



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

MEMOIRE

Présenté à

FACULTE DES SCIENCES – DEPARTEMENT DE CHIMIE

Pour l'obtention du diplôme de :

MASTER

Filière : **Chimie**

Option : Chimie Théorique et Computationnelle

Par :

M^{elle} LARBAOUI Djazia Fatiha

Sur le thème

Etude écotoxicologique des composés organiques sur les espèces aquatiques, accumulation et impact sur la santé humaine. Modélisation QSAR



Soutenu publiquement le 10 Septembre 2020 devant le jury composé de :

Mme CHEMOURI Hafida	Maître de Conférences A	Université de Tlemcen	Président
Mr MEKELLECHE Sidi Mohamed	Professeur	Université de Tlemcen	Examineur
Mr CHARIF Imad Eddine	Maître de Conférences A	Université de Tlemcen	Examineur
Mme BELLIFA Khadidja	Maître de Conférences B	Université de Tlemcen	Encadrant
Mme BENCHOUK Wafaa	Maître de Conférences A	Université de Tlemcen	Co-Encadrant

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

Dédicaces

Je dédie ce mémoire



A ma très chère maman

Aucune dédicace ne pourrait exprimer le degré d'amour et d'affection que j'éprouve pour toi. Tu n'as cessé de me soutenir et de m'encourager durant toutes les années de mes études. Tu as toujours été présente à mes côtés pour me consoler quand il fallait. Tu m'as comblé avec ta tendresse et ton affection. Que Dieu t'accorde santé, bonheur et longue vie afin que je puisse te combler à mon tour.



A mon très cher papa

Ton encouragement, ta compréhension et ta patience sont pour moi un soutien indispensable, tu as su m'inculquer le sens de la responsabilité et la confiance en soi face aux difficultés de la vie. Je te dois ce que je suis aujourd'hui et ce que je serai demain et je ferai toujours de mon mieux pour rester ta fierté et ne jamais te décevoir. J'espère que ce travail sera le fruit de tous tes sacrifices, tes peines et tes efforts. Que Dieu le tout puissant te préserve pour moi, t'accorde santé, bonheur et te protège de tout mal.



A mes chers frères Fethallah, Imad et ma petite sœur Hanene

Puisse Allah vous protéger et renforcer notre fraternité. Je vous souhaite beaucoup de succès, de prospérité et une vie pleine de joie et de bonheur.



A mon très cher fiancé

Je ne saurais exprimer ma profonde reconnaissance pour le soutien continu dont tu as toujours fait preuve. Tu m'as toujours incité à faire de mon mieux et m'encourager. Je prie Dieu tout puissant de préserver notre attachement mutuel, et d'exaucer tous nos rêves.



A mes meilleures amies Zineb et Rym que j'aime tant et avec qui je passe des moments inoubliables. Le destin m'a offert la chance de vous côtoyer, le bonheur que vous m'apportez n'a aucun prix.



A ma chère cousine Chahrazed, en souvenir des moments heureux passés ensemble. Je te dois beaucoup d'affection et amour, que dieu te procure santé et joie et protège ton petit ange Khalil.



A mon cher oncle Hamid et ma tante Lamia, je vous dédie ce travail avec mes vœux de prospérité et de bonheur.



A ma chère cousine Yasmine, pour tous les moments drôles et joyeux passés ensembles.



A la mémoire de mes grands-parents qui auraient tant souhaité assister à ce jour, que dieu vous accueille dans son vaste paradis.



A tous ceux que j'aime.

Merci !

Remerciement

Je souhaite tout d'abord exprimer ma reconnaissance à Mademoiselle NEGADI Latifa professeur et responsable du laboratoire de thermodynamique appliquée et modélisation moléculaire (LATA2M) de l'université A. Belkaid de Tlemcen, là où mon travail a été réalisé.

Je tiens bien évidemment à remercier le département de chimie de m'avoir appris à aimer l'univers de la chimie.

J'exprimer ma profonde gratitude et mes chaleureux remerciements à mon encadreur Mme Errahoui-Bellifa Khadidja qui m'a laissé une large part d'autonomie dans ce travail tout en me mettant sur les bonnes rails de réflexions riches et porteuses, j'ai pu acquérir grâce à son expérience beaucoup de connaissances. Je salue également son amabilité, sa patience, sa disponibilité et sa souplesse d'esprit.

Je voudrais remercier également Mme BENCHOUK Wafaa pour m'avoir fait l'honneur d'être Co-encadreur de ce mémoire et de répondre à tout questionnement.

Je dédie un merci particulier pour monsieur le professeur MEKELLECHE Sidi Mohamed qui a réussi à m'inspirer et me donner l'envie d'apprendre et de toujours tenter de se surpasser.

J'adresse mes sincères remerciements à Mme. CHEMOURI Hafida pour avoir accepté d'être directrice de ce mémoire et pour son temps consacré à l'examiner.

Mes vifs remerciements vont également à monsieur MEKELLECHE Sidi Mohamed et à monsieur CHARIF Imad Eddine d'avoir accepté de participer au jury.

Je remercie tous mes enseignants pour la qualité de l'enseignement qu'ils m'ont fournis au cours de ces années passées à l'université A. Belkaid Tlemcen.

Je tiens également à remercier mes camarades de promotion avec qui j'ai eu le plaisir de partager mes 2 années de master.

Enfin un grand merci à l'ensemble des mes proches qui m'ont aidé et motivé durant ce cursus rempli d'embuches et pour leur soutien au quotidien.

Sommaire

Introduction générale	1
<i>Références bibliographiques</i>	3

Chapitre I : Généralités-Composés Persistants, Bioaccumulables et Toxiques

Introduction.....	4
I.1 Généralités.....	4
I.1.1 Polluant.....	4
I.1.2 Polluants Organiques Persistants.....	5
I.1.3 Les PBTs dans les organismes aquatiques.....	6
1. La persistance.....	7
2. La bioaccumulation.....	7
2.1 La bioconcentration.....	8
2.2 La bioamplification.....	8
3. Facteurs influençant la bioaccumulation.....	9
4. Le potentiel de bioaccumulation.....	9
4.1 Facteur de Bioconcentration (BCF).....	9
4.2 Facteur de bioaccumulation (BAF).....	10
4.3 Coefficient de partage log Kow	10
I.1.4 La Toxicité.....	10
1. Un poison ou un toxique.....	10
2. Toxicité aquatique aiguë.....	11
3. Toxicité aquatique chronique.....	11
I.1.5 Effets des PBT sur l'organisme humain.....	11
I.1.6 Poissons utilisés pour l'étude du BCF/Toxicité des PBTs.....	12
I.2 Etudes QSAR du BCF et de toxicité des PBTs vis-à-vis des poissons.....	13
<i>Références bibliographiques</i>	15

Chapitre II : Bases Théoriques-Partie A : Méthodes QSAR

Introduction.....	17
1. Définition d'une méthode QSAR.....	17
2. Principe des méthodes QSAR.....	17
3. Méthodologie d'une étude QSAR.....	17

Sommaire

3.1 Développement du modèle.....	17
3.1.1 Base de donnée données.....	18
3.1.2 Descripteurs moléculaires.....	18
3.1.2.1 Descripteurs théoriques.....	19
a. Les descripteurs 1D.....	19
b. Les descripteurs 2D.....	19
c. Les descripteurs 3D.....	19
3.1.2.2 Descripteurs empiriques.....	20
3.1.2.2.1 Coefficient de partage (octanol-eau) K_{ow} , $\log P$	20
a. Détermination expérimentale de $\log P$	20
b. Détermination théorique de $\log P$	20
3.2 Validation du modèle.....	20
3.2.1 Validation interne.....	20
3.2.1.1 Validation croisée (Cross Validation CV).....	21
3.2.1.2 Y-Randomisation.....	21
3.2.2 Validation externe.....	21
3.3 Domaine d'applicabilité.....	22
<i>Références Bibliographiques</i>	25

Chapitre II : Bases Théoriques-Partie B : Méthodes d'analyse statistique

Introduction.....	27
1. Méthodes de régression linéaire (SLR, MLR).....	27
1.1 Régression linéaire simple (SLR).....	27
1.2 Régression linéaire Multiple (MLR).....	29
1.3 Signification et qualité de la régression linéaire.....	31
1.3.1 Test de la signification globale de la régression.....	31
1.3.2 Test de signification de chaque descripteur (t-Student).....	32
1.4 Indicateurs de qualité d'une régression linéaire.....	33
2. Méthode des moindres carrés partiels PLS.....	34
2.1 Définition.....	34
2.2 Principe.....	35
2.3 Etapes de la régression PLS.....	35
3. Méthode d'analyse en composante principale (PCA).....	37

Sommaire

3.1	Principe.....	37
3.2	Etape de la PCA.....	37
	<i>Références Bibliographiques.....</i>	<i>39</i>

Chapitre II : Bases Théoriques-Partie C : Méthodes Quantiques

	Introduction.....	40
1.	Approximation de Born-Oppenheimer.....	40
2.	Méthode Hartree-Fock(HF).....	41
2.1	Approximation du champ moyen de Hartree.....	41
2.2	Équations de Hartree-Fock.....	42
3.	Méthode de Hartree-Fock-Roothaan.....	42
4.	Méthodes Post-SCF.....	44
5.	Théorie de la Fonctionnelle de Densité (DFT).....	44
5.1	Théorèmes de Hohenberg-Kohn.....	45
5.1.1	Premier théorème.....	45
5.1.2	Second théorème.....	46
5.2	Approche de Kohn-Sham.....	47
5.3	Approximation de la densité locale (LDA:Local Density Approximation).....	48
5.4	Approximation de la densité de spin locale (LSDA).....	48
5.5	Approximation du gradient généralisé (GGA).....	48
5.6	Fonctionnelle hybride B3LYP.....	49
	<i>Références Bibliographiques.....</i>	<i>50</i>

Chapitre III : Résultats et discussion

	Introduction.....	51
	Méthodologie.....	51
III.1	Etude du facteur de bioconcentration.....	57
III.1.1	Modèle avec un seul descripteur.....	57
III.1.2	Modèles avec plusieurs descripteurs.....	59
III.1.2.1	Modèles avec deux descripteurs.....	59
III.1.2.2	Modèles avec trois descripteurs.....	60
III.1.3	Validation du meilleur modèle.....	60
III.1.3.1	Validation interne.....	60

Sommaire

III.1.3.2 Validation externe.....	62
III.1.4 Domaine d'applicabilité.....	63
III.1.5 Analyse en composantes principales PCA.....	64
III.1.6 Interprétation mécanistique du meilleur modèle QSAR.....	67
Conclusion.....	68
III.2 Etude de la toxicité.....	69
III.2.1 Elaboration des modèles.....	70
III.2.2 Validation du meilleur modèle QSAR.....	71
III.2.2.1 Validation interne.....	71
III.2.2.2 Validation externe.....	71
III.2.3 Domaine d'applicabilité.....	73
III.2.4 Interprétation mécanistique du meilleur modèle.....	74
Conclusion.....	75
<i>Références Bibliographiques.....</i>	<i>76</i>
Conclusion générale.....	77

LISTE DES ABREVIATIONS

BAF	BioAccumulation Factor
BCF	BioConcentration Factor
BMF	BioMagnification Factor
B3LYP	Becke3-ParameterLee-Yang-Parr
DA	Applicability Domaine
DFT	Density Functional Theory
GGA	Generalized Gradient Approximation
HF	Hartree-Fock
HOMO	Highest Occupied Molecular Orbital
IC₅₀	Inhibition Concentration
IGC₅₀	Inhibition Growth Concentration
LDA	Local Density Approximation
LD₅₀	Lethal Dose
LC₅₀	Lethal Concentration
LMO	Leave Many Out
LOO	Leave One Out
LUMO	Lowest Unoccupied Molecular Orbital
MLR	Multiple Linear Regression
OA	Orbitale Atomique
OECD	Organization for Economic Cooperation and Development
OM	Orbitale Moléculaire
PBT	Persistent Bioaccumulable Toxique
PCA	Principal Composante Analysis
PLS	Partial Least Squares
POP	Persistent Organic Polluant
QSAR	Quantitative Structure-Activity Relationships
SCF	Self Consistent Field
SLR	Simple Linear Regression
SSE	Sum of Squares Error
SSR	Sum of Squares Regression
SST	Sum of squares Total

Introduction générale

Introduction générale

L'un des facteurs le plus prépondérant contribuant principalement au développement de nouveaux produits chimiques est l'évolution du mode de vie. En effet, le confort et le bien-être que procure l'industrie chimique par la diversité de ses activités a entraîné le recours à de nouvelles molécules, qui sont perpétuellement synthétisées et mises sur le marché en vue d'applications les plus diverses.

A ce jour, de nombreux polluants organiques entrent dans la composition de produits industriels. Les diverses activités anthropiques peuvent générer des quantités non négligeables de polluants chimiques indésirables dans tous les compartiments de l'environnement, dans l'air, le sol, l'eau, le sédiment et le biote. Suite à des utilisations massives et variées, ces polluants sont émis et dispersés dans l'environnement et dans différentes étapes du cycle de l'eau. Ces polluants présentent un risque sanitaire et environnemental majeur pour les espèces aquatiques et pour la santé humaine. Le fonctionnement des écosystèmes peut être déséquilibré et certains milieux deviennent alors très vulnérables à la réception de ces polluants. Les hydrocarbures aromatiques (HA), les polychlorobiphényles (PCB), les pesticides, les phtalates issus des matières plastiques, les phénols, les produits pharmaceutiques et d'hygiène corporels (PPCP) figurent parmi les polluants organiques toxiques, émergents et/ou prioritaires [1]. Ces substances sont préoccupantes, en raison de leur résistance à la biodégradation et leur capacité à persister durant plusieurs années et se déplacer sur de longues distances dans l'atmosphère avant de se déposer. Elles ont généralement une faible solubilité dans l'eau (lipophile), ce qui entraîne leur bioaccumulation dans les tissus adipeux et donc peuvent s'accumuler dans les organismes et puis dans la chaîne alimentaire. Cela a conduit à une préoccupation internationale croissante pour identifier et étudier les produits chimiques ayant un potentiel de persistance, de bioaccumulation et de toxicité (PBTs) [2].

La tendance des produits chimiques à se bioconcentrer/s'accumuler dans les espèces aquatiques est généralement exprimée par le facteur de bioconcentration (BCF), défini comme le rapport entre la concentration chimique dans l'espèce et celle de son environnement à l'état d'équilibre [3]. En règle générale, les poissons sont utilisés pour les évaluations du BCF [4].

Le facteur de bioconcentration (BCF) est utile pour caractériser le comportement environnemental d'un produit chimique, en particulier pour voir s'il a un effet cumulatif. L'accumulation de polluants organiques dans l'écosystème aquatique est particulièrement préoccupante car les poissons servent de nourriture à de nombreuses espèces, y compris les humains.

Cependant, dans les évaluations des risques humains et environnementaux, les données BCF indispensables ne sont généralement pas facilement disponibles [5]. Étant donné que la détermination expérimentale des valeurs du BCF est coûteuse et prend du temps, de nombreux chercheurs ont tendance à utiliser des méthodes d'estimation pour l'étude de l'accumulation (la bioconcentration) des composés PBTs.

Le développement croissant de la technologie des moyens informatiques a permis à la chimie de s'enrichir d'outils de simulation numérique et de modélisation moléculaire. Ceci peut surmonter les problèmes rencontrés dans l'expérience en faisant appel aux études théoriques par modélisation QSAR.

Jusqu'à présent, de nombreux travaux de recherche [6-9] ont été menés pour simuler et prédire le facteur de bioconcentration par différentes voies, et la plupart des QSAR ont été dérivées de $\log K_{ow}$. Cependant, tous les BCF n'avaient pas toujours une bonne corrélation avec le K_{ow} pour tous types de composés [10]. En outre, la robustesse du modèle obtenu est l'un des problèmes restants [11-13].

L'objectif de ce présent travail est de réaliser une étude QSAR du facteur de bioconcentration BCF et de toxicité d'une série de composés persistants vis-à-vis les poissons en respectant toute la méthodologie QSAR et tous les critères d'OECD [14].

Le travail présenté dans ce manuscrit est articulé sur trois chapitres :

- Le premier chapitre englobe des généralités sur les composés Persistants Bioaccumulables et Toxiques (PBTs).
- Le second chapitre est consacré à une étude bibliographique sur les méthodes utilisées et se divise en 3 parties :
 - ✓ Méthodes QSAR.
 - ✓ Méthodes d'analyse statistique.
 - ✓ Méthodes de chimie quantique.
- le troisième chapitre sera dédié aux résultats obtenus et leurs discussions.

Enfin, nous clôturons ce mémoire par une conclusion générale.

Références bibliographiques

- [1] D. Sopheak Net, Contaminants organiques en milieux aquatiques : développements analytiques, techniques et applications, Thèse doctorat, Université de Lille 1, **2016**.
- [2] A.A. Toropov, A.P. Toropova, A. Lombardo, A. Roncaglioni, E. Benfenati and G. Gini, CORAL: Building up the model for bioconcentration factor and defining its applicability domain, European Journal of Medicinal Chemistry. 46, 1400-1403, **2011**.
- [3] X. Lu, S. Tao, H. Hu and R.W. Dawson, Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors, Chemosphere. 41, 1675-1688, **2000**.
- [4] T.H. Miller, M.D. Gallidabino, J.I. Macrae, S.F. Owen, N.R. Bury and L.P. Barron, Prediction of bioconcentration factors in fish and invertebrates using machine learning, Science of the Total Environment. 648, 80-89, **2019**.
- [5] P.N.H. Wassenaar, E. Verbruggen, E. Cieraad, W. Peijnenburg and M. Vijver, Variability in fish bioconcentration factors: Influences of study design and consequences for regulation, Chemosphere. 239, 124-133, **2020**.
- [6] P. Gramatica and E. Papa, QSAR Modeling of Bioconcentration Factor by theoretical molecular descriptors, QSAR Combinatorial Science. 22, 374-385, **2003**.
- [7] R. Garg and J. Smith, Predicting the bioconcentration factor of highly hydrophobic organic chemicals, Food and Chemical Toxicology. 69, 252-259, **2014**.
- [8] F. Grisoni, V. Consonni, S. Villa, M. Vighi and R. Todeschini, QSAR models for bioconcentration: is the increase in the complexity justified by more accurate predictions, Chemosphere. 127, 171-179, **2015**.
- [9] A. Kumar, P. Singh and V. Kumar, DFT-Based Prediction of Bioconcentration Factors of Polychlorinated Biphenyls in Fish Species Using Molecular Descriptors, Advances in Biological Chemistry. 10, 1-15, **2020**.
- [10] P. Gramatica and E. Papa, An update of the BCF QSAR Model Based on Theoretical Molecular Descriptors, QSAR and Combinatorial Science. 24, 953-960, **2005**.
- [11] X. Lu, S. Tao, H. Hu and R. Dawson, Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors, Chemosphere. 41, 1675-1688, **2000**.
- [12] A. Gissi, A. Lombardo, A. Roncaglioni, D. Gadaleta, G. Nicolotti and E. Benfenati, Evaluation and comparison of benchmark QSAR models to predict a relevant REACH endpoint: The bioconcentration factor (BCF), Environmental Research. 137, 398-409, **2015**.
- [13] M. Nendza, R. Kühne, A. Lombardo, S. Stempel and G. Schüürmann, PBT assessment under REACH: Screening for low aquatic bioaccumulation with QSAR classifications based on physicochemical properties to replace BCF in vivo testing on fish, Science of the Total Environment. 616, 97-106, **2018**.
- [14] OECD, Guidance Document on the Validation of (Quantitative) Structure-activity Relationships QSAR Models, OECD Environment Health and Safety Publications, **2007**.

Chapitre I

Généralités-Composés Persistants

Bioaccumulables Toxique (PBTs)

Introduction

Au cours des derniers siècles, l'accroissement exponentiel de la population associé à l'urbanisation et à l'industrialisation ont impacté sévèrement l'environnement. En effet, cette croissance engendre une hausse des activités industrielles, urbaines et agricoles et s'accompagne généralement d'une augmentation considérable des polluants et en particuliers les polluants persistants bioaccumulables et toxiques (PBTs), qui constituent par leur présence un risque sanitaire et environnemental majeur.

La pollution de l'environnement et en particulier de l'eau est un réel problème qui ne cesse de prendre de l'importance partout dans le monde. L'estimation de la bioaccumulation est très importante dans l'évaluation scientifique des risques que constituent ces substances pour l'environnement et pour la santé humaine et font l'objet de la préoccupation actuelle des efforts de réglementation [1] et l'ardeur des chercheurs quant à la dangerosité potentielle de ces produits.

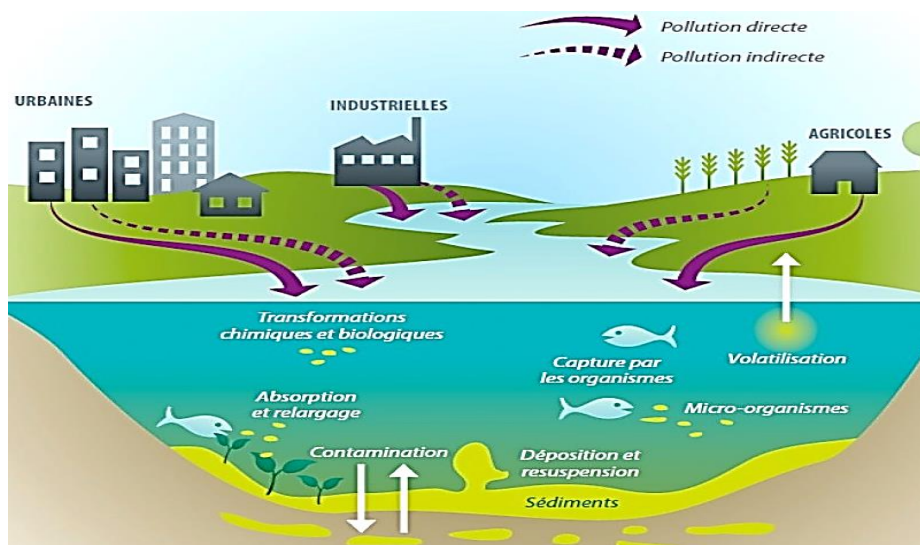


Figure 1 : Les PBTs dans le milieu aquatique.

I.1 Généralités

I.1.1 Polluant

Une substance polluante (contaminante) peut être définie comme une substance qui, au-delà d'un certain seuil, a un impact négatif sur un être vivant, un milieu ou d'une manière générale l'environnement.

Les polluants peuvent être classés notamment selon le compartiment qu'ils affectent (air, eau, sol), selon le seuil de concentration impactant (micro ou macro-polluants), ou selon leur durabilité dans le milieu [2].

On peut distinguer également :

- Les substances polluantes présentes à l'état naturel :
 - Gaz des volcans
 - Minéraux dans les sols et les eaux
 - Produits de dégradation
 - Hydrocarbures (incendies)
- Les polluants d'origine anthropique (produits ou sous-produits de synthèse) :
 - Activités industrielles (chimie, métallurgie, pharmacie, papeterie, ...)
 - Production d'énergie (nucléaire, pétrole, gaz, charbon)
 - Activités du secteur agricole (engrais, pesticides...)
 - Activités domestiques
 - Décharges
 - Pollution urbaine

I.1.2 Polluants Organiques Persistants

Les polluants organiques persistants (POPs) peuvent être produits de manière **intentionnelle** (industrie du pesticide par exemple) ou **non intentionnelle** (par la combustion de biomasse et l'incinération de déchets par exemple) [3].

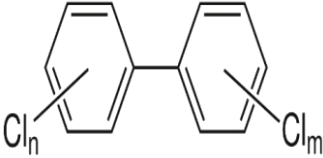
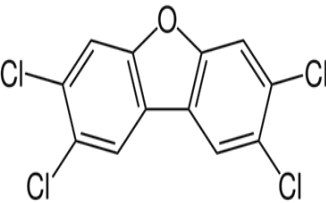
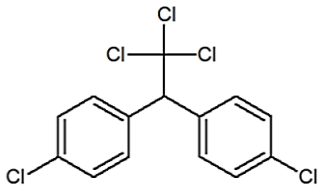
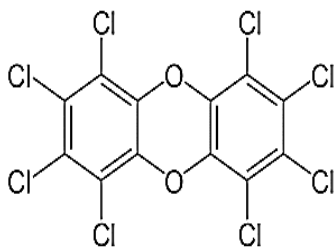
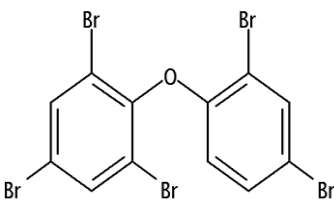
Les POPs constituent un groupe de polluants qui sont regroupés non pas en fonction de leurs propriétés chimiques mais plutôt parce qu'ils vérifient quatre grandes propriétés (qui sont détaillées ci-dessous) :

- 1) **Persistance** : les POPs se dégradent lentement dans les différents milieux.
- 2) **Transport longue-distance** : on peut retrouver des concentrations non négligeables de ces polluants très loin de leurs sources d'émissions.
- 3) **Bioaccumulation** : les POPs ont tendance à s'accumuler tout au long de la chaîne trophique.
- 4) **Toxicité** : l'exposition à ces substances peut causer des effets nocifs.

