REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABOU BEKR BELKAÏD DE TLEMCEN

FACULTE DES SCIENCES
DEPARTEMENT DE CHIMIE

Laboratoire de Thermodynamique Appliquée et Modélisation Moléculaire (LATA2M)

T H E S E

Pour l'obtention du diplôme de

Doctorat en Chimie Physique

Option : Chimie Théorique et Modélisation Moléculaire

Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires. « Modélisation QSAR »

Présentée par :

Mme. ERRAHOUI née BELLIFA KHADIDJA

Soutenu publiquement le : 08 /10 /2015 devant le jury composé de

Melle. Amina NEGADI	**Professeur**	**U. A. Belkaïd -Tlemcen**	**Présidente**
Mr. Meziane BRAHIMI	**Professeur**	**U.S.T.H.B - Alger**	**Examinateur**
Mme. Yamina AIT-MEBAREK	**MC.A**	**U. M'hamed Bouguera- Boumerdes**	**Examinateur**
Mr. Bachir MOSTEFA-KARA	**Professeur**	**U. A. Belkaïd - Tlemcen**	**Examinateur**
Mr. Sidi Mohamed MEKELLECHE	**Professeur**	**U. A. Belkaïd - Tlemcen**	**Directeur de thèse**

A mes très chers parents

A mon mari et à mes enfants

A mes frères et à ma sœur

A mes amies

A tous ceux qui me sont chers

° ° ° *Remerciements* ° ° °

Ce travail a été effectué au sein de laboratoire de Thermodynamique appliquée et modélisation moléculaire, ***(LATA2M), de l'Université A. Belkaïd de Tlemcen****, dirigé par Mademoiselle la professeur* ***NEGADI LATIFA.***

Cette thèse a été dirigée par Monsieur ***MEKELLECHE SIDI MOHAMED****, Professeur à la faculté des sciences,* ***Université A. Belkaïd de Tlemcen****, à qui je tiens à exprimer toute ma gratitude pour son encadrement, sa disponibilité, ses conseils et pour m'avoir fait bénéficié de l'étendue de ses connaissances. Je lui remercie très chaleureusement pour son engagement, sa persévérance et les encouragements qu'il a su me prodiguer jusqu'au dernier jour. Merci de m'avoir appris à structurer mes idées, à mieux valoriser mon travail même si j'ai encore largement du travail dans ce domaine.*

J'exprime ma profonde et respectueuse gratitude à Mademoiselle ***NEGADI AMINA****, Professeur à* ***l'Université A. Belkaïd de Tlemcen****, qui nous a fait l'honneur d'accepter de présider le jury de cette thèse.*

Je suis très honorée par la présence de Monsieur ***MEZIANE BRAHIMI****, Professeur à* ***l'Université USTHB d'Alger****, à qui j'adresse mes remerciements les plus sincères et l'expression de mon profond respect pour avoir accepté d'examiner ce travail.*

Je tiens à adresser mes vifs remerciements et l'expression de mon profond respect à Madame ***AIT-MEBAREK YAMINA****, professeur à* ***l'Université M'hamed Bouguera de Boumerdes****, pour l'intérêt qu'elle a porté à ce travail en acceptant d'en être l'un des rapporteur de cette thèse.*

Mes vifs et sincères remerciements vont également à Monsieur ***MOSTEFA-KARA BACHIR,*** *Professeur à* ***l'Université A. Belkaïd de Tlemcen****, pour l'honneur qu'il nous a fait d'accepter de juger notre travail.*

Enfin, j'adresse mes remerciements à tous les collègues de l'équipe «chimie théorique et modélisation moléculaire» du laboratoire de recherche «Thermodynamique Appliquée et Modélisation Moléculaire» et à tous ceux qui ont contribué de prés ou de loin à la réalisation de ce travail.

LISTE DES ABREVIATIONS

ADN Acide Désoxyribo Nucléique

AM1 Austin Model 1

B3LYP Becke3-ParameterLee-Yang-Parr

CGTO Contracted Gaussian Type Orbital Configuration

CLOA Combinaison Linéaire d'Orbitales Atomiques

CT Charge transfer

DA Domaine d'applicabilité

DFT Density Functional Theory

DM Dipole Moment

GA Genetic Algorithm

GGA Generalized Gradient Approximation

GTO GaussianTypeOrbital

HF Hartree-Fock

HOMO Highest Occupied Molecular Orbital

IGC_{50}^{-1} Inhibition of 50% Growth Concentration

KS Kohn et Sham

LDA Local Density Approximation

LD Lethal Dose

LMO Leave Many Out

LOO Leave One Out

LUMO Lowest Unoccupied Molecular Orbital

MLR Multiple Linear Regression

MOA Mode of Action

MSD Mulliken Spin density

NA Nucleic Acid

NN Neurone Network

OA Orbitale Atomique

OM Orbitale Moléculaire

PCA Principal Composante Analysis

PLS Partial Least Squares

PM3 Parametric Method 3

PM6 Parametric Method 6

QSPR Quantitative Structure-Property Relationships

QSAR Quantitative Structure Activity Relationship

SCF : Self Consistant Field

STO Slater Type Orbital

SLR SimpleLinear Regression

SOMO Singly Occupied Molecular Orbital

SSR Sum of Squares Regression

SSE Sum of Squares Error

SVM Support Vector Machine

SOMMAIRE

CHAPITRE III : METHODES DE LA CHIMIE QUANTIQUE

CHAPITRE IV : GENERALITES SUR LA TOXICOLOGIE - MODELISATION QSTR

APPLICATION III

Estimation de la toxicité aiguë des phénols halogénés en utilisant les paramètres d'hydrophobie et d'électrophilie

INTRODUCTION GENERALE

Les produits chimiques font partie de notre quotidien, qu'ils soient d'origine naturelle ou synthétique. Ils représentent une partie intégrante de notre environnement, dans nos maisons, nos vêtements ou encore notre alimentation. L'utilisation des produits chimiques est aujourd'hui un facteur essentiel du développement de notre société et contribue à la prospérité économique que connait le monde. Depuis les années 1930, la production mondiale de substances chimiques a été multipliée par 400. Le plastique, les conservateurs, les détergents, les peintures, etc., nous rendent d'innombrables services. Cela dit, il n'en reste pas moins qu'ils peuvent également présenter des risques pour l'homme, les autres êtres vivants et l'environnement dans sa globalité. Certaines substances peuvent avoir des effets nocifs importants sur l'environnement et la santé, même à faible dose. D'autres suscitent des inquiétudes par leur caractère persistant dans les milieux, d'autres encore sont difficiles à mesurer et les effets à long terme d'un grand nombre sont méconnus. Face à l'omniprésence des substances chimiques dans notre quotidien et devant l'importance des incertitudes qui demeurent pour conduire l'évaluation complète de leurs effets, l'enjeu est de déterminer l'équilibre acceptable par nos sociétés entre les bénéfices apportés et la prise de risque pour la santé humaine et l'environnement. Sur les 100 000 substances chimiques recensées, moins de 3000 (moins de 3%) ont fait l'objet d'analyses approfondies quant à leurs propriétés dangereuses et l'évaluation quantifiée des risques toxiques et écotoxiques **[1]**. Néanmoins ces dernières années, les connaissances sur les effets sur la santé se sont affinées grâce à plusieurs études axées sur certaines grandes familles de substances (métaux lourds, dioxines, éthers de glycol, hydrocarbures aromatiques, etc.). L'effet sur la santé peut être produit par une dose faible ou élevée, par une exposition unique ou répétée ; il peut être immédiat ou différé, réversible ou irréversible. Il dépend aussi de la cible biologique.

Parmi les substances toxiques, celles aux propriétés cancérogènes, mutagènes ou toxiques pour la reproduction sont les plus préoccupantes. Il en existe dans de nombreuses familles : composés organiques aromatiques halogénés, phénols, métaux, phtalates, huiles minérales, hydrocarbures, substances complexes dérivées du pétrole.

En raison de leur utilisation universelle, les hydrocarbures aromatiques tels que les dibenzofuranes, les dérivés du nitrobenzène et du phénol sont des polluants ubiquitaires dans presque tous les écosystèmes aquatiques et terrestres. Parce que la plupart d'entre eux produisent des effets néfastes sur les espèces vivantes, la contamination de toute notre planète par ces composés représente donc une menace, tant pour l'équilibre des écosystèmes que pour la santé humaine. Ces composés (xénobiotiques) ont inspiré un grand nombre d'études expérimentales [2-5] pour évaluer leur toxicité. Cependant, ces études sont très couteuses et consomment beaucoup de temps parfois des années et elles sont impossibles pour certains cas, pour cette raison le recours aux méthodes alternatives moins couteuses et rapides est indispensable.

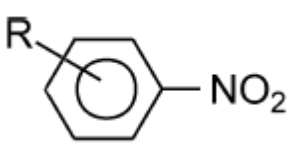
L'utilisation de méthodes alternatives à l'expérimentation, parmi lesquelles les relations quantitatives structure propriété/activité (QSPR/QSAR) sont devenues d'un grand intérêt et sont même recommandées dans les nouvelles règlementations [6, 7] afin d'obtenir les données nécessaires à l'enregistrement des substances.

L'élaboration des modèles mathématiques QSPR/QSAR reliant les propriétés physico-chimiques et les activités biologiques à la structure moléculaire permet, d'une part, d'expliquer l'origine de ces activités/propriétés et, d'autre part, de les prédire pour des molécules pour lesquelles les données expérimentales ne sont pas disponibles.

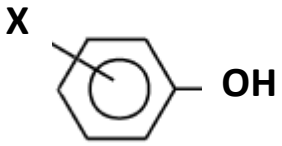
Ces méthodes sont à l'heure actuelle principalement utilisées pour le criblage des composés toxiques par exemple. La procédure de criblage consiste à utiliser ces méthodes pour prédire a priori les performances (ou la dangerosité) des substances pour des applications données. Ainsi, il est possible d'identifier les molécules présentants le plus fort potentiel pour la propriété/ l'activité attendue. Les premières applications de ces méthodes, basées sur des analyses statistiques, ont concerné principalement des applications biologiques [8-10], toxicologiques [11, 12] ou pharmaceutiques [13-15].

Il s'agit, dans cette thèse, de développer et d'évaluer le potentiel de tels modèles QSAR pour l'explication et la prédiction de la toxicité de certaines familles de molécules aromatiques toxiques, en l'occurrence, les nitrobenzènes, les nitro-aromatiques et les phénols

halogénés en utilisant une méthodologie de type QSAR reliant les activités expérimentales aux structures moléculaires, exprimées par des descripteurs quantiques et non quantiques.



Dérivé du nitrobenzène



Phénol halogéné

Dans la première partie nous avons élaboré des modèles QSAR pour la toxicité d'une série constituée de 50 nitrobenzènes substitués vis-à-vis d'une espèce aquatique qui est le *Tetrahymena Pyriformis,* cette base de donnée a été prise des travaux de Schultz et ses collaborateurs [**16, 17**] et qui sont regroupés dans la base de données TETRATOX [**18**]**.** Les valeurs de la toxicité sont considérées très fiables parce que le même protocole a été utilisé pour faire les mesures de cette toxicité.

Les descripteurs choisis et utilisés reflètent le phénomène de transport de la molécule toxique de la phase aqueuse aux lipides membranaires et le caractère électronique et électrophile de ces molécules.

Dans la deuxième application nous avons essayé d'élaborer des modèles en se basant sur le mode d'action (MOA) des nitro-aromatiques. La base de données a été divisée en deux séries, la première série constituée de 32 molécules qui reflètent un mécanisme d'attaque nucléophile en utilisant des descripteurs convenables pour ce type de mécanisme, tandis que la deuxième série constituée de 43 molécules qui reflètent un mécanisme succession Redox « *Redox-Cycling*» en utilisant d'autres descripteurs qui peuvent nous informer sur ce mécanisme.

Dans la troisième application, nous avons élaboré des modèles QSAR pour étudier la toxicité d'une série de phénols halogénés constituée de 45 molécules en respectant tous les critères concernant un modèle QSAR fiable, robuste et prédictif.

La stratégie générale adoptée pour élaborer les modèles QSAR dans les trois applications est celle qui utilise des descripteurs ciblés en se basant sur la nature des molécules de nos bases de données et les différents mécanismes possibles pour expliquer la toxicité. Cette stratégie est plus rationnelle parce que les modèles élaborés contiennent un nombre réduit de descripteurs, ces derniers ont une signification chimique et sont appropriés pour expliquer la toxicité de ces familles de molécules toxiques. En revanche, la stratégie adoptée par les nouveaux programmes de modélisation QSPR/QSAR comme CODESSA[**19**], DRAGON [**20**] et CORAL [**21**] ou les modèles élaborés peuvent contenir un grand nombre de descripteurs, ces modèles ont une bonne interprétation statistique mais il est difficile et parfois impossible de les interpréter d'un point de vue chimique et biologique.

Le manuscrit de cette thèse est articulé sur cinq chapitres :

- Dans le premier chapitre nous présentons les méthodes QSPR/QSAR, leur principe, méthodologie, les avantages et les applications seront ainsi détaillées.

- Dans le deuxième chapitre, nous exposons les différentes bases théoriques des outils d'analyse de données statistiques nécessaires à la mise en œuvre des modèles QSAR.

- Le troisième chapitre est consacré à la description des méthodes de chimie quantique utilisées pour le calcul des structures (descripteurs) moléculaires ainsi que pour l'étude des mécanismes de la toxicité.

- Le quatrième chapitre sera dédié à la présentation de notions de toxicologie, une synthèse bibliographique sur les études QSAR en toxicologie.

- Dans le cinquième chapitre nous présentons et nous discutons les résultats obtenus pour les applications effectuées :

Application 1 **:** Etude QSAR de la toxicité des nitrobenzènes vis-à-vis *Tetrahymena-pyriformis* à l'aide des descripteurs quantiques

Application 2 : Estimation de la toxicité des nitroaromatiques : Modèles QSAR pour les mécanismes « *Redox-cycling* » et « Nucleophilic Attack»

Application 3 : Estimation de la toxicité aiguë des phénols halogénés en utilisant les paramètres d'hydrophobie et d'électrophilie (Article soumis)

Nous terminerons par une conclusion générale et les perspectives envisagées pour ce travail.

Références bibliographiques

[1] J.C. Dearden, Prediction of Environmental Toxicity and Fate Using Quantitative Structure-Activity Relationships (QSARs), J. Braz. Chem. Soc **2002,**13, 754-762.

[2] M.W.Toussaint, T.R. Shedd, W.H.Vanderschalie, G.R.Leather, Environ. Toxicol.Chem. **1995**, 14, 907.

[3] J.C.Dearden, In Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology; Karcher, W.; Devillers, J.eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, **1990**, p. 25.

[4] C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, American Chemical Society:Washington DC, USA, **1995**.

[5] J.Dearden, ECVAM Workshop on The Use of Computer Models as Alternatives to Animal Experiments in Chemical Risk Assessment, October 3-4, Praha, Czech **2002.**

[6] Règlement (CE) n° 1907/2006 du Parlement Européen et du Conseil du 18 décembre **2006** concernant l'enregistrement, l'évaluation et l'autorisation des substances chimiques, ainsi que les restrictions applicables à ces substances (REACH), instituant une agence européenne des produits chimiques, modifiant la directive 1999/45/CE et abrogeant le règlement (CEE) n° 793/93 du Conseil et le règlement (CE) n° 1488/94 de la Commission ainsi que la directive 76/769/CEE du Conseil et les directives 91/155/CEE, 93/67/CEE, 93/105/CE et 2000/21/CE de la Commission.

[7] N. Margossian, *Le règlement REACH - La règlementation européenne sur les produits chimiques*, Dunod / L'Usine Nouvelle, Paris, **2008**.

[8] V.K. Agrawal, P.V. Khadikar, *Bioorg. Med. Chem.*, **2001**, *9*, 3035-3040.

[9] D.A. Winkler, *Brief. Bioinf.* **2002**, *3*, 73-86.

[10] H. Gao, J.A. Katzenellenbogen, R. Garg, C. Hansch, *Chem. Rev.*, **1999**, *99*, 723-744.

[11] C.D. Selassie, R. Garg, S. Kapur, A. Kurup, R.P. Verma, S.B. Mekapati, C. Hansch, *Chem. Rev.* **2002**, *102*, 2585-2606.

[12] S.P. Bradbury, *Toxicol. Lett.*, **1995**, *79*, 229-237.

[13] R. Garg, S.P. Gupta, H. Gao, M.S. Babu, A.K. Debnath, C. Hansch, *Chem. Rev.*, **1999**, *99*, 3525-3602.

[14] M. Grover, B. Singh, M. Bakshi, S. Singh, *Pharm. Sci. Tech. Today*, **2000**, *3*, 50-57.

[15] M. Grover, B. Singh, M. Bakshi, S. Singh, *Pharm. Sci. Tech. Today*, 2000, *3*, 28-35.

