



Faculté des Sciences  
Département de Chimie

## THESE

Présentée pour obtenir le diplôme de **Doctorat En Sciences**

**Option : Chimie Analytique et Environnement**

## THEME

Modélisation de la rétention chromatographique et du facteur de bioaccumulation d'une série de polluants toxiques (Pesticides ; Polychlorobiphényles(PCBs); Hydrocarbures Aromatiques Polycycliques(HAPs).

Par : M<sup>elle</sup> . Khadidja Amirat

Devant le jury :

**Président :**

M<sup>f</sup>.Nasr Eddine Chakri      Professeur      Université Badji Mokhtar Annaba

**Directeur de thèse :**

M<sup>f</sup>.Djelloul Messadi      Professeur      Université Badji Mokhtar Annaba

**Examineurs :**

M<sup>me</sup>.Salima Ali Mokhnache      Professeur      Université Badji Mokhtar Annaba

M<sup>f</sup>. Abdelhak Gheid      Professeur      Université Cherif Messadia Souk Ahras

M<sup>f</sup>.Rachid Merdés      Professeur      Université 8 Mai 1945 Guelma

M<sup>f</sup>.Noureddine Zenati      M.C.A      Université Cherif Messadia Souk Ahras

# Remerciements

D'abord je remercie Mon Dieu qui me donne la volonté et le courage pour réaliser ce travail.

Cette thèse n'aurait pas vu le jour

Sans la confiance, la patience et la générosité du responsable de laboratoire LASEA Monsieur

Le Professeur **D. MESSADI** que je remercie vivement pour avoir accepté la direction de cette thèse. Je voudrais aussi le remercier pour le temps et la patience qu'il m'a accordés tout au long des années d'étude et de recherche en Magister et en Doctorat.

Je tiens également à remercier les membres de jury:

- ❖ **Pr : Nasr Eddine Chakri** pour avoir accepté la présidence de ce jury ; et
- ❖ **Pr : Salima Ali Mokhnache**
- ❖ **Pr : Abdelhak Gheid**
- ❖ **Pr : Rachid Merdés**
- ❖ **Dr : Nouredine Zenati**

pour avoir accepté d'examiner ce travail.

Enfin, je ne saurais oublier mes camarades de laboratoire et également tous ceux qui par leur présence ou par leur aide m'ont permis de mener à bien ce travail, surtout Nadia, Fatiha, Khalil, Youcef, Abdelkrim, Rana, Soumaya, Nabila.

## *Dédicace*

*Je dédie ce modeste travail à :*

- ♣ *La mémoire de mes grands parents et mes deux oncles.*
- ♣ *Mes très chers parents (Salah, Fattoum) et mon grand père Saïd.*
- ♣ *Mes sœurs (Warda, Noura, Fouzia)*
- ♣ *Mes frères (Abd el hafid, Abidine, Bilel, Ali Mohamed Cherif).*
- ♣ *Ma belle sœur Ghania et mon beau frère Brahim.*
- ♣ *Mes nièces (Mounia, Abir, Malék).*
- ♣ *Mes neveux (Raouf, Fateh, Omar).*
- ♣ *A toute ma famille.*
- ♣ *Mes professeurs.*
- ♣ *Mes collègues et mes amies surtout (Nassima, Zohra, Radia).*
- ♣ *Enfin, toute l'équipe de labo [Lasea](#).*

*Kh Adidja-Amirat*



**TABLE DES  
MAT ÈRES**

## Table des Matières

Titres	Page
REMERCEMENTS	
DÉDICACE	
TABLE DES MATIÈRES	I
LISTE DES TABLEAUX	V
LISTE DES FIGURES	VI
SYMBOLES ET ABRÉVIATIONS	VIII
INTRODUCTION GÉNÉRALE	1
PARTIE I. GÉNÉRALITÉS SUR LES POLLUANTS TOXIQUES	
PARTIE I.A. LES PESTICIDES	
A.I. Introduction	4
A.II. Histoire et définitions	4
A.III. Classification	4
A.III.1. Premier système de classification	5
A.III.2. Deuxième système de classification	6
A.IV. Les pesticides dans notre environnement	6
A.V. Les pesticides et la santé	7
A.V.1. Toxicité aiguë	7
A.V.2. Pesticides et perturbation hormonale	7

A.V.3. Problème de développement du fœtus et pesticides : les effets des pesticides perturbateurs endocriniens	7
A.V.4. Impact des pesticides sur le système immunitaire	8
PART E I. B. LES POLYCHLOROB PHENYLES (PCBS)	
B.I. Identité des PCBS	9
B.II. Dans quels milieux rencontre-t-on les PCBs ?	9
B .III. L'exposition humaine et la toxicité aux PCBs	10
B.IV. Quels sont les effets sur la santé des PCBs ?	10
B.IV.1. En toxicité aiguë	10
B.IV.2.En toxicité chronique	11
PART E I.C.LES HYDROCARBURES AROMAT QUES POLYCYCL QUES (HAPs)	
C.I. Définition et caractéristiques des hydrocarbures aromatiques polycycliques	12
Références bibliographiques de la partie I	13
PART E II	
PART E II .A. OPT M SAT ON DE LA GÉOMETR E DE LA MOLÉCULE	
A.I. Généralités	15
A. II. Méthodes semi-empiriques utilisées	17
A.II. 1. Le cadre Hartree - Fock – Roothaan	17
A.II.2.Les méthodes semi-empiriques	20
A.II.3.Champ de force	26

A.II.3. 1 – Définition	26
A.II.3. 2 .Quelques exemples	26
A.II.3. 3 .Représentation simple d'un champ de force	27
A.II.3. 4. Exemple de calcul	29
A.II.3. 5. Champs de force MM2 et MM+	30
A.II.3 .5 .1 .Champ de force MM2	30
A.II.3. 5. 2 .Champ de force MM+	35
PART E II .B.DÉVELOPPEMENT ET ÉVALUATION DES MODÈLES	
B. I. Les Modèles	37
B.I.1.La régression linéaire Simple	37
B.I.2.La régression linéaire Multiple	37
B.II. Développement et évaluation de modèle	41
B.II.1.Robustesse du modèle	41
B.II.2.Domaine d'application	41
B.II. 3.Test de randomisation	41
B.II.4.Validation externe	42
B.II.5 .Sélection d'un sous ensemble de descripteurs	42
B.II.5.1- Algorithme de Kennard et Stone (CADEX)	43
B.II.6.ALGORITHME GÉNÉRIQUE	43
B.II.6. 1.Les origines	44

B.II .6.2..Principe	44
B.II .6.3. Initialisation aléatoire du modèle	44
B.II.6.4 .Etape de croisement	45
B.II. 6.5. Etape de mutation	45
B.II. 6.6.Conditions d'arrêt	46
Références bibliographiques de la partie II	47
PART E III.RÉSULTATS ET D SCUSS ONS	
III. 1.Méthodologie	51
III.2. Modélisation des temps de rétention pour un grand ensemble hétérogène constitué de (Pesticides et PCBs).	53
III. 3.Modélisation du facteur de bioaccumulation d'un ensemble de PCBs.	63
III .4.Modélisation des indices de rétentions des HAPs.	69
CONCLUS ON GÉNÉRALE	
ANNEXE DES PEST C DES	81
ANNEXE DES PCBS	97
ANNEXE DES HAPS	108
RÉSUMÈ	





**L STE DES  
TABLEAUX**

## L STE DES TABLEAUX

<b>Tableau</b>	<b>Titre</b>	<b>Page</b>
Tableau II. 1	Etude comparative des techniques ab initio,semi –empirique et mécanique moléculaire.	36
Tableau III.1	Les valeurs de En, nR06, ATS1m, ATS7v, GATS2e, EEig05d et log tr pour un ensemble de 84 pesticides ou toxiques. Les 17 derniers composés constituent l'ensemble de test.	53
Tableau III.2	Classe des descripteurs sélectionnés	57
Tableau III.3	Caractéristiques des descripteurs sélectionnés pour le meilleur modèle AG / RLM optimal.	58
Tableau III.4	Matrice de corrélation entre les descripteurs sélectionnés et log tr.	59
Tableau III.5	Valeurs de log BCF (exp), pour un ensemble de 58 PCB. Les 28 derniers PCBs constituent l'ensemble de test.	63
Tableau III.6	Le seul descripteur sélectionné pour la modélisation de log BCF.	66
Tableau III.7	Valeurs de Ir, la masse moléculaire, l'énergie de solvation pour un ensemble de 93 HAP. Les 23 derniers produits chimiques constituent l'ensemble de test.	69
Tableau III.8	Liste des descripteurs sélectionnés pour la modélisation de Ir.	74
Tableau III.9	Caractéristiques des descripteurs sélectionnés pour le meilleur modèle AG / MLR.	74
Tableau III.10	Matrice de corrélation entre les indices de rétention et les descripteurs sélectionnés.	75



**L I S T E D E S  
F I G U R E S**

## LISTE DES FIGURES

Figure	Titre	Page
Figure I. 1	Structures chimiques des principales familles de pesticides	6
Figure I.2	La formule de la structure générale et les positions de substitution des PCBs.	9
Figure II. 1	Représentation schématique des contributions ; à un champ de force de MM.	28
Figure II.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.	29
Figure II.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.	31
Figure II. 4	Deux façons pour modéliser les contributions de variation d'angle extra – planaire.	33
Figure II. 5	Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces.	34
Figure III.1	Les valeurs prédites de log tr en fonction des valeurs expérimentales de log tr.	59
Figure III.2	Test de randomisation associé au modèle QSRR précédent.	60
Figure III.3	Diagramme de Williams du modèle QSRR sélectionné.	61
Figure III.4	Contributions relatives des descripteurs sélectionnés dans le modèle AG / RLM.	62

Figure III.5	Les valeurs prédites de logarithme décimal BCF en fonction des valeurs expérimentales pour l'ensemble des données.	66
Figure III. 6	Test de randomisation associé au modèle QSPR précédent. Les carrés représentent les propriétés ordonnées au hasard et le cercle correspond aux propriétés réelles.	67
Figure III.7	Diagramme de Williams du modèle QSPR développé.	68
Figure III.8	Les valeurs des indices de rétention prédites en fonction des indices de rétention expérimentales.	76
Figure III.9	Le diagramme de Williams pour le modèle QSRR optimal.	77
Figure III.10	Test de randomisation associé au modèle de QSRR développé.	78



**SYMBOLES ET  
ABRÉVATIONS**

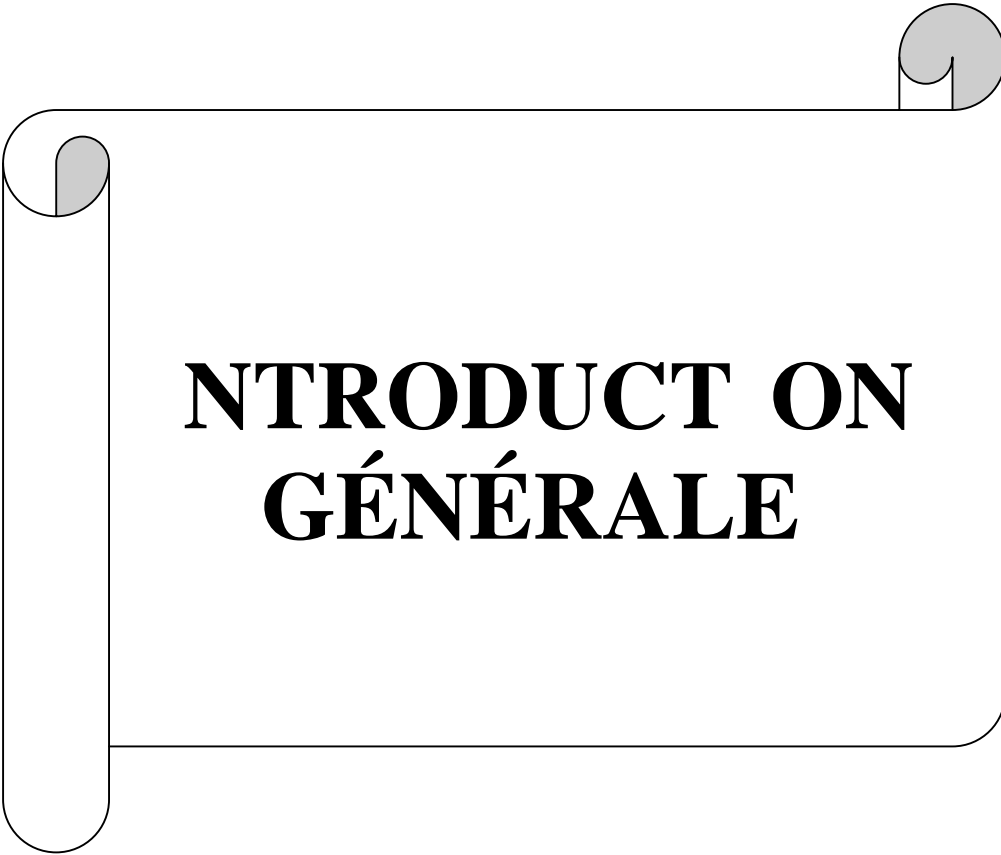
## SYMBOLES ET ABREVIATIONS

<b>SYMBOLE</b>	<b>ABREVIATION</b>
ACP	Analyse en composantes principales
AG	Algorithme génétique
AED	Détecteur à émission électronique.
AM1	Austin Model 1
ATS1m	Broto-Moreau autocorrélation d'une structure topologique - lag 1 / pondérée par les masses atomiques.
ATS7v	Broto-Moreau autocorrélation d'une structure topologique - lag 7 / pondérée par les volumes atomiques de van der Waals.
<b>C</b>	Matrice des coefficients des OM
CHARMM	Chemistry at Harvad Macromolecular Mechanics.
CIRC	Centre international de recherche sur le cancer
CITEPA	Centre interprofessionnel technique d'études de la pollution atmosphérique.
CLHP	Chromatographie en phase liquide haute performance
Cos	Cosinus
CPG	Chromatographie en phase gazeuse
DDE	Dichlorodiphényldichloroéthylène
DDT	Dichlorodiphényltrichloroéthane
DFT	Théorie de la fonctionnelle de la densité
ECD	Détecteur à capture d'électrons.
EPA	Agence de protection de l'environnement.
EEig05d.	La valeur propre 05 de la matrice d'adjacence des cotés pondérée par les moments dipolaires.
$e_i$	Résidu : différence entre les valeurs observée et estimée.
$e_{i \text{ std}}$	Résidu de prédiction standardisé.
$E_{el}$	Energie électronique.
EN	Energie totale de la molécule.
EN de solvation	L'énergie libérée lors la dissolution d'un composé.
EQMC	Erreur quadratique moyenne sur l'ensemble de calibrage.
EQMP	Erreur quadratique moyenne sur l'ensemble de prédiction.
EQMP <sub>ext</sub>	Erreur quadratique moyenne sur l'ensemble de prédiction externe.
<b>F</b>	Matrice de Fock
F	Statistique de Fisher.
FIV	Facteur d'inflation de la variance.
FPD	Détecteur à photométrie de flamme
GATS2e	Autocorrélation Geary - lag 2 / pondéré par les électronégativités atomiques de Sanderson.
GTO	Gaussian Type Orbitals.

H	Matrice de projection, ou matrice chapeau.
HATS0v	Autocorrélation de levier pondéré en fonction du décalage 0 / pondéré par les volumes atomique de vander Waals .
HF	Hartree Fock
HFR	Hartree- Fock -Roothaan
$h_{ii}$	Eléments diagonaux de la matrice chapeau.
HOMO	Plus haute orbitale moléculaire occupée
$H_{\mu\nu}^{ON}$	Intégrales mono-électroniques.
LMO	Validation croisée par omission d'un ensemble d'observations.
LogP	Logarithme de Coefficient de partage octanol / eau
log tr	Logarithme décimal du temps de rétention.
LOO	Validation croisée par omission d'une observation.
LUMO	Plus basse orbitale moléculaire inoccupée
MCO	Méthode des moindres carrés ordinaires
MM	Mécanique moléculaire.
MM2	Mécanique moléculaire 2
MMFF	Merck Molecular Force Field.
MNDO	Molecular Neglected of Differential Overlap.
MS	Spectrométrie de masse
MSD	Détecteur par spectroscopie de masse.
MW (la masse moléculaire)	La somme de tout les poids atomiques des éléments qui constituent le composé.
n	Dimension de la population.
n-p	Nombre de degrés de liberté de la somme des carrés des résidus.
NPD	Détecteur Azote-Phosphore.
nR06	Nombre de cycles à 6 chaînons.
n (val)	Nombre d'électrons de valence du système.
OM	Orbitale moléculaire
OM-CLOA	Orbitale moléculaire-Combinaison linéaire d'orbitales atomiques
OMS	Organisation Mondiale de la Santé.
P	Probabilité
p	Nombre de descripteurs en comptant la constante (nombre de paramètres).
PCBs	Les polychlorobiphényles.
PM3	Parametrization Method 3
PM6	Parametrization Method 6
POP	Polluants Organiques Persistants.
PRESS	Somme des carrés des erreurs de prédiction.
Q <sup>2</sup>	Coefficient de prédiction.



QSAR	Quantitative Structure/ Activity - Relationships.
QSPR	Quantitative Structure/ Property Relationships.
QSRR	Quantitative Structure/ Retention - Relationships.
R <sup>2</sup>	Coefficient de détermination.
RDN	Recouvrement différentiel nul.
RLM	Régression linéaire multiple.
RLS	Régression linéaire simple.
S	Erreur standard.
<b>S</b>	Matrice de recouvrement
SCE	Somme des carrés des écarts.
SCF	Champ auto-Cohérent.
SCT	Somme des carrés totale.
STO	Slater Type Orbitals
T	Valeurs réelles du t de Student.
$\hat{V}_{ee}$	Opérateur d'interaction électronique.
V(i)	Énergie potentielle de l'électron i.
X	Matrice des valeurs observées des variables explicatives.
X'	Matrice transposée de X
v	Matrice diagonale comportant les énergies orbitales.
	Fonction d'onde
	Distribution de la densité électronique
	Constante de Hammett



**INTRODUCTION  
GÉNÉRALE**

## *INTRODUCTION GÉNÉRALE*

---

Les études QSRR sont dérivées d'études QSAR (Quantitative Structure-Activity Relationships) et QSPR (Quantitative Structure-Property Relationships). Le principe de ces études est d'établir une corrélation entre des données structurales des molécules et son activité biologique dans le cas d'études QSAR ou leur propriété physico-chimique dans le cas d'une étude QSPR. Hansch et Fujita établirent, en 1964, les premières corrélations entre les propriétés physico-chimiques ( $\log P$ ,  $pK_a$ , paramètres stériques et électroniques) et l'activité biologique (activités enzymatiques, pharmacologiques) [1]. En 1971, ils réalisent une étude de relation structure-activité sur différentes familles d'antifongiques : benzoquinones, sels d'alkylpyridinium, imidazoles et phénols. Ils observent que quels que soient la famille et le champignon utilisé, l'activité antifongique dépend du  $\log P$  (coefficient de partage eau-octanol) expérimental ou calculé [2]. Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physico-chimiques des analytes avec les temps de rétention obtenus expérimentalement [3] : c'est l'étude quantitative des relations structure-temps de rétention notée QSRR. Toutes ces études s'appuient sur le concept postulant que des structures similaires ont des propriétés similaires. Ce type d'étude permet d'une part, d'expliquer les paramètres moléculaires impliqués dans l'activité biologique, une propriété physicochimique ou l'élution sur une phase stationnaire et de prévoir d'autre part, l'influence de certaines modifications structurales dans l'activité biologique, une propriété physicochimique ou la rétention d'un composé sur une colonne.

La modélisation par ordinateur d'une molécule implique généralement une présentation graphique de la géométrie ou de la configuration des atomes de la molécule, suivie de l'application d'une méthode théorique [4]. La modélisation moléculaire est un terme général qui englobe différentes techniques de graphisme moléculaire et de chimie computationnelle permettant d'afficher, simuler, analyser, calculer et stocker les propriétés des molécules [5].

La modélisation et la simulation s'imposent souvent lorsque l'expérience réelle est :

- trop difficile,
- trop dangereuse,
- trop coûteuse,
- trop longue ou trop rapide,
- éthiquement inacceptable,
- impossible à réaliser.

## *INTRODUCTION GÉNÉRALE*

---

Dans notre thèse, nous avons modélisé et estimé avec succès le temps et l'indice de rétention ainsi que le facteur de bioaccumulation d'un ensemble de produits toxiques (Pesticides, Polychlorobiphenyles (PCBs) et Hydrocarbures aromatiques polycycliques (HAPs)) en utilisant des approches hybrides : algorithmes génétiques /régression linéaire multiple pour le temps et l'indice de rétention, et régression simple pour le facteur de bioaccumulation (AG/RML et AG/RLS).

Cette thèse comporte en plus de la bibliographie, des annexes, d'une introduction et d'une conclusion générale, deux grandes parties distinctes :

Dans la première partie, nous avons défini les différentes molécules étudiées comme des bases de données dans notre travail (Pesticides, PCBs et HAPs).

Dans la deuxième partie nous avons décrit les différentes méthodes d'optimisation des géométries des molécules (MNDO, MM, AM1, PM3, PM6) ; puis nous avons développé et évalué les différents modèles (algorithmes génétiques, robustesse des modèles; validation externe ; test de randomisation).

Dans la partie résultats et discussions, nous présentons et discutons les trois meilleurs modèles développés :

AG/RLM, pour le temps et l'indice de rétention et AG/RLS, pour le facteur de bioaccumulation.

## *INTRODUCTION GÉNÉRALE RÉFÉRENCES BIBLIOGRAPHIQUES*

---

[1] Hansch, C., Fujita T. (1964), “p-s-p Analysis : A method for the correlation of biological activity and chemical structure”, *Journal of the American Chemical Society*, Vol .86 No. 8 , pp. 1616-1662. .

[2] Hansch, C., Lien E.J. (1971),”Structure activity relationships in antifungal agents. A survey”, *Journal of Medicinal Chemistry*, Vol. 14, pp. 653-670.

[3] Tham, S.Y., Agatonovic-Kustrin, S. (2002), “Application of the artificial neural network in quantitative structure-gradient elution retention relationship of phenylthiocarbamyl amino acids derivaties”, *Journal of Pharmaceutical Biomedical Analysis*, Vol. 28 No. 3-4 , pp.581-590.

[4] Clark, T., (1985), *Handbook of Computational Chemistry*, Wiley Edition,London.

[5] Kollman, P. (1996), “Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules”, *Accounts of chemical Research*, Vol. 29 No. 10, pp. 462-469.

# PART I

## *Généralités sur les polluants toxiques:*

Les Pesticides.

Les polychlorobiphényles (PCBs).

Les hydrocarbures Aromatiques

Polycycliques (HAPs).

## A.I.Introduction

Dans les pays industrialisés, la révolution verte des années soixante du siècle précédent a considérablement augmenté la productivité agricole en jouant sur l'augmentation des surfaces cultivées, la mécanisation, la plantation de cultures sélectionnées et hybrides aux rendements plus élevés, le remembrement et la lutte contre toutes les nuisances. Cette lutte passe notamment par le recours massif aux pesticides, qui sont des produits chimiques dangereux destinés à repousser ou tuer les rongeurs, champignons, maladies, insectes et "mauvaises herbes" qui fragilisent le mode de culture intensif. Les pesticides ne sont pas seulement utilisés dans l'agriculture mais aussi dans le jardin du particulier, dans les parcs ouverts au public, pour l'entretien de la voirie, des voies ferrées, des aires de loisirs (golfs, hippodromes...). Les pesticides sont des **Polluants Organiques Persistants** qui perdurent dans l'environnement, s'accumulent dans les graisses et sont, d'une manière générale, dangereux pour la santé : cancers, altération du système immunitaire, problèmes de reproduction,...etc.Les pesticides touchent aussi massivement les zones rurales des PVD (Pays en Voie de Développement) où malformations, cancers, maladies congénitales, désordres du système nerveux déciment la population [1].

## A.II.Histoire et définitions

Avant la seconde Guerre Mondiale, les pesticides employés en agriculture étaient des dérivés de composés minéraux ou de plantes : arsenic, cuivre, zinc, manganèse, plomb, pyrèthre, roténone, sulfate de nicotine... que l'on retrouve en partie dans les cigarettes actuelles.

Ce sont les armes chimiques de la première Guerre Mondiale comme le fameux gaz moutarde (composé de chlore) qui assureront un nouveau débouché industriel pour les pesticides, une fois le conflit terminé. Ainsi, les organochlorés firent leur apparition avec de nombreuses déclinaisons qui ont connu un énorme succès comme le célèbre DDT, interdit en Europe depuis 1972 [1].

## A.III. Classification

Les pesticides disponibles aujourd'hui sur le marché sont caractérisés par une telle variété des structures chimiques, de groupes fonctionnels et d'activités que leur classification est complexe .D'une manière générale, les substances actives peuvent être classées soit en fonction de la nature de l'espèce à combattre (1<sup>er</sup> système de classification), soit en fonction

de la nature chimique de la principale substance active qui les composent (2<sup>ème</sup> système de classification) [2].

### A.III.1. Premier système de classification

Le premier système de classification repose sur le type de parasites à contrôler .Il existe principalement trois grandes familles d'activités que sont les herbicides, les fongicides et les insecticides.

**Les herbicides** représentent les pesticides les plus utilisés dans le monde ,toutes cultures confondues .Ils sont destinés à éliminer les végétaux rentrant en concurrence avec les plantes à protéger en ralentissant leur croissance . Les herbicides possèdent différents modes d'actions sur les plantes, ils peuvent être des perturbateurs de la régulation d'une hormone, (l'auxine) (principale hormone agissant sur l'augmentation de la taille de cellule), de la photosynthèse ou encore des inhibiteurs de la division cellulaire, de la synthèse des lipides, de cellulose ou des acides aminés.

**Les fongicides** permettent quant à eux de combattre la prolifération des maladies des plantes provoquées par des champignons ou encore des bactéries .Ils peuvent agir différemment sur les plantes soit en inhibant le système respiratoire ou la division cellulaire, soit en perturbant la biosynthèse des acides aminés, des protéines ou le métabolisme des glucides.

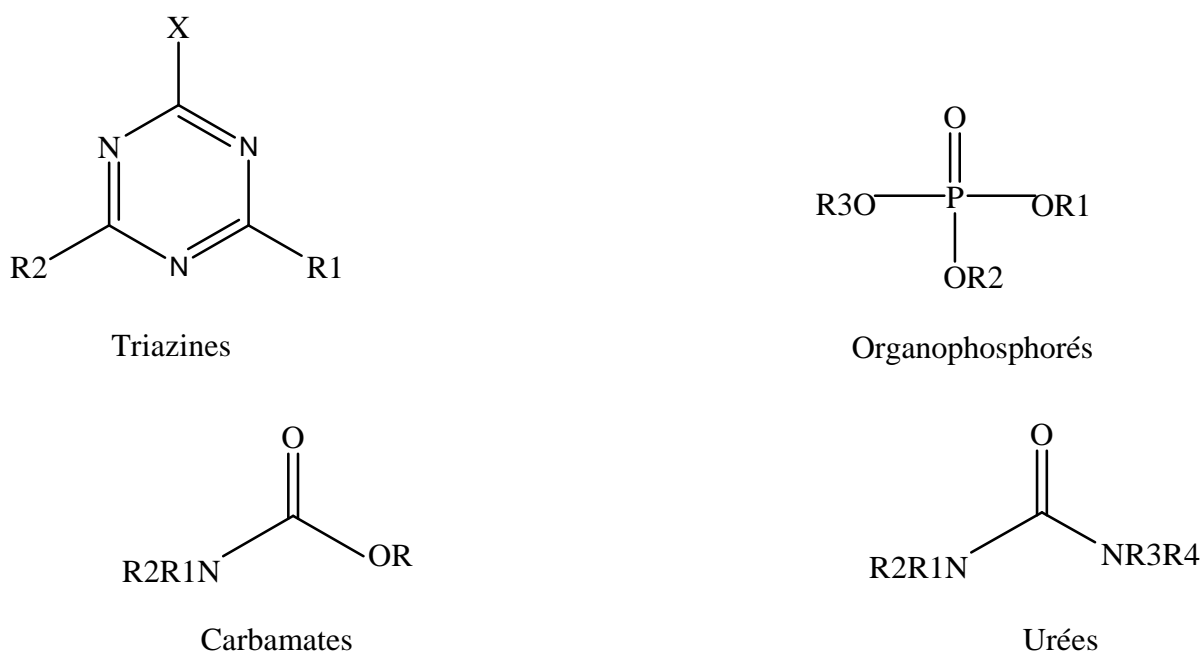
**Les insecticides** sont utilisés pour la protection des plantes contre les insectes .Ils interviennent en les éliminant ou en empêchant leur reproduction, différents types existent : les neurotoxiques, les régulateurs de croissance et ceux agissant sur la respiration cellulaire.

Outre les grandes familles mentionnées ci –dessus, d'autres peuvent être citées en exemple :les acaricides ,contre les acariens ;les nématicides ,contre les vers du groupe des nématodes ;les rodenticides,contre les rongeurs ,les taupicides ,contre les taupes ,les molluscicides ,contre les limaces et escargots ou encore les corvicides et corvifuges ,respectivement contre les corbeaux et les autre oiseaux ravageurs de culture[2].



### A.III.2. Deuxième système de classification

Le deuxième système de classification tient compte de la nature chimique de la substance active qui compose majoritairement les produits phytosanitaires. Compte tenu de la variété des propriétés physico-chimiques des pesticides disponibles sur le marché, il existe un très grand nombre de familles chimiques. Les plus anciens et principaux groupes chimiques sont les organochlorés, les organophosphorés, les carbamates, les triazines et les urées substituées. Les structures chimiques caractéristiques de certaines de ces familles sont présentées en figure I.1



**Figure I.1. Structures chimiques des principales familles de pesticides**

Ce deuxième système de classification ne permet pas de définir de manière systématique un composé. Certains pesticides peuvent en effet être composés de plusieurs fonctionnalités chimiques. Ils peuvent alors être classés dans une ou plusieurs familles chimiques [2].

### A.IV. Les pesticides dans notre environnement :

Déversés dans notre environnement lors des traitements, les pesticides y sont présents partout [3].

## **A.V. Les pesticides et la santé :**

### **A.V.1. Toxicité aiguë :**

Les intoxications aiguës par les pesticides sont celles où, quelques heures après une exposition importante, des symptômes apparaissent rapidement. Les personnes les plus fréquemment victimes d'intoxications aiguës par les pesticides sont bien sûr les agriculteurs, qui manipulent et appliquent ces pesticides sur leurs cultures. L'Organisation Mondiale de la Santé (OMS) a estimé qu'il y a chaque année dans le monde 1 million de graves empoisonnements par les pesticides, avec quelque 220 000 décès. Les jeunes enfants sont aussi très fréquemment victimes d'empoisonnement par les pesticides, habituellement suite à des ingestions accidentelles ou à des atteintes dermatologiques [4].

### **A.V.2. Pesticides et perturbation hormonale :**

Les conséquences de l'exposition à des pesticides perturbateurs endocriniens peuvent être très diverses :

1. Des anomalies congénitales.
2. Des déficits immunitaires.
3. Des problèmes de reproduction.
4. Le développement de certains cancers.
5. Des problèmes neurologiques, cognitifs et comportementaux.

De nombreux pesticides sont soupçonnés d'être des perturbateurs endocriniens.

### **A.V.3. Problème de développement du fœtus et pesticides : les effets des pesticides perturbateurs endocriniens**

Le fœtus en développement et le bébé sont extrêmement sensibles aux effets des pesticides, L'exposition du fœtus à des pesticides à certaines périodes de la grossesse peut conduire à un avortement spontané, à des retards de croissance, des handicaps à la naissance ...

L'exposition du fœtus à des perturbateurs endocriniens (comme certains pesticides) est même soupçonnée de modifier le sexe de l'enfant à naître.

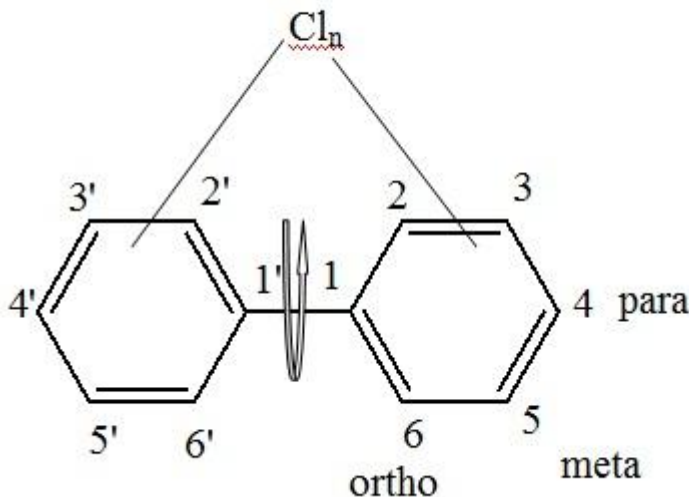
#### **A.V.4. Impact des pesticides sur le système immunitaire :**

Dans la littérature scientifique, l'exposition à certains pesticides à été liée chez l'homme à :

1. Des cancers associés à la suppression immunitaire.
2. Des réactions allergiques (dermites, asthme, anaphylaxie).
3. Des réponses auto-immunes.
4. La suppression de la fonction immunitaire et une plus grande sensibilité aux agents pathogènes.

### B.I. Identité des Polychlorobiphényles (PCBs)

Les polychlorobiphényles (PCBs) sont des composés organiques de 1 à 10 atomes de chlore attachés au biphényle, répondant à la formule chimique générale  $C_{12}H_{10-x}Cl_x$  (figure I.2). Ils étaient fabriqués pour la première fois par Monsanto en 1929, la production de PCBs a été interdite dans les années 1970 en raison de la toxicité élevée de la plupart des (209) PCBs et leurs mélanges [5]. Les PCBs ont été utilisés comme des fluides pour les transformateurs et les condensateurs industriels isolants, et sont connus comme des polluants organiques persistants. Même la production des PCBs a été arrêtée. Ils ont encore une influence sur la santé de l'être humain [6-8] et les animaux en raison de leur accumulation dans l'environnement [9].



**Figure I. 2.** La formule de la structure générale et les positions de substitution des PCBs.

### B.II. Dans quels milieux rencontre-t-on les PCBs ?

Les PCBs sont aujourd'hui très répandus dans l'environnement. Pourtant, ils n'existent pas à l'état naturel et leur production et utilisation sont désormais interdites. Les principales sources d'émission sont donc historiques. Il s'agissait à l'époque principalement des activités industrielles, plus rarement de l'incinération [10]. Aujourd'hui, seules quelques sources de pollution diffuses demeurent, celles qui échappent à la réglementation.

Une fois émis, les PCBs peuvent être aisément transportés sur de longues distances, adsorbés sur des particules en suspension, dans l'air ou dans l'eau ou intégrés dans la chaîne alimentaire. Ils sont ainsi principalement retrouvés dans l'air, l'eau, les sols, les sédiments et les organismes vivants (dont l'organisme humain).

### **B.III. L'exposition humaine et la toxicité aux PCBs**

L'alimentation constitue la principale voie de contamination de la population générale, représentant plus de 90 % de l'exposition totale [10]. L'air ne représente que 3% des apports en PCBs. Les PCBs sont très lipophiles, c'est-à-dire qu'ils présentent une affinité particulière pour les graisses. Ils s'accumulent donc au fur et à mesure de la chaîne alimentaire et sont retrouvés préférentiellement dans les tissus graisseux des animaux et le tissu adipeux des organismes humains. Ils peuvent aussi être retrouvés dans le lait maternel et ils sont capables de franchir la barrière placentaire [10]. Les aliments les plus riches en PCBs sont donc, en premier lieu, les aliments d'origine animale riches en graisses tels que les poissons, les fruits de mer, les crustacés, puis la viande, les produits laitiers et, enfin, les œufs et les végétaux.

Les populations principalement exposées sont les pêcheurs qui consomment régulièrement les produits de leur capture provenant de zones d'eau contaminées.

### **B.IV. Quels sont les effets sur la santé des PCBs ?**

Les effets sont essentiellement liés à la charge corporelle en PCBs c'est-à-dire à l'accumulation de molécules dans l'organisme au cours du temps.

#### **B.IV.1. En toxicité aiguë :**

C'est-à-dire pour des expositions à haute dose (rejets accidentels, activités professionnelles), ce sont généralement des affections de la peau et des effets cutanés (chloracnée, éruptions, pigmentation des ongles et de la peau), des troubles oculaires (hypersécrétion) et hépatiques (altération transitoire de l'activité d'enzymes hépatiques) qui sont observés.

**B.IV.2. En toxicité chronique :**

C'est à dire pour des niveaux d'exposition plus faibles mais récurrents, les manifestations les plus préoccupantes sont des effets neurocomportementaux, des ictères et des dérèglements hépatiques. De tels troubles ont aussi été observés chez le jeune enfant dont la mère avait été fortement exposée aux PCBs pendant la grossesse et la période d'allaitement (diminution du quotient intellectuel de l'enfant, des capacités mnésiques et d'apprentissage, des fonctions neuromusculaires, des capacités visuelles, altération de la peau).

Concernant les effets cancérogènes des PCBs, dès 1987, le Centre international de recherche sur le cancer (CIRC) considère que ce sont des cancérogènes possibles (groupe 2B) pour l'humain (au vu des données restreintes montrant qu'une exposition professionnelle prolongée à des concentrations élevées peut induire une incidence accrue des cancers du foie). Une exposition ponctuelle au travers d'un aliment très contaminé aux PCBs aurait alors peu d'impact sur la santé. En 1997, l'Agence de protection de l'environnement américaine (EPA) confirme le caractère de cancérogène probable.

### C.I.Définition et caractéristiques des hydrocarbures aromatiques polycycliques

Les hydrocarbures benzenoïdes plans ,ou hydrocarbures aromatiques polycycliques (HAPs) sont des molécules organiques issues de la combustion incomplète des matières carbonées [11] suite à des processus naturels ,mineurs [12-14],et des processus anthropiques,majoritaires [15].Les HAPs sont constitués d'atomes de carbone et d'hydrogène formant aux moins deux anneaux aromatiques condensés [16 ,17].Ils sont libérés dans tous les compartiments de l'environnement [18,19].Le nombre des HAPs identifiés à ce jour est de l'ordre de 130 [20].Certains posent des problèmes environnementaux majeurs du fait de leur toxicité.

Les HAPs sont étudiés depuis plus d'un siècle et les propriétés cancérigènes et mutagènes de nombreux HAP sont bien établies, alors que celles de plusieurs autres sont en cours d'investigation. À cause de la pollution générée par l'émission croissante de HAPs dans l'atmosphère (imbrûlés du haut fourneau, de l'entreprise sidérurgique Annaba par exemple), il est impératif de disposer de méthodes qui, à la fois, permettent une identification fiable, et une quantification précise de ces composés.

La chromatographie gazeuse(CG) est largement utilisée pour la séparation et l'analyse des HAPs .Cependant l'analyse d'échantillons renfermant des composés pour lesquels on ne dispose pas d'étalons (particulièrement les isomères) reste complexe, du fait du déficit d'information sur leurs caractéristiques chromatographiques. Aussi, les méthodes de calcul à l'avance des grandeurs de rétention de HAPs à partir de leurs structures sont –elle importantes.

- [1] notre-planete.info, <http://www.notre-planete.info/ecologie/alimentation/pesticides.php>
- [2] El Mrabet, K.(2008), “Développement d’une méthode d’analyse de résidus de pesticides par dilution isotopique associée à la chromatographie en phase liquide couplée à la spectrométrie de masse en tandem dans les matrices céréalières après extraction en solvant chaud pressurisé ”, thèse de doctorat , université Pierre et Marie Curie,France,pp.17-19.
- [3] <https://fr.wikipedia.org/wiki/Pesticide>.
- [4] [www.generations-futures.com/2sommpeostos.html](http://www.generations-futures.com/2sommpeostos.html)
- [5] National Research Council (U.S.) (1979), Committee on the Assessment of Polychlorinated Biphenyls in the Environment. Polychlorinated biphenyls: a report; National Academy of Sciences: Washington.
- [6] Angulo Lucena, R., Farouk Allam, M., Serrano Jiménez, S. and Luisa Jodral Villarejo, M. A. (2007), “A Review of environmental exposure to persistent organochlorine residuals during the last fifty years”. *Current Drug Safety*, Vol. 2 No. 2, pp. 163-172.
- [7] Roveda, A. M., Veronesi, L., Zoni, R., Colucci, M. E. and Sansebastiano, G. (2006), “Exposure to polychlorinated biphenyls (PCBs) in food and cancer risk: recent advances”,*Igiene e Sanita Pubblica*,Vol. 62 No. 6, pp. 677-696.
- [8] Lundqvist, C., Zuurbier, M., Leijns, M., Johansson, C., Ceccatelli, S., Saunders, M., Schoeters, G., Ten Tusscher, G. and Koppe, J. G. (2006),” The effects of PCBs and dioxins on child health”, *Acta Paediatrica*,Vol. 95 No. 453, pp. 55-64.
- [9] Poppenga, R. H. (2000), “Current environmental threats to animal health and productivity”, *The Veterinary Clinics of North America. Food Animal Practice* ,Vol. 16 No. 3, pp. 545-558.
- [10] Organisation mondiale de la santé (2004), « substances chimiques dangereuses : les principaux risques pour les enfants « aide- mémoire » euro/02/04.
- [11] Samanta,S.,Singh,O.V.,Jain,R.k. (2002),“Polycyclic aromatic hydrocarbons:environmental pollution and biomeridation —a review”,*Trends in biotechnology* ”, Vol .20,pp.243-248.



- [12] Wilcke,W.,(2000), “Synopsis polycyclic aromatic hydrocarbons (PAHs) in soil—a review”, *Journal of Plant Nutrition and soil science*,Vol.163 No 3 ,pp.229-248.
- [13] Juhasz,A.L., Naidu,R.(2000), “ Bioremediation of high molecular weight polycyclic aromatic hydrocarbons—a review of the microbial degradation of benzo[ a ]pyrene”, *International Biodeterioration and Biodegradation*,Vol.45,pp.57-88.
- [14] Loi canadienne sur la protection de l’environnement (LCPE)(1994), liste des substances d’intérêt prioritaire –Rapport d’évaluation –hydrocarbures aromatiques polycycliques,pp.69.
- [15] Hill ,A.J.,Ghoshal,S.(2002), “Micellar solubilization of naphthalene and phenanthrene from nonaqueous-phase liquids , *Environmental science & technology*,Vol.36 No18,pp.3901-3907.
- [16] Menzie, C.A., Potocki,B.B.,Santodonato,J.(1992), “Exposure to carcinogenic PAHs in the environment , *Environmental Science & Technology* ,Vol.26 No 7,pp.1278-1284.
- [17] Li, J., Chen, B.H., (2002) “Solubilization of model polycyclic aromatic hydrocarbons by nonionic surfactants”,*Chemical Engineering Science*,Vol.57 No.14 ,pp.2825-2835.
- [18] Rababah, A., Matsuzawa, S. (2002), “Treatment system for solid matrix contaminated with fluoranthene. II—Recirculating photodegradation technique ,*Chemosphere*,Vol.46 No1,pp. 49-57.
- [19] Gabet, S. (2004), “ Remobilisation d’Hydrocarbures Aromatiques Polycycliques (HAP) présents dans les sols contaminés à l’aide d’un tensioactif d’origine biologique’”, Thèse No.12, université de Limoges, France, pp.176.
- [20] Hydrocarbures aromatiques polycycliques, Rapport d’étude (2005),INERIS-DRC-66244-DESP-R01,pp.85.