I.1.3 Les PBTs dans les organismes aquatiques

(Persistance, bioaccumulation, bioconcentration et bioamplification)

Tableau 1 : Exemples des PBTs.

Noms	structures	utilisations
PCB (PolyChloroBiphényles)		<ul style="list-style-type: none"> * Plastifiants * Peintures * Combustion des déchets médicaux * Produits de soudure * Adhésifs
PCDF(PolyChloroDibenzofuranes)		<ul style="list-style-type: none"> * Pesticides * Fluides réfrigérants
DDT(1,1,1-trichloro-2,2-bis(4-chlorophenyl)éthane)		<ul style="list-style-type: none"> * Pesticides * Insecticides
Les dioxines (PCDD) (Les PolyChloroDibenzodioxines)		<ul style="list-style-type: none"> * Combustions incomplètes * incinérateurs défoliants
PBDE (DiphénylEthersPolyBromés)		<ul style="list-style-type: none"> * retardateurs de flammes

1. La persistance

La persistance est la résistance d'une substance chimique à toutes les dégradations, chimique, biologique ou photolytique (décomposition chimique par la lumière) dans l'environnement. Elle est mesurée par la demi-vie globale dans le milieu correspondant et en tenant compte de tous les modes de dégradation.

La demi-vie peut être remplacée par DT_{50} qui désigne le temps de disparition de la substance dans le compartiment [4].

Tableau 2 : Demi-vie des POPs dans l'eau et dans les sédiments [5].

<i>Milieu</i>	<i>Temps de demi-vie</i>
Eau	>180 jours
Sédiments	>360-720 jours

2. La bioaccumulation

La bioaccumulation désigne la capacité des organismes à concentrer et accumuler des substances chimiques à des concentrations supérieures à celles dans l'eau [6], via toutes les voies d'exposition possible : eau, sédiments et ingestion de nourriture [7].

La bioaccumulation est un phénomène dynamique qui résulte de nombreux processus biologiques propres à chaque espèce : la respiration et l'alimentation agissent sur **l'apport** de contaminant à l'organisme ; tandis que l'excrétion, la métabolisation (biotransformation), la reproduction et la dilution par la croissance participent à **l'élimination** du contaminant [8].

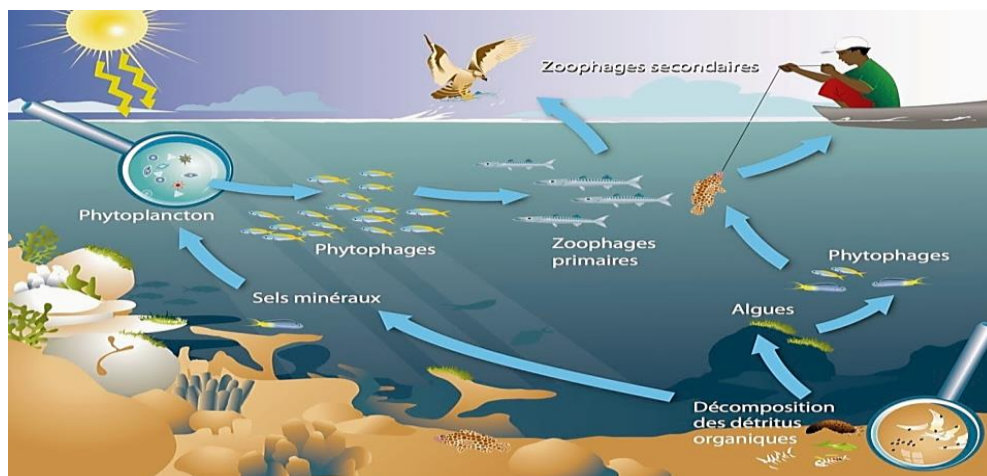


Figure 2 : La bioaccumulation (Bioconcentration + bioamplification) dans le milieu aquatique.

La bioaccumulation est une combinaison de **bioconcentration** et de **bioamplification** [9].

2.1 La bioconcentration

La bioconcentration est le processus par lequel la concentration chimique dans un organisme aquatique atteint un niveau qui dépasse celui dans l'eau à la suite de son exposition à un produit chimique dans l'eau mais n'inclut pas l'exposition via l'alimentation [9].

2.2 La bioamplification

Dans le cas des molécules persistantes et peu métabolisées, le processus de bioaccumulation se réitère à chaque niveau trophique, on parle alors de bioamplification [2].

Le terme bioamplification désigne l'accumulation de substances chimiques, le long de la chaîne trophiques, via la nourriture. Elle peut être constatée, en relevant une corrélation entre le niveau trophique et la teneur en contaminant le long de la chaîne alimentaire [2].

Pour les contaminants persistants, la bioamplification se traduit donc par des concentrations, dans les organismes de hauts niveaux trophiques importants, pouvant dépasser les seuils admissibles pour l'Homme. En étant le consommateur final de nombreux produits de la mer, l'Homme est en effet le récepteur final des contaminants bioaccumulés le long des chaînes trophiques marines [2].

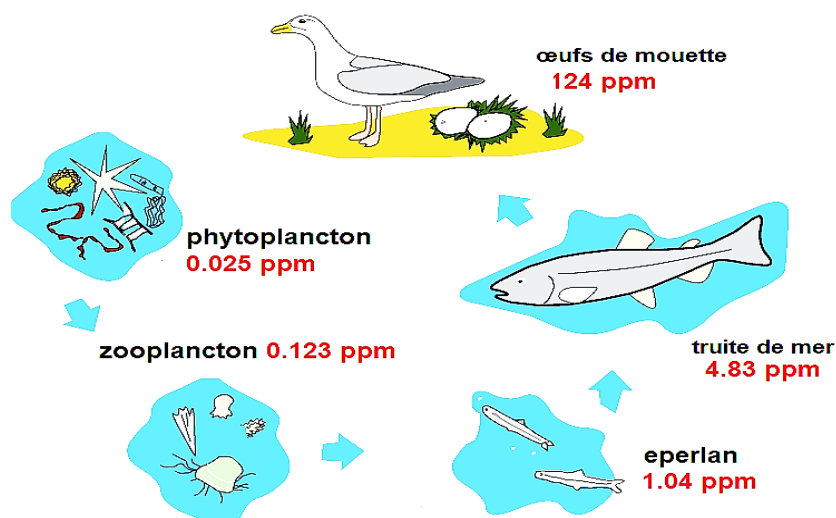


Figure 3 : La bioamplification dans l'écosystème aquatique.

3. Facteurs influençant la bioaccumulation

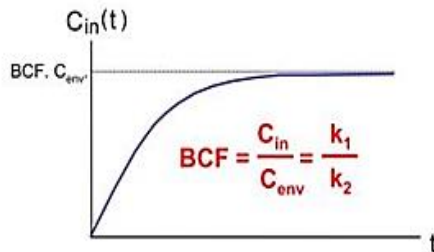
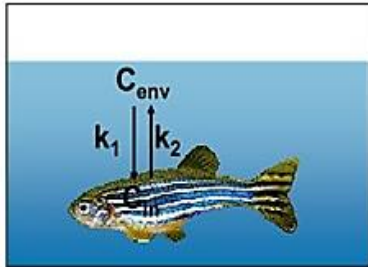
La bioaccumulation dépend :

- ▶ De la composition de l'eau (sa dureté par exemple), du pH et de la teneur en carbone organique dissous ou sous forme particulaire, qui diminue la disponibilité de la substance pour l'organisme testé.
- ▶ De l'organisme, en particulier, sa capacité à métaboliser la substance chimique, ainsi que sa teneur en graisses.
- ▶ De la substance chimique (PBT) en particulier ses propriétés physico-chimiques (comme sa solubilité dans l'eau et dans les graisses) ainsi que la facilité avec laquelle elle peut traverser les membranes biologiques et être métabolisée ou dégradée.

4. Le potentiel de bioaccumulation

Ce potentiel peut être mesuré par trois facteurs :

4.1 Facteur de Bioconcentration (BCF)



$$\frac{dC_B}{dt} = K_1 C_W - K_2 C_B \quad (1)$$

K_1 et K_2 désignent les constantes de vitesse d'accumulation et d'élimination de contaminants [8].

L'équation (1) traduit un bilan de masse entre apport et élimination par l'organisme. Une fois l'équilibre atteint, la concentration dans l'organisme reste constante et le facteur de bioconcentration (BCF) sera défini comme suit :

$$BCF = C_B / C_W = K_1 / K_2 \quad (2)$$

- C_B : la concentration en contaminant dans l'organisme.
- C_W : la concentration en contaminant dissous dans l'eau.

4.2 Facteur de bioaccumulation (BAF)

Le facteur de bioaccumulation (BAF) est considéré comme la combinaison du facteur de bioconcentration (BCF), qui représente l'accumulation par un organisme d'une substance chimique uniquement via l'eau et du facteur de bioamplification (BMF), qui représente l'accumulation par un organisme d'une substance chimique via la nourriture. Le BMF sera alors dépendant de la complexité de la chaîne trophique (eau douce ou eau marine) et de la position de l'organisme étudié dans la chaîne alimentaire [7]. De ce fait, il peut être calculé selon l'équation suivante :

$$BAF = BCF \times \prod_{i=1}^n BMF_i \quad (3)$$

Le BAF s'exprime sur le poids humide de tissus des organismes : en L.kg⁻¹ P.H.

4.3 Coefficient de partage log K_{ow}

Certains contaminants peuvent avoir une affinité plus forte pour l'eau que pour la matière vivante, ils ne sont donc pas ou peu bioaccumulables et sont à écarter [2].

Un des paramètres déterminant la répartition d'un contaminant entre l'eau et les autres compartiments est le caractère lipophile, décrit par le coefficient de partage octanol-eau (K_{ow}), qui correspond au rapport entre la concentration en contaminant dans l'octanol et celle dans l'eau d'une solution octanol/eau à l'équilibre [2].

$$K_{ow} = \frac{C \text{ substance dans l'octanol}}{C \text{ substance dans l'eau}} \quad (4)$$

Le K_{ow} est généralement exprimé sous la forme de son logarithme en base 10, log K_{ow} [2]. Plus le **log K_{ow}** de la substance étudiée est grand, plus le composé est lipophile (affinité pour les tissus vivants) et donc susceptible de se bioaccumuler [2].

I.1.4 La Toxicité

1. Un poison ou un toxique

Un poison, ou toxique, est une substance capable de perturber le fonctionnement normal d'un organisme vivant. Il peut être de source naturelle, ou de nature chimique ou biologique [10].

Les effets toxiques peuvent être classés de différentes façons, selon, par exemple :

- La durée : aiguë, chronique
- Le type d'action : locale, systémique,...
- Le mécanisme d'action : stimulant, inhibiteur, ...
- La voie de pénétration : respiratoire, cutanée, digestive,...
- Le tissu ou l'organe affecté : sang (hématotoxique), foie (hépatotoxique), rein (néphrotoxique), le système nerveux (neurotoxique) [10].

2. Toxicité aquatique aiguë

Signifie la propriété intrinsèque d'une substance de provoquer des effets néfastes sur des organismes lors d'une exposition de courte durée (minutes, heures, jours).

3. Toxicité aquatique chronique

Désigne les propriétés potentielles ou réelles d'une substance de provoquer des effets néfastes sur des organismes aquatiques, qu'après un temps d'exposition relativement long et de façon permanente (semaines, mois, années).

La toxicité peut être quantifiée par :

- **DL₅₀** : Dose Létale, correspond à la dose d'une substance pouvant causer la mort de 50 % d'une population animale.
- **CL₅₀** : Concentration Létale, désigne les concentrations du produit chimique qui causent la mort de 50 % des animaux au cours de la période d'observation.
- **IC₅₀** : Concentration Inhibitrice, désigne la concentration à laquelle un toxique capable d'inhiber un processus particulier de 50 %.
- **IGC₅₀** : Concentration Inhibitrice de la Croissance, la concentration qui peut inhiber la croissance de 50% de population.

I.1.5 Effets des PBTs sur l'organisme humain

L'exposition aux PBTs augmente le risque de:








- Troubles du système nerveux.
- Tumeur, diabète, hypertension, perturbateur endocrinien.
- Problèmes cardiovasculaire, problèmes de reproduction
- Résistance à l'insuline chez les non-diabétiques.

- Troubles métaboliques à partir de faibles doses.
- Dommages mutagènes à l'ADN [11].

I.1.6. Poissons utilisés pour l'étude du BCF/Toxicité des PBTs

Plusieurs poissons sont utilisés et recommandés par l'OECD [12] pour l'étude de la bioaccumulation des produits organiques dans l'écosystème aquatique.

Tableau 3 : Poissons recommandés par l'OECD.

Nom	Nom scientifique Latin	Image
Rainbow trout	<i>Salmo Gairdneri</i>	
Bluegill	<i>Lepomis Macrochirus</i>	
Guppy	<i>Poecilia Reticulata</i>	
Carp	<i>Cyprinus Carpio</i>	
Fathead minnow	<i>Pimephalse Promelas</i>	
Red killifish	<i>Oryzias Latipes</i>	
Zebrafish	<i>Brachydanio Rerio</i>	

I.2 Etudes QSAR du BCF et de toxicité des PBTs vis-à-vis des poissons

Aujourd'hui, les QSAR s'établissent et sont acceptés pour l'évaluation des risques des produits chimiques. Plusieurs études récentes décrivent l'utilisation des QSAR pour l'évaluation des risques environnementaux et toxicologiques.

Un grand nombre de travaux sur la toxicité et l'accumulation quantifiée par le facteur de bioconcentration en utilisant la méthode QSAR est publié [13-16].

Kabiruddin Khan et al. [13] ont étudié la toxicité des biocides vis à vis les poissons et les daphnies (*Daphnia magna*) selon les principes de modélisation QSAR recommandés par l'OCDE. Les modèles ont été développés à l'aide de descripteurs 2D (les indices constitutionnels et topologiques) simples et interprétables et ont été calculés à l'aide du logiciel DRAGON (version 7) [17].

Les modèles dérivés sont statistiquement solides et testés à l'aide de diverses méthodes de validation (coefficient de détermination R^2 (0,800 et 0,648), validation croisée R_{cv}^2 (0,760 et 0,602) et prédiction R^2_{pred} (0,875 et 0,817)) pour les jeux de données de toxicité sur les poissons (nTraining = 66, nTest = 22) et *Daphnia magna* (nTraining = 100, nTest = 33), montrant leurs excellentes performances pour la prédiction de composés externes dans le domaine d'applicabilité.

A.A. Toropov et al. [14] ont construit un modèle dépendant du facteur de bioconcentration (log BCF) et ont défini son domaine d'applicabilité en utilisant des descripteurs basés sur SMILES avec le logiciel CORAL (CORrelation And Logic). Pour améliorer le modèle, la prédiction et la fiabilité sur de nouveaux composés, ils ont introduit une nouvelle fonction, qui utilise le Delta (obs) = log BCF (expr) - log BCF (calc) et les résultats statistiques sont donnés comme suit :

n = 502, $R^2 = 0.553$, $Q^2_{LOO} = 0.5500$, RMSE = 0.897, F = 620 (série d'apprentissage).

n = 322, $R^2 = 0.7780$, $R^2_{pred} = 0.7751$, RMSE = 0.627, F = 1122 (série de calibration).

n = 165, $R^2 = 0.8277$, $R^2_{pred} = 0.8241$, RMSE = 0.545, F = 783 ; k = 1.0321, k' = 0.9206, $R^2_m = 0.795$ (série de test)

Ils ont ensuite éliminé les valeurs aberrantes et cela a augmenté la prédiction du modèle.

X. Lu et al. [15] ont développé un modèle du facteur de bioconcentration (BCF) pour une large gamme de composés organiques non ioniques chez les poissons, basé sur les indices de connectivité moléculaire et sur les facteurs de correction de polarité. La modélisation topologique non linéaire utilisant des facteurs de correction de polarité a abouti à la meilleure qualité d'estimation du BCF pour tous les 239 composés étudiés, avec une erreur d'estimation absolue moyenne de 0,478 unités log.

Les indices de connectivité moléculaire se sont avérés être de bons descripteurs du BCF pour les composés non polaires, mais pas pour les composés polaires. Lorsque des facteurs de correction de polarité ont été introduits dans le modèle de connectivité moléculaire linéaire, l'estimation du BCF pour les composés polaires a été beaucoup plus élevée.

La comparaison statistique entre le modèle basé sur le MCI (indices de connectivité moléculaire) et un modèle basé sur le Kow (coefficient de partage octanol / eau) a révélé que l'estimation du BCF basée sur les paramètres topologiques était aussi bonne que celle obtenue par Kow.

T.H. Miller et al. [16] ont développé plusieurs outils de modélisation basés sur l'apprentissage automatique pour la prédiction du facteur de bioconcentration (BCF) chez les poissons (*Cyprinus carpio*) et les invertébrés (*Gammarus pulex*).

14 descripteurs moléculaires ont été utilisés comprenant 6 descripteurs topologiques, 4 descripteurs constitutionnels, 3 descripteurs électrotopologiques et une propriété physico-chimique. Ceux qui ont influencé le plus le facteur de bioconcentration étaient la masse moléculaire (Mw), le coefficient de distribution octanol-eau (log D), la surface polaire topologique (TPSA) et le nombre d'atomes d'azote (nN).

Parmi les 24 modèles construits pour 352 composés, le meilleur a été appliqué pour la prédiction du facteur de bioconcentration (BCF) chez les poissons et les invertébrés d'eau douce, *Gammarus pulex*. Le modèle pour *G. pulex* a montré de bonnes performances avec R^2 de 0,99 et 0,93 pour les données de vérification et de test, respectivement, R^2 et l'erreur quadratique moyenne (RMSE) pour les données de test (n = 110 cas) variaient respectivement de (0.23, 0.73) et (0.34, 1.20).

Références bibliographiques

- [1] J.A. Arnot and F.A.P.C. Gobas, A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms, *Environmental Reviews*. 14, 257-297, **2006**.
- [2] C. Meinesz, Contamination chimique des chaînes trophiques marines. Recommandations pour un futur réseau de surveillance sur la façade méditerranéenne, 1-45, **2011**.
- [3] V. Loizeau, Prise en compte d'un modèle de sol multi-couches pour la simulation multi-milieu à l'échelle européenne des Polluants Organiques Persistants, Thèse Doctorat, Université Paris-Est, **2014**.
- [4] A. Surchamp, Emissions potentielles de polluants organiques persistants à partir du milieu urbain et par les activités de traitement des déchets : Impact sur la qualité de l'air au voisinage des sources, Thèse Doctorat, Université Pierre et Marie Curie, **2016**.
- [5] R. Papp, Evaluation de l'impact sur la santé et sur l'environnement des sites industriels, collège nationale d'experts en environnement et l'industrie chimique, **2010**.
- [6] P. Grégoire, N.E. Abriak and A. Zri, Bioaccumulation dans les tissus des espèces marines fréquentant les sites d'immersion, *Déchets, sciences et techniques* N° 51, 11-17, **2008**.
- [7] A. Isabelle and S. Aliz, Méthodologie de détermination d'un facteur de bioaccumulation (BAF) sur les mollusques en milieu marin. BAF opérationnel déterminé dans le contexte DCE, Ifremer, **2016**.
- [8] A. Abarnou, Approche méthodologique permettant la correspondance entre mesure de contaminants dans l'eau et mesure dans d'autre matrice intégratives, Ifremer, **2012**.
- [9] J.A. Arnot and F.A.P.C. Gobas, A Generic QSAR for Assessing the Bioaccumulation Potential of Organic Chemicals in Aquatic Food Webs, *QSAR & Combinatorial Science*. 22, 337-345, **2003**.
- [10] G. Lapointe, *Notion de toxicologie*, Deuxième édition revue et augmentée, Québec, **2004**.
- [11] I. Lessigiarska, A.P. Worth ORTH, B. Sokull-kluttgen, S. Jeram, J.C. Dearden, T.I. Netzeva and M.T.D. Cronin, QSAR investigation of a large data set for fish, algae and daphnia toxicity, *SAR and QSAR in Environmental Research*. 15, 413-431, **2004**.
- [12] OECD. Bioaccumulation in fish: aqueous and dietary exposure. OECD guidelines for testing of chemicals. Paris: OECD. **2012**, <http://www.oecdilibrary>

- [13] K. Khan, P.M. Khan, G. Lavado, C. Valsecchi, J. Pasqualini, D. Baderna, M. Marzo, A. Lombardo, K. Roy and E. Benfenati, QSAR modeling of *Daphnia magna* and fish toxicities of biocides using 2D descriptors, *Chemosphere*. 229, 8-17, **2019**.
- [14] A.A. Toropov, A.P. Toropova, A. Lombardo, A. Roncaglioni, E. Benfenati and G.Gini, CORAL: Building up the model for bioconcentration factor and defining its applicability domain, *European Journal of Medicinal Chemistry*. 46, 1400-1403, **2011**.
- [15] X. Lu, S. Tao, H. Hu and R.W. Dawson, Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors, *Chemosphere*. 41, 1675-1688, **2000**.
- [16] T.H. Miller, M.D. Gallidabino, J.I. Macrae, S.F. Owen, N.R. Bury and L.P. Barron, Prediction of bioconcentration factors in fish and invertebrates using machine learning, *Science of the Total Environment*. 648, 80-89, **2019**.
- [17] R. Todeschini, V. Consonni, A. Mauri and M. Pavan, DRAGON-software for calculation of Molecular Descriptors, **2004**.

Chapitre II
Bases Théoriques

Chapitre II

Partie A-Méthodes QSAR

Introduction

Le coût, le temps nécessaire et la disponibilité des laboratoires équipés pour la réalisation des synthèses et des tests rendent l'expérience particulièrement difficile, car il serait impossible qu'elle fournisse les valeurs des activités de tous les composés organiques. Il est donc crucial de faire appel aux approches théoriques telles que les méthodes QSAR (relation quantitative structure-activité) pour compenser les contraintes de l'expérience.

Ces méthodes ont permis aux chercheurs de justifier les données expérimentales disponibles et prédire les activités pour de nouveaux composés ou des composés pour lesquels les données expérimentales ne sont pas disponibles, en un temps plus court.

Dans ce chapitre, une étude bibliographique sur la méthodologie QSAR a été présentée, les différentes étapes de développement, de validation et d'application sont aussi mises en œuvre.

1. Définition d'une méthode QSAR

Une relation quantitative structure-activité (QSAR) est un modèle mathématique qui associe un ou plusieurs paramètres quantitatifs (descripteurs) dérivés de la structure moléculaire des composés étudiés, à une grandeur macroscopique (activité).

2. Principe des méthodes QSAR

Le principe d'une étude QSAR consiste à trouver une relation mathématique, reliant de manière quantitative une activité biologique mesurée pour une série de composés similaires, dans les mêmes conditions expérimentales, avec des propriétés moléculaires appelées descripteurs, à l'aide des méthodes d'analyse statistique.

La forme générale d'un modèle QSAR est la suivante : $\text{Activité} = f(D_1, D_2, \dots, D_n, \dots)$

D_1, D_2, \dots, D_n sont des descripteurs moléculaires.

3. Méthodologie d'une étude QSAR

3.1 Développement du modèle

Les étapes pour élaborer un modèle QSAR fiable et adéquat sont les suivantes :

a/ Collecter une base de données expérimentales fiable et en nombre le plus important possible.

b/ Choisir des descripteurs adéquats pour l'activité étudiée.

c/ Choisir une méthode d'analyse des données afin de relier l'activité étudiée aux descripteurs choisis. (Chapitre II. Partie B)

- Régression linéaire simple et multiple (SLR, MLR).
- Régression des moindres carrés partiels (PLS).
- Analyse en composantes principales (PCA).

d/ Validation du modèle

- Validation interne
- Validation externe

e/ Validation externe

- Vérification des paramètres statistiques sur la série de test
- Critère de Tropsha

f/ Domaine d'applicabilité

3.1.1 Base de donnée données

Le choix de la base de données constitue une étape cruciale, elle a une forte influence sur le développement du modèle final, puisque ce dernier sera ajusté par rapport à ces données de référence. En effet, cette base expérimentale doit être de qualité, fiable avec des barres d'erreur faibles et obtenue suivant un même protocole expérimental.

3.1.2 Descripteurs moléculaires

Les descripteurs ont pour but de décrire de manière numérique l'information contenue dans la structure d'une molécule, c'est à dire les caractéristiques physicochimiques des molécules à partir de leurs représentations structurales. Ils peuvent être obtenus expérimentalement ou calculés à partir de la structure moléculaire. Une fois qu'ils sont disponibles et à l'aide des outils de modélisation classiques, il est possible de les lier avec l'activité étudiée.