[16] M.T.D. Cronin, B.W. Gregory and T.W. Schultz, *Chem. Res. Toxicol.* **1998**, *11*, 902-908, Quantitative structure-activity analysis of nitrobenzene toxicity to *Tetrahymena Pyriformis*..

[17] M.T.D.Cronin, N.Manga, J.R.Seward, G.D.Sinks and T.W. Schultz, *Chem. Res. Toxicol*, Parametrization of electrophilicity for the prediction of the toxicity of

aromatic compounds. **2001**, *14*, 1498-1505

[18] TETRATOX http://www.vet.utk.edu/TETRATOX/index.php

[19] CODESSA PRO, University of Florida, www.codessa-pro.com

[20] R .Todeschini, V. Consonni, A. Mauri, M. Pavan, 2005. *Logiciel DRAGON*, *Version 5.3*.Milano.

[21] CORAL www.insilico.eu/coral

CHAPITRE I

MODELISATION QSPR/QSAR

Introduction

La connaissance des propriétés et des activités est d'une importance capitale pour pouvoir classer et utiliser les composés chimiques. La caractérisation expérimentale complète est difficile, voire impossible, pour des raisons de temps, de coût, de dangerosité de certains essais ou d'éthique (limitations des essais sur les animaux). L'utilisation des méthodes alternatives à l'expérience est devenue plus qu'indispensable. Parmi ces méthodes, on trouve les méthodes de modélisation moléculaire qui permettent de justifier les données expérimentales disponibles et prédire les propriétés/activités pour des composés nouveaux ou des composés pour lesquels les données expérimentales ne sont pas disponibles. Parmi ces méthodes de modélisation les plus utilisées, on peut citer les méthodes QSPR (*Quantitative Structure-Property Relationships*) et QSAR (*Quantitative Structure-Activity Relationships*). Ces méthodes s'appuient sur le principe que les propriétés physico-chimiques et les activités biologiques des molécules dépendent fortement de leurs structures chimiques.

I.1. Généralités sur la modélisation QSPR/QSAR

Une relation QSPR/QSAR est un modèle ou une formule mathématique qui permet de relier, d'une manière quantitative, la structure d'une molécule à une propriété ou à une activité donnée. Les méthodes QSPR/QSAR sont de plus en plus utilisées, du fait de la croissance des moyens de calculs (machines, logiciels,...). Récemment, on assiste à la mise en place d'un nouveau règlement REACH [1] (*Registration, Evaluation, Authorisation and Restriction of Chemicals)* qui recommande l'utilisation des méthodes alternatives pour limiter le recours à l'expérimentation.

En fait, les premiers travaux QSPR/QSAR remontent au 19ème siècle. En effet, dès 1868, Crum-Brown et Fraser [2] ont postulé l'existence de relations entre les activités physiologiques et les structures chimiques en reliant les variations de l'activité biologique à des modifications structurales. Cependant, à cette époque, les structures moléculaires n'étaient pas encore connues.

Une avancée importante vers les modèles QSPR/QSAR proprement dits a été réalisée grâce au développement des équations de Hammett [3] dans lesquelles les constantes σ caractérisent de manière quantitative les vitesses de réactions pour les composés organiques.

$$\log\frac{K}{K_0} = \sigma\rho \qquad (1)$$

Ou K et K_0 sont les constantes respectives de la réaction étudiée et de celle d'une référence et ρ une constante de réaction dépendant du type de réaction [3,4].

Les premiers travaux utilisant la méthodologie QSPR/QSAR telle qu'employée actuellement sont dus à Hansch [5] et Free et Wilson [6]. D'un côté, Hansch a proposé des modèles reliant directement l'activité biologique des composés avec les propriétés hydrophobes, électroniques et stériques à l'échelle moléculaire. D'un autre côté, Free et Wilson ont développé des modèles empiriques, dits de contributions de groupes, pour l'étude de l'activité biologique.

Au cours de ces dernières décennies, l'utilisation des méthodes QSPR/QSAR n'a pas cessé de progresser. Elle est même devenue indispensable en chimie pharmaceutique, en toxicologie et pour la conception de médicaments [7-9].

I.2. Principe des méthodes QSPR/QSAR

Le principe des méthodes QSPR/QSAR est d'établir une relation mathématique reliant de manière quantitative des propriétés moléculaires, appelées descripteurs, avec une observable macroscopique (activité biologique, toxicité, propriété physico-chimique, etc.), pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données. La forme générale d'un tel modèle est la suivante :

$$Propriété/\ Activité = f\,(D_1, D_2, \ldots D_n, \ldots) \qquad (2)$$

$D_1, D_2, \ldots D_n$ sont des descripteurs des structures moléculaires.

L'objectif d'une telle méthode est d'analyser les données structurales afin de détecter les facteurs déterminants pour la propriété /activité mesurée. Pour ce faire, différents types d'outils statistiques peuvent être employés :

- Régressions linéaires simples et multilples [10] (voir chapitre 2),
- Régressions aux moindres carrées partielles (PLS) [11],
- Arbres de décision [12],
- Réseaux de neurones [13-15],
- Algorithmes génétiques [16].
- Vecteurs Machines [15].

Une fois cette relation est établie et validée, elle peut alors être employée pour la prédiction de la propriété /activité de nouvelles molécules, pour lesquelles les valeurs expérimentales ne sont pas disponibles. De tels modèles peuvent être également utilisés pour mieux comprendre les mécanismes et les modes d'action.

I.3. Méthodologie générale d'une étude QSPR/QSAR

La méthodologie générale d'une étude QSAR/QSPR est la suivante :

a- Constituer une base de données à partir des mesures expérimentales fiables de la propriété ou de l'activité de chaque composé.

b- Sélectionner les descripteurs en relation avec la propriété ou l'activité étudiée.

c- Diviser cette base de données, aléatoirement, en une série d'apprentissage (*training set*) qui contient généralement les 2/3 de la base de données et une série de test (*test set*) constituée par le 1/3 restant.

d- Etablir des modèles mathématiques en utilisant la série d'apprentissage.

e- Caractériser les modèles élaborés par leurs indices de validation internes et vérifier leur robustesse par un test de hasardisation (*Randomization*) de la variable dépendante Y (réponse).

f-Valider les modèles élaborés en utilisant la série de test et calculer leurs paramètres statistiques de validation externe.

g- Elaborer le domaine d'applicabilité du modèle retenu.

h- Explorer et exploiter les modèles validés pour comprendre les mécanismes et les modes d'action.

Remarque:

La stratégie adoptée dans les programmes CODESSA et CORAL est différente de celle présentée ci-dessus. Ces programmes utilisent toutes les informations contenues dans la structure chimique et le meilleur modèle obtenu est celui qui possède les meilleures données statistiques. **Cependant, bien que ces modèles aient un bon pouvoir prédictif, ils ne sont pas toujours utiles pour interpréter et expliquer les mécanismes et les modes d'action.**

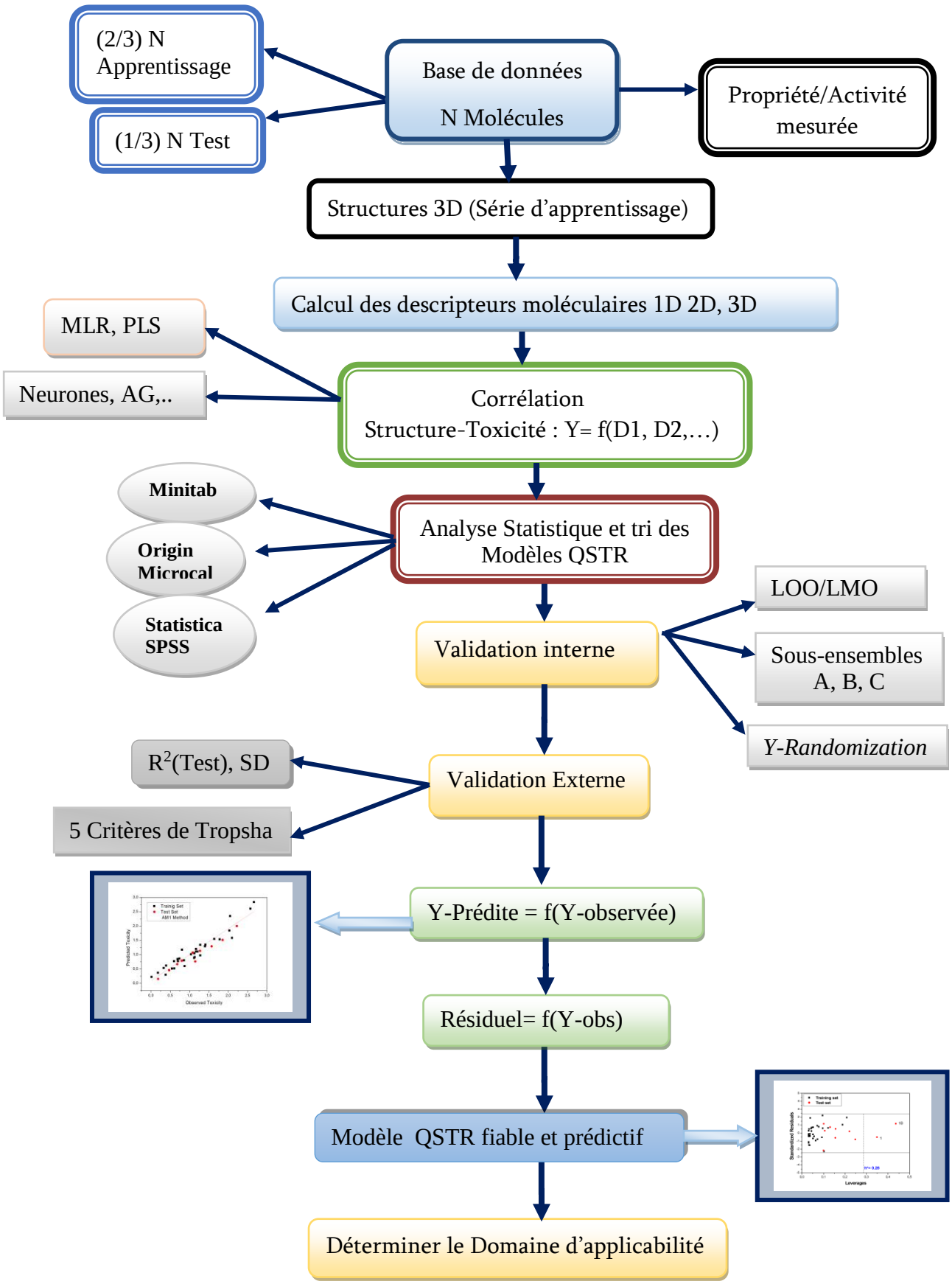


Figure.1. Schéma d'élaboration et validation d'un modèle QSTR

I.3.1. Bases de données

Un modèle QSTR est très dépendant des données expérimentales de référence. Le choix de la base de données est décisif dans le développement de tel modèle. Pour être de qualité, une base de données doit être composée de données expérimentales aussi fiables que possible obtenues en suivant un protocole unique puisque les erreurs sur celles-ci se propageront dans le modèle final. Il y'a plusieurs éléments à vérifier dans les étapes de nettoyage d'une base de données. Il faut tout d'abord vérifier que les structures sont correctes d'un point de vue chimique (règle de valence, …), des structures erronées entrainent la génération de mauvais descripteurs et donc de mauvais modèles. Plusieurs bases de données de toxicité sont disponibles sur le net, on peut citer :

- TETRATOX : http://www.vet.utk.edu/TETRATOX/index.php
- TOXNET : http://toxnet.nlm.nih.gov/
- ECHA : http://echa.europa.eu/fr

I.3.2. Descripteurs moléculaires

De nombreuses recherches ont été menées, au cours de ces dernières décennies, pour trouver la meilleure façon de représenter l'information contenue dans la structure des molécules et ces structures elles-mêmes en un ensemble de nombres réels appelés descripteurs qui sont soit théoriques ou empiriques :

I.3.2.1. Les descripteurs moléculaires théoriques

Les descripteurs 1D : sont accessibles à partir de la formule brute de la molécule et décrivent des propriétés globales du composé comme le nombre d'atomes et la masse moléculaire,…etc.

Ces descripteurs sont couramment utilisés du fait de leur extrême simplicité. Cependant, ils peuvent poser problème pour une bonne interprétation des mécanismes d'interaction du fait qu'ils ne permettent pas de tenir en compte des effets stériques et d'isomérie.

Les descripteurs 2D *:* sont calculés à partir de la formule développée de la molécule. On distingue :

- *Les indices 2D constitutionnels* : qui caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles,…etc.
- *Les indices 2D topologiques :* peuvent être obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Exemples: indice de Wiener [17], indice de Randić [18], indice de connectivité de valence de Kier-Hall [19], indice de Balaban [20], …etc

Ces descripteurs 2D permettent de prédire les propriétés physiques mais sont insuffisantes pour expliquer certaines propriétés et activités biologiques comme la toxicité.

Les descripteurs 3D *:* sont évalués à partir des positions relatives des atomes dans l'espace, et décrivent des caractéristiques plus complexes. Leurs calculs nécessitent donc de connaître la géométrie 3D de la molécule.

- Les *descripteurs 3D géométriques :* les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie.
- Les *descripteurs 3D électroniques* : permettent de quantifier les différents types d'interactions inter- et intramoléculaires, de grande influence sur l'activité biologique des molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie est minimale, et fait souvent appel à **la chimie quantique**.

Tableau.1. Exemples de descripteurs électroniques de la chimie quantique:

Descripteur	*Définition*
E_{LUMO}	Energie de l'orbitale la plus basse vacante.
E_{HOMO}	Energie de l'orbitale la plus haute occupée
μ	Potentiel chimique électronique
η	Dureté chimique
S	Mollesse chimique
ω	Indice d'électrophilie globale
AEI	Indice de l'énergie d'activation
Amax	Super-délocalisabilité maximale
MSD	Densité de spin de Mulliken
E_{SOMO}	Energie de l'orbitale occupée par un électron singulier

Parmi les descripteurs cités ci-dessus, on trouve les descripteurs de réactivité. Se sont des descripteurs utilisés pour étudier la réactivité chimique des molécules et ils sont dérivés de la DFT conceptuelle.

Descripteurs de réactivité issus de la DFT conceptuelle :

La DFT conceptuelle permet de caractériser les propriétés de réactivité des composés chimiques [**21**]. Les descripteurs qui en sont issus représentent un moyen simple de rationaliser le comportement chimique des molécules, sur la base du principe HSAB (*Hard and Soft Acids and Bases*) de Pearson [**22**]**.** Leur fiabilité a d'ailleurs été démontrée via différentes analyses théoriques, dédiées principalement à la réactivité chimique [**23-27**].

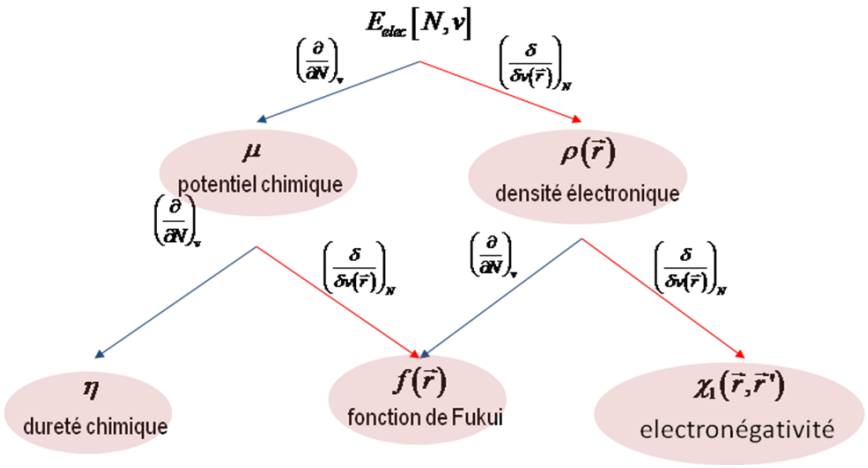


Figure.2. les descripteurs dérivant de la DFT

Les dérivées de l'énergie donnent deux catégories distinctes d'indices de réactivité

1) ***les indices globaux :*** sont obtenus par les dérivées de l'énergie par rapport à N

- le potentiel chimique « μ »
- l'électronégativité χ
- la dureté η
- la mollesse S
- Indice d'électrophilie $\omega = \frac{\mu^2}{2\eta}$

2) ***les indices locaux*** : se sont les dérivées de l'énergie qui ne dépendent que d'une coordonnée spatiale tels que, la densité électronique, les indices de Fukui (f^-, f^+, f^0) la dureté locale, la mollesse locale

Les descripteurs 3D spectroscopiques : les molécules peuvent être caractérisées par des mesures spectroscopiques, par exemples par leurs fonctions d'onde vibrationnelles. Les spectres infrarouges peuvent être obtenus par calcul théorique, après recherche de la géométrie optimale de la molécule. Ces spectres sont alors codés en vecteurs de descripteurs de taille fixe. Les descripteurs de type *MoRSE (Molecule Representation of Structures based on Electron diffraction)* [**28**] sont calculés à partir d'une simulation du spectre infrarouge. Par ailleurs, le calcul de certains descripteurs demande une étape

préliminaire d'alignement des molécules, pour les placer dans une orientation commune. Ces descripteurs, qui sont souvent spécifiques à certaines méthodes *(CoMFA, CoMSIA...).*

I.3.2.2. Les descripteurs moléculaires empiriques

Il existe plusieurs descripteurs qui peuvent être mesurer expérimentalement tels que le coefficient de partage (octanol-eau) (log *P*), la polarisabilité (α).