# PART E II

*Aspects théoriques de la modélisation moléculaire*

Optimisation de la géométrie de la molécule.

Développement et évaluation de la qualité d'un modèle.

## A.I. GÉNÉRAL TÉS

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  $\Psi$ , alors que dans la théorie de la fonctionnelle de la densité (DFT), la distribution de la densité électronique ( $\rho$ ) 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 (M M), appelée parfois : calcul par champ de force empirique, (empirical Force Field, EFF, en anglais), permet le calcul de la structure et de l'énergie d'entités moléculaires [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].

## A.II. MÉTHODES SEM -EMP R QUES UT L SÉES

Les méthodes AM1, PM3 et PM6 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.

### A.II. 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 d'un système à N électrons est représentée par un déterminant de Slater  $\Psi_0$  de spin orbitales unique. Les spin orbitales consistent en des produits d'orbitales moléculaires (OM) et de fonctions de spin (ou  $\alpha$  ou  $\beta$ ),  $\Phi_a = \varphi_a \alpha$ ,  $\bar{\Phi} = \varphi_a \beta$ .

On représentera  $\Psi_0$  par :

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

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

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

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

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 [10,11]). Elles servent dans la description des états excités.

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

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

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  $\mathbf{V}$  une matrice diagonale comportant les énergies orbitales.

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

$$F_{\mu\nu} = H_{\mu\nu} + J_{\mu\nu} - \frac{1}{2} K_{\mu\nu} \quad (4)$$

$$F_{\mu\nu} = H_{\mu\nu} + \sum_{\rho}^n \sum_{\sigma}^m P_{\rho\sigma} \left[ \langle \mu\nu/\rho\sigma \rangle - \frac{1}{2} \langle \mu\sigma/\rho\nu \rangle \right] \quad (5)$$

Avec :

$$H_{\mu\nu} = \int \chi_{\mu}^*(1) \hat{h} \chi_{\nu}(1) d\tau_1 \quad (6)$$

$$\langle \mu\nu/\rho\sigma \rangle = \iint \chi_{\mu}^*(1) \chi_{\nu}(1) \frac{1}{r_{12}} \chi_{\rho}^*(2) \chi_{\sigma}(2) d\tau_1 d\tau_2 \quad (7)$$

$$\text{et } P_{\rho\nu} = 2 \sum_a^m C_{\rho a}^* C_{\nu a} \quad (8)$$

où  $\mu, \nu, \rho$  et  $\sigma$  désignent des orbitales atomiques, et  $H_{\mu\nu}$  des intégrales mono-électroniques représentant les valeurs moyennes de l'opérateur associé à l'énergie cinétique et

l'opérateur énergie potentielle d'interaction noyau – électron ( $\widehat{V}_{en}$ ). Les  $\langle \mu\nu/\rho \sigma \rangle$  sont des intégrales de répulsion bi-électroniques représentant  $\widehat{V}_{ee}$  (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  $\widehat{J}$  et d'échange  $\widehat{K}$  respectivement.

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

$$E_{el} = 2 \sum_a^m \varepsilon_a - \frac{1}{2} \sum_{\mu \nu}^m P_{\mu \nu} \left( J_{\mu \nu} - \frac{1}{2} K_{\mu \nu} \right) \quad (9)$$

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) [12].

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) [13,14].

Avec 
$$\mathbf{F}^{\rangle} \mathbf{C}^{\rangle} = \mathbf{S}^{\rangle 1/2} \mathbf{F} \quad (10)$$

$$\mathbf{F}^{\rangle} = \mathbf{S}^{\rangle 1/2} \mathbf{F} \mathbf{S}^{\rangle 1/2} \quad (11)$$

et :

$$\mathbf{C}^{\rangle} = \mathbf{S}^{1/2} \mathbf{C} \quad (12)$$

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

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

### A.II.2 .Les méthodes semi-empiriques

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

- 1) Dans la construction de  $\rho_0$ : 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  $\mathbf{F}^j$  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) [15], 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 [16] 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 :

$$\hat{H}_{\text{val}} = \sum_{i=1}^{n(\text{val})} \left[ -\frac{1}{2} \nabla_i^2 + V(i) \right] + \sum_{i=1}^{n(\text{val})} \sum_{j>i} \frac{1}{r_{ij}} \quad (13)$$

qui se simplifie en :

$$\hat{H}_{\text{val}} = \sum_{i=1}^{n(\text{val})} \hat{H}_{\text{val}}^c(i) + \sum_{i=1}^{n(\text{val})} \sum_{j>i} \frac{1}{r_{ij}} \quad (14)$$

Où :

$$\hat{H}_{\text{val}}^c(i) = \left[ -\frac{1}{2} \nabla_i^2 + V(i) \right] \quad (15)$$

$n(\text{val})$  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  $\hat{H}_{\text{val}}$ .

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

$$F_{\text{val},rs} = H_{\text{val},rs}^c + \sum_{t=1}^b \sum_{u=1}^b P_{tu} \left[ (rs|tu) - \frac{1}{2} (ru|ts) \right] \quad (16)$$

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

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

$$H_{\mu_A \mu_B}^c = \frac{1}{2} \left( \beta_{\mu_A} + \beta_{\nu_B} \right) S_{\mu_A \nu_B} \quad ; \quad A \neq B \quad (17)$$

Où : les  $\beta$  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  $\beta_{C2s}$  et  $\beta_{C2p}$ .

Les éléments de la matrice de cœur à partir d'orbitales atomiques différentes centrées sur le même atome sont fournis par l'équation (15) :

$$H^c(1) = -\frac{1}{2} \nabla_1^2 + V(1) \quad , \quad \text{où } V(1) \text{ est l'énergie potentielle de l'électron de}$$

valence 1 dans le champ du cœur. Décomposant  $V(1)$  en contributions individuelles de cœurs atomiques, il vient :

$$H^c(1) = -\frac{1}{2} \nabla_1^2 + V_A(1) + \sum_{B \neq A} V_B(1) \quad (18)$$

Ainsi :

$$H_{\mu_A \nu_B}^c = \left\langle \mu_A \left| -\frac{1}{2} \nabla_1^2 + V_A \right| \nu_B \right\rangle + \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| V_B \right\rangle \quad (19)$$

Des considérations de la théorie des groupes [17] permettent d'annuler

$\langle \mu_A | -\frac{1}{2} \nabla^2 + V_A | \nu_A \rangle$ , de telle sorte que :

$$H_{\mu_A \nu_B}^c = \sum_{B \neq A} \langle \mu_A | \nu_B | \nu_A \rangle \quad (20)$$

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

$$V_B = -\frac{C_B}{r_{1B}} \quad (21)$$

$$\langle \mu_A | \nu_B | \nu_A \rangle = -C_B \langle \mu_A | \frac{1}{r_{1B}} | \nu_A \rangle \quad (22)$$

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 centrée sur l'atome B :

$$H_{\mu_A \nu_B}^c = \sum_{B \neq A} \langle \mu_A | \nu_B | \nu_A \rangle = -\sum_{B \neq A} C_B \langle \mu_A \nu_A | s_B s_B \rangle; \mu_A \neq \nu_A \quad (23)$$

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

$$H_{\mu_A \mu_A}^c = \langle \mu_A | -\frac{1}{2} \nabla^2 + V_A | \nu_A \rangle + \sum_{B \neq A} \langle \mu_A | \nu_B | \nu_A \rangle \quad (24)$$

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

$$H_{\mu_A \nu_A}^c = U_{\mu_A \mu_A} \sum_{B \neq A} C_B \langle \mu_A \nu_A | s_B s_B \rangle \quad (25)$$

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 intégral mono-centre et la distance inter - nucléaire en utilisant une procédure d'expansion multipole [18].

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

$$V_{CC} = \sum_{B \neq A} \sum_A [C_A C_B (s_A s_B / s_B s_B) + f_{AB}] \quad (26)$$

où :

$$f_{AB} = f_{AB}^{MNDO} = \left[ C_A C_B (s_A s_B / s_B s_B) \left( e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} \right) \right] \quad (27)$$

$\alpha_A$  et  $\alpha_B$  sont les paramètres des atomes A et B. Pour les paires O-H et N-H, par exemple, on aura :

$$f_{AH}^{MNDO} = \left[ (C_A C_H (s_A s_H) s_H s_H) \left( R_{AH} e^{-\alpha_A R_{AH}} + e^{-\alpha_H R_{AH}} \right) \right] \alpha_A \alpha_H \quad (28)$$

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  $\xi$  de la STO. Pour la MNDO  $\xi_s = \xi_p$ .

3)  $\beta_s$  et  $\beta_p$ . 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 :

$$f_{AB}^{AM1} = f_{AB}^{MNDO} + \frac{C_A C_B}{R_{AB}} \left[ \sum_k a_{kA} \exp \left[ -b_{kA} (R_{AB} - C_{BA})^2 \right] \right] + \frac{C_A C_B}{R_{AB}} \left[ \sum_k a_{kB} \exp \left[ -b_{kB} (R_{AB} - C_{kB})^2 \right] \right] \quad (29)$$

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.

❖ La méthode PM6 :

Exploite la plupart des approximations utilisées dans AM1 et PM3 tout en faisant intervenir de nombreuses modifications.

Ainsi la proposition de Voityuk et Roch [19] de faire intervenir des paramètres diatomiques dans le terme d'interaction cœur-cœur pour la méthode AM1(d) (qui est une extension de AM1 à une base d'orbitales spd par addition d'orbitales d dans la paramétrisation du Molybdène) a-t-elle été adoptée par Stewart [20] pour l'interaction cœur-cœur dans PM6, à la place des expressions utilisées dans les méthodes MNDO et AM1.

Ainsi, l'expression de la fonction de répulsion de cœur est donnée par :

$$E_{AB}^{PM6} = Z_A Z_B (s_A s_A / s_B s_B) (1 + u_{AB} e^{-r_{AB}} (R_{AB} + 0.0003 R_{AB}^6)) \quad (30)$$

Pour de nombreuses interactions diatomiques, la forme générale de la fonction de répulsion de cœur présentée a été modifiée quand des erreurs spécifiques ont été détectées par suite de l'inadéquation de l'approximation introduite.

Après optimisation, il a été constaté que l'énergie d'interaction de liaison hydrogène était trop faible ce qui fut corrigé en modifiant la fonction de répulsion de cœur uniquement pour les interactions C-H et O-H

$$E_{AH}^{PM6} = Z_A Z_H (s_A s_A / s_H s_H) (1 + u_{AH} e^{-r_{AH}} R_{AH}^2) \quad (31)$$

Où A représente l'atome de carbone ou d'oxygène.

D'autres modifications sont apportées dans plusieurs autres cas [21].

Notons que la méthode PM6 a été paramétrée sur 70 éléments ce qui permet un vaste domaine d'applications. La modification du terme d'interaction cœur-cœur a permis une amélioration significative pour les éléments des groupes principaux, alors que l'utilisation d'orbitales *d* a permis l'extension de la méthode aux métaux de transition.

### A.II.3. Champ de force

#### A.II.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.

#### A.II.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/mm2mm3.html) introduit par Allinger et al [22-25], 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 [26] très largement utilisé dans le traitement des protéines et des acides nucléiques.
- ❖ **CHARMM**: [http:// yuri. Harvard. Edu](http://yuri.harvard.edu) (Chemistry at Harvard Macromolecular Mechanics) développé par Mackerall, Karplus et al. [24] 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 [ 27-29], 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.

### **A.II.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

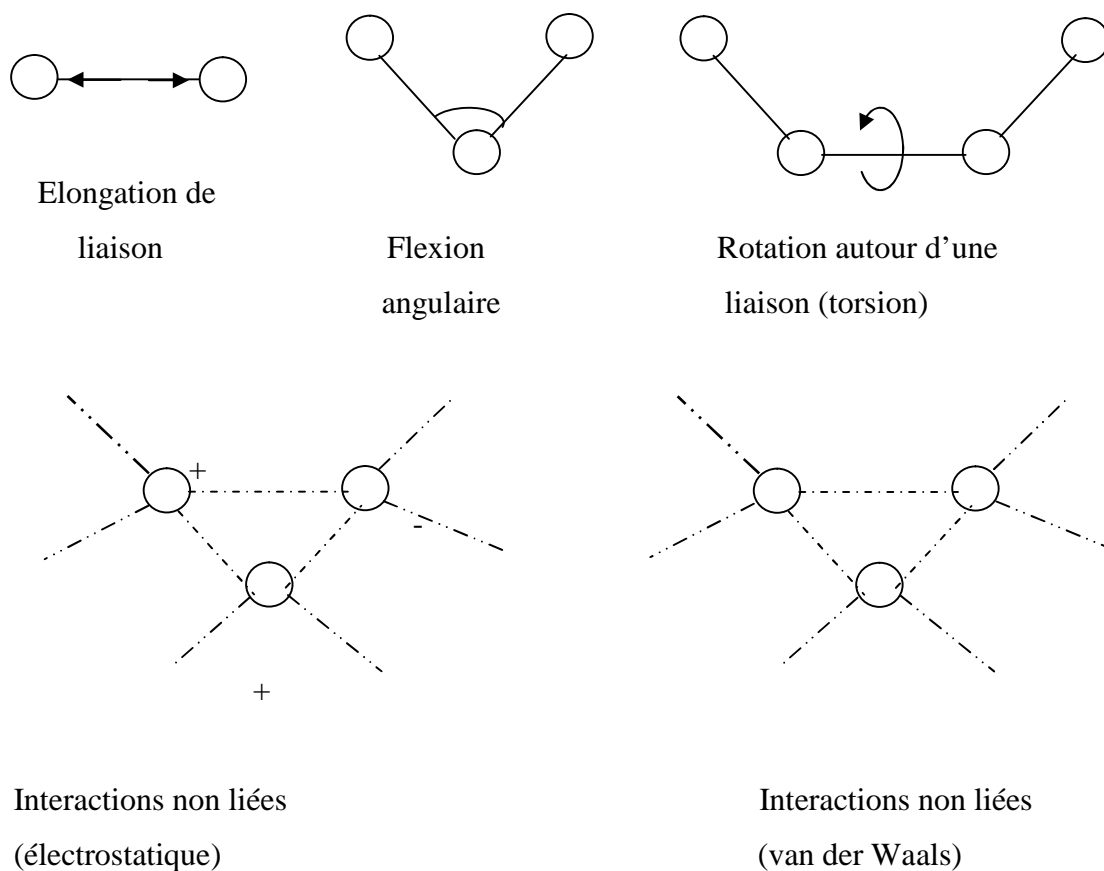
du champ de force affectent ses performances, et aident également dans le processus de paramétrisation.

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

$$\begin{aligned}
 V(\mathbf{r}^N) = & \sum_{\text{liaisons}} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} (1 - \cos(n\omega - \gamma)) \\
 & + \sum_{i=1}^N \sum_{j=i+1}^N \left( 4 \epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \quad (32)
 \end{aligned}$$

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



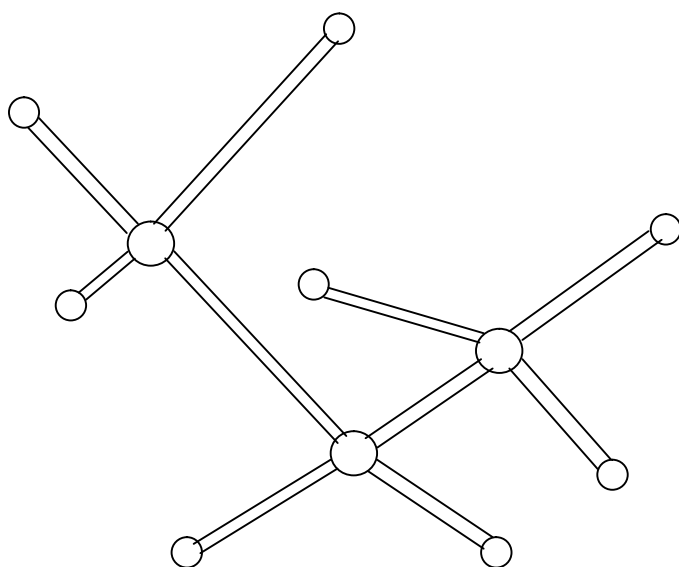
**Fig. II.1 :** Représentation schématique des contributions à un champ de force de MM.



Le premier terme de l'équation (32) 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 (32) 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.

#### A.II.3. 4. Exemple de calcul : énergie d'une conformation du propane.

A titre d'illustration nous montrons comment la relation (32) peut être utilisée pour calculer l'énergie de conformation du propane (Fig.II.2).



**Fig .II.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 ( $\text{CH}_2$ ) et un groupe correspondant aux 6 hydrogènes liés aux carbones des groupements méthyl.

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  $\sigma$  et  $\epsilon$  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.

### A.II.3 .5 . Champs de force MM2 et MM+

#### A.II.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 :

$$v(r) = \frac{k}{2} (r - r_0)^2 [1 - k'(r - r_0)] \quad (33)$$

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

$$v(\theta) = \frac{k}{2} (\theta - \theta_0)^2 \quad (34)$$

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.

$$v(\theta) = \frac{k}{2} (\theta - \theta_0)^2 \left[ 1 - k'(\theta - \theta_0)^2 \right] \quad (35)$$

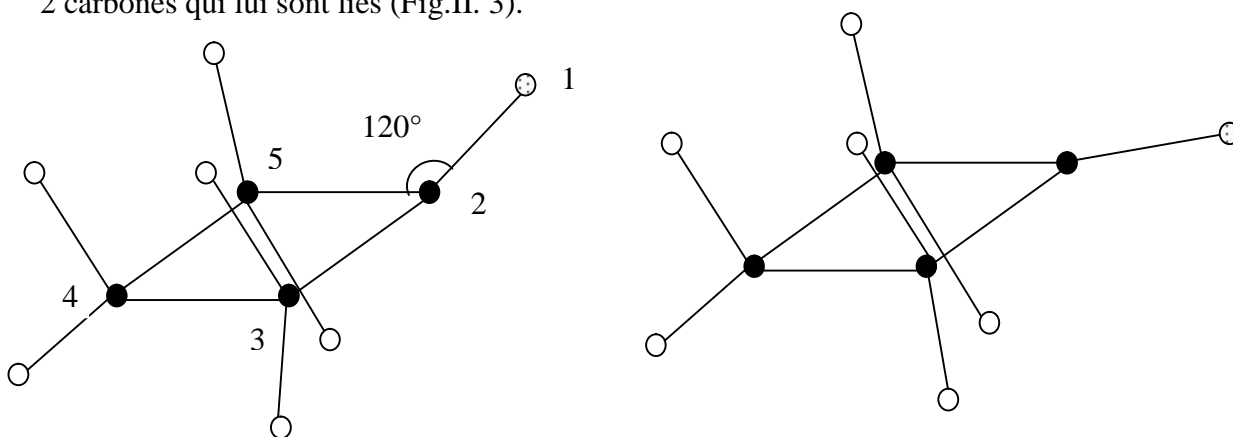
- ❖ Torsion des angles dièdres : correspond à la rotation d'une liaison selon l'angle dièdre formé par 4 atomes. Le champ de force MM2 utilise 3 termes d'une série de Fourier :

$$v(\omega) = \frac{V_1}{2} (1 + \cos \omega) + \frac{V_2}{2} (1 - \cos 2\omega) + \frac{V_3}{3} (1 + \cos 3\omega) \quad (36)$$

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 (36). 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.II. 3).



**Fig.II.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  $120^\circ$ . Expérimentalement, il est établi que l'atome d'oxygène demeure dans le plan du cyclobutane bien que les angles C-C=O soit grands ( $133^\circ$ ). Ceci

parce que l'énergie de liaison, 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 .

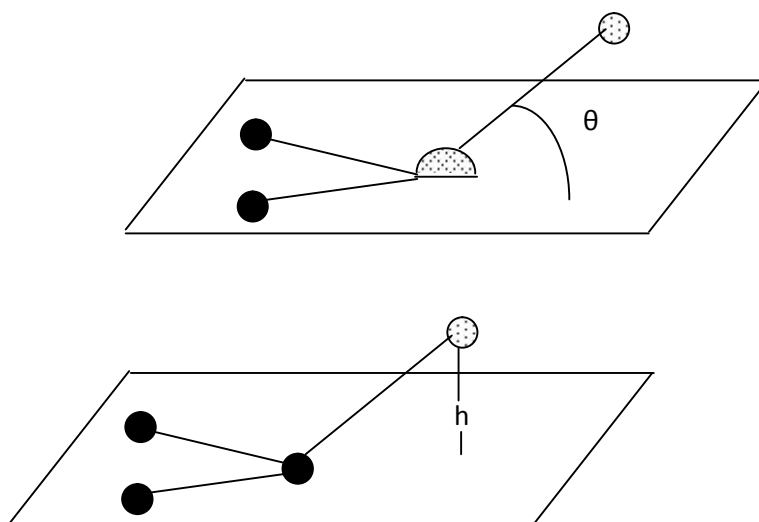
Un potentiel de torsion de la forme suivante :

$$v(\omega) = k (1 - \cos 2\omega) \quad (37)$$

peut être utilisé pour maintenir l'angle de rotation impropre à  $0^\circ$  ou  $180^\circ$ .

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  $0^\circ$  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 :

$$v(\theta) = \frac{k}{2} \theta^2 \quad ; \quad v(h) = \frac{k}{2} h^2 \quad (38)$$



**Fig.II.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 :

$$v_{\Delta\theta} = \frac{1}{2} k_{12} (\Delta l_1 + \Delta l_2) \Delta \theta \quad (39)$$

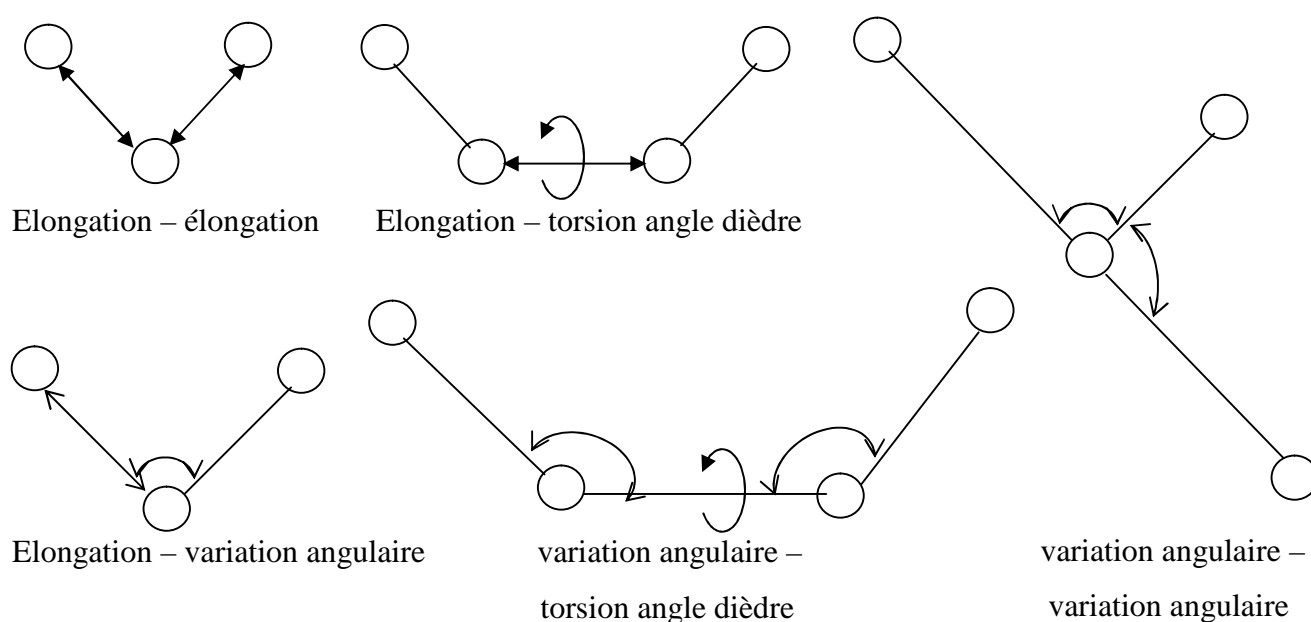
avec  $l_1 = l_1 - l_{10}$  ;  $l_2 = l_2 - l_{20}$  et  $\theta = \theta - \theta_0$

$l_{10}$ ,  $l_{20}$  et  $\theta_0$  représentent les valeurs de référence pour  $l_1$ ,  $l_2$  et  $\theta$  respectivement.

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

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

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



**Fig.II.5 :** Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces.

- ❖ Interactions électrostatiques : Le terme électrostatique  $v_{es}$  est généralement défini par la somme des interactions électrostatiques entre toutes les paires d'atomes exceptées les paires 1, 2 et 1, 3 :  $v_{es} = \sum_{1, \geq 4} v_{es, ij}$ , où les atomes  $i, j$  vérifient la relation  $(1, \geq 4)$ .

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

MM2 n'utilise pas cette procédure mais associe un moment dipolaire avec chaque liaison puis calcule  $V_{es}$  comme somme des énergies potentielles d'interactions entre

moments de liaisons. Chaque moment dipolaire étant localisé au centre, et dirigé le long de cette liaison. Les valeurs des moments de liaisons de certains types de liaisons ont été choisies pour ajuster les moments dipolaires expérimentaux de petites molécules.

L'équation (40) décrit un modèle de charge, basé essentiellement sur les dipôles centrés, utilisé dans MM2 [30].

$$v_{es} = \frac{\mu_i \mu_j}{k r^3} (\cos \chi - 3 \cos \alpha_i \cos \alpha_j) \quad (40)$$

$\chi$  et  $\alpha_i$ ,  $\alpha_j$  désignent respectivement les angles entre les dipôles et les angles entre chaque dipôle et le vecteur de connexion.

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

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

$$v_{vdw} = A e^{-\alpha r} - \frac{B}{r^6} \quad (41)$$

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

### A.II.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 compare les trois techniques computationnelles majeures évoquées.

**Tableau II.1** : Étude comparative des techniques ab initio, semi-empirique et mécanique moléculaire (d'après [31]).

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



**B.I.Les modèles :****B.I.1.La régression linéaire simple (RLS) :**

La plus simple et la plus populaire des techniques de régression, est un modèle particulier de régression dans lequel :

- Il n'y a qu'une seule variable explicative (numérique).
- Le modèle est linéaire dans la variable, dans les paramètres.

Comme toutes les techniques prédictives, elle a deux objectifs :

- Construire un modèle dont les paramètres soient interprétables par le praticien en termes de propriétés de la population dont est extrait l'échantillon. On espère bien entendu, que les paramètres du modèle soient de bons estimateurs des paramètres de la population sous-jacente.
- Utiliser le modèle pour faire des prédictions.

RLS traite la question suivante :

- Une grandeur  $y$  est mesurée.
- Pour un certain nombre de valeurs d'une autre grandeur  $X$ .
- Dans un premier temps, la RLS cherche à matérialiser le fait que les points expérimentaux sont approximativement alignés. Elle le fait en identifiant la meilleure droite » passant au travers du nuage de points .Cette droite, dite « droite des moindres carrés »(DMC) sera caractérisée par une pente  $b$  et une ordonnée à l'origine  $a$ . Ces grandeurs seront les deux premiers paramètres(ou coefficients) du modèle de la RLS.

**B.I.2.La régression linéaire multiple :**

Un modèle de régression multiple entre une variable expliquée  $Y$  et  $p$  variables explicatives  $X_1, \dots, X_p$ , s'écrit pour tout  $i=1, \dots, n$  :

$$y_i = S_0 + \sum_{j=1}^p S_j x_{ij} + V_i \quad (42)$$

où les  $y_i, x_{i1}, x_{i2}, \dots, x_{ip}$  sont des données respectivement relatives aux variables  $Y, X_1, \dots, X_p$ .

Les estimateurs des coefficients  $\beta_j$  sont calculés en utilisant la méthode des moindres carrés ordinaires. Les variables aléatoires  $\varepsilon_i$  représentent les termes d'erreur non observables du modèle. On peut estimer ces erreurs par les résidus ordinaires  $e_i$ , différences entre les valeurs observées  $y_i$  et les valeurs estimées  $\hat{y}_i$ .

Deux paramètres statistiques sont couramment utilisés pour l'évaluation de la qualité du modèle :

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (43)$$

Où  $\bar{y}$  est la valeur moyenne des valeurs observées.

- La racine de l'écart quadratique moyen de prédiction :

$$\dagger_n \equiv EQMP = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2} = \sqrt{\frac{PRESS}{n}} \quad (44)$$

La validation croisée par "leave-one-out" (LOO) consiste à recalculer le modèle sur (n-1) objets, et à utiliser le modèle ainsi obtenu pour calculer les variables à expliquer du composé écarté, notée  $\hat{y}_{(i)}$ . On répète le procédé pour chacun des n-objet. La somme des carrés des erreurs de prédiction, désignée par le symbole PRESS (éq. (44)), est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \quad (45)$$

Contrairement à  $R^2$ , qui augmente avec le nombre de paramètres du modèle, le facteur  $Q^2_{LOO}$  affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au coefficient  $Q^2_{LOO}$ . Une valeur de  $Q^2_{LOO} > 0,5$  est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [32]. Si de petites valeurs de  $Q^2$  indiquent des modèles peu robustes, caractérisés par de faibles capacités prédictives internes, le contraire n'est pas nécessairement vrai. En fait, si une forte valeur de  $Q^2$  est une condition nécessaire de robustesse et d'une possible capacité prédictive élevée d'un modèle, cette condition seule n'est pas suffisante, et peut conduire à une surestimation de la capacité prédictive du modèle, lorsqu'il est appliqué à des composés réellement externes.

En fait, si une forte valeur de  $Q^2_{LOO}$  est une condition nécessaire d'une possible capacité prédictive élevée d'un modèle, cette condition seule n'est pas suffisante.

Pour éviter une surestimation de la capacité prédictive du modèle on a également appliqué la procédure « leave-more-out » (LMO), répétée 8000 fois, en excluant 50% des objets à chaque étape ( $Q^2_{LMO}/50$ ). Les modèles QSPR/QSRR, à cause (souvent) de leur complexité et de la sophistication des outils de chimiométrie employés, peuvent constituer une source de corrélations fortuites. Dans le but d'établir que le modèle n'est pas dû au hasard, on a appliqué le test de randomisation de Y (Y-scrambling) [33]. Ce test consiste à générer un vecteur de la propriété étudiée par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu un modèle QSRR, selon la méthode habituelle. Ce procédé est répété 100 fois dans ce travail. Dans la technique de validation par bootstrap on simule de nouveaux échantillons de taille(n), par tirages aléatoires avec remise. De cette façon l'ensemble de calibrage, qui conserve sa taille initiale (n), se compose, en général, d'objets répétés, l'ensemble de test rassemblant les objets exclus [34]. Le modèle est calculé sur l'ensemble de calibrage et les réponses prédites pour l'ensemble de test. Tous les carrés des différences entre les valeurs prédites et observées des objets de l'ensemble de test sont collectés dans le PRESS. Cette procédure de construction des ensembles de calibrage et d'évaluation est répétée plusieurs milliers de fois (8000 dans notre cas), les PRESS sont additionnés, et une capacité de prédiction moyenne est calculée [35].

Pour éviter des modèles présentant des problèmes de colinéarité et sans réelle capacité de prédiction, nous avons appliqué la règle QUIK (Q Under Influence of K) de Todechini et al. [36], basé sur l'indice de corrélation multivariable K. Cette règle est déduite de l'hypothèse que la corrélation totale dans l'ensemble formé par les prédicteurs X du modèle plus la réponse Y,  $K(x,y)$  doit toujours être plus grande que celle uniquement mesurée sur l'ensemble des prédicteurs  $K(xx)$ . On rejette les modèles qui ne vérifient pas la condition :

$$D(K) = K_{xy} - k_{xx} > 0 \quad ( )$$

Certains modèles, parmi les 100 calculés par le logiciel MOBY-DIGS, peuvent présenter des performances similaires et conduire, plus au moins, aux mêmes capacités prédictives. Dans ce cas on sélectionne les modèles avec les DK les plus élevés, que l'on vérifie par la suite.

En général, les meilleurs modèles sont sélectionnés en maximisant DK et en essayant de trouver un compromis de la prédictivité externe ( $Q^2_{Loo} > 0,7$  ;  $Q^2_{Boot} > 0,6$ ).

Le  $Q^2_{ext}$  externe pour l'ensemble de test est déterminée par l'équation suivante:

$$Q^2_{ext} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_{tra})^2} \quad (47)$$

La validation des modèles est complétée en utilisant l'ensemble de test. d'après Golbraikh et al. [37] Un model QSRR /QSPR possède une capacité de prédiction acceptable s'il vérifie les conditions suivantes :

$$Q^2_{ext} > 0,5 \quad ; \quad r^2 > 0,6 \quad (48)$$

$$(r^2 - r^2_0) / r^2 \quad \text{ou} \quad (r^2 - r^2_0) / r^2 < 0,1 \quad (49)$$

$$0,85 \leq k \leq 1,1 \quad \text{ou} \quad 0,85 \leq k' \leq 1,1 \quad (50)$$

$Q^2_{\text{ext}}$  est le coefficient de prédiction pour l'ensemble de test ;  $r$  est le coefficient de corrélation entre les valeurs calculées et expérimentales de l'ensemble de test,  $r^2_0$  et  $r'^2_0$  sont les coefficients de détermination,  $r^2_0$  (valeurs calculées en fonction de celle observées) et  $r'^2_0$  (valeurs observées en fonction de celles calculées);  $K$  et  $K'$  sont les pentes des droites de régression passant par l'origine, respectivement des valeurs calculées en fonction de celles observées, et des valeurs observées en fonction de celles calculées.

## B.II.DÉVELOPPEMENT ET ÉVALUATION DE MODÈLE

### B.II. 1.Robustesse du modèle :

La stabilité du modèle a été explorée en utilisant la « validation croisée par omission d'une observation (LOO : cross validation by leave one out) [38].

### B.II.2.Domaine d'application :

Le domaine d'application a été discuté à l'aide du diagramme de Williams (traité en détail dans [32,39], représentant les résidus de prédiction standardisés en fonction des valeurs des leviers  $h_i$ . L'équation (51) définit le levier d'un composé dans l'espace original des variables indépendantes ( $x_i$ ):

$$h_i = \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \quad (i = 1 \dots n), \quad (51)$$

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

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

### B.II. 3.Test de randomisation :

Ce test permet de mettre 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 QSPR, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 fois dans notre cas) [40].

#### B.II. 4. Validation externe :

En plus du test de randomisation, il est intéressant [41], pour juger la qualité du modèle, de considérer la racine de l'écart quadratique moyenne (RMSE pour Root Mean Squared Error ), calculée sur différents ensembles :

- Ensemble d'estimation (appelée EQMC).
- Ensemble de validation croisée (appelé EQMP).
- Ensemble de prédiction externe (désignée par EQMP<sub>ext</sub>).
- Ces valeurs RMSE sont mieux adaptées, pour juger de la qualité d'un modèle que les valeurs de R<sup>2</sup> et Q<sup>2</sup> seules, qui constituent de bons tests uniquement pour les données réparties régulièrement.

$$EQMC = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (52)$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_{(i)})^2}{n_{ext}}} \quad (53)$$

#### B.II.5. Sélection d'un sous ensemble de descripteurs :

Des logiciels spécialisés permettent le calcul plus de 6000 descripteurs moléculaires appartenant à différentes classes .Plutôt que de rechercher à expliquer la variable dépendante (grandeur d'intérêt) par tous les régresseurs qui donnent 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 évolutifs et génétiques.

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

### **B.II.5.1- Algorithme de Kennard et Stone (CADEX):**

L'algorithme Kennard et Stone [42] 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 (éch i), l'algorithme :

- ❖ calcule la distance vers chaque échantillon déjà sélectionné ;
- ❖ attribue à (éch i) la plus petite des distances.

L'échantillon (éch i) associé à la plus grande distance est donc le plus éloigné de tous les échantillons déjà sélectionnés ; c'est donc lui 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 de Kennard et Stone est considéré parmi les meilleures méthodes de construction des ensembles de calibrage et de test pour les différentes bases de données [43].

### **B.II.6.ALGORITHME GÉNÉTIQUE:**

Les algorithmes génétiques appartiennent à la famille des algorithmes évolutionnistes (un sous-ensemble des métaheuristiques). Leur but est d'obtenir une solution approchée, en un temps correct, à un problème d'optimisation, lorsqu'il n'existe pas (ou qu'on ne connaît pas) de méthode exacte pour le résoudre en un temps raisonnable. Les algorithmes génétiques utilisent la notion de sélection naturelle développée au XIX<sup>e</sup> siècle par le scientifique Darwin et l'appliquent à une population de solutions potentielles au problème donné. On se rapproche par "bonds" successifs d'une solution, comme dans une procédure de séparation et évaluation, à ceci près que ce sont des formules qui sont recherchées et non plus directement des valeurs.

**B.II.6. 1. Les origines:**

L'utilisation d'algorithmes génétiques, dans la résolution de problèmes, est à l'origine le fruit des recherches de John Holland et de ses collègues et élèves de l'Université du Michigan qui ont, dès 1960, travaillé sur ce sujet. La nouveauté introduite par ce groupe de chercheurs a été la prise en compte de l'opérateur d'enjambement en complément des mutations. Et c'est cet opérateur qui permet le plus souvent de se rapprocher de l'optimum d'une fonction en combinant les gènes contenus dans les différents individus de la population. Le premier aboutissement de ces recherches a été la publication en 1975 de « *Adaptation in Natural and Artificial System* » [44].

**B.II .6.2. 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" ), 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.

**B.II .6.3. 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  $P$ ) ;



**B.II.6.4. Étape 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 le cas contraire, il n'est pas pris en compte. Cette procédure est répétée pour de nombreuses paires (100 fois par exemple).

**B.II.6. 5. Étape 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. Cette procédure est répétée pour tous les chromosomes, c'est-à-dire  $P$  fois.

**B.II. 6.6. Conditions d'arrêt**

Les étapes 2 et 3 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 [45], calculée selon :

$$\text{FIT} = \frac{R^2}{1 - R^2} \frac{n - p - 1}{n + p^2} \quad (54)$$

$p$  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 variable  $p$  différent.

- [1] Hohenberg , P., Kohn,W. (1964), “ Inhomogeneous Electron Gas”, *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”, *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”, *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”, *The journal of chemical physics* , Vol. 41 No. 12, pp. 3663.
- [7] Sutcliffe , B. T. (1997), “ The Nuclear Motion Problem in Molecular Physics”, *Advances in Quantum Chemistry*, Vol. 28 , pp. 65.
- [8] Roothan, C.C.J. (1951), “New Developments in Molecular Orbital 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”,*Proceeding of the Royal Society Of London A*, Vol.205 No 1083,pp.541.
- [10] Koopmans, T. A. (1933), “The distribution of wave function and characteristic value among the individual electrons of an atom”, *Physica*, Vol. 1 , pp.104-113.
- [11] Koopmans, T. A. (1933), “ Ordering of Wave Functions and Eigenenergies to the Individual Electrons of an Atom”, *Physica*, Vol. 1 , pp.104-113.
- [12] Blinder, S. M. (1965), “Basic Concepts of Self-Consistent-Field Theory”, *American Journal of Physics*, Vol.33,pp.431.
- [13] Löwdin, P. O. (1950), “On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals”,*The journal of Chemical Physics*, Vol. 18,pp.365.
- [14] Löwdin P. O., (1970),“ On the Orthogonality problem”,*Advances in Quantum Chemistry*, Vol 5, pp 185-199.

- [15] Dewar, M. J. S. , Thiel ,W. ( 1977), “Ground States of Molecules.38. The MNDO Method. Approximations and Parameters ”, *Journal of the American Chemical Society* , Vol .99 No. 15, pp.4899-4907.
- [16] 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”, *Journal of the American Chemical Society*, Vol .107,pp.3902-3909.
- [17] Stewart J. J. P. (1989), “Optimization of parameters for semiempirical methods I. Method”, *Journal of Computational Chemistry*,Vol. 10 No. 2, pp. 209–220.
- [18] Dewar, M. J. S. , Thiel, W. (1977), “A Semiempirical Model for the Two-Center Repulsion Integrals in the NDDO Formalism”,*Theoretica chimica acta*, Vol .46,pp. 89-104
- [19] Voityuk, A.A.,Rösch, N.(2000),“ AM1/d Parameters for Molybdenum”, *The Journal of Physical Chemistry A*, Vol. 104 No. 17,pp. 4089-4094.
- [20] Stewart, J.J.P. (2007), “Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements”, *Journal of Molecular Modeling*, Vol. 13, pp. 1173-1213.
- [21] Hakan, K . (2009), “Parameterization of the AM1\* semiempirical molecular orbital method for the first-row transition metals and other elements”, thèse de doctorat . Friedrich-Alexander-Universität Erlangen-Nürnberg.
- [22] Allinger ,N.L., (1977), “Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms”, *Journal of the American Chemical Society*, Vol. 99 No. 25, pp. 8127-8134.
- [23] Burkert, U., Allinger, N.L., (1982, 1986) “ Molecular Mechanics, ACS Monograph No.177, *American Chemical Society*, Washington, DC.
- [24] Allinger, N. I., Yuh, Y.H, , Lii, J.H., (1989), “ Molecular mechanics. The MM3 force field for hydrocarbons”, *Journal of the American Chemical Society*, Vol . 111, pp. 8551-8565.
- [25] Allinger, N.I., Chem, k., Lii, J.H., J. (1996), “An improved force field (MM4) for saturated hydrocarbons”, *journal of Computational Chemistry*, Vol. 17, pp . 642-668 .
- [26] 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), *Journal of Chemical Physics*, Vol .102, pp. 3586-3616.
- [27] Halgren T.A., (1996), “Merck Molecular Force Field: I. Basis, Form, Scope, Parameterization and Performance of MMFF94”, *Journal of Computational Chemistry*, Vol. 17 No. 5, pp. 490-519.
- [28] Halgren ,T.A, (1996), “Merck Molecular Force Field: II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions”, *Journal of Computational Chemistry* , Vol. 17 No. 5, pp. 520-552, 553-586, 616- 641.
- [29] Halgren, T.A., Nckbar, R.B., (1996), “ Merck Molecular Force Field IV. Conformational Energies and Geometries for MMFF94,” *Journal of Computational Chemistry*, Vol. 17 No. 5-6, pp. 587-615.
- [30] Allinger, N.L. (1977), “ Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms”,*Journal of the American Chemical Society*,Vol.99 No. 25,pp.8127-8134.
- [31] Ramachandran ,K.I. , Deepa, G., Namboori, K. (2008).“ *Computational chemistry and Molucular Modeling : Principles and Applications*”. DOI 10.1007/978-3-540-77304-7.
- [32] 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 No.10, pp.1361-1375.
- [33] Wold, S. , Eriksson, L. (1995), *Chemometric Methods in Molecular Design*, VCH Publisher, Weinheim.
- [34] Efron, B. (1994), *The Jacknife, the Bootstrap and Other Resampling Planes* , Society for Industrial and Applied Mathematics.Philadelphia.
- [35] Wehrens, R., Putter, H. , Buydens, L.M.C. (2000), “The bootstrap: a tutorial”, *Chemometrics and Intelligent Laboratory Systems*, Vol. 54 No 1, pp. 35-52.
- [36] Todeschini, R., Consonni, V., Maiocchi, A. (1999), “The K correlation index: theory development and its applications in chemometrics”, *Chemometrics and Intelligent Laboratory Systems*, Vol.46, pp. 13-29.