3.1.2.1 Descripteurs théoriques

a. Les descripteurs 1D

Ce sont des descripteurs simples, qui nécessitent peu de connaissances sur la structure moléculaire et sont directement liés à la formule brute de la molécule. Ces descripteurs ne permettent pas d'élaborer des modèles complexes, il faut alors recourir à d'autres classes de descripteurs.

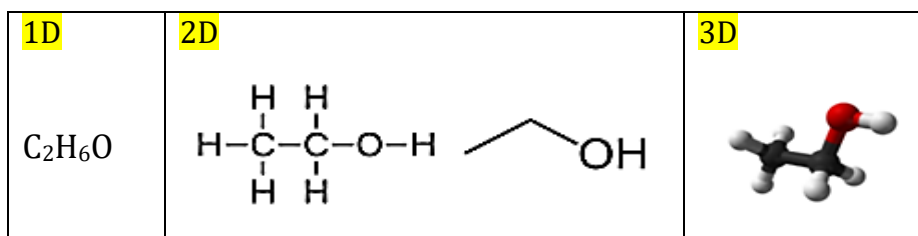
b. Les descripteurs 2D

Les descripteurs 2D sont accessibles à partir de la formule développée de la molécule. Ils constituent deux types d'indice : constitutionnels et topologiques.

c. Les descripteurs 3D

Ces descripteurs décrivent des caractéristiques plus complexes, donc leur calcul nécessite la connaissance de la géométrie 3D de la molécule, c'est-à-dire les positions relatives des atomes dans l'espace et l'utilisation des méthodes de modélisation moléculaire empirique ou ab-initio.

Ces descripteurs s'avèrent relativement coûteux en temps de calcul, mais apportent davantage d'informations pour l'interprétation mécanistique des modèles QSAR élaborés.



➤ Les descripteurs 3D géométriques

On peut citer : la surface accessible au solvant, le volume moléculaire, le moment d'inertie, Les distances, les angles de liaisons ou angles dièdres de la molécule.

➤ Les descripteurs 3D électroniques (quantique)

Ces descripteurs décrivent les propriétés électroniques. Ils sont très utilisés dans la plus part des études de l'activité biologique, car ils permettent de quantifier les différents types d'interactions inter et intramoléculaires.

Le calcul de ces descripteurs nécessite une bonne optimisation de la géométrie (énergie minimale → stabilité maximale) en faisant appel à la chimie quantique.

➤ **Les descripteurs thermodynamiques**

Les descripteurs thermodynamiques sont calculés à partir de la fonction de partition totale Q de la molécule. Parmi ces descripteurs on peut citer : $\Delta G, \Delta U, \Delta S, \Delta H$, la température critique CT , la pression critique CP ...

3.1.2.2 Descripteurs empiriques

3.1.2.2.1 Coefficient de partage (octanol-eau) K_{ow} , $\log P$

Le coefficient de partage octanol-eau (K_{ow}) est un descripteur physico-chimique largement utilisé dans des études (QSAR), il décrit le caractère lipophile (hydrophobe) d'une molécule et il est donné par la relation suivante :

$$K_{ow} = \frac{C \text{ substance dans l'octanol (phase lipidique)}}{C \text{ substance dans l'eau (phase aqueuse)}}$$

a. Détermination expérimentale de $\log P$

La méthode expérimentale utilisée pour mesurer le coefficient de partage d'une molécule est la méthode des flacons agités [1] d'une solution octanol/eau. Cette méthode est préconisée comme procédure standard de caractérisation par l'OCDE [2] (Organisation de Coopération et de Développement Economique).

b. Détermination théorique de $\log P$

Le coefficient de partage peut également être déterminé théoriquement en utilisant quelques logiciels tels que: ACD/ChemSketch, HyperChem, Ecosar, ChemSpider (online)... qui sont basés sur certaines méthodes d'estimation. Parmi ces méthodes on peut citer : La Méthode de Hansch [3], méthode de Rekker [4], méthode de Ghose et Viswanadhan [5], méthode de Klopman et Iroff [6], La méthode de Bodor [7].

3.2 Validation du modèle

3.2.1 Validation interne

Une série d'apprentissage constituée de 2/3 de la base de données est utilisée afin de vérifier la stabilité et le pouvoir explicatif du modèle final.

3.2.1.1 Validation croisée (Cross Validation CV)

La technique la plus employée pour déterminer la stabilité du modèle prédictif est de tester l'influence de chaque échantillon sur le modèle final.

Ce processus consiste à extraire un certain nombre k des N molécules de la base de données initiale et à construire un nouveau modèle avec les $N-k$ molécules restantes à l'aide des descripteurs choisis.

Ce processus est ensuite réitéré pour prédire les valeurs de toutes les molécules du jeu d'entraînement, en fonction du nombre de molécules retirées à chaque itération, on parlera de Leave-One-Out (LOO) ou de Leave-Many-Out (LMO) [8] selon qu'une ou plusieurs molécules retirées.

Cette méthode est quantifiée par R^2_{cv} (LOO/LMO). Si le R^2_{cv} est proche de R^2 , le modèle est jugé stable.

3.2.1.2 Y-Randomisation

Il est possible d'obtenir un modèle présentant de bons paramètres statistiques par un pur hasard.

Dans le but de prouver qu'une valeur élevée du coefficient de corrélation (R^2 ou R^2_{cv}) du modèle établis n'a pas été résultat d'une corrélation due au hasard [9], une technique appelée Y-randomisation est appliquée sur la série d'apprentissage.

Cette approche consiste à redistribuer aléatoirement les valeurs de l'activité expérimentale sur l'ensemble de la série d'apprentissage en gardant les mêmes descripteurs et un nouveau modèle est dérivé.

L'opération est répétée plusieurs fois et les modèles aléatoires obtenus doivent présenter de faibles performances c'est-à-dire la moyenne des coefficients de corrélation résultants doit être très faible par rapport à celui obtenu avec le vrai modèle ($R^2_r \ll R^2$).

R^2_r : coefficient de détermination des modèles obtenus avec les Y-randomisées.

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

3.2.2 Validation externe

Afin de vérifier le pouvoir prédictif du modèle QSAR élaboré, une série de test (les molécules non employées dans le développement du modèle) généralement constituée de 1/3 de la base de données est utilisée.

Le but d'un bon modèle QSAR est non seulement de prédire l'activité des composés d'ensemble d'apprentissage, mais aussi de prévoir les activités des molécules de la série de test [10].

Cette validation est basée sur le R^2 (test) et SD (test) entre les activités observées et les activités prédites pour l'ensemble de test. Cependant, plusieurs études [11,12] ont montré que ces deux paramètres sont insuffisants pour vérifier de manière fiable le pouvoir prédictif des modèles QSAR. Par conséquent, d'autres paramètres doivent être vérifiés pour cet objectif. Ces paramètres sont connus sous le nom «critères de validation externe» ou souvent appelés « critères de Trophsa » [13].

Les critères de Trophsa sont les suivants :

- $R^2 > 0.7$
- $R_{cv}^2 > 0.6$
- $\frac{R^2 - R_0^2}{R^2} < 0.1$ et $0.85 \leq k \leq 1.15$
- $\frac{R^2 - R_0'^2}{R^2} < 0.1$ et $0.85 \leq k' \leq 1.15$
- $|R^2 - R_0^2| \leq 0.3$

Avec :

- R^2 : Coefficient de corrélation pour les molécules de la série de test.
- R_0^2 : Coefficient de corrélation entre les valeurs prédites et expérimentales pour la série de test.
- $R_0'^2$: Coefficient de corrélation entre les valeurs expérimentales et prédites pour la série de test.
- K : Constante de la droite (à l'origine) de corrélation (valeurs prédites en fonction des valeurs expérimentales).
- K' : Constante de la droite (à l'origine) de corrélation (valeurs expérimentales en fonction des valeurs prédites).

3.3 Domaine d'applicabilité

Un modèle QSAR idéal est celui qui est capable de prédire l'activité de n'importe quelle molécule imaginable. Cependant cela est souvent loin d'être possible car même les modèles les plus exhaustifs, dignes de confiance et validés, ne peuvent prédire des activités de manière fiable pour l'intégralité des composés chimiques

existants. Le nombre limité du jeu d'entraînement rend l'espace chimique des modèles élaborés limité. Par conséquent, lorsqu'une molécule se situe en dehors de cet espace chimique, la prédiction ne sera plus fiable [14].

Afin d'éviter cette extrapolation hasardeuse et prévenir ce type de problème, un domaine d'applicabilité (DA) doit être déterminé.

Le DA correspond donc à la région de l'espace chimique incluant les composés du jeu d'apprentissage et les composés similaires, proches dans ce même espace [15].

La détermination des DA est donc d'une grande importance. Cette partie de l'analyse est d'ailleurs explicitement demandée par l'OCDE [14,16].

Le DA est discuté à l'aide du diagramme de Williams qui représente la variation des résidus standardisés de la variable dépendante avec la distance entre les valeurs des descripteurs et leurs moyennes appelée *Leverage* [13,17].

La valeur seuil de Leverage est donnée par Williams : $h^* = 3\bar{P}/N$

Avec :

- $\bar{P} = P + 1$
- N : nombre d'observations
- P : nombre de descripteurs contenus dans le modèle

Les composés ayant un résidu et un levier qui dépasse le seuil h^* , seront considérés en dehors du domaine d'applicabilité du modèle élaboré. Ce sont les observations aberrantes (*outliers*).

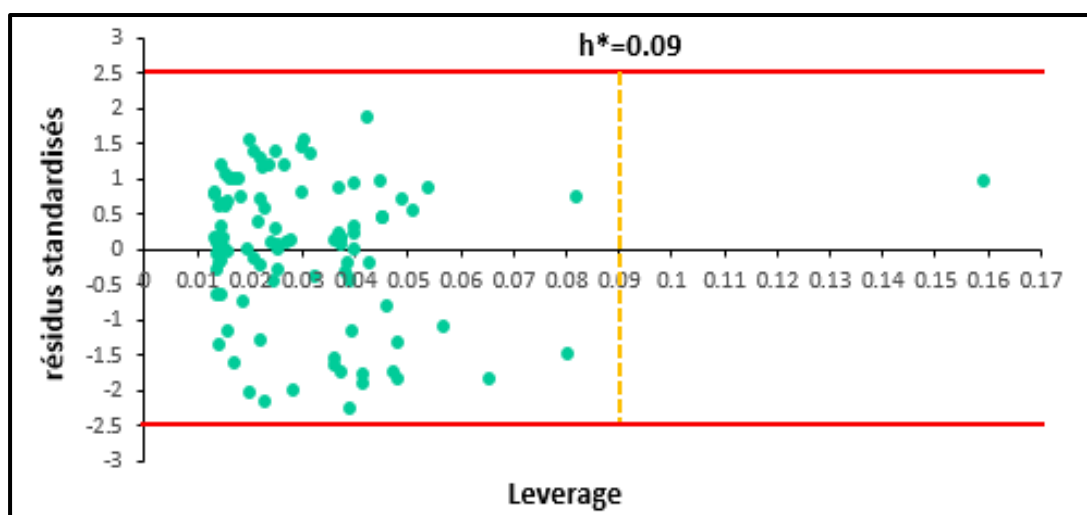


Figure 1 : Diagramme de Williams.

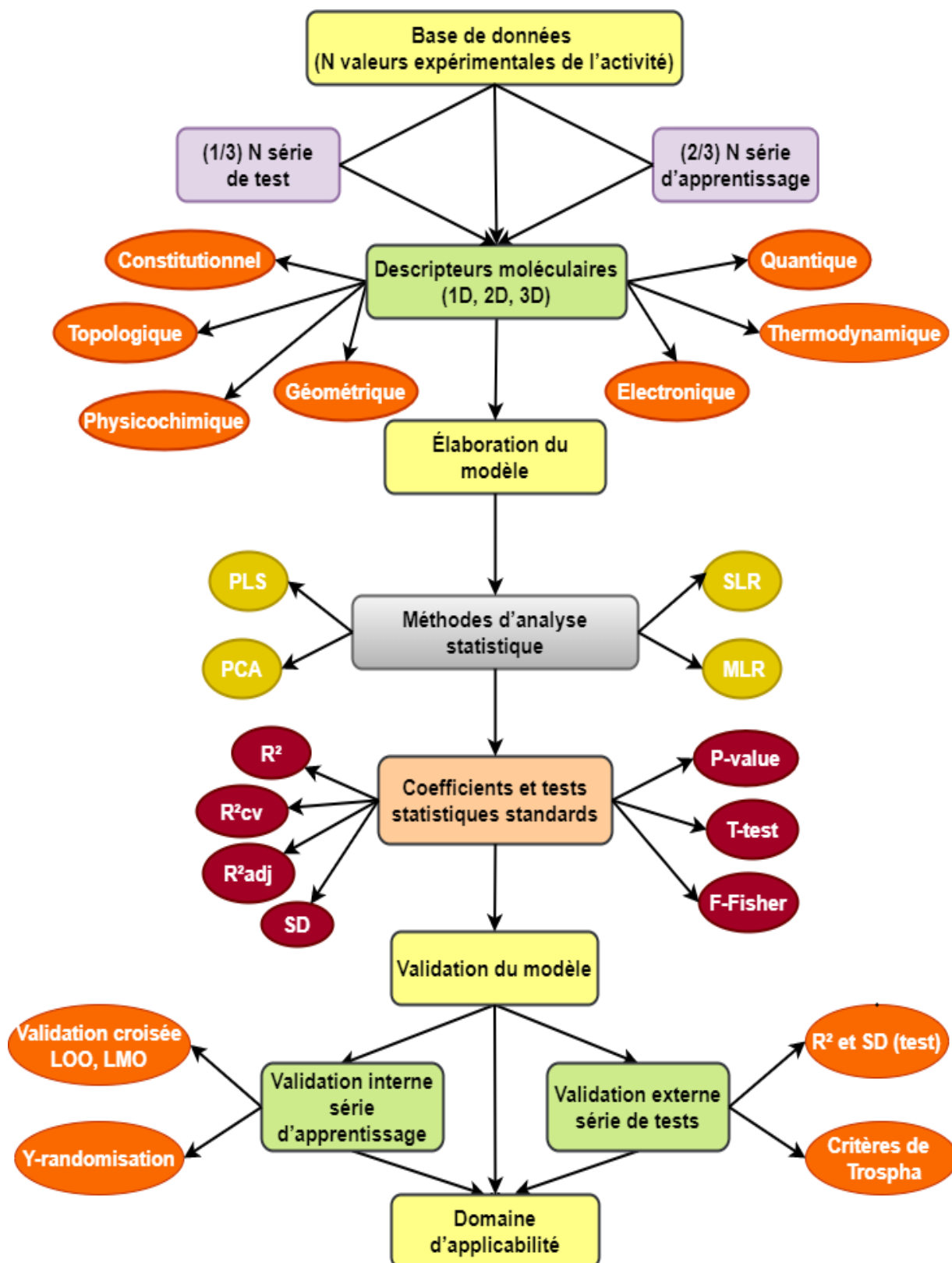


Figure 2 : Méthodologie QSAR

Références Bibliographiques

- [1] B. Lee and F.M. Richards, The interpretation of protein structures: estimation of static accessibility, *Journal of molecular biology*. 55, 379-490, **1971**.
- [2] OECD Guidelines for Testing of Chemicals N°107, OECD, Paris, **1992**.
- [3] R.S. Pearlman, W.J. Dunn and J.H. Block, Partition Coefficient Determination and Estimation, Editions. Pergamon, New York, 3-20, **1986**
- [4] R. F. Rekker, H. M.de Kort, the hydrophobic fragmental constant, *European Journal of Medicinal Chemistry, ChimTher.* 14, 479-488, **1979**.
- [5] A.K. Ghose and G.M. Crippen, Atomic physicochemical parameters for three dimensional structure directed quantitative structure activity relationships, modeling dispersive and hydrophobic interactions, *Journal of Chemical Information & Computer Sciences.* 27, 21-35, **1987**.
- [6] G. Klopman and L.D. Iroff, Calculation of partition coefficients by the charge density method, *Journal of Computational Chemistry.* 2, 157-160, **1981**.
- [7] N. Bodor and P. Buchwald, Molecular size based approach to estimate partition properties for organic solutes, *Journal of Physical Chemistry.* 101, 3404-3412, **1997**.
- [8] L. Zhang, H. Zhu, T.I. Oprea, A. Golbraikh and A. Tropsha, QSAR modeling of the blood-brain barrier permeability for diverse organic compounds, *Pharmaceutical Research.* 25, 1902-1914, **2008**.
- [9] M. Clark and R.D. Cramer, The probability of chance correlation using partial least squares (PLS), *Molecular Informatics.* 12, 137-145, **1993**.
- [10] S. Ekins, G. Bravi, S. Binkley, J.S. Gillespie, B.J. Ring, J.H. Wikel and S.A. Wrighton, Three and four dimensional quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors, *Drug Metabolism & Disposition.* 28, 994-1002, **2000**.
- [11] A. Golbraikh and A. Tropsha, Beware of q^2 !, *Journal of Molecular Graphics & Modeling.* 20, 269-276, **2002**.
- [12] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu and A. Tropsha, Does rational selection of training and test sets improve the outcome of QSAR modeling?, *Journal of Chemical Information & modeling.* 52, 2570-2578, **2012**.
- [13] A. Tropsha, P. Gramatica and V.K. Gombar, the importance of Being Earnest: Validation is the Absolute Essential for Successful Application and interpretation of QSPR Models, *QSAR & Combinatorial Sciences.* 22, 69-77, **2003**.
- [14] Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models, Organisation de Coopération et de Développement Economique, Paris, 13, **2009**.

[15] T.I. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D. Roberts, T. Schultz, D.W. Stanton, J.J.M. Van De Sandt, W. Tong, G. Veith and C. Yang, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships, The report and recommendations of ECVAM Workshop, *Atla.* 33,155-173, **2005**.

[16] J. Tunkel, K. Mayo, C. Austin, A. Hickerson and P. Howard, Practical Considerations on the Use of Predictive Models for Regulatory Purposes, *Environmental Science & Technology.* 39, 188-2199, **2005**.

[17] L. Eriksson, J. Jaworska, A. Worth, M. Cronin, R.M. Mc Dowell and P. Gramatica, Methods for reliability, uncertainty assessment and applicability evaluations of classification and regression based QSARs, *Environmental Health Perspectives.* 111, 1361-1375, **2003**.

Chapitre II
Partie B- Méthodes d'analyse statistique

Introduction

Une méthode d'analyse de données est nécessaire pour la mise au point d'un modèle QSAR. Elle permet en effet, de rechercher une relation entre une activité et une ou plusieurs variables quantitatives (descripteurs moléculaires).

Plusieurs approches sont envisageables, il s'agit alors de choisir la plus adaptée et celle permettant au mieux caractériser le système pour obtenir un modèle fiable.

Dans l'ensemble de notre étude, nous avons principalement utilisé comme techniques pour l'analyse des données et la construction des modèles QSAR, la régression linéaire simple et multiple (SLR, MLR), la régression des moindres carrés partiels (PLS) et l'analyse en composantes principales (PCA).

Dans ce sous chapitre, le principe et les différentes étapes de chaque méthode sont mis en œuvre.

1. Méthodes de régression linéaire (SLR, MLR)

Les méthodes de régression linéaire sont les plus utilisées, elles permettent de mettre en évidence une relation linéaire entre une réponse ou variable à expliquer (dépendante), notée Y , et une ou plusieurs variables explicatives (indépendantes) notées X .

1.1 Régression linéaire simple (SLR)

La régression linéaire simple est une méthode de régression permettant de relier linéairement une variable dépendante Y avec une variable indépendante X .

La relation entre ces deux variables s'écrit de la manière suivante : $Y = a_0 + aX$ (1)

Cependant, ce modèle est une forme simplifiée, car en réalité, il est perturbé par un terme d'erreur (résidu) noté ε que l'on doit introduire.

La relation devient alors : $Y = a_0 + aX + \varepsilon$ (2)

L'intercepte a_0 et a sont les coefficients de régression, constantes inconnues qu'on cherche à estimer et ε est le terme d'erreur.

Pour déterminer les paramètres a et a_0 , n observations de la variable X , notées x_i ($i=1, \dots, n$) et n valeurs de la variable Y notées y_i sont alors mesurées, c'est-à-dire une collecte de n couples de données $(x_i ; y_i)$. Cela se traduit par l'équation suivante:

$$Y_i = a_0 + aX_i + \varepsilon_i \quad (i = 1, \dots, n) \quad (3)$$

• **Critère des Moindres Carrés**

Pour trouver la meilleure droite, il faut chercher les meilleures valeurs des coefficients de régression \mathbf{a} et \mathbf{a}_0 .

$$\widehat{Y}_i = \widehat{a}_0 + \widehat{a}X_i \quad (4)$$

\widehat{a}_0 et \widehat{a} sont des estimateurs de a_0 et a

Les écarts ou les différences entre les valeurs \widehat{Y}_i estimées (obtenues à partir de l'équation de régression) et Y_i observées (expérimentales) sont appelés les moindres (résidus). $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$ (5)

Si l'on prend la somme des écarts (moindres), ces derniers se compensent et la somme totale peut être nulle, ce qui ne reflète pas la réalité.

Pour éviter ce problème de compensation, il faut prendre le carré des écarts (moindres) d'où l'appellation moindres carrés.

La somme des carrés des écarts (SCE) est donnée par :

$$S(\widehat{a}, \widehat{a}_0) = \sum_{i=1}^n \widehat{\varepsilon}_i^2 = \sum_{i=1}^n (Y_i - \widehat{a}_0 - \widehat{a}X_i)^2 \quad (6)$$

Le principe de la méthode des moindres carrés consiste à minimiser la quantité S c'est-à-dire la somme des moindres carrés. En effet cette fonction est minimale lorsque ses dérivées par rapport à \mathbf{a}_0 et \mathbf{a} s'annulent.

$$\frac{\partial S}{\partial \widehat{a}_0} = -2 \sum_{i=1}^n (Y_i - \widehat{a}_0 - \widehat{a}X_i) = 0 \quad (7)$$

$$\frac{\partial S}{\partial \widehat{a}} = -2 \sum_{i=1}^n X_i(Y_i - \widehat{a}_0 - \widehat{a}X_i) = 0 \quad (8)$$

L'équation (7) donne :

$$\sum_{i=1}^n Y_i - n\widehat{a}_0 - \widehat{a} \sum_{i=1}^n X_i = 0 \quad (9)$$

En utilisant la formule de la moyenne $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ et en divisant par n, on obtient :

$$\widehat{a}_0 = \bar{Y} - \widehat{a}\bar{X} \quad (10)$$

L'équation (8) donne :

$$\sum_{i=1}^n X_i Y_i - \widehat{a}_0 \sum_{i=1}^n X_i - \widehat{a} \sum_{i=1}^n X_i^2 = 0 \quad (11)$$

En remplaçant \hat{a}_0 par sa formule obtenue en équation (10), on obtient :

$$\sum_{i=1}^n X_i Y_i - (\bar{Y} - \hat{a} \bar{X}) \sum_{i=1}^n X_i - \hat{a} \sum_{i=1}^n X_i^2 = 0 \quad (12)$$

A partir de l'équation (12), on peut tirer l'expression du deuxième coefficient de régression \hat{a} .

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

$$\hat{a} = \frac{\sum (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad (13)$$

$$\hat{a}_0 = \bar{Y} - \frac{\sum (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \bar{X} \quad (14)$$



Figure 1 : Représentation graphique de la régression linéaire simple

1.2 Régression linéaire Multiple (MLR)

La régression linéaire multiple est une méthode de régression permettant de relier linéairement une réponse ou variable à expliquer (dépendante), notée Y_i , et plusieurs variables explicatives (indépendantes) notées X_i . Cela se traduit par l'équation suivante :

$$Y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + \dots + a_p X_{ip} + \varepsilon_i \quad i = 1, \dots, n \quad (15)$$

Dans notre étude, Y_i et X_i représentent les activités observées et les descripteurs calculés respectivement.

Les n échantillons (observations) des variables dépendantes et indépendantes sont connues. Il s'agit donc de considérer un système d'équations qui peut être donné sous la forme matricielle suivante :

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_p \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ 1 & X_{31} & X_{32} & \dots & X_{3p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_p \end{pmatrix}$$

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

$$S(a_0, \dots, a_p) = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - a_0 - a_1 X_{i1} - \dots - a_p X_{ip})^2 \quad (16)$$

La valeur prédite de la variable dépendante Y (estimée par le modèle de régression)

$$\text{s'écrit : } \hat{Y} = X\hat{a} = Xa \quad (17)$$

Les valeurs des a qui minimisent ce critère seront les solutions a_0, a_1, a_p du système linéaire de (p+1) équations à (p+1) inconnues obtenues comme suit :

$$a = (X'X)^{-1} X'Y \quad (18)$$

Avec

- X' : la matrice transposée de la matrice des variables explicatives X.
- $(X'X)^{-1}$: la matrice inverse de la matrice $(X'X)$.
- Y : vecteur des valeurs de la variable à expliquer.