- **Coefficient de partage (octanol-water) Kow, log *P***

Le coefficient de partition est défini comme le rapport de la concentration du soluté dans la phase huileuse C' à la concentration du soluté non ionisé dans la phase aqueuse C, à l'équilibre.

C' > C ⇒P >1 ⇒log P >0 : le soluté est dit lipophile (hydrophobe)

C' < C ⇒P <1 ⇒log P <0 : le soluté est dit hydrophile

Le partage d'une molécule entre une phase aqueuse et une phase lipidique conditionne en partie ses propriétés biologiques telles que le transport, le passage à travers les membranes, l'affinité pour un récepteur et la fixation par une protéine, l'activité pharmacologique ou encore la toxicité. S'agissant de contaminants, ce même partage conditionne leur devenir dans notre environnement en particulier leur accumulation dans les organismes aquatiques. Depuis les travaux de Collander **[29]**, puis ceux du groupe de Hansch **[30]** quelques années plus tard, le coefficient de partage *P* d'une molécule dans un système biphasique constitué de deux solvants non-miscibles (le plus souvent le système *n*-octanol/eau), est reconnu pour sa faculté à mimer le passage de cette molécule à travers les membranes biologiques. Le partage est donc une propriété physico-chimique importante qui peut être utilisée pour représenter la nature lipophile ou hydrophile d'une molécule.

Le coefficient de partage est largement utilisé dans des études de relations structure-activité quantitatives (QSARs) dans les sciences pharmaceutiques, biochimiques, toxicologiques et dans les sciences de l'environnement. La lipophilie intéresse donc tout autant la communauté qui étudie les problèmes de santé humaine que celle qui est impliquée dans les problèmes de l'environnement.

- Détermination expérimentale de log *P*

La méthode expérimentale utilisée pour déterminer le coefficient de partage d'une molécule est la méthode des flacons agités ou «*shake-flask* ». Cette méthode reste cependant la méthode de choix pour les molécules organiques et de ce fait, elle est préconisée comme procédure standard de caractérisation par l'OCDE [31] (Organisation de Coopération et de Développement Economique).

Toutefois, elle tend à être supplantée par les méthodes chromatographiques. En particulier, la chromatographie liquide haute performance à polarité de phase inversée, adaptée aux études de criblage, elle est aussi préconisée par l'OCDE [32]. Dans ce cas, on utilise comme indice de lipophilie, une valeur déduite de la mesure des temps de rétention.

- Méthodes de calcul et d'estimation de log *P*

La plupart des méthodes expérimentales de détermination de log *P* souffrent du même inconvénient, à savoir que leur domaine d'application est relativement étroit. D'autre part, du fait de la nature intrinsèque de certaines molécules, leurs log *P* sont inaccessibles à l'expérience. C'est le cas en particulier des surfactants qui ont tendance à s'accumuler à l'interface du système bi-phasique au lieu de se disperser dans les deux phases. Enfin, dans le domaine de la conception assistée par ordinateur ou dans le domaine de la chimie combinatoire, les chercheurs travaillent sur des modèles moléculaires avant même que les molécules aient été synthétisées. Ceci explique le succès de nombreuses méthodes d'estimation de log *P* qui ont été décrites dans la littérature depuis plusieurs décennies [33]. Les plus anciennes sont des méthodes fragmentales dans lesquelles une molécule est divisée en fragments prédéfinis et les contributions correspondantes sont sommées pour conduire à une valeur estimée du log *P*. Parmi ces méthodes on peut citer

- Méthode de Hansch [34]
- Méthode de Rekker [35]
- La méthode de Ghose et Viswanadhan [36]
- La méthode de Klopman et Iroff [37]

- La méthode de Bodor [**38**]

Parmi les logiciels utilisés pour l'estimation de log *P* on peut citer : ACD/LABS, Clog P, Hyperchem, KOWWIN

I. 3.3. Méthodes d'analyse des données :

Pour élaborer un modèle QSPR/QSAR nous avons besoin d'une méthode d'analyse de données, cette méthode permet de quantifier la relation qui existe entre la propriété/Activité et la Structure (descripteurs).

Il existe plusieurs méthodes pour construire un modèle et analyser les données statistiques de ce dernier, certaines sont linéaires telles que la régression linéaire multiple (MLR), la régression aux moindres carrés partiels (PLS), d'autres sont non linéaires comme les arbres de décisions, les réseaux de neurones… ces méthodes sont disponibles dans des logiciels tels que, Excel, Origin Microcal, **Minitab**, Statistica, SPSS, R,…

La méthode utilisée dans nos études est la méthode de Régression Linéaire Multiple (MLR) implémentée sur Minitab et qui est présentée en détail dans le chapitre II.

I.3.4. Interprétation et validation d'un modèle QSPR/QSAR

Une fois développé, le modèle doit être interprété en analysant tous les paramètres statistiques de ce modèle (Chapitre II), sa qualité doit être aussi étudiée, cette qualité est vérifiée par ce que l'on appelle validation. Sa robustesse, c'est-à-dire l'influence des composés de la série d'apprentissage sur le modèle, est estimée par des méthodes de validation interne. Afin d'estimer son pouvoir prédictif, des données expérimentales supplémentaires sont nécessaires afin de déterminer la capacité du modèle à prédire ces valeurs c'est ce que l'on appelle validation externe. Enfin, il est important de savoir quel type de molécules utilisées avec quel modèle. On parle alors de domaine d'applicabilité.

I.3.4.1.Validation interne

Dans le passé, la validation interne d'un modèle QSPR/QSAR a été réalisée en utilisant la validation croisée LOO (*leave-one out*) ou LMO (leave-many out) qui est quantifiée par

le coefficient R^2_{cv} (chapitre II). Ce processus consiste à extraire un certain nombre k de molécules du jeu initial à N molécules et à construire un nouveau modèle avec les (N-k) molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent). Ce processus est ensuite réitéré pour retirer et prédire les valeurs de toutes les molécules de la série d'apprentissage. En fonction du nombre de molécules retirées à chaque itération, on parlera de Leave-One-Out (LOO) ou de Leave-Many-Out (LMO) selon qu'une ou plusieurs molécules est (sont) retirée(s) [**39**]. Dans ces dernières années, d'autres méthodes sont utilisées pour faire la validation interne, tel que la hasardisation de la réponse (Y-*Randomization*).

- ***Y-Randomization***

Afin de s'assurer qu'un modèle QSPR est fiable, les tests de *Y-randomization* [**40**] sont une des techniques les plus employées. En effet, il n'est pas rare d'obtenir des corrélations fortuites (ou « *chance corrélation* »), c'est-à-dire un modèle affichant de bons résultats statistiques (R^2, SD) pour l'apprentissage, mais impliquant des descripteurs qui ne sont pas reliés à la propriété/activité modélisée. Ces modèles aléatoires peuvent être détectés par la procédure *Y-randomization*. Elle consiste à mélanger aléatoirement les propriétés/activités expérimentales pour la série d'apprentissage en utilisant les mêmes descripteurs, de nouveaux modèles sont obtenus. Ces derniers doivent avoir des performances très faibles.

Cependant, la validation interne est insuffisante pour étudier le pouvoir prédictif d'un modèle. Pour cette raison la validation externe du modèle est devenue une norme et une partie obligatoire dans la modélisation QSPR/QSAR [**41, 42**].

I.3. 4. 2. Validation externe

Cette méthode consiste à prédire la propriété/activité d'une série de molécules appelée généralement série de test qui ne sont pas dans la série de développement du modèle, cette validation est caractérisée par les paramètres R^2(test) R^2_{cv} (test). Récemment plusieurs études [**43, 44**] ont montré l'insuffisance des paramètres R^2, R^2_{CV} pour vérifier 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 » (*Tropsha criteria*) [**43**].

Critères de validation Externe (série de test)

- $R^2 > 0.7$ (critère 1)
- $R^2_{cv} > 0.6$ (critère 2)

- $\frac{R^2 - R_0^2}{R^2} < 0.1$ et $0.85 \leq k \leq 1.15$ (critère 3)
- $\frac{R^2 - R'^2_0}{R^2} < 0.1$ et $0.85 \leq k' \leq 1.15$ (critère 4)

- $\left|R^2 - R_0^2\right| \leq 0.3$ (critère 5)

Avec

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

R^2_0 coefficient de corrélation entre les valeurs prédites et expérimentales pour la série de test.

R'^2_0 coefficient de corrélation entre les valeurs expérimentales et prédites pour la série de test.

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

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

I.3.4. 3. Domaine d'applicabilité

Comme les modèles QSAR sont utilisés pour la prise de décisions relatives aux augmentations de gestion des produits chimiques, la fiabilité des prévisions du modèle est une préoccupation croissante.

Un modèle QSAR/QSPR ne peut pas être considéré comme un modèle universel, parce qu'il est développé sur un nombre limité de composés qui ne couvrent pas tout l'espace

chimique. Par conséquent l'activité/propriété prédite d'un composé, chimiquement dissimilaire au jeu d'apprentissage, ne pourra pas être considérée fiable [**45, 46**]. Le domaine d'applicabilité (DA) permet de définir la zone dans laquelle un composé pourra être prédit avec confiance. 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 [**47**]. Ils existent plusieurs méthodes pour la détermination de domaine d'applicabilité d'un modèle QSPR/QSAR parmi ces méthodes on trouve la méthode de « *leverage ».*

Méthode de « *leverage* » *:* cette méthode est basée sur la variation des résiduels standardisés de la variable dépendante avec « *leverage* » (la distance entre les valeurs des descripteurs et leurs moyennes). Si un composé a un résiduel et un leverage qui dépasse le seuil **$h^*=3p/n$** (ou p est le nombre de descripteurs plus 1 et n le nombre d'observations), ce composé est considéré en dehors du domaine d'applicabilité du modèle élaboré.

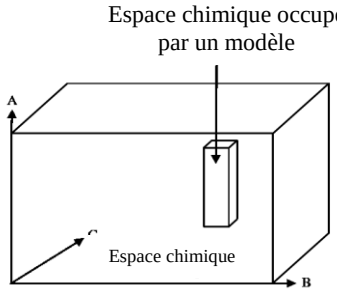


Figure.3. représentation schématique du domaine d'applicabilité d'un modèle QSPR/QSAR

I.4. Applications des méthodes QSPR/QSAR

Les applications des méthodes QSPR/QSAR sont très nombreuses, elles touchent tous les domaines ou la structure chimique intervient, entre autre on peut citer :

- *Propriétés physico-chimiques* : point d'ébullition, Point de fusion, densité, indice de réfraction, température critique, viscosité, solubilité, pression de vapeur, tension superficielle, Coefficients de partition : eau/octanol, air/eau, huile/air, lait/plasma **[48]**…

- *Activités biologiques* :

  Anti VIH, Anti malaria, Anti Diabète, Anti Cancer, Anti-oxydante, Anti-inflammatoire…

- *Autres propriétés/activités* :

  - Prédiction de la toxicité aquatique des composés chimiques vis-à-vis des espèces environnementales.
  - Toxicité des nanoparticules
  - Toxicité des pesticides et des colorants
  - Propriétés inhibitrices de corrosion
  - Concentration micellaire critique
  - Prédiction de plusieurs propriétés dangereuses telle que l'explosibilité et l'inflammabilité de certaines familles de molécules chimiques.
  - Conception des médicaments et de nombreux autres produits tels que les agents tensio-actifs, parfums, les colorants et les produits de la chimie fine .

  …

Références Bibliographiques

[1] N. Margossian, Le règlement REACH - La règlementation européenne sur les produits chimiques, Dunod / L'usine Nouvelle, Paris, **2007**.

[2] A. Crum Brown, T.R. Fraser, On the connection between chemical constitution and physiological action. Part 1. On the physiological action of salts of the ammonium bases, derived from strychnia, brucia, thebia, codeia, morphia and nicotia. *Trans. Roy.Soc. Edinburgh*, **1868**, 25, 151-203.

[3] L.P. Hammett, The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc*. **1937**, 59, 96-103.

[4] A.C. Brown, and T.R. Fraser, On the connection between chemical constitution and Physiological Action; with special reference to the physiological action of the salts of the ammonium bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia, *J. Anat. Physiol*. **1868**, 2, 224-242.

[5] C. Hansch, A. Leo, D. Hoekmann, Exploring QSAR: hydrophobic, electronic and steric constants.Washington, DC: *Am. Chem. Soc*.**1995**.

[6] S.M. Free, J.W. Wilson, A mathematical contribution to structure–activity studies. *J. Med. Chem.***1964**, 7, 395-399.

[7] M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure- Property relationships in pharmaceutical research Part2. *Pharm. Sci. Tech. Today*, **2000**, 3, 50-57.

[8] M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure–property relationships in pharmaceutical research–Part 1*Pharm. Sci. Tech. Today*, **2000**, *3*, 28-35.

[9] A.R.D. Katritzky, C. Fara, R.O. Petrukhin, D.B. Tatham, U. Maran, A. Lomaka, M.Karelson, The present utility and future Potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors, *Curr. Top. Med. Chem.* **2002**, 2,1333-1356.

[10] J. Ghasemi, S. Saaidpour, S.D. Brown, QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *J. Mol. Struct. (Theochem)*. **2007**, 805, 27-32.

[11] P. Geladi, B.R. Kowalski, Partial Least Squares Regression: a Tutorial, *Anal. Chim. Acta*. **1986**, 185, 1-17.

[12] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, An introduction to decision tree modeling, *J. Chemom.* **2004**, 18, 275-285.

[13] A.F. Duprat, T. Huynh, G. Dreyfus, Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of LogP*J. Chem. Inf. Comput. Sci.* **1998**, 38, 586-594.

[14] I.V. Tetko, A.E.P. Villa, D.J. Livingstone, An Enhancement of Generalization Ability in Cascade Correlation Algorithm by Avoidance of Overfitting/Overtraining, *J. Chem. Inf. Comput. Sci.* **1996**, 36, 794-803.

[15] J. Gasteiger, J. Zupan, Neural Networks in Chemistry, *Angew. Chem. Int. Ed. Engl.* **1993**, 32, 503-527.

[16] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review *J. Chemometr.* **2001**, 15, 559-569.

[17] H. Wiener, Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, 69, 17–20.

[18] M. Randic, Characterization of molecular branching. *J. Am. Chem. Soc*. 97, 6609–6615, *J. Am. Chem. Soc.* **1975**, 97, 6609-6615.

[19] L.B. Kier, L.H. Hall, W.J. Murray, Molecular connectivity I: Relationship to nonspecific local anesthesia. *J. Pharm. Sci.* **1975**, 64, 1971–1974

[20] A.T. Balaban, Highly Discriminating Distance-Based Topological Index, *Chem. Phys. Lett.* **1982**, 89, 399-404.

[21] P. Geerlings, F. De Proft, W. Langenaeker, Conceptual density functional theory, *Chem. Rev.* **2003**, 103, 1793-1874.

[22] R.G. Pearson, Hard and Soft Acids and Bases, *J. Am. Chem. Soc.* **1963**, 85, 3533-3539.

[23] C.A. Caro, J.H. Zagal, F. Bedioui, C. Adamo, G.I. Cardenas-Jiron, *J. Phys. Chem. A.* **2004**, 108, 6045-6051.

[24] G.I. Cardenas-Jiron, S. Gutierrez-Oliva, J. Melin, A. Toro-Labbe, Relations between Potential Energy, Electronic Chemical Potential, and Hardness Profiles *J. Phys. Chem. A*. **1997**, 101, 4621-4627.

[25] P.K. Chattaraj, P. Perez, J. Zevallos, A. Toro-Labbe, Ab Initio SCF and DFT Studies on Solvent Effects on Intramolecular Rearrangement Reactions, *J. Phys. Chem. A.*

2001, 104, 4272-4283.

[26] R.G. Parr, R.A. Donnelly, M. Levy, W.E. Palke, Electronegativity: The Density Functional Viewpoint, *J. Chem. Phys.***1978**, 68, 3801-3807.