- [37] Golbraikh, A., Tropsha, A. (2002), “Beware of  $q^2!$ ”, *Journal of Molecular Graphics and Modelling*, Vol.20 No.4, pp.269-276.
- [38] Draper N.R., Smith H., (1998), *Applied Regression Analysis*, Third Edition, Wiley series in Probability and Statistics, New York.
- [39] Tropsha, A., Gramatica, P., Gombar, V.K. (2003), “The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models”, *QSAR and Combinatorial Science*, Vol. 22 No 1, pp. 69-76.
- [40] Sharma, B.K., Singh, P., Piloni, P., Sarbhai, K., Yenamandra, S., Prabhakar. (2001), “CP-MLR/PLS directed QSAR study on apical Sodium-codependent bile acid transporter inhibition activity of benzothiepins”, *Molecular Diversity*, Vol. 15 No 1, pp. 135-147.
- [41] Todeschini R., Cossoni V., (2000), *Handbook of Molecular Descriptors*, Mannhold, R., Kubinyi, H. Timmerman eds., Wiley-VCH, Verlag GMBH, Weinheim.
- [42] Kennard, R., Stone, L.A. (1969), “Computer aided design of experiments”, *Technometrics*, Vol. 11 No.1, pp.137-148.
- [43] Tropsha, A., Gramatica, P., Gombar, V.K. (2003), “The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models”, *QSAR and Combinatorial Science*, Vol. 22, pp. 69-76.
- [44] [http://fr.wikipedia.org/wiki/Les\\_algorithmes\\_génétiques](http://fr.wikipedia.org/wiki/Les_algorithmes_génétiques).
- [45] Kubinyi, H., (1994), “Variable Selection in QSAR Studies. I. An Evolutionary Algorithm”, *Quantitative Structure Activity Relationship*, Vol. 13, pp. 285-294.



**RÉSULTATS ET  
DISCUSSIONS**

### III. 1.Méthodologie :

Les données prélevées dans la littérature des trois propriétés d'intérêt ( log tr ; log BCF ; Ir) de trois bases de données (ensemble hétérogène des pesticides et PCBs; les PCBs ; les HAPs) ont été éclatées en deux sous ensembles disjoints : l'un de calibrage pour la sélection des descripteurs par algorithme génétique (AG) et pour la construction du modèle ,et l'autre de test comportant de 15 à 48 % du total des données, uniquement utilisé pour la validation statistique externe.

Le choix aléatoire (commande sample du logiciel MINITAB) et le choix raisonné (algorithme CADEX de Kennard et Stone) ont été, selon les cas, utilisés.

La boîte à outils TOMCAT version 1.01 (nécessitant pour son fonctionnement l'environnement MATLAB) se compose de plusieurs dossiers, dont le dossier Subset – sélection qui contient l'algorithme CADEX.

Les géométries des molécules, représentées à l'aide de l'un des programmes HYPERCHEM 6.03 ou SPARTAN 1.1.0, ont été prés-optimisées par des calculs de mécanique moléculaire, puis les géométries finales ont été optimisées par l'une des méthodes semi-empiriques : PM3 ou PM6.

Les géométries optimisées sont transférées dans le logiciel Dragon pour le calcul, pour chaque molécule, de 1664 descripteurs appartenant à différentes classes. Des descripteurs quantochimiques (tels que HOMO, LUMO, énergie de solvation, masse moléculaire), calculables par les deux logiciels HYPERCHEM et SPARTAN ont été ajoutés et utilisés lors de la sélection des descripteurs pour le développement des modèles.

Les descripteurs muets (écarts-types inférieurs à 0,0001) ou redondants (R 0,95) sont éliminés, ce qui permet une prés - réduction du nombre de descripteurs. (De deux descripteurs hautement corrélés on exclut automatiquement celui qui est corrélé avec le plus grand nombre de descripteurs).

Les algorithmes génétiques (AG) basés sur la recherche stochastique, constituent une méthode de choix pour la sélection de sous –ensembles de variables explicatives.

Dans le logiciel MOBY-DIGS les processus de croisement et de mutation sont contrôlés par un paramètre T variant entre 0 et 1.

Les paramètres de l'AG ont été fixés comme suit : population des modèles : Pop =100 ; valeur de T : choisie égale à 0,5 pour équilibrer les rôles joués par les deux processus de croisement et de mutation.

On commence avec une variable explicative, puis on augmente ce nombre (à 2, à 3,...) jusqu'à ce que l'ajout d'un descripteur n'améliore plus les statistiques du modèle : ce qui indique le nombre optimal de descripteurs.



L'analyse de régression multilinéaire (RML) par les moindres carrés ordinaires, a été réalisée par le logiciel MOBY-DIGS.

### III.2. Modélisation des temps de rétention pour un grand ensemble hétérogène constitué de (Pesticides et PCBs).

L'approche hybride algorithme génétique /régression multilinéaire (AG/RLM) a été utilisée pour modéliser le temps de rétention d'un ensemble hétérogène de 84 pesticides et PCBs (tableau III. 1) éclaté par l'algorithme CADEX, en deux ensembles de calibrage (de 67 éléments) et de validation externes (de 17 éléments).

Tableau III.1. Les valeurs de En, nR06, ATS1m, ATS7v, GATS2e, EEig05d et log tr pour un ensemble de 84 pesticides ou toxiques. Les 17 derniers composés constituent l'ensemble de test.

Composé	log tr	En	nR06	ATS1m	ATS7v	GATS2e	EEig05d
a-666	1,275	-258,84	1	3,207	0	1,222	1,904
Chlorbufan	1,313	-99,54	1	2,959	2,197	0,733	1,278
Atrazine	1,323	78,72	1	2,924	1,386	0,305	1,226
Trietaezine	1,326	238,59	1	2,997	1,609	0,936	1,229
Fonofos	1,337	-434,06	1	3,49	1,609	1,212	1
PCB15	1,344	111,06	2	2,991	1,792	0,758	1
Carbofuran	1,36	-463,49	1	2,995	1,858	0,911	1,687
4,4'-DDM	1,39	88,01	2	3,04	2,197	0,769	1
PCB31	1,399	75,77	2	3,129	1,792	0,795	1,234
Benoxacor	1,401	-276,99	2	3,153	1,099	0,847	2,097
Phosphamidon	1,43	-1235,2	0	3,451	2,18	0,516	1,855
Benfuresate	1,432	-661,33	1	3,379	2,313	0,772	2,013
Aldrin	1,434	110,23	2	3,518	0	0,763	2,287
PCB52	1,443	41,37	2	3,251	1,609	0,833	1,904
Metolachlor	1,464	-415,16	1	3,155	2,2	0,637	1,987
Trichoronate	1,472	-678	1	3,607	2,361	1,173	1,809
Methoprene	1,476	-733,24	0	3,164	2,743	1,051	2,345

Composé	log tr	En	nR06	ATS1m	ATS7v	GATS2e	EEig05d
Chlorpyriphos	1,476	-848,13	1	3,673	2,574	0,91	1,812
Thiobencarb	1,478	-209,2	1	3,141	2,417	0,668	2,153
Methiocarb	1,483	-333,36	1	3,029	1,843	0,512	2,147
Isodrin	1,485	314,67	2	3,518	0	0,763	2,299
Fenthion	1,501	-770,14	1	3,542	2,138	0,548	1,88
Allethrin	1,503	-463,44	0	3,232	2,972	0,76	2,604
Isocarbophos	1,52	-958,97	1	3,509	2,27	0,826	1,586
Isofenphos	1,523	-1069,5	1	3,627	3,03	0,957	2,181
Triadimenol	1,533	-133,17	1	3,264	2,727	0,97	1,939
Procymedone	1,544	-265,24	2	3,261	2,303	0,64	2,623
Quinalphos	1,549	-639,59	2	3,579	2,624	0,8	1,618
Alpha-endosulfan	1,549	-538,27	1	3,739	1,114	0,776	2,719
Phenthoate	1,551	-953,31	1	3,688	2,655	0,628	1,811
Chlorbenside	1,557	112,67	2	3,229	2,398	0,825	2,075
Crotoxyphos	1,561	-1255,5	1	3,519	2,601	0,492	1,939
Prothiofos	1,572	-718,96	1	3,735	2,687	1,091	1,977
Tetrachlorvinphos	1,576	-933,67	1	3,652	2,506	0,549	2,149
Chinomethionate	1,577	120,53	2	3,246	1,414	0,671	2,286
PCB 87	1,581	31,71	2	3,359	2,079	0,872	2,155
4;4'-DDE	1,582	112,9	2	3,325	2,565	0,63	2,067
Methidathion	1,587	-837,39	0	3,718	2,136	0,613	2,134
Buprofezin	1,59	-80,05	2	3,332	2,594	0,525	2,522
Fenamiphos	1,59	-844,63	1	3,511	2,643	0,58	2,17
2,4'-DDD	1,6	63,83	2	3,325	2,303	0,63	1,848

Composé	log tr	En	nR06	ATS1m	ATS7v	GATS2e	EEig05d
PCB149	1,608	3	2	3,457	2,197	0,911	2,268
Endrin	1,612	47,94	3	3,594	0	0,785	2,728
2,4'-DDT	1,627	17,68	2	3,426	2,398	0,59	1,854
4,4'-DDT	1,652	10,03	2	3,426	2,708	0,59	2,177
Benalaxyl	1,652	-430,6	2	3,326	3,335	0,552	2,255
PCB187	1,653	-17,33	2	3,546	2,398	0,95	2,484
Resmethrin	1,658	-348,91	1	3,39	3,023	0,758	2,603
PCB167	1,661	7,27	2	3,457	2,485	0,911	2,373
Bifenthrin	1,663	-859,34	2	3,603	3,231	0,475	2,612
Famphur	1,664	-1113,8	1	3,763	2,39	0,572	1,734
Edifenphos	1,664	-453,02	2	3,721	2,833	0,68	1,774
PCB202	1,665	-33,33	2	3,627	2,303	0,99	2,633
Bromopropylate	1,67	-377,13	2	3,601	3,151	0,919	2,413
Fenpropathrin	1,673	-181,69	2	3,428	3,36	0,953	2,947
Phenothrin	1,678	-322,16	2	3,423	3,191	0,891	2,603
Dicofol	1,678	-158,88	2	3,469	2,708	0,75	2,177
Tetramethrin	1,679	-669,63	1	3,373	3,033	0,481	2,665
Pyridaphenthion	1,68	-752,41	2	3,673	2,866	0,793	1,781
Leptophos	1,687	-435,46	2	3,797	2,923	0,96	2,069
Imidan	1,687	-841,72	1	3,717	2,613	0,536	2,371
Tertradifon	1,688	-230,03	2	3,618	2,485	0,916	2,515
Pyrazophos ethyl	1,695	-980,22	1	3,747	2,922	0,83	2,113
Fenarimol	1,697	137,13	3	3,398	2,878	0,856	1,92
Permethrin	1,7	-305,86	2	3,543	3,191	0,804	2,722

Composé	log tr	En	nR06	ATS1m	ATS7v	GATS2e	EEig05d
Azinphos-methyl	1,7	-517,14	2	3,726	2,572	0,623	1,921
PCB194	1,701	-16,85	2	3,627	2,773	0,99	2,641
Lindane	1,34	-258,84	1	3,207	0	1,222	1,904
Fenchlorphos	1,429	-810,07	1	3,612	1,962	0,79	1,78
Pentanochlor	1,451	-273,01	1	2,976	2,352	0,751	1,878
Bromophos-methyl	1,504	-757,78	1	3,707	2,07	0,758	1,769
PCB 70	1,522	51,64	2	3,251	2,079	0,833	2,124
PCB101	1,545	19,14	2	3,359	2,079	0,872	2,214
2,4'-DDE	1,545	119,53	2	3,325	2,303	0,63	1,756
PCB118	1,614	29,77	2	3,359	2,303	0,872	2,15
PCB153	1,627	-2,81	2	3,457	2,398	0,911	2,262
Beta-endosulfan	1,637	-491,71	1	3,739	1,114	0,776	2,719
4,4'-DDD	1,637	29,37	2	3,325	2,565	0,63	2,177
PCB141	1,639	10,66	2	3,457	2,303	0,911	2,478
Sulprophos	1,647	-659,2	1	3,721	2,647	1,113	2,389
PCB138	1,65	9,71	2	3,457	2,398	0,911	2,256
PCB185	1,661	-3,25	2	3,546	2,303	0,95	2,713
PCB128	1,666	22,13	2	3,457	2,398	0,911	2,243
PCB180	1,674	-11,09	2	3,546	2,565	0,95	2,487

Parmi les descripteurs pouvant être en relation avec log tr, les six descripteurs du tableau III.2 sont les mieux adaptés pour la modélisation par RLM.

Tableau III. 2 : Les descripteurs, leurs classes et significations

Descripteur	Definition	Classe
En	Énergie totale de la molécule	Descripteur Quanto-chimique
nR06	Nombre de cycles à 6 chaînons	Descripteurs constitutionnels
ATS1m	Broto-Moreau autocorrélation d'une structure topologique - lag 1 / pondérée par les masses atomiques.	Indices d'autocorrélation 2D
ATS7v	Broto-Moreau autocorrélation d'une structure topologique - lag 7 / pondérée par les volumes atomiques de van der Waals.	
GATS2e	Autocorrélation Geary - lag 2 / pondérée par électronégativités atomiques de Sanderson.	
EEig05d	Valeur propre 05 de la matrice d'adjacence des cotés pondérée par les moments dipolaires.	Indices d'adjacence des côtés.

Le modèle basé sur ces descripteurs a pour équation :

$$\begin{aligned} \log tr = & 0.32095 \pm (0.08595) + 0.00006735 \pm (0.00002276) \text{ En} \\ & + 0.03612 \pm (0.01117) \text{ nR06} + 0.27856 \pm (0.03128) \text{ ATS1m} \\ & + 0.070972 \pm (0.006836) \text{ ATS7v} - 0.11332 \pm (0.02614) \text{ GATS2e} \\ & + 0.08315 \pm (0.01173) \text{ EEig05d} \end{aligned} \quad (1)$$

$R^2=90,54$  ;  $Q^2_{\text{LOO}}=88,15$  ;  $Q^2_{\text{BOOT}}=86,58$  ;  $Q^2_{\text{ext}}=87,15$  ;  $R^2_{\text{adj}}=89,59$  ;  $K_{\text{xx}}= 29.62$  ;  $K_{\text{xy}}= 40,01$

$F=95,6528$  ;  $S=0,0381$  unité log.

Les valeurs de  $R^2$  et  $R^2_{\text{adj}}$  montrent la qualité de l'ajustement, alors que la petite différence entre  $R^2$  et  $Q^2_{\text{LOO}}$  renseigne sur la robustesse du modèle qui est, en outre, hautement significatif (grande valeur de la statistique F de Fisher). La validation par bootstrap confirme tout à la fois la bonne capacité de prédiction interne et la stabilité du modèle.

Les paramètres statistiques reproduits dans le tableau III.3 permettent d'évaluer les descripteurs de ce modèle.

Tableau III.3 : Caractéristiques des descripteurs sélectionnés pour le modèle AG / RLM optimal.

Prédicteur	Coef	SE Coef	t	P	FIV
Constante	0,32095	0,08595	3,73	0,000	-
En	0,00006735	0,00002276	2,96	0,004	3,913
nR06	0,03612	0,01117	3,23	0,002	2,472
ATS1m	0,27856	0,03128	8,90	0,000	2,390
ATS7v	0,070972	0,006836	10,38	0,000	1,268
GATS2e	-0,11332	0,02614	-4,34	0,000	1,126
EEig05d	0,08315	0,01173	7,09	0,000	1,264

La valeur de t pour un descripteur est liée à sa signification statistique. Les valeurs absolues élevées de t indiquent que chaque coefficient de régression est plus grand que l'écart type qui lui est associé. La probabilité de t(p) pour un descripteur donne sa signification statistique lorsqu'il est impliqué dans un modèle QSPR 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 le fait du hasard. Les valeurs des probabilités de t pour les six descripteurs sont toutes très inférieures à 0,05, ce qui indique qu'ils sont très significatifs. Les valeurs des facteurs d'inflation de la variance (FIV, toutes inférieures à 5) et la matrice de corrélation reproduite dans le tableau III.4 (page suivante) suggèrent que ces descripteurs sont faiblement corrélés entre eux.

Tableau III.4. Matrice de corrélation entre les descripteurs sélectionnés et log tr.

	En	nR06	ATS1m	ATS7v	GATS2e	EEig05d	log tr
En	1.000						
nR06	0.614	1.000					
ATS1m	-0.501	0.063	1.000				
ATS7v	-0.369	-0.068	0.249	1.000			
GATS2e	0.169	0.067	0.086	-0.097	1.000		
EEig05d	-0.021	0.159	0.363	0.202	-0.018	1.000	
log tr	-0.126	0.385	0.658	0.583	-0.134	0.639	1.000

La figure III.1 reproduit les valeurs de log tr prédites en fonction de log tr expérimental avec l'équation (1) pour les composés de calibration et de validation.

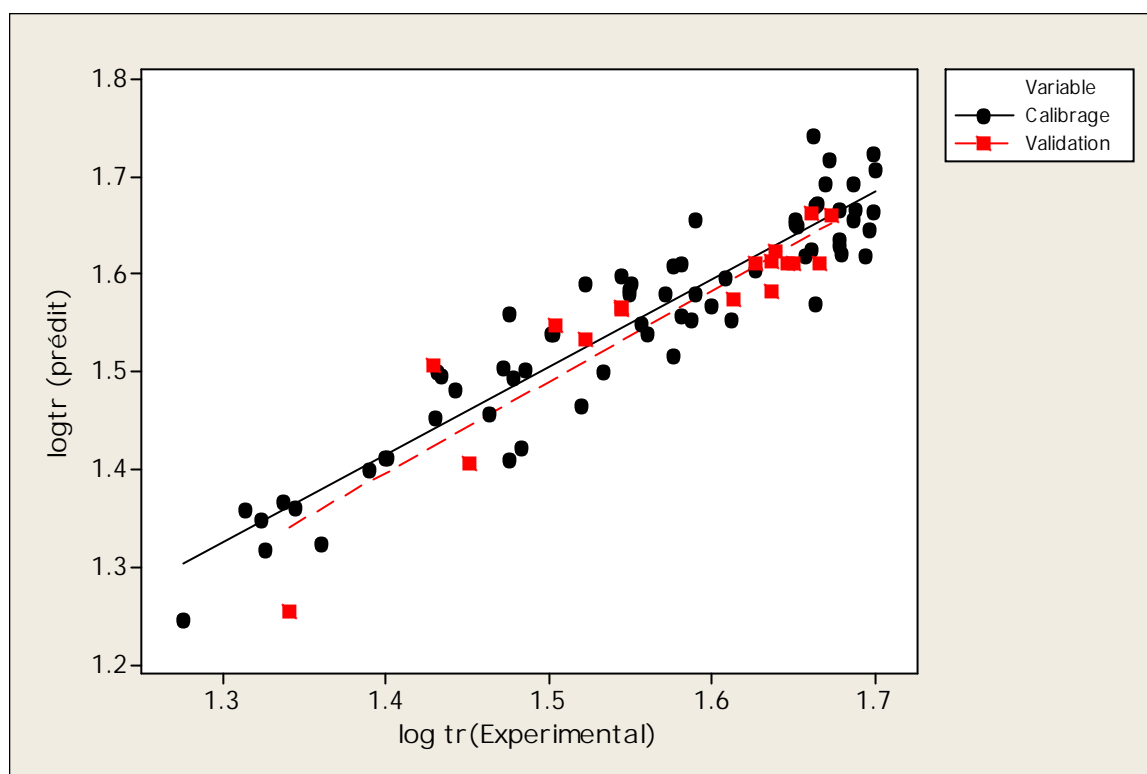


Figure III.1.: Les valeurs prédites de log tr en fonction des valeurs expérimentales de log tr.



Le test de randomisation (Fig III.2) permet de s'assurer que le modèle établi a une base réelle, et qu'il n'est pas dû au hasard.

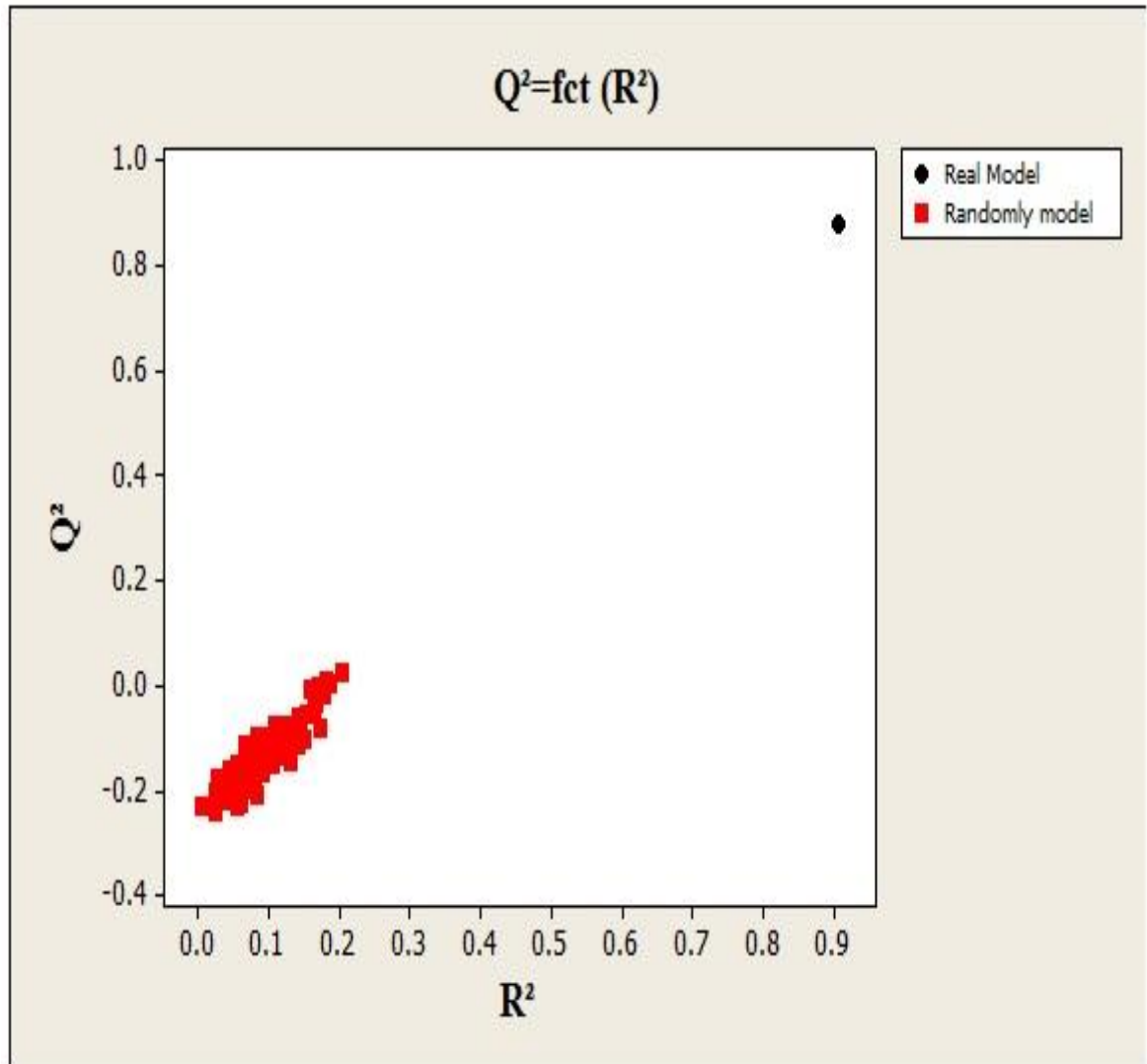


Figure III.2.: Test de randomisation associé au modèle QSRR précédent. Les carrés représentent les rétentions aléatoires et le cercle correspond à la rétention réelle.

Le domaine d'application du modèle a été discuté à l'aide de diagramme de Williams (Figure III.3), on constate que tous les résidus sont situés dans l'intervalle de trois écarts-types, et que tous les composés ont un  $h_i < h^*$  ce qui met en évidence l'absence de point aberrant et /ou influent.

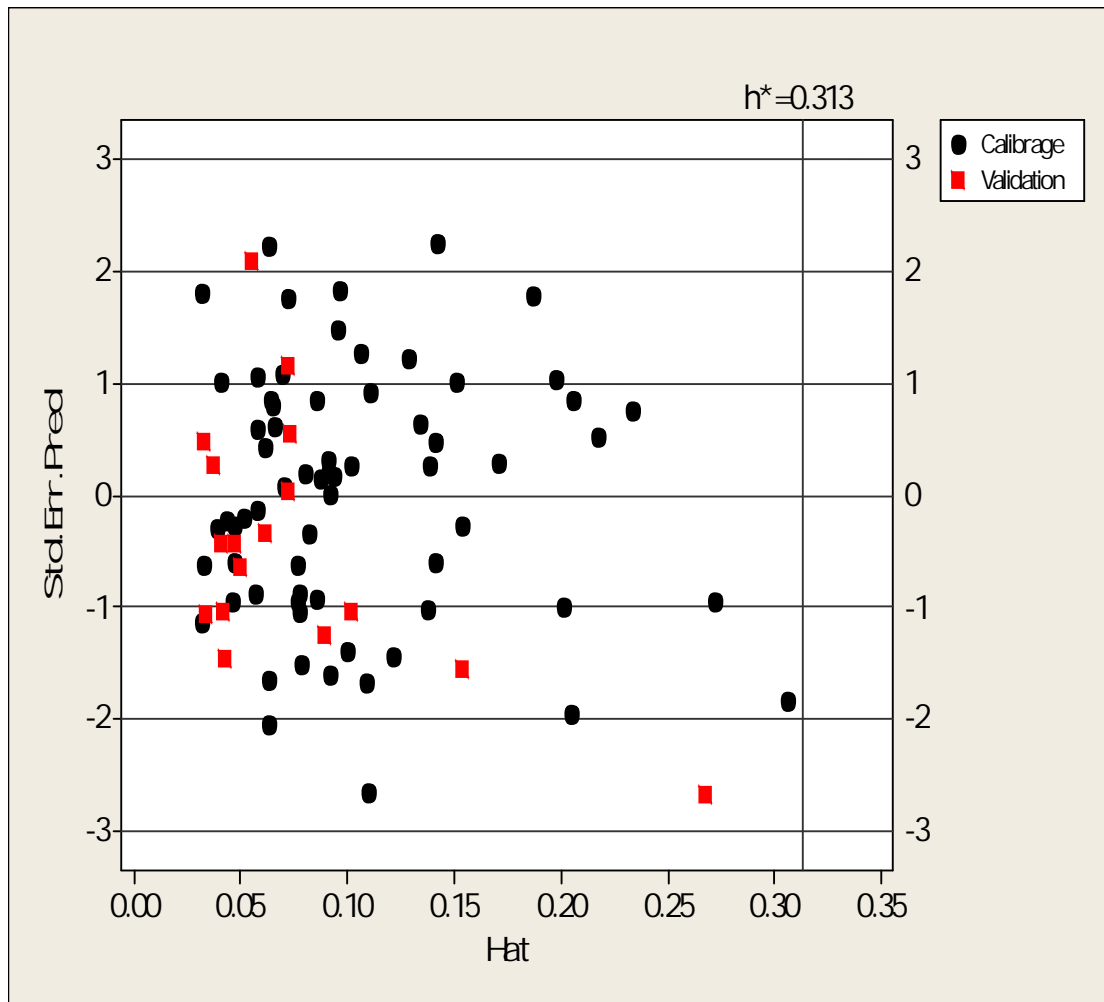


Figure III.3: Diagramme de Williams du modèle QSRR sélectionné.

Les contributions relatives des six descripteurs au modèle (figure III.4) ont été calculées à l'aide d'une procédure décrite dans la littérature. Ces contributions décroissent dans l'ordre :

ATS7v (20,7701 %) > ATS1m (19,6349%) > EEig05d (17,4652%) > GATS2e (14,4739%) > nR06 (14,0646%) > En (13,5912%).

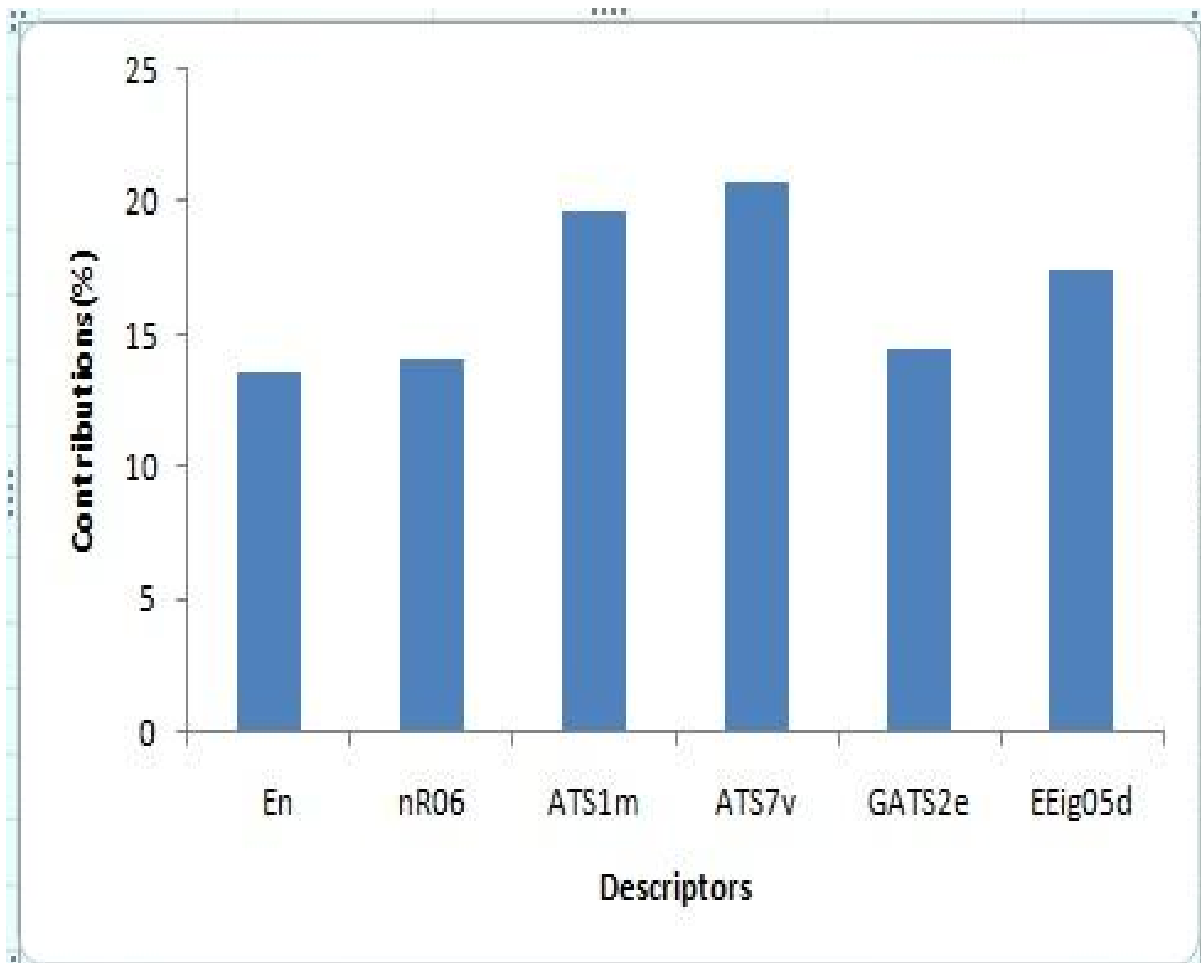


Figure III.4: Contributions relatives des descripteurs sélectionnés dans le modèle AG / RLM.

Notons que la différence dans les contributions au modèle de deux descripteurs quelconques n'est pas significative, ce qui prouve que les six descripteurs sont indispensables pour la génération du modèle prédictif.

**III. 3. Modélisation du facteur de bioaccumulation d'un ensemble de PCBs :**

Les 58 PCBs (tableau III.5) ont été séparés par algorithme CADEX en deux ensembles de calibrage (30 composés) et de validation externe (28 composés), le logiciel HYPERCHEM a été utilisé pour la représentation des géométries des molécules, optimisées par la méthode semi-empirique PM3.

Tableau III.5 : Valeurs de log BCF (exp), pour un ensemble de 58 PCB. Les 28 derniers PCBs constituent l'ensemble de test.

Nomenclature UAPAC	Numéro de Congénère	logBCF <sub>exp</sub>	HATS0v
Biphenyl	PCB 0	2,64	0,063
4-Chlorobiphenyl	PCB 3	2,77	0,085
4,4'-Dichlorobiphenyl	PCB15	3,28	0,106
2,2'-Dichlorobiphenyl	PCB 4	3,38	0,099
2,4'-Dichlorobiphenyl	PCB 8	3,57	0,106
2,2',6,6'-Tetrachlorobiphenyl	PCB 54	3,85	0,129
2,2',5-Trichlorobiphenyl	PCB 18	4,11	0,132
2,4,4'-Trichlorobiphenyl	PCB28	4,2	0,126
2,2',3,3'-Tetrachlorobiphenyl	PCB 40	4,23	0,156
2,4,5-Trichlorobiphenyl	PCB29	4,26	0,142
3,3',4,4'-Tetrachlorobiphenyl	PCB 77	4,59	0,174
2,3,4',6-Tetrachlorobiphenyl	PCB 64	4,6	0,151
2,2',5,5'-Tetrachlorobiphenyl	PCB 52	4,63	0,16
2,2',3,4,5'-Pentachlorobiphenyl	PCB 87	5,38	0,181
2,3,3',4,4',5'-Hexachlorobiphenyl	PCB 157	5,39	0,216
2,2',4,5,5'-Pentachlorobiphenyl	PCB 101	5,4	0,18

Nomenclature UAPAC	Numéro de Congénère	logBCF <sub>exp</sub>	HATS0v
2,2',3,3',6,6'-Hexachlorobiphenyl	PCB 136	5,43	0,176
Decachlorobiphenyl	PCB 209	5,44	0,251
2,2',3,3',4,5,5',6,6'-Nonachlorobiphenyl	PCB 208	5,71	0,237
2,2',3,3',4,5,6'-Heptachlorobiphenyl	PCB 174	5,8	0,215
2,2',3,4,4',5,5'-Heptachlorobiphenyl	PCB 180	5,8	0,225
3,3',4,4',5-Pentachlorobiphenyl	PCB 126	5,81	0,209
2,2',3,4,5,5'-Hexachlorobiphenyl	PCB 141	5,81	0,208
2,2',3,3',4,4',5,5'-Octachlorobiphenyl	PCB 194	5,81	0,248
2,2',3,3',5,5',6,6'-Octachlorobiphenyl	PCB 202	5,82	0,222
2,2',3,3',4,5,5',6-Octachlorobiphenyl	PCB 198	5,88	0,237
2,2',3,3',4,4',5,6-Octachlorobiphenyl	PCB 195	5,92	0,231
2,2',3,3',4,4',5,6'-Octachlorobiphenyl	PCB 196	5,92	0,232
3,3',4,4',5,5'-Hexachlorobiphenyl	PCB 169	5,97	0,238
3,5-Dichlorobiphenyl	PCB14	3,78	0,143
2,4-Dichlorobiphenyl	PCB 7	3,55	0,106
2,3'-Dichlorobiphenyl	PCB 6	3,8	0,12
2,5-Dichlorobiphenyl	PCB 9	3,89	0,121
2,3-Dichlorobiphenyl	PCB 5	4,11	0,117
2,4',5-Trichlorobiphenyl	PCB31	4,23	0,143
2,3',4',5-Tetrachlorobiphenyl	PCB 70	4,77	0,175
2,2',3,5'-Tetrachlorobiphenyl	PCB 44	4,84	0,162

Nomenclature UAPAC	Numéro de Congénère	logBCF <sub>exp</sub>	HATS0v
2,2',4,5'-Tetrachlorobiphenyl	PCB 49	4,84	0,154
2,2',4,4'-Tetrachlorobiphenyl	PCB 47	4,85	0,142
2,2',4,4',6,6'-Hexachlorobiphenyl	PCB 155	4,93	0,168
2,2',4,5-Tetrachlorobiphenyl	PCB 48	5	0,153
2,2',3,4',5-Pentachlorobiphenyl	PCB 90	5	0,184
2,2',4,4',5-Pentachlorobiphenyl	PCB 99	5	0,172
2,3,3',4,4'-Pentachlorobiphenyl	PCB 105	5	0,186
2,3,3',4,6-Pentachlorobiphenyl	PCB 109	5	0,177
2,3',4,4',5-Pentachlorobiphenyl	PCB 118	5	0,193
2,2',3,4,4',5'-Hexachlorobiphenyl	PCB 138	5,39	0,199
2,2',3,4',5,6'-Hexachlorobiphenyl	PCB 148	5,39	0,194
2,3,3',4,4',5-Hexachlorobiphenyl	PCB 156	5,39	0,219
2,2',3,4',5'-Pentachlorobiphenyl	PCB 97	5,43	0,181
2,2',3,5,5',6-Hexachlorobiphenyl	PCB 151	5,54	0,199
2,2',4,4',5,5'-Hexachlorobiphenyl	PCB 153	5,65	0,198
2,2',3,3',4,4'-Hexachlorobiphenyl	PCB 128	5,77	0,194
2,2',3,4,4',5,6'-Heptachlorobiphenyl	PCB 182	5,8	0,211
2,2',3,4',5,5',6-Heptachlorobiphenyl	PCB 187	5,8	0,217
2,2',3,4,4',5,6-Heptachlorobiphenyl	PCB 183	5,84	0,211
2,3,3',4,4',5,6-Heptachlorobiphenyl	PCB 191	5,84	0,223
2,2',3,4,4',5-Hexachlorobiphenyl	PCB 137	5,88	0,201

Le seul descripteur utilisé pour la modélisation du facteur de bioaccumulation est présenté dans le tableau III.6

Tableau III.6 : Le seul descripteur sélectionné pour la modélisation de log BCF :

Descripteur	Signification	Classe
HATS0v	Autocorrélation de levier en fonction du décalage 0 / pondéré par les volumes atomiques de van der Waals .	GETAWAY

L'équation de régression calculée est :

$$\log \text{BCF} = 1,5779 \pm (0,1958) + 18,538 \pm (1,065)\text{HATS0v.} \quad (2)$$

$R^2=91,53$ ;  $Q^2_{\text{LOO}}=90,28$ ;  $Q^2_{\text{BOOT}}=89,23$  ;  $Q^2_{\text{ext}}= 90,44$  ;  $R^2_{\text{adj}}= 91,23$  ;  $K_{xx}= 0$  ;  $K_{xy}= 95,67$

$F=302,7423$  ;  $S=0,3106$  unité log.

La figure III.5 reproduit Les valeurs expérimentales de logarithme décimal BCF en fonction des valeurs prédites pour l'ensemble des données.

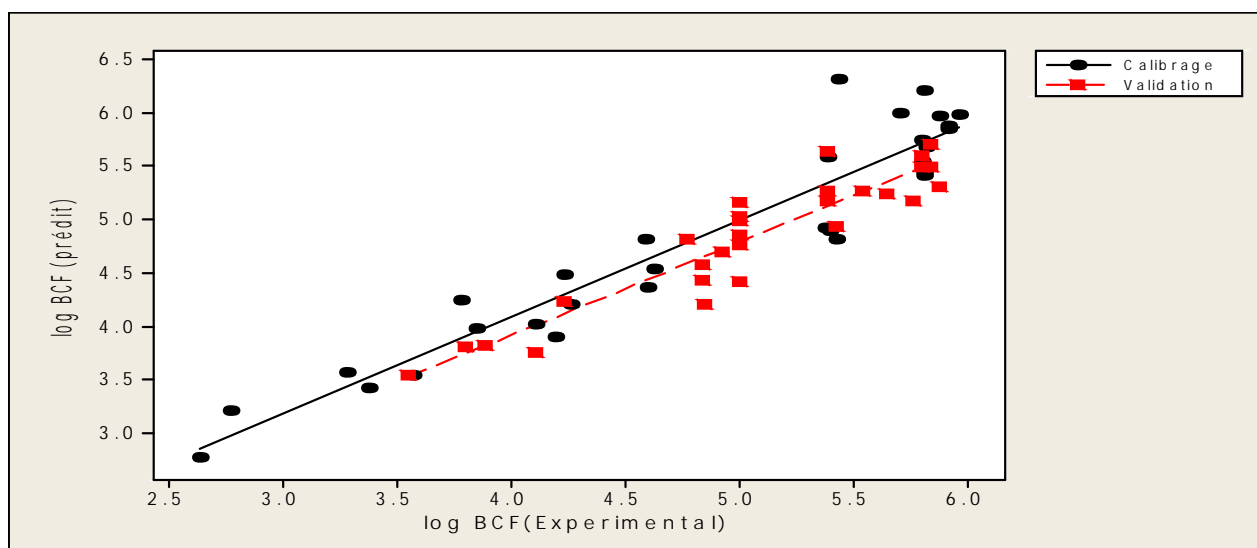


Figure. III.5 : Les valeurs prédites de logarithme décimal BCF en fonction des valeurs expérimentales pour l'ensemble des données.