➤ **Analyse de la variance :**

La relation fondamentale de l'analyse de la variance est donnée par :

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2 \quad (19)$$

Ces diverses sommes des carrés sont définies comme suit :

$$SST = \sum_i (Y_i - \bar{Y})^2 \quad (20)$$

SST (Sum of Squares Total) représente la somme des carrés total et se décompose en :

$$SSR = \sum_i (\hat{Y}_i - \bar{Y})^2 \quad (21)$$

SSR (Sum of Squares Regression) Somme des carrés expliquée par le modèle.

$$SSE = \sum_i (Y_i - \hat{Y}_i)^2 \quad (22)$$

SSE (Sum of Squares Error) Somme des carrés des résidus qui représente la partie non expliquée par le modèle.

L'équation de l'analyse de la variance est donc la suivante : $SST = SSR + SSE$ (23)

Cette formule montre que la variation totale de Y (SST), peut être expliquée par le modèle grâce à SSR, et SSE est la partie qui ne peut pas être expliquée par le modèle.

➤ **Représentation de l'analyse de la variance**

Les sommes citées ci-dessus sont présentées dans la table d'analyse de variance ou souvent appelée table d'ANOVA (ANalysis Of VAriance).

Tableau 1 : Analyse de la variance

Source de Variation	Degrés de Liberté DF	Somme des Carrées SS	Moyenne des carrées MS
Model	P	$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$MSR = \frac{SSR}{P}$
Error	n-1-p	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n - 1 - P}$
Total	n-1	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$MST = \frac{SST}{n - 1}$

SS : Sum of squares : somme carrée

MS : Mean square : le rapport SS/DF

DF: degré de liberté

1.3 Signification et qualité de la régression linéaire

La confirmation de l'existence du modèle, la contribution de chaque descripteur dans l'explication de la réponse Y ainsi que la qualité du modèle nécessitent la connaissance de plusieurs paramètres.

1.3.1 Test de la signification globale de la régression :

a. Test de Fisher-Snedecor

Ce test est surtout utilisé dans le cadre de la régression multiple et couramment employé afin de mesurer le niveau de signifiante statistique du modèle, c'est-à-dire la qualité du choix du jeu de paramètres.

$$F = \frac{MS_{model}}{MS_{error}} \tag{24}$$

Avec :

$$MS_{\text{model}} = \frac{SS_{\text{Model}}}{p} \quad (25)$$

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{n - p - 1} \quad (26)$$

Le test Fisher-Snedecor permet de tester l'hypothèse nulle selon laquelle chaque coefficient est significativement différent de zéro.

Après le calcul de F (observé) on le compare avec le F théorique obtenu à partir des tables statistiques usuelles (la table de Fisher).

Si la quantité F observée dépasse le seuil ($F > \text{seuil}$), on rejette l'hypothèse H_0 (au niveau $\alpha=5\%$) dans le cas contraire (Si F observé est plus petit que le F théorique : acceptation de l'hypothèse nulle H_1).

La statistique F est liée au coefficient de détermination par la relation suivante :

$$F = \frac{R^2}{1 - R^2} \frac{n - P - 1}{P} \quad (27)$$

b. La valeur "P-Value"

Un résultat statistiquement significatif est un résultat qui serait improbable si l'hypothèse nulle était vérifiée.

La p-value est utilisée pour quantifier la significativité statistique d'un résultat dans le cadre d'une hypothèse nulle. L'idée générale est de prouver que l'hypothèse nulle n'est pas vérifiée, car dans le cas où elle le serait, le résultat observé serait fortement improbable.

On compare la p-value au risque α choisi (par exemple $\alpha=5\%$), Si $P\text{-value} < \alpha$ alors l'hypothèse nulle ($a_1 = \dots = a_p = 0$) est rejetée.

1.3.2 Test de signification de chaque descripteur (t-Student)

Le t-test de Student est utilisé afin d'évaluer la contribution de chaque variable dans l'explication de la réponse Y.

La formule utilisée pour calculer le t-test pour chaque paramètre a_i est la suivante :

$$t\text{-test}(\hat{a}_i) = \frac{\hat{a}_i}{S(\hat{a}_i)} \quad (28)$$

Avec $S(\hat{a}_i)$ est l'erreur type du paramètre \hat{a}_i .

- Si $t\text{-test} < t\text{-test seuil}$: H_0 vérifiée.
- Si $t\text{-test} > t\text{-test seuil}$: H_0 rejetée.

1.4 Indicateurs de qualité d'une régression linéaire

a. Coefficient de détermination R^2

R^2 est défini par :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (29)$$

Interprétation

Plus le R^2 tend vers 1, plus le nuage de points se resserre autour de la droite de régression. Quand les points sont exactement alignés sur la droite de régression, alors $R^2 = 1$ (cas idéal) et on peut dire que les valeurs prédites et observées sont corrélées.

R^2 est un indicateur qui permet de juger la qualité d'une régression linéaire **simple**. Cependant, il n'est pas recommandé de l'utiliser pour comparer des modèles complexes, c'est à dire en régression multiple, car la valeur de ce dernier peut augmenter lors d'une nouvelle variable ajoutée à l'équation de régression, même si elle ne contribue pas dans l'explication de la réponse Y .

Un autre indicateur statistique peut être utilisé en régression multiple, appelé R^2 ajusté.

b. Coefficient de Régression ajusté R^2_{adj}

Le coefficient de détermination ajusté R^2_{adj} est défini par :

$$R^2_{adj} = 1 - \frac{SSE/n - 1 - p}{SST/n - 1} \quad (30)$$

Ce coefficient est utilisé en régression multiple, car il tient compte du nombre de variables explicatives.

c. Cross validated R^2_{cv}

La procédure statistique cross-validation est utilisée pour évaluer le pouvoir explicatif des modèles QSAR. Le coefficient qui décrit la validation est donné par la formule suivante :

$$R_{CV}^2 = 1 - \frac{\sum_i (Y_i^{\text{pred}} - Y_i^{\text{obs}})^2}{\sum_i (Y_i^{\text{obs}} - Y_i^{\text{mean}})^2} \quad (31)$$

Ce coefficient peut être calculé à partir de PRESS comme suit :

$$R_{CV}^2 = 1 - \frac{\text{PRESS}}{\text{SS Total}} \quad (32)$$

d. Critère de validation croisée (PRESS)

La somme des erreurs de prédiction « **PRE**diction **S**um of **S**quares » (**PRESS**) est définie par :

$$\text{PRESS} = \sum_i \varepsilon_i^2 \quad (33)$$

Un PRESS faible signifie un bon pouvoir prédictif du modèle.

e. Déviation standard (SD)

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

$$s^2 = \text{MSE} = \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n - p - 1} = \frac{\text{SSE}}{n - p - 1} \quad (34)$$

RMSE (Root Mean Square Error) = SD = s

Une valeur faible de SD signifie un bon ajustement statistique du modèle et une forte fiabilité de la prédiction.

2. Méthode des moindres carrés partiels PLS

2.1 Définition

La régression des moindres carrés partiels PLS, est une généralisation et combinaison de la régression linéaire multiple et de l'analyse en composantes principales. Elle peut être utilisée lorsque le nombre de descripteurs est élevé et que ceux-ci sont fortement corrélés.

2.2 Principe

Le principe de La régression PLS consiste à réduire le nombre de prédicteurs à un nombre plus petit de composantes non corrélées et qui effectue la régression par les moindres carrées sur ces composantes plutôt que sur les données initiales.

2.3 Etapes de la régression PLS

Etape 1 : Construction de la première composante t_1 ($h=1$)

- **Maximiser la corrélation** entre la première composante t_1 et la variable à expliquer Y.
- **Maximiser la variance** de la première composante t_1 afin qu'elle représente au mieux toutes les variables explicatives.

$$t_1 = WX = W_{11}X_1 + W_{12}X_2 + W_{13}X_3 + \dots W_{1p}X_p \quad (35)$$

W : poids de chaque variable explicative X dans la première composante t_1 , il peut être obtenu comme suit :

$$W_{1j} = \frac{\text{Cov}(X_j, Y)}{\sqrt{\sum_{k=1}^p \text{Cov}^2(X_k, Y)}} \quad (36)$$

L'équation de la régression simple de Y sur t_1 est la suivante :

$$Y = C_1 t_1 + Y^{[1]} \quad (37)$$

Avec :

- C_1 : coefficient de régression de Y sur la première composante t_1 .
- $Y^{[1]}$: résidu non expliqué par la première composante t_1 .

Etape 2 : Construction des composantes suivantes

- ✓ $h = 2$ (deuxième composante t_2)

Les résidus de la 1^{ère} composante vont servir comme données pour construire la deuxième composante t_2 , en faisant p régressions simples des variables $X_1 \dots X_p$ sur t_1 .

De la même manière, la deuxième composante sera construite en utilisant les résidus des X et de Y provenant de la première composante.

$$w_{2j} = \frac{\text{Cov}(X_j^{[1]}, Y^{[1]})}{\sqrt{\sum_{k=1}^p \text{Cov}^2(X_k^{[1]}, Y^{[1]})}} \quad (38)$$

$$Y = C_1 t_1 + C_2 t_2 + Y^{[2]} \quad (39)$$

Avec :

- $Y^{[2]}$: le résidu de Y non expliqué par la deuxième composante.
- C_1 et C_2 sont les coefficients de régression de Y sur t_1 et sur t_2 respectivement.

La procédure est répétée jusqu'à obtenir **h = A composantes = ?**

Etape 3 : Choix du nombre de composantes

Le problème qui se pose après avoir construit un modèle est de connaître son aptitude à prédire les réponses de nouvelles variables d'entrée [X]. Pour le cas de PLS, ceci consiste surtout à évaluer le nombre optimal de composantes à inclure dans le modèle. En effet, si **beaucoup** de composantes sont incluses, il peut y avoir un phénomène de surévaluation (**over-fitting**) : les composantes sans importance peuvent fausser les prédictions.

Par contre, si **trop peu** de composantes sont incluses, on risque d'avoir peu d'information pour expliquer le Y.

Etape 4 : Critère de choix de nombre de composantes:

Pour déterminer à quel moment le nombre de composantes est suffisant pour traduire la variance du système, une démarche de validation interne est réalisée et la composante est considérée utile si elle contribue de manière significative à améliorer la robustesse de l'analyse.

Critère de « validation croisée »

1. Calcul de R^2_{cv} après l'ajout de chaque composante.

- Si $R^2_{cv}(h) > R^2_{cv}(h - 1)$, cela signifie que la nouvelle composante ajoutée a un effet sur l'explication de Y.

- Si $R^2_{cv}(h) \approx R^2_{cv}(h - 1)$, cela signifie que la composante ajoutée n'a pas d'effet sur l'explication de Y et l'équation doit contenir dans ce cas (h-1) composantes.

2. calcul du PRESS

$$PRESS_h = \sum_{i=1}^n (Y_i - \hat{Y}_{(-1)}^h)^2 \tag{40}$$

Cette quantité décroît en fonction du nombre de composantes pour atteindre une valeur minimale et se stabiliser par la suite. C'est ce minimum qui détermine le nombre de composantes à retenir pour le modèle.

3. Méthode d'analyse en composante principale (PCA)

3.1 Principe

L'analyse en composantes principales (PCA: Principal Component Analysis) est une méthode qui consiste à transformer des variables corrélées (liées entre elles) en nouvelles variables non corrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet de réduire le nombre de variables et de rendre l'information moins redondante.

3.2 Etape de la PCA

1- centrer et réduire les variables.

- Center : retrancher la moyenne ($X - \bar{X}$). Cette opération a pour but de donner le même poids à toutes les variables.
- Réduire : diviser la variable par l'écart type $(X - \bar{X})/\bar{V}$ dans le but d'analyser des variables (données) sans unité.

2- matrice de corrélation

$$\begin{pmatrix} 1 & R_{X_1 X_2} & R_{X_1 X_3} \\ R_{X_2 X_1} & 1 & R_{X_2 X_3} \\ R_{X_3 X_1} & R_{X_3 X_2} & 1 \end{pmatrix}$$

3- diagonaliser la matrice de corrélation.

Cette étape va nous permettre d'avoir les valeurs et les vecteurs propres.

A partir des vecteurs propres, on construit les composantes.

4- construction des composantes principales.

<i>composantes</i>	<i>Valeurs propres</i>	<i>Vecteurs propres</i>	<i>proportion</i>
1	λ_1	$V_1 \begin{pmatrix} a_1 \\ b_1 \end{pmatrix}$	%
2	λ_2	$V_2 \begin{pmatrix} a_2 \\ b_2 \end{pmatrix}$	%

$$PC_1 = a_1X_1 + b_1X_2 \quad \text{et} \quad PC_2 = a_2X_1 + b_2X_2$$

La première composante principale traduit la plus grande part de la variance globale du système, les composantes successives n'ayant pour but que d'expliquer la variance résiduelle, non expliquée par les composantes qui les précèdent.

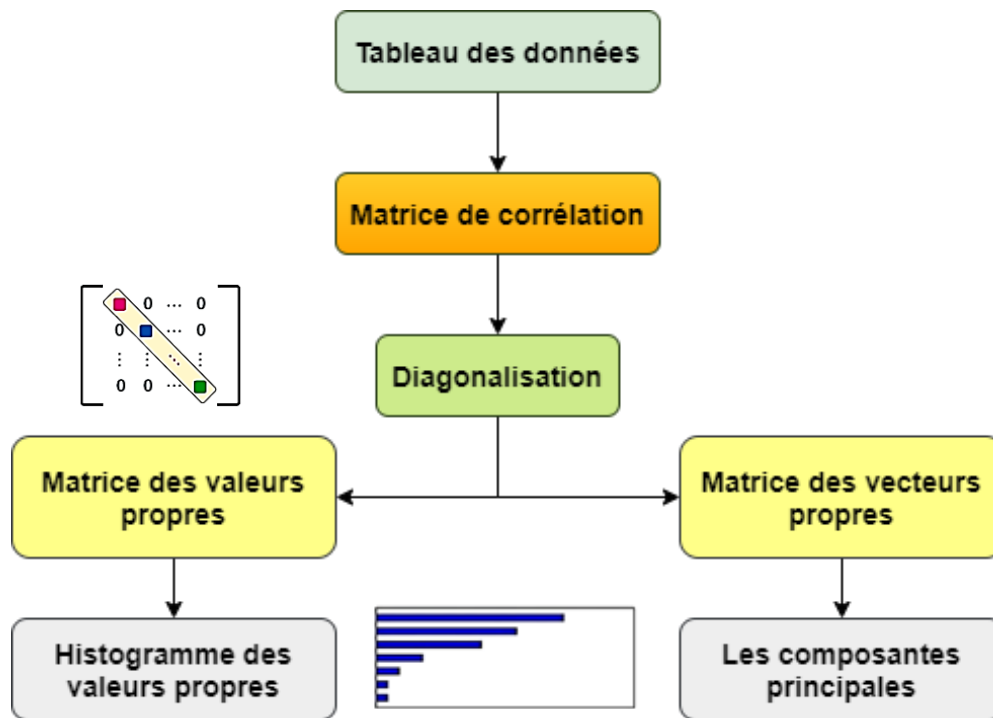


Figure 2 : Etapes de la méthode PCA

Des représentations à 2 dimensions suivant les composantes PC1 et PC2 peuvent être utilisées, ces deux composantes étant celles qui caractérisent la plus grande part de la variance dans le système.

La matrice des coordonnées nous permet d'analyser la dispersion des individus dans le nouvel espace défini. Ainsi, deux échantillons proches graphiquement portent une information très similaire.

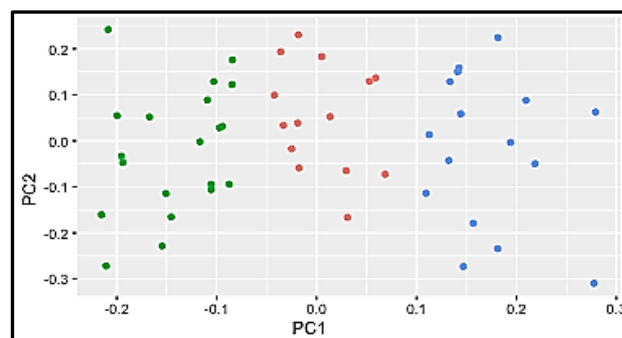


Figure 3 : Matrice des coordonnées dans le plan des 2 premières composantes principales d'une analyse PCA.

Références Bibliographiques

- [1] P.A. Cornillon, É. Matzner-Løber, Régression: théorie et applications, Springer, 2007.
- [2] K. BELLIFA, Etude des relations quantitatives structure-toxicité des composés chimiques à l'aide des descripteurs moléculaires. "Modélisation QSAR", 2015
- [3] W.C. Parr and M.A. Smith, Developing case-based business statistics courses, The American Statistician, 52, 330-337, 1998.
- [4] R.D. Cook and S. Weisberg, An introduction to regression graphics, John Wiley & Sons, New York, 2009.
- [5] P. Besse, Apprentissage statistique & data mining, INSA, Toulouse, 2009.
- [6] G.j. McLachlan, Discriminant analysis and statistical pattern recognition, John Wiley & Sons, New York, 2004.
- [7] J.P. Nakache, J. Confais, Statistique explicative appliquée: analyse discriminante, modèle logistique, segmentation par arbre, Editions Technip, 2003.
- [8] G. Saporta, Probabilités, analyse des données et statistique, Editions Technip, 2006.
- [9] S. Frontier, Étude de la décroissance des valeurs propres dans une analyse en composantes principales : Comparaison avec le modèle du baton brisé, Journal of Experimental Marine Biology and Ecology. 25, 67-75, 1976.

Chapitre II

Partie C-Méthodes Quantiques

Introduction

Le comportement électronique et nucléaire des molécules étant responsable des propriétés chimiques, il ne peut être décrit adéquatement qu'à partir de l'équation de Schrödinger et des autres postulats fondamentaux de la mécanique quantique.

Dans le présent sous-chapitre, on expose différentes méthodes de résolution de l'équation de Schrödinger, celles basées sur la théorie de Hartree-Fock (HF), ainsi celles basées sur la théorie de la fonctionnelle de la densité (DFT) qui permettent d'atteindre des solutions précises de l'équation et qui sont parmi les principaux outils de la chimie computationnelle actuelle.

$$H\Psi = E\Psi \quad (1)$$

- Ψ : fonction d'onde décrivant le système à N noyaux et à n électrons.
- E : valeurs propres de H.
- H : opérateur Hamiltonien $\hat{H} = \hat{T} + \hat{V}$

$$(2)$$

L'équation de Schrödinger ne peut être résolue de manière exacte que pour l'atome d'hydrogène et pour les systèmes hydrogénoïdes, en effet, pour les systèmes poly-électroniques, il est nécessaire d'introduire différentes approximations.

La première approximation en chimie quantique est de considérer que l'équation de Schrödinger [1] non relativiste indépendante du temps, où l'hamiltonien H total est défini par la somme des 5 termes suivants : $H = T_e + T_N + V_{ee} + V_{NN} + V_{eN}$

$$(3)$$

$$H = -\frac{\hbar^2}{2m_e} \sum_i^n \Delta_i - \frac{\hbar^2}{2M_K} \sum_K^N \Delta_K + \sum_{i>j}^n \frac{e^2}{r_{ij}} + \sum_{K>L}^N \frac{Z_K Z_L e^2}{r_{KL}} - \sum_{K=1}^N \sum_{i=1}^n \frac{Z_K e^2}{R_{Ki}} \quad (4)$$

Terme cinétique des électrons	Terme cinétique des noyaux	Terme de répulsion électrons- électrons	Terme de répulsion noyaux- noyaux	Terme d'attraction électrons- noyaux
-------------------------------------	----------------------------------	--	--	---

1. Approximation de Born-Oppenheimer

L'approximation de Born-Oppenheimer [2] permet de décomposer l'hamiltonien en deux contributions distinctes : l'une électronique et l'autre nucléaire. Étant donné que la masse des noyaux est beaucoup plus importante que celle des électrons, leur mouvement est négligeable devant celui des électrons, il est donc possible de considérer les noyaux

fixes. Il s'en suit donc que leur énergie cinétique est négligeable devant celle des électrons et que l'énergie d'interaction entre noyaux est constante.

L'hamiltonien H peut se réduire à la forme suivante : c'est l'hamiltonien électronique.

$$H_{\text{elec}} = T_e + V_{ee} + V_{eN} \quad (5)$$

L'énergie totale du système (pour une configuration nucléaire donnée) sera alors obtenue en ajoutant à l'énergie électronique E_{elec} du système, la répulsion nucléaire.

$$E_{\text{tot}} = E_{\text{elec}} + \sum_k \sum_{k < l} \frac{Z_k Z_l}{r_{kl}} \quad (6)$$

2. Méthode Hartree-Fock (HF)

2.1 Approximation du champ moyen de Hartree

L'approximation du champ moyen, proposée par Hartree [3] en 1927, stipule que la répulsion d'un électron i avec les autres électrons $j, (j \neq i)$ est remplacée par l'interaction de celui-ci avec un champ moyen formé par la totalité des autres électrons.

Le potentiel bi-électronique est alors remplacé par un potentiel mono-électronique

$$\text{moyen de l'électron } i. \sum_j e^2 / r_{ij} = U(i) \quad (7)$$

L'approximation de Hartree permet d'exprimer l'hamiltonien total comme suit :

$$H = \sum_{i=1}^n \left(-\frac{\hbar^2}{2m} \Delta_i - \sum_{k=1}^N \frac{Z_k e^2}{r_{ik}} \right) + \sum_{i=1}^n U(i) \quad (8)$$

$$h^c(i) = \sum_{i=1}^n \left(-\frac{\hbar^2}{2m} \Delta_i - \sum_{k=1}^N \frac{Z_k e^2}{r_{ik}} \right) \quad (9)$$

$h^c(i)$: Hamiltonien mono-électronique de cœur.

En se basant sur le théorème des électrons indépendants, la fonction d'onde totale peut être écrite comme le produit de fonctions d'onde mono-électroniques.

$$\Psi_{\text{Tot}} = \prod_{i=1}^n \phi_i(i) = \Psi_1(1) \Psi_2(2) \dots \Psi_n(n) \quad (10)$$

L'énergie de Hartree est donnée par :

$$E^H = \sum_{i=1}^n H_{ii}^c + \frac{1}{2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n J_{ij} \quad (11)$$

$$H_{ii}^c = \int \Psi_i h^c(i) \Psi_i d\tau \quad (12)$$

H_{ii}^c : Intégrale mono-électronique de cœur.

$$J_{ij} = \iint \Psi_i^2(i) \frac{e^2}{r_{ij}} \Psi_j^2(j) d\tau_i d\tau_j \quad (13)$$

J_{ij} : Intégrale bi-électronique coulombienne.

2.2 Équations de Hartree-Fock

La fonction d'onde poly-électronique de Hartree (Eq.10) ne vérifie ni le principe d'indiscernabilité des électrons ni le principe d'exclusion de Pauli [4]. Pour tenir compte de ces deux principes, Fock [5] a proposé d'écrire la fonction d'onde totale sous forme de déterminant de Slater [6].

$$\Psi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_1(2) & \dots & \phi_1(N) \\ \phi_2(1) & \phi_2(2) & \dots & \phi_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_N(1) & \phi_N(2) & \dots & \phi_N(N) \end{vmatrix} \quad \begin{aligned} \phi_1(1) &= \Psi_1(1)\alpha_1(1) \\ \phi_1(2) &= \Psi_1(2)\alpha_1(2) \end{aligned} \quad (14)$$

La fonction d'onde construite sous la forme de déterminant de Slater est utilisée pour calculer l'énergie électronique qui se décompose en une somme de termes mono et bi électroniques :

$$E^{HF} = 2 \sum_i H_{ii} + \sum_i \sum_{j>i} (2J_{ij} - K_{ij}) \quad (15)$$

$$K_{ij} = \iint \Psi_i(i)\Psi_j(i) \frac{e^2}{r_{ij}} \Psi_i(j)\Psi_j(j) d\tau_i d\tau_j \quad (16)$$

K_{ij} : Intégrale bi-électronique d'échange résulte de la nature antisymétrique de la fonction d'onde multiélectronique. L'ajout de ce terme à l'énergie de Hartree a permis sa diminution ($E^{HF} < E^H$), car Fock a tenu compte du principe de Pauli et d'indiscernabilité.

3. Méthode de Hartree-Fock-Roothaan

Pour calculer l'énergie d'une molécule avec la méthode HF, il est nécessaire de connaître les expressions analytiques des intégrales H_{ii}, J_{ij}, K_{ij} qui sont en fonction des orbitales moléculaires inconnues en 1928. Pour remédier à ce point, Roothaan a débloqué

la situation en utilisant la méthode variationnelle OM-CLOA (Combinaison Linéaire d'Orbitales Atomiques).