[27] F. De Vleeschouwer, P.Jaque, P.Geerlings, A. Toro-Labbe, F. De Proft, Regioselectivity of Radical Additions to Substituted Alkenes: Insight from Conceptual Density Functional Theory. *J. Org. Chem*. **2010**, 75, 4964–4974.

[28] J. Schuur, P. Selzer, J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci*. **1996**, 36,334–344.

[29] R. Collander, Permeability of Plant Cells*, Ann. Rev. Plant. Physiol*. **1957**, 8, 335-348.

[30] A. Leo, C. Hansch, D. Elkins, Partition coefficients and their uses , Chemical reviews, Departement of chemistry, Pomona college, Claremont, California,**1971**.

[31] OECD Guidelines for Testing of Chemicals No. 107, OECD,Paris, **1992**.

[32] OECD Guidelines for Testing of Chemicals No. 117, OECD, Paris, **1992**.

[33] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol*. **1971**, 55, 379–384.

[34] R.S. Pearlman, In: Partition Coefficient Determination and Estimation, Dunn, W. J.; Block, J. H.; Pearlman, R. S. Eds., Pergamon, New York, **1986**, 3-20.

[35] R. F. Rekker, H. M. de Kort. (1979). The hydrophobic fragmental constant: An extension to a 1000 data point set. *Eur. J. Med. Chem.* **1979**, 14, 479.

[36] A.K. Ghose, G.M. Crippen. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci*. **1987**, 27, 21–35.

[37] G.Klopman, L.D. Iroff, Calculation of partition coefficients by the charge density method, *J. Comput. chem.* **1981**, 2, 157-160.

[38] N. Bodor, Buchwald P. Molecular size based approach to estimate partition properties for organic solutes. *J. Phys. Chem B*. **1997**, 101, 3404.

[39] L. Zhang, H. Zhu, T.I. Oprea, A. Golbraikh, A. Tropsha, QSAR Modeling of the

Blood-Brain Barrier Permeability for Diverse Organic Compounds*, Pharm. Res.* **2008**, 25, 1902–1914.

[**40**] L. He, P.C. Jurs, Assessing the reliability of a QSAR model's predictions. *J. Mol. Graph. Model.* **2005**, 23, 503-523.

[**41**] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb.Sci.* **2003**, 22, 69-77.

[**42**] P.P. Roy, S. Paul, I. Mitra, K. Roy, On Two Novel Parameters for Validation of Predictive QSAR models, *Molecules*, **2009**, 14*,* 1660-1701.

[**43**] A. Golbraikh, A. Tropsha, Beware of q^2! *J. Mol. Graph. Model.* **2002***,* 20*,* 269–276*.*

[**44**] 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? *J. Chem. Inf. Model.* **2012**, 52, 2570–2578.

[**45**] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, Varnek A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena-pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem.Inf. Model.* **2008**, 48, 1733

[**46**] J. Jaworska, N.N. Jeliazkova, T. Aldenberg: QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim* **2005**, 33, 445-459.

[**47**] T. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M. Cronin, P. Gramatica, J. Jaworska, S. G. Kahn, C. Klopman, G. Marchant, N.N. Myatt, G.Jeliazkova, R. Patlewicz, D. Roberts, T. Schultz, D. Stanton, J. Sandt, W. Tong, G. Veith, C. Yang, Current status of methods for defining the applicability domain of (Quantitative) Structure–Activity Relationships. *Altern. Lab. Anim.* **2005**, 33, 1-19.

[**48**] A. R. Katritzky, M. Kuanar, S. Slavov, and C. D. Hall, Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction, *Chem. Rev.* **2010,** 110*,* 5714–5789

CHAPITRE II

STATISTIQUES ET ANALYSE DES DONNEES

Introduction

L'élaboration de modèles QSPR/QSAR n'est pas une chose facile. La première difficulté réside dans la différence d'échelles existant entre les données à corréler. La structure étant à une échelle moléculaire alors que les activités /propriétés à prédire sont à une échelle macroscopique. Un des problèmes importants réside également dans le traitement de données. En fait, de nombreux outils existent et il s'agit de trouver le moyen le plus adapté pour obtenir un modèle fiable à partir des données disponibles.

L'analyse de la régression est l'outil le plus utilisé en modélisation QSPR/QSAR. L'idée est de décrire et d'évaluer la relation entre une variable (dite variable à expliquer ou variable dépendante) souvent notée Y, et une (ou plusieurs) variable(s), dite(s) variable(s) explicative(s) ou indépendante (s) notée (s) X.

Dans ce chapitre, nous présentons la régression linéaire ceci dit la régression linéaire simple (SLR) et la régression linéaire multiple (MLR) utilisées pour élaborer les modèles QSPR/QSAR.

II.1. Régression linéaire simple (RLS)

Pour étudier le lien entre les variables Y et X, on doit tout d'abord donner un rappel sur quelques notions de statistique

- ***Espérance:*** désigne la moyenne des valeurs prises par *X*, et pondérées par leurs probabilités de réalisation.

$$\mu = E(X) = \sum_{i=1}^{i=k} x_i P(x_i = X)$$

- ***Variance*** *:* est la moyenne des écarts par rapport à la moyenne arithmétique, en terme mathématique elle peut être considérée comme une mesure servant à caractériser la dispersion d'une distribution. Elle est donnée par :

$$V(X) = \frac{1}{n}\sum_{i}^{n}(X_i - \overline{X})^2 \text{, avec } \quad \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{la moyenne de la variable X}$$

- ***Covariance*** : elle permet d'étudier les variations simultanées de deux variables X et Y par rapport à leurs moyennes respectives.

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

- ***Ecart-type****:* est une mesure de la dispersion d'une variable aléatoire, c'est la racine carrée de la variance.

$$\sigma = \sqrt{V(X)}$$

La relation linéaire entre les deux variables, explicative (X) et à expliquer (Y) est de la forme suivante :

$$Y=a+bX$$

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

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

Avec β_0 (appelé *Intercept*) et β_1 sont les paramètres inconnus, appelés coefficients de régression, que l'on cherche à déterminer. ε *,* le terme d'erreur, qui est supposé être une variable aléatoire et vérifie :

- La variance des résidus est constante, $V(\varepsilon_i)$ = constante.

- La covariance des résidus est nulle, $Cov(\varepsilon_i, \varepsilon_j) = 0$ (si $i \neq j$), ce qui implique la non corrélation des résidus. Ces hypothèses sont généralement appelées **hypothèses faibles**. Les **hypothèses fortes** supposent en plus la normalité des résidus (ce qui implique donc leur indépendance puisqu'ils sont non corrélés), qui nous permettra par la suite d'effectuer des tests sur le modèle de régression linéaire.

On suppose que l'on dispose d'un échantillon de n observations pour Y et X, soit

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$\{(y_1, x_1),\ldots,(y_n, x_n)\}$. Ainsi, pour chaque observation i, on peut écrire :

D'un point de vue matriciel le modèle de régression linéaire s'écrit

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

L'objectif est de déterminer des valeurs pour les paramètres β_0 et β_1 à partir des n observations sur les variables X et Y.

Graphiquement, il s'agit de trouver les paramètres de la droite de régression, qui passe au milieu du nuage de points dessiné dans le plan (X, Y).

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

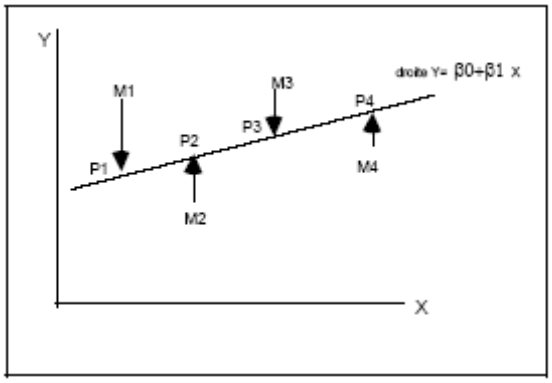


Figure.1. Représentation graphique de la régression

Pour trouver la droite qui passe « au plus près» de tous les points il faut se donner un critère d'ajustement. On projette les points de M1 a M4 parallèlement à l'axe Y. Sur la droite on obtient les points P1 à P4, comme le montre la Figure 1. Le critère retenu pour déterminer la droite passant au plus près de tous les points sera tel que : « La somme des carrés des écarts (SCE) des points observés M_i a la droite solution soit minimum »

La droite solution sera appelée droite de régression de Y sur X. Le critère d'ajustementest le «critère des Moindres Carrées Ordinaires » (Ordinary Least Squares). Les écarts sont calculés en projetant les points M_i parallèlement à l'axe Y.

II.1.1- Méthode d'estimation des paramètres β_0 et β_1

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

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

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

La valeur de cette fonction S est minimum lorsque les dérivées de S par rapport à β_0 et β_1s'annulent. La solution est obtenue en résolvant le système :

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

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

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

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

L'équation **(3)** nous donne

$$\sum Y_i - n\hat{\beta}_0 - n\hat{\beta}_1 \sum X_i = 0 \qquad (5)$$

En utilisant la formule de la moyenne $\overline{X} = \sum_{i}^{n} \frac{X_i}{n}$ et en divisant par n nous obtenons l'équation ci-dessous

$$\hat{\beta}_0 = \overline{Y} - \beta_1 \overline{X} \qquad (6)$$

L'équation **(4)** nous donne

$$\sum Y_i X_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2 = 0 \quad (7)$$

On remplace $\hat{\beta}_0$ obtenu dans l'équation **(6)** dans l'équation **(7)** nous aboutissons à l'équation ci-dessous

$$\sum Y_i X_i - (\overline{Y} - \hat{\beta}_1 \overline{X}) \sum X_i - \hat{\beta}_1 \sum X_i^2 = 0 \quad (8)$$

Cette équation nous permet d'obtenir la formule du deuxième coefficient de l'équation de régression

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X_i})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \quad (9)$$

En remplaçant l'expression de $\hat{\beta}_1$ dans l'équation (6) on obtient

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

$$\hat{Y} = \beta_0 + \beta_1 X \quad (10)$$

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

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

$$Y_i - \hat{Y}_i = \hat{\varepsilon} i \quad (11)$$

la variation résiduelle appelée σ^2 est donnée par :

$$S^2 = \frac{1}{n-2} \sum_i^n \hat{\varepsilon}_i^2 \quad (12)$$

II.1.2- Décomposition de la variance et qualité de la régression :

A partir de l'équation de la droite de régression (modèle retenu), on peut pour tout point i d'abscisse X_i calculer son estimation (ordonnée) Y_i. A partir des équations **(6)** et **(10)** on obtient :

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_1(X_i - \bar{X}) \quad \textbf{(13)}$$

Ou encore

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_1(X_i - \bar{X}) \quad \textbf{(14)}$$

$$Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y}) \quad \textbf{(15)}$$

On élève les deux membres au carré et on somme sur les observations i on obtient

$$\sum(Y_i - \hat{Y}_i)^2 = \sum (Y_i - \bar{Y})^2 + (\hat{Y}_i - \bar{Y})^2 - 2\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) \quad \textbf{(16)}$$

En utilisant l'équation **(14)** on aboutit à

$$\sum(Y_i - \hat{Y}_i)^2 = \sum (Y_i - \bar{Y})^2 + (\hat{Y}_i - \bar{Y})^2 - 2\sum(Y_i - \bar{Y})\hat{\beta}_1(\hat{X}_i - \bar{X}) \quad \textbf{(17)}$$

En utilisant la transformation de l'équation **(9)** $\hat{\beta}_1 \sum(X_i - \bar{X})^2 = \sum(X_i - \bar{X})(Y_i - \bar{Y})$ on obtient

$$\sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \bar{Y})^2 + \sum(\hat{Y}_i - \bar{Y}) - 2\hat{\beta}_1^2 \sum(X_i - \bar{X})^2 \quad \textbf{(18)}$$

En utilisant l'équation **(14)** on aboutit à la relation

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

Finalement on obtient la relation fondamentale de l'analyse de la variance

$$\sum_i (Y_i - \overline{Y})^2 = \sum_i (\hat{Y}_i - \overline{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2 \quad (20)$$

Ces diverses sommes des carrés sont définies comme suit :

* La quantité SST (Sum of Squares Total) représente la somme des carrés des écarts à la moyenne (SCE total).

$$SST = \sum (Y_i - \overline{Y})^2 \quad (21)$$

Cette variation totale à expliquer, peut se décomposer en :

-Somme des carrés expliquée par le model, ou plus précisément par la variable (X)

$$SSR = \sum (\hat{Y}_i - \overline{Y})^2 \quad (22)$$

-Somme des carrés des résidus (partie que le modèle n'explique pas)

$$SSE = \sum \hat{\varepsilon}_i^2 = \sum (Y_i - \hat{Y})^2 \quad (23)$$

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

$$SST = SSR + SSE \quad (24)$$

Cette dernière formule montre que les variations de Y de sa moyenne, c'est-à-dire SCE totale (SS Total) peuvent être expliquées par le modèle grâce à SCE (Modèle) (SS Model ou SS régression) et ce que ne peut être explique par le modèle est dans SCE (Erreurs) (SS Error).

II.1.3. Représentation de l'analyse de la variance :

Toutes ces quantités sont présentées habituellement sur un tableau appelé table d'analyse de variance ou table d'ANOVA (*Analysis Of Variance*) faisant apparaitre les sources de variation : le total (en 3ème ligne de la table) qui se décompose en deux parties : la partie modèle et la partie erreur. A chaque source de variation correspond un nombre de degrés de liberté (DF) respectivement égal à n-1, p, n-p-1 (ou n est le nombre d'observations, p le nombre de variables régresseurs (la variable X_0, constante égale à 1, correspondant au paramètre β_0, n'est pas comprise).

Nous présentons le tableau général de l'analyse de variance, tous les logiciels effectuant des calculs de régression, donnent comme sortie ce tableau.

Tableau .1. *Analyse de la variance*

Source de Variation	*Degrés de liberté DF*	*Somme des carrés SS*	*Moyenne des carrées MS*
Model	p	$\sum(\hat{Y}_i - \overline{Y})^2$	$\frac{\sum_i^n (\hat{Y}_i - \overline{Y})^2}{p}$
Error	n-1-p	$\sum(Y_i - \hat{Y}_i)^2$	$\frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-p-1}$
Total	n-1	$\sum_i^n (Y_i - \overline{Y})^2$	$\frac{\sum_i^n (\hat{Y}_i - \overline{Y})^2}{p} + \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-p-1}$

DF : degré de liberté

SS : Sum of squares : somme carrée des écarts

MS : Mean square, est le rapport SS/DF

MSE: Mean Square Error

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

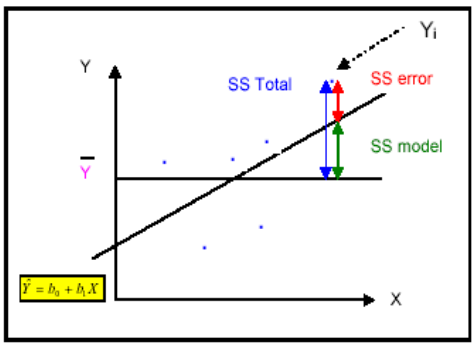


Figure.2. *Décomposition des différents écarts*

Une fois le modèle de régression établi, il convient dans un premier temps de vérifier si les hypothèses faites lors de l'estimation par moindres carrées sont respectées (normalités des résidus, Non-corrélation des résidus). Dans un second temps, nous évaluons la qualité du modèle de régression et nous testons sa validité.

II.1.4. Hypothèses de l'analyse de régression linéaire

Normalité des résidus

- Loi normale

En théorie des probabilités, on dit qu'une variable aléatoire réelle x suit une loi normale (ou loi normale gaussienne, loi de Laplace-Gauss) d'espérance μ et d'écart type σ strictement positif (donc de variance σ^2) si cette variable aléatoire réelle x admet pour densité de probabilité la fonction $p(x)$ définie, pour tout nombre réel x, par :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad (25)$$

Une telle variable aléatoire est alors dite variable gaussienne. On note habituellement cela de la manière suivante: $x \approx N(\mu, \sigma^2)$

- **Loi normale centrée réduite**

Cette loi est un cas particulier de la loi normale, ou la variable x est centrée réduite.

Une variable centrée réduite a

- Une moyenne nulle

- Une variance égale à 1
- Un écart type égal à 1

La fonction de densité de probabilité devient alors

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x)^2} \quad (26)$$

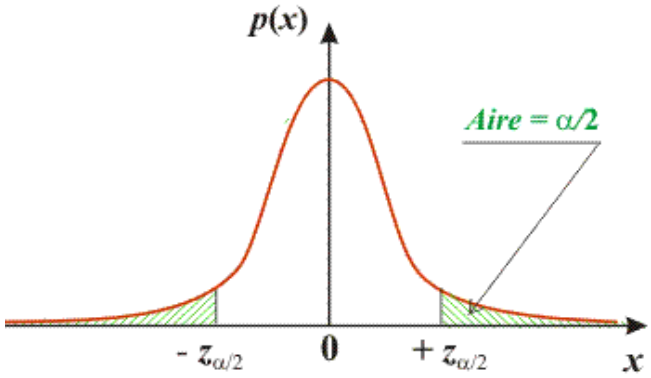


Figure. 3. Représentation graphique de la fonction de densité de la loi normale centrée réduite

II.1.5. Qualité de la régression linéaire

Les objectifs d'une modélisation statistique peuvent être de natures différentes, les objectifs explicatifs et les objectifs prédictifs. Ces objectifs sont déterminés par plusieurs coefficients.