Les paramètres statistiques suivants obtenus pour l'ensemble de validation vérifient les conditions d'acceptabilité de Golbraikh et al. (inéquations 48 – 50 du chapitre développement et évaluation du modèles) démontrant ainsi la capacité prédictive du présent modèle :

$$Q^2_{\text{ext}} = 0,7793 > 0,5 \quad ; \quad r^2 = 0,8707 > 0,6$$

$$(r^2 - r'^2_0) / r^2 = 0,0014 < 0,1 \quad \text{ou} \quad (r^2 - r'^2_0) / r^2 = -0,0046 < 0,1$$

$$0,85 < k = 1,0447 < 1,15 \quad \text{ou} \quad 0,85 < k' = 0,9552 < 1,15$$

Les résultats des modèles randomisés sont comparés à ceux du modèle réel initial à l'aide d'un graphique représentant les paramètres statistiques  $R^2$  et  $Q^2$ . La figure III.6 montre une nette séparation entre les statistiques des réponses randomisées et celle du modèle de départ. Ce qui permet d'affirmer qu'une relation structure-propriété réelle a été établie.

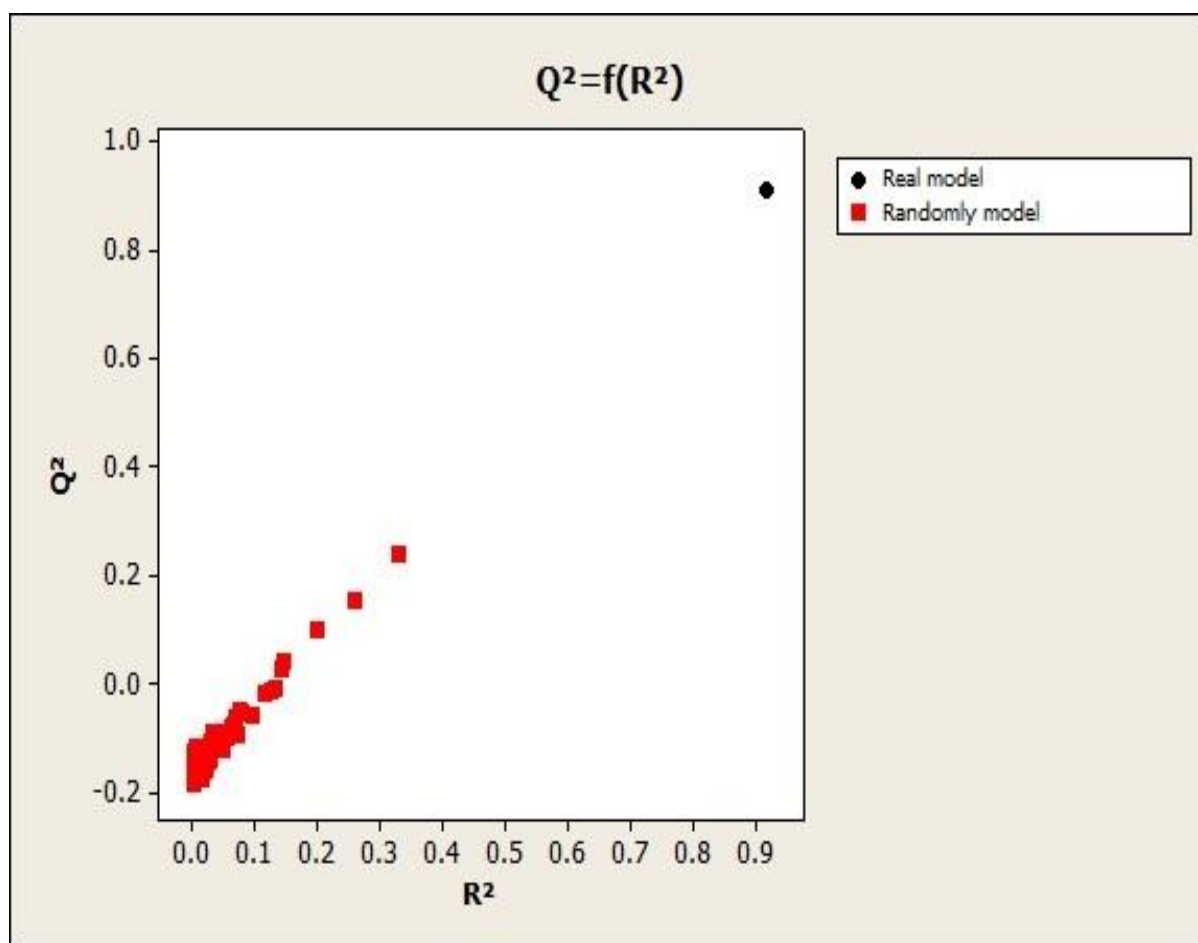


Figure III. 6 : Test de randomisation associé au modèle QSPR précédent. Les carrés représentent les propriétés ordonnées au hasard et le cercle correspond aux propriétés réelles.



Le diagramme de Williams (figure III.7) a été utilisé pour définir le domaine d'application du modèle RLS. Nous notons l'absence de point aberrant et /ou influent.

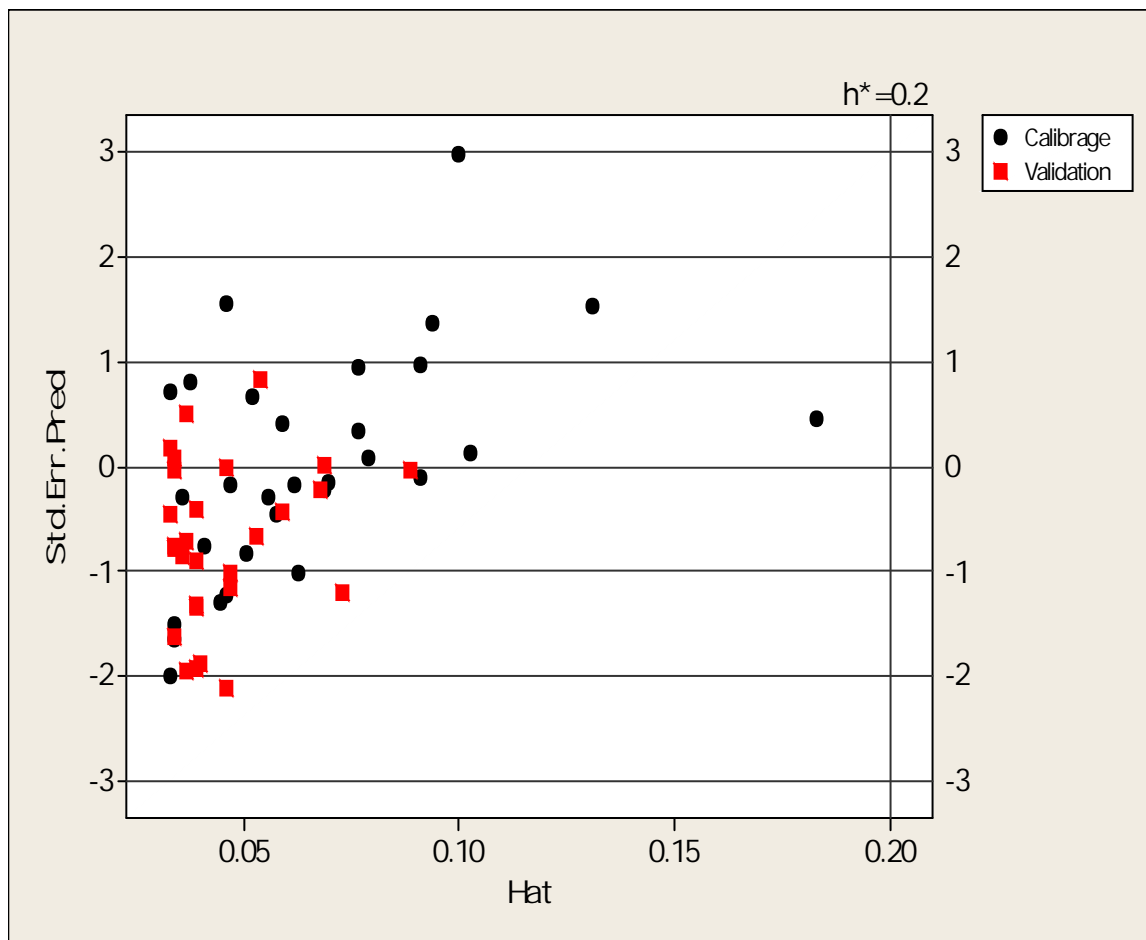


Figure III. 7 : Diagramme de Williams du modèle QSPR développé.

### III.4. MODÉLISATION DES INDICES DE RÉTENTION DES HAPS :

93 HAPs séparés aléatoirement (tableau III.7) en  $n_{\text{calib}} = 70$  et  $n_{\text{test}} = 23$  ; on utilise le logiciel SPARTAN pour la représentation des géométries des molécules puis on les optimise par la méthode semi-empirique PM6.

Tableau III.7: Valeurs de  $I_r$ , la masse moléculaire, l'énergie de solvation pour un ensemble de 93 HAP. Les 23 derniers produits chimiques constituent l'ensemble de test.

Composé	$I_r$	Energie de solvation (solv en)	Masse moléculaire (MW)
Naphthalene	200	-11,24	128,174
1-Methylnaphthalene	221,04	-11,61	142,201
2-Ethylnaphthalene	236,08	-9,99	156,228
1-Ethylnaphthalene	236,56	-10,86	156,228
2,7-Dimethylnaphthalene	237,71	-11,75	156,228
1,3-Dimethylnaphthalene	240,25	-12,03	156,228
1,7-Dimethylnaphthalene	240,66	-11,6	156,228
1,6-Dimethylnaphthalene	240,72	-11,97	156,228
1,4-Dimethylnaphthalene	243,57	-11,62	156,228
Acenaphthelene	244,63	-15,77	152,196
1,5-Dimethylnaphthalene	244,98	-11,75	156,228
1,2-Dimethylnaphthalene	246,49	-11,95	156,228
2,3,6-Trimethylnaphthalene	263,31	-11,84	170,255
1-Methylacenaphthelene	265,24	-16,66	166,223
2,3,5-Trimethylnaphthalene	265,9	-12,12	170,255
Phenanthrene	300	-16,21	178,234

Composé	Ir	Energie de solvation (En de solv)	Masse moléculaire (MW)
1-Phénylnaphthalene	315,19	-11,69	204,272
3-Méthylphenanthrene	319,46	-16,56	192,261
2-Méthylantracene	321,57	-15,64	192,261
2-Méthylphenanthrene	321,57	-16,3	192,261
4-Méthylphenanthrene	323,17	-16,97	192,261
1-Méthylantracene	323,33	-15,67	192,261
1-Méthylphenanthrene	323,9	-16,73	192,261
9-Méthylantracene	329,13	-17,14	192,261
9-Ethylphenanthrene	337,05	-15,98	206,288
2-Ethylphenanthrene	337,5	-14,89	206,288
3,6-Diméthylphenanthrene	337,83	-16,89	206,288
2,7-Diméthylphenanthrene	339,23	-16,32	206,288
9-Isopropylphenanthrene	345,78	-13,62	220,315
1,8-Diméthylphenanthrene	346,26	-17,19	206,288
9-n-Propylphenanthrene	350,3	-14,47	220,315
Pyrene	351,22	-20,51	202,256
9-Méthyl-10-Ethylphenanthrene	359,91	-15,5	220,315
1-Méthyl-7-isoprppylphenanthrene	368,67	-14	234,342
4-Méthylpyrene	369,54	-21	216,283
1-Méthylpyrene	373,55	-21,3	216,283

Composé	Ir	Energie de solvation (En de solv)	Masse moléculaire (MW)
9,10-Dimethyl-3-ethylphenanthrene	381,85	-16,26	234,342
1-Ethylpyrene	385,35	-20,48	230,31
2,7-Dimethylpyrene	386,34	-20,64	230,31
Benzo (c) phenanthrene	391,39	-19,54	228,294
9-Phenylanthracene	396,38	-14,51	254,332
Cyclopenta (cd) pyrene	396,54	-24,09	226,278
Benzo (a) anthracene	398,5	-19,93	228,294
Triphenylene	400	-21,24	228,294
9-Phenylphenanthrene	406,9	-15,76	254,332
11-Methylbenzo (a) anthracene	412,72	-20,37	242,321
1-Methylbenzo (a) anthracene	414,37	-20,87	242,321
1-n-Butylpyrene	414,87	-18,24	258,364
1-Methyltriphenylene	416,32	-20,86	242,321
9-Methylbenzo (a) anthracene	416,5	-20,17	242,321
9-Methyl-10-phenylphenanthrene	417,16	-15,77	268,354
8-Methylbenzo (a) anthracene	417,56	-20,35	242,321
6-Methylbenzo (a) anthracene	417,57	-20,39	242,321
3-Methylchrysene	418,1	-21,11	242,321
2-Methylchrysene	418,8	-20,83	242,321
12-Methylbenzo (a) anthracene	419,39	-20,52	242,321

Composé	Ir	Energie de solvation (En de solv)	Masse moléculaire (MW)
4-Methylbenzo (a) anthracene	419,67	-20,51	242,321
5-Methylchrysene	419,68	-20,35	242,321
4-Methylchrysene	420,83	-20,42	242,321
1-Phenylphenanthrene	421,66	-16,02	254,332
1-Methylchrysene	422,87	-21,24	242,321
7-Methylbenzo (a) anthracene	423,14	-21,88	242,321
1,12-Dimethylbenzo (a) anthracene	436,82	-17,49	256,348
Benzo (j) fluoranthene	440,92	-25,17	252,316
Benzo (b) fluoranthene	441,74	-23,86	252,316
Benzo (k) fluoranthene	442,56	-23,42	252,316
1,6,11-Trimethyltriphenylene	446,24	-21,44	270,375
Benzo (e) pyrene	450,73	-25,11	252,316
Benzo (a) pyrene	453,44	-24,4	252,316
Perylene	456,22	-24,93	252,316
Pentacene	486,81	-22,65	278,354
Dibenzo (a,c) anthracene	495,01	-24,59	278,354
Dibenzo (a,h) anthracene	495,45	-24,35	278,354
Picene	500	-25,02	278,354
Dibenzo (def,mno) chrysene	503,89	-27,94	276,338
2-Methylnaphthalene	218,14	-11,4	142,201

Composé	Ir	Energie de solvation ( En de solv)	Masse moléculaire (MW)
2,6-Dimethylnaphthalene	237,58	-11,34	156,228
2,3-Dimethylnaphthalene	243,55	-11,73	156,228
1,8-Dimethylnaphthalene	249,52	-12,71	156,228
Anthracene	301,69	-15,34	178,234
9-Methylphenanthrene	323,06	-16,71	192,261
2-Phenylnaphthalene	332,59	-13,02	204,272
Fluoranthene	344,01	-20,27	202,256
9,10-Dimethylanthracene	355,49	-18,19	206,288
2-Methylpyrene	370,15	-20,63	216,283
Chrysene	400	-20,72	228,294
2-Methylbenzo (a) anthracene	413,78	-20,26	242,321
3-Methylbenzo (a) anthracene	416,63	-20,04	242,321
5-Methylbenzo (a) anthracene	418,72	-20,54	242,321
6-Methylchrysene	420,61	-21,61	242,321
1,3-Dimethyltriphenylene	432,32	-21,37	256,348
7,12-Dimethylbenzo (a) anthracene	443,38	-21,4	256,348
Benzo (b) chrysene	497,66	-24,29	278,354

Dimension du modèle RLM=2

Tableau III.8 : Liste des descripteurs sélectionnés pour la modélisation de Ir.

Descripteur	Signification	Classe
Energie de solvation	La somme de tout les poids atomiques des éléments qui constituent le composé.	Descripteurs quato-chimiques.
Masse moléculaire	L'énergie libérée lors la dissolution d'un composé.	

L'équation RLM calculée:

$$Ir = - 50,00 \pm (4,400) - 4,63 \pm (0,281) \text{ En de solv} + 1,52 \pm (0,031) MW \quad (3)$$

$$R^2 = 99,27 ; Q^2_{\text{LOO}} = 99,18 ; Q^2_{\text{BOOT}} = 99,12 ; Q^2_{\text{ext}} = 99,17 ; R^2_{\text{adj}} = 99,25 ; K_{xx} = 75,68 ;$$

$$K_{xy} = 86,61 ; F = 4569,847 ; S = 6,2428.$$

Les paramètres statistiques reproduits dans le tableau III.9 permettent d'évaluer les deux descripteurs de ce modèle.

Tableau III.9 : Caractéristiques des descripteurs sélectionnés pour le meilleur modèle AG / RLM.

Prédicteur	Coef	SE Coef	t	P
Constante	-49,968	4,400	-11,36	0,000
En de solvation	-4,6285	0,2803	-16,51	0,000
Masse moléculaire (MW)	1,52264	0,03104	49,05	0,000

Vérification des conditions de Golbraikh et al.

$$Q^2_{\text{ext}} = 0,9954 > 0,5 \quad ; \quad r^2 = 0,9969 > 0,6$$

$$(r^2 - r^2_0) / r^2 = -0,0012 < 0,1 \quad \text{ou} \quad (r^2 - r'^2_0) / r^2 = -0,0012 < 0,1$$

$$0,85 < k = 1,0101 < 1,15 \quad \text{ou} \quad 0,85 < k' = 0,9899 < 1,15$$

La matrice de corrélation reproduite dans le tableau III.10 suggère que ces deux descripteurs sont faiblement corrélés entre eux et fortement corrélés avec la variable à expliquer (Ir).

Tableau III.10 : Matrice de corrélation entre les indices de rétention et les descripteurs sélectionnés.

	Ir	En de solvation	Masse moléculaire (MW)
Ir	1,000		
En de solvation	-0,855	1,000	
Masse moléculaire (MW)	0,981	-0,757	1,000



La figure III. 8 reproduit les valeurs des indices de rétention prédites en fonction des indices de rétention expérimentales, fait ressortir une faible dispersion caractéristique d'un bon ajustement.

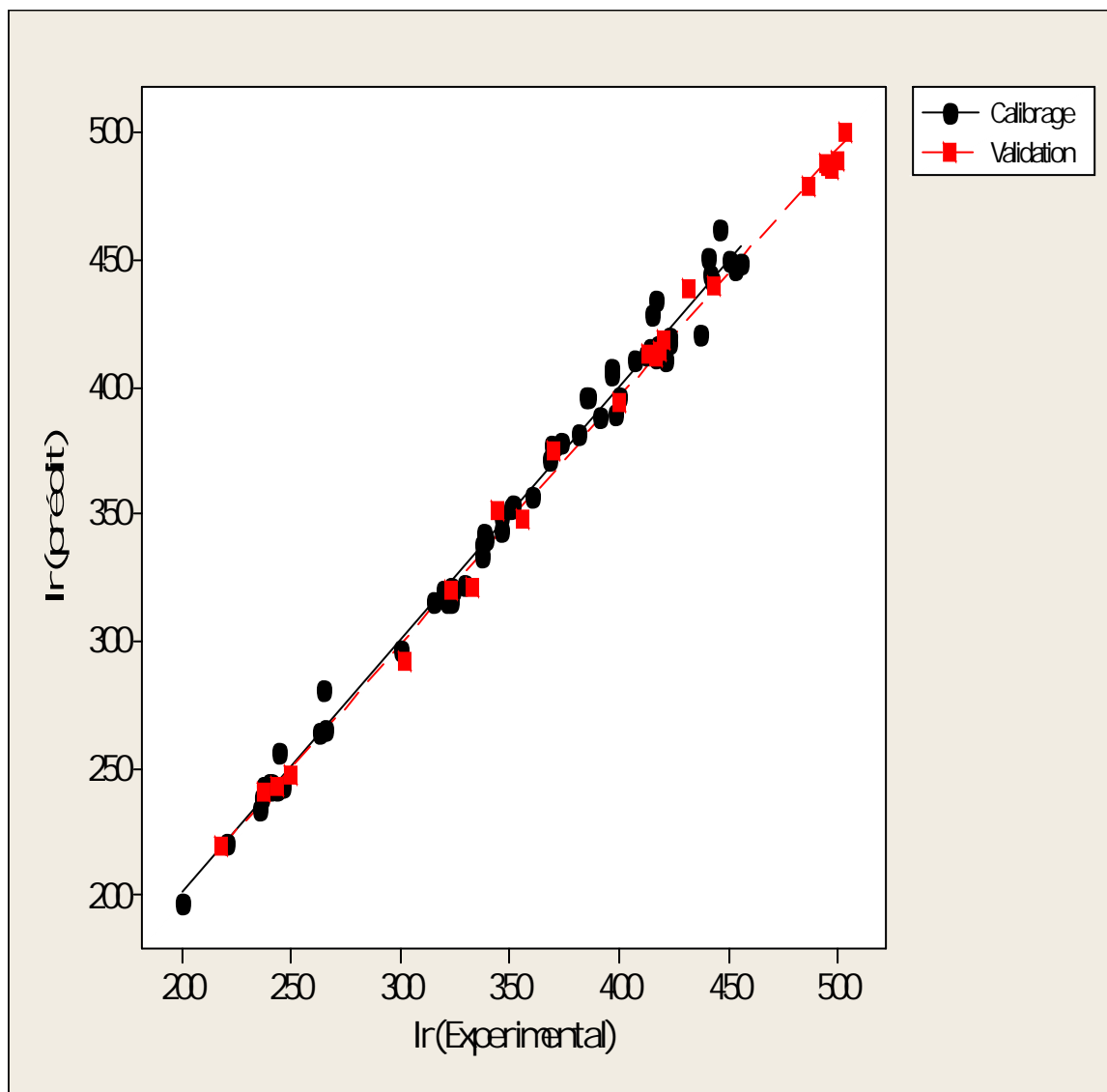


Figure III.8 : Les valeurs des indices de rétention prédites en fonction des indices de rétention expérimentales.

Les contributions des deux descripteurs impliqués diminuent dans l'ordre suivant: Masse moléculaire (MW) (72,66 %) > En de solvation (27,34%).

Le diagramme de Williams (figure III.9) a été utilisé pour la définition du domaine d'application du modèle RLM. Tous les composés sont situés dans l'intervalle de  $\pm 3$  et présentent  $h_i < h^*$  caractéristiques de l'absence de composé aberrant /ou influent.

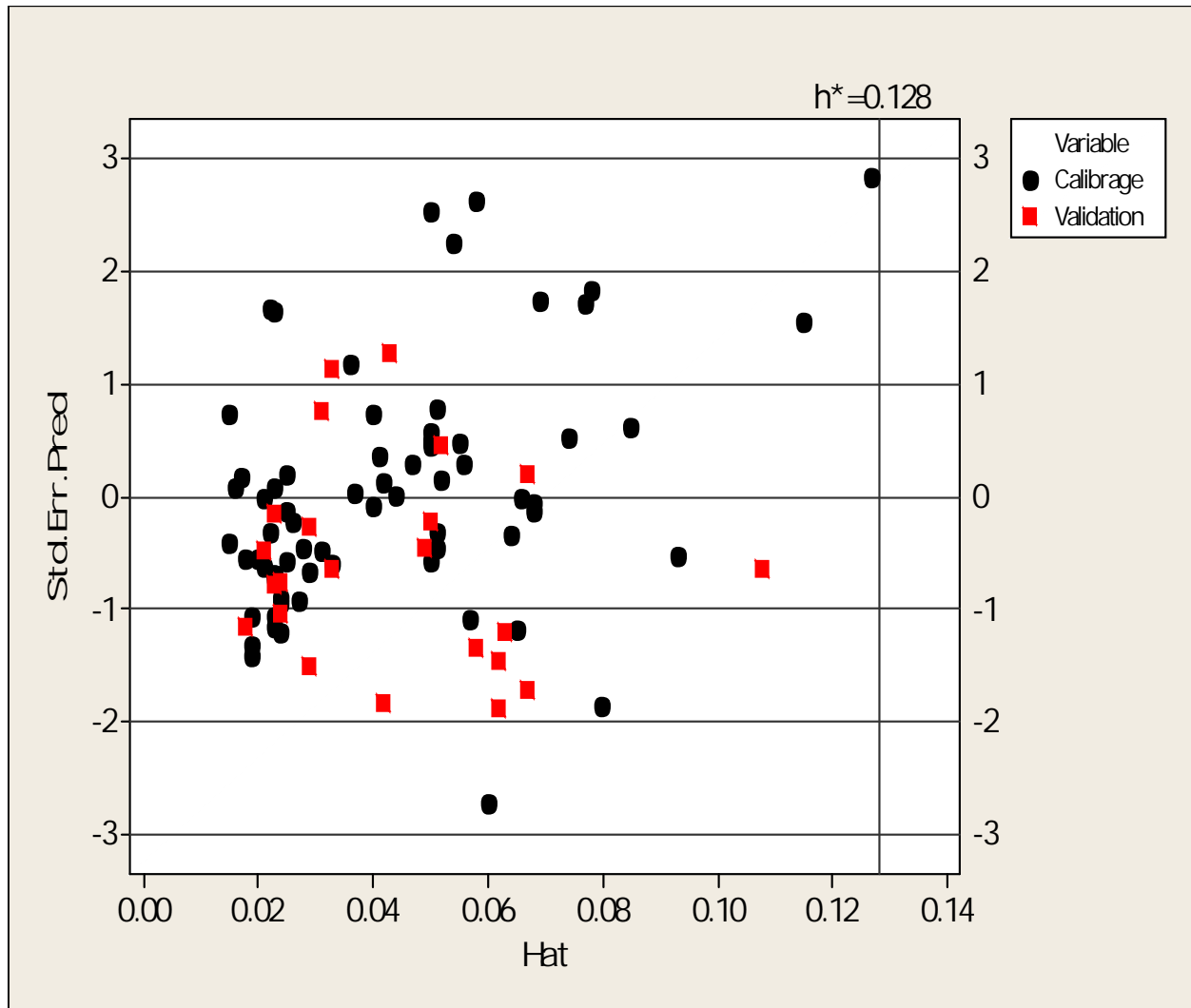


Figure III .9 : Diagramme de Williams pour le modèle QSRR optimal.

Le test de randomisation (Fig III.10) permet de s'assurer que le modèle établi a une base réelle, et qu'il n'est pas dû au hasard.

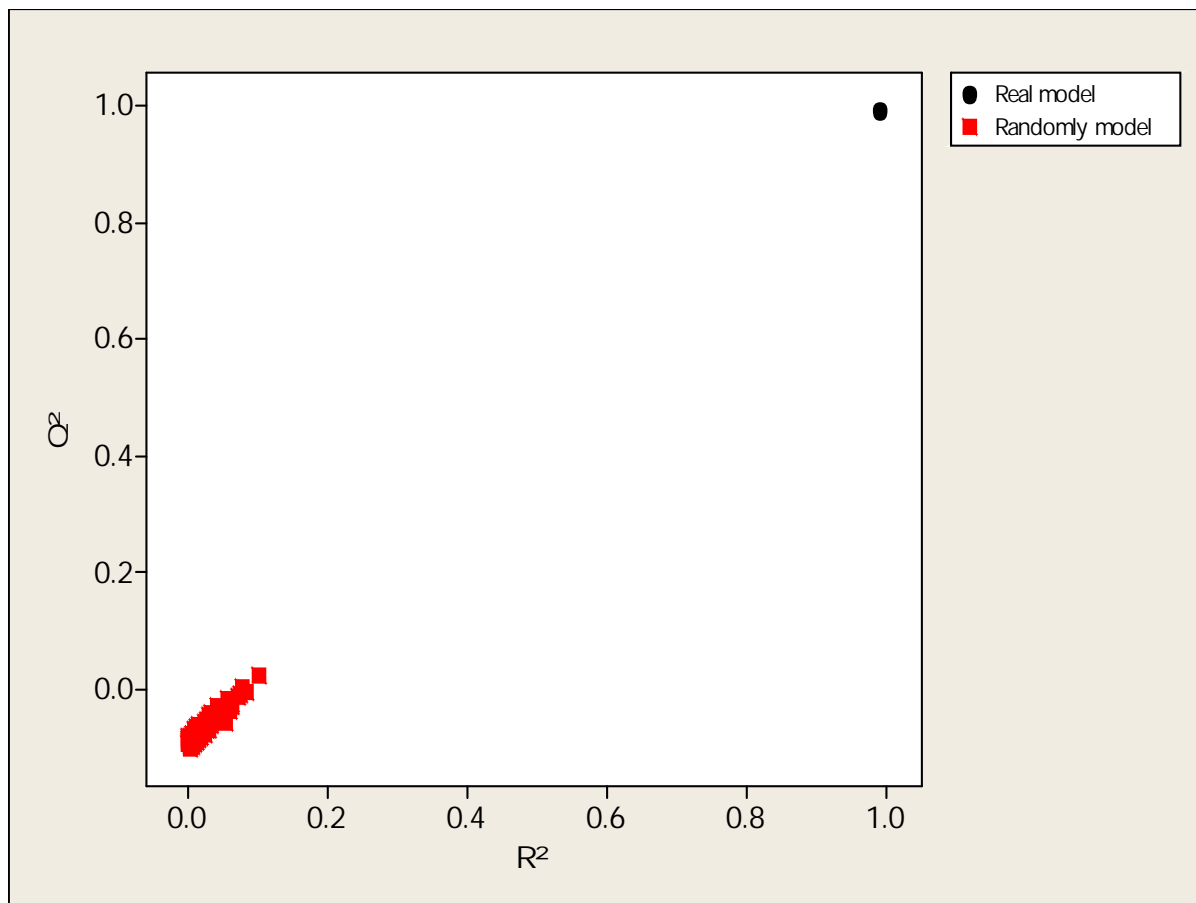


Figure III.10 : Test de randomisation associé au modèle de QSRR précédent. Les carrés présentent les rétentions au hasard et le cercle correspond à la rétention réelle.



**CONCLUS ON  
GÉNÉRALE**

## CONCLUSION GÉNÉRALE

---

La modélisation QSXR (X : activité ; propriété ; rétention); est le procédé par lequel des liens quantitatifs sont établis entre la structure moléculaire d'un ensemble de composés avec les variables à expliquer. Elle présente un enjeu industriel important, car elle permet de réduire les délais et les coûts de développement.

Nous avons appliqué cette méthodologie QSXR pour relier le temps et l'indice de rétention de 84 produits toxiques (pesticides, PCBs) et de 93 HAPS, et le facteur de bioaccumulation de 58 PCBs à des descripteurs moléculaires théoriques reflétant certaines particularités des molécules considérées.

Les deux bases de données : 84 composés (Pesticides et PCBs) et les 58 PCBs ont été divisées selon l'algorithme Kennard et Stone comme suit :

- Pour la base de données (Pesticides et PCBs), les 84 données ont été divisées en deux sous ensembles disjoints, le premier contenant 67 composés pour la construction de modèle, et le deuxième contenant 17 composés pour la validation externe.
- Pour la base de données des PCBs, les 58 données ont été divisées en deux sous ensembles disjoints, la première contenant 30 composés pour la construction de modèle, et le deuxième contenant 28 composés pour la validation externe.
- Pour la base de données des HAPs, les 93 données ont été divisées en deux sous ensembles disjoints, la première contenant 70 composés pour la construction de modèle, et le deuxième contenant 23 composés pour la validation externe.

Trois modèles ont été établis en appliquant une approche hybride algorithme génétique / régression linéaire multiple pour la première base de données et simple pour la deuxième base de données.

La sélection des variables explicatives a été réalisée par algorithme génétique, dans la version MOBYDIGS de TODESCHINI, en maximisant  $Q^2_{L00}$ .

Les statistiques réunies ci-après permettent de faire des comparaisons, et de tirer plusieurs conclusions.

- Pour l'ensemble des (Pesticides et PCBs) : (modélisation de temps de rétention :

## CONCLUSION GÉNÉRALE

$n_{tr}$	$n_{ext}$	$Q^2_{LOO}(\%)$	$R^2(\%)$	$Q^2_{LMO/50}(\%)$	$Q^2_{BOOT}(\%)$	$R^2_{adj}(\%)$	$Q^2_{ext}(\%)$
67	17	88,15	90,54	86,45	86,58	89,59	87,15
<b>EQMC</b>	<b>EQMP</b>	$EQMP_{ext}$	$S$	$F$			
0,036	0,04	0,042	0,0381	95,6528			

➤ Pour l'ensemble des PCBs (modélisation de facteur de bioaccumulation) :

$n_{tr}$	$n_{ext}$	$Q^2_{LOO}(\%)$	$R^2(\%)$	$Q^2_{LMO/50}(\%)$	$Q^2_{BOOT}(\%)$	$R^2_{adj}(\%)$	$Q^2_{ext}(\%)$
30	28	90,28	91,53	89,52	89,23	91,23	90,44
<b>EQMC</b>	<b>EQMP</b>	$EQMP_{ext}$	$S$	$F$			
0,3	0,321	0,319	0,3106	302,7423			

➤ Pour l'ensemble des HAPs (modélisation de l'indice de rétention) :

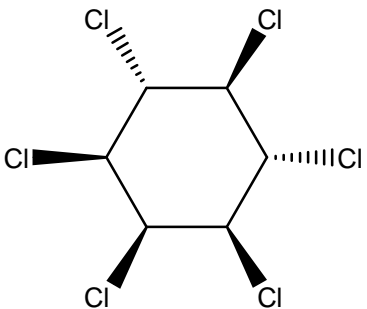
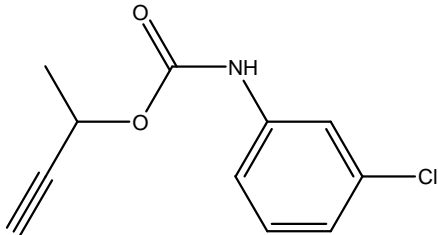
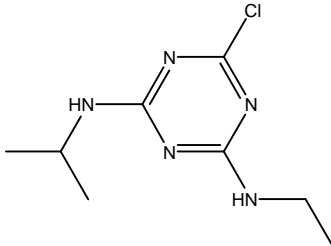
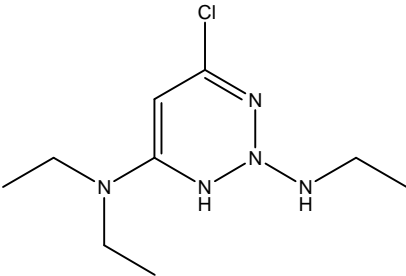
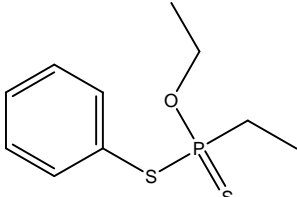
$n_{tr}$	$n_{ext}$	$Q^2_{LOO}(\%)$	$R^2(\%)$	$Q^2_{LMO/50}(\%)$	$Q^2_{BOOT}(\%)$	$R^2_{adj}(\%)$	$Q^2_{ext}(\%)$
70	23	99,18	99,27	99,1324	99,12	99,25	99,17
<b>EQMC</b>	<b>EQMP</b>	$EQMP_{ext}$	$S$	$F$			
6,108	6,475	6,519	6,2428	4569,847			

Pour les trois modèles développés :

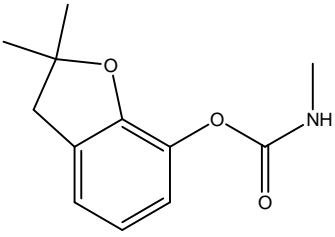
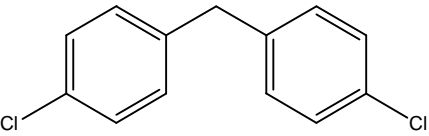
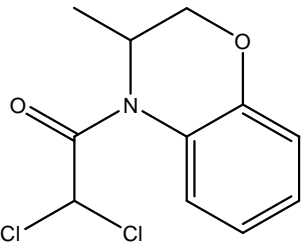
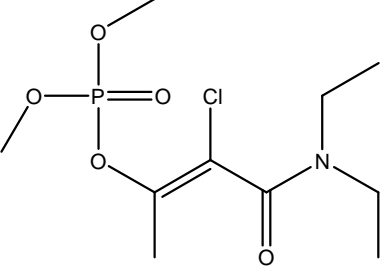
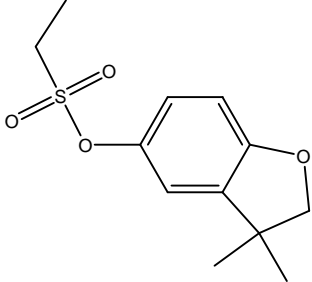
- ❖ Les valeurs de  $R^2$  et de  $R^2_{adj}$  montrent, à chaque fois, la qualité de l'ajustement.
- ❖ La très faible différence entre  $R^2$  et  $Q^2$  renseigne sur la robustesse du modèle qui est, en outre, très hautement significatif (valeur élevée de la statistique F de Fisher).
- ❖ La similitude d'EQMC et EQMP signifie que la capacité de prédiction interne du modèle n'est pas trop dissemblable de son pouvoir d'ajustement.
- ❖ Le très faible écart entre  $Q^2$  et  $Q^2_{LMO/50}$  démontre la bonne stabilité dans la validation interne, et la validation par bootstrap ( $Q^2$  Bootstrap) confirme tout à la fois la capacité de prédiction interne et la stabilité des modèles.
- ❖ La validation statistique externe ( $Q^2_{EXT}$ ;  $EQMP_{EXT}$ ) atteste de la bonne capacité prédictive des composés n'ayant pas participé au calcul des modèles, mais qui appartiennent cependant au domaine chimique de l'ensemble d'essai.