La technique OM-CLOA consiste à chercher une fonction inconnue (Orbitale Moléculaire Ψ_k) en utilisant une combinaison linéaire de fonctions connues (Orbitales Atomiques φ_r) et des coefficients à faire varier, d'où l'appellation méthode de variation. Dans le cadre de cette approximation, il s'agit de trouver les meilleurs coefficients C_{kr} qui minimisent l'énergie électronique.

$$\Psi_k = \sum_{k=1}^N C_{kr} \varphi_r \quad (17)$$

Après certaines manipulations algébriques, on aboutit aux équations de Roothaan définies par le système séculaire suivant [7]:

$$\sum_{s=1}^N C_{ks} (F_{rs} - \varepsilon_k S_{rs}) = 0 \quad r = 1, \dots, N \quad (18)$$

N : Nombre d'orbitales atomiques combinées.

C_{ks} : Coefficients à faire varier.

S_{rs} : Intégrale de recouvrement.

$$S_{rs} = \int \varphi_{r(i)} \varphi_{s(i)} d\tau \quad (19)$$

F_{rs} : Élément matriciel de Fock.

$$F_{rs} = h_{rs}^c + \sum_{p=1}^n \sum_{q=1}^n P_{pq} \{2 \langle rs|tu \rangle - \langle rt|su \rangle\} \quad (20)$$

Où r, s, p et q symbolisent les OA, P_{pq} est l'élément de la matrice densité.

h_{rs}^c : Hamiltonien mono-électronique de cœur.

$$h_{rs}^c = \int \varphi_{r(i)}^* h^c \varphi_{s(i)} d\tau \quad (21)$$

$\langle rs|tu \rangle$: Intégrale bi-électronique coulombienne.

$$\langle rs|tu \rangle = \iint \varphi_{r(i)} \varphi_{s(i)} \frac{e^2}{r_{ij}} \varphi_{t(j)} \varphi_{u(j)} d\tau_i d\tau_j \quad (22)$$

$\langle ru|ts \rangle$: Intégrale bi-électronique d'échange.

$$\langle ru|ts \rangle = \iint \varphi_{r(i)} \varphi_{u(i)} \frac{e^2}{r_{ij}} \varphi_{t(j)} \varphi_{s(j)} d\tau_i d\tau_j \quad (23)$$

4. Méthodes Post-SCF

Le principal problème posé par la méthode Hartree-Fock-Roothaan, découle du fait que la corrélation existant entre les mouvements des électrons n'est pas prise en compte. Ceci rend cette méthode relativement restreinte dans le calcul quantitatif des propriétés thermodynamiques, telles que l'enthalpie d'activation, l'énergie de Gibbs de réactions, énergies de dissociation.

Ces limitations ont mené au développement de nouvelles méthodes, permettant de calculer ces propriétés d'une manière efficace, tout en prenant compte la corrélation électronique. C'est le cas des méthodes post-SCF (interaction de configurations (CI) [8,9] et la théorie de perturbation Møller-Plesset d'ordre n (MP_n)) ainsi que les méthodes DFT. L'énergie de corrélation d'un système, correspond à la différence entre l'énergie exacte non-relativiste et l'énergie Hartree-Fock obtenue avec la base la plus étendue possible.

$$E_{\text{corr}} = E_{\text{exacte}} - E_{\text{HF}} \quad (24)$$

Les techniques Post-HF s'avèrent très coûteuses en temps de calcul pour des systèmes de grande taille, d'où la nécessité de faire appel à la théorie de la fonctionnelle de densité.

5. Théorie de la Fonctionnelle de Densité (DFT)

La théorie de la fonctionnelle de la densité (DFT, *Density Functional Theory*) est l'une des méthodes les plus utilisées dans les calculs quantiques en raison de son application possible à des systèmes de tailles très variées, allant de quelques atomes à plusieurs centaines d'atomes, tout en incluant les effets de corrélation électronique.

La théorie de DFT diffère conceptuellement des approches précédentes, en prenant pour propriété fondamentale, la densité électronique. Cette approche a pour origine le postulat de Thomas et Fermi, selon lequel l'énergie du système, dans son état fondamental, peut être décrite sous la forme d'une fonctionnelle de la densité électronique $\rho(\vec{r})$ qui peut, elle-même, être reliée à la fonction d'onde Ψ .

$$E_0 = E[\rho] \quad (25)$$

$$\rho(\vec{r}) = \Psi^*(\vec{r}) \Psi(\vec{r}) = |\Psi^2(\vec{r})| \quad (26)$$

L'objectif principal de la théorie DFT est de remplacer la fonction d'onde multiélectronique qui dépend de $4N$ variables (3 coordonnées d'espace (x,y,z) et une variable de spin $(\alpha$ ou $\beta)$), par la densité électronique en tant que quantité de base pour

les calculs, dépendant de trois variables (x, y, z) seulement. Il s'agit donc d'une quantité plus facile à traiter tant mathématiquement que conceptuellement.

5.1 Théorèmes de Hohenberg-Kohn

5.1.1 Premier théorème

« L'énergie moléculaire, la fonction d'onde et toutes les autres propriétés électroniques de l'état fondamental sont déterminées à partir de la densité électronique de l'état fondamental $\rho_0(x, y, z)$ » [10].

$$\int \rho_0(\mathbf{r}) \, d\mathbf{r} = n \quad (27)$$

$$\text{Avec : } \rho = \rho(\vec{r}) \quad (28)$$

$\rho_0(\mathbf{r})$: exprime la densité ponctuelle au point \mathbf{r} , son intégrale sur tout l'espace donne le nombre d'électrons. La densité électronique peut être alors utilisée comme variable de base pour la résolution de l'équation de Schrödinger électronique.

L'hamiltonien électronique d'un système poly-électronique s'écrit :

$$H = -\frac{1}{2} \sum_i^n \Delta_i + \sum_{i>j}^n \frac{1}{r_{ij}} + \sum_i^n V(\mathbf{r}_i) \quad (29)$$

Avec :

$$V(\mathbf{r}_i) = - \sum_{\alpha} \frac{Z_{\alpha} e^2}{r_{i\alpha}} \quad (30)$$

$V(\mathbf{r}_i)$: Potentiel externe de l'électron i .

Ce potentiel correspond à l'attraction de l'électron (i) avec tous les noyaux qui sont externes par rapport au système d'électrons.

L'énergie totale $E[\rho(\mathbf{r})]$ peut s'écrire comme une somme de trois fonctionnelles :

$$E[\rho] = T[\rho] + V_{ee}[\rho] + V_{eN}[\rho] \quad (31)$$

Où :

$$T[\rho] = \int \left[-\frac{1}{2} \nabla^2 \rho(\mathbf{r}) \right] d\mathbf{r} \quad (32)$$

Énergie cinétique des électrons.

$$V_{eN}[\rho] = \int \rho(\mathbf{r}) v(\mathbf{r}) d\mathbf{r} \quad (33)$$

Énergie potentielle d'attraction noyau-électron.

L'énergie potentielle de répulsion électron-électron $V_{ee}[\rho]$ est composée de deux parties :

Interaction coulombienne classique

$$J[\rho] = \frac{1}{2} \iint \frac{1}{r_{12}} \rho(r_1)\rho(r_2) dr_1 dr_2 \quad (34)$$

Energie d'échange et de corrélation

$$K[\rho] = \frac{1}{4} \iint \frac{1}{r_{12}} \rho(r_1, r_2)\rho(r_1, r_2) dr_1 dr_2 \quad (35)$$

Par conséquent, la fonctionnelle de l'énergie peut s'écrire :

$$E_0[\rho] = \int \rho_0(r) v(r) dr + F[\rho_0] \quad (36)$$

Où :

$$F[\rho_0] = T[\rho_0] + V_{ee}[\rho_0] \quad (37)$$

$F[\rho_0]$ est la fonctionnelle universelle de Hohenberg et Kohn, indépendante du potentiel externe et valable quel que soit le système étudié, mais à l'heure actuelle, sa formule exacte est inconnue.

5.1.2 Second théorème

« Si on suppose que l'énergie de l'état fondamental E_0 correspond à la densité ρ_0 , toutes les autres densités d'essai ($\rho' \neq \rho_0$) donnent des énergies supérieures à E_0 telles que $E[\rho'] > E_0[\rho_0]$ »

Ce théorème applique le principe variationnel : la meilleure densité est celle qui donne l'énergie minimale, c'est à dire que l'état fondamental est caractérisé par une densité unique qui correspond à l'état le plus stable.

5.2 Approche de Kohn-Sham

En raison que Hohenberg et Kohn n'ont pas donné une méthode pour calculer l'énergie E_0 à partir de la densité électronique ρ_0 , ni comment déterminer ρ sans déterminer la fonction d'onde, Kohn et Sham (1955) [11] ont résolu le problème en introduisant un système fictif de référence, noté s , constitué de n électrons non interagissant et ayant la même densité électronique que le système réel.

Étant donné que les électrons n'interagissent pas entre eux dans ce système de référence, donc le terme V_{ee} est éliminé et l'Hamiltonien s'écrit:

$$\hat{H}_s = \sum_{i=1}^n \left[-\frac{1}{2} \int \nabla_i^2 + v_s(r_i) \right] = \sum_{i=1}^n h_i^{KS} \quad (38)$$

$$h_i^{KS} = -\frac{1}{2} \int \nabla_i^2 + v_s(r_i) \quad (39)$$

Les équations de Kohn et Sham, pour l'électron i , s'écrivent alors comme suit :

$$h_i^{KS} \phi_i^{KS} = \epsilon_i^{KS} \phi_i^{KS} \quad (40)$$

ϕ_i^{KS} : Sont les orbitales de Kohn et Sham de l'électron i .

Terme d'échange-corrélation:

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

$$E_{xc}[\rho] = \Delta T[\rho] + \Delta V_{ee}[\rho] \quad (41)$$

ΔT Représente la différence de l'énergie cinétique entre le système réel (électrons interagissant) et le système fictif (électrons non interagissant).

$$\Delta T = T[\rho] - T_s[\rho] \quad (42)$$

ΔV_{ee} Représente la différence entre la vraie répulsion et la répulsion coulombienne.

$$\Delta V_{ee} = V_{ee}[\rho] - \frac{1}{2} \iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 \quad (43)$$

5.3 Approximation de la densité locale (LDA : Local Density Approximation)

Dans le cadre de cette approximation, Hohenberg et Kohn ont montré que la densité varie très lentement avec la position. L'énergie d'échange-corrélation est donnée par la formule suivante :

$$E_{xc}^{LDA}[\rho] = \int \rho(r) \epsilon_{xc} \rho(r) dr \quad (44)$$

ϵ_{xc} est l'énergie d'échange-corrélation par électron dans un gaz électronique homogène (Gellium), cette quantité est exprimée comme la somme des deux contributions : énergie d'échange ϵ_x et énergie de corrélation ϵ_c .

$$\epsilon_{xc}(\rho) = \epsilon_x(\rho) + \epsilon_c(\rho) \quad (45)$$

$$\epsilon_x(\rho) = -\frac{3}{4} \left(\frac{3}{\pi} \right)^{\frac{1}{3}} (\rho(r))^{\frac{1}{3}} \quad (46)$$

Le terme de corrélation $\varepsilon_c(\rho)$ est exprimé par la formule de Vosko, Wilk, et Nusair (VWN) [12].

5.4 Approximation de la densité de spin locale (LSDA)

Pour les molécules à couches ouvertes et les géométries des molécules près de leur état de dissociation, l'approximation LSDA donne des résultats meilleurs que LDA.

Dans LDA, les électrons de spins opposés ont les mêmes orbitales KS spatiales. Par contre, LSDA distingue entre les orbitales des électrons de spins opposés :

$\Phi_{i\alpha}^{KS}$ Pour les électrons de spin α et $\Phi_{i\beta}^{KS}$ pour les électrons de spin β .

$$E_{xc}^{LSDA} = E_{xc}[\rho^\alpha, \rho^\beta] \quad (47)$$

5.5 Approximation du gradient généralisé (GGA)

Les approximations LDA et LSDA sont basées sur le modèle du gaz électronique uniforme dans lequel la densité électronique varie très lentement avec la position.

- Dans LDA, E_{xc}^{LDA} est une fonction de ρ uniquement.
- Dans LSDA, E_{xc}^{LSDA} est une fonction de ρ^α et ρ^β .

L'approximation GGA (Generalized Gradient Approximation) a pu améliorer LDA et LSDA en introduisant les gradients des densités des spins ρ^α et ρ^β dans l'expression de l'énergie d'échange-corrélation, elle peut s'écrire alors comme suit :

$$E_{xc}^{GGA}[\rho^\alpha, \rho^\beta] = \int f(\rho^\alpha(r), \rho^\beta(r)) \nabla(\rho^\alpha(r), \rho^\beta(r)) dr \quad (48)$$

f est une fonction des densités de spin et de leurs gradients.

La difficulté réside dans la recherche d'expressions analytiques de E_{xc}^{GGA} , elle est divisée en deux contributions : échange et corrélation.

$$E_{xc}^{GGA} = E_x^{GGA} + E_c^{GGA} \quad (49)$$

Les contributions d'échange et de corrélation sont communément traitées de manière distincte, pour être par la suite combinées de façon à donner la fonctionnelle complète.

parmi les fonctionnelles développées les plus connues et utilisées, on peut citer :

- * Les fonctionnelles d'échange de Becke (B88) [13] et de Perdew et Wang (PW91) [14].
- * Les fonctionnelles de corrélation de Perdew (P86) [14], de Lee, Yang et Parr (LYP) [15] et de Perdew et Wang (PW91) [14].
- * Les fonctionnelles hybrides telles que B3LYP [15, 16].

Toutes ces fonctionnelles permettent une amélioration vis-à-vis des fonctionnelles LDA au niveau de plusieurs propriétés moléculaires telles que les longueurs et les énergies de liaisons.

5.6 Fonctionnelle hybride B3LYP :

La fonctionnelle B3LYP (Becke-3-parametres-Lee-Young-Parr) [16] est une fonctionnelle à trois paramètres d'ajustement combinant les fonctionnelles d'échange local, d'échange de Becke et d'échange HF, avec les fonctionnelles de corrélation locale de Vosko, wilk, Nusair (VWN) et corrigée du gradient de Lee, Yang et Parr (LYP).

L'énergie de cette fonctionnelle s'écrit :

$$E_{xc}^{B3LYP} = (1 - a_0 - a_x)E_x^{LDA} + a_0 E_x^{HF} + a_x E_x^{B88} + a_c E_c^{LYP} + (1 - a_c)E_c^{VWN} \quad (50)$$

Les 3 paramètres a_0 , a_x et a_c ont été ajustés respectivement à 0.20, 0.72 et 0.81 [16], pour reproduire les meilleurs lissages avec des valeurs expérimentales des énergies d'atomisation des molécules.

Cette fonctionnelle donne des résultats remarquablement précis pour un grand nombre de systèmes [17].

Références bibliographiques

- [1] E. Schrödinger, Quantization as an Eigenvalue Problem, *Annalen der Physik. Leipzig.* 79, 361-376, **1926**.
- [2] M. Born and J.R. Oppenheimer, on the quantum theory of molecules, *Annals of Physics.* 84, 457, **1927**.
- [3] V. Minkine, B. Simkine and R. Minaev, *Théorie de la structure moléculaire*, Edition Mir, Moscou, **1982**.
- [4] W. Pauli, Über den Zusammenhang des Abschlusses der Elektronen gruppe nim Atom mit der Komplexstruktur der Spektren, *Zeitschrift für Physik.* 31, 765-783, **1925**.
- [5] V. Fock, Näherungsmethode zur Losung des quanten-Mechanischen Mehrkörper probleme, *Zeitschrift für Physik.* 61, 126-148, **1930**.
- [6] J.C. Slater, The theory of Complex Spectra, *Physical Review.* 34, 1293, **1929**.
- [7] C.C.J. Roothaan, New developments in molecular orbital theory, *Reviews of Modern Physics.* 23, 69, **1951**.
- [8] I. Shavitt, *Methods of Electronic Structure Theory*, H. F. Shaefer (Ed), Plenum Press, New York. 189, **1977**.
- [9] A. Julg, *Chimie Quantique Structurale et Eléments de Spectroscopie Théorique*, Alger : O.P.U, 431, **1978**.
- [10] P. Hohenberg and W. Kohn, Inhomogeneous Electron Gas, *Physical Review B.* 136, 864-871, **1964**.
- [11] W. Kohn and L.J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects, *Physical Review A.* 140, 1133, **1965**.
- [12] S.H. Vosko, L. Wilk and M. Nusair, Accurate spin dependent electron liquid correlation energies for local spin density calculations: A critical analysis, *Canadian Journal of Physics.* 58, 1200, **1980**.
- [13] A.D. Becke, Density functional exchange energy approximation with correct Asymptotic behavior, *Physical Review A.* 38, 3098, **1988**.
- [14] (a) J.P. Perdew, P. Ziesche and H. Eschrig (Eds), Akademie Verlag, Berlin, *Electronic structure of solids.* 11-20, **1991**. (b) J.P. Perdew, Density-functional approximation for the correlation energy of the inhomogeneous electron gas, *Physical Review B.* 33, 8822, **1986**.
- [15] C. Lee, W. Yang and R.G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, *Physical Review B.* 37, 785, **1988**.
- [16] A.D. Becke, Density-functional thermochemistry. III. The role of exact exchange *Journal of Chemical Physics.* 98, 5648, **1993**.
- [17] V. Barone, Inclusion of Hartree-Fock exchange in the density functional approach. Benchmark computations for di atomic molecules containing H, B, C, N, O, and F atoms, *Chemical Physics Letters.* 226, 392-398, **1994**.

Chapitre III
Résultats et Discussion

Introduction

Le facteur de bioconcentration (BCF) est un paramètre écotoxicologique important décrivant l'accumulation de produits chimiques principalement dans les organismes aquatiques, qui vivent dans des environnements contaminés. L'accumulation de polluants organiques dans les poissons est particulièrement préoccupante car les poissons servent de nourriture à de nombreuses espèces, y compris les humains. Cependant, dans les évaluations des risques humains et environnementaux, les données sur le BCF ne sont généralement pas facilement disponibles étant donné que la détermination expérimentale des valeurs du BCF est coûteuse et prend du temps, d'où la nécessité de faire appel aux modèles QSAR qui sont simples d'utilisation et permettent d'obtenir une valeur de BCF fiable et à moindre coût en utilisant différentes propriétés physico-chimiques.

L'objectif de cette étude est de développer un modèle QSAR d'estimation du facteur de bioconcentration BCF pour une large gamme de composés organiques PBTs en utilisant différents descripteurs et identifiant ceux qui expliquent le mieux ce facteur. Cette étude sera suivie par la prédiction de la toxicité d'une série constituée de 33 composés vis-à-vis les poissons (*Fathead Minnow*) en suivant toute la méthodologie QSAR et en respectant tous les critères d'OECD [1].

Méthodologie

* Les composés étudiés comprenaient différentes familles : MAH (Hydrocarbures Aromatiques Monocycliques), PAH (Hydrocarbures Aromatiques Polycycliques), CB (Chloro-benzène), BB (Bromo-Benzène), CN (Chloro-Naphtalène), PCB (poly-Chloro-Biphényle), PBB (Poly-Bromo-Biphényle), CDF (Chloro-Dibenzo-Furan) et PHENOL.

* Les structures ont été optimisées par le logiciel Gaussian 09 [2], à l'aide de la fonctionnelle hybride B3LYP et la base 6-31 G(d,p), suivie d'un calcul de fréquence afin de s'assurer qu'aucune des structures optimisées ne présentent de fréquences imaginaires.

* Une base de données constituée de 145 composés avec leur valeur de log K_{ow} et log BCF a été collectée à partir des articles disponibles dans la littérature [3,4] et à partir de la base de données online ECHA [5] pour l'étude du facteur de bioconcentration.

Chapitre III : Résultats et discussion

* Une autre base de données constituée de 33 composés (contenus dans la première base du BCF) avec leurs valeurs de pLC₅₀ a été collectées à partir des bases de données online ECOTOX [6] ECHA [5] et les références [7,8] pour l'étude de la toxicité.

* Les techniques d'analyse de données utilisées pour la construction des modèles QSAR sont réalisées sur le programme Minitab 17 [9] et sont les suivantes :

- SLR : pour les modèles à un seul descripteur.
- MLR : pour les modèles à plusieurs descripteurs non corrélés.
- PLS : pour les modèles à plusieurs descripteurs fortement corrélés.
- PCA : pour voir les ressemblances entre variables et individus.

* Les structures ont été réécrites en langage smiles afin de calculer quelques descripteurs moléculaires de type 2D.

* Les descripteurs utilisés pour l'étude QSAR du BCF et de la toxicité sont donnés dans le tableau suivant :

Tableau 1: Descripteurs utilisés pour l'étude QSAR du BCF et de la toxicité.

<i>Descripteurs</i>	<i>signification</i>	<i>logiciels /Expressions</i>
Log Kow	Coefficient de partage octanol/eau	Littérature [3]
TPSA	La surface polaire topologique	molinspiration (Online) [10]
Mw	Poids moléculaire	molinspiration (Online) [10]
Vm	Volume molaire	ChemSpider (Online) [11]
α	polarisabilité	ChemSpider (online) [11]
E_{HOMO}	Energie la plus haute occupée	Gaussian 09W [2]
E_{LUMO}	Energie la plus basse vacante	Gaussian 09W [2]
μ	Potentiel chimique électronique	$\mu = (\epsilon_{HOMO} + \epsilon_{LUMO})/2$
η	Dureté	$\eta = \epsilon_{LUMO} - \epsilon_{HOMO}$
S	Mollesse	$S=1/ (2* \eta)$
Nu	Nucléophilie	$Nu = \epsilon_{HOMO}(Nu) - \epsilon_{HOMO}(TCE)$
ω	Electrophilie	$\omega = \mu^2/ (2* \eta)$

Chapitre III : Résultats et discussion

Tableau 2 : Base de données et descripteurs calculés.