II.1.5.1 Coefficient de détermination R^2

Pour évaluer la précision avec laquelle la dépendance trouvée décrit la variance de la variable dépendante (c'est à dire la qualité de l'ajustement statistique), le carré du coefficient de détermination (R^2) est utilisé.

On définit alors le coefficient de détermination, qui mesure la part de la variance expliquée par le modèle dans la variance totale.

Un modèle, que l'on qualifie de bon, possède des estimations proches des vraiesvaleurs de Y. Les deux quantités SCE totale (SST) et SCE modèle (SSR) sont des sommes des carrés

donc toujours positives ou nulles et telles que : SSR SST. Le rapport de ces deux quantités nous donne le coefficient de corrélation R^2:

$$\boldsymbol{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}} \quad \textbf{(27)}$$

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

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{b_1^2 \sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{cov}(x_i, y_i)^2}{\text{var}(x_i)\,\text{var}(y_i)} = r_{xy}^2 \quad \textbf{(28)}$$

II.1.5.2 Déviation standard (SD)

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

$$s^2 = MSE = \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n - p - 1} = \frac{SSE}{n - p - 1} \quad \textbf{(29)}$$

Root MSE= SD= s **(30)**

L'estimation de l'erreur-type appelé aussi déviation standard SD est une mesure de la dispersion des valeurs observées de la variable dépendante sur la droite de régression (de surface). Les petites valeurs de SD signifient un bon ajustement statistique du modèle et une forte fiabilité de la prédiction.

II.1.6.Confiance à accorder aux résultats

On cherche à tester la validité d'hypothèses concernant les (vrais) paramètres du modèle.

On se donne une hypothèse nulle, notée H0 que l'on teste, et une hypothèse alternative, notée H1.

La procédure est basée sur la construction d'une statistique, calculée sur l'échantillon aléatoire afin de décider, avec un niveau de confiance raisonnable, si on peut supposer que les données de l'échantillon suivent l'hypothèse nulle (c'est-`a-dire si on peut supposer que l'hypothèse nulle est acceptable).

La statistique retenue dépend de l'hypothèse que l'on teste (une statistique de Student quand on ne teste qu'un contrainte (un seul paramètre) une statistique de Fisher quand on teste plusieurs contraintes (tous les paramètres).

On détermine ensuite une règle de décision pour savoir si l'hypothèse nulle doit être acceptée ou rejetée. Plus précisément, on détermine des intervalles de confiance autour des valeurs estimées.

On cherche à :

- Tester la signification globale de la régression
- Tester la signification de chaque paramètre

II.1.6.1.Test de la signification globale de larégression :

- **Test de Fisher-Snedecor**

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

Ce test permet de connaître l'apport global de l'ensemble des variables X_1, …..Xp à la détermination de Y.

On teste l'hypothèse nulle : H0 : $\beta_{1=.....}$ $\beta_p=0$ contre Ha : il existe au moins un β_j parmi $\beta_{1....}B_p$ non égal à 0.

$$F = \frac{MSmodel}{MSerror} \qquad (31)$$

avec

$$MS\,mod\,el = \frac{SSModel}{p} \qquad (32)$$

et

$$\text{MSerror} = \frac{\text{SSerror}}{\text{n} - \text{p} - 1} \tag{33}$$

Si H0 est vraie et sous réserve des suppositions suivantes :

- Ce rapport F est une valeur observée d'une variable à p et n-p-1 degrés de liberté.
- Si les ε_i sont indépendants et suivent une loi normale de même variance $\varepsilon_i \approx N(0, \sigma^2)$

Alors la statistique F suit une loi de Fisher-Snedecor

$$\text{F} = \frac{\text{MS mod el}}{\text{MSerror}} \approx \text{F}(\text{p}, \text{n} - \text{p} - 1)$$

Une équation de régression est considérée comme statistiquement significative si la valeur observée de F est supérieure à une valeur tabulée pour le niveau de signification choisi (typiquement 95%) et les degrés de liberté correspondants de F. Les degrés de liberté de F sont égaux à p et n-p-1. L'importance de l'équation au niveau de 95% signifie qu'il n'y a seulement une probabilité de 5% que la dépendance trouvée est obtenue grâce à une chance numérique entre les variables, c'est à dire qu'il n'y a pas de relation réelle entre la variable dépendante et les variables indépendantes.

Règle de décision :

Si la quantité F observée dépasse seuil, on rejette l'hypothèse H0 (au niveau α=0,05) et dans le cas contraire, on conserve H0.

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

alors H0 : $\beta_{1=\ldots\ldots} \beta_p$=0 doit être rejetée au niveau α. Cependant, la plupart des logiciels ne demandent pas de fixer a priori un niveau α pour effectuer le test : ils donnent le niveau à partir duquel notre décision aurait changé :

c'est la notion de **p-value**.

En d'autres termes, la p-valeur du test est la probabilité, si H0 était la bonne hypothèse, d'avoir observé une valeur pour F qui ait dépassé le F que nous avons observé au niveau décisionnel, on rejettera donc H0 lorsque la p-valeur est faible.

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

Remarque : Pour la régression simple, ce test porte uniquement sur le paramètre β_1 .

Ce test fournit un moyen d'apprécier la régression dans son ensemble, ce qui ne signifie pas que chacun des coefficients de la régression soit significativement différent de 0.

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

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{P} \qquad \textbf{(34)}$$

- *Test de Student*

- *Statistiques liées au paramètre β_1*

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

$\hat{\beta}_1$ est le rapport de la covariance entre X et Y divisé par la variance de X.

$$MSE = s^2 = \frac{\sum (Y_i - \hat{Y})^2}{n-p-1}$$

Remarque : pour la régression simple l'estimateur de l'écart type de $\hat{\beta}_1$ devient:

$$s(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (X_i - \overline{X})^2}} = \frac{\sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n-2}}}{\sqrt{\sum (X_i - \overline{X})^2}} \qquad \textbf{(35)}$$

On calcule alors le t-test

$$t - \text{observé} = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \quad (36)$$

Remarque :

Si on veut améliorer la précision de β_1 il faut augmenter la taille de l'échantillon (n).

La variance de β_1 est inversement proportionnelle à la dispersion des Xi autour de la moyenne. Donc, si on veut améliorer la précision de β_1 il faut, si possible, augmenter la variance empirique des Xi.

Raisonnement

On compare la p-value associée à T observé, au risque α choisi (par exemple α=0,05)

Si p-value $\leq\alpha$ alors on rejette l'hypothèse $\beta_1 = 0$ ceci dit que β_1 est significativement différent de zéro au niveau α

On peut calculer un intervalle de confiance (IC de niveau 1- α) autour de $\hat{\beta}_1$, ce qui permet de statuer sur le paramètre β_1.

$IC_{1-\alpha}(\beta_1)=[\hat{\beta}_1 - t_{1-\alpha/2}.s(\hat{\beta}_1) ; \hat{\beta}_1 + t_{1-\alpha/2}.s(\hat{\beta}_1)]$

- ***Statistiques liées au paramètre β_0***

Le t-test pour le paramètre est donnée par la formule ci-dessous

$$t\text{- observé} = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \quad (37)$$

$s(\hat{\beta}_0)$ est l'erreur type du paramètre β_0

Si p-value $\leq\alpha$ **alors on rejette l'hypothèse** $\beta_0 = 0$

Remarque :

La valeur que doit atteindre le test de Student pour l'on puisse rejeter l'hypothèse nulle dépend du nombre d'observations (n) et de niveau de confiance ou de précision recherchée (de 90% à 99% en général).

En pratique si on choisit le risque α =5% et si n est grand (n>30) pour approcher la loi de Student par la loi normale la valeur critique de t-test oscille le plus souvent autour de 2 [1, 7], alors l'intervalle de confiance de β_0 à 95% est donné par

$IC_{0,95}(\beta_0)=[\hat{\beta}_0 - 1{,}96.s(\hat{\beta}_0)\ ;\ \hat{\beta}_0 + 1{,}96.s(\hat{\beta}_0)]$

II.2. Régression linéaire multiple (MLR)

Tout comme en régression linéaire simple, la régression multiple cherche à approximer une relation trop complexe en général, par une fonction mathématique simple. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante (à expliquer) Y (ici, la toxicité) et une série de p variables indépendantes (explicatives) Xi (ici, les descripteurs). L'objectif est d'obtenir une équation de la forme suivante :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \qquad (38)$$

Cette équation est linéaire par rapport aux paramètres (coefficients de régression) β_0, β_1..., β_P.

La détermination de l'équation (38) se fait alors à partir d'une base de données de n échantillons pour laquelle à la fois les variables indépendantes et la variable dépendante sont connues. Il s'agit donc de considérer un système d'équations :

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{2,1} + \ldots + \beta_p X_{n,1} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{1,2} + \beta_2 X_{2,2} + \ldots + \beta_p X_{n,2} + \varepsilon_2 \\ &\ldots \\ Y_n &= \beta_0 + \beta_1 X_{1,p} + \beta_2 X_{2,p} + \ldots + \beta_p X_{n,p} + \varepsilon_n \end{aligned} \qquad (39)$$

Ou les résidus ε_i représentent l'erreur du modèle, constituée par l'incertitude sur la variable dépendante Yi d'une part, sur les variables indépendantes Xi d'autre part, mais aussi par les informations contenues dans les variables indépendantes mais non exprimées via les variables dépendantes.

Ce système d'équation peut être donné sous la forme matricielle suivante :

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

III.2.1.Estimation des paramètres statistiques du modèle

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

$$S(\beta_0,\dots\beta_p)=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_{1i}-\dots\beta_pX_{pi})^2$$

Les valeurs des β qui minimisent ce critère seront les solutions β_0, β_1, … β_p du système linéaire de (p+1) équations à (p+1) inconnues.

La méthode consiste alors à choisir les coefficients du vecteur b en faisant en sorte de minimiser la somme des carrées des écarts entre les valeurs prédites et les valeurs réelles sur l'intégralité de la base de données et ceci sous couvert de certaines hypothèses de départ.

En premier lieu, les variables indépendantes X_i, comme leur nom l'indique, sont supposées indépendantes entre elles et leur incertitude est négligeable. Ensuite, les différents échantillons Y_i sont supposés indépendants entre eux. Enfin, par nature, la dépendance de Y vis-à-vis des X_i est supposée linéaire. La valeur prédite de la variable dépendante Y (estimé par le modèle de régression) s'écrit :

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

Les résidus peuvent donc être définis comme la différence entre les valeurs prédites et observées de Y.

$$Y_i-\hat{Y}_i=\hat{\varepsilon}_i$$

II.2. 2.Tests sur le modèle linéaire

Comme pour le modèle linéaire simple, les hypothèses de régression linéaire doivent être vérifiées pour un modèle de régression multiple.

II.2.3.Test de la signification globale de la régression (F-Fisher)

Ce test permet de connaître l'apport global de l'ensemble des variables X_1, …..Xp à la détermination de Y.

On veut tester l'hypothèse nulle:

H0: β1 = …= βp = 0 contre Ha: il existe au moins un βj parmi β1 ,…, βp non égal à 0.

On calcule la statistique de test $F = \frac{MS\ \mathrm{mod}\ el}{MSerror}$

II. 2. 4. Test de signification de chaque paramètre (chaque descripteur) t-Student

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

A partir de cette statistique, il est possible de tester un à un la nullité des différents paramètres du modèle de régression linéaire multiple et de construire des intervalles de confiance sur ces paramètres, très utiles lors de la phase d'interprétation du modèle.

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

$t - observé = \frac{\hat{\beta}_i}{s(\hat{\beta} i)}$, avec $s(\hat{\beta} i)$ est l'erreur type du paramètre β_i

II.2. 5. Sélection de variables et choix du modèle

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

II.2.5.1 Critères de comparaison de modèle

II.2.5.1.1. Limitation du coefficient de détermination R^2

R^2 qui varie entre 0 et 1, mesure la proportion de variation totale de Y autour de la moyenne expliquée par la régression. Plus la valeur de R^2 sera proche de 1 (cas idéal) et plus les valeurs prédites et observées sont corrélées. Un R^2 faible signifie que le modèle a un faible pouvoir explicatif et les descripteurs (certains d'eux) sont sans effet sur la réponse.

Le jugement sur la valeur de R^2 est très subjectif. Bien que ce coefficient soit très facile à comprendre, il faut se garder d'y attacher trop d'importance car il est loin de fournir un critère suffisant pour juger de la qualité d'une régression. Il n'est pas recommandé d'utiliser R^2 pour comparer des modèles avec un nombre différent de descripteurs, le coefficient R^2 nous dira toujours de choisir le modèle avec le plus grand nombre de descripteurs car son R^2 sera plus important (on projette sur un espace plus grand), même si les variables sont sans effets sur la réponse Y.

La valeur de R^2 dépend de la taille de l'échantillon et le nombre de variables prédictives dans l'équation. Il garde la même valeur ou augmente lors d'une nouvelle variable de prédiction est ajoutée à l'équation de régression, même si la variable ajoutée ne contribue pas à la réduction de la variance inexpliquée. Par conséquent, un autre paramètre statistique peut être utilisé, appeléR^2ajusté (R^2ajusté).

II.2.5.1.2. Coefficient de détermination ajusté $R^2_{ajusté}$:

Il est obtenu par une expression semblable à celle deR^2, mais la SSR et SST sont divisés par leurs degrés de liberté correspondants. Ce coefficient est utilisé en régression multiple par ce qu'il tient compte du nombre de paramètres du modèle.

$$R^2_{adj} = 1 - \frac{(n - int\ ercept)(1 - R^2)}{n - p} \qquad (40)$$

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

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

II.2.5.1.3. Critère de validation croisée : PRESS

La somme des erreurs quadratiques de prédiction « *prediction sum of squares* » (PRESS) est définie par

$$PRESS = \sum_{i=1}^{n} \varepsilon_{(i)}^2 \qquad (41)$$

Ce critère permet de sélectionner les modèles ayant un bon pouvoir prédictif. (On cherche toujours le PRESS le plus petit).

Ces informations permettent d'interpréter les tables d'analyse de variance complètes données par tout logiciel mettant en œuvre la régression linéaire. La table complète est du type suivant :

Tableau. 2. Analyse de variance

Source	DF	SS	MS	F	p-value
Regression	p	SSR	MSR	MSR/MSE	p-value
Error	n-p-1	SSE	MSE		
Total	n-1				
R^2					
R^2_{adj}					

III.2.5.2. Validation d'un modèle

- **Cross validated R^2**

La procédure statistique cross-validation peut être utilisée pour évaluer le pouvoir prédictif des modèles QSAR.

Par exemple la procédure « *Leave-One-Out* » retire successivement une molécule de la série d'apprentissage contenant n molécules. Un modèle QSAR/QSPR est construit sur un ensemble *n-1* de composés et la molécule retirée est prédite par le modèle. Cette

procédure est répétée « n » fois afin de prédire les activités/propriétés de toutes les molécules.

Le coefficient qui décrit la validation est donné par l'équation ci-dessous

$$R^2_{CV} = 1 - \frac{\sum_i (Y_i^{pred} - Y_i^{obs})^2}{\sum (Y_i^{obs} - Y^{mean})^2} \quad (42)$$

Ce coefficient peut être calculé à partir de PRESS comme suit

$$R^2_{CV} = 1 - \frac{PRESS}{SStotal} \quad (43)$$

II.2.6. Colinéarité des variables explicatives

Quand les variables sont très corrélées et donc quasi-colinéaires, ce qui rend le déterminent de X'X proche de 0 : on dit que le système est mal conditionné. L'inversion de la matrice conduit alors à des estimations ayant une variance très importante, voir même parfois à des problèmes numériques. Il est donc important de diagnostiquer de tels problèmes.

Nous essayons de donner des outils de diagnostics. Les solutions sont d'autres méthodes tel que la régression « Ridge », la méthode PLS (partial least-square) et la régression sur composante principale.

