. (2004), *Support Vector Machine in Chemistry*, World Scientific Publishing Company, New York, NY.
- Ramsey, L.F. and Schafer, W.D. (1997), *The Statistical Sleuth*, Wadsworth Publishing Company, Belmont.
- Roy, K. and Leonard, J.T. (2005), "QSAR analyses of 3-(4-benzylpiperidin-1-yl)-N-phenylpropylamine derivatives as potent CCR5 antagonists", *Chemical Information and Modeling*, Vol. 45 No. 5, pp. 1352-1368.
- Sharma, B.K., Singh, P., Pilania, P., Sarbhai, K., Yenamandra, S. and Prabhakar, C.P. (2011), "MLR/PLS directive QSAR study on apical sodium-codependent bile acid transporter inhibition activity of benzothiepinines", *Molecular Diversity*, Vol. 15, pp. 135-147.
- Shen, M., Béguin, C., Golbraikh, A., Stables, J.P., Kohn, H. and Tropsha, A.J. (2004), "Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds", *Medicinal Chemistry*, Vol. 47 No. 9, pp. 2356-2364.
- Todeschini, R. and Consonni, V. (2000), *Handbook of Molecular Descriptors*, Wiley-VCH Verlag GmbH, Weinheim.
- Todeschini, R., Consonni, V., Mauri, A. and Pavan, M. (2005), "Dragon Software for the calculation of molecular descriptors – Version 5.3 for Windows", Talete srl, Milano.
- Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. and Pavan, M. (2009), "Moby Digs Software for multilinear regression analysis and variable subset selection by genetic algorithm", Release 1.1 for Windows, Milano.
- Tong, W., Hong, H., Xie, Q., Shi, L., Fang, H. and Perkins, R. (2005), "Assessing QSAR limitations – a regulatory perspective", *Current Computer-Aided Drug Design*, Vol. 1, pp. 195-205.

- Tropsha, A. and Golbraikh, A. (2007), "Predictive QSAR modeling workflow, model applicability domains, and virtual screening", *Current Pharmaceutical Design*, Vol. 13 No. 34, pp. 3494-3504.
- Tropsha, A., Gramatica, P. and Gombar, V.K. (2003), "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models", *QSAR & Combinatorial Science*, Vol. 22, pp. 69-77.
- Wehrens, R., Putter, H. and Buydens, L.M.C. (2000), "The bootstrap: a tutorial", *Chemometrics and Intelligent Laboratory Systems*, Vol. 54 No. 1, pp. 35-52.
- Wold, S. and Eriksson, L. (1995), *Chemometric Methods in Molecular Design*, VCH Publisher, Weinheim.
- Worth, A.P., Bassan, A., De Bruijn, J., Gallegos Saliner, A., Netzeva, T., Pavan, M., Patlewicz, G., Tsakovska, I. and Eisenreich, S. (2007), "The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods", *SAR & QSAR in Environmental Research*, Vol. 18 Nos 1-2, pp. 111-125.
- Wu, W., Walczak, B., Massart, D.L., Heurding, S., Erni, F., Last, I.R. and Prebble, K.A. (1996), "Artificial neural networks in classification of NIR spectral data: design of the training set", *Chemometrics and Intelligent Laboratory Systems*, Vol. 33 No. 1, pp. 35-46.
- Xu, J., Wang, L., Liu, L., Bai, Z. and Wang, L. (2011), "QSPR study of the absorption maxima of azobenzene dyes", *Bulletin of the Korean Chemical Society*, Vol. 32 No. 11, pp. 3865-3872.
- Zarei, K., Atabati, M. and Kor, K. (2014), "Bee algorithm and adaptive neuro-fuzzy inference system as tools for QSAR study toxicity of substituted benzenes to *Tetrahymena pyriformis*", *Bulletin of Environmental Contamination and Toxicology*, Vol. 92 No. 6, pp. 642-649.
- Zheng, F., Bayram, E., Sumithran, S.P., Ayers, J.T., Zhan, C.G., Schmitt, J.D., Dwoskin, L.P. and Crooks, P.A. (2006), "QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release", *Bioorganic & Medicinal Chemistry*, Vol. 14 No. 9, pp. 3017-3037.
- Zupan, J. and Gasteiger, J. (1993), *Neural Networks for Chemists: An Introduction*, VCH Publishers, Weinheim.

Corresponding author

Nadia Ziani can be contacted at: ziani_nadia84@yahoo.fr