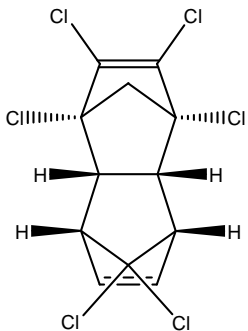
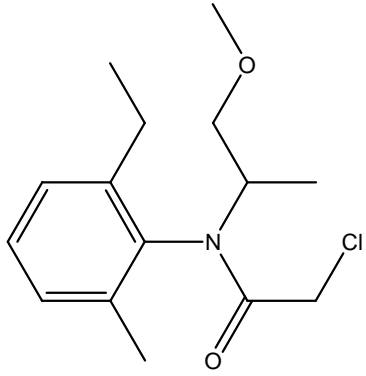
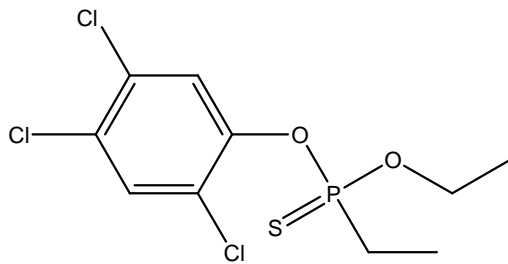
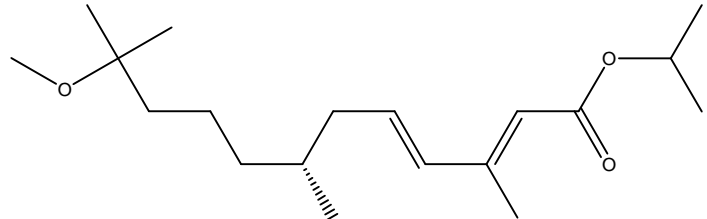


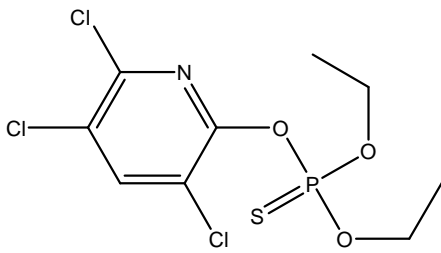
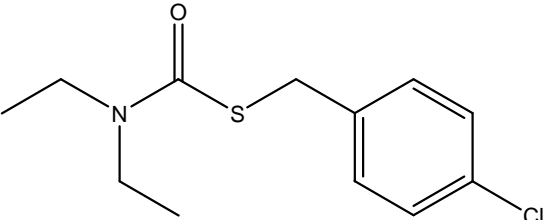
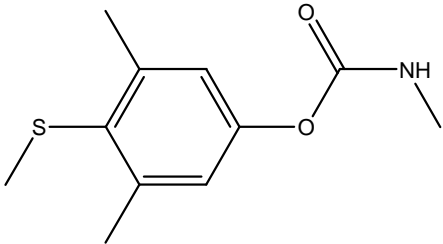
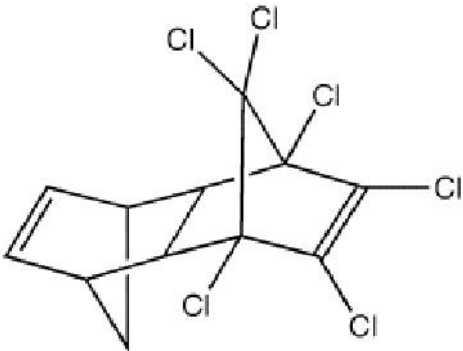
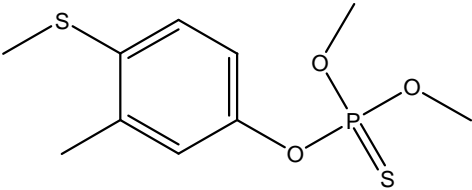
**ANNEXE DES  
PEST C DES**

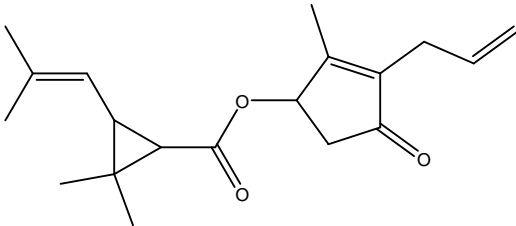
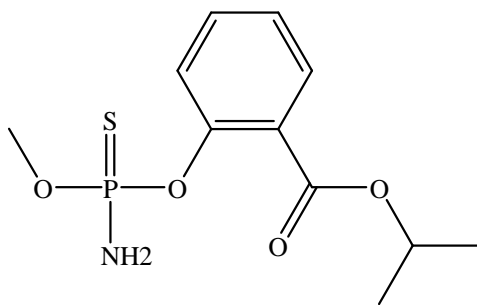
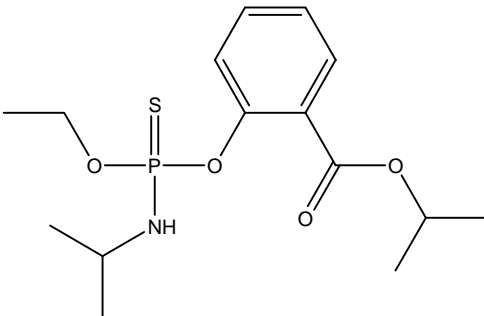
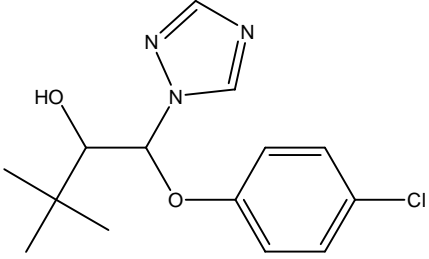
Composé et N° de cas	Structure et Nomenclature UAPAC
a-666 319-84-6	 <p data-bbox="751 647 1222 683">1,2,3,4,5,6-Hexachloro-cyclohexane</p>
Chlorbufam 1967-16-4	 <p data-bbox="676 965 1294 994">(3-Chloro-phenyl)-carbamic acid 1-methyl-prop-2-ynyl ester</p>
Atrazine 1912-24-9	 <p data-bbox="730 1319 1241 1348">6-Chloro-<i>N</i>-ethyl-<i>N'</i>-isopropyl-[1,3,5]triazine-2,4-diamine</p>
Trietazine 1912-26-1	 <p data-bbox="767 1700 1203 1729">2-Ethylamino-4-Diethylamino-6-chlorotriazine;</p>
Fonofos (Dyfonate) 944-22-9	 <p data-bbox="794 1960 1177 1989">O-Ethyl S-phenyl ethylphosphonodithioate</p>

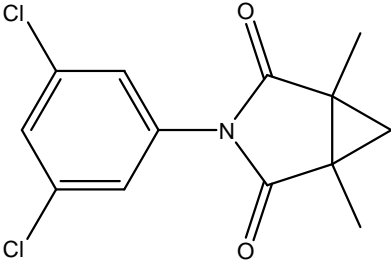
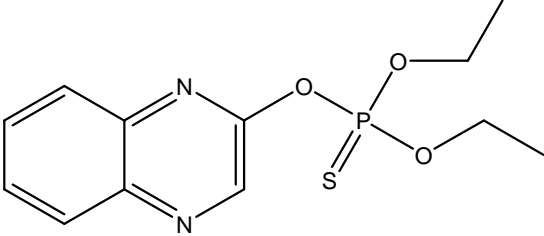
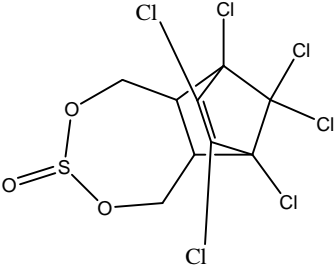
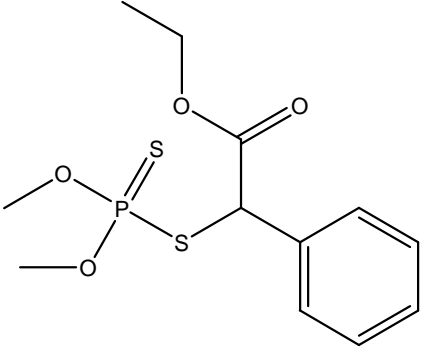


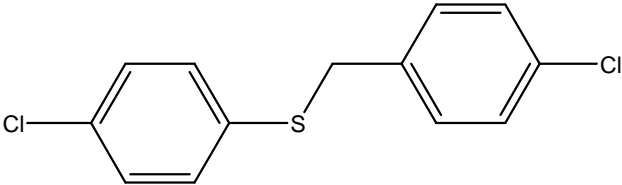
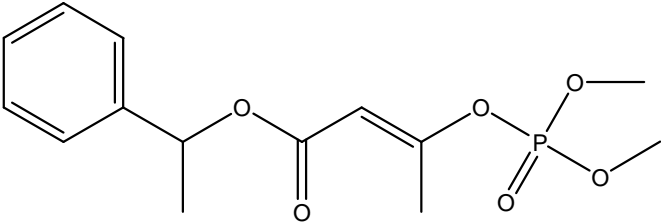
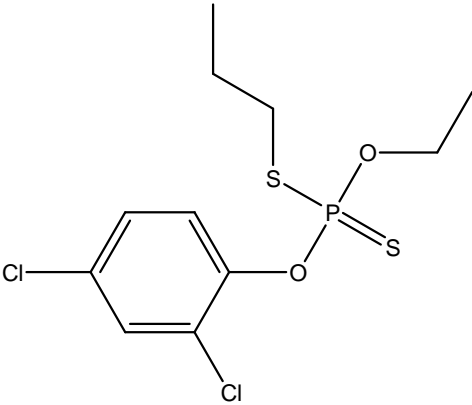
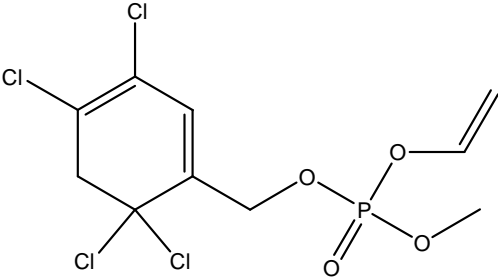
Composé et N° de cas	Structure et Nomenclature UAPAC
Carbofuran  1563-66-2	 <p data-bbox="726 571 1252 604">2,3-dihydro-2,2-dimethylbenzofuran-7-yl methylcarbamate</p>
4,4'-DDM  101-76-8	 <p data-bbox="845 806 1133 840">Bis (4-chlorophenyl) methane</p>
Benoxacor  98730-04-2	 <p data-bbox="614 1176 1364 1209">2,2-Dichloro-1-(3-methyl-2,3-dihydro-benzo[1,4]oxazin-4-yl)-ethanone</p>
Phosphamidon  113171-21-6	 <p data-bbox="566 1556 1412 1590">Phosphoric acid 2-chloro-2-diethylcarbamoyl-1-methyl-vinyl ester dimethyl ester</p>
Benfuresate  68505-69-1	 <p data-bbox="686 1948 1284 1982">Ethanesulfonic acid 3,3-dimethyl-2,3-dihydro-benzofuran-5-yl ester</p>

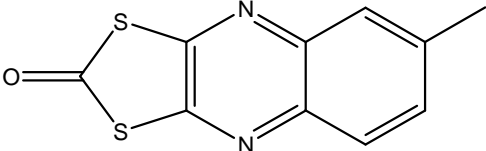
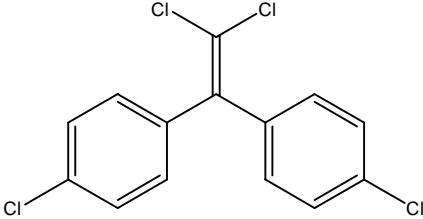
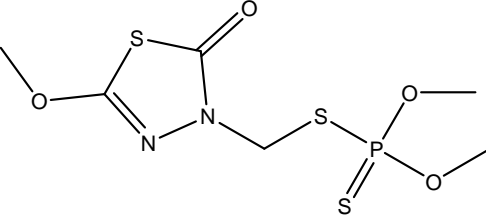
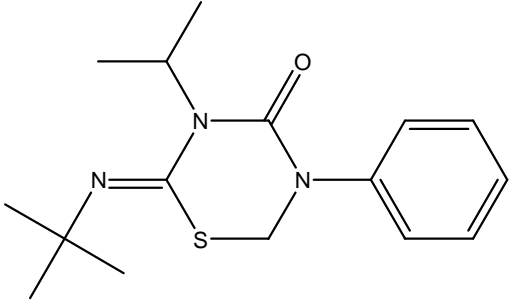
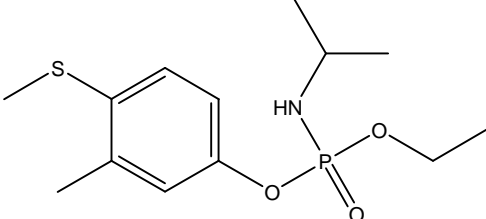
Composé et N° de cas	Structure et Nomenclature UAPAC
Aldrin  309-00-2	 <p>(1R,4S,4aS,5S,8R,8aR)-1,2,3,4,10,10-hexachloro-1,4,4a,5,8,8a-hexahydro-1,4:5,8-dimethanonaphthalene</p>
Metolachlor  51218-45-2	 <p>2-Chloro-<i>N</i>-(2-ethyl-6-methyl-phenyl)-<i>N</i>-(2-methoxy-1-methyl-ethyl)-acetamide</p>
Trichloronate  327-98-0	 <p>Ethyl-phosphonothioic acid <i>O</i>-ethyl ester <i>O</i>-(2,4,5-trichloro-phenyl) ester</p>
Methoprene  40596-69-8	 <p>isopropyl (E,E)-(RS)-11-methoxy-3,7,11-trimethyldodeca-2,4-dienoate</p>

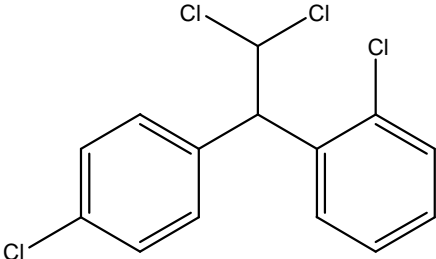
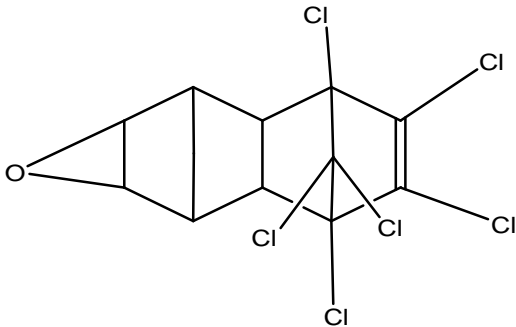
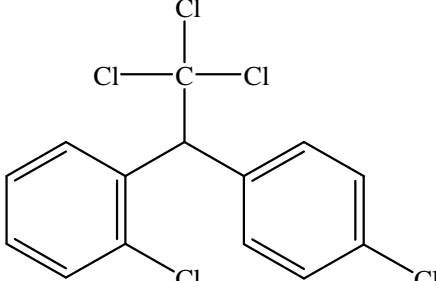
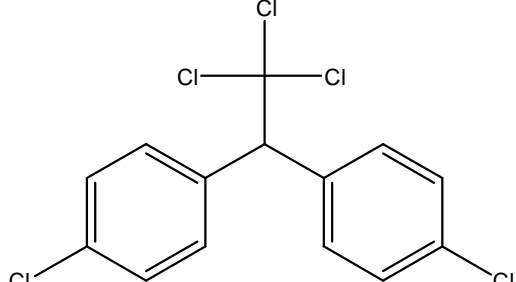
Composé et N° de cas	Structure et Nomenclature UAPAC
Chlorpyrifos  2921-88-2	 <p>O,O-diethyl O-3,5,6-trichloro-2-pyridyl phosphorothioate</p>
Thiobencarb  28249-77-6	 <p>S-4-chlorobenzyl diethyl(thiocarbamate)</p>
Methiocarb  2032-65-7	 <p>4-methylthio-3,5-xyllyl methylcarbamate</p>
Isodrin  465-73-6	 <p>(1alpha,4alpha,4abeta,5beta,8beta,8abeta)-1,2,3,4,10,10-hexachloro-1,4,4a,5,8,8a-hexahydro-1,4:5,8-Dimethanonaphthalene</p>
Fenthion  55-38-9	 <p>Thiophosphoric acid <i>O,O'</i>-dimethyl ester <i>O''</i>-(3-methyl-4-methylsulfanyl-phenyl) ester</p>

Composé et N° de cas	Structure et Nomenclature UAPAC
<p>Allethrin</p> <p>584-79-2</p>	 <p>2-methyl-4-oxo-3-(2-propenyl)-2-cyclopenten-1-yl 2,2-dimethyl-3-(2-methyl-1-propenyl)cyclopropanecarboxylate</p>
<p>Isocarbophos</p> <p>24353-61-5</p>	 <p>2-(Amino-methoxy-thiophosphoryloxy)-benzoic acid isopropyl ester</p>
<p>Isofenphos</p> <p>18181-70-9</p>	 <p>isopropyl (RS)-O-[ethoxy(isopropylamino)phosphinothioyl]salicylate</p>
<p>Triadimenol</p> <p>55219-65-3</p>	 <p>b-(4-chlorophenoxy)-a-(1,1-dimethylethyl)-1H-1,2,4-triazole-1-ethanol</p>

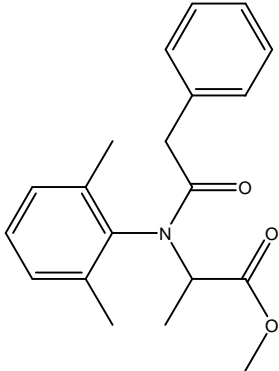
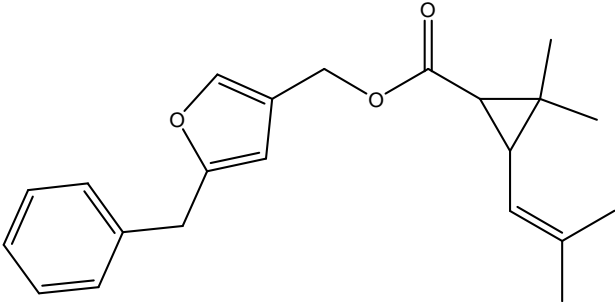
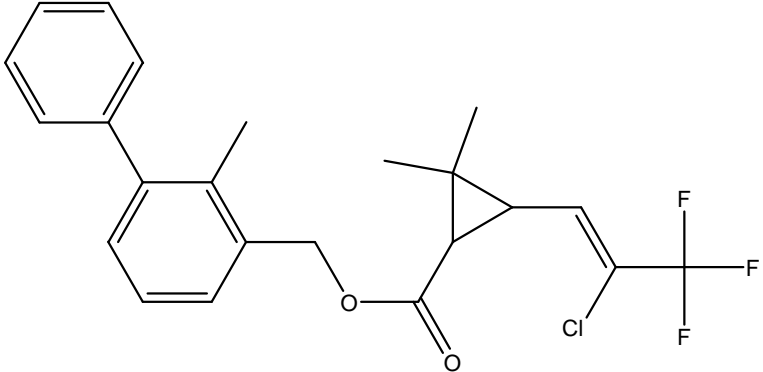
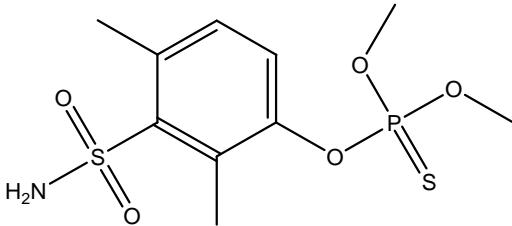
Composé et N° de cas	Structure et Nomenclature UAPAC
Procymidone 32809-16-8	 <p data-bbox="488 663 1302 689">3-(3,5-dichlorophenyl)-1,5-dimethyl-3-azabicyclo[3.1.0]hexane-2,4-dione</p>
Quinalphos 13593-03-8	 <p data-bbox="707 1048 1254 1075">O,O-diethyl O-quinoxalin-2-yl phosphorothioate</p>
Alpha-endosulfan 959-98-8	 <p data-bbox="488 1503 1294 1529">1,9,10,11,12,12-Hexachloro-4,6-dioxo-5-thia-tricyclo[7.2.1.0<sup>2,8</sup>]dodec-10-ene 5-oxide</p>
Phenthoate 2597-03-7	 <p data-bbox="488 1928 1222 1955">(Dimethoxy-thiophosphorylsulfanyl)-phenyl-acetic acid ethyl ester</p>

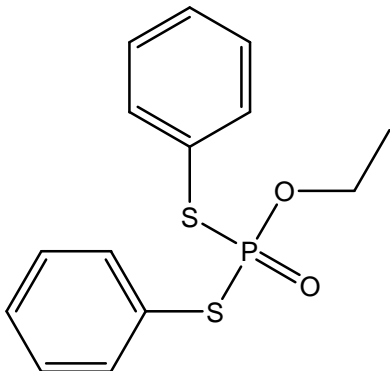
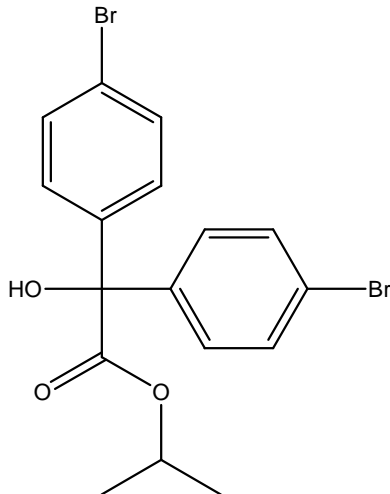
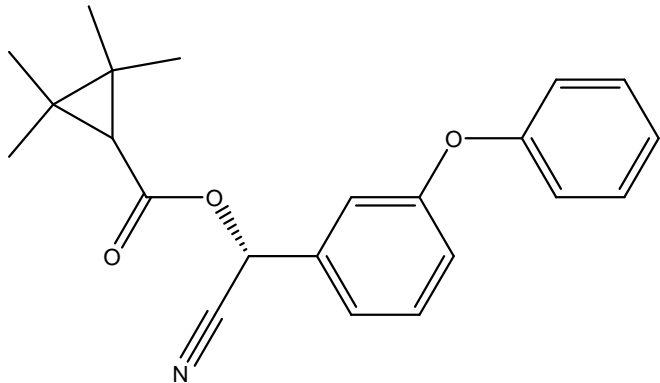
Composé et N° de cas	Structure et Nomenclature UAPAC
Chlorbenseide 103-17-3	 <p data-bbox="772 533 1200 562">4-chlorobenzyl 4-chlorophenyl sulfide</p>
Crotoxyphos (Ciodrin) 7700-17-6	 <p data-bbox="604 871 1369 900">3-(Dimethoxy-phosphoryloxy)-but-2-enoic acid 1-phenyl-ethyl ester</p>
Prothifos 34643-46-4	 <p data-bbox="488 1350 1398 1379">Dithiophosphoric acid <i>O</i>-(2,4-dichloro-phenyl) ester <i>O'</i>-ethyl ester <i>S</i>-propyl ester</p>
Tetrachlorvinphos 22248-79-9	 <p data-bbox="488 1776 1485 1805">Phosphoric acid methyl ester 3,4,6,6-tetrachloro-cyclohexa-1,3-dienylmethyl ester vinyl ester</p>

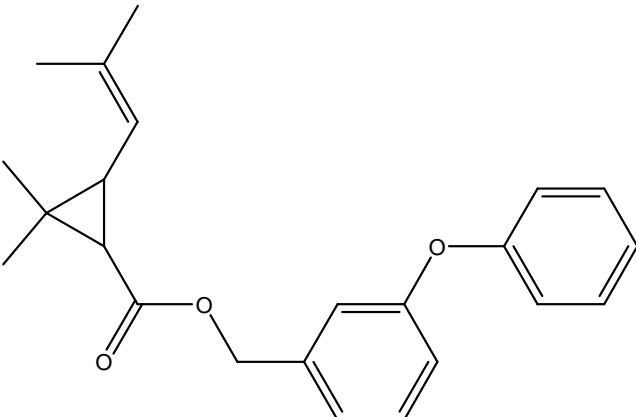
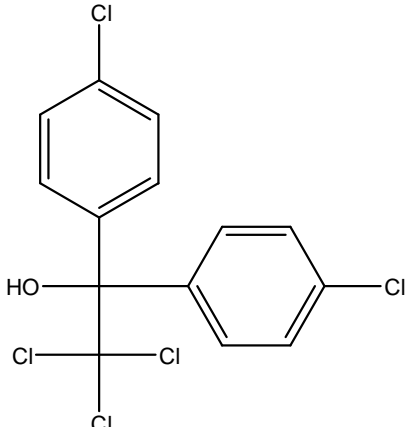
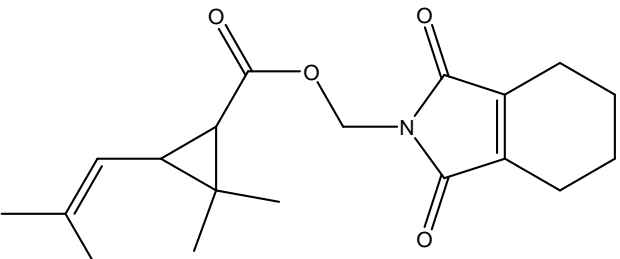
Composé et N° de cas	Structure et Nomenclature UAPAC
Chinomethionat  2439-01-2	 <p data-bbox="734 539 1241 568">6-methyl-1,3-dithiolo[4,5-b]quinoxalin-2-one</p>
4,4'-DDE  72-55-9	 <p data-bbox="766 891 1209 920">1,1'-(Dichloroethylenylidene)bis(4-chlorobenzene)</p>
Methidathion  950-37-8	 <p data-bbox="486 1182 1489 1211">S-2,3-dihydro-5-methoxy-2-oxo-1,3,4-thiadiazol-3-ylmethyl O,O-dimethyl phosphorodithioate</p>
Buprofezin  69327-76-0	 <p data-bbox="609 1637 1362 1666">(Z)-2-tert-butylimino-3-isopropyl-5-phenyl-1,3,5-thiadiazinan-4-one</p>
Fenamiphos  22224-92-6	 <p data-bbox="639 1935 1335 1964">ethyl 3-methyl-4-(methylthio)phenyl (1-methylethyl)phosphoramidate</p>

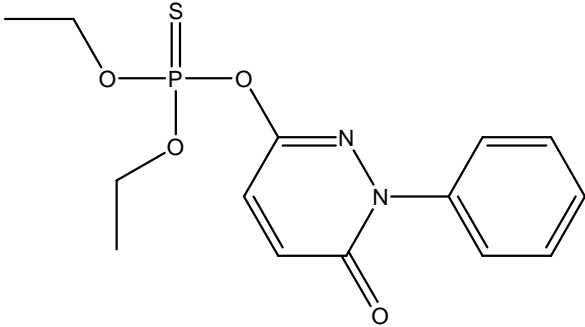
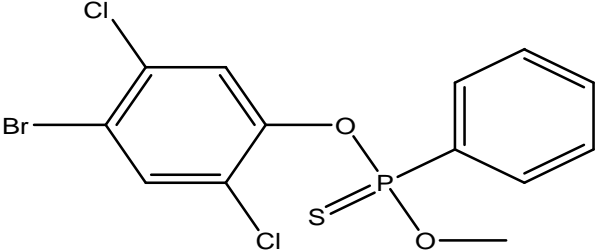
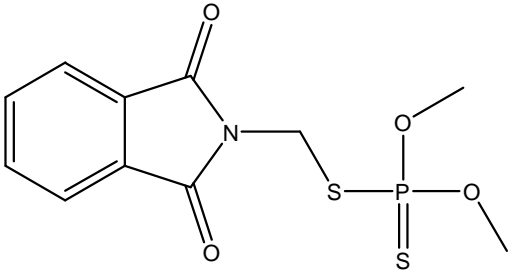
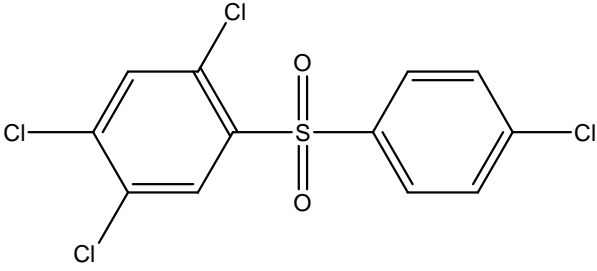
Composé et N° de cas	Structure et Nomenclature UAPAC
2,4'-DDD  53-19-0	 <p data-bbox="667 604 1316 638">1,1-Dichloro-2-(o-chlorophenyl)-2-(p-chlorophenyl)ethane</p>
Endrin  72-20-8	
2,4 -DDT  789-02-6	
4,4'-DDT  50-29-3	 <p data-bbox="702 1736 1220 1769">1,1,1-Trichloro-2,2-bis-(4'-chlorophenyl)ethane</p>

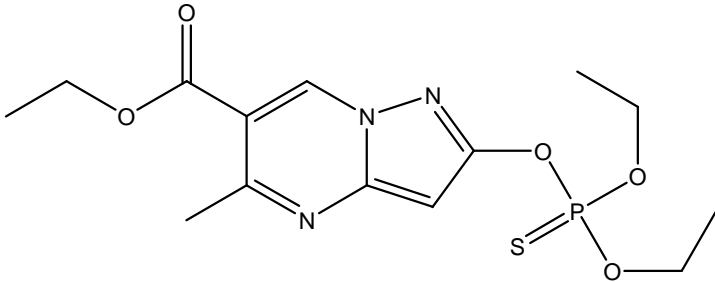
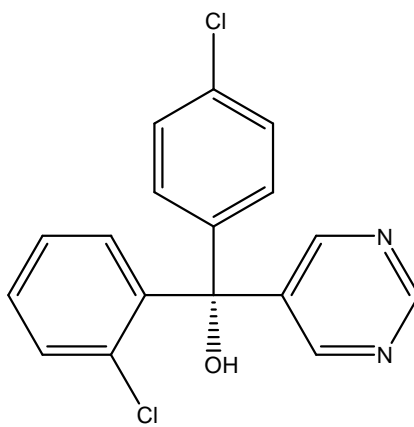
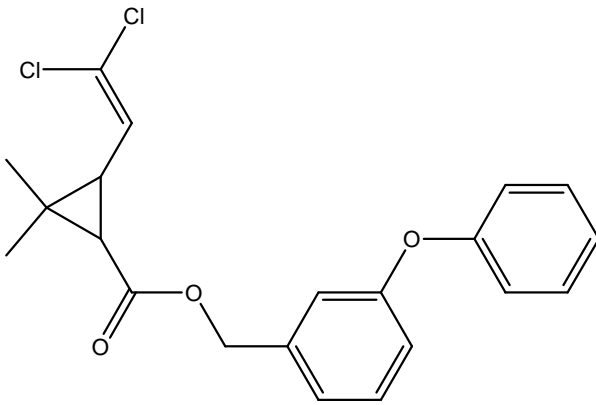
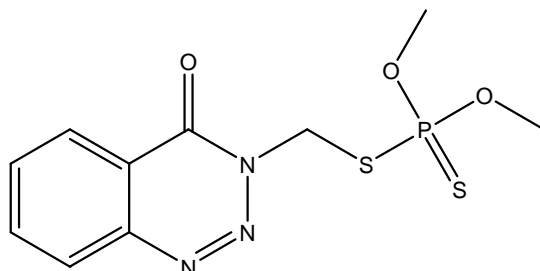


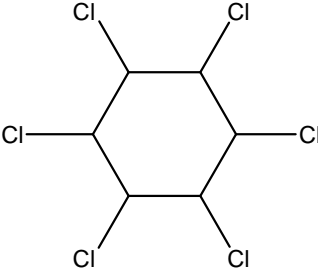
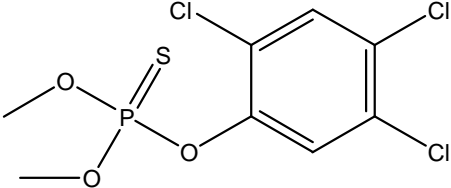
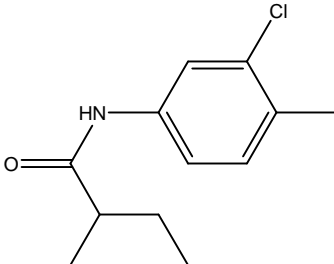
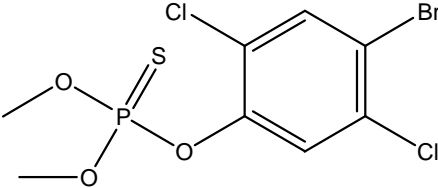
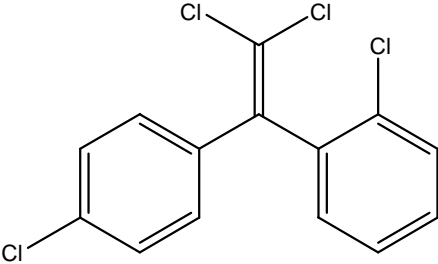
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>Benalaxyl 71626-11-4</p>	 <p>methyl N-(2,6-dimethylphenyl)-N-(phenylacetyl)-DL-alaninate</p>
<p>Resmethrin 10453-86-8</p>	 <p>[5-(phenylmethyl)-3-furanyl]methyl 2,2-dimethyl-3-(2-methyl-1-propen-1-yl)cyclopropanecarboxylate</p>
<p>Bifenthrin 82657-04-3</p>	 <p>2-methylbiphenyl-3-ylmethyl (1RS,3RS)-3-[(Z)-2-chloro-3,3,3-trifluoroprop-1-enyl]-2,2-dimethylcyclopropanecarboxylate</p>
<p>Famphur 52-85-7</p>	 <p>O-4-dimethylsulfamoylphenyl O,O-dimethyl phosphorothioate</p>

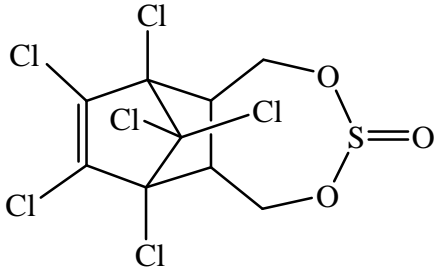
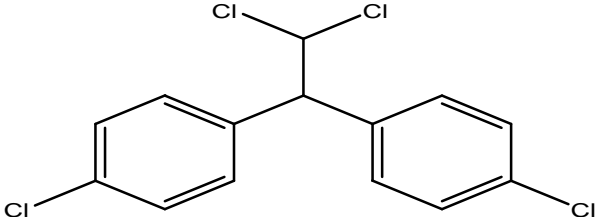
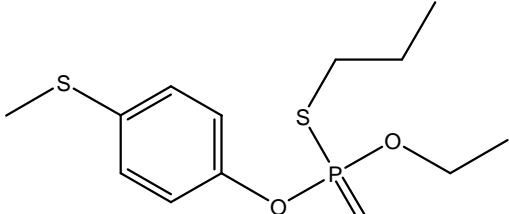
Composé et N° de cas	Structure et Nomenclature UAPAC
Edifenphos  17109-49-8	 <p data-bbox="678 716 1292 748">Dithiophosphoric acid <i>O</i>-ethyl ester <i>S,S'</i>-diphenyl ester</p>
Bromopropylate  18181-80-1	 <p data-bbox="486 1288 1252 1319">1-methylethyl 4-bromo-<math>\alpha</math>-(4-bromophenyl)-<math>\alpha</math>-hydroxybenzeneacetate</p>
Fenpropathrin  39515-41-8	 <p data-bbox="486 1742 1332 1774">(RS)-<math>\alpha</math>-cyano-3-phenoxybenzyl 2,2,3,3-tetramethylcyclopropanecarboxylate</p>

Composé et N° de cas	Structure et Nomenclature UAPAC
Phenothrin  26002-80-2	 <p data-bbox="486 840 1484 871">(3-phenoxyphenyl)methyl 2,2-dimethyl-3-(2-methyl-1-propen-1-yl)cyclopropanecarboxylate</p>
Dicofol (Kelthane)  115-32-2	 <p data-bbox="726 1344 1244 1377">2,2,2-trichloro-1,1-bis(4-chlorophenyl)ethanol</p>
Tetramethrin  7696-12-0	 <p data-bbox="486 1691 1484 1747">(1,3,4,5,6,7-hexahydro-1,3-dioxo-2H-isoindol-2-yl)methyl 2,2-dimethyl-3-(2-methyl-1-propen-1-yl)cyclopropanecarboxylate</p>

Composé et N° de cas	Structure et Nomenclature UAPAC
Pyridafenthion  119-12-0	 <p data-bbox="724 674 1246 734">O,O-Diethyl O-(2,3-dihydro-3-oxo-2-phenyl-6-pyridazinyl)phosphorothioate</p>
Leptophos  21609-90-5	 <p data-bbox="533 1104 1442 1133">O-(4-bromo-2,5-dichlorophenyl) O-methyl P-phenylphosphonothioate</p>
Imidan  732-11-6	 <p data-bbox="655 1520 1270 1550">O,O-dimethyl S-phthalimidomethyl phosphorodithioate</p>
Tedion (Tetradifon)  116-29-0	 <p data-bbox="679 1861 1294 1890">1,2,4-Trichloro-5-(4-chloro-benzenesulfonyl)-benzene</p>

Composé et N° de cas	Structure et Nomenclature UAPAC
Pyrazophos  13457-18-6	 <p data-bbox="486 622 1460 656">ethyl 2-diethoxyphosphinothioxy-5-methylpyrazolo[1,5-a]pyrimidine-6-carboxylate</p>
Fenarimol  60168-88-9	 <p data-bbox="638 1115 1332 1144">(4-Chloro-phenyl)-(2-chloro-phenyl)-pyrimidin-5-yl-methanol</p>
Permethrin  52645-53-1	 <p data-bbox="510 1585 1420 1615">3-(2,2-Dichloro-vinyl)-2,2-dimethyl-cyclopropanecarboxylic acid 3-phenoxy-benzyl ester</p>
Azinphos-methyl (Guthion)  86-50-0	 <p data-bbox="518 1926 1460 1960">S-3,4-dihydro-4-oxo-1,2,3-benzotriazin-3-ylmethyl O,O-dimethyl phosphorodithioate</p>

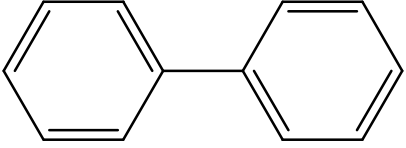
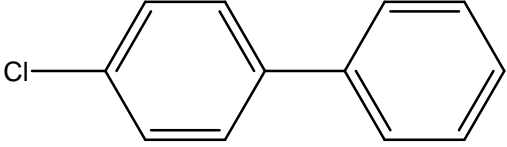
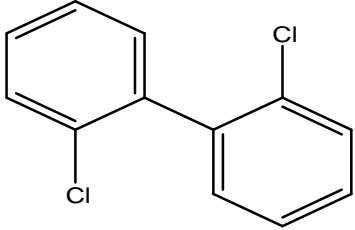
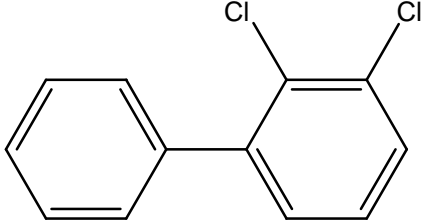
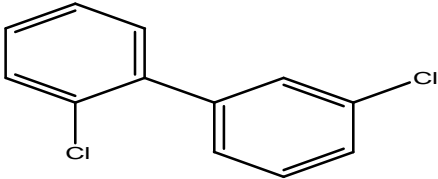
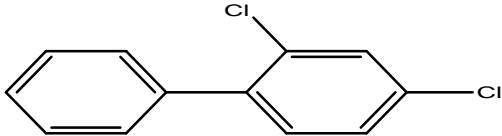
Composé et N° de cas	Structure et Nomenclature UAPAC
Lindane(BHC)  58-89-9	 <p>1,2,3,4,5,6-Hexachloro-cyclohexane</p>
Fenchlorphos  299-84-3	 <p>O,O-dimethyl O-2,4,5-trichlorophenyl phosphorothioate</p>
Pentanochlor  2307-68-8	 <p>N-(3-chloro-4-methylphenyl)-2-methylpentanamide</p>
Bromophos-methyl  2104-96-3	 <p>4-Bromo-2,5-dichlorophenyl dimethyl phosphorothionate</p>
2,4'-DDE  3424-82-6	 <p>1,1-Dichloro-2-(o-chlorophenyl)-2-(p-chlorophenyl)ethene</p>

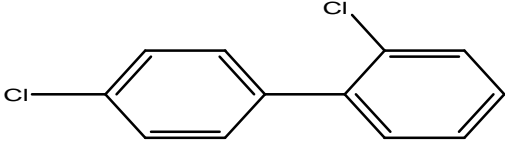
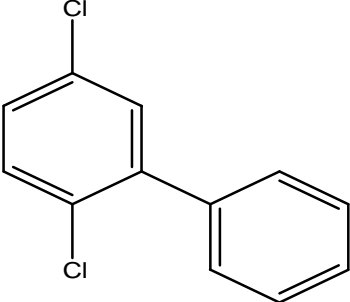
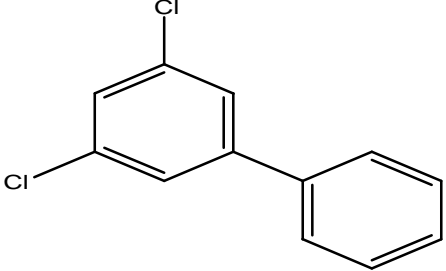
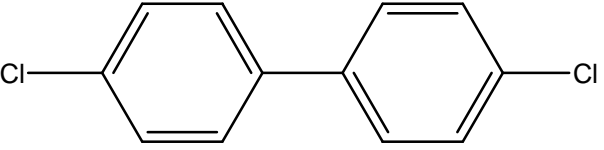
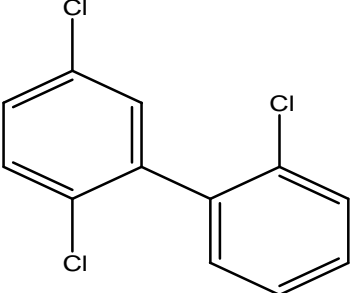
Composé et N° de cas	Structure et Nomenclature UAPAC
Beta -endosulfan  33213-65-9	 <p data-bbox="488 607 1410 640">1,9,10,11,12,12-Hexachloro-4,6-dioxa-5-thia-tricyclo[7.2.1.0<sup>2,8</sup>]dodec-10-ene 5-oxide</p>
4,4'-DDD  72-54-8	 <p data-bbox="708 927 1275 960">1,1-Bis(4-chlorophenyl)-2,2-dichloroethane</p>
Sulprofos  35400-43-2	 <p data-bbox="691 1346 1283 1368">O-ethyl O-[4-(methylthio)phenyl] S-propyl phosphorodithioate</p>

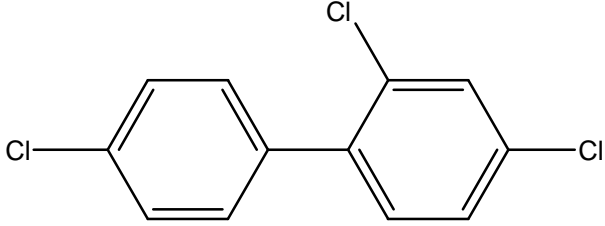
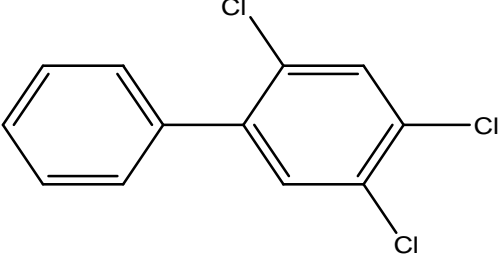
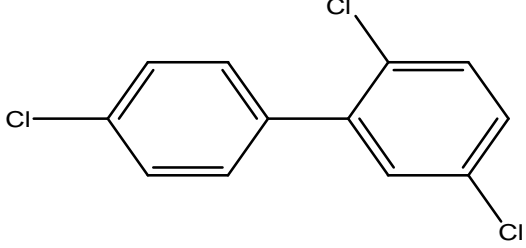
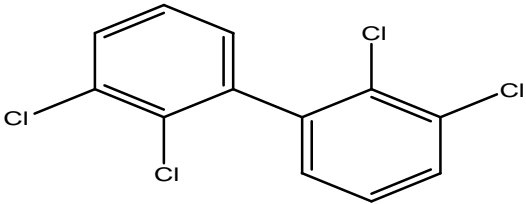
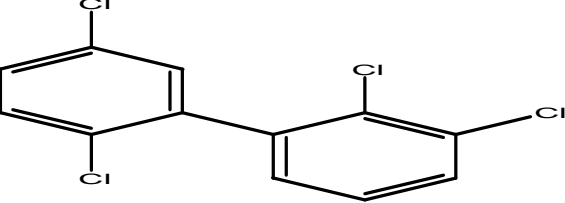


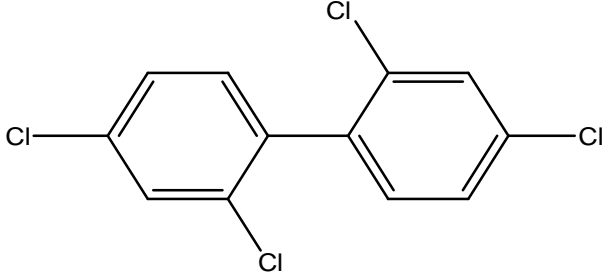
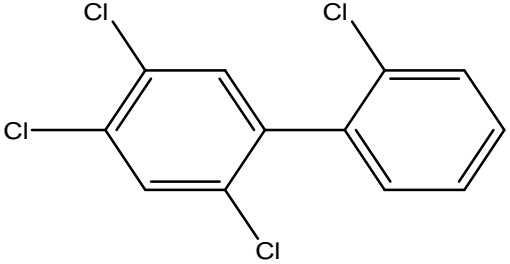
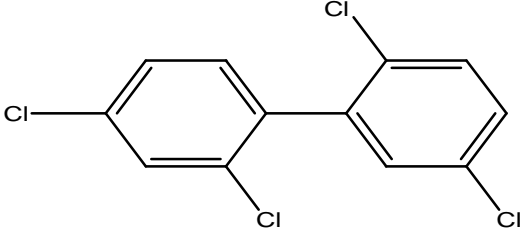
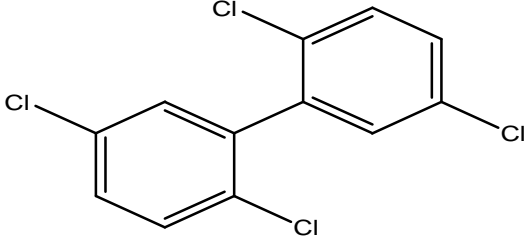
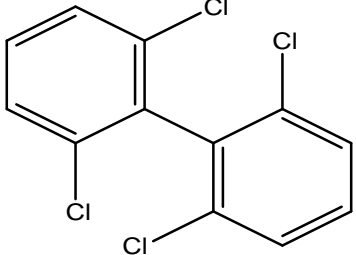
**ANNEXE  
DES PCBs**