La détection d'une colinéarité se fait à l'aide de :

- Matrice de corrélation : elle permet de détecter des fortes corrélations entre deux variables.
- Facteur d'inflation de la variance VIF : On définit le facteur d'inflation de la variance (VIF) par

$$VIF = \frac{1}{1 - R^2_j} \quad (44)$$

Où R_j^2 est le coefficient de détermination de la régression de la variable X_j sur les autres variables. Plus Xj est linéairement proche des autres variables, plus R_j^2 est proche de 1 et le

VIF grand. L'avantage du VIF par rapport à la matrice de corrélation est qu'il prend en compte des corrélations multiples. **[8-14].**

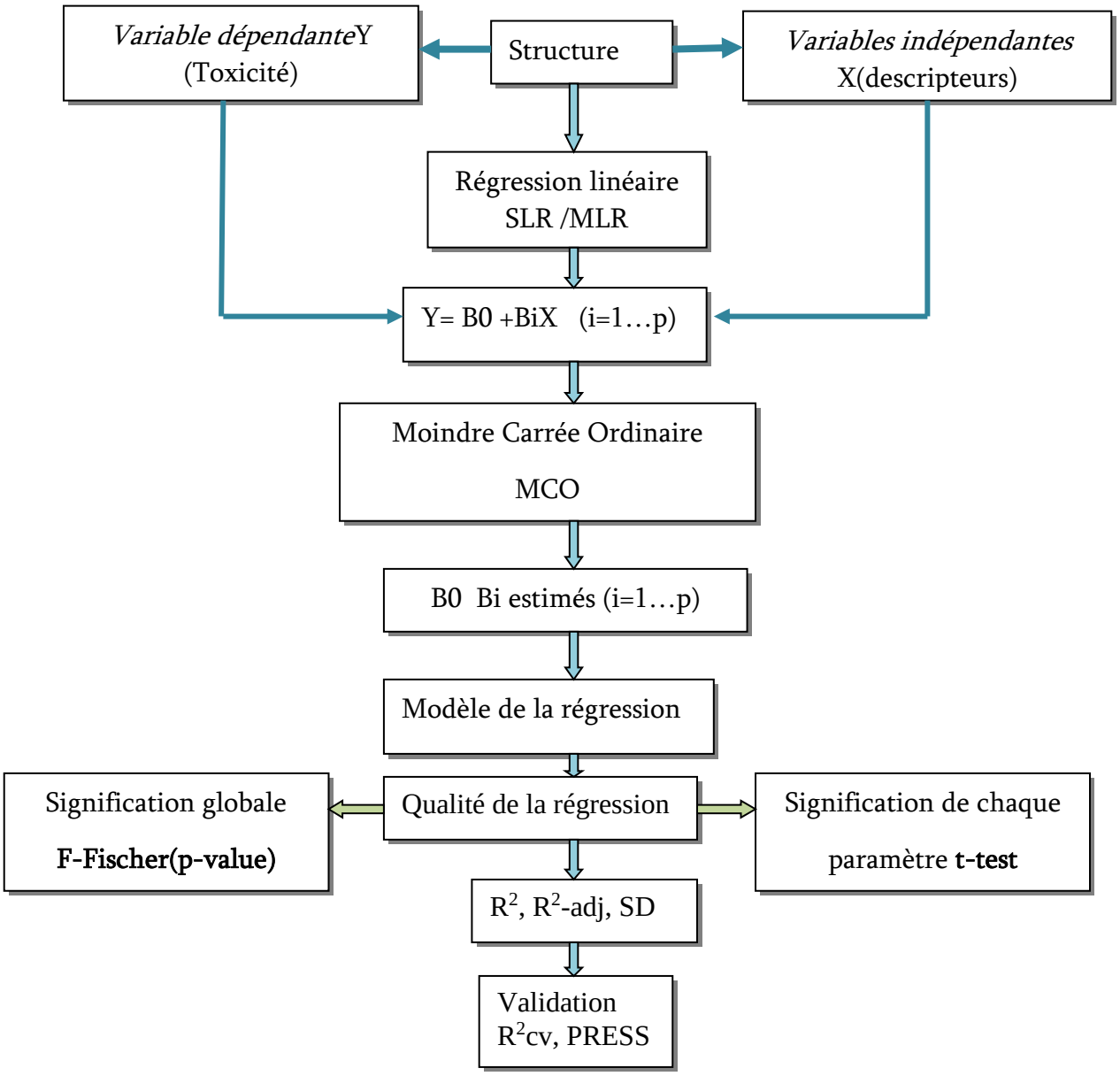


Figure. 4. *Schéma récapitulatif de la méthode de régression linéaire*

Tableau. 3. *Tableau récapitulatif des statistiques relatives à l'analyse de la variance*

Statistique	Formule	Signification
Mean square	SS/DF	Le rapport d'une somme des carrés des écarts (SS) divisée par le rapport de degrés de liberté (DF)
F value	$F = \frac{MS\ \mathrm{mod}\ el}{MSerror}$	Statistique de Fisher-Snedecor pour tester si tous les paramètres B sont nuls
s=SD	$s^2 = MSE = \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-p-1} = \frac{SSE}{n-p-1}$ Root MSE= SD= s	Standard déviation, l'écart moyen résiduel.
R^2	$R^2 = \frac{SCE_{Modèle}}{SCE_{Totale}} = \frac{SSR}{SST} = 1 - \frac{SCE_{résidu}}{SCE_{Totale}}$	F et R square (R^2) sont liés par la relation suivante
$R^2_{ajusté}$	$R^2_{ajusté} = 1 - \frac{(n - \mathrm{int}\ ercept)(1-R^2)}{n-p}$	R^2ajusté en fonction du nombre des régresseurs du modèle
R^2cv	$R^2_{CV} = 1 - \frac{\sum_i (Y_i^{pred} - Y_i^{obs})^2}{\sum_i (Y_i^{obs} - Y^{mean})^2}$	R^2 Cross-validated
Mean Value	$\overline{Y} = \sum_i^n \frac{Y_i}{n}$	Moyenne de la variable réponse Y.
t-test	t- observé = $\frac{\hat{\beta}_0}{s(\hat{\beta}_0)}$	Test de student pour tester si chaque paramètre est différent de zéro.
PRESS	$PRESS = \sum_{i=1}^n \varepsilon_{(i)}^2$	Prediction sum of squares
VIF	$V_j = \frac{1}{1-R_j^2}$	Facteur d'inflation de la variance

Références bibliographiques

[1] S. Chatterjee and A. S. Hadi, " Regression Analysis by example", fourth Edition, a John Wiley & Sons, Inc., Hoboken, New Jersey **2006**.

[2] S. Weisberg, "Applied Linear Regression", thirth Edition, John Wiley & Sons, IncHoboken, New Jersey **2005**.

[3] A. C. Rencher, G. B. Schaalje, "Linear Models in Statistics", Second Edition, John Wiley & Sons, Inc., Hoboken, New Jersey, **2008**.

[4] P. A. Cornillon, E.M.atznerLøber, "Régression théorie et Applications ", Springer Verlag France, Paris **2007**.

[5] M. Lejeune, "Statistiques : la théorie et ses applications", Springer Verlag, Paris 2004.

[6] A.F. Siegel "Practical Business Statistics", IRWIN, **1997** (3rd edition)

[7] R. D. Cook, S .Weisberg, "An introduction to Regression Graphics", Wiley Series in Probability and Statistics, **1994**.

[8] J. JACQUES,"Modélisation Statistique" http ://labomath.univ-lille1.fr/~jacques

[9] P. Besse. "Pratique de la modélisation statistique , Publications du laboratoire de statistique et Probabilités", **2003**. Disponible sut http:// www.math.univ-toulouse.fr

[10] P. Besse. "Apprentissage Statistique Data mining , Publications du laboratoire de statistique et Probabilités", **2009**.

[11] G.J.Mclachlan, "Discriminant analysis and Statistical Pattern Recognition. Wiley, New-York, 1992.

[12] J.P.Nakache et J.Confais, "Statistique explicative appliquée", Edition Technip **2003.**

[13] G.Saporta, "Probabilité, Analyse de données et statistique", 2ème Edition, Edition Technip, **2006**.

[14] D.Laffly , "Régression multiple : principes et exemples d'application", **2006.**

CHAPITRE III

METHODES DE LA CHIMIE QUANTIQUE

L'élaboration des modèles QSPR/QSAR repose sur le calcul des structures (descripteurs), ce dernier est assuré en faisant appel aux outils de la modélisation moléculaire.

Différentes approches sont envisageables dans le cadre des outils de modélisation moléculaire. Les méthodes quantiques en l'occurrence les méthodes *ab initio*, la théorie de la fonctionnelle de la densité et les méthodes semi-empiriques sont capables de calculer plusieurs propriétés des systèmes. C'est pour cette raison que ces approches ont été employées dans le cadre de cette étude. Dans cette partie, il s'agit d'expliciter les méthodes de chimie quantique, utilisées non seulement pour le calcul des structures moléculaires (descripteurs) nécessaires à la mise en place de modèles prédictifs mais aussi pour l'explication des mécanismes de toxicité des séries de composés étudiées.

Introduction :

L'état d'un système à N noyaux et n électrons est d'écrit en mécanique quantique par une fonction d'onde ψ satisfaisant l'équation de Schrödinger [1].

$$H\,\Psi = E\,\Psi \qquad (1)$$

ψ : sont les fonctions propres de H

E : sont les valeurs propres de H

L'hamiltonien H total d'une molécule comportant N noyaux et n électrons, est défini par la somme de cinq termes (terme cinétique des électrons, terme cinétique des noyaux, terme de répulsions électrons-électrons, terme de répulsions noyaux-noyaux et terme d'attractions electrons-noyaux).

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

Born et Oppenheimer [**2**] ont proposé l'approximation des noyaux fixes qui consiste à séparer l'hamiltionien électronique de l'hamiltonien nucléaire. Dans le cadre de cette approximation (et en se plaçant dans le cadre non relativiste), l'hamiltonien H peut se réduire à la forme suivante :

$$H = -\frac{\hbar^2}{2m_e}\sum_{i=1}^{n}\Delta_i - \sum_{K=1}^{N}\sum_{i=1}^{n}\frac{Z_K e^2}{R_{Ki}} + \sum_{i\rangle j}^{n}\frac{e^2}{r_{ij}} \qquad (3)$$

La résolution exacte de l'équation (1) n'est possible que pour l'atome d'hydrogène et les systèmes hydrogénoïdes. Pour les systèmes poly-électroniques, il est nécessaire de faire appel aux méthodes d'approximation pour résoudre l'équation de Schrödinger d'une manière approchée.

Les propriétés moléculaires qui peuvent être calculées par la résolution de l'équation de Schrödinger sont multiples. On peut citer entre autres :

- Structures et énergies moléculaires
- Energies et structures des états de transition
- Fréquences de vibration
- Spectres IR et Raman
- Propriétés thermochimiques
- Energies de liaison
- Chemins réactionnels
- Orbitales moléculaires
- Charges atomiques
- Moments multipolaires
- Déplacements chimiques RMN et susceptibilités magnétiques
- Affinités électroniques et potentiels d'ionisation
- Polarisabilités et hyperpolarisabilités
- Potentiels électrostatiques et densités électroniques
- etc.

III.1. Méthode de Hartree-Fock-Roothaan

III.1.1. Approximation du champ moyen de Hartree

L'approximation du champ moyen, proposée par Hartree [**3**] en 1927, consiste à remplacer l'interaction d'un électron avec les autres électrons par l'interaction de celui-ci avec un champ moyen créé par la totalité des autres électrons ; ce qui permet de remplacer le potentiel biélectronique $\sum_j e^2 / r_{ij}$ qui exprime la répulsion instantanée entre l'électron i et les autres électrons j≠i par un potentiel monoélectronique moyen de l'électron i de la forme U(i). Par conséquent et en se basant sur le théorème des électrons indépendants, nous pouvons écrire la fonction d'onde totale comme le produit de fonctions d'onde mono-électroniques:

$$\Psi = \Psi_1(1) \ . \ \Psi_2(2) . \Psi_2(2) ... \Psi_n(n) \tag{4}$$

III.1.2. Méthode de Hartree-Fock

La fonction d'onde polyélectronique de Hartree (Eq. 4) ne vérifie ni le principe d'indiscernabilité des électrons ni le principe d'exclusion de Pauli. Pour tenir-compte de ces deux principes, Fock [**4**] a proposé d'écrire la fonction d'onde totale Ψ sous forme d'un déterminant, appelée déterminant de Slater [**5**], dont la forme abrégée pour un système à couches fermées est:

$$\Psi(1,2,...,n) = \frac{1}{(n!)^{1/2}} \left| \Phi_1(1) \overline{\Phi_1}(2) ... \Phi_m(2m-1) \overline{\Phi}_m(2m) \right| \tag{5}$$

avec :

$$\Phi_1(1) = \Phi_1(1)\alpha(1) \tag{6}$$

$$\overline{\Phi}_1(2) = \Phi_1(2)\beta(2) \tag{7}$$

Φ est une orbitale moléculaire monoélectronique. α et β et sont les fonctions de spin.

III.1.3. Méthode de Hartree-Fock-Roothaan

Les expressions analytiques des orbitales moléculaires Φ_i n'ont pas été définies dans le cadre de la méthode de Hartree-Fock. C'est Roothaan **[6]** qui a utilisé la technique OM-CLOA pour construire les OM. Cette méthode consiste à exprimer l'orbitale moléculaire Φ_i par une combinaison linéaire d'orbitales atomiques φ_μ :

$$\Phi_i = \sum_{\mu=1}^{N} C_{i\mu}\, \phi_\mu \qquad (8)$$

$C_{i\mu}$ sont les coefficients à faire varier. N étant le nombre d'OA combinées.

Les meilleurs coefficients sont ceux qui minimisent l'énergie. En procédant par la méthode des variations et après certaines manipulations algébriques, on aboutit aux équations de Roothaan définies par le système séculaire suivant **[6]**:

$$\sum_{r=1}^{N} C_{kr}\left(F_{rs} - \varepsilon_k\, S_{rs}\right) = 0 \qquad s = 1,2,\ldots N \qquad (9)$$

avec:

$$\begin{cases} F_{rs} = h^c_{rs} + \sum_{p=1}^{n}\sum_{q=1}^{n} P_{pq}\left\{2\,\langle rs \mid pq\rangle - \langle rq \mid ps\rangle\right\} \\ S_{rs} = \langle \phi_r \mid \phi_s \rangle \\ h^c_{rs} = \int \phi_r^*(i)\, h^c\, \phi_s(i)\, d\tau_i \end{cases} \qquad (10)$$

Où r, s, p et q symbolisent les OA. P_{pq} est l'élément de la matrice densité. Les termes $\langle rs \mid pq\rangle$ et $\langle rq \mid ps\rangle$ représentent les intégrales biélectroniques coulombiènne et d'échange respectivement. Srs est une intégrale de recouvrement.

III.2. Méthodes Post-SCF

La méthode Hartree-Fock-Roothaan présente l'inconvénient majeur de ne pas tenir compte de la **corrélation électronique** qui existe entre le mouvement des électrons. Ceci

rend cette méthode relativement restreinte dans le calcul quantitative des propriétés thermodynamiques telles que l'enthalpie d'activation, l'énergie de Gibbs de réactions, énergies de dissociation,…

Ces propriétés peuvent être calculées d'une manière efficace par les méthodes Post-SCF en tenant-compte de la corrélation électronique . Les deux familles importantes de méthodes qui ont été développées sont celles d'interaction de configurations (CI) et la théorie des perturbations Moller-Plesset d'orde n (MPn) et les méthodes DFT.

III.2.1. Méthode d'interaction de configuration (CI)

La méthode CI [**7,8**], utilise une combinaison linéaire de déterminants de Slater pour décrire l'état fondamental. Cette combinaison représente les différentes excitations de un ou plusieurs électrons des orbitales moléculaires occupées vers les orbitales moléculaires vides

$$\Psi = \sum_{k}^{A} C_k \, \Phi_k \qquad (11)$$

Où les déterminants Φ_k , k = 1, 2, 3, …, décrivent respectivement l'état fonda mental et les états mono, bi et triexcités, …, etc. A est le nombre de configurations prises en considération. Pour obtenir un résultat satisfaisant, il est nécessaire d'avoir une combinaison très étendue des déterminants. Une valeur exacte de l'énergie demandera, à priori, une infinité de déterminants.