N°	Class	Chemical	log BCF	log Kow	Mw (g/mol)	α (Å ³)	Vm (cm ³ /mol)	TPSA (Å ²)
1	Phenol	1,2,3-trihydroxybenzene	0.1	0.29	126.11	12.6	84.7	60.68
2	Phenol	benzene-1,3-diol	0.42	0.76	110.11	11.9	86.3	40.46
3	Phenol	Catechol	0.55	0.88	110.11	11.9	86.3	40.46
4	MAH	Benzene	0.64	2.13	78.11	10.4	89.4	0
5	Phenol	Pentachlorophenol	1.01	4.78	266.34	20.9	147.6	20.23
6	Phenol	Chlorohydroquinone	1.02	1.52	144.56	13.8	98.2	40.46
7	Phenol	3-methoxyphenol	1.06	1.51	124.14	13.8	111.9	29.46
8	MAH	Toluene	1.12	2.73	92.14	12.3	105.7	0
9	Phenol	2,3,6-Trichlorophenol	1.12	3.33	197.45	17	123.7	20.23
10	MAH	Styrene	1.13	2.95	104.15	14.7	115.4	0
11	MAH	Ethylbenzene	1.19	3.15	106.17	14.2	122.3	0
12	MAH	o-Xylene	1.24	3.12	106.17	14.2	122	0
13	Phenol	Phenol	1.24	1.46	94.11	11.2	87.9	20.23
14	Phenol	1,4-dimethoxybenzene	1.24	2.1	138.17	15.7	137.4	18.47
15	Phenol	3-Chlorophenol	1.25	2.5	128.56	13.1	99.8	20.23
16	MAH	m-Xylene	1.27	3.2	106.17	14.2	122	0
17	MAH	p-Xylene	1.27	3.15	106.17	14.2	122	0
18	Phenol	2,3,4,6-Tetrachlorophenol	1.29	4.17	231.89	18.9	135.7	20.23
19	Phenol	2-methylphenol	1.35	1.94	108.14	13.1	104.1	20.23
20	Phenol	3-methylphenol	1.35	1.94	108.14	13.1	104.1	20.23
21	Phenol	4-methylphenol	1.35	1.94	108.14	13.1	104.1	20.23
22	Phenol	2,6-Dichlorophenol	1.41	2.61	163	15	111.8	20.23
23	MAH	p-Methylstyrene	1.5	3.37	118.18	16.7	131.7	0
24	Phenol	2,4-Dichlorophenol	1.5	3.3	163	15	111.8	20.23
25	MAH	m-Methylstyrene	1.55	3.37	118.18	16.7	131.7	0
26	MAH	Isopropylbenzene	1.55	3.72	120.19	16	139.5	0
27	Phenol	1-hydroxy-2,4-dimethylbenzene	1.56	2.4	122.17	15	120.4	20.23
28	Phenol	1-hydroxy-3,4-dimethylbenzene	1.56	2.4	122.17	15	120.4	20.23
29	Phenol	Hydroquinone	1.6	0.55	110.11	11.9	86.3	40.46
30	Phenol	1-hydroxy-2,6-dimethylbenzene	1.62	2.4	122.17	15	120.4	20.23
31	PAH	Naphthalene	1.64	4.7	128.17	17.5	123.5	0
32	Phenol	4-chlorophenol	1.66	2.43	128.56	13.1	99.8	20.23
33	BB	Bromobenzene	1.7	2.99	157.01	13.5	105.6	0
34	Phenol	2,3,5-Trichlorophenol	1.71	3.69	197.45	17	123.7	20.23
35	Phenol	2,5-Dichlorophenol	1.83	2.88	163	15	111.8	20.23
36	CB	Chlorobenzene	1.85	2.84	112.56	12.3	101.4	0
37	Phenol	2,3-Dichlorophenol	1.85	2.83	163	15	111.8	20.23
38	Phenol	2,3,4,5-Tetrachlorophenol	1.85	4.39	231.89	18.9	135.7	20.23
39	MAH	1-chloro-2-methyl-4-hydroxybenzene	1.92	2.89	142.59	15	116.1	20.23
40	Phenol	2,4,5-Trichlorophenol	2.12	3.71	197.45	17	123.7	20.23
41	Phenol	2,3,4-Trichlorophenol	2.13	3.66	197.45	17	123.7	20.23
42	Phenol	2,3,5,6-Tetrachlorophenol	2.15	4.39	231.89	18.9	135.7	20.23
43	Phenol	3,5-Dichlorophenol	2.21	3.33	163	15	111.8	20.23
44	Phenol	3,4-Dichlorophenol	2.22	3.22	163	15	111.8	20.23

Chapitre III : Résultats et discussion

45	MAH	1-chloro-4-methylbenzene	2.25	3.27	126.59	14.3	117.7	0
46	Phenol	2-Chlorophenol	2.33	2.16	128.56	13.1	99.8	20.23
47	Phenol	2,4,6-Trichlorophenol	2.43	3.06	197.45	17	123.7	20.23
48	CB	1,2-Dichlorobenzene	2.48	3.71	147	14.3	113.3	0
49	Phenol	3,4,5-Trichlorophenol	2.51	4.02	197.45	17	123.7	20.23
50	CB	1,4-Dichlorobenzene	2.52	3.37	147	14.3	113.3	0
51	CDF	Benzo[b]furan	2.56	2.86	118.14	14.4	106.3	13.14
52	PAH	Acenaphthylene	2.58	3.97	152.2	20.3	128.2	0
53	PAH	Acenaphthene	2.59	3.92	154.21	20.5	134.9	0
54	PAH	Biphenyl	2.64	3.88	154.21	20.2	154.7	0
55	MAH	2-Phenyldodecane	2.65	8.19	246.44	32.6	288.1	0
56	CB	1,3-Dichlorobenzene	2.65	3.44	147	14.3	113.3	0
57	PCB	4-Chlorobiphenyl	2.69	4.63	188.66	24	178.6	0
58	MAH	1,2-dichloro-4-methylbenzene	2.78	3.74	161.03	16.2	129.6	0
59	BB	1,3-Dibromobenzene	2.8	3.78	235.91	16.5	121.8	0
60	PAH	Anthracene	2.83	4.54	178.23	24.6	157.7	0
61	BB	1,4-Dibromobenzene	2.83	3.89	235.91	16.5	121.8	0
62	MAH	1,3-dichloro-4-methylbenzene	2.86	3.88	161.03	16.2	129.6	0
63	CDF	Octachlorodibenzofuran	2.94	8.2	443.75	37	236.1	13.14
64	BB	Hexabromobenzene	3.04	6.07	551.49	28.7	186.5	0
65	BB	1,2-Dibromobenzene	3.1	3.64	235.91	16.5	121.8	0
66	CB	1,2,3-Trichlorobenzene	3.11	4.27	181.45	16.2	125.3	0
67	PAH	2-Methylnaphthalene	3.2	4.11	142.2	19.4	139.8	0
68	PAH	Fluorene	3.23	4.38	166.22	21.3	148.3	0
69	CB	1,2,4-Trichlorobenzene	3.26	4.04	181.45	16.2	125.3	0
70	PCB	2,2'-Dichlorobiphenyl	3.26	5	223.1	24	178.6	0
71	PCB	4,4'-Dichlorobiphenyl	3.28	5.58	223.1	24	178.6	0
72	CDF	Dibenzofuran	3.34	4.21	168.19	21.5	140.5	13.14
73	CB	1,2,3,5-Tetrachlorobenzene	3.36	4.65	215.89	18.2	137.2	0
74	PCB	2,2',4,4',6-Pentachlorobiphenyl	3.37	6.23	326.44	29.9	214.5	0
75	CB	1,3,5-Trichlorobenzene	3.38	4.08	181.45	16.2	125.3	0
76	PAH	Phenanthrene	3.42	4.46	178.23	24.6	157.7	0
77	PAH	Benzo[a]pyrene	3.42	5.97	252.32	35.8	196.1	0
78	PAH	Pyrene	3.43	4.88	202.26	28.7	162	0
79	CN	Octachloronaphthalene	3.44	6.42	403.73	33	219.2	0
80	PAH	2-Methylphenanthrene	3.48	4.86	192.26	26.5	173.9	0
81	CDF	2,3,7,8-Tetrachlorodibenzofuran	3.53	6.53	305.98	29.3	188.3	13.14
82	PCB	2,4'-Dichlorobiphenyl	3.55	5.1	223.1	24	178.6	0
83	CN	1,4-Dichloronaphthalene	3.56	4.88	197.06	21.4	147.5	0
84	CDF	1,2,3,4,6,7,8-Heptachlorodibenzofuran	3.62	7.92	409.31	35.1	224.1	13.14
85	PAH	2-Chlorophenanthrene	3.63	4.07	212.68	26.5	169.6	0
86	CN	2-Monochloronaphthalene	3.63	3.9	162.62	19.4	135.5	0
87	PAH	9-Methylanthracene	3.66	5.07	192.26	26.5	173.9	0
88	BB	1,2,4-Tribromobenzene	3.66	4.54	314.8	19.6	138	0
89	PCB	2,4',5-Trichlorobiphenyl	3.75	5.67	257.55	26	190.6	0

Chapitre III : Résultats et discussion

90	CB	1,2,4,5-Tetrachlorobenzene	3.76	4.67	215.89	18.2	137.2	0
91	CB	1,2,3,4-Tetrachlorobenzene	3.77	4.65	215.89	18.2	137.2	0
92	PCB	3,5-Dichlorobiphenyl	3.78	5.37	223.1	24	178.6	0
93	BB	1,2,4,5-Tetrabromobenzene	3.79	5.13	393.7	22.6	154.2	0
94	CN	1,8-Dichloronaphthalene	3.79	4.41	197.06	21.4	147.5	0
95	BB	1,3,5-Tribromobenzene	3.85	4.51	314.8	19.6	138	0
96	PCB	2,2',6,6'-Tetrachlorobiphenyl	3.85	5.21	291.99	27.9	202.5	0
97	CB	Pentachlorobenzene	3.86	5.18	250.34	20.1	149.2	0
98	CB	2,4,5-Trichlorotoluene	3.87	4.8	195.48	18.1	141.6	0
99	PCB	3,3',4,4'-Tetrachlorobiphenyl	3.9	6.36	291.99	27.9	202.5	0
100	PBB	2,4,6-Tribromobiphenyl	3.93	6.03	390.9	29.3	203.3	0
101	PBB	2,2',4,4',6,6'-Hexabromobiphenyl	3.96	7.2	627.59	38.4	251.8	0
102	PAH	Benzo[a]anthracene	4	5.61	228.29	31.6	191.8	0
103	PCB	2,2',4,4'-Tetrachlorobiphenyl	4.02	6.11	291.99	27.9	202.5	0
104	PCB	2,4,5-Trichlorobiphenyl	4.02	5.9	257.55	26	190.6	0
105	PCB	2,2',3,3',4,4',5,5',6,6'-decachlorobiphenyl	4.02	8.18	498.66	39.6	274.2	0
106	CDF	2,3,4,7,8-Pentachlorodibenzofuran	4.03	6.92	340.42	31.2	200.2	13.14
107	CN	2,3-Dichloronaphthalene	4.04	4.71	197.06	21.4	147.5	0
108	CN	2,7-Dichloronaphthalene	4.04	4.81	197.06	21.4	147.5	0
109	CN	1,2,3,4-Tetrachloronaphthalene	4.1	5.5	265.95	25.2	171.4	0
110	PBB	4,4'-Dibromobiphenyl	4.19	5.72	312	26.3	187.1	0
111	PCB	2,5-Dichlorobiphenyl	4.2	5.16	223.1	24	178.6	0
112	PCB	2,2',3,3'-Tetrachlorobiphenyl	4.23	6.18	291.99	27.9	202.5	0
113	PCB	2,3-Dichlorobiphenyl	4.25	5.2	223.1	24	178.6	0
114	CB	Hexachlorobenzene	4.26	5.73	284.78	22.1	161.1	0
115	PCB	2,2',5-Trichlorobiphenyl	4.27	5.6	257.55	26	190.6	0
116	CN	1,3,5,8-Tetrachloronaphthalene	4.4	5.96	265.95	25.2	171.4	0
117	CN	1,3,7-Trichloronaphthalene	4.43	5.59	231.51	23.3	159.4	0
118	MAH	Octachlorostyrene	4.52	6.29	379.71	29.7	211	0
119	CN	1,3,5,7-Tetrachloronaphthalene	4.53	6.38	265.95	25.2	171.4	0
120	PCB	2,4,4'-Trichlorobiphenyl	4.63	5.62	257.55	26	190.6	0
121	PCB	2,3',4',5-Tetrachlorobiphenyl	4.77	5.6	291.99	27.9	202.5	0
122	PBB	2,2',5,5'-Tetrabromobiphenyl	4.8	6.5	469.8	32.4	219.5	0
123	PCB	2,2',4,4',5,5'-Hexachlorobiphenyl	4.83	6.92	360.88	31.8	226.4	0

Chapitre III : Résultats et discussion

124	PCB	2,2',3,5'-Tetrachlorobiphenyl	4.84	5.75	291.99	27.9	202.5	0
125	PCB	2,2',4,5'-Tetrachlorobiphenyl	4.84	5.85	291.99	27.9	202.5	0
126	PCB	2,2',5,5'-Tetrachlorobiphenyl	4.87	6.1	291.99	27.9	202.5	0
127	PCB	2,2',4,4',6,6'-Hexachlorobiphenyl	4.93	6.54	360.88	48	226.4	0
128	PCB	2,2',4,5'-Tetrachlorobiphenyl	5	5.85	291.99	27.9	202.5	0
129	PCB	2,2',3,3',4,4',5,5'-Octachlorobiphenyl	5.08	7.8	429.77	35.7	250.3	0
130	PCB	2,2',3,4,5'-Pentachlorobiphenyl	5.38	6.29	326.44	29.9	214.5	0
131	PCB	2,2',4,5,5'-Pentachlorobiphenyl	5.4	6.38	326.44	29.9	214.5	0
132	PCB	2,2',3',4,5'-Pentachlorobiphenyl	5.43	6.29	326.44	29.9	214.5	0
133	PCB	2,2',3,3',6,6'-Hexachlorobiphenyl	5.43	6.22	360.88	31.8	226.4	0
134	PCB	2,2',3,5,5',6'-Hexachlorobiphenyl	5.54	6.64	360.88	31.8	226.4	0
135	PCB	2,2',3,3',4,4',5,5',6-Nonachlorobiphenyl	5.71	8.09	464.22	37.6	262.3	0
136	PCB	2,2',3,3',4,4'-Hexachlorobiphenyl	5.77	6.74	360.88	31.8	226.4	0
137	PCB	3,3',4,4',5'-Pentachlorobiphenyl	5.81	6.89	326.44	29.9	214.5	0
138	PCB	2,2',3,4,5,5'-Hexachlorobiphenyl	5.81	6.82	360.88	31.8	226.4	0
139	PCB	2,2',3,3',5,5',6,6'-Octachlorobiphenyl	5.82	7.24	429.77	35.7	250.3	0
140	PCB	2,2',3,4,4',5',6-Heptachlorobiphenyl	5.84	7.2	395.33	33.7	238.4	0
141	PCB	2,2',3,4,4',5'-Hexachlorobiphenyl	5.88	6.83	360.88	31.8	226.4	0
142	PCB	2,2',3,3',4,5,5',6-Octachlorobiphenyl	5.88	7.62	429.77	35.7	250.3	0
143	PCB	2,2',3,3',4,4',5,6-Octachlorobiphenyl	5.92	7.56	429.77	35.7	250.3	0
144	PCB	2,2',3,4,5,5',6'-Heptachlorobiphenyl	5.93	7.11	395.33	33.7	238.4	0
145	PCB	3,3',4,4',5,5'-Hexachlorobiphenyl	5.97	7.42	360.88	31.8	226.4	0

Les 145 composés de la base de données classés par ordre croissant de log BCF se divisent en deux sous séries, une série d'apprentissage constituée de 109 composés et les 36 molécules restantes forment la série de test.

III.1 Etude du facteur de bioconcentration

Relation entre le facteur de bioconcentration BCF et le coefficient de partage K_{ow}

Lier la bioconcentration aux paramètres de l'hydrophobicité comme le coefficient de partage octanol/eau (K_{ow}) est la méthodologie habituelle pour l'estimation du BCF, son utilisation reste la méthode la plus répandue, la plus commune et la plus importante.

La bioconcentration est supposée être un processus de partition entre l'eau et le lipide des poissons, et est donc modélisée en utilisant K_{ow} comme substitut des lipides biologiques.

III.1.1 Modèle avec un seul descripteur

A/ Pour toute la base de données

$$\log BCF = -0.085 + 0.7062 \log K_{ow} \quad (n=145) \quad (1)$$

Le modèle présente un bon coefficient de détermination avec une faible valeur de la déviation standard ($R^2 = 72.79 \%$, $SD=0.767376$), le coefficient de partage explique donc assez bien le facteur de bioconcentration.

Cependant, pour certaines familles de PBTs, le coefficient de partage est insuffisant pour expliquer le BCF, le R^2 entre le $\log BCF$ et $\log K_{ow}$ est très faible comme on le remarque sur le tableau ci-dessous :

Tableau 3 : Relation entre $\log K_{ow}$ et $\log BCF$.

<i>famille</i>	<i>MAH</i>	<i>PAH</i>	<i>CB-BB</i>	<i>CN</i>	<i>PCB-PBB</i>	<i>CDF</i>	<i>Phénols</i>
R^2	0.4928	0.2042	0.6346	0.1671	0.4481	0.2629	0.3584

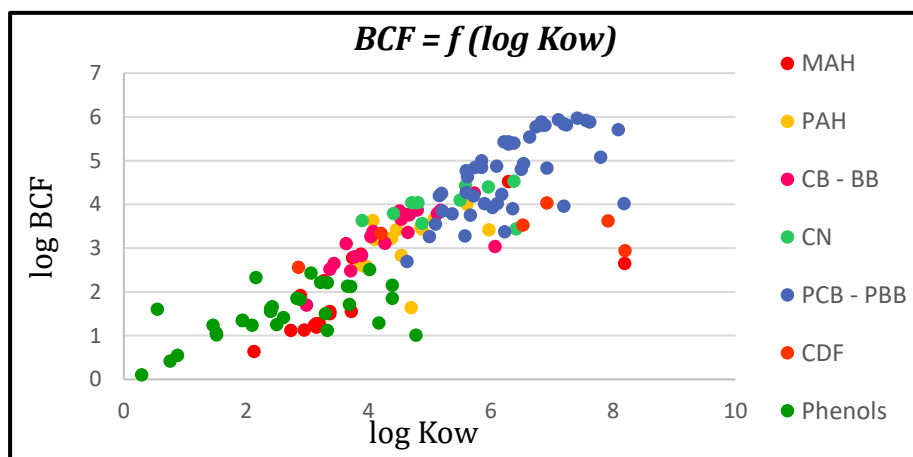


Figure 1 : Relation entre $\log K_{ow}$ et $\log BCF$ pour chaque famille de composés PBTs.

B/ Pour la série d'apprentissage

$$\log \text{BCF} = 0.036 + 0.6770 \log K_{ow} \quad (\mathbf{n=109}) \quad (2)$$

Le modèle obtenu donne de bon paramètres statistiques ($R^2=71.19\%$, $SD=0.794887$, $F\text{-Ficher}=264.35$) mais l'analyse des résidus standardisés a montré l'existence de trois valeurs aberrantes (outliers) avec un résiduel qui dépasse 2.5, Il s'agit de Pentachlorophenol (N°5), 2-Phenyldodecane (N°55) et Octachloro-dibenzofuran (N°63). Leur élimination améliore forcément la qualité du modèle.

➤ **Élimination des observations aberrantes avec un résidu standardisé élevé**

Tableau 4 : Modèles QSAR du BCF avec un seul descripteur (SLR) sans observations aberrantes.

<i>n</i>	<i>Modèle</i>	<i>R²</i>	<i>SD</i>	<i>F</i>
106	$\log \text{BCF} = -0.169 + 0.7379 \log K_{ow}$	80.36 %	0.658265	425.64

On remarque qu'après l'élimination des trois composés dont le résidu standardisé dépasse la valeur seuil **2,5** (en valeur absolue) exigée par MINITAB, le modèle obtenu a pu présenter de bons paramètres statistiques, mais on peut encore améliorer sa qualité en éliminant les molécules hydrophiles et hyper-lipophiles.

- Une substance hydrophobe et liposoluble est potentiellement bioaccumulable dans les tissus des espèces, cela veut dire que les molécules hydrophiles ne participent pas au processus de bioconcentration.
- Plusieurs travaux [12-14] ont montré que pour les composés hyper-lipophiles avec $\log K_{ow}$ élevé, le phénomène de bioconcentration ne se présente pas.

➤ **Élimination des molécules hydrophiles et hyper-lipophiles**

Tableau 5 : Modèles QSAR du BCF avec un seul descripteur (SLR) sans molécules hydrophiles, sans molécules hyper-lipophiles et sans observations aberrantes.

<i>n</i>	<i>Modèle</i>	<i>R²</i>	<i>SD</i>	<i>F</i>
98	$\log \text{BCF} = -0.361 + 0.7870 \log K_{ow}$	82.12 %	0.592200	440.78

L'élimination des molécules ayant un $\log K_{ow} < 1$ (molécules hydrophiles), $\log K_{ow}$ élevé > 8 (molécules hyper-lipophiles) [12] et les outliers (résiduel standardisé $> 2,5$) a permis d'améliorer d'une manière très significative la qualité du modèle.

Dans le but d'obtenir des modèles plus explicatifs, il est nécessaire d'élaborer des modèles avec plusieurs descripteurs (2 et 3 descripteurs).

III.1.2 Modèles avec plusieurs descripteurs (lipophilie + paramètres taille, effet stérique, polarité...)

Pour certaines familles de POPs/PBTs, la bioaccumulation ne dépend pas uniquement du coefficient de partage, il existe d'autres facteurs (taille, polarité ...) [15,16] qui déterminent ce phénomène d'accumulation.

Pour élaborer des modèles avec plus d'un descripteur, l'analyse de la matrice de corrélation est indispensable afin de vérifier les colinéarités des descripteurs.

Matrice de corrélation

	log Kow	Mw	α	Vm
Mw	0.876			
α	0.910	0.857		
Vm	0.940	0.877	0.957	
TPSA	-0.621	-0.401	-0.488	-0.532

D'après la matrice de corrélation, les descripteurs sont pratiquement tous corrélés entre eux, on ne peut alors les combiner en utilisant la régression linéaire multiple (MLR), d'où la nécessité d'utiliser la PLS (partial least square).

III.1.2.1 Modèles avec deux descripteurs

Tableau 6 : Modèles QSAR du BCF avec deux descripteurs en utilisant la méthode PLS.

Modèles (n=98)	R²	R² cv	F
$\log BCF = -0.362588 + 0.790848 \log Kow - 0.000069 Mw$	82.12 %	80.13 %	218.10
$\log BCF = -0.354407 + 0.799008 \log Kow - 0.002793 \alpha$	82.12 %	81.04 %	218.15
$\log BCF = -0.426569 + 0.742914 \log Kow + 0.001696 Vm$	82.15 %	81.02 %	218.60
$\log BCF = -0.042828 + 0.735002 \log Kow - 0.014392 TPSA$	82.69 %	81.42 %	226.87

Le meilleur modèle QSAR obtenu avec la méthode d'analyse PLS est celui qui combine les deux descripteurs : le coefficient de partage ($\log K_{ow}$) et la surface polaire topologique (TPSA) qui est très utilisé dans l'étude des médicaments (propriétés ADMET) [17]. Ce dernier modèle a donné des paramètres statistiques très satisfaisants.

Tableau 7: Contribution des composantes dans l'explication du facteur de bioconcentration.

log BCF = f (log K_{ow}, TPSA)			
Composantes	R²	R² cv	PRESS
1	76.48 %	75.20 %	46.6767
2	82.69 %	81.42 %	34.9743

L'ajout d'une deuxième composante a fait augmenter le R² cv et diminuer le PRESS, donc elle est importante et le modèle QSAR est obtenu avec la combinaison des deux composantes.

III.1.2.2 Modèles avec trois descripteurs

Tableau 8 : Modèles QSAR du BCF avec trois descripteurs en utilisant la méthode PLS.

Modèles (n=98)	R²	R² cv	F
log BCF = -0.043602 + 0.709227 log K _{ow} + 0.005541 α - 0.014956 TPSA	82.70 %	81.14 %	149.81
log BCF = -0.136994 + 0.647981 log K _{ow} + 0.003168 V _m - 0.015664 TPSA	82.80 %	81.20 %	150.86

La combinaison de trois descripteurs a donné pratiquement les mêmes paramètres statistiques obtenus dans le **tableau 6**, cela veut dire que l'ajout d'un troisième descripteur n'a pas amélioré la qualité du modèle, on peut donc se limiter au modèle à deux variables qui combine le coefficient de partage et la surface polaire topologique et étudier ensuite sa validation.

III.1.3 Validation du meilleur modèle

III.1.3.1 Validation interne

Deux techniques sont employées sur la série d'apprentissage pour vérifier la stabilité et le pouvoir explicatif du modèle final (log BCF = f (log K_{ow} + TPSA)).

➤ **Validation croisée (Cross Validation CV) :**

LOO (Leave One Out): Cette étape consiste à tester l'influence de chaque échantillon sur le modèle final et est quantifiée par le R^2_{cv} .

Dans le meilleur modèle QSAR, le $R^2_{cv}=81.42\%$ est assez proche de $R^2=82.69\%$, cela prouve qu'il n'est pas sensible à l'opération de mettre à part une molécule et de la remettre dans la série d'apprentissage. C'est une première indication sur la stabilité du modèle QSAR retenu.

➤ **Y-Randomisation :**

Dans le but de prouver que le meilleur modèle obtenu ne présente pas de bons paramètres statistiques par un pur hasard, la technique Y-randomisation est appliquée sur la série d'apprentissage.

Le tableau ci-dessous regroupe les 10 premières randomisations pour le modèle QSAR retenu ($\log BCF=f(\log Kow + TPSA)$).

Tableau 9 : Randomisation pour le meilleur modèle retenu dans l'étude du BCF.

$\log BCF = -0.042828 + 0.735002 \log Kow - 0.014392 TPSA$					
Itération	R^2_r	$R^2_{r cv}$	Itération	R^2_r	$R^2_{r cv}$
1	3.69 %	0	6	4.67 %	0
2	1.67 %	0	7	0.99 %	0
3	1.02 %	0	8	2.22 %	0
4	1.12 %	0	9	2.95 %	0
5	2.64 %	0	10	2.13 %	0

Les modèles élaborés avec les valeurs randomisées du log BCF ont donné des R^2_r inférieur à 10%. Ces résultats indiquent que le meilleur modèle n'a pas été résulté d'une corrélation due au hasard et il est très stable.

D'après les résultats de la validation interne on peut conclure que le modèle est stable, fiable et possède un grand pouvoir explicatif.

III.1.3.2 Validation externe

Afin de vérifier le pouvoir prédictif du meilleur modèle QSAR élaboré, une série de test est utilisée. Cette validation se fait en deux étapes :

- Vérification des paramètres statistiques de la série de test.
- Vérification des critères de Tropsha.

L'analyse des résidus standardisés a montré l'existence de deux valeurs aberrantes (outliers : observations aberrantes). L'élimination de ces deux molécules de la série de test a permis d'obtenir les résultats ci-dessous :

Tableau 10 : Régression entre $\log BCF_{exp}$ et $\log BCF_{cal}$.

Equation de Régression (série de test)	R^2	$R^2 cv$	SD
$\log BCF_{cal} = f(\log BCF_{exp})$	90.70 %	89.51 %	0.360512

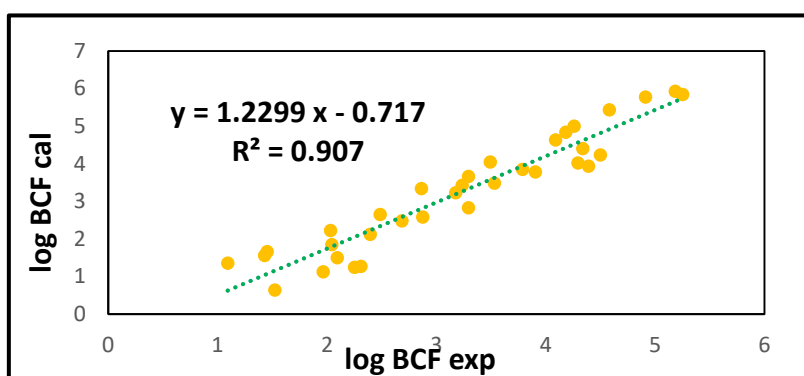


Figure 2 : Régression (série de test) entre $\log BCF_{cal}$ et $\log BCF_{exp}$.