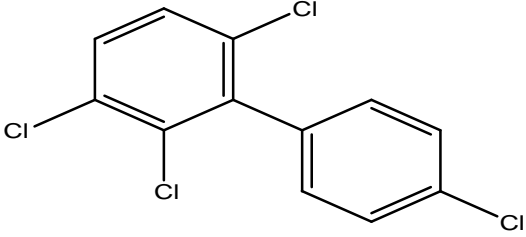
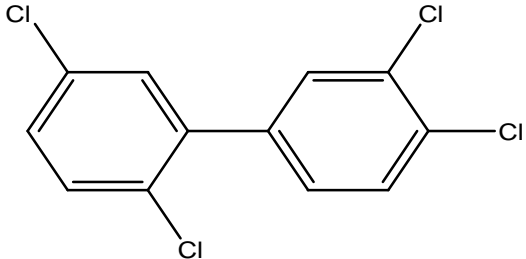
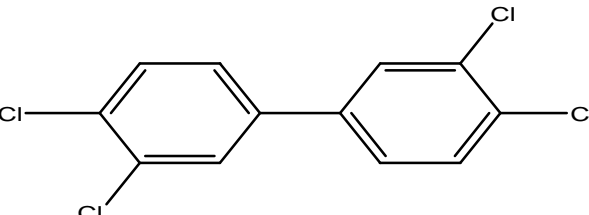
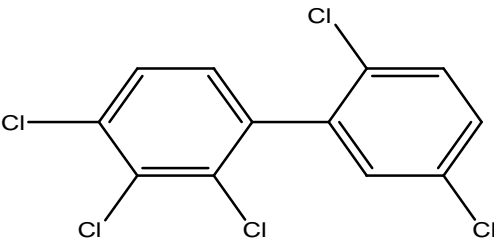
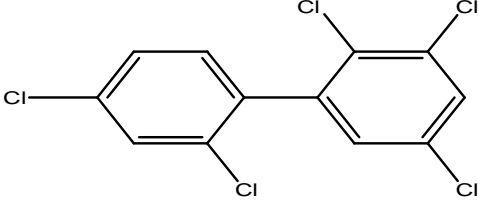


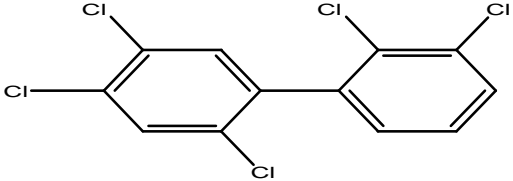
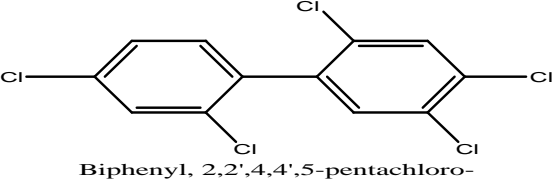
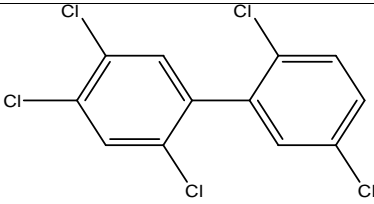
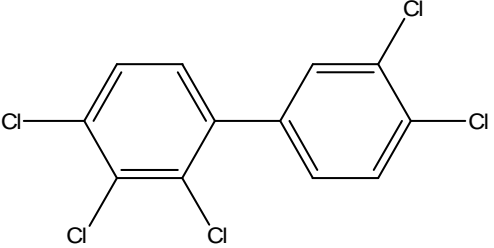
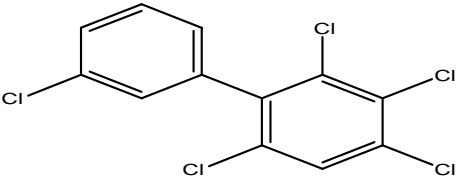
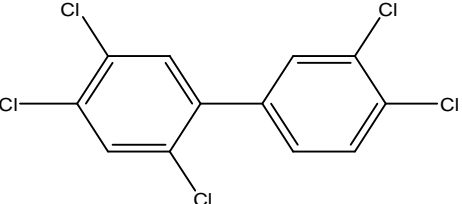
Composé et N° de cas	Structure et Nomenclature UAPAC
PCB 0 92-52-4	 Biphenyl
PCB 3 2051-62-9	 Biphenyl, 4-chloro-
PCB 4 13029-08-8	 Biphenyl, 2,2'-dichloro-
PCB 5 16605-91-7	 Biphenyl, 2,3-dichloro-
PCB 6 25569-80-6	 Biphenyl, 2,3'-dichloro-
PCB 7 33284-50-3	 Biphenyl, 2,4-dichloro-

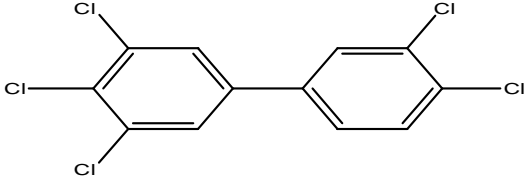
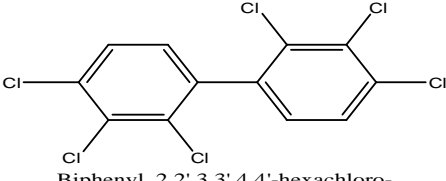
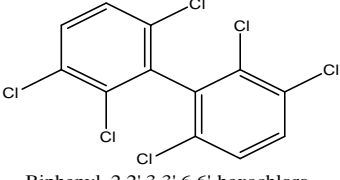
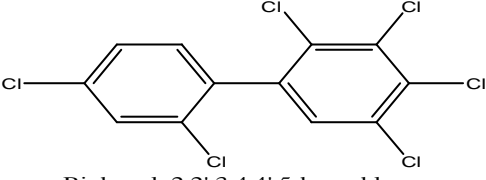
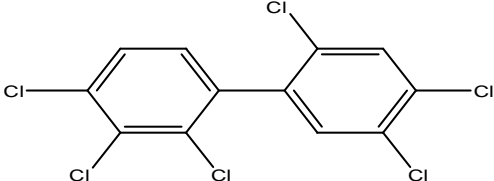
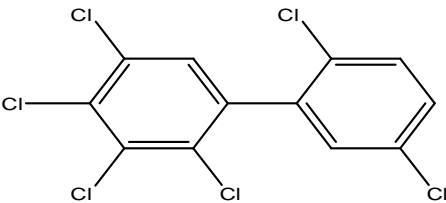
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 8</p> <p>34883-43-7</p>	 <p>Biphenyl, 2,4'-dichloro-</p>
<p>PCB 9</p> <p>34883-39-1</p>	 <p>Biphenyl, 2,5-dichloro-</p>
<p>PCB14</p> <p>34883-41-5</p>	 <p>Biphenyl, 3,5-dichloro-</p>
<p>PCB15</p> <p>2050-68-2</p>	 <p>Biphenyl, 4,4'-dichloro-</p>
<p>PCB 18</p> <p>37680-65-2</p>	 <p>Biphenyl, 2,2',5-trichloro-</p>

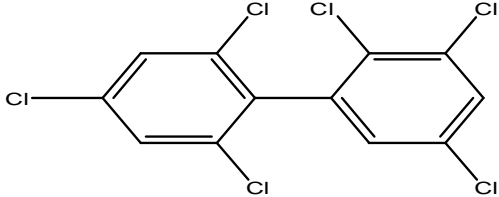
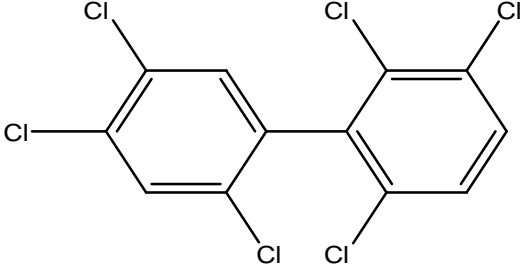
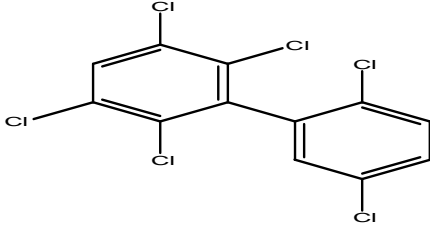
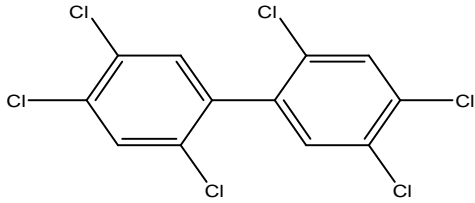
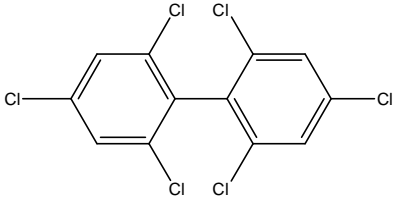
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 28</p> <p>7012-37-5</p>	 <p>Biphenyl, 2,4,4'-trichloro-</p>
<p>PCB 29</p> <p>15862-07-4</p>	 <p>Biphenyl, 2,4,5-trichloro-</p>
<p>PCB31</p> <p>16606-02-3</p>	 <p>Biphenyl, 2,4',5-trichloro-</p>
<p>PCB 40</p> <p>38444-93-8</p>	 <p>Biphenyl, 2,2',3,3'-tetrachloro-</p>
<p>PCB 44</p> <p>41464-39-5</p>	 <p>Biphenyl, 2,2',3,5'-tetrachloro-</p>

Composé	Structure et Nomenclature UAPAC
<p>PCB 47 2437-79-8</p>	 <p>Biphenyl, 2,2',4,4'-tetrachloro-</p>
<p>PCB 48 70362-47-9</p>	 <p>Biphenyl, 2,2',4,5-tetrachloro-</p>
<p>PCB 49 41464-40-8</p>	 <p>Biphenyl, 2,2',4,5'-tetrachloro-</p>
<p>PCB 52 35693-99-3</p>	 <p>Biphenyl, 2,2',5,5'-tetrachloro-</p>
<p>PCB 54 15968-05-5</p>	 <p>Biphenyl, 2,2',6,6'-tetrachloro-</p>

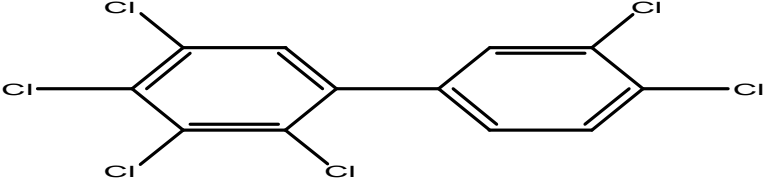
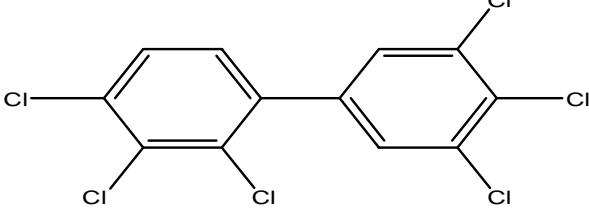
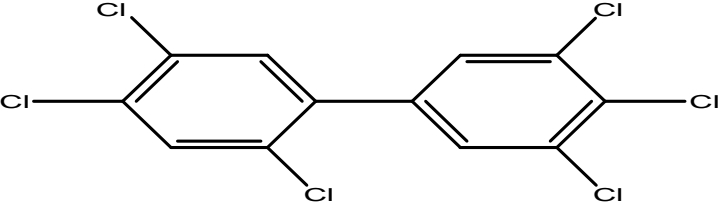
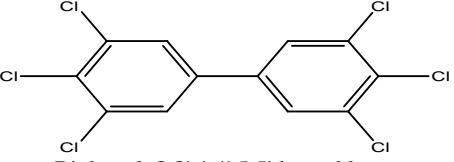
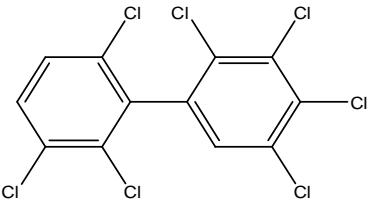
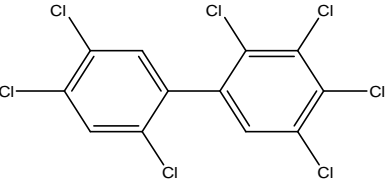
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 64 52663-58-8</p>	 <p>Biphenyl, 2,3,4',6-tetrachloro-</p>
<p>PCB 70 32598-11-1</p>	 <p>Biphenyl, 2,3,4',5-tetrachloro-</p>
<p>PCB 77 32598-13-3</p>	 <p>Biphenyl, 3,3',4,4'-tetrachloro-</p>
<p>PCB 87 38380-02-8</p>	 <p>Biphenyl, 2,2',3,4,5'-pentachloro-</p>
<p>PCB 90 68194-07-0</p>	 <p>Biphenyl, 2,2',3,4',5-pentachloro-</p>

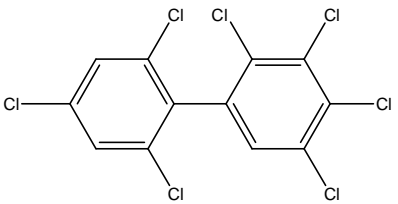
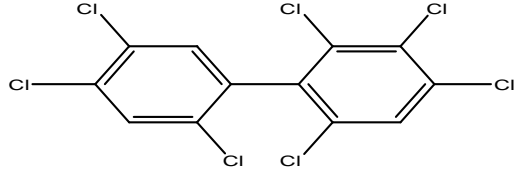
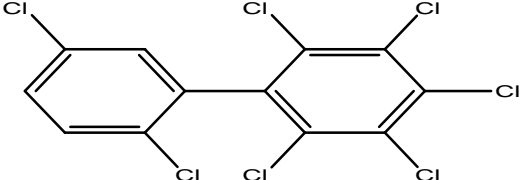
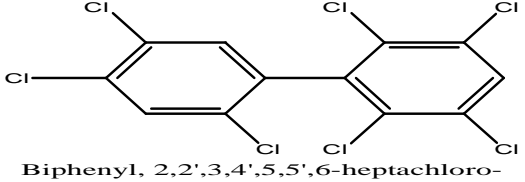
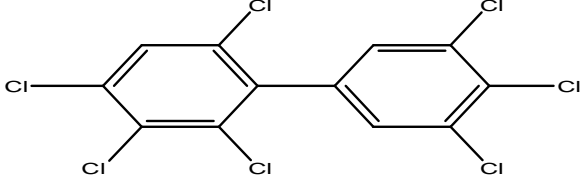
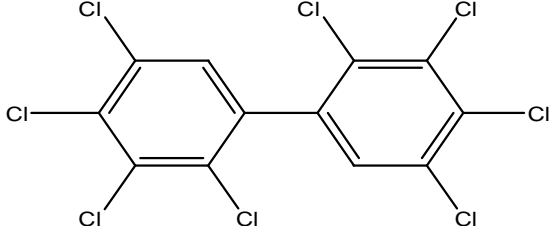
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 97 41464-51-1</p>	 <p>Biphenyl, 2,2',3',4,5-pentachloro-</p>
<p>PCB 99 38380-01-7</p>	 <p>Biphenyl, 2,2',4,4',5-pentachloro-</p>
<p>PCB 101 37680-73-2</p>	 <p>Biphenyl, 2,2',4,5,5'-pentachloro-</p>
<p>PCB 105 32598-14-4</p>	 <p>Biphenyl, 2,3,3',4,4'-pentachloro-</p>
<p>PCB 109 74472-35-8</p>	 <p>Biphenyl, 2,3,3',4,6-pentachloro-</p>
<p>PCB 118 31508-00-6</p>	 <p>Biphenyl, 2,3',4,4',5-pentachloro-</p>

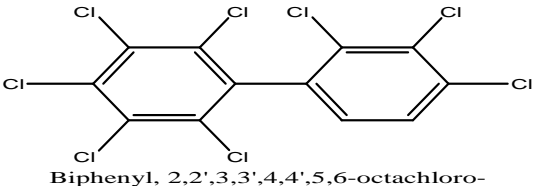
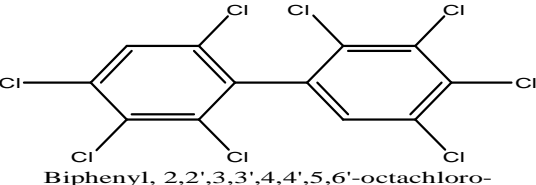
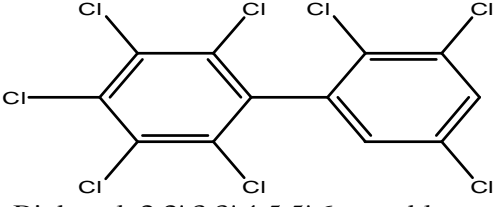
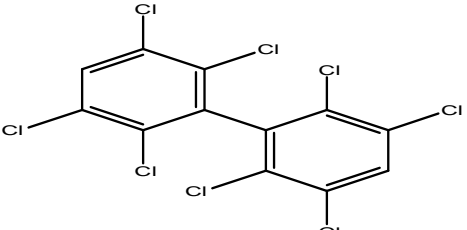
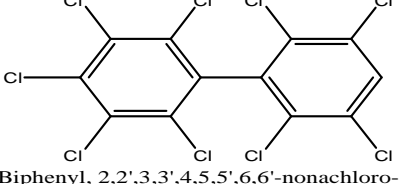
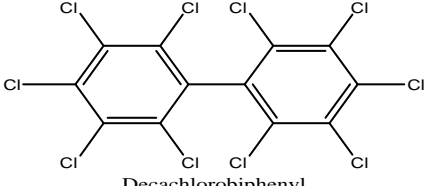
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 126 57465-28-8</p>	 <p>Biphenyl, 3,3',4,4',5-pentachloro-</p>
<p>PCB 128 38380-07-3</p>	 <p>Biphenyl, 2,2',3,3',4,4'-hexachloro-</p>
<p>PCB 136 38411-22-2</p>	 <p>Biphenyl, 2,2',3,3',6,6'-hexachloro-</p>
<p>PCB 137 35694-06-5</p>	 <p>Biphenyl, 2,2',3,4,4',5-hexachloro-</p>
<p>PCB 138 35065-28-2</p>	 <p>Biphenyl, 2,2',3,4,4',5'-hexachloro-</p>
<p>PCB 141 52712-04-6</p>	 <p>Biphenyl, 2,2',3,4,5,5'-hexachloro-</p>

Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 148 74472-41-6</p>	 <p>Biphenyl, 2,2',3,4',5,6'-hexachloro-</p>
<p>PCB 149 38380-04-0</p>	 <p>2,2',3,4',5,6-Hexachlorobiphenyl</p>
<p>PCB 151 52663-63-5</p>	 <p>Biphenyl, 2,2',3,5,5',6-hexachloro-</p>
<p>PCB 153 35065-27-1</p>	 <p>Biphenyl, 2,2',4,4',5,5'-hexachloro-</p>
<p>PCB 155 33979-03-2</p>	 <p>Biphenyl, 2,2',4,4',6,6'-hexachloro-</p>



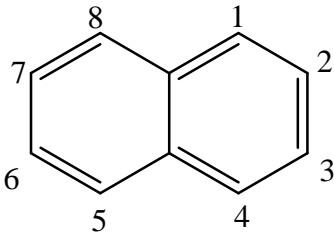
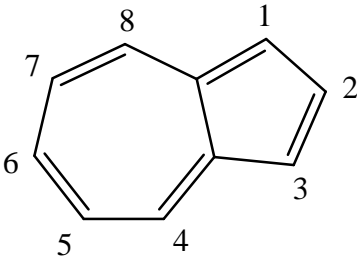
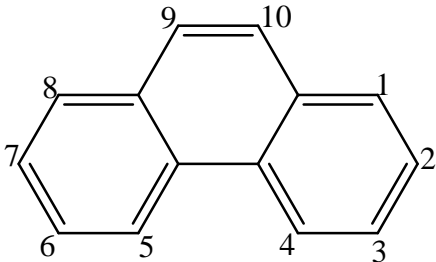
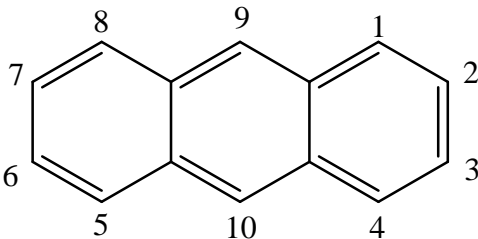
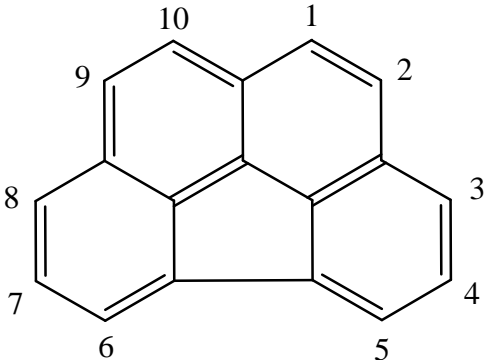
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 156 38380-08-4</p>	 <p>Biphenyl, 2,3,3',4,4',5-hexachloro-</p>
<p>PCB 157 69782-90-7</p>	 <p>Biphenyl, 2,3,3',4,4',5'-hexachloro-</p>
<p>PCB 167 52663-72-6</p>	 <p>2,3,4,4',5,5'-Hexachlorobiphenyl</p>
<p>PCB 169 32774-16-6</p>	 <p>Biphenyl, 3,3',4,4',5,5'-hexachloro-</p>
<p>PCB 174 38411-25-5</p>	 <p>Biphenyl, 2,2',3,3',4,5,6'-heptachloro-</p>
<p>PCB 180 35065-29-3</p>	 <p>Biphenyl, 2,2',3,4,4',5,5'-heptachloro-</p>

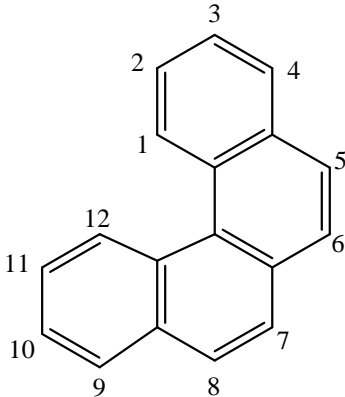
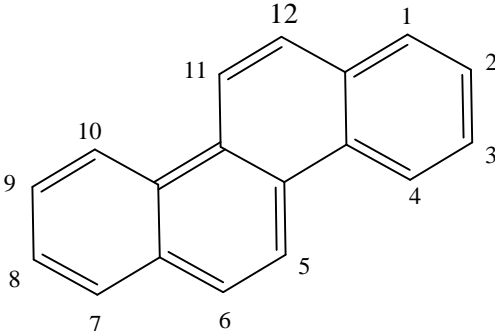
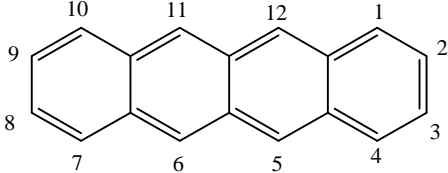
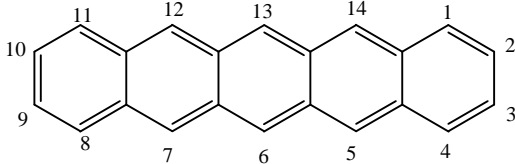
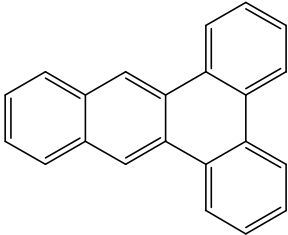
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 182 60145-23-5</p>	 <p>Biphenyl, 2,2',3,4,4',5,6'-heptachloro-</p>
<p>PCB 183 52663-69-1</p>	 <p>Biphenyl, 2,2',3,4,4',5',6'-heptachloro-</p>
<p>PCB 185 52712-05-7</p>	 <p>2,2',3,4,5,5',6'-Heptachlorobiphenyl</p>
<p>PCB 187 52663-68-0</p>	 <p>Biphenyl, 2,2',3,4',5,5',6'-heptachloro-</p>
<p>PCB 191 74472-50-7</p>	 <p>Biphenyl, 2,3,3',4,4',5',6'-heptachloro-</p>
<p>PCB 194 35694-08-7</p>	 <p>Biphenyl, 2,2',3,3',4,4',5,5'-octachloro-</p>

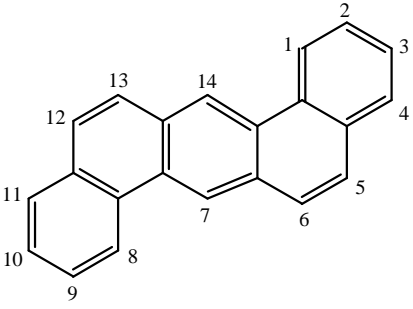
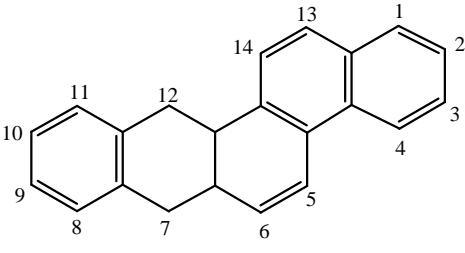
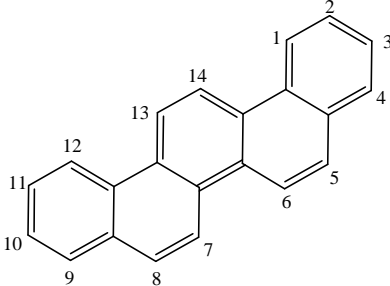
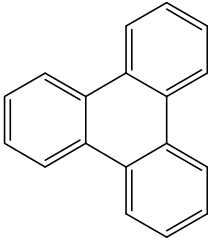
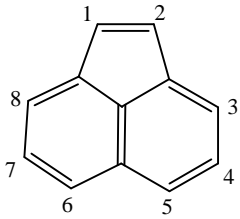
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>PCB 195 52663-78-2</p>	 <p>Biphenyl, 2,2',3,3',4,4',5,6-octachloro-</p>
<p>PCB 196 42740-50-1</p>	 <p>Biphenyl, 2,2',3,3',4,4',5,6'-octachloro-</p>
<p>PCB 198 68194-17-2</p>	 <p>Biphenyl, 2,2',3,3',4,5,5',6-octachloro-</p>
<p>PCB 202 2136-99-4</p>	 <p>Biphenyl, 2,2',3,3',5,5',6,6'-octachloro-</p>
<p>PCB 208 52663-77-1</p>	 <p>Biphenyl, 2,2',3,3',4,5,5',6,6'-nonachloro-</p>
<p>PCB 209 2051-24-3</p>	 <p>Decachlorobiphenyl</p>

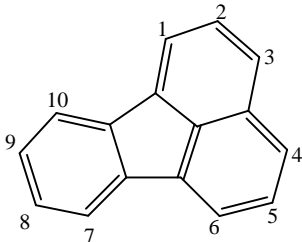
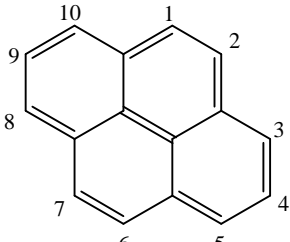
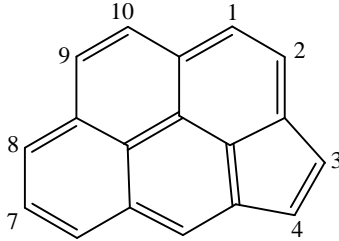
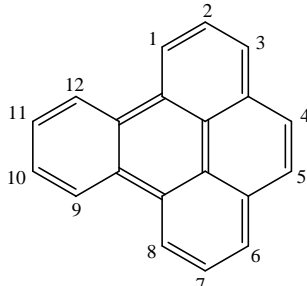
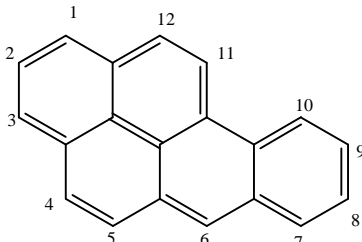


**ANNEXE  
DES HAPs**

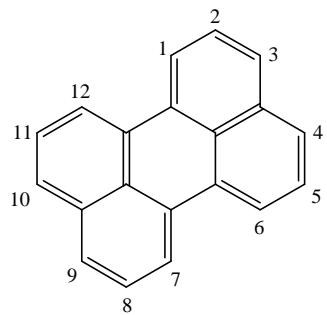
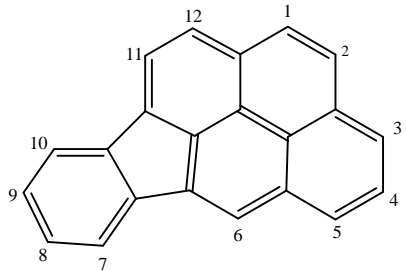
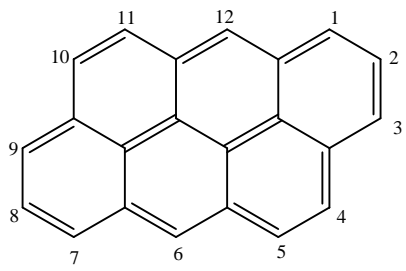
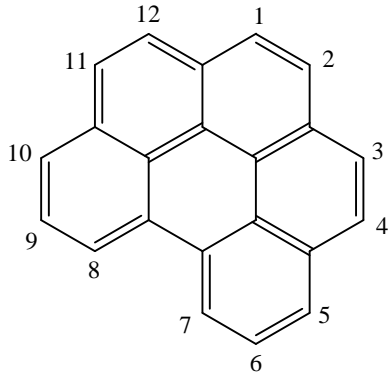
Composé et N° de cas	Structure et Nomenclature UAPAC
<p>Naphtalène</p> <p>91-20-3</p>	 <p>Naphthalene</p>
<p>Azuléne</p> <p>275-51-4</p>	 <p>Azulene</p>
<p>Phenanthréne</p> <p>85-01-8</p>	 <p>Phenanthrene</p>
<p>Anthracéne</p> <p>120-12-7</p>	 <p>Anthracene</p>
<p>Benzo[ghi]fluoranténe</p> <p>203-12-3</p>	 <p>Benzo[ghi]fluoranthene</p>

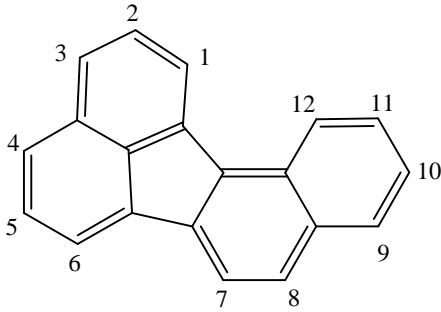
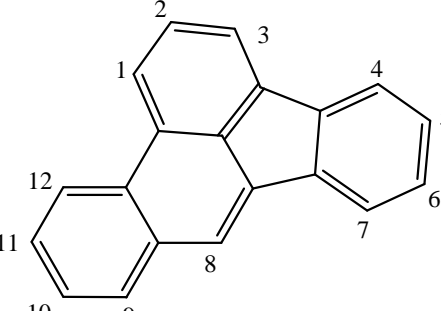
Composé et N° de cas	Structure et Nomenclature UAPAC
<p data-bbox="277 389 571 423">Benzo[c]phenanthréne</p> <p data-bbox="363 501 485 535">195-19-7</p>	 <p data-bbox="871 651 1114 680">Benzo[c]phenanthrene</p>
<p data-bbox="360 792 488 826">Chrysène</p> <p data-bbox="363 904 485 938">218-01-9</p>	 <p data-bbox="946 1064 1038 1093">Chrysene</p>
<p data-bbox="347 1167 504 1200">Naphtacéne</p> <p data-bbox="370 1279 481 1312">92-24-0</p>	 <p data-bbox="932 1355 1054 1384">Naphthacene</p>
<p data-bbox="357 1458 494 1491">Pentacéne</p> <p data-bbox="363 1570 488 1603">135-48-8</p>	 <p data-bbox="948 1668 1038 1697">Pentacene</p>
<p data-bbox="284 1765 564 1798">Dibenzo [a,c]anthracéne</p> <p data-bbox="363 1877 485 1910">215-58-7</p>	 <p data-bbox="900 2040 1086 2069">Dibenzo[a,c] anthracene</p>

Composé et N° de cas	Structure et Nomenclature UAPAC
<p>Dibenzo[a,h]anthracène</p> <p>53-70-3</p>	 <p>Dibenzo[a,h]anthracene</p>
<p>Benzo[b] chrysène</p> <p>214-17-5</p>	 <p>Benzo[B] Chrysene</p>
<p>Picéne</p> <p>213-46-7</p>	 <p>Picene</p>
<p>Triphenylène</p> <p>217-59-4</p>	 <p>Triphenylene</p>
<p>Acenaphthelène</p> <p>208-96-8</p>	 <p>Acenaphthylene</p>

Composé et N° de cas	Structure et Nomenclature UAPAC
Fluoranthène 206-44-0	 <p style="text-align: center;">Fluoranthene</p>
Pyrène 129-00-0	 <p style="text-align: center;">Pyrene</p>
Cyclopenta[cd]pyrène 27208-37-3	 <p style="text-align: center;">Cyclopenta[cd]pyrene</p>
Benzo[e]pyrène 192-97-2	 <p style="text-align: center;">Benzo[e]pyrene</p>
Benzo[a]pyrène 50-32-8	 <p style="text-align: center;">Benzo[A] Pyrene</p>



Composé et N° de cas	Structure et Nomenclature UAPAC
<p>Perylène</p> <p>198-55-0</p>	 <p>Perylene</p>
<p>Indeno[1,2,3-cd]pyrène</p> <p>193-39-5</p>	 <p>Indeno[1,2,3-cd]pyrène</p>
<p>Dibenzo[def,mno]chrysène</p> <p>191-26-4</p>	 <p>Dibenzo[def,mno]chrysene</p>
<p>Benzo[ghi]perylène</p> <p>191-24-2</p>	 <p>Benzo[ghi]perylene</p>

Composé et N° de cas	Structure et Nomenclature UAPAC
<p data-bbox="293 315 555 349">Benzo[j]fluorantène</p> <p data-bbox="363 394 485 427">205-82-3</p>	 <p data-bbox="884 562 1094 595">Benzo[j]fluoranthene</p>
<p data-bbox="293 846 555 880">Benzo[b]fluorantène</p> <p data-bbox="363 958 485 992">205-99-2</p>	 <p data-bbox="903 1084 1139 1117">Benzo[B] Fluoranthene</p>



## Management of Environmental Quality: An International

Chemometric modeling to predict retention times for a large set of pesticides or toxicants using hybrid genetic algorithm/multiple linear regression approach

Khadija Amirat Nadia Ziani Djelloul Messadi

### Article information:

To cite this document:

Khadija Amirat Nadia Ziani Djelloul Messadi , (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

Permanent link to this document:

<http://dx.doi.org/10.1108/MEQ-05-2015-0080>

Downloaded on: 05 April 2016, At: 03:46 (PT)

References: this document contains references to 20 other documents.

To copy this document: [permissions@emeraldinsight.com](mailto:permissions@emeraldinsight.com)

The fulltext of this document has been downloaded 10 times since 2016\*

### Users who downloaded this article also downloaded:

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

(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

Token:JournalAuthor:644DA30D-11D7-475D-BEBB-FC5A86A9D2EB:

### 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](http://www.emeraldinsight.com/authors) for more information.

### About Emerald [www.emeraldinsight.com](http://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 retention times for a large set of pesticides or toxicants using hybrid genetic algorithm/multiple linear regression approach

Hybrid  
GA/MLR  
approach

313

Received 11 May 2015  
Revised 20 July 2015  
1 September 2015  
Accepted 12 October 2015

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

## Abstract

**Purpose** – The purpose of this paper is to predict the retention times of 84 pesticides or toxicants.

**Design/methodology/approach** – Quantitative structure – retention relationship analysis was performed on a set of 84 pesticides or toxicants using a hybrid approach genetic algorithm/multiple linear regression (GA/MLR).

**Findings** – A model with six descriptors was developed using as independent variables. Theoretical descriptors derived from Spartan and Dragon softwares when applying GA/MLR approach.

**Originality/value** – A six parameter linear model developed by GA/MLR, with  $R^2$  of 90.54,  $Q^2$  of 88.15 and S of 0.0381 in Log value. Several validation techniques, including leave-many-out cross-validation, randomization test, 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 retention times of pesticides or toxicants.

**Keywords** Hybrid GA/MLR model, Molecular descriptors, Pesticides or toxicants, QSRR, Retention times

**Paper type** Research paper

## 1. Introduction

Recently computational methods have been used to solve complex problems in many aspects of science. One particularly useful method of the development of quantitative structure-activity relationships (QSARs) has found application in environmental chemistry and ecotoxicology (Deweese and Schultz, 2001; Leblond *et al.*, 2000; Cotescu and Diudea, 2006; Li *et al.*, 2009; Lu *et al.*, 2008). QSAR approach systematization which has to be associated to the work of Hansch and Fujita (1964) is based on the assumption that the structure of a molecule must contain the features responsible for its physical, chemical and biological properties and on the possibility of representing a molecule by numerical descriptors. The underlying hypothesis for QSAR models is that all molecules interact with the receptor in same or similar mode of action (Nendza and Wenze, 2006). The descriptors most used in the early QSAR analyses are the octanol/water partition coefficient ( $\text{Log}P$ ), the Hammett  $\sigma$  constant (Hammett, 1937; Hammett, 1940), acting as an electronic effect descriptor and the lipophilicity parameter  $\pi$ , which is defined by analogy to the electronic descriptor. Together with these empirical descriptors, the classical models employ other physical-chemical properties as parameters; some of them derived from quantum-chemical calculations, namely: partial charges, HOMO/LUMO energies, etc.



Pesticides are necessary and essential in agricultural production (General Inspectorate for Health Protection, 1996; Chun and Kang, 2003). The detection of pesticides or toxicants residues in agricultural production is one of the most important questions in our daily life. Multi-residues detect method is difficult to develop, due to the fact that compounds of different polarities, solubilities, volatilities and  $pK_a$  values have to be simultaneously extracted and analyzed. Now, most of pesticides or toxicants residues were detected by gas chromatography with selective and sensitive detectors (ECD, NPD, FPD, AED or MSD). Mass spectrometry is a very selective and sensitive technique for both multi-residues determination and trace-level identification of a wide range of pesticides (Stajnbaher and Zupancic-Kralj, 2003).

## 2. Materials and methods

### 2.1 Data set

In this work a set of 84 pesticides or toxicants were studied. The logarithm values of retention times used here (Table I) were taken from Chinese Academy of Inspection and Quarantine. These logarithm values span between 1.275 and 1.701.

### 2.2 Descriptor generation

To develop a quantitative structure/retention relationship (QSRR) model the main ingredients are the molecular descriptors, a final result of a logical and mathematical procedure with a transformation of a chemical structure encoded inside a symbolic representation in a molecule into a number. The chemical structure of each compound was sketched using the Spartan 10 software (Spartan software, 2011). The final geometry of the minimum energy conformation was obtained by the semi-empirical method PM6. So the final geometries were used as inputs for the generation of more than 1,600 descriptors of different kinds including constitutional, topological, 2D auto correlation, connectivity indices, geometrical, RDF, 2DMoRSE, WHIM, GETAWAY, using Dragon software (version 5.3) (Todeschini *et al.*, 2006). Frontier orbital energies, dipolar moment, bonding energy "En" and volume were calculated using the Spartan software. Using the corresponding options in Dragon software, we first eliminated the descriptors that provide no information (standard deviations less than 0.0001); then we removed high correlated descriptors ( $r > 0.98$ ), where the variable with the most high cross-correlations with other descriptors was deleted.

### 2.3 Data splitting

The whole data set ( $n = 84$ ) was divided into training and prediction set in order to develop and validate the model for its external predictive capacity. The splitting was based on Kennard and Stone (1969) algorithm.

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 (in general by Euclidian metric) and that the no validation samples fall outside the measured region. Using Kennard and Stone algorithm the entire set was divided into two subsets: a training set of 67 compounds ( $n_{tr} = 67$ ), and a test set including the remaining 17 compounds ( $n_{pr} = 17$ ).