Remarque :

L'état correspondant à k = 0 ou état fondamental dans les méthodes CI, représente en fait le niveau HF. L'énergie du système et les coefficients sont obtenus par la méthode variationnelle.

$$\sum_{k}^{A} C_k \left(H_{kl} - e S_{kl} \right) = 0 \qquad (12)$$

III.2.2. Méthode de Möller-Plesset d'ordre 2 (MP2)

Cette approche, proposée par Moller-Plesset [**9**], tient compte de la corrélation électronique en utilisant la théorie des perturbations. L'hamiltonien polyélectronique s'écrit :

$$H = H^0 + \lambda V \qquad (13)$$

H^0, représente l'hamiltonien d'ordre zéro, pris comme une somme d'opérateurs monoélectroniques de Fock :

$$H^0 = \sum_i F(i) = \sum_i \left\{ h^c(i) + \sum_j [J_j(i) - K_j(i)] \right\} \qquad (14)$$

λV est la perturbation (λ est un paramètre qui varie entre 0 et 1) définie par :

$$\lambda V = \sum_i \sum_{j \rangle i} \frac{1}{r_{ij}} - \sum_i \sum_j [J_j(i) - K_j(i)] \qquad (15)$$

La fonction d'onde et l'énergie du n^{éme} état du système ont la forme :

$$E_n = E_n^0 + \lambda E_n^1 + \lambda^2 E_n^2 + \ldots \qquad (16)$$

$$\Psi_n = \Psi_n^0 + \lambda \Psi_n^1 + \lambda^2 \Psi_n^2 + \ldots \qquad (17)$$

Ou E_0^1, E_0^2 et E_0^3 sont respectivement les corrections énergétiques au premier, second et troisième ordre. La correction énergétique d'ordre n s'obtient en appliquant la méthode des perturbations de Rayleigh-Schrödinger. Celle d'ordre 1 et qui correspond à l'énergie Hartree-Fock est donnée par :

$$E_{HF} = E_0^0 + E_0^1 \qquad (18)$$

L'énergie de corrélation est donnée par la somme des corrections énergétiques d'ordre supérieur à un. Celle du deuxième ordre est définie par [10] :

$$E_n^2 = \frac{1}{4} \sum_r \sum_s \sum_t \sum_u \frac{\langle rs / tu \rangle^2}{e_r + e_s + e_t + e_u}$$

$$E_{MP2} = E_n^{HF} + E_n^2 \qquad (19)$$

$$\Rightarrow E_{MP2} = \sum_k e_k - \frac{1}{2} \sum_i \sum_j \left(J_{ij} - K_{ij} \right) + \frac{1}{4} \sum_r \sum_s \sum_t \sum_u \frac{\langle rs / tu \rangle^2}{e_r + e_s + e_t + e_u}$$

Où :

$$\langle rs / tu \rangle = \iint \phi_r(1) \quad \phi_s(1) \quad \frac{1}{r_{12}} \quad \phi_t(1) \quad \phi_u(2) \; d\tau_1 \; d\tau_2$$
$$- \iint \phi_r(1) \quad \phi_s(2) \quad \frac{1}{r_{12}} \quad \phi_u(1) \quad \phi_t(2) \; d\tau_1 \; d\tau_2 \qquad (20)$$

De la même manière, on obtient les autres ordres de perturbation.

La fonction d'onde électronique de n électrons dépend de 3n coordonnées d'espace et de n coordonnées de spin. L'opérateur hamiltonien est constitué de termes mono et biélectroniques, par conséquent, l'énergie moléculaire est développée en terme d'intégrales introduisant 6 coordonnées d'espace. En d'autres termes, la fonction d'onde électronique devient de plus en plus complexe avec l'augmentation du nombre d'électrons, ceci a inspiré la recherche de fonctions qui mettent en jeu moins de variables que la fonction d'onde.

Remarque :

Dans les méthodes décrites précédemment (HF, CI et MP2), un système à *n* électrons est décrit par une fonction d'onde qui dépend de 4n variables (3n variables d'espace et n variables de spin). De plus, ces méthodes sont très coûteuses en temps de calcul et en mémoire CPU, en particulier pour des systèmes de grandes tailles. L'idée fondamentale de la théorie de la fonctionnelle de la densité (DFT) est de réduire le

nombre de variables en remplaçant la fonction d'onde par une fonctionnelle qui est 'la densité électronique' $\rho(x,y,z)$ qui ne dépend de 3 variables seulement.

III.3. Théorie de la fonctionnelle de densité (DFT)

III.3.1. Fondement de la théorie DFT :

Historiquement, les premiers à avoir exprimé l'énergie en fonction de la densité furent Thomas (1927), Fermi (1927, 1928) et Dirac (1930) sur le modèle du gaz uniforme d'électrons non interagissants. Le but des méthodes DFT est de déterminer des fonctionnelles qui permettent de relier la densité électronique à l'énergie **[11].** Cependant, la DFT a véritablement débuté avec les théorèmes fondamentaux de Hohenberg et Kohn en 1964 **[12]** qui établissent une relation fonctionnelle entre l'énergie de l'état fondamental et sa densité électronique.

- **1er *théorème* de Hohenberg et Kohn :**

Enoncé : « *L'énergie moléculaire, la fonction d'onde et toutes les autres propriétés électroniques de l'état fondamental sont déterminées à partir de la densité électroniques de l'état fondamental* $\rho_0(x,y,z)$ ». [12]

Rappelons l'expression de l'Hamiltonien électronique d'un système polyélectronique :

$$H = -\frac{1}{2}\sum_{i}^{n} \Delta_i + \sum_{i\rangle\, j}^{n} \frac{1}{r_{ij}} + \sum_{i}^{n} \nu(r_i) \qquad (21)$$

avec $$v(r_i) = -\sum_{\alpha} \frac{Z_\alpha}{r_{i\alpha}} \qquad (22)$$

$\nu(r_i)$: potentiel externe de l'électron i :

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

$\rho_0(r)$: exprime la densité électronique au point r (nombre d'électrons). En intégrant cette densité ponctuelle sur toute l'espace, on obtient le nombre total d'électrons :

$$\int \rho_0(r)\, dr = n \tag{23}$$

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

$$E_0[\rho_0] = V_{ne}[\rho_0] + T[\rho_0] + V_{ee}[\rho_0] \tag{24}$$

avec

$$V_{ne}[\rho_0] = \int \rho_0(r) v(r) dr \tag{25}$$

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

$$E_0[\rho] = \int \rho_0(r) v(r) dr + F[\rho_0] \tag{26}$$

avec
$$F[\rho_0] = T[\rho_0] + V_{ee}[\rho_0] \tag{27}$$

La fonctionnelle $F[\rho_0]$ est inconnue

- **2ème *théorème* de Hohenberg et Kohn :**

Enoncé : « *Pour une densité d'essai* $\tilde{\rho}(r)$, *telle que* $\tilde{\rho}(r) \geq 0$ et $\int \tilde{\rho}(r)\, dr = n$, l'inégalité suivante est vérifiée :

$$E_0 \leq E[\tilde{\rho}] \tag{28}$$

Ce théorème est l'équivalent du principe variationnel.

III.3.2 Méthode de Kohn et Sham :

Le théorèmes de Hohenberg et Kohn ne donnent pas une procédure pour calculer l'énergie E_0 à partir de ρ_0, ni comment déterminer ρ_0 sans déterminer, au préalable, la fonction d'onde. C'est Khon et Sham, en 1965, qui ont élaboré une méthode pratique pour trouver E_0 a partir de ρ_0 **[13].** Ils ont considéré un système fictif de référence, noté s, constitué par les n électrons non interagissants.

Le système de référence est choisi de telle façon à avoir :

$$\rho_s(r) = \rho_0(r) \tag{29}$$

Etant donné que les électrons n'interagissent pas entre eux dans le système de référence, l''hamiltonien de système de référence s'écrit

$$\hat{H}_s = \sum_{i=1}^{n}\left[-1/2\nabla_i^2 + v_s(r_i)\right] = \sum_{i=1}^{n} h_i^{KS} \tag{30}$$

avec

$$h_i^{KS} = -1/2\nabla_i^2 + v_s(r_i) \tag{31}$$

Par conséquent, les équations de Kohn et Sham, pour l'électron i, peuvent s'écrire comme suit :

$$h_i^{KS}\theta_i^{KS} = \varepsilon_i^{KS}\theta_i^{KS} \tag{32}$$

θ_i^{KS} : Orbitale de Kohn et Sham de l'électron i.

- **Terme d'échange-corrélation**

Soit ΔT la différence de l'énergie cinétique entre le système réel (électrons interagissants) et le système fictif (électrons non-interagissants)

$$\Delta T = T[\rho] - T_s[\rho] \tag{33}$$

donc

$$\Delta V = V_{ee}[\rho] - 1/2 \iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 \tag{34}$$

Δv est la différence entre la vrais répulsion électron-électron et la répulsion colombienne entre deux distributions de charge ponctuelle. L'énergie s'écrit alors :

$$E_v[\rho] = \int \rho(r) v(r) dr + T_s[\rho] + 1/2 \iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + \Delta T[\rho] + \Delta V_{ee}[\rho] \tag{35}$$

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

$$E_{xc}[\rho] = \Delta T[\rho] + \Delta V_{ee}[\rho] \tag{36}$$

$$E_v[\rho] = \int \rho(r) v(r) dr + T_s[\rho] + 1/2 \iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + E_{xc}[\rho] \tag{37}$$

Le problème majeur pour les calcules DFT, selon le schéma de Kohn et Sham, est de trouver une bonne approximation pour l'énergie échange- corrélation E_{xc}.

Les orbitales de KS permettent de calculer la densité électronique ρ_0 à l'aide de la formule suivante :

$$\rho_0 = \rho_s = \sum_{i=1}^{n} \left| \theta_i^{KS} \right|^2 \tag{38}$$

Les orbitales de KS permettent également de calculer le cinétique du système de référence T_S. De cette manière, l'énergie E_0 peut s'écrire :

$$E_0 = -\sum_{\alpha} Z_{\alpha} \int \frac{\rho(r_1)}{r_{1\alpha}} dr_1 - 1/2 \sum_{i=1}^{n} \left\langle \theta_i^{KS}(1) \middle| \nabla_1^2 \middle| \theta_i^{KS}(1) \right\rangle + 1/2 \iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} d\, r_1 dr_2 + E_{xc}[\rho] \quad (39)$$

L'équation aux valeurs propres correspondante est de la forme :

$$\left[-1/2\nabla_1^2 - \sum_{\alpha} \frac{Z_{\alpha}}{r_{1\alpha}} + \int \frac{\rho(r_2)}{r_{12}} dr_2 + v_{xc}(1) \right] \theta_i^{KS}(1) = \varepsilon_i^{KS} \theta_i^{KS}(1) \quad (40)$$

Le potentiel d'échange-correlation V_{xc} est défini comme la dérivée de l'énergie échange-correlation E_{xc} par rapport à la densité électronique :

$$v_{xc}(r) = \frac{\partial E_{xc}[\rho(r)]}{\partial \rho(r)} \quad (41)$$

Il existe plusieurs approximations de ce potentiel d'écange-corrélation.

III.3.3 Approximation de la densité locale LDA :

Hohenberg et Khon ont montré que si ρ varie extrêmement lentement avec la position, l'énergie d'échange-corrélation $E_{xc}[\rho_s]$ peut s'écrie comme suit :

$$E_{xc}^{LDA}[\rho] = \int \rho(r) \varepsilon_{xc}(\rho) dr \quad (42)$$

ε_{xc} : étant l'énergie d'échange-corrélation par électron. Cette quantité est exprimée comme la somme des deux contributions :

$$\varepsilon_{xc}(\rho) = \varepsilon_x(\rho) + \varepsilon_c(\rho) \quad (43)$$

avec

$$\varepsilon_x(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}(\rho(r))^{1/3} \quad (44)$$

Donc

$$E_x^{LDA} = \int \rho\varepsilon_x dr = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int[\rho(r)]^{4/3} dr \quad (45)$$

Le terme de corrélation $\varepsilon_c(\rho)$ est exprimé par la formule de Vosko, Wilk, et Nusair (VWN) [14]. Cette formule assez compliquée est donnée dans la référence [15, page 183].

- **Fonctionelles E_x et E_c**

La fonctionnelle E_{xc} peut s'écrire comme la somme de deux fonctionnelles d'échange E_x et de corrélation Ec :

$$E_{xc} = E_x + E_c \quad (46)$$

E_x est définit par la même formule utilisée pour l'énergie d'échange dans le cadre de la méthode de Hrtree-Fock en remplaçant les orbitales de Hartree Fock par les orbitales de Kohn et Sham.

$$E_x = -\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\langle \theta_i^{KS}(1)\theta_j^{KS}(2)\left|1/r_{12}\right|\theta_j^{KS}(1)\theta_i^{KS}(2)\right\rangle \quad (47)$$

L'énergie de corrélation est calculée comme la différence entre E_{xc} et E_c.

$$E_c = E_{xc} - E_x \quad (48)$$

III.3.4 Méthode Xα :

Dans cette méthode, développée pas Slater en 1951 [16], E_{xc} est exprimée par la seule contribution de l'échange. Cette méthode néglige donc la contribution de la corrélation.

$$E_{xc} \approx E_x^{X\alpha} = -\frac{9}{8}\left(\frac{3}{\pi}\right)^{1/3} \alpha \int [\rho(r)]^{4/3} dr \qquad (49)$$

α est un paramètre ajustable, compris entre 2/3 et 1.

III.3.5 Approximation de la densité de spin locale LSDA :

Pour les molécules à couches ouvertes et les géométries des molécules prés de leur état de dissociation, l'approximation LSDA donne des résultats meilleurs que l'approximation LDA. Dans LDA, les électrons ayant des spins opposés ont les mêmes orbitales KS spatiales. En revanche, LSDA distingue entre les orbitales des électrons de spins opposés ($\theta_{i\alpha}^{KS}$ pour les e- de spin α et $\theta_{i\beta}^{KS}$ pour les e- de spin ß. Par conséquent, on aura :

$$E_{xc} = E_{xc}\left[\rho^{\alpha}, \rho^{\beta}\right] \qquad (50)$$

C'est l'équivalent de la méthode UHF (Unrestricted Hartree-Fock) pour les chaînes ouvertes.

III.3.6. Approximation du Gradient Généralisé (GGA) :

Les approximations LDA et LSDA sont basées sur le modèle du gaz électronique uniforme dans lequel la densité électronique ρ varie très lentement avec la position. La correction de cette approximation, plus au moins grossière, nécessite l'inclusion des gradients des densités des spin ρ^{α} et ρ^{β}. L'énergie d'échange-corrélation, dans le cadre de l'approximation du gradient généralisé GGA (Generalized –gradient approximation), s'écrit alors:

$$E_{xc}^{GGA}\left[\rho^{\alpha},\rho^{\beta}\right]=\int f\left(\rho^{\alpha}(r),\rho^{\beta}(r),\nabla\rho^{\alpha}(r),\nabla\rho^{\beta}(r)\right)dr \qquad (51)$$

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

E_{xc}^{GGA} est divisé en deux contributions : échange et corrélation

$$E_{xc}^{GGA}=E_{x}^{GGA}+E_{c}^{GGA} \qquad (52)$$

Terme d'échange :

En 1988, Becke [17] a utilisé le terme d'échange pour apporter une correction de l'approximation LSDA :

$$E_{x}^{B88}=E_{x}^{LSDA}-b\sum_{\sigma=\alpha,\beta}\int\frac{(\rho^{\sigma})^{4/3}\chi_{\sigma}^{2}}{1+6b\chi_{\sigma}\sinh^{-1}\chi_{\sigma}}dr \qquad (53)$$

avec

$$\chi_{\sigma}=\left|\nabla\rho^{\sigma}\right|/(\rho^{\sigma})^{4/3}$$
$$\sinh^{-1}x=\ln\left[x+(x^{2}+1)^{1/2}\right] \qquad (54)$$

et
$$E_{x}^{LSDA}=-\frac{3}{4}\left(\frac{6}{\pi}\right)^{1/3}\int\left[(\rho^{\alpha})^{4/3}+(\rho^{\beta})^{4/3}\right]dr \qquad (55)$$

Terme de corrélation:

La fonctionnelle de l'énergie de corrélation $E_c[\rho]$, corrigé à l'aide de l'approximation GGA, est exprimée à l'aide de la formule de Lee-Yang-Parr [18] :

$$E_{c}^{GGA}=E_{c}^{LYP} \qquad (56)$$

Cette formule assez compliquée est donnée également dans Réf. [15, page 185].

Remarque :

Il est fort intéressent de noter que n'importe quelle fonctionnelle d'échange (ou de corrélation) d'une méthode peut être combinée avec n'importe quelle fonctionnelle d'échange (ou de corrélation) d'une autre méthode. Par exemple, la notation BLYP/6-31G* indique qu'il s'agit d'un calcul DFT avec la fonctionnelle d'échange de Becke 1988 et la fonctionnelle de corrélation de Lee-Yang-Parr, avec les orbitales de Kohn et Sham (KS) développées sur base gaussienne de type 6-31G*.

III.3.7 Fonctionnelle hybride B3LYP :

La fonctionnelle hybride B3LYP (Becke 3-parameters Lee-Yang-Parr) consiste à une hybridation (mélange) de plusieurs fonctionnelles de différentes méthodes comme le montre l'expression suivante :

$$E_{xc}^{B3LYP} = (1 - a_0 - a_x) E_x^{LSDA} + a_0 E_x^{exact} + a_x E_x^{B88} + (1 - a_c) E_c^{VWN} + a_c E_c^{LYP} \qquad (57)$$

E_x^{exact} est donnée par l'équation (47).