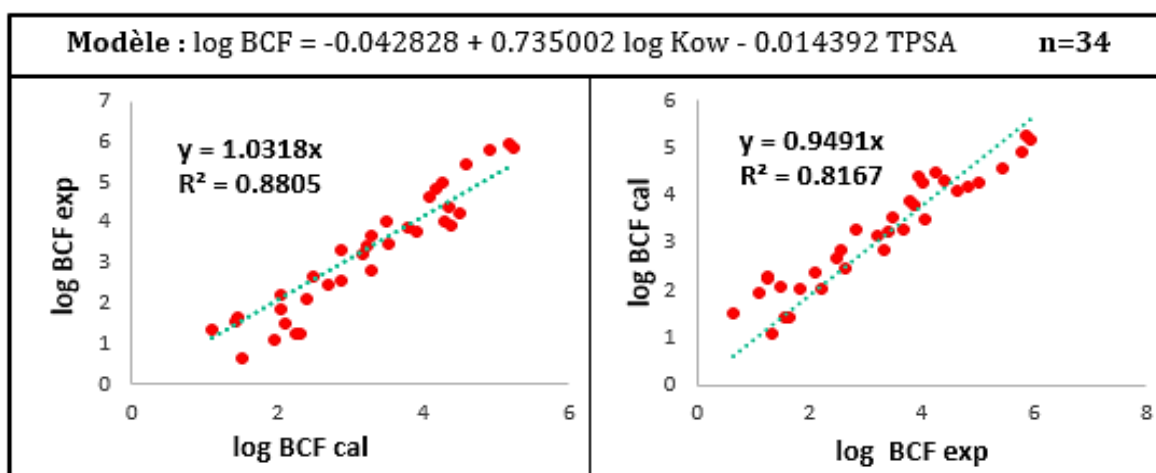


Figure 3 : a- Variation de $\log BCF_{exp}$ en fonction de $\log BCF_{cal}$ (avec intercepte =0).
b- Variation de $\log BCF_{cal}$ en fonction de $\log BCF_{exp}$ (avec intercepte =0).

Tableau 11 : Critères de Tropsha pour la série de test (Etude du BCF).

Modèle : $\log \text{BCF} = -0.042828 + 0.735002 \log \text{Kow} - 0.014392 \text{TPSA}$								n=34
R^2	R^2_{cv}	R_0^2	$R_0'^2$	k	K'	$\frac{R^2 - R_0^2}{R^2}$	$\frac{R^2 - R_0'^2}{R^2}$	$ R^2 - R_0^2 $
90.70 %	89.33 %	88.05 %	81.67 %	1.03	0.95	0.029	0.099	0.0265

- $R^2 > 0.7$
- $R^2_{cv} > 0.6$
- $\frac{R^2 - R_0^2}{R^2} < 0.1$ et $0.85 \leq k \leq 1.15$
- $\frac{R^2 - R_0'^2}{R^2} < 0.1$ et $0.85 \leq k' \leq 1.15$
- $|R^2 - R_0^2| \leq 0.3$

Les résultats de la validation externe du meilleur modèle QSAR montrent que :

- ✓ Les paramètres statistiques (R^2 , R^2_{cv} et SD) de la série de test sont très satisfaisants.
- ✓ Tous les critères de Tropsha sont vérifiés.
- ✓ Les valeurs de k et K' sont assez proches de 1, ce qui indique que la valeur du facteur de bioconcentration calculée en utilisant l'équation correspondant au meilleur modèle est très proche à la valeur expérimentale.
- ✓ Le meilleur modèle QSAR possède un fort pouvoir prédictif.

III.1.4 Domaine d'applicabilité

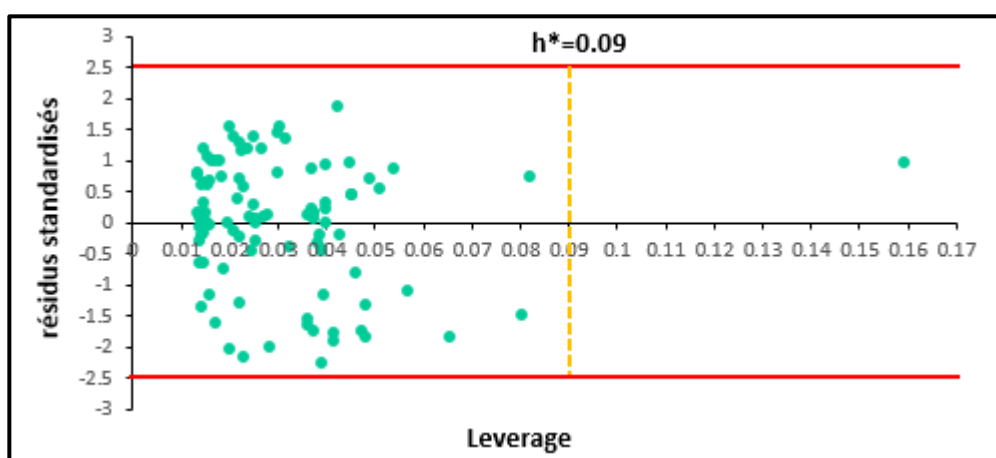


Figure 4 : Domaine d'applicabilité du meilleur modèle.

D'après le graphique du domaine d'applicabilité, on constate la présence d'une seule observation aberrante ayant un *leverage* supérieur à la valeur seuil $h^*=3(P+1)/n$. Avec : $P=2$ nombre de variables explicatives du modèle et $n=98$ nombre d'observations. Ce composé (Chlorohydroquinone N° 6) est considéré comme *outlier*, hors du domaine d'applicabilité du modèle élaboré et donc la prédiction de sa valeur du facteur de bioconcentration est impossible.

III.1.5 Analyse en composantes principales PCA

Dans le but de voir les ressemblances entre variables (explicatives et à expliquer) et entre variables et individus, l'analyse en composantes principales est utilisée.

Tableau 12 : Valeurs propres de la matrice de corrélation.

Composante	PC1	PC2	PC3
Valeur Propre	2,442	0,464	0,094
Proportion	0,814	0,155	0,031
Cumule	0,814	0,969	1,000

L'analyse du tableau 12 (uniquement la série d'apprentissage) a montré que :

- Trois composantes sont obtenues.
- La première composante est la plus importante (proportion 81,4 %).
- La troisième composante n'a aucune importance (proportion de 3,1 %).
- Le tableau des données (variables (log BCF, log K_{ow}, TPSA) et individus (composés 1,2...)) peut être donc représenté par deux composantes principales PC1 et PC2 ce qui est confirmé par la figure ci-dessous :

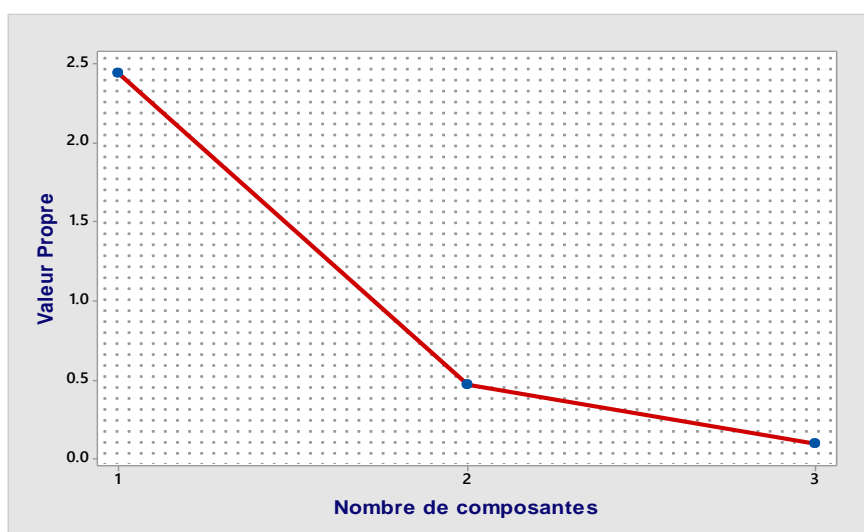


Figure 5 : Variation de la valeur propre avec le nombre de composantes.

La figure 5 indique que la 1^{ère} et la 2^{ème} composante principale expliquent la plus grande part de la variance des données. Par conséquent, les composantes principales restantes expliquent une très faible proportion de la variance (proche de zéro) et sont certainement sans importance.

Les coefficients (vecteurs propres) des variables dans chaque composante sont donnés dans le tableau 13.

Tableau 13 : Coefficients des variables dans chaque composante principale.

<i>Variables</i>	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>
log BCF	0,604	0,367	-0,708
Log Kow	0,604	0,369	0,707
TPSA	-0,520	0,854	-0,002

Les résultats du **tableau 13** montrent que la 1^{ère} composante principale PC1 présente une forte association positive avec log BCF et log Kow, on peut alors l'interpréter comme étant une mesure de la bioconcentration fortement liée à la lipophilie. La deuxième composante présente une forte association positive avec TPSA et mesure donc principalement la polarité des individus (les composés PBTs étudiés).

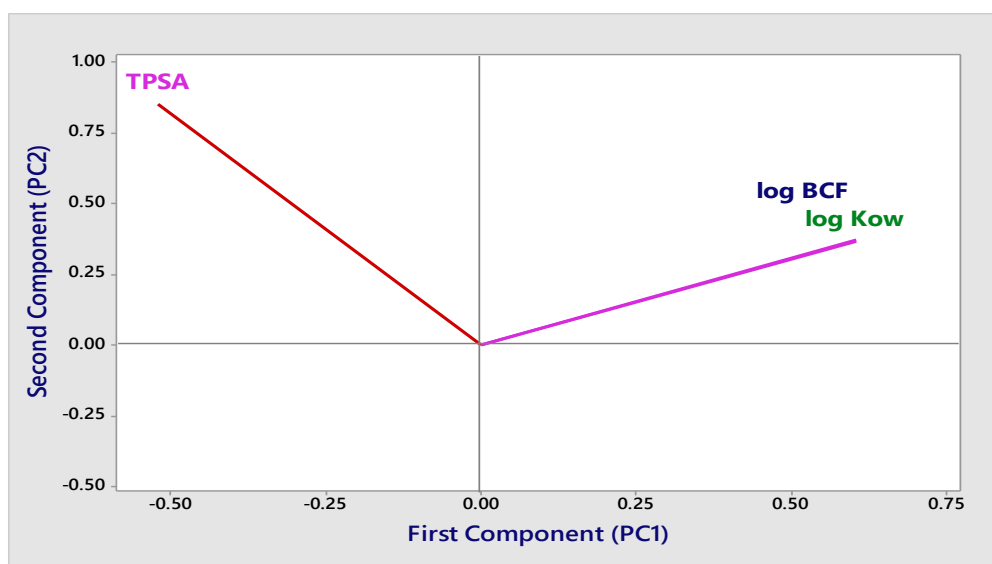


Figure 6 : Projection des variables dans le plan des deux premières composantes.

D'après la figure 6 qui représente le diagramme de contribution des variables dans les composantes, log BCF et log Kow présentent une forte contribution positive à la 1^{ère} et la 2^{ème} composante.

Par conséquent, $\log K_{ow}$ contribue positivement dans l'explication du $\log BCF$ (bioconcentration /accumulation) et ils sont fortement corrélés, ce qui indique qu'une augmentation de $\log K_{ow}$ (lipophilie) est accompagnée par une forte bioconcentration/accumulation.

En revanche, TPSA présente une forte contribution négative à la 1^{ère} composante (PC1) et une contribution positive à la 2^{ème} composante (PC2). Par conséquent, TPSA a une relation négative avec $\log K_{ow}$ et contribue négativement dans l'explication du $\log BCF$.

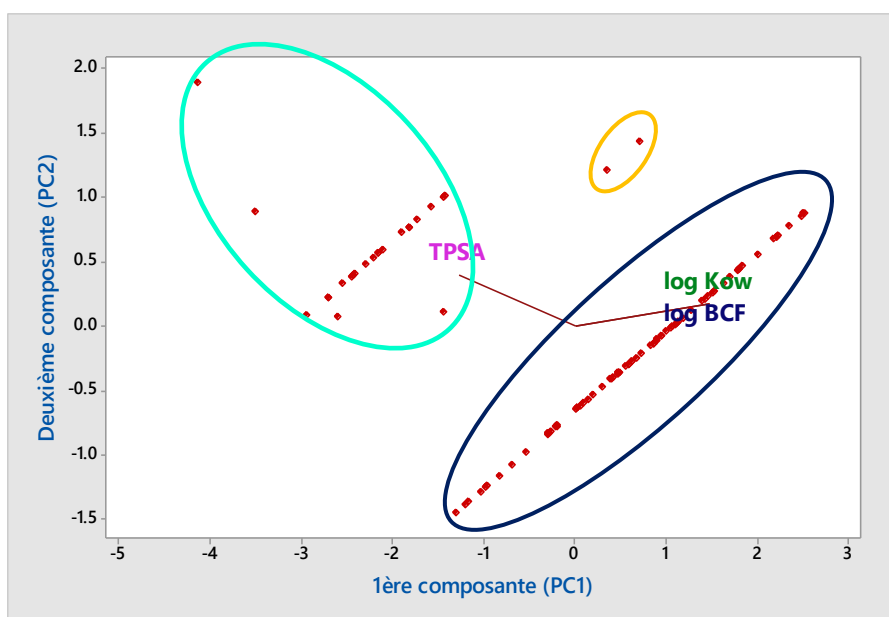


Figure 7 : Présentation des variables et des individus dans le plan des 2 composantes.

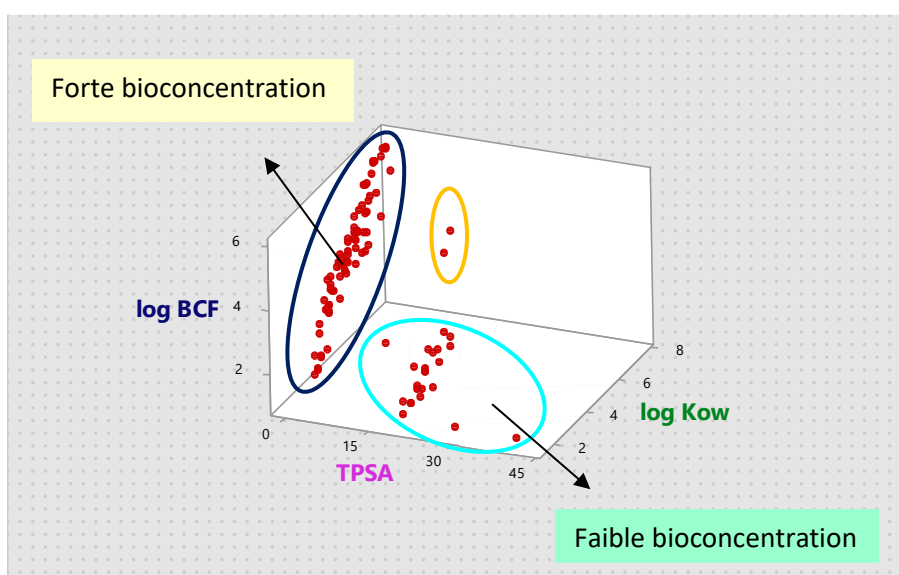


Figure 8 : Relation entre variables et individus en 3D.

Les figures 7 et 8 examinent la relation entre la variable à expliquer (log BCF), les deux variables explicatives (log K_{ow} et TPSA) et les individus (les composés étudiés).

Il s'avère que la série d'apprentissage est constituée principalement de deux catégories de composés PBTs (individus) :

- Les composés ayant une forte bioconcentration (accumulation) sont caractérisés par un grand log BCF, une forte lipophilie et une polarité nulle (TPSA=0).
- Les composés qui montrent une faible accumulation sont caractérisés par un petit log BCF, une faible lipophilie et une forte polarité.
- Deux composés ont montré une accumulation intermédiaire, il s'agit de 2,3,7,8-tetrachlorodibenzofurane et 2,3,4,7,8-pentachlorodibenzofurane.

III.1.6 Interprétation mécanistique du meilleur modèle QSAR

Modèle : $\log \text{BCF} = -0.042828 + 0.735002 \log K_{ow} - 0.014392 \text{TPSA}$ (3)

Les principaux descripteurs pouvant expliquer le facteur de bioconcentration sont le paramètre de lipophilie, présenté par log K_{ow} et la surface polaire topologique TPSA.

Le facteur de bioconcentration est proportionnel à la lipophilie et inversement proportionnel à la polarité, ce qui signifie que le composé le plus bioaccumulable est caractérisé par une forte lipophilie et une faible polarité.

D'une manière générale, le processus de bioconcentration est contrôlé par des interactions polaires et non polaires entre les produits chimiques, l'eau et les poissons. Si le produit est apolaire, il sera soluble dans la phase organique et non pas dans la phase aqueuse puisque l'eau est polaire, le soluté sera donc d'autant plus lipophile qu'il sera apolaire et donc peut être conservé à l'intérieur des couches lipidiques des membranes cellulaires, ce qui indique qu'il y a une forte interaction entre le produit chimique et les tissus du poisson plutôt qu'entre le produit chimique et l'eau.

Pour les composés hyper-hydrophobes, la fraction biodisponible de la concentration chimique dans l'eau, diminue avec un log K_{ow} croissant en raison de l'augmentation du coefficient de sorption des produits chimiques sur le carbone organique particulaire et dissous [4,12,14], ce qui fait diminuer la quantité de PBTs dans les espèces aquatiques et par conséquent une diminution de bioconcentration/ accumulation de ces composés dans les poissons.

Conclusion

Le facteur de bioconcentration de 145 composés a été étudié en respectant tous les critères d'OECD et toute la méthodologie QSAR.

Les descripteurs 2D utilisés pour l'étude du facteur de bioconcentration sont calculés avec Molinspiration (Online) et ChemSpider (Online) et les valeurs du logKow ont été prises de la littérature.

La PLS était la méthode statistique utilisée pour relier l'activité mesurée quantifiée par log BCF qui exprime l'accumulation et la structure présentée par les descripteurs moléculaires.

La combinaison de deux descripteurs (log Kow et TPSA) a permis d'obtenir un bon modèle QSAR expliquant d'une manière très significative le phénomène de bioconcentration/accumulation des composés PBTs étudiés dans le milieu aquatique (poissons).

Le modèle développé est simple, stable ayant non seulement une signification statistique mais aussi un pouvoir explicatif (R^2_{cv} LOO proche de R^2) et un fort pouvoir prédictif (R^2 test est élevé et tous les critères de Tropsha sont vérifiés).

Ce modèle peut être utilisé pour la prédiction du facteur de bioconcentration pour de nouveaux composés PBTs vis-à-vis les poissons.

III.2 Etude de la toxicité

Tableau 14 : Base de données et descripteurs calculés.

N°	Chemical	pLC ₅₀ (mg/L)	log Kow	E(Homo) ev	E(Lumo) ev	μ (ev)	η (ev)	S (ev)	Nu (ev)	w (ev)	TPSA (Å ²)
1	benzene-1,3-diol	3.04	0.76	-5.78	0.18	-2.80	5.96	0.08	3.34	0.66	40.46
2	1,4-dimethoxybenzene	3.07	2.1	-6.22	-0.11	-3.17	6.12	0.08	2.89	0.82	18.47
3	3-methoxyphenol	3.21	1.51	-5.66	0.28	-2.69	5.94	0.08	3.46	0.61	29.46
4	3-methylphenol	3.29	1.94	-5.88	0.04	-2.92	5.92	0.08	3.24	0.72	20.23
5	Toluene	3.32	2.73	-6.41	0.12	-3.14	6.53	0.08	2.71	0.76	0
6	Benzene	3.4	2.13	-6.72	0.07	-3.32	6.79	0.07	2.40	0.81	0
7	o-Xylene	3.48	3.12	-6.24	0.19	-3.03	6.44	0.08	2.87	0.71	0
8	4-methylphenol	3.58	1.94	-5.74	0.05	-2.85	5.79	0.09	3.38	0.70	20.23
9	1-hydroxy-2,6-dimethylbenzene	3.75	2.4	-5.72	0.28	-2.72	6.00	0.08	3.40	0.62	20.23
10	2-methylphenol	3.77	1.94	-5.83	0.14	-2.85	5.98	0.08	3.28	0.68	20.23
11	Chlorobenzene	3.77	2.84	-6.71	-0.37	-3.54	6.34	0.08	2.41	0.99	0
12	3-Chlorophenol	3.84	2.5	-6.29	-0.38	-3.34	5.91	0.08	2.83	0.94	20.23
13	1-hydroxy-2,4-dimethylbenzene	3.86	2.4	-5.63	0.20	-2.71	5.82	0.09	3.49	0.63	20.23
14	Bromobenzene	3.89	2.99	-6.59	-0.37	-3.48	6.22	0.08	2.53	0.97	0
15	1-hydroxy-3,4-dimethylbenzene	3.9	2.4	-5.65	0.18	-2.73	5.83	0.09	3.47	0.64	20.23
16	2-Chlorophenol	4.02	2.16	-6.25	-0.37	-3.31	5.88	0.08	2.86	0.93	20.23
17	p-Xylene	4.21	3.15	-6.14	0.16	-2.99	6.30	0.08	2.98	0.71	0
18	1-chloro-2-methyl-4-hydroxybenzene	4.27	2.89	-5.99	-0.25	-3.12	5.73	0.09	3.13	0.85	20.23
19	1,3-Dichlorobenzene	4.3	3.44	-6.92	-0.74	-3.83	6.18	0.08	2.19	1.19	0
20	2,4-dichlorophenol	4.3	2.99	-6.35	-0.77	-3.56	5.58	0.09	2.77	1.13	20.23
21	2,4,6-Trichlorophenol	4.33	3.06	-6.56	-1.06	-3.81	5.51	0.09	2.55	1.32	20.23
22	1-chloro-4-methylbenzene	4.33	3.27	-6.45	-0.31	-3.38	6.15	0.08	2.66	0.93	0
23	1,2-Dichlorobenzene	4.4	3.71	-6.85	-0.68	-3.76	6.17	0.08	2.27	1.15	0
24	1,3-dichloro-4-methylbenzene	4.54	3.88	-6.70	-0.65	-3.67	6.05	0.08	2.42	1.12	0
25	1,4-Dichlorobenzene	4.62	3.37	-6.75	-0.77	-3.76	5.98	0.08	2.37	1.18	0
26	1,2-dichloro-4-methylbenzene	4.74	3.74	-6.65	-0.64	-3.64	6.01	0.08	2.47	1.11	0
27	1,3,5-Trichlorobenzene	4.74	4.08	-7.25	-1.06	-4.16	6.18	0.08	1.87	1.40	0
28	1,2,3-Trichlorobenzene	4.89	4.27	-7.10	-0.96	-4.03	6.14	0.08	2.02	1.32	0
29	1,2,4-Trichlorobenzene	5	4.04	-6.92	-1.04	-3.98	5.89	0.08	2.19	1.35	0
30	1,2,3,4-Tetrachlorobenzene	5.43	4.65	-7.05	-1.24	-4.15	5.81	0.09	2.07	1.48	0
31	1,2,4,5-Tetrachlorobenzene	5.47	4.67	-7.02	-1.30	-4.16	5.71	0.09	2.10	1.52	0
32	2,3,4,5-Tetrachlorophenol	5.72	4.39	-6.70	-1.24	-3.97	5.46	0.09	2.41	1.44	20.23
33	Pentachlorophenol	6.06	4.78	-6.81	-1.44	-4.13	5.37	0.09	2.31	1.59	20.23

Les 33 composés de la base de données qui sont classés par ordre croissant de pLC₅₀ se divisent en deux sous séries, une série d'apprentissage constituée de 25 composés et les 8 molécules restantes forment la série de test.

Matrice de corrélation

	log Kow	E _{HOMO}	E _{LUMO}	μ	η	S	Nu	w
E(HOMO)	-0.788							
E(LUMO)	-0.850	0.837						
μ	-0.857	0.953	0.964					
η	-0.285	-0.081	0.479	0.226				
S	0.296	0.063	-0.492	-0.242	-0.997			
Nu	-0.788	1.000	0.837	0.953	-0.081	0.063		
w	0.864	-0.873	-0.995	-0.979	-0.411	0.428	-0.873	
TPSA	-0.636	0.731	0.371	0.563	-0.496	0.494	0.731	-0.420

D'après la matrice de corrélation, certains descripteurs sont fortement corrélés entre eux et d'autres ne le sont pas, ceci dit :

- Pour les descripteurs corrélés, on utilise PLS.
- Pour les descripteurs non corrélés, on utilise MLR.