Chemical	Log tr	En	nR06	ATS1m	ATS7v	GATS2e	EEig05d
a-666	1.275	-258.84	1	3.207	0	1.222	1.904
Chlorbufan	1.313	-99.54	1	2.959	2.197	0.733	1.278
Atrazine	1.323	78.72	1	2.924	1.386	0.305	1.226
Trietaezine	1.326	238.59	1	2.997	1.609	0.936	1.229
Fonofos	1.337	-434.06	1	3.49	1.609	1.212	1
PCB15	1.344	111.06	2	2.991	1.792	0.758	1
Carbofuran	1.36	-463.49	1	2.995	1.858	0.911	1.687
4,4'-DDM	1.39	88.01	2	3.04	2.197	0.769	1
PCB31	1.399	75.77	2	3.129	1.792	0.795	1.234
Benoxacor	1.401	-276.99	2	3.153	1.099	0.847	2.097
Phosphamidon	1.43	-1,235.2	0	3.451	2.18	0.516	1.855
Benfuresate	1.432	-661.33	1	3.379	2.313	0.772	2.013
Aldrin	1.434	110.23	2	3.518	0	0.763	2.287
PCB52	1.443	41.37	2	3.251	1.609	0.833	1.904
Metolachlor	1.464	-415.16	1	3.155	2.2	0.637	1.987
Trichoronate	1.472	-678	1	3.607	2.361	1.173	1.809
Methoprene	1.476	-733.24	0	3.164	2.743	1.051	2.345
Chlorpyriphos	1.476	-848.13	1	3.673	2.574	0.91	1.812
Thiobencarb	1.478	-209.2	1	3.141	2.417	0.668	2.153
Methiocarb	1.483	-333.36	1	3.029	1.843	0.512	2.147
Isodrin	1.485	314.67	2	3.518	0	0.763	2.299
Fenthion	1.501	-770.14	1	3.542	2.138	0.548	1.88
Allethrin	1.503	-463.44	0	3.232	2.972	0.76	2.604
Isocarbophos	1.52	-958.97	1	3.509	2.27	0.826	1.586
Isofenphos	1.523	-1,069.49	1	3.627	3.03	0.957	2.181
Triadimenol	1.533	-133.17	1	3.264	2.727	0.97	1.939
Procymedone	1.544	-265.24	2	3.261	2.303	0.64	2.623
Quinalphos	1.549	-639.59	2	3.579	2.624	0.8	1.618
Alpha-endosulfan	1.549	-538.27	1	3.739	1.114	0.776	2.719
Phenthoate	1.551	-953.31	1	3.688	2.655	0.628	1.811
Chlorbenside	1.557	112.67	2	3.229	2.398	0.825	2.075
Crotoxyphos	1.561	-1,255.53	1	3.519	2.601	0.492	1.939
Prothiofos	1.572	-718.96	1	3.735	2.687	1.091	1.977
Tetrachlorvinphos	1.576	-933.67	1	3.652	2.506	0.549	2.149
Chinomethionate (oxythioquinox)	1.577	120.53	2	3.246	1.414	0.671	2.286
PCB 87	1.581	31.71	2	3.359	2.079	0.872	2.155
4,4'-DDE	1.582	112.9	2	3.325	2.565	0.63	2.067
Methidathion	1.587	-837.39	0	3.718	2.136	0.613	2.134
Buprofezin	1.59	-80.05	2	3.332	2.594	0.525	2.522
Fenamiphos	1.59	-844.63	1	3.511	2.643	0.58	2.17
2,4'-DDD	1.6	63.83	2	3.325	2.303	0.63	1.848
PCB149	1.608	3	2	3.457	2.197	0.911	2.268
Endrin	1.612	47.94	3	3.594	0	0.785	2.728
2,4'-DDT	1.627	17.68	2	3.426	2.398	0.59	1.854
4,4'-DDT	1.652	10.03	2	3.426	2.708	0.59	2.177
Benalaxyl	1.652	-430.6	2	3.326	3.335	0.552	2.255
PCB187	1.653	-17.33	2	3.546	2.398	0.95	2.484
Resmethrin	1.658	-348.91	1	3.39	3.023	0.758	2.603
PCB167	1.661	7.27	2	3.457	2.485	0.911	2.373
Bifenthrin	1.663	-859.34	2	3.603	3.231	0.475	2.612

(continued)

**Table I.**  
Values of En, nR06,  
ATS1m, ATS7v,  
GATS2e, EEig05d  
and Log tr for a set  
of 84 pesticides or  
toxicants

Chemical	Log tr	En	nR06	ATS1m	ATS7v	GATS2e	EEig05d
Famphur	1.664	-1,113.8	1	3.763	2.39	0.572	1.734
Edifenphos	1.664	-453.02	2	3.721	2.833	0.68	1.774
PCB202	1.665	-33.33	2	3.627	2.303	0.99	2.633
Bromopropylate	1.67	-377.13	2	3.601	3.151	0.919	2.413
Fenpropathrin	1.673	-181.69	2	3.428	3.36	0.953	2.947
Phenothrin	1.678	-322.16	2	3.423	3.191	0.891	2.603
Dicofol	1.678	-158.88	2	3.469	2.708	0.75	2.177
Tetramethrin	1.679	-669.63	1	3.373	3.033	0.481	2.665
Pyridaphenthion	1.68	-752.41	2	3.673	2.866	0.793	1.781
Leptophos	1.687	-435.46	2	3.797	2.923	0.96	2.069
Imidan	1.687	-841.72	1	3.717	2.613	0.536	2.371
Tertradifon	1.688	-230.03	2	3.618	2.485	0.916	2.515
Pyrazophos ethyl	1.695	-980.22	1	3.747	2.922	0.83	2.113
Fenarimol	1.697	137.13	3	3.398	2.878	0.856	1.92
Permethrin	1.7	-305.86	2	3.543	3.191	0.804	2.722
Azinphos-methyl	1.7	-517.14	2	3.726	2.572	0.623	1.921
PCB194	1.701	-16.85	2	3.627	2.773	0.99	2.641
Lindane	1.34	-258.84	1	3.207	0	1.222	1.904
Fenclorphos	1.429	-810.07	1	3.612	1.962	0.79	1.78
Pentanochlor	1.451	-273.01	1	2.976	2.352	0.751	1.878
Bromophos-methyl	1.504	-757.78	1	3.707	2.07	0.758	1.769
PCB 70	1.522	51.64	2	3.251	2.079	0.833	2.124
PCB101	1.545	19.14	2	3.359	2.079	0.872	2.214
2,4'-DDE	1.545	119.53	2	3.325	2.303	0.63	1.756
PCB118	1.614	29.77	2	3.359	2.303	0.872	2.15
PCB153	1.627	-2.81	2	3.457	2.398	0.911	2.262
Beta-endosulfan	1.637	-491.71	1	3.739	1.114	0.776	2.719
4,4'-DDD	1.637	29.37	2	3.325	2.565	0.63	2.177
PCB141	1.639	10.66	2	3.457	2.303	0.911	2.478
Sulprophos	1.647	-659.2	1	3.721	2.647	1.113	2.389
PCB138	1.65	9.71	2	3.457	2.398	0.911	2.256
PCB185	1.661	-3.25	2	3.546	2.303	0.95	2.713
PCB128	1.666	22.13	2	3.457	2.398	0.911	2.243
PCB180	1.674	-11.09	2	3.546	2.565	0.95	2.487

**Table I.** Note: The last 17 chemicals are the test set

#### 2.4 Model development and validation

Multiple linear regression analysis and variable subset selection were performed by package MobyDigs for windows/PC (Todeschini *et al.*, 2009), using ordinary least squares regression method and, as previously indicated, genetic algorithm/variable subset selection (GA-VSS). The goodness of fit of the calculated model was assessed by means of the multiple determination coefficient  $R^2$  and the standard deviation error in calculation (SDEC) defined as:

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

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

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

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

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

The 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):

$$\sigma_n \equiv SDEP = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2} = \sqrt{\frac{PRESS}{n}} \quad (3)$$

A value  $Q^2 > 0.5$  is generally regarded as a good result and  $Q^2 > 0.9$  as excellent (Eriksson *et al.*, 2003; Tropsha *et al.*, 2003).

However, studies have indicated that while  $Q^2$  is a necessary condition for high predictive power in a model, its alone is not sufficient. To avoid overestimating the predictive power of the model the leave-more out (LMO up to 50 percent of perturbation: LMO/50) procedure (repeated 8,000 times in this study) was also performed. In a typical LMO validation, objects of the data set are divided in  $G$  cancellation groups of equal size,  $m_i = (n/G)$ . Based on the value of  $n$ ,  $G$  is generally selected between 2 and 10. A large number of models are developed with each of the  $n-m_i$  objects in the training set and  $m_i$  objects in the validation set. For each corresponding model  $m_i$  objects are predicted and  $Q_{LMO}^2$  computed (as average value of the number of validation runs).

In order to evidence the existence of fortuitous correlations, the randomization test (Y-scrambling) (Wold and Eriksson, 1995) was adopted. This test consists of building a property vector whose components are the components of the actual property vector, but randomly permuted in their positions. This new activity vector is used as if it was really an experimental one, and a QSRR model is computed in the usual way. This process was repeated 100 times, in order to test the capacity factor of the model to extract actual structure/retention relationships.

By bootstrap validation technique, the original size of the data set ( $n$ ) is preserved for the training set, by the selection of  $n$  objects with repetition; in this way the training set usually consists of repeated objects and the evaluation set of the objects left out (Efron, 1994).

The model is calculated on the training set and responses are predicted on the evaluation set. All the squared differences between the true response and the predictive response of the objects of evaluation set are collected in PRESS. This procedure of building training sets and evaluation sets is repeated 8,000 times in this study, PRESS are summed and the average predictive power is calculated.



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

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

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

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

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

where the sum runs over the test set objects  $n_{ext}$ . The external  $R_{CV_{ext}}^2$  for the test set is determined by the following equation:

$$R_{CV_{ext}}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_{tra})^2} \quad (6)$$

where  $y_i$  and  $\hat{y}_i$  are the observed and the calculated response values, respectively; and  $\bar{y}_{tr}$  is the averaged value for the response variable of the training set; and the summation runs over the test set.

### 2.5 QSAR applicability domain (AD)

The AD was discussed by the Williams plot (Eriksson *et al.*, 2003; Tropsha *et al.*, 2003) of jackknifed residuals vs leverages (hat diagonal values  $h_i$ ). The jackknifed residuals (or studentized residuals) are the standardized cross-validated residuals. Each residual is divided by its standard deviation, which is calculated without the  $i$ -th observation. The leverage ( $h_i$ ) value of a chemical in the original variable space is defined as:

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

where  $x_i$  is the descriptor row-vector of the query compound, and  $X$  is the  $n$  ( $p+1$ ) matrix of  $p$  model parameter values for  $n$  training set compounds. The superscript  $T$  refers to the transpose of the matrix/vector. The warning leverage value ( $h^*$ ) is defined as  $3(p+1)/n$ . When  $h$  value of a compound is lower than  $h^*$ , the probability of accordance between predicted and actual values is as high as that for the compounds in the training set. A chemical with  $h_i > h^*$  will reinforce the model if the chemical is in the training set. But such a chemical in the validation set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residual may be low. Thus the leverage and the jackknifed residual should be combined for the characterization of the AD.

### 3. Results and discussion

Application of the GA-VSS led to several good models for the prediction of  $\text{tr}_s$  based on different sets of molecular descriptors. The best six dimensional model was constructed using descriptors: En, nR06, ATS1m, ATS7v, GATS2e and EEig05d descriptors. All data concerning value of descriptors and the logarithm of retention times are summarized in Table I.

The equation of the optimal model can be written as:

$$\begin{aligned} \text{Log tr} = & 0.32095 \pm (0.08595) + 0.00006735 \pm (0.00002276) \text{ En} \\ & + 0.03612 \pm (0.01117) \text{ nR06} + 0.27856 \pm (0.03128) \text{ ATS1m} \\ & + 0.070972 \pm (0.006836) \text{ ATS7v} - 0.11332 \pm (0.02614) \text{ GATS2e} \\ & + 0.08315 \pm (0.01173) \text{ EEig05d} \end{aligned} \quad (8)$$

Here En (bonding energy) is a descriptor calculated with Spartan software, nR06, ATS1m, ATS7v, GATS2e, EEig05d were calculated in Dragon software. More information about these descriptors can be found in Dragon software user's guide (Todeschini *et al.*, 2006) and the references therein. All relevant statistical parameters of the proposed model are reported hereafter (Table II).

The correlation matrix as shown in Table III suggests that these descriptors are weakly correlated with each other. Thus; the model can be regarded as an optimal regression equation. Values of  $R^2$  and  $R^2_{\text{adj}}$  attest the good fitting performances of the model which, moreover, is very highly significant (great value of the Fisher parameter  $F$ ).

The model is robust, the difference between  $R^2$  and  $Q^2$  is small (2.4 percent). The model demonstrates a very good stability in internal validation (difference between  $Q^2_{\text{LOO}}$  and  $Q^2_{\text{LMO}}$  is about 2.7 percent), while bootstrapping confirms the internal predictivity and stability of the model.  $\text{SDEP}_{\text{ext}}$  is a little bit different from SDEP. The model was also verified by Y-scrambling. Figure 2 clearly ensures the existence of a linear relationship between Log tr and the descriptors: En, nR06, ATS1m, ATS7v, GATS2e, EEig05d. As can be observed the permuted responses yield poor predictive

$n_{tr}$	$n_{ext}$	$Q^2_{\text{LOO}}$	$R^2$	$Q^2_{\text{LMO}/50}$	$Q^2_{\text{BOOT}}$	$R^2_{\text{adj}}$	$Q^2_{\text{ext}}$
67	17	88.15	90.54	86.45	86.58	89.59	87.15
SDEC	SDEP	$\text{SDEP}_{\text{ext}}$	S	F			
0.036	0.04	0.042	0.0381	95.6528			

**Table II.**  
Statistical  
parameters of a  
developed model

	En	nR06	ATS1m	ATS7v	GATS2e	EEig05d	Log tr
En	1.000						
nR06	0.614	1.000					
ATS1m	-0.501	0.063	1.000				
ATS7v	-0.369	-0.068	0.249	1.000			
GATS2e	0.169	0.067	0.086	-0.097	1.000		
EEig05d	-0.021	0.159	0.363	0.202	-0.018	1.000	
Log tr	-0.126	0.385	0.658	0.583	-0.134	0.639	1.000

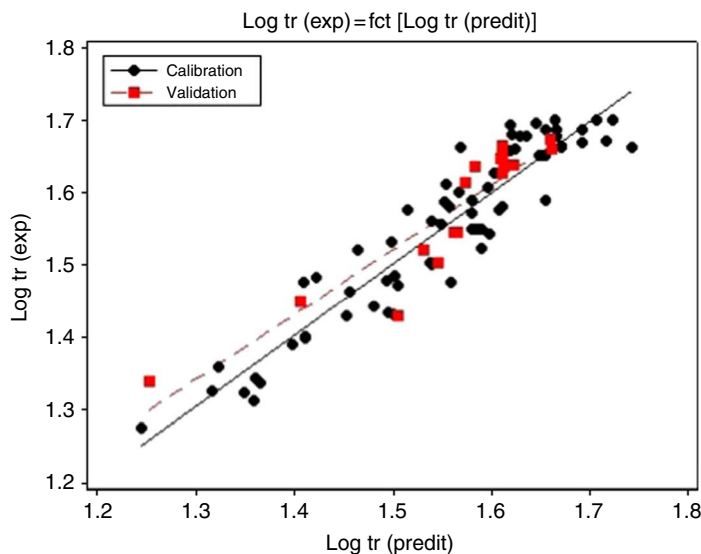
**Table III.**  
Correlation  
matrix between the  
selected descriptors  
and Log tr

models, all having  $Q^2 < 0.2$  percent. On the other hand, the correctly ordered Log tr yield good statistical parameters, and therefore it is located isolated in the plot. The relative contributions of the six descriptors to the model were determined and are plotted in Figure 4. Six descriptors were needed in the QSRR model. The significance of the descriptors involved in the model decreases in the following order: ATS7v (20.77 percent) > ATS1m (19.64 percent) > EEig05d (17.46 percent) > GATS2e (14.47 percent) > nR06 (14.06 percent) > En (13.6 percent). It should be noted that the difference in the descriptor contribution between any two descriptors used in the model is not significant, indicating that all of the descriptors are indispensable in generating the predictive model.

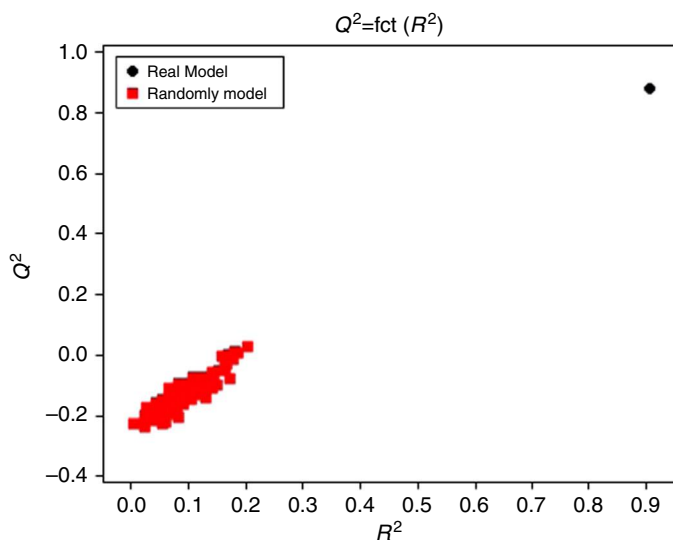
The positive coefficients of ATS1m, ATS7v, EEig05d, nR06, En (Equation 8) indicates that the pesticides or toxicants with larger values for these descriptors would have larger Log tr values. GATS2e has a smaller correlation coefficient with the experimental Log tr values ( $R = -0.134$ ). The negative sign of GATS2e in Equation (8) indicates that the pesticides or toxicants containing atoms with larger atomic Sanderson electronegativities would possess lower Log tr because this descriptor increases with increased atomic electronegativities (Table IV) (Figures 1-4).

**Table IV.**  
Characteristics of  
descriptors selected  
for the best model  
AG/MLR

Predictor	Coef	SE Coef	<i>T</i>	<i>P</i>	VIF
Constant	0.32095	0.08595	3.73	0.000	–
En	0.00006735	0.00002276	2.96	0.004	3.913
nR06	0.03612	0.01117	3.23	0.002	2.472
ATS1m	0.27856	0.03128	8.90	0.000	2.390
ATS7v	0.070972	0.006836	10.38	0.000	1.268
GATS2e	-0.11332	0.02614	-4.34	0.000	1.126
EEig05d	0.08315	0.01173	7.09	0.000	1.264

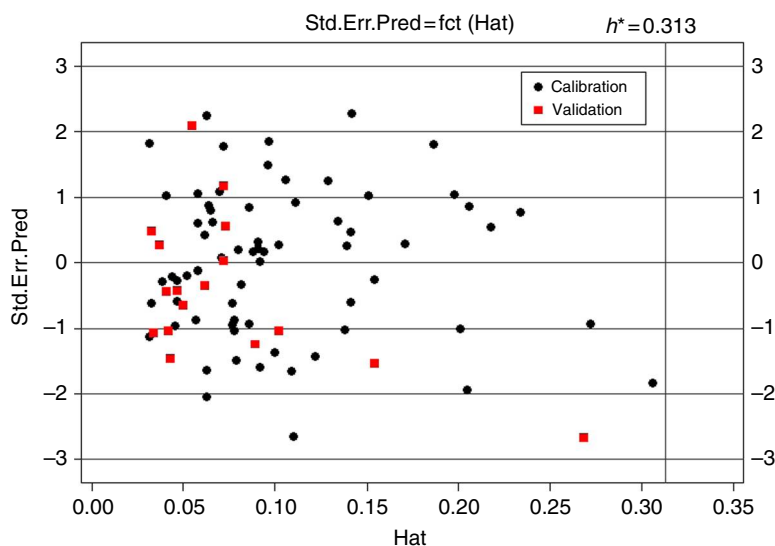


**Figure 1.**  
Experimental vs  
cross-validated  
Log tr



**Note:** Square represent the randomly ordered retentions and the circle corresponds to the real retentions

**Figure 2.**  
Randomization test  
associated to the  
previous QSRR  
model



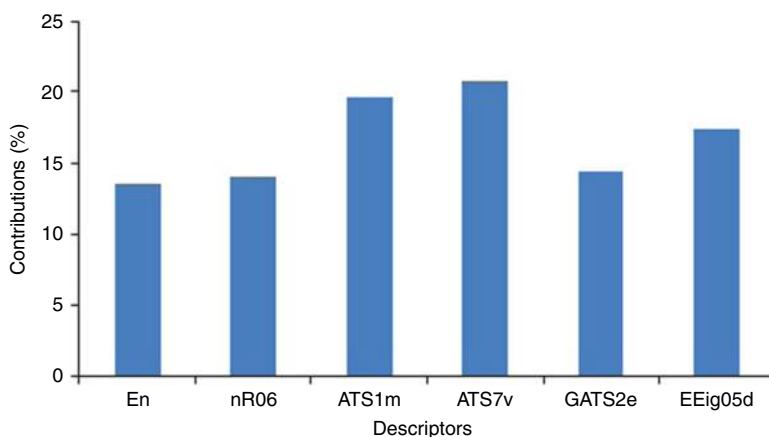
**Figure 3.**  
Williams plot of the  
current QSRR model

### 3.1 Mechanistic interpretation

The descriptors and their definitions and classes are gathered in the Table V.

The total energy (En) is the sum of the electronic and the nuclear energy in a molecule, this descriptor describes the system state and gives a precious information about the stability of the molecule, so molecules which have low energies are stables as

**Figure 4.**  
Relative contributions of the selected descriptors to the GA/MLR model



**Table V.**  
The descriptors and their definitions and classes

Descriptor	Meaning	Class
En	Total energy of the molecule	Quantum-chemical descriptor
nR06	Number of 6-membered rings	Constitutional descriptors
ATS1m	Broto-Moreau autocorrelation of a topological structure – lag 1/weighted by atomic masses	2D autocorrelation indices
ATS7v	Broto-Moreau autocorrelation of a topological structure – lag 7/weighted by atomic Van der Waals volumes	
GATS2e	Geary autocorrelation – lag 2/weighted by atomic Sanderson electronegativities	
EEig05d	Eigenvalue 05 from edge adj. matrix weighted by dipole moments	Edge adjacency indices

a result their chromatographic elution is difficult, and the molecules with high energies are active they go out easily.

nR06 is selected as “constitutional descriptor” that is dependent basically on the composition of molecule rather than on geometry and topology.

2D autocorrelations are molecular descriptors which describe how a considered property is distributed along a topological molecular structure.

ATS1m and ATS7v are Broto-Moreau autocorrelation descriptors which are calculated as:

$$ATS1m = \sum_{i=1}^{nSk-1} \sum_{j>1} m_i \cdot m_j \cdot \delta_{ij} \quad (9)$$

$$ATS7v = \sum_{i=1}^{nSk-1} \sum_{j>1} v_i \cdot v_j \cdot \delta_{ij} \quad (10)$$

where  $m$  and  $v$  are respectively the atomic masses and the atomic Van der Waals volumes which are used to weight the molecular graph and  $k$  the lag.  $nSK$  is the number of non-hydrogen atoms,  $\delta_{ij}$  is the Kronecker delta ( $\delta_{ij} = 1$  if  $d_{ij} = k$ , zero otherwise,  $d_{ij}$  being the topological distance between two considered atoms).  $\Delta$  is the sum of the Kronecker deltas, i.e. the number of atom pairs at distance equal to  $k$ .

GATS2e belongs to the Geary autocorrelations, it is defined by Equation (11), where  $e$  is the atomic Sanderson electronegativities,  $\bar{e}$  is its average value on the molecule:

$$\text{GATS } 2e^{\frac{\frac{1}{2\Delta} \cdot \sum_{i=1}^{nSk} \sum_{j=1}^{nSk} \delta_{ij} (e_i - e_j)^2}{\frac{1}{(nSk-1)} \cdot \sum_{i=1}^{nSk} \delta_{ij} (e_i - \bar{e})^2}} \quad (11)$$

These descriptors indicate the roles of atomic masses, atomic Van der Waals volumes, and atomic Sanderson electronegativities in retention mechanism.

Edge adjacency indices are molecular descriptors calculated from the edge adjacency matrix of a molecule. The edge adjacency matrix is derived from the H-depleted molecular graph and encodes the connectivity between graph edges. It is a square symmetric matrix of dimension  $B \times B$ , where  $B$  is the number of bonds between non-hydrogen atom pairs. The entries of the matrix equal one if the considered bonds are adjacent and zero otherwise. The weighting of the EEig05d by the dipole moment highlights the important role of specific interactions solute-stationary phase in the chromatographic retention.

### 3.2 Applicability domain

On analyzing the model applicability domain from Williams plot, all residuals were located within the range of three standard deviations, and there is no influential compound both for training or prediction set (Figure 3), which means that the model has a good external predictivity.

## 4. Conclusion

A QSRR model for the estimation of the logarithm of retention times of 84 pesticides or toxicants was established in the following six steps: molecular structure input and generation of the files containing the chemical structures; quantum mechanics geometry optimization with a semi-empirical method; structural descriptors computation; structural descriptors selection; structure model generation with a multivariate method and statistical analysis.

According to obtained results it is concluded that the En, nR06, ATS1m, ATS7v, GATS2e and EEig05d can be used successfully for modeling retention property (Log tr) of the under study compounds. High correlation coefficient (0.9054) and low prediction error (SDEP = 0.04; SDEP<sub>ext</sub> = 0.042 in Log unit) obtained confirm good predictive ability of the model. The QSRR model proposed with the simply calculated molecular descriptors can be used to estimate the chromatographic retention times for new compounds even in the absence of the standard candidates.

## References

- Chun, O.K. and Kang, H.G. (2003), "Estimation of risks of pesticide exposure by food intake to Koreans", *Food and Chemical Toxicology*, Vol. 41 No. 8, pp. 1063-1076.
- Cotescu, A. and Diudea, M.V. (2006), "QSTR study on aquatic toxicity against *Poecilia reticulata* and *Tetrahymena pyriformis* using topological indices", *Internet Electronic Journal of Molecular Design*, Vol. 5 No. 2, pp. 116-134.
- Deweese, A.D. and Schultz, T.W. (2001), "Structure-activity relationships for aquatic toxicity to tetrahymena: halogen substituted aliphatic esters", *Environmental Toxicology*, Vol. 16 No. 1, pp. 54-60.
- Efron, B. (1994), *The Jackknife, the Bootstrap and Other Resampling Planes*, Society for Industrial and Applied Mathematics, Philadelphia, PA.
- Eriksson, L., Jaworska, J., Worth, A., Cronin, M., Mc Dowell, R.M. and Gramatica, P. (2003), "Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs", *Environmental Health Perspectives*, Vol. 111 No. 10, pp. 1361-1375.
- General Inspectorate for Health Protection (1996), *Analytical Methods for Pesticide Residues in Foodstuffs*, 6th ed., Ministry of Health Welfare and Sport, Amsterdam.
- Hammett, L.P. (1937), "The effect of structures upon the reactions of organic compounds, benzene derivatives", *Journal of the American Chemical Society*, Vol. 59 No. 1, pp. 96-103.
- Hammett, L.P. (1940), *Physical Organic Chemistry*, McGraw Hill, New York, NY.
- Hansch, C. and Fujita, T. (1964), " $\rho$ - $\sigma$ - $\pi$  analysis: a method for the correlation of biological activity and chemical structure", *Journal of the American Chemical Society*, Vol. 86 No. 8, pp. 1616-1626.
- Kennard, R. and Stone, L.A. (1969), "Computer aided design of experiments", *Technometrics*, Vol. 11 No. 1, pp. 137-148.
- Leblond, J.D., Applegate, B.M., Menn, F.M., Schultz, T.W. and Saylor, G.S. (2000), "Structure-toxicity assessment of metabolites of the aerobic bacterial transformation of substituted naphthalenes", *Environmental Toxicology and Chemistry*, Vol. 19 No. 5, pp. 1235-1246.
- Li, F., Chen, J., Wang, Z., Li, J. and Qia, X. (2009), "Determination and prediction of xenoestrogens by recombinant yeast-based assay and QSAR", *Chemosphere*, Vol. 74 No. 9, pp. 1152-1157.
- Lu, G.H., Wang, C. and Guo, X.L. (2008), "Prediction of toxicity of phenols and anilines to algae by quantitative structure-activity relationship", *Biomedical and Environmental Sciences*, Vol. 21 No. 3, pp. 193-196.
- Nendza, M. and Wenzel, A. (2006), "Discriminating toxicant classes by mode of action-1.(Eco) toxicity profiles", *Environmental Science and Pollution Research*, Vol. 13 No. 3, pp. 192-203.
- Spartan software (2011), "Release for Window, Macintosh and Linux version 1.1.0", Molecular Modeling System, Pittsburg, CA.
- Stajnbaher, D. and Zupancic-Kralj, L. (2003), "Multiresidue method for determination of 90 pesticides in fresh fruits and vegetables using solid-phase extraction and gas chromatography mass spectrometry", *Journal of Chromatography A*, Vol. 1015 Nos 1-2, pp. 185-198.
- Todeschini, R., Consonni, V. and Pavan, M. (2006), "DRAGON Software for the Calculation of Molecular Descriptors", Release 5.4 for Windows, Talete s.r.l., Milano.
- Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. and Pavan, M. (2009), "MOBYDIGS, software for multilinear regression analysis and variable subset selection by genetic algorithm", Release 1.1 for Windows, Milano.

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 and Combinatorial Science*, Vol. 22 No. 1, pp. 69-76.

Wold, S. and Eriksson, L. (1995), *Chemometrics Methods in Molecular Design*, VCH Publishers, Weinheim.

#### About the authors

Khadidja Amirat Phd Student at the Badji Mokhtar University. Her interests are analytical chemistry, bromatology and chemiometry.

Nadia Ziani Phd Student at the Badji Mokhtar University. Her interests are analytical chemistry, bromatology and chemiometry.

Dr Djelloul Messadi Head of the Environmental and Food Safety Laboratory. Faculty of Sciences, the Badji Mokhtar University, Annaba, Algeria. Led supported several doctorates. He is the Co-author of several papers published in international scientific journals. His interests are analytical chemistry, bromatology and chemiometry. Dr Djelloul Messadi is the corresponding author and can be contacted at: [d\\_messadi@yahoo.fr](mailto:d_messadi@yahoo.fr)



# Modeling of the Retention indices of a set of polycyclic aromatic hydrocarbons using a hybrid approach

Khadija Amirat<sup>1</sup>; Fatiha Mebarqi<sup>2</sup>; Nadia Ziani<sup>3</sup>; Djelloul Messadi<sup>4</sup>

Environmental and food Security laboratory, Badji Mokhtar University, BP 12, 23000, Annaba, Algeria

1: khadija\_amirat@yahoo.fr

2: fatiha\_mebarki@yahoo.fr

3: Ziani\_nadia84@yahoo.fr

4: D\_messadi@yahoo.fr

## Abstract

A structure/retention indices relationship was searched for 93 PAHs while promoting the hybrid genetic algorithm/multilinear regression approach, the structural parameters being calculated with the software Spartan and DRAGON. Among about a hundred of 2 regressor models gotten, we selected the one that present best values of the prediction parameter ( $Q^2$ ) and of the determination coefficient ( $R^2$ ). The reliability of the proposed model was further illustrated using various evaluation techniques: leave-many-out, cross-validation procedure, randomization test, and validation through the test set.

**Keywords:** structure/ retention indices; PAHs; software; molecular descriptors.

## 1. INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) are important classes of organic compounds, which usually have two to six fused benzene rings, with occasional incorporation of cyclopentene rings. A wide variation of alkyl substituents gives rise to thousands of different PAHs and many have been identified in environmental samples. PAHs are generally highly toxic and carcinogenic compounds [1] and ubiquitous contaminants of aquatic and atmospheric ecosystems, where they are present as a result of natural processes such as forest fires, volcanic emissions, but the predominant PAH sources in the environment are related to human activities such as oil spills, burning fossil fuels and domestic wastes, transport emissions.

In past decades, Quantitative Structure-Activity properties Relationships / (QSAR / QSPR/ QSRR/) or QSXR has become a powerful theoretical tool, alternative to quantum mechanics, for the description and prediction of properties of complex molecular systems in different environments. The approach QSXR proceeds from the assumption of a one correspondence between any physical property, chemical affinity or biological activity of a chemical compound and its molecular structure [2]. The latter can be represented by the

chemical composition, the connectivity of the atoms, the potential energy surface, and the electron wave function of a compound. Different physicochemical molecular descriptors reflecting the structure can be determined empirically or by using theoretical and computational methods of different complexities. It is emphasized that the knowledge of the exact chemical constitution and / or the three-dimensional molecular structure of the studied compounds is a prerequisite to the application of QSXR approach. The success of the approach QSXR depends critically on the precise definition and the appropriate use of molecular descriptors. We distinguish arbitrarily empirical molecular descriptors theoretical molecular descriptors. The empirical descriptors can be divided into two general classes, the first reflects the intramolecular electronic interactions (structural descriptors) while the second takes into account the intermolecular interactions in condensed media such as liquids and solutions (solvation descriptors). The objective of this work is to develop a robust QSRR model that could predict the retention indices for a diverse set of PAHs, using the general molecular descriptors and to seek the important features related to the retention indices value.

## 2. Materials and methods

### 2.1. Dataset:

The experimental retention indices values for 93 PAHs were taken from the article published by Jujun Kang *et al* [3]. A complete list of the compounds name and their corresponding retention indices of the 93 PAHs is shown in Table 1. The data set was randomly divided into two subsets: a training set of 70 compounds and a test set of 23 compounds.

### 2.2. Descriptor Generation:

The chemical structure of each compound was sketched on a PC using Spartan 10 [4] program and optimized using PM6 semi empirical method. The resulted geometry was transferred into the soft ware Dragon version 5.3[5], to calculate 1600

descriptors of the type geometrical and Getaway(Geometry, Topology and Atom Weighted Assembly).descriptors with constant or near constant values inside each group were discarded .For each pair of correlated descriptors (with correlation coefficient  $r \geq 0.95$ ),the one showing the highest pair correlation with the other descriptors was excluded. The GA (Genetic Algorithm) [6] has been considered superior to other methods of variable selection techniques .So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini [7] by maximizing the cross-validated explained variance  $Q^2_{LOO}$ .

The chemical structure of each compound was sketched on a PC using Spartan10[4] program and optimized using PM6 semi empirical method .the resulted geometry was transferred into the soft ware Dragon version 5.3[5], to calculate 1600 descriptors of the type geometrical and Getaway(Geometry, Topology and Atom Weighted Assembly).descriptors with constant or near constant values inside each group were discarded .For each pair of correlated descriptors (with correlation coefficient  $r \geq 0.95$ ),the one showing the highest pair correlation with the other descriptors was excluded. The GA (Genetic Algorithm) [6] has been considered superior to other methods of variable selection techniques .So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini [7] by maximizing the cross-validated explained variance  $Q^2_{LOO}$ .

#### 2.4. Model Development and Validation:

Multiple linear regression analysis and variable selection were performed by package MobyDigs for windows/PC [7][31] using the Ordinary Least Square regression (OLS) method. 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).

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

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 every one. 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.

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

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).

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

A value  $Q^2 > 0.5$  is generally regarded as a good result and  $Q^2 > 0.9$  as excellent [32, 33][8,9].

However, studies [10, 11][34, 35] 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

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

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.

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:

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

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

According to Golbraikh and Tropsha[11]. A QSPR model is successful if it satisfies several criteria as follows :

$$R^2_{CVext} > 0.5 \quad (6)$$

$$r^2 > 0.6 \quad (7)$$

$$(r^2 - r^2_0)/r^2 < \text{or } (r^2 - r^2_0)/r^2 < 0.1 \quad (8)$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \quad (9)$$

Here:

$$r = \frac{\sum (y_i - \tilde{y}_i)(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (10)$$

$$r_0^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \quad (11)$$

$$r_0'^2 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y})^2} \quad (12)$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \quad (13)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad (14)$$

$$T1 = \frac{(r^2 - r^2_0)}{r^2} \quad (15)$$

$$T2 = \frac{(r^2 - r'^2_0)}{r^2} \quad (16)$$

$$Ab = [r^2_0 - r'^2_0] \quad (17)$$

where  $r$  is the correlation coefficient between the calculated and experimental values in the test set;  $r^2_0$  (calculated versus observed values) and  $r'^2_0$  (observed versus calculated values) are the coefficients of determination;  $k$  and  $k'$  are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively

$y_i^{r_0}$ ,  $\tilde{y}_i^{r_0}$ ; are defined as  $y_i^{r_0} = k\tilde{y}_i$  and  $\tilde{y}_i^{r_0} = k'y_i$  and the summations runs over the test set.

## 2.5. QSAR AD (Applicability Domain)

The AD was discussed by the Williams plot [8, 9] of jackknifed residuals versus leverages (hat diagonal values ( $h_i$ )). The jackknifed residuals (or Studentized residuals) are the standardized cross-validated residuals. Each residuals is divided by its standard deviation, which is calculated without the  $i$ -th observation. The leverage( $h_i$ ). value of a chemical in the original variable space is defined as :

$$h_i = x_i (X^T X)^{-1} x_i^T \quad (i=1, \dots, n) \quad (18)$$

Where  $x_i$  is the descriptor row-vector of the query compound, and  $X$  is the  $n \times (p+1)$  matrix of  $p$  model parameter values for  $n$  training set compounds. The superscript  $T$  refers to the transpose of the matrix/vector. The warning leverage value ( $h^*$ ) is defined as  $3(p+1)/n$ . When  $h$  value of a compound is lower than  $h^*$ , the probability of accordance between predicted and actual values is as high as that for the compounds in the training set. A chemical with  $h_i > h^*$  will reinforce the model if the chemical is in the training set. But such a chemical in the validation set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residual may be low. Thus the leverage and the jackknifed residual should be combined for the characterization of the AD.

## 3. Results and Discussion

Application of the GA-VSS led to several good models for the prediction of based on different sets of molecular descriptors. The best biparametric model was constructed using: Molecular weight; salvation energy .All data concerning value of descriptors and the retention indices are summarized in Table 1.

The equation of the optimal model can be written as:

$$R_i = -50. \pm(4.400) -4.63 \pm(0.281) \text{SOLV EN} +1.52 \pm(0.031) \text{MW} \quad (19)$$

Here Molecular weight (MW); solvation energy( SOLV EN) are descriptors calculated with Spartan software, All relevant statistical parameters are reported in Table 2.

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

The model is robust, the difference between  $R^2$  and  $Q^2$  is small 0.09(%). The model demonstrates a very good stability in internal validation (difference between  $Q^2_{LOO}$  and  $Q^2_{LMO}$  is about 0.04(%). While bootstrapping confirms the internal predictivity and stability of the model.  $SDEP_{ext}$  is a little bit different from  $SDEP$ . Some important statistical parameters (as given in Table 3) were used to evaluate the involved descriptors. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in Table 3 express that the regression coefficients of the descriptors involved in the GA/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 QSRR model (i.e., descriptors' interactions). Descriptors with t-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance[12][13]. The smaller t-probability suggests the more significant descriptor. The t-probability values of the tree descriptors are very small, indicating that all of them are highly significant descriptors. The VIF values and the correlation matrix as shown in Table 4 suggest that these descriptors are weakly correlated with each other. The distributions of errors for the entire dataset are given in Figure 1. As the errors are distributed on both sides of the zero line, one may conclude that there is no systematic error in the model development. The model was also verified by Y-scrambling. Figure 2 clearly ensures the existence of a linear relationship between  $R_i$  and the descriptors Molecular weight. As can be observed the permuted responses yield poor predictive models, all having  $Q^2 < 0.2$ . On the other hand, the correctly ordered  $R_i$  yield good statistical parameters, and therefore it is located isolated in the plot.

The statistical parameters of Tropsha et al reported in Table 5 were obtained for the test set, which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the present model:

$$R^2_{cv\_ext} = >0.5$$

$$r^2 = >0.6$$

$$T1 = (r^2 - r^2_0) / r^2 = -0.0012 < 1$$

$$\text{Or } T2 = (r^2 - r^2_0) / r^2 = -0.0012 < 0.1$$

$$0.85 \leq k = 1.0101 < 1.15 \text{ or } 0.85 k' = 0.9899 \leq 1.15$$

**Table 1: Values of  $R_i$ , Molecular weight, solvation energy for a set of 93 PAHs. The last 23 chemicals are the test set.**

Chemical	$R_i$	Solvation energy (Solv En)	Molecular weight (MW)
Naphthalene	200	-11.24	128.174
1-Methylnaphthalene	221.04	-11.61	142.201
2-Ethyl naphthalene	236.08	-9.99	156.228
1-Ethyl naphthalene	236.56	-10.86	156.228
2,7-Dimethylnaphthalene	237.71	-11.75	156.228
1,3-Dimethylnaphthalene	240.25	-12.03	156.228
1,7-Dimethylnaphthalene	240.66	-11.6	156.228
1,6-Dimethylnaphthalene	240.72	-11.97	156.228
1,4-Dimethylnaphthalene	243.57	-11.62	156.228
Acenaphthelene	244.63	-15.77	152.196
1,5-Dimethylnaphthalene	244.98	-11.75	156.228

1,2-Dimethylnaphthalene	246.49	-11.95	156.228
2,3,6-Trimethylnaphthalene	263.31	-11.84	170.255
1-Methylacenaphthelene	265.24	-16.66	166.223
2,3,5-Trimethylnaphthalene	265.9	-12.12	170.255
Phenanthrene	300	-16.21	178.234
1-Phenylnaphthalene	315.19	-11.69	204.272
3-Methylphenanthrene	319.46	-16.56	192.261
2-Methylanthracene	321.57	-15.64	192.261
2-Methylphenanthrene	321.57	-16.3	192.261
4-Methylphenanthrene	323.17	-16.97	192.261
Chemical	$R_i$	Solvation energy (Solv En)	Molecular weight (MW)
1-Methylanthracene	323.33	-15.67	192.261
1-Methylphenanthrene	323.9	-16.73	192.261
9-Methylanthracene	329.13	-17.14	192.261
9-Ethylphenanthrene	337.05	-15.98	206.288
2-Ethylphenanthrene	337.5	-14.89	206.288
3,6-Dimethylphenanthrene	337.83	-16.89	206.288
2,7-Dimethylphenanthrene	339.23	-16.32	206.288
9-Isopropylphenanthrene	345.78	-13.62	220.315
1,8-Dimethylphenanthrene	346.26	-17.19	206.288
9-n-Propylphenanthrene	350.3	-14.47	220.315
Pyrene	351.22	-20.51	202.256
9-Methyl-10-Ethylphenanthrene	359.91	-15.5	220.315
1-Methyl-7-isopropylphenanthrene	368.67	-14	234.342
4-Methylpyrene	369.54	-21	216.283
1-Methylpyrene	373.55	-21.3	216.283
9,10-Dimethyl-3-ethylphenanthrene	381.85	-16.26	234.342
1-Ethylpyrene	385.35	-20.48	230.31
2,7-Dimethylpyrene	386.34	-20.64	230.31
Benzo(c)phenanthrene	391.39	-19.54	228.294
9-Phenylanthracene	396.38	-14.51	254.332
Cyclopenta(cd)pyrene	396.54	-24.09	226.278
Benzo(a)anthracene	398.5	-19.93	228.294
Triphenylene	400	-21.24	228.294
9-Phenylphenanthrene	406.9	-15.76	254.332
11-Methylbenzo(a)anthracene	412.72	-20.37	242.321

1-Methylbenzo(a)anthracene	414.37	-20.87	242.321
1-n-Butylpyrene	414.87	-18.24	258.364
1-Methyltriphenylene	416.32	-20.86	242.321
9-Methylbenzo(a)anthracene	416.5	-20.17	242.321
9-Methyl-10-phenylphenanthrene	417.16	-15.77	268.354
8-Methylbenzo(a)anthracene	417.56	-20.35	242.321
6-Methylbenzo(a)anthracene	417.57	-20.39	242.321
3-Methylchrysene	418.1	-21.11	242.321
2-Methylchrysene	418.8	-20.83	242.321
<b>Chemical</b>	<b>Ri</b>	<b>Solvation energy (Solv En)</b>	<b>Molecular weight (MW)</b>
12-Methylbenzo(a)anthracene	419.39	-20.52	242.321
4-Methylbenzo(a)anthracene	419.67	-20.51	242.321
5-Methylchrysene	419.68	-20.35	242.321
4-Methylchrysene	420.83	-20.42	242.321
1-Phenylphenanthrene	421.66	-16.02	254.332
1-Methylchrysene	422.87	-21.24	242.321
7-Methylbenzo(a)anthracene	423.14	-21.88	242.321
1,12-Dimethylbenzo(a)anthracene	436.82	-17.49	256.348
Benzo(j)fluoranthene	440.92	-25.17	252.316
Benzo(b)fluoranthene	441.74	-23.86	252.316
Benzo(k)fluoranthene	442.56	-23.42	252.316
1,6,11-Trimethyltriphenylene	446.24	-21.44	270.375
Benzo(e)pyrene	450.73	-25.11	252.316
Benzo(a)pyrene	453.44	-24.4	252.316
Perylene	456.22	-24.93	252.316
Pentacene	486.81	-22.65	278.354
Dibenzo(a,c)anthracene	495.01	-24.59	278.354
Dibenzo(a,h)anthracene	495.45	-24.35	278.354
Picene	500	-25.02	278.354
Dibenzo(def,mno)chrysene	503.89	-27.94	276.338
2-Methylnaphthalene	218.14	-11.4	142.201
2,6-Dimethylnaphthalene	237.58	-11.34	156.228
2,3-Dimethylnaphthalene	243.55	-11.73	156.228

1,8-Dimethylnaphthalene	249.52	-12.71	156.228
Anthracene	301.69	-15.34	178.234
9-Methylphenanthrene	323.06	-16.71	192.261
2-Phenylnaphthalene	332.59	-13.02	204.272
Fluoranthene	344.01	-20.27	202.256
9,10-Dimethylanthracene	355.49	-18.19	206.288
2-Methylpyrene	370.15	-20.63	216.283
Chrysene	400	-20.72	228.294
2-Methylbenzo(a)anthracene	413.78	-20.26	242.321
3-Methylbenzo(a)anthracene	416.63	-20.04	242.321
5-Methylbenzo(a)anthracene	418.72	-20.54	242.321
<b>Chemical</b>	<b>Ri</b>	<b>Solvation energy (Solv En)</b>	<b>Molecular weight (MW)</b>
6-Methylchrysene	420.61	-21.61	242.321
1,3-Dimethyltriphenylene	432.32	-21.37	256.348
7,12-Dimethylbenzo(a)anthracene	443.38	-21.4	256.348
Benzo(b)chrysene	497.66	-24.29	278.354

Table 2 statistical parameters of a developed model.

$n_{tr}$	$n_{ext}$	$Q^2_{LOO}$	$R^2$	$Q^2_{LMO/50}$	$Q^2_{BOOT}$	$R^2_{adj}$	$Q^2_{t}$
70	23	99.18	99.27	99.1324	99.12	99.25	99.7
$SDEC$	$SDEP$	$SDEP_{ext}$	$S$	$F$			
6.108	6.475	6.519	6.2428	4569.847			

Table 3. Characteristics of the selected descriptors in the best GA/ MLR model.

Predictor	Coef	SE Coef	T	P	VIF
Constant	-49.968	4.400	-11.36	0.000	-
SOLV EN	-4.6285	0.2803	-16.51	0.000	2.341
WEIGHT	1.52264	0.03104	49.05	0.000	2.341

Table 4: correlation matrix between retention indices and the selected descriptors.

	Ri	SOLV EN	Molecular WEIGHT
Ri	1.000		
SOLV EN	-0.855	1.000	
Molecular WEIGHT	0.981	-0.757	1.000

Table 5: The statistical parameters of Tropsha *et al.*

$R^2_{cv\ ext}$	$r^2$	k	k'	$r^2_o$
0.9954	0.9969	1.0101	0.9899	0.9980
$r^2_o$	Ab	T1	T2	
0.9980	0.000	-0.0012	-0.0012	

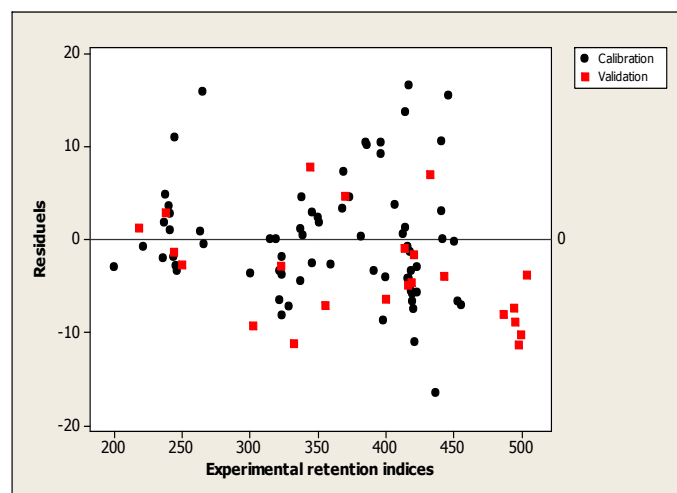


Fig1 :Residuals versus Ri (exp) for the entire dataset.

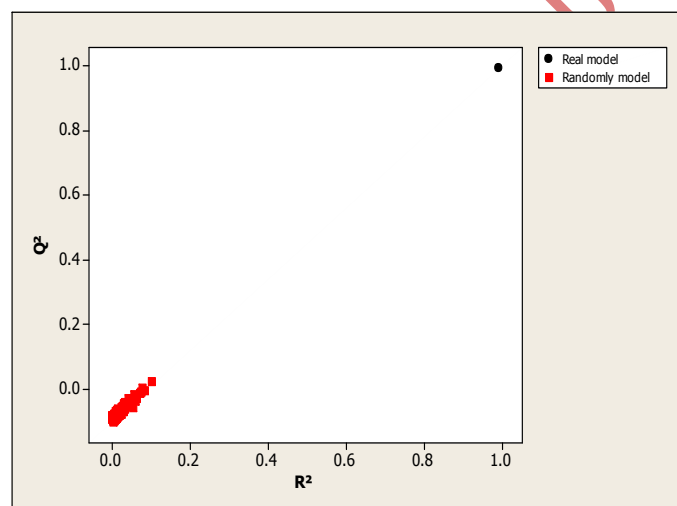


Fig 2 :Randomization test associated to the previous QSRR model. Square represent the randomly ordered retention and the circle corresponds to the real retention.

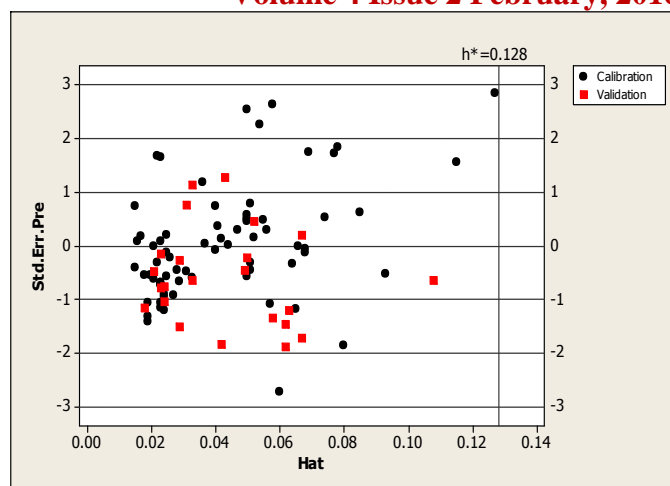


Fig 3: Williams plot of the current QSRR model.

### 3.3 Applicability Domain

On analyzing the model applicability domain from Williams plot, all residuals were located within the range of three Standard deviations, and there is no influential compound both for training or prediction set (Figure.3), which means that the model has a good external predictivity.

### 4. Conclusion

A QSRR model for the estimation of the retention indices for 93 PAHs was established in the following six steps: molecular structure input and generation of the files containing the chemical structures; quantum mechanics geometry optimization with a semi-empirical method; structural descriptors computation; structural descriptors selection; structure model generation with a multivariate method and statistical analysis.

According to obtained results it is concluded that the salvation energy and the molecular weight can be used successfully for modeling retention indices ( $R_i$ ) of the under study compounds. High correlation coefficient (0.9927) and low prediction error ( $SDEP=6.475$ ;  $SDEP_{ext}=6.519$ ) obtained confirm good predictive ability of the model. The QSRR model proposed with the double calculated molecular descriptors can be used to estimate retention indices for new compounds even in the absence of the standard candidates.

### References

- [1] National Research Council (1979), Committee on the Assessment of Polychlorinated Biphenyls in the Environment. Polychlorinated biphenyls: a report; National Academy of Sciences: Washington, U.S.A.
- [2] Angulo Lucena, R., Farouk Allam, M., Serrano Jiménez, S. and Luisa Jodral Villarejo, M. A. (2007), "review of environmental exposure to persistent organochlorine residuals during the last fifty years". Current Drug Safety, Vol. 2 No.2, pp. 163-172.
- [3] Roveda, A. M., Veronesi, L., Zoni, R., Colucci, M. E. and Sansebastiano, G. (2006), "Exposure to polychlorinated

biphenyls (PCBs) in food and cancer risk: recent advances”, *Igiene e Sanita Pubblica*, Vol. 62 No.6, pp. 677-696.

[4] Lundqvist, C.,Zuurbier, M., Leijs, M., Johansson, C.,Ceccatelli, S.,Saunders, M.,Schoeters, G., Ten Tusscher, G. and Koppe, J. G. (2006),” The effects of PCBs and dioxins on child health”, *Acta Paediatrica*,Vol.95 No.453,pp.55-64.

[5] Poppenga, R. H. (2000), “Current environmental threats to animal health and productivity”, *The Veterinary Clinics of North America. Food Animal Practice* ,Vol. 16 No.3, pp.545-558.

[6] Bren, U.,Zupan, M., Guengerich, F. P. and Mavri, J.( 2006)” Chemical Reactivity as a Tool to Study Carcinogenicity: Reaction between Chloroethylene Oxide and Guanine”, *The Journal of Organic Chemistry* ,Vol. 71 No.11,pp. 4078-4084.

[7] Lebeuf, M., Noël, M.,Trottier, S. and Measures, L. (2007), “Temporal trends (1987-2002) of persistent,bioaccumulative and toxic (PBT) chemicals in beluga whales (*Delphinapterus leucas*) from the St.Lawrence Estuary, Canada”, *Sciences of the Total Environment*, Vol.383 No. (1-3), pp.216-231.

[8]]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.

[9] 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

[10] Kubinyi H., Hamprecht F.A. & Mietzner T., 1998. Three-dimensional quantitative similarity-activity relationships (3D QSiAR ) from SEAL similarity matrices, *Journal of Medicinal Chemistry*, Vol.41(14), 2553-2564.

[11] Golbraikh, A. and Tropsha, A. (2002), “Beware of  $q^2!$ ”,*Journal of Molecular Graphics and Modelling*, Vol.20 No.4, pp.269-276.

[12] Ramsey, L. F.; Schafer, W. D. *The Statistical Sleuth*; Wadsworth Publishing Company: USA, 1997.

# Estimation and modeling of bioaccumulation factor for a set of (PCBs): a QSPR study

Khadija Amirat<sup>1</sup> ; Fatiha Mebarki<sup>2</sup> ; Nadia Ziani<sup>3</sup> ; Djelloul Messadi<sup>4</sup>

Environmental and food Security laboratory, Badji Mokhtar University, BP 12, 23000,  
Annaba, Algeria

1:khadija\_amirat@yahoo.fr

2:Fatiha\_mebarki@yahoo.fr

3:Ziani\_nadia84@yahoo.fr

4:d\_messadi@yahoo.fr

## Abstract

A structure/ bioaccumulation factor relationship was searched for 58 PCBs while promoting the hybrid genetic algorithm/simple linear regression approach, the structural parameters being calculated with the software Hyperchem and DRAGON. Among about a hundred of single regressor models gotten, we selected the one that present best values of the prediction parameter ( $Q^2$ ) and of the determination coefficient ( $R^2$ ). The reliability of the proposed model was further illustrated using various evaluation techniques: leave-many-out, cross-validation procedure, randomization test, and validation through the test set.

**Keywords:** PCBs, bioaccumulation factor, QSPR, molecular descriptors, software.

## 1. INTRODUCTION

Polychlorinated biphenyls (PCBs), organic compounds with 1 to 10 chlorine atoms attached to biphenyl, have the general chemical formula  $C_{12}H_{10-x}Cl_x$  (Figure 1). First manufactured by Monsanto in 1929, the PCBs production was banned in the 1970<sup>th</sup> due to the high toxicity of most PCBs (209) and mixtures [1]. PCBs were used as insulating fluids for industrial transformers and capacitors, and are known as persistent organic pollutants. Even if the production of the PCBs was stopped, they still have an influence on the human [2-4] and animal [5] health due to their accumulation in the environment. Moreover, the toxicity and carcinogenicity of PCBs could be related to mechanistic studies of their truncated analogue vinyl chloride [6]. Ecological and toxicological aspects of polychlorinated biphenyls (PCBs) in the environment are under investigation due to their worldwide distribution [7-10]. Starting with the 20th century, several mathematical approaches, that link chemical structure and property/activity in a quantitative manner, have been introduced [11]. Nowadays, quantitative structure-property/activity relationships (QSPRs/QSARs) are currently used in pharmaceutical chemistry, toxicology and other related fields [12].

Bioaccumulation of chemicals is quantitatively expressed in terms of BCF, defined as the equilibrium of its concentration

inside an organism (or in a certain tissue of the organism, usually in the fat) to that in the ambient environment [13]. The concentrations in tissues and in the environment are measured at steady-state after chronic exposure. However, the real test period may be too short to achieve steady-state. In addition, metabolism and chemical degradation may occur and large molecules may not permeate sufficiently through membranes into the organism, often lowering BCF values. Thus, experimental determination of BCF may underestimate the environmental risk [14]. In ideal case, the measured value of BCF should be strongly related to the high complexity of bioaccumulation process, taking into account such factors as metabolism, organ-specific bioconcentration, irreversible binding onto proteins, incomplete depuration, and kinetic effects [15]. Fish with an average lipid content of 4.8% are preferred model animals for bioconcentration studies due to their relevance as food for many species, including humans [16], and to the availability of standardized testing protocols. Bioaccumulation is a thermodynamically driven partitioning process between aquatic environment and the lipid tissues of fish, thus, n-octanol is often a satisfactory surrogate for biological lipids [17]. As demonstrated earlier, it is important to know the BCF of all PCBs congeners. The literature data on experimental BCF of PCBs are limited and their measurement is difficult and expensive. Thus, quantitative structure-property relationship (QSPR) methods based on the descriptors derived directly from the molecular structure are vital to supply the missing data independently of experimentation.

The BCF of a chemical is most commonly estimated from established correlations between  $\log BCF$  and  $\log KOW$  [17, 18]. However, multilinear QSAR/QSPR models including  $\log KOW$  are valid only for compounds with  $\log KOW$  values  $< 6$  [19, 20]. For highly hydrophobic chemicals ( $\log KOW > 6$ ) non-linear [18, 21] bilinear [18] and polynomial [22] equations relate  $\log BCF$  and  $\log KOW$ . While  $\log KOW$  models indicate priorities for assessing dangerous substances, they may not provide reliable predictions for unknown BCF.

The objective of this work is to develop a robust QSPR model that could predict the bioaccumulation factor for a diverse set of PCBs, using the general molecular descriptors and to seek



the important features related to the bioaccumulation factor values.

## II Methodology

### II.1. Dataset:

Known experimental logarithmic BCF values of the 58 PCBs were taken from the literature [23]. The whole set of experimentally observed values lies in the range from 2.64 to 5.97 (<6) clearly indicating that the application of the linear QSAR approach to the studied property is relevant [19, 20]. A complete list of the compounds name and their corresponding experimental values of PCBs is shown in Table 1. The data set was divided into two subsets: a training set of 30 compounds and a test set of 28 compounds according to Kennard and Stones algorithm. The training set was used to build the genetic algorithm /SLR and the test set was used to evaluate its prediction ability of the model.

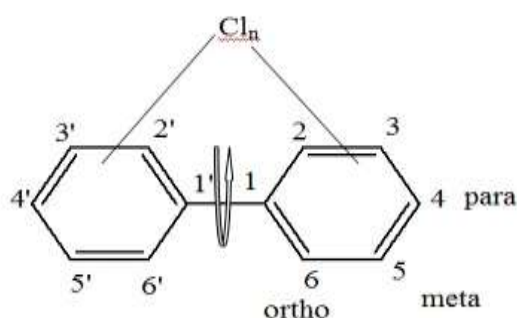


Figure 1. General structural formulae and substitution positions of the PCBs.

Table 1: Values of LogBCF(exp), HATS0v for a set of 58 PCBs. The last 28 chemicals are the test set.

Congener Number	LogBCFexp	HATS0v
PCB 0	2.64	0.063
PCB 3	2.77	0.085
PCB15	3.28	0.106
PCB 4	3.38	0.099
PCB 8	3.57	0.106
PCB 54	3.85	0.129
PCB 18	4.11	0.132
PCB28	4.2	0.126
PCB 40	4.23	0.156
PCB29	4.26	0.142
PCB 77	4.59	0.174
PCB 64	4.6	0.151
PCB 52	4.63	0.16
PCB 87	5.38	0.181
PCB 157	5.39	0.216

PCB 101	5.4	0.18
PCB 136	5.43	0.176
PCB 209	5.44	0.251
Congener Number	LogBCFexp	HATS0v
PCB 208	5.71	0.237
PCB 174	5.8	0.215
PCB 180	5.8	0.225
PCB 126	5.81	0.209
PCB 141	5.81	0.208
PCB 194	5.81	0.248
PCB 202	5.82	0.222
PCB 198	5.88	0.237
PCB 195	5.92	0.231
PCB 196	5.92	0.232
PCB 169	5.97	0.238
PCB14	3.78	0.143
PCB 7	3.55	0.106
PCB 6	3.8	0.12
PCB 9	3.89	0.121
PCB 5	4.11	0.117
PCB31	4.23	0.143
PCB 70	4.77	0.175
PCB 44	4.84	0.162
PCB 49	4.84	0.154
PCB 47	4.85	0.142
PCB 155	4.93	0.168
PCB 48	5	0.153
PCB 90	5	0.184
PCB 99	5	0.172
PCB 105	5	0.186
PCB 109	5	0.177
PCB 118	5	0.193
PCB 138	5.39	0.199
PCB 148	5.39	0.194
PCB 156	5.39	0.219
PCB 97	5.43	0.181
PCB 151	5.54	0.199
PCB 153	5.65	0.198
PCB 128	5.77	0.194
PCB 182	5.8	0.211
PCB 187	5.8	0.217

PCB 183	5.84	0.211
PCB 191	5.84	0.223
PCB 137	5.88	0.201

## 2.2. Descriptor Generation:

The structures of the molecules were drawn using Hyperchem 6.03 software [24]. The final geometries were obtained with the semi empirical method PM3. All calculations were carried out at the RHF (restricted Hartree-Fock) level with non configuration interaction. The molecular structures were optimized using the algorithm Polak-Ribiere and a gradient norm limit of 0.001 kcal.A<sup>o</sup><sup>-1</sup>.mol<sup>-1</sup>. The resulted geometry was transferred into the software Dragon version 5.3 [25] to calculate 1600 descriptors of the type Geometrical and GETAWAY (Geometry, Topology and Atoms Weighted Assembly). Descriptors with constant or near constant values inside each group were discarded. For each pair of correlated descriptors (with correlation coefficient  $r \geq 0.95$ ), the one showing the highest pair correlation with the other descriptors was excluded. The GA (Genetic Algorithm) [26] has been considered superior to other methods of variable selection techniques. So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini [27] by maximizing the cross-validated explained variance  $Q^2_{LOO}$ .

## 2.3. Kennard and Stone algorithm

Kennard and Stone's algorithm [28] has been widely used for splitting datasets 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 dataset 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 Stone's algorithm has been considered as one of the best ways to build training and test sets [29,30]. Using Kennard and Stone's algorithm the entire set divided into two subsets: training set of 40 compounds, and a test set including the remaining 18 compounds.

## 2.4. Model Development and Validation:

Models with one variable were performed by the software MOBYDYGS [31] using the Ordinary Least Square regression (OLS) method.

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).

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

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 every one. 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.

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

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).

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

A value  $Q^2 > 0.5$  is generally regarded as a good result and  $Q^2 > 0.9$  as excellent [32, 33].

However, studies [34, 35] 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

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

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 was divided according to Kennard and Stone algorithm into a training set (30 objects) used to develop the QSAR models and a validation set (28 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:

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

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

According to Golbraikh and Tropsha[35]. A QSPR model is successful if it satisfies several criteria as follows :

$$R^2_{cv_{ext}} > 0.5 \quad (6)$$

$$r^2 > 0.6 \quad (7)$$

$$(r^2 - r_0^2) / r^2 < \text{or } (r^2 - r_0'^2) / r^2 < 0.1 \quad (8)$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \quad (9)$$

Here:

$$r = \frac{\sum (y_i - \tilde{y}_i)(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (10)$$

$$r_0^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \quad (11)$$

$$r_0'^2 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y})^2} \quad (12)$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \quad (13)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad (14)$$

$$T1 = \frac{(r^2 - r_0^2)}{r^2} \quad (15)$$

$$T2 = \frac{(r^2 - r_0'^2)}{r^2} \quad (16)$$

$$Ab = [r^2_0 - r_0'^2] \quad (17)$$

where  $r$  is the correlation coefficient between the calculated and experimental values in the test set;  $r^2_0$  (calculated versus observed values) and  $r_0'^2$  (observed versus calculated values) are the coefficients of determination;  $k$  and  $k'$  are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively

$y_i^{r_0}$ ,  $\tilde{y}_i^{r_0}$ ; are defined as  $y_i^{r_0} = k\tilde{y}_i$  and  $\tilde{y}_i^{r_0} = k'y_i$  and the summations runs over the test set.

## 2.5. QSAR AD (Applicability Domain)

The AD was discussed by the Williams plot [31,32] of jackknifed residuals versus leverages (hat diagonal values ( $h_i$ )). The jackknifed residuals (or Studentized residuals) are the standardized cross-validated residuals. Each residual is divided by its standard deviation, which is calculated without the  $i$ -th observation. The leverage ( $h_i$ ), value of a chemical in the original variable space is defined as :

$$h_i = x_i (X^T X)^{-1} x_i^T \quad (i=1, \dots, n) \quad (18)$$

Where  $x_i$  is the descriptor row-vector of the query compound, and  $X$  is the  $n \times (p+1)$  matrix of  $p$  model parameter values for  $n$  training set compounds. The superscript  $T$  refers to the transpose of the matrix/vector. The warning leverage value ( $h^*$ ) is defined as  $3(p+1)/n$ . When  $h$  value of a compound is lower than  $h^*$ , the probability of accordance between predicted and actual values is as high as that for the compounds in the training set. A chemical with  $h_i > h^*$  will reinforce the model if the chemical is in the training set. But such a chemical in the validation set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residual may be low. Thus the leverage and the jackknifed residual should be combined for the characterization of the AD.

## 3. Results and Discussion

Application of the GA-VSS led to several good models for the prediction of based on different sets of molecular descriptors. The best single dimensional model was constructed using the descriptor HATS0e. All data concerning value of this descriptor and the dependent variable (LogBCF) are summarized in Table 1.

The equation of the optimal model can be written as:

LOG BCF = 1.5779 ± (0.1958) + 18.538± (1.065) HATS0v.  
(19)

HATS0v is calculated in Dragon software. More information about this descriptor can be found in Dragon software user's guide [25] and the references therein.

All relevant statistical parameters are reported in Table 2.

Values of R<sup>2</sup> and R<sup>2</sup><sub>adj</sub> attest the good fitting performances of the model which, moreover, is very highly significant (great value of the Fisher parameter F).

The model is robust, the difference between R<sup>2</sup> and Q<sup>2</sup> is small (1.25%). The model demonstrates a very good stability in internal validation (difference between Q<sup>2</sup><sub>LOO</sub> and Q<sup>2</sup><sub>LMO</sub> is about 0.76%). While bootstrapping confirms the internal predictivity and stability of the model. SDEP<sub>ext</sub> is a little bit different from SDEP. The model works slightly worse in external prediction than in internal prediction.

Some important statistical parameters (as given in Table 3) were used to evaluate the involved descriptor. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in Table 3 express that the regression coefficients of the descriptors involved in the GA/SLR 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% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [36]. The smaller t-probability suggests the more significant descriptor. The t-probability values of the two descriptors are equal to zero, indicating that all of them are highly significant descriptors. The VIF values and the correlation matrix as shown in Table 4 suggest that these descriptors are weakly correlated with each other. The distributions of errors for the entire dataset are given in Figure 2. As the errors are distributed on both sides of the zero line, one may conclude that there is no systematic error in the model development. The model was also verified by Y-scrambling. Figure 3 clearly ensures the existence of a linear relationship between LogBCF and the descriptor HATS0v; As can be observed the permuted responses yield poor predictive models, all having Q<sup>2</sup> < 0.2. On the other hand, the correctly ordered LogBCF yield good statistical parameters, and therefore it is located isolated in the plot.

The statistical parameters of Tropsha et al reported in Table 5 were obtained for the test set, which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the present model:

$$R^2_{cv_{ext}} = 0.7793 > 0.5$$

$$r^2 = 0.8707 > 0.6$$

$$T1 = \frac{(r^2 - r'^2_0)}{r^2} = 0.0014 < 1$$

$$\text{Or } T2 = \frac{(r^2 - r'^2_0)}{r^2} = -0.0046 < 0.1$$

$$0.85 \leq k = 1.0447 < 1.15 \text{ or } 0.85 \leq k' = 0.9552 \leq 1.15$$

**Table 2 statistical parameters of a developed model.**

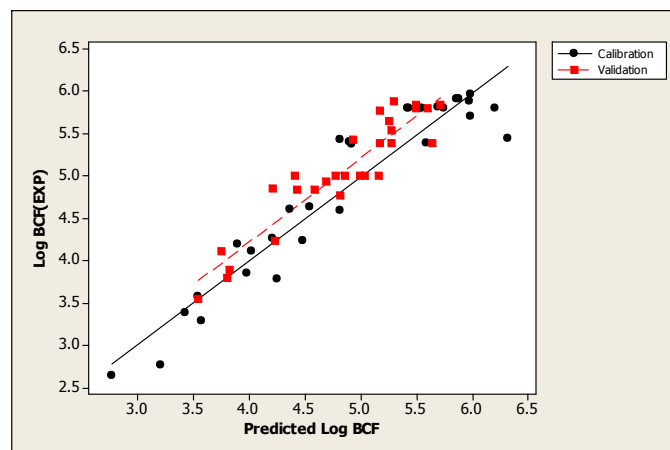
$n_{tr}$	$n_{ext}$	$Q^2_{LOO}$ (%)	$R^2$ (%)	$Q^2_{LMO/50}$ (%)	$Q^2_{BOOT}$ (%)	$R^2_{adj}$ (%)	$Q^2_{e_{xt}}$ (%)
30	28	90.28	91.53	89.52	89.23	91.23	90.44
$SD_{EC}$	$SD_{EP}$	$SDEP_{ext}$	$S$	$F$			
0.3	0.3	0.319	0.310	302.742			
	21		6	3			

**Table 3. Characteristics of the selected descriptor in the best GA/SLR model.**

Predictor	Coef	SE Coef	T	P
Constant	1.5779	0.1958	8.06	0.000
HATS0v	18.538	1.065	17.40	0.000

**Table 4: The statistical parameters of Tropsha et al.**

$R^2_{cv_{ext}}$	$r^2$	k	k'	$r^2_0$
0.7793	0.8707	1.0447	0.9552	0.8695
$r^2_0$	Ab	T1	T2	
0.8747	0.0948	0.0014	-0.0046	



**Fig:2 Predicted Log BCF versus experimental LogBCF for the entire dataset.**

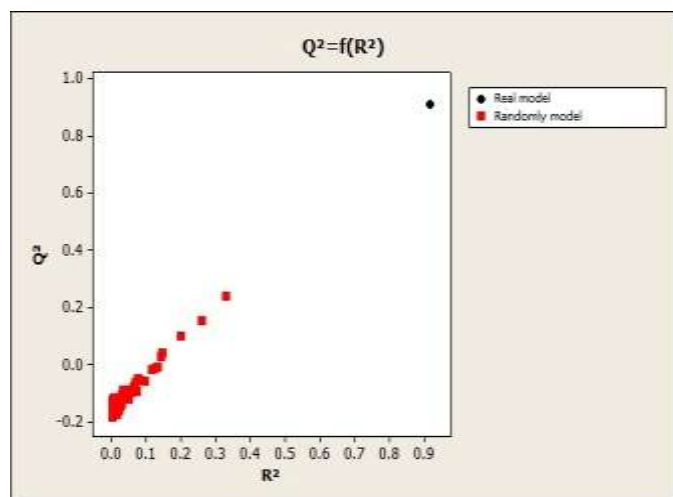


Fig:2 Randomization test associated to the previous QSPR model. Square represent the randomly ordered properties and the circle corresponds to the real properties.

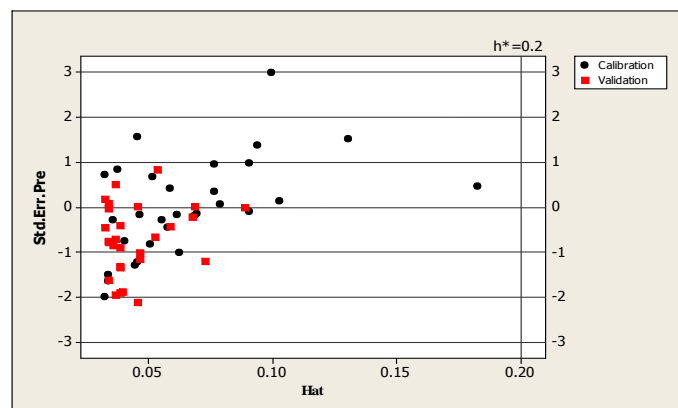


Fig: 3 Williams plot of the current QSRR model

### 3.2 Mechanistic Interpretation

The selected descriptor and its class and meaning are gathered in the Table 5.

Table 5: selected descriptor and its meaning and class for the best GA/ SLR model.

Descriptor	meaning	class
HATS0v	leverage-weighted autocorrelation of lag 0 / weighted by atomic van der Waals volumes	GETAWAY

HATS0v is a GETAWAY descriptor and correlates with the experimental Log BCF values of 0.957. The GETAWAY descriptors [37,38] have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix, which is based on the spatial autocorrelation formulas, weighting the molecule atoms by the physico-chemical properties  $w$  together with 3D information encoded by the elements of the molecular influence matrix  $H$  and influence/distance matrix  $R$ . These descriptors, as based on spatial autocorrelation, encode information on the effective position of substituents and fragments in the molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties. The positive sign of HATS0v (equation 20) means that the increase in this descriptor increases the Log BCF.

$$HATS0v = \sum_{i=1}^{nat-1} \sum_{j>1} (V_i \cdot h_{ij}) \cdot (V_j \cdot h_j) \delta(0, d_j) \quad (20)$$

Where;  $nAT$  is the number of molecule atoms;  $d_{ij}$  is the topological distance between atoms  $i$  and  $j$ ;  $V$  is the atomic van der Waals volumes;  $d$  is the topological diameter;  $\delta(k; d_{ij})$  is a Dirac-delta function ( $\delta = 1$  if  $d_{ij} = k$ , zero otherwise);  $\delta(k; d_{ij}; h_{ij})$  is another Dirac-delta function ( $\delta = 1$  if  $d_{ij} = k$  and  $h_{ij} > 0$ , zero otherwise).

### 3.3 Applicability Domain

On analyzing the model applicability domain from Williams plot, all residuals were located within the range of three Standard deviations, and there is no influential compound both for training or prediction set (Figure.3), which means that the model has a good external predictivity.

### 4. Conclusion

A QSPR model for the estimation of the logarithm of bioaccumulation factor for 58 PCBs was established

According to obtained results it is concluded that the HATS0v can be used successfully for modeling bioaccumulation factor (Log BCF) of the under study compounds. High correlation coefficient (0.9153) and low prediction error (SDEP=0.321; SDEPext=0.319 in Log unit) obtained confirm good predictive ability of the model. The QSPR model proposed with the simply calculated molecular descriptor can be used to estimate bioaccumulation factor for new compounds

### References

[1] National Research Council (1979), Committee on the Assessment of Polychlorinated Biphenyls in the Environment. Polychlorinated biphenyls: a report; National Academy of Sciences: Washington, U.S.A.

- [2] Angulo Lucena, R., Farouk Allam, M., Serrano Jiménez, S. and Luisa Jodral Villarejo, M. A.(2007), "review of environmental exposure to persistent organochlorine residuals during the last fifty years". *Current Drug Safety*,Vol. 2 No.2, pp. 163-172.
- [3] Roveda, A. M., Veronesi, L., Zoni, R., Colucci, M. E. and Sansebastiano, G. (2006), "Exposure to polychlorinated biphenyls (PCBs) in food and cancer risk: recent advances", *Igiene e Sanita Pubblica*,Vol. 62 No.6, pp. 677-696.
- [4] Lundqvist, C.,Zuurbier, M., Leijds, M., Johansson, C.,Ceccatelli, S.,Saunders, M.,Schoeters, G., Ten Tusscher, G. and Koppe, J. G. (2006)," The effects of PCBs and dioxins on child health", *Acta Paediatrica*,Vol.95 No.453,pp.55-64.
- [5] Poppenga, R. H. (2000), "Current environmental threats to animal health and productivity", *The Veterinary Clinics of North America. Food Animal Practice* ,Vol. 16 No.3, pp.545-558.
- [6] Bren, U.,Zupan, M., Guengerich, F. P. and Mavri, J.( 2006)" Chemical Reactivity as a Tool to Study Carcinogenicity: Reaction between Chloroethylene Oxide and Guanine", *The Journal of Organic Chemistry* ,Vol. 71 No.11,pp. 4078-4084.
- [7] Lebeuf, M., Noël, M.,Trottier, S. and Measures, L. (2007), "Temporal trends (1987-2002) of persistent,bioaccumulative and toxic (PBT) chemicals in beluga whales (*Delphinapterus leucas*) from the St.Lawrence Estuary, Canada", *Sciences of the Total Environment*, Vol.383 No. (1-3), pp.216-231.
- [8] Tan, J.,Cheng, S. M., Loganath, A.,Chong, Y. S.and Obbard, J. P. (2007), "Selected organochlorine pesticide and polychlorinated biphenyl residues in house dust in Singapore",*Chemosphere* ,Vol. 68 No.9, pp.1675-1682.
- [9] Borrell, A., Cantos, G., Aguilar, A., Androukaki, E. and Dendrinou, P. (2007) "Concentrations and patterns of organochlorine pesticides and PCBs in Mediterranean monk seals (*Monachus monachus*) from Western Sahara and Greece", *Science of the Total Environment* , Vol.381 No. (1-3), pp.316-325.
- [10] Klánová, J.,Kohoutek, J., Kostrohounová, R. and Holoubek, I. (2007),"Are the residents of former Yugoslavia still exposed to elevated PCB levels due to the Balkan wars?. Part 1: air sampling in Croatia, Serbia, Bosnia and Herzegovina", *Environment International*, Vol. 33 No. 6, pp.719- 726.
- [11] Hansch, C. (1969),"Quantitative approach to biochemical structure-activity relationships", *Accounts of Chemical Research*, Vol.2 No.8, pp.232-239.
- [12] Hansch, C.and Leo, A. (1979),"Substituent Constants for Correlation Analysis in Chemistry and Biology", John Wiley & Sons,New York.
- [13] Voutsas, E., Magoulas, K. and Tassios, D.( 2002), "Prediction of the bioaccumulation of persistent organic pollutants in aquatic food webs", *Chemosphere* ,Vol.48,pp. 645- 651.
- [14] Franke, C. (1996), "How meaningful is the bioconcentration factor for risk assessment?" *Chemosphere*,Vol. 32 No.10,pp. 1897-1905.
- [15] Franke, C., Studinger, G., Berger, G., Bohling, S., Bruckmann, U., Cohors-Fresenborg, D.and Jthnck, U. (1994), "The assessment of bioaccumulation",*Chemosphere* ,Vol.29 No.7,pp.1501-1514.
- [16] Schechter, A., Cramer, P., Boggess, K., Stanley, J., Pöpke, O., Olson, J., Silver, A. and Schmitz, M. (2001), "Intake of dioxins and related compounds from food in the U.S. population". *Journal of Toxicology and Environmental Health, Part A*,Vol. 63 No.1,pp. 1-18.
- [17] Barron, M. G. (1990), "Bioconcentration. Will waterborne organic chemicals accumulate in aquatic animals? ", *Environmental science & technology*. Vol.24, pp. 1612-1618.
- [18] Bintein, S., Devillers, J. and Karcher, W. (1993), "Nonlinear dependence of fish bioconcentration on n-octanol/water partition coefficient", SAR and QSAR in Environmental Research,Vol. 1, pp.29-39.
- [19] Isnard, P. and Lambert, S. (1988)," Estimating bioconcentration factors from octanol-water partition coefficient and aqueous solubility", *Chemosphere*,Vol. 17,pp. 21-34.
- [20] Mackay, D. (1982), "Correlation of bioconcentration factors", *Environmental Science &Technology*,Vol. 16,pp.274-278.
- [21] Dimitrov, S. D., Mekenyan, O. G. and Walker, J. D. (2002), "Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals",SAR and QSAR in Environmental Research,Vol. 13 No.1, pp.177-184.
- [22] Connell, D. W. and Hawker, D. W. (1988)," Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals by fish",*Ecotoxicology & Environmental Safety*, Vol.16 No.3, pp.242-257.
- [23] Ivanciuc, T., Ivanciuc, O. and Klein, D. J. (2006), "Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quantitative super-structure/activity relationships (QSSAR)", *Molecular Diversity*,Vol. 10,pp.133-145.
- [24] Hyperchem™ (2000), Release 7, Hypercube for Windows, Molecular Modeling System.
- [25] Todeschini, R., Consonni, V.and Pavan, M. (2006),DRAGON Software for the Calculation of Molecular Descriptors. Release 5.4 for Windows,Taete s.r.l.,Milano, Italy.
- [26] Leardi, R., Boggia, R. and Terrile, M. (1992) "Genetic algorithms as a strategy for feature selection", *Journal of Chemometrics*, Vol.6, pp. 267-281.
- [27] Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. and Pavan, M. (2009), "MOBYDIGS, Software for Multilinear Regression Analysis and Variable Subset

Selection by Genetic Algorithm”, Release 1.1 for windows, Milano, Italy.

[28] Kennard, R. and Stone, L.A (1969), Technometrics, Vol.11, p.137.

[29] 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 and Combinatorial Science, Vol.22, pp.69.

[30] 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”, Chemometrics and Intelligent Laboratory Systems, Vol.33, pp. 35-46.

[31] MOBYDIGS – Models BY Descriptors In Genetic Selection – ver. 1.1 for Windows, Talete S.r.l., Milano, Italy.

[32] 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.

[33] 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

[34] Kubinyi H., Hamprecht F.A. & Mietzner T., 1998. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices, Journal of Medicinal Chemistry, Vol.41(14), 2553-2564.

[35] Golbraikh, A. and Tropsha, A. (2002), “Beware of  $q^2!$ ”, Journal of Molecular Graphics and Modelling, Vol.20 No.4, pp.269-276.

[36] Ramsey, L. F. and Schafer, W. D. (1997), “The Statistical Sleuth”, Wadsworth Publishing Company, U.S.A.

[37] Consonni, V.; Todeschini, R.; Pavan, M. J. Chem. Inf. Comput. Sci. 2002, 42, 682

[38] Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. J. Chem. Inf. Comput. Sci. 2002, 42, 693.

## ملخص:

تم تطوير ثلاثة نماذج: بنية / زمن استبقاء; بنية / عامل التراكم البيولوجي وبنية / دليل استبقاء لمجموعة من المركبات السامة (المبيدات الحشرية; المركبات متعددة الكلور ثنائية الفينيل; المركبات العطرية المتعددة الحلقات)، باستعمال طريقة الخوارزمية الجينية الهجينة / الانحدار الخطي المتعدد بالنسبة للمتغيرين زمن و دليل الاستبقاء و البسيط بالنسبة للمتغير عامل التراكم البيولوجي، تمت عملية تقسيم قاعدة البيانات (المبيدات الحشرية; المركبات متعددة الكلور ثنائية الفينيل) بواسطة خوارزمية Kennard و Stone; اما بالنسبة للمركبات العطرية المتعددة الحلقات فكان التقسيم عشوائيا، إلى مجموعتين فرعيتين: الاولى استخدمت لبناء النموذج والثانية للتحقق الخارجي من صحة النموذج لكل المتغيرات زمن و دليل الاستبقاء و عامل التراكم البيولوجي، وقد تم حساب الوسائط الاحصائية بواسطة البرنامج المعلوماتية Spartan و Hyperchem و DRAGON. من بين مئات النماذج التي تم الحصول عليها، اخترنا أفضل نموذج من حيث قيمة معامل التنبؤ ( $Q^2$ ) ومعامل التحديد ( $R^2$ ) للمتغيرات زمن و دليل الاستبقاء و عامل التراكم البيولوجي. و تم التحقق من صحة النماذج المقترحة باستخدام تقنيات مختلفة منها التقييم: leave-Many-out، التحقق من صحة الاختبار العشوائي و التحقق من صحة الاختبار المتصالب.

**الكلمات الدالة:** بنية / زمن استبقاء; بنية / عامل التراكم البيولوجي وبنية / دليل استبقاء; المواد السامة; خوارزمية Kennard و Stone الوسائط الاحصائية البرنامج المعلوماتية.



**Résumé:**

Trois modèles : structure / temps de rétention ; structure /facteur de bioaccumulation et structure / indice de rétention ont été recherchés pour des composés toxiques (Pesticides, PCBs,HAPs ), tout en favorisant des approches hybrides : algorithme génétique / régression linéaire multiple pour les temps et les indices de rétention et simple pour le facteur de bioaccumulation , Les trois ensembles des données ont été divisés en deux sous-ensembles: un ensemble de calibrage pour la construction du modèle et un ensemble de test selon l'algorithme Kennard et Stone pour les deux variables temps de rétention et facteur de bioaccumulation et un choix aléatoire pour la variable indice de rétention des HAPs. les paramètres structurels étant calculées avec les logiciels Spartan ;Hyperchem et DRAGON. Parmi une centaine de modèles obtenues, nous avons choisi celle qui présentent les meilleures valeurs du paramètre de prédiction ( $Q^2$ ) et du coefficient de détermination ( $R^2$ ) pour les trois variables à expliquer .La fiabilité du deux modèles proposés a également été illustré en utilisant diverses techniques d'évaluation: leave- Many-out, la procédure de validation croisée, test de randomisation, et la validation par l'ensemble de test.

**Mots-clés:**

Mots-clés: Structure / temps de rétention; structure /facteur de bioaccumulation ; structure / indice de rétention ;composés toxiques; l'algorithme Kennard et Stone ; les paramètres structurels ; logiciels.

## **Abstract**

Three models : structure/retention time; structure /bioaccumulation factor and structure/retention indice relationships were searched for toxicants compounds (Pesticides, PCBs, PAHs) while promoting the hybrid genetic algorithm/multilinear for the retention indice and time and simple regression approach for bioaccumulation factor, the structural parameters being calculated with the softwares Spartan;Hyperchem and DRAGON.The three data set were divided into two subsets: a set of calibration for the model building and a test set according to Kennard and Stone algorithm for both variable retention time and bioaccumulation factor and a random choice for the the variable retention indices.Among about a hundred models gotten, we selected three models that present best values of the prediction parameter ( $Q^2$ ) and of the determination coefficient ( $R^2$ ) for the three variables: retention time and retention indices ; bioaccumulation factor .The reliability of the proposed models was further illustrated using various evaluation techniques : leave-many-out, cross-validation procedure, randomization test, and validation through the test set.

**Keywords:** structure/ retention time; structure /bioaccumulation factor; toxicants compounds; Kennard and Stone algorithm ; structural parameters ;softwares .