Résultats et discussion

III.2.1 Elaboration des modèles

➤ **Etude de la toxicité pLC₅₀ avec un seul descripteur (log K_{ow})**

$$pLC_{50} = 1.995 + 0.7154 \log K_{ow} \quad (n=25) \quad (4)$$

Le coefficient de partage contribue à l'explication de la toxicité et le modèle obtenu donne de bons paramètres statistiques ($R^2=83.10\%$, $SD=0.326063$).

La toxicité augmente proportionnellement avec la lipophilie.

Dans le but d'obtenir un modèle plus explicatif et de lier le facteur de bioconcentration à la toxicité, on va essayer de combiner les descripteurs qui expliquent ce dernier avec un autre descripteur électronique.

➤ **Etude de la toxicité pLC₅₀ avec trois descripteurs**

Le meilleur modèle QSAR obtenu en utilisant la régression des moindres carrés partiels (PLS) est représenté dans le tableau ci-dessous.

Tableau 15: Le meilleur modèle QSAR avec 3 descripteurs pour l'étude de la toxicité.

Modèles (n=25)	R ²	R ² cv	F
pLC ₅₀ = - 4.4209 + 0.5082 log K _{ow} - 0.0116 TPSA + 86.1660 S	92.89 %	89.21 %	143.65

D'après les résultats du **tableau 15**, la combinaison du facteur de bioconcentration présenté par log K_{ow} et TPSA avec un descripteur électronique a donné un modèle avec des paramètres statistiques très satisfaisants, cela veut dire que la toxicité peut être liée à la bioconcentration et à la mollesse de la molécule.

Tableau 16: Contribution des composantes à l'explication de la toxicité.

pLC ₅₀ = f (log K _{ow} , TPSA, S)			
composantes	R ²	R ² cv	PRESS
1	91.47 %	85.99 %	2.02711
2	92.89 %	89.21 %	1.56054
3	93.21 %	88.99 %	1.59335

L'ajout de la troisième composante a fait augmenter le PRESS et diminuer le R² cv donc elle n'est pas nécessaire et on peut se limiter à la deuxième composante. Le meilleur modèle QSAR sera élaboré qu'avec les deux premières composantes.

III.2.2 Validation du meilleur modèle QSAR

III.2.2.1 Validation interne

Pour vérifier la stabilité et le pouvoir explicatif du modèle final, on procède comme suit :

➤ **Validation croisée (Cross Validation CV) :**

LOO (Leave One Out): $R^2_{cv}=89.21\%$ est proche de $R^2=92.89\%$, cela montre que le modèle n'est pas sensible à l'opération de mettre à part une molécule et de la remettre dans la série d'apprentissage. C'est une première indication sur la stabilité du modèle retenu.

➤ **Y-Randomisation :**

Tableau 17 : Randomisation pour le modèle QSAR retenu dans l'étude de la toxicité.

<i>Itération</i>	R^2_r	$R^2_{r cv}$	<i>Itération</i>	R^2_r	$R^2_{r cv}$
1	4.81 %	0	6	3.59 %	0
2	21.84 %	2.05 %	7	12.41 %	0
3	1.72 %	0	8	22.14 %	1.55 %
4	1.66 %	0	9	12.11 %	0
5	17.54 %	0	10	9.36 %	0

L'analyse des résultats du **tableau 17** montre que :

- Les modèles élaborés avec les valeurs randomisées de la toxicité pLC₅₀ ont donné des R^2_r inférieur à 25%.
- le meilleur modèle n'a pas été résulté d'une corrélation due au hasard.
- Le modèle élaboré existe réellement et il est stable.

III.2.2.2 Validation externe

Afin de vérifier le pouvoir prédictif du modèle QSAR obtenu, on passe par les deux étapes suivantes :

- Vérification des paramètres statistiques de la série de test.

Tableau 18 : Régression entre $pLC_{50_{exp}}$ et $pLC_{50_{cal}}$.

n	Equation de Régression (série de test)	R ²	R ² cv	SD
8	$pLC_{50_{cal}} = f(pLC_{50_{exp}})$	92.83 %	90.01 %	0.188949

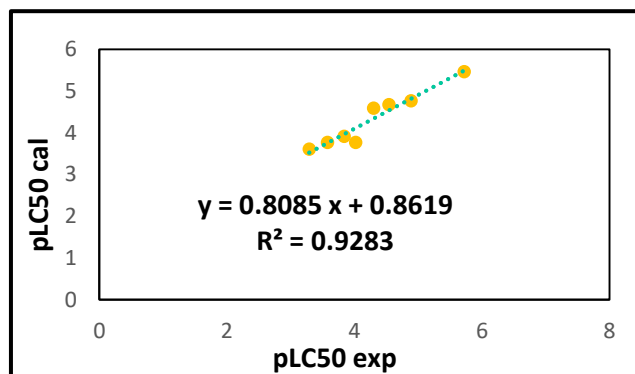


Figure 9 : Régression (série de test) entre $pLC_{50_{cal}}$ et $pLC_{50_{exp}}$

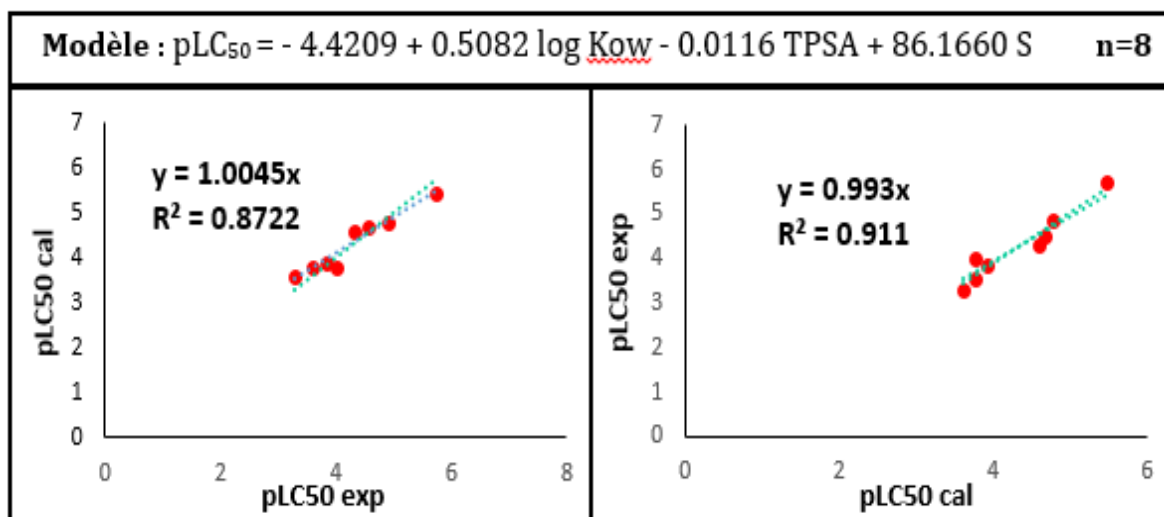


Figure 10 : a- Variation de $pLC_{50_{cal}}$ en fonction de $pLC_{50_{exp}}$ (avec intercepte =0).

b- Variation de $pLC_{50_{exp}}$ en fonction de $pLC_{50_{cal}}$ (avec intercepte =0).

➤ Vérification des critères de Tropsha.

Tableau 19 : Critères de Tropsha pour la série de test (Etude de la toxicité).

Modèle : $pLC_{50} = -4.4209 + 0.5082 \log Kow - 0.0116 TPSA + 86.1660 S$ n=8								
R ²	R ² cv	R ₀ ²	R ₀ ' ²	k	K'	$\frac{R^2 - R_0^2}{R^2}$	$\frac{R^2 - R_0'^2}{R^2}$	$ R^2 - R_0^2 $
92.90 %	87.14 %	87.22 %	91.10 %	1.005	0.993	0.0604	0.0186	0.0561

Les deux tableaux 18 et 19 regroupent les résultats de la validation externe et montrent que :

- ✓ Les paramètres statistiques (R^2 , R^2_{cv} et SD) de la série de test sont très satisfaisants.
- ✓ Tous les critères de Tropsha sont vérifiés.
- ✓ Les valeurs de k et K' sont assez proches de 1, ce qui indique que l'activité calculée, en utilisant l'équation de notre meilleur modèle est très proche à la valeur expérimentale.
- ✓ Le meilleur modèle QSAR possède un fort pouvoir prédictif.

III.2.3 Domaine d'applicabilité

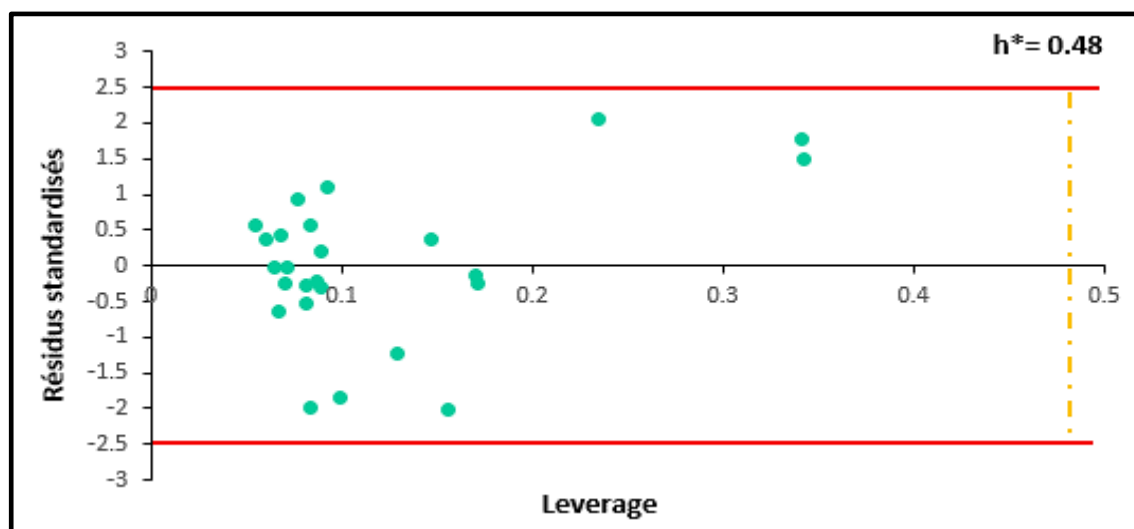


Figure 11 : Domaine d'applicabilité du meilleur modèle.

La figure 11 montre que toutes les observations ont des résidus standardisés compris entre $[-2,2]$ et aucun composé ne présente un *leverage* supérieur à la valeur seuil $h^*=0,48$ ($h^*=3(P+1)/n$) donc il n'y a aucune observation aberrante (outlier).

Les résultats de la validation externe et le domaine d'applicabilité montrent que le modèle QSAR élaboré dans ce travail peut être utilisé pour prédire la toxicité, quantifiée par pLC_{50} pour de nouvelles molécules polluantes organiques persistantes.

III.2.4 Interprétation mécanistique du meilleur modèle

Modèle : $pLC_{50} = - 4.4209 + 0.5082 \log Kow - 0.0116 TPSA + 86.1660 S$ (5)

Les principaux descripteurs pouvant expliquer la toxicité observée pour cette série de PBTs, sont le facteur de bioconcentration présenté par ($\log Kow + TPSA$) et la mollesse.

La toxicité est d'autant plus importante que les substances sont caractérisées par une forte lipophilie, une faible polarité et une grande mollesse.

Plus la molécule est molle plus elle se déforme facilement et plus elle a la capacité de faire un transfert de charge, c'est à dire des interactions entre la molécule en question et l'organisme.

La toxicité d'une substance dépend donc de sa capacité à s'accumuler dans l'organisme (bioconcentration) et sa capacité à interagir avec l'organe cible de l'espèce étudiée (mollesse).

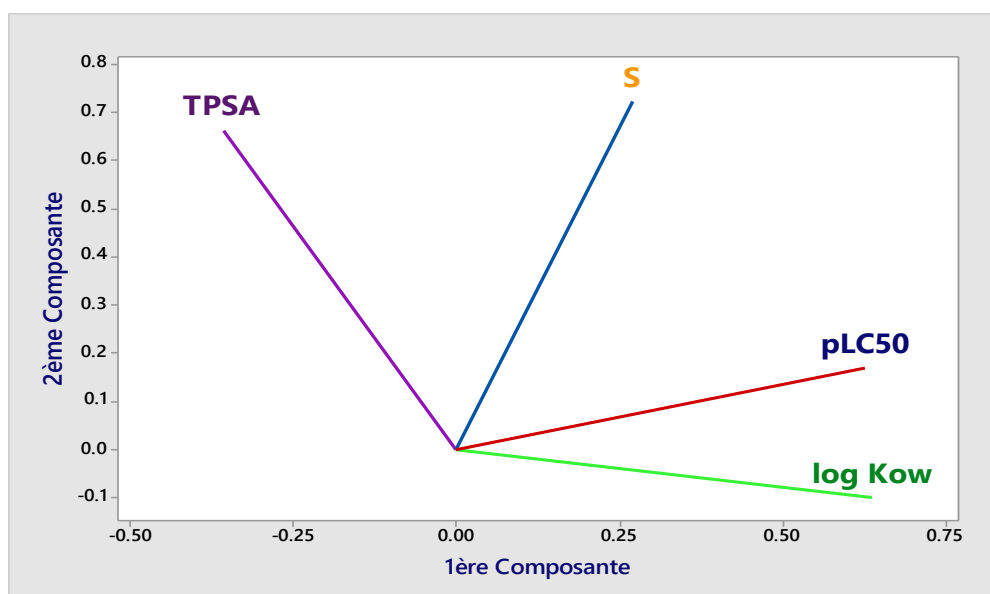


Figure 12 : Projection des variables dans le plan des deux premières composantes.

D'après la figure 12 de l'analyse en composantes principales, il est clair que la toxicité des PBTs étudiés exprimée par pLC_{50} est :

- Liée fortement et positivement au coefficient de partage $\log Kow$.
- Liée positivement à la mollesse (softness S).
- Liée négativement à la TPSA (le caractère polaire d'une substance PBT).

Conclusion

La toxicité d'une série de 33 composés a été étudiée en respectant tous les critères d'OECD et toute la méthodologie QSAR.

Les descripteurs électroniques ont été calculés après avoir optimisé les molécules avec la méthode B3LYP/6-31G(d,p).

La PLS était la méthode statistique utilisée pour élaborer le meilleur modèle. La combinaison de trois descripteurs ($\log K_{ow} + \text{TPSA} + S$) a permis d'obtenir un très bon modèle.

Le modèle développé est simple, stable ayant non seulement une signification statistique mais aussi un pouvoir explicatif (R^2_{cv} LOO proche de R^2 et $R^2_r \ll R^2$ (Y-randomisation)) et un fort pouvoir prédictif (R^2 (test) et SD (test) sont élevés et tous les critères de Tropsha sont vérifiés).

Ce modèle peut être utilisé pour prédire la toxicité pour de nouvelles molécules persistantes, bioaccumulables et toxiques PBTs vis-à-vis les poissons.

Références bibliographique

- [1] OECD, <http://www.oecd.org/chemicalsafety/risk-assessment/>
- [2] Gaussian 09, Revision A.1, M.J. Frisch et al., Gaussian, Inc, Wallingford CT, **2009**.
- [3] X. Lu, S. Tao, H. Hu and R.W. Dawson, Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors, *Chemosphere*. **41**, 1675-1688, **2000**.
- [4] A. Jon Arnot and A.P.C. Gobas. Frank, A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms, *Environmental Reviews*. **14**, 257-297, **2006**.
- [5] ECHA, <https://echa.europa.eu/fr/information-on-chemicals>
- [6] ECOTOX, <https://cfpub.epa.gov/ecotox/search.cfm>
- [7] L. Hall and E. Maynard, QSAR investigation of benzene toxicity to Fathead Minnow using molecular connectivity, *Environmental Toxicology and Chemistry*. **8**, 783-788, **1989**.
- [8] B.D. Gute and S.C. Basak, predicting acute toxicity (LC50) of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach, SAR and QSAR in *Environmental Research*. **7**, 117-131, **1997**.
- [9] MINITAB, State College, PA Minitab, Inc, **2006**.
- [10] Molinspiration, <https://molinspiration.com/cgi-bin/properties>
- [11] Chemspider, <https://www.chemspider.com/>
- [12] D. Mackay, J.A. Arnot, E.P. Petkova, K.B. Wallace, D.J. Call, L.T. Brooke and G.D. Veith, The physicochemical basis of QSARs for baseline toxicity, SAR and QSAR in *Environmental Research*. **20**, 393-414, **2009**.
- [13] A. Gissi, A. Lombardo, A. Roncaglioni, D. Gadaleta, G. Mangiatordi, O. Nicolotti and E. Benfenati, Evaluation and comparison of benchmark QSAR models to predict a relevant REACH endpoint: The bioconcentration factor, *Environmental Research*. **137**, 398-409, **2015**.
- [14] X.H. Wang, Y. Yu, T. Huang, W.C. Qin, L.M. Su and Y.H. Zhao, Comparison of Toxicities to *Vibrio fischeri* and Fish Based on Discrimination of Excess Toxicity from Baseline Level, *PLOS*. **11**, 1-17, **2016**.
- [15] M. Nendza, R. Kühne, A. Lombardo, S. Stempel and G. Schüürmann, PBT assessment under REACH: Screening for low aquatic bioaccumulation with QSAR classifications based on physicochemical properties to replace BCF in vivo testing on fish, *Science of the Total Environment*. **616**, 97-106, **2018**.
- [16] X. Bodiguel, Caractérisation et modélisation des processus de bioaccumulation des PCB chez le merlu (*Merluccius merluccius*) du golfe du Lion, Thèse Doctorat, l'Université Montpellier I, **2008**.
- [17] S. Boudergua, M. Alloui, S. Belaidi, M. Mogren Al Mogren, U.A. Ibrahim and M. Hochlaf, QSAR Modeling and Drug-Likeness Screening for Antioxidant Activity of Benzofuran Derivatives, *Journal of Molecular Structure*. **1189**, 307-314, **2019**.

Conclusion Générale

Conclusion générale

L'utilisation des approches de modélisation *in silico* rapides et rentables sont désormais essentielle pour identifier les substances dangereuses et pour aider à répondre à la demande d'évaluations d'un nombre potentiellement important de composés polluants organiques persistants, en particulier pour éviter le recours aux essais sur les animaux car l'évaluation complète par voie expérimentale est impossible.

L'objectif de ce présent travail était de développer des modèles QSAR fiables pour la prédiction du facteur de bioconcentration et de la toxicité pour certaines familles de polluants organiques persistants (POPs) vis-à-vis les poissons.

Les logiciels utilisés pour le calcul des descripteurs étaient ChemSpider (online), Molinspiration (online) et Gaussian09 et pour l'élaboration des modèles : Minitab17.

Dans la première application, nous avons utilisé une série de 145 composés et avons établi des modèles reliant certains descripteurs moléculaires (le coefficient de partage log Kow, polarisabilité α , volume molaire V_m , le poids moléculaire M_w et la surface polaire topologique TPSA) au facteur de bioconcentration, quantifié par log BCF. Les techniques d'analyse de données utilisées pour la construction des modèles étaient la SLR (Régression linéaire Simple) pour le modèle à un seul descripteur ($\log \text{BCF} = f(\log \text{kow})$) et la PLS (Partial Least Square) pour les modèles à 2 et 3 variables corrélées. Les résultats obtenus montrent que le meilleur modèle était celui qui combine deux descripteurs : le coefficient de partage et la surface polaire topologique. Ce modèle a montré après la validation interne un bon pouvoir explicatif et bonne stabilité (validation croisée ($R^2_{cv} = 81.42\%$ assez proche de $R^2 = 82.69\%$) et Y-randomisation ($R^2_r \ll R^2$)), ainsi qu'un bon pouvoir prédictif après validation externe (R^2 et SD de la série de test sont élevés et tous les critères de Tropsha sont vérifiés). L'analyse du domaine d'applicabilité montre la présence d'une seule observation aberrante ayant un *leverage* supérieur à la valeur seuil $h^* = 3(P+1)/n$, le composé en question est donc considéré comme outlier, hors du domaine d'applicabilité du modèle élaboré et la prédiction de sa valeur du facteur de bioconcentration est donc impossible. Le modèle obtenu montre que la bioconcentration est d'autant plus importante que les molécules sont faiblement polaires et fortement lipophiles.

Conclusion générale

Dans la deuxième application, une série de 33 composés a été utilisée et nous avons relié les descripteurs expliquant le facteur de bioconcentration obtenu dans le meilleur modèle de la première application et aussi un descripteur électronique calculé par la méthode B3LYP et la base 6-31G(d,p) à la toxicité afin de voir comment l'accumulation des PBTs et leur interactions avec l'organisme font varier la toxicité, quantifiée par la pLC_{50} .

La méthode statistique utilisée pour la construction du meilleur modèles était la PLS (Partial Least Square) et les résultats obtenus montrent que le meilleur modèle était celui qui combine trois descripteurs : le coefficient de partage log Kow, la surface polaire topologique TPSA et la mollesse S.

Ce modèle a montré après la validation interne un bon pouvoir explicatif et bonne stabilité (validation croisée ($R^2_{cv} = 89.21\%$ proche de $R^2 = 92.89\%$) et Y-randomisation ($R^2_r \ll R^2$)), ainsi qu'un bon pouvoir prédictif après validation externe (R^2 et SD de la série de test sont élevés et tous les critères de Tropsha sont vérifiés).

L'analyse du domaine d'applicabilité a montré l'absence de toute observation aberrante.

Les deux modèles retenus dans ce travail sont à la fois simples, fiables, prédictifs, et interprétables et sont aussi en accord avec les règles mises en place par l'OCDE quant à la validation des modèles QSAR. Ils peuvent cependant être utilisés pour concevoir et prédire la bioconcentration et la toxicité pour de nouveaux composés PBTs.

ملخص

الهدف من هذا العمل الحالي هو تطوير نماذج QSAR موثوقة ومستقرة وقابلة للتنبؤ من أجل تحديد نشاطين لعائلات معينة من الملوثات العضوية الثابتة تجاه الأسماك.

1-التنبؤ بعامل التركيز الحيوي لسلسلة مكونة من 145 مركب من الملوثات العضوية الثابتة.

2-التنبؤ بسمية لسلسلة مكونة من 33 مركب من الملوثات العضوية الثابتة.

المنتجات الملوثات العضوية الثابتة هي ملوثات منتشرة في البيئة بسبب استخدامها على نطاق واسع في الصناعة والزراعة. تم تطوير أفضل نماذج QSAR التي تم الحصول عليها باستخدام طريقة (PLS)، وقد تم التحقق من ثباتها وقوتها التفسيرية من خلال طريقة (LOOCv) و-Y. Randomisation تم التحقق من قوتهم التنبؤية من خلال التحقق الخارجي من سلسلة الاختبارات غير المدرجة في مجموعة التدريب وعن طريق التحقق من جميع معايير Tropsha.

يتم التحقق من مجال تطبيق النماذج المختارة من خلال طريقة leverage وتظهر النتائج التي تم الحصول عليها أن نماذج QSAR التي تم تطويرها موثوقة ومستقرة ولديها قدرة تنبؤية جيدة.

Résumé

L'objectifs de ce présent travail est d'élaborer des modèles QSAR fiables, stables et prédictifs pour la prédiction de deux activités de certaines familles de polluants persistants, bioaccumulables et toxique (PBTs) vis-à-vis les poissons.

1- La prédiction du facteur de bioconcentration d'une série de 145 composés PBTs.

2- La prédiction de la toxicité d'une série de 33 composés PBTs.

Ces produits Polluants Organiques Persistants sont des contaminants omniprésents dans l'environnement en raison de leur large utilisation dans l'industrie et l'agriculture.

Les meilleurs modèles QSAR obtenus sont élaborés avec la méthode PLS (Partial least square), leur stabilité et leur pouvoir explicatif ont été vérifiés par la validation croisée Leave-One-Out (LOOCv) et l'Y-Randomisation. Leur pouvoir prédictif a été vérifié par la validation externe sur la série de test non incluse dans le jeu d'entraînement et en vérifiant tous les critères de Tropsha.

Le domaine d'applicabilité des modèles retenus est vérifié par la méthode de 'leverage' et les résultats obtenus montrent que les modèles QSAR élaborés sont fiables, stables et ont une bonne capacité prédictive.

Abstract

The objective of this present work is to develop reliable, stable and predictive QSAR models for the prediction of two activities of some families of persistent, bioaccumulative and toxic pollutants (PBTs) with respect to fish.

1- The prediction of the bioconcentration factor of a series of 145 PBTs compounds.

2- The prediction of the toxicity of a series of 33 PBTs compounds.

These Persistent Organic Pollutants products are ubiquitous contaminants in the environment due to their wide use in industry and agriculture.

The best QSAR models obtained are developed with the PLS (Partial Least Square) method, their stability and explanatory power have been verified by Leave One-Out (LOOCv) cross-validation and Y Randomization. Their predictive power was verified by external validation on the test series not included in the training set and by verifying all of the Tropsha criteria.

The field of applicability of the selected models is verified by the "leverage" method and the results obtained show that the QSAR models developed are reliable, stable and have good predictive capacity.