Les valeurs des 3 paramètres d'ajustement sont **[19]**:

$a_0 = 0.20$

$a_x = 0.72$

$a_c = 0.81$

III.3.8. Processus SCF de résolution des équations de Kohn et Sham :

Etape 1 : La densité initiale est prise usuellement comme la superposition de densités électronique des atomes individuels pour une géométrie bien choisie. Cette densité initiale permet d'obtenir le terme d'échange-corrélation et résoudre les équations de Kohn et

Sham (eq. 40). On note que les orbitales moléculaires de Kohn et Sham θ_i^{KS} sont généralement exprimées à l'aide d'orbitales atomiques χ_r:

$$\theta_i^{KS} = \sum_{r=1}^{b} C_{ri}\ \chi_r \qquad (58)$$

En procédant par la méthode de variation, on obtient un système séculaire qui ressemble à celui de Roothaan.

$$\sum_{s=1}^{b} C_{si}\left(h_{rs}^{KS} - \varepsilon_i^{KS} S_{rs}\right) = 0, \qquad r = 1, 2, ..., b \qquad (59)$$

Etape 2 : Les orbitales KS obtenues dans l'étape 1 sont utilisées pour calculer la nouvelle densité ρ donnée par la formule (38).

Les itérations (étapes 1 et 2) seront répétées jusqu'à atteindre la convergence, c'est-à-dire jusqu'à l'obtention d'un champ auto-cohérent (Self-Consistent Field).

En conclusion, on peut dire que le succès des méthodes de la DFT se justifie par le fait que ces méthodes permettent souvent d'obtenir, à plus faible coût, des résultats d'une précision comparable à celle obtenue avec des calculs post-Hartee–Fock comme CI ou MP2. D'autre part, les méthodes DFT combinées avec des méthodes de niveaux inférieurs commencent à être utilisées pour des systèmes de grandes tailles et pour les molécules biologiques. C'est le cas de la méthode ONIOM [20-25]. Par exemple, dans un calcul de type ONIOM(B3LYP/6-31G(d,p):AM1:AMBER), trois méthodes AMBER, AM1 et B3LYP sont combinées lors du traitement de la molécule.

III. 4. Comparaison des temps de calcul des différentes méthodes :

Dans le tableau ci-dessous, nous avons présenté une comparaison des temps de calcul relatifs pour 3 systèmes :

- Camphre $C_{10}H_{16}O$
- Morphine $C_{17}H_{19}NO_2$
- Triacetyldiminicine $C_{36}H_{25}NO_{12}$.

Les calculs ont été effectués avec différentes méthodes et différentes bases d'OA en utilisant le programme SPARTAN 04 [26].

Tabeau.1. Comparaison des temps de calcul des différentes méthodes

Méthode/base d'OA	Camphre		Morphine		Triacetyldiminicine	
	Single point	Optimisation de la géométrie	Single point	Optimisation de la géométrie	Single point	Optimisation de la géométrie
HF/3-21G	**1***	5	**1***	12	**1***	40
HF/6-31G*	7	30	8	110	7	360
HF/6-311+G**	42	180	-	-	-	-
B3LYP/6-31G*	13	65	12	160	10	400
B3LYP/6-311+G**	85	370	76	-	-	-
MP2/6-31G*	27	270	80	2000	320	-
MP2/6-311+G**	260	-	650	-	-	-

* ***Le calcul single point HF/3-21G est pris comme référence pour les calculs de niveaux supérieurs.***

III.5 Comparaison des performances des différentes méthodes de calcul:

Afin de mettre en évidence la performance des méthodes de chimie quantique (semi-empérique, Hartree-Fock, DFT, MP2) et la mécanique moléculaire, nous présentons une comparaison entre ces différentes méthodes en se référant à « *A Guide to Molecular Mechanics and Quantum Chemical Comparisons* » du manuel du programme Spartan PC Pro. **[26].**

Tableau. 2. Comparaison des performances des différentes méthodes de calcul

méthodes Types de calculs	Mécanique moléculaire	semi-empirique	Hartree-Fock	DFT	MP2
Géométries des composés organiques	Bonne avec précautions	Bonne	Bonne	Bonne	Bonne
Géométries des métaux	-	Bonne	Mauvaise	Bonne	Mauvaise
Géométries de l'état de transition	-	Bonne avec précautions	Bonne	Bonne	Bonne
conformation	Bonne	Mauvaise	Bonne	Bonne	Bonne
thermochimie (générale)	-	Mauvaise	Bonne	Bonne	Bonne
thermochimie (isodesmique)	-	Mauvaise	Bonne	Mauvaise	Bonne
Coût	Faible ----------> Elevé				

III.6. Bases d'orbitales atomiques :

Les éléments de la matrice de Fock sont des fonctions de variables C_{kr}.. C'est pourquoi la solution des équations de Roothan (Eqs.(9-10)) implique une procédure itérative pour laquelle il faut définir les coefficients C_{kr} de départ.

La première étape qui précède le déclenchement de ce processus consiste à calculer toutes les intégrales moléculaires mono et biélectroniques sur une base d'orbitales atomiques (OA). Il y a deux sortes de fonctions de base qui sont d'un usage courant. Le premier type de bases sont les orbitales de type Slater STO **[27]** qui sont Les meilleures OA analytiques définies par:

$$\Psi_{nlm} = N_n \, r^{n^*-1} \exp\left(-\zeta r\right) Y_{lm}\left(\theta, \phi\right) \qquad (60)$$

où N_n est le facteur de normalisation et ζ est l'exponentielle orbitale (exposant de Slater, déterminant la taille de l'orbitale.), $Y_{lm}(\theta,\varphi)$ sont les harmoniques sphériques.

Les fonctions de types Slater (STOs) présentent une forme analytique simple mais elles ne sont pas utilisées à grande échelle dans les programmes moléculaires *ab initio*. Cela est dû à la complexité du calcul d'intégrales moléculaires sur la base STO.

Les programmes *ab initio* de chimie quantique (Gaussian par exemple), utilisent les le second type de bases, fonctions gaussiennes (GTOs) proposées par Boys [**28**].

$$g(\alpha,\vec{r}) = c\ x^n\ y^l\ z^m \exp\left(-\alpha r^2\right) \qquad (61)$$

Dans cette équation, α est une constante déterminant la taille de la fonction. La somme (n+l+m) définie le type de l'orbitale atomique.

n+l+m= 0 (OA de type s)

n+l+m= 1 (OA de type p)

n+l+m= 2 (OA de type d)

Les fonctions gaussiennes sont largement utilisées dans les calculs *ab initio* [**29**]**.** Cela peut être justifié par le fait que « *Le produit de deux gaussiennes centrées en deux points A et B est équivalent à une gaussienne centrée au point C* ». Cette propriété mathématique permet de faciliter considérablement le calcul d'intégrales moléculaires multicentriques.

En pratique les orbitales atomiques OA de Slater (STO) sont approchées par une combinaison de plusieurs OA gaussiennes (GTO).

La plus simple est la base STO-3G encore appelée base minimale. Ceci signifie que les orbitales de type Slater sont représentées par trois fonctions gaussiennes. Dans la base minimale **STO-3G**, on utilise 3 gaussiennes pour approcher chacune des orbitales de type Slater.

Si cette base donne une assez bonne description de la densité électronique aux distances éloignées du noyau ($r \to \infty$), la description du comportement de la fonction d'onde exacte au voisinage du noyau ($r \to 0$) est assez mauvaise. Pour cette raison, plusieurs bases gaussiennes étendues ont été élaborées. Ces dernières diffèrent par le nombre des fonctions contractées et les coefficients de contraction. On appelle une fonction gaussienne contractée (**CGTO**) une combinaison linéaire de gaussiennes primitives (PGTOs) :

$$G^{CGTO} = \sum_{\lambda=1}^{k} d_\lambda \, g_\lambda^{PGTO} \qquad (62)$$

d_λ étant le coefficient de contraction de la gaussienne primitive g_λ . k est le degré de contraction.

La base **3-21G** est une *Split Valence-Double Zeta* (SV-DZ), où chaque orbitale atomique des couches internes est décrite par une contraction de 3 gaussiennes primitives. Les orbitales de la couche de valence sont réparties en deux groupes : les orbitales proches du noyau sont décrites par une contraction de 2 primitives, et les orbitales éloignées par une seule gaussienne primitive.

La base **6-311G** est une *Split Valence-Triple Zeta* (SV-TZ) dans laquelle les orbitales de cœur (couches internes) sont exprimées par une contraction de 6 gaussiennes primitives. Les orbitales de la split couche de valence sont exprimées par des contractions de 3, 1 et 1 primitives respectivement.

L'utilisation des bases de fonctions provenant d'un calcul atomique dans le traitement des molécules reste insatisfaisante, même si les exposants sont réoptimisés. En effet, il faut tenir compte du fait que dans la molécule, les atomes subissent une déformation du nuage électronique, et des distorsions dues à l'environnement. Ce phénomène peut être pris en compte par l'introduction de fonctions supplémentaires dans

la base atomique, dites de **polarisation**. L'ajout de ces fonctions est très utile dans le but d'avoir une bonne description des grandeurs telles que l'énergie de dissociation, les moments dipolaires et multipolaires,…etc. Ces fonctions nous permettent d'augmenter la flexibilité de la base en tenant compte de la déformation des orbitales de valence lors de la déformation de la molécule. Ces orbitales sont de type p, d pour l'hydrogène ; d, f et g pour les atomes de la 2^{éme} et 3^{éme} période, …, etc. Les orbitales de polarisation, qui sont des OA de nombre quantique *l* plus élevé que celui des OA de valence, sont très utiles pour la localisation des états de transitions. En effet, dans une réaction, des liaisons se coupent, d'autres se créent. Il est donc essentiel de pouvoir bien décrire les déformations du nuage électronique.

Un autre type de fonctions est indispensable à inclure dans la base d'orbitale atomique chaque fois que le phénomène physique décrivant la propriété étudiée nécessite une bonne description de l'espace situé au-delà des orbitales de valence (espace diffus). Ce sont les fonctions **diffuses**, qui augmentent la taille du nuage électronique. Pour les espèces ayant des doublets libres et les espèces chargées (anions), la présence d'orbitales diffuses est indispensable. On note par le signe +, signifiant la présence d'orbitales diffuses, celle des orbitales de polarisation est notée par un astérisque (*). Par exemple la base **6-31+G*** désigne une base SV-DZ 6-31G avec des orbitales diffuses, et de polarisation sur les atomes lourds ; **6-311++G*** est une base SV-TZ 6-311G avec des orbitales diffuses sur tous les atomes, et des orbitales de polarisation uniquement sur les atomes lourds. D'autres bases gaussiennes ont été proposées par Dunning et Huzinaga [**30**,**31**]. Malgré les divers perfectionnements apportés à la base gaussienne, l'utilisation de ces bases présente plusieurs inconvénients [**32**]. Pour cette raison, la recherche d'une base plus fiable et plus pratique reste toujours un centre d'intérêt de première importance des chimistes théoriciens, et on assiste ces dernières années à un retour, même s'il est un peu timide, vers les orbitales de Slater de qualité supérieure à celle des GTOs [**32**]**.** On note également que plusieurs programmes moléculaires utilisant les STOs commencent à faire leur apparition. A titre d'exemple, nous citons les programmes ALCHEMY [**33**]**,** STOP [**34**] et ADF (*Amsterdam Functional Theory*) [**35**]**.**

III.5. Les méthodes semi-empiriques

Ayant utilisé les méthodes AM1 PM3 et PM6 pour l'optimisation de nos composés nous rappelons ici les éléments essentiels de ces méthodes semi-empiriques, leurs avantages et défauts à travers un court historique.

Contrairement aux méthodes ab initio, les méthodes semi-empiriques utilisent des données ajustées sur des résultats expérimentaux afin de simplifier les calculs. La longueur et la difficulté des calculs est en grande partie due aux intégrales biélectroniques qui apparaissent aux cours du processus de résolution.

Pour réduire ce temps de calcul, la base d'orbitales atomiques est d'abord réduite au minimum. Ensuite, les méthodes quantiques semi-empiriques font certaines approximations sur ces intégrales parmi lesquelles, certaines intégrales bi-électroniques sont négligées et d'autres sont paramétrées grâce à des données expérimentales. Plusieurs méthodes pour le traitement des intégrales bi-électroniques existent comme la méthode CNDO (Complete Neglect of Differential Overlap). De plus, les intégrales de recouvrement S_{ij} sont exprimées par :

$$S_{ij} = \delta_{ij} \begin{cases} = 1 & si \quad i = j \\ = 0 & si \quad non \end{cases} \qquad (63)$$

En principe, seuls les électrons de valence sont considérés diminuant ainsi le nombre de termes calculés. Le reste de l'atome est traité comme un « coeur » de charge « Z - le nombre d'électrons de coeur ». Les électrons de coeur sont pris en compte par une fonction de répulsion « coeur-coeur » en même temps que la répulsion nucléaire. Cette énergie positive s'ajoute à l'énergie SCF. Durant l'évolution des méthodes semi-empiriques, cette fonction a été améliorée pour mieux rendre compte notamment, des liaisons hydrogènes.

De nombreuses méthodes semi-empiriques basées sur ces approximations ont été développées notamment les méthodes MNDO **[36, 37],** AM1 **[38]**, PM3 **[39]**, **PM6 [40].** L'avantage majeur de ces méthodes est la rapidité de calcul (méthode d'ordre b3) au

sacrifice de la qualité de la description énergétique. L'incorporation des effets de dispersion et de liaison hydrogène passe par l'ajout de fonctions empiriques de type « champ de force ». En revanche, l'usage de paramètres basés sur des propriétés expérimentales permet de prendre en compte, bien que partiellement, des effets de corrélation électronique (absente dans la méthode de Hartree-Fock).

III.5.1. La méthode MNDO

La méthode MNDO (Modified neglect of differential overlap) est une méthode reportée par Dewar et Thiel en 1977 **[36, 37].** À l'origine, MNDO ne considérait que les atomes de carbone, d'hydrogène, d'oxygène et d'azote. Puis, par la suite, la paramétrisation s'est étendue à un plus grand nombre d'atomes.

Un des désavantages de cette méthodologie est qu'elle décrit mal les liaisons hydrogènes **[40]** qui sont pourtant essentielles dans de nombreux complexes protéines-ligands.

III.5.2. La méthode AM1

Afin de corriger le problème de la représentation des liaisons hydrogènes, Dewar *et al.* ont développé, en 1985, la méthode AM1 (Austin Model 1) **[38]**. Pour cela, des fonctions Gaussiennes ont été ajoutées à la méthodologie MNDO pour représenter les interactions noyau-noyau.

III.5.3. La méthode PM3

Malgré les efforts effectués dans le développement de la méthodologie AM1, certains problèmes de paramétrisation persistent. Stewart *et al.* ont donc proposé en 1989 une nouvelle méthodologie nommée « Parametrized Model 3 » (PM3) **[39].** Dans cette méthodologie, la paramétrisation atomique a été effectuée par groupe d'élément. Deux fonctions Gaussiennes par atome sont utilisées pour le calcul de la répulsion coeur-coeur. De plus, des paramètres pour les éléments du groupe *d* font partie de cette méthode.

- **La méthode PM3-PIF**

En dépit de la correction effectuée au niveau de la répulsion coeur-coeur pour la liaison hydrogène, les méthodologies MNDO et PM3 sont malgré tout déficientes dans la représentation de ce type d'interaction. La méthodologie PM3 a tendance à surestimer l'interaction liaison hydrogène [41]. C'est pourquoi Bernal-Uruchurtu et Ruiz-Lopez ont mis en place en 2000 une correction à la méthode PM3 nommée PIF (Parametrized Interaction Functions) [41]. Cette correction permet de mieux décrire la surface d'énergie potentielle des liaisons hydrogènes impliquant les atomes d'oxygène. La fonction pour le calcul d'interaction coeur-coer de la méthode PM3, est remplacée par la fonction suivante :

$$V_{AB}^{PIF} = \alpha_{AB} e^{-\beta_{AB} R_{AB}} + \frac{\chi_{AB}}{R_{AB}^{6}} + \frac{\delta_{AB}}{R_{AB}^{8}} + \frac{\varepsilon_{AB}}{R_{AB}^{10}} \qquad (64)$$

Cette fonction a été paramétrée pour un nombre restreint d'interaction (principalement les interactions oxygène-oxygène, oxygène-hydrogène et hydrogène-hydrogène).

III.5.4. La méthode PM6

Stewart *et al.* ont développé, en 2007, une nouvelle méthode s'appuyant sur PM3 nommée PM6 dans laquelle a été incorporé un nouveau paramétrage coeur-coeur avec un accent sur les composés d'intérêt biologique [40]. Pour cela, ils ont modifié l'interaction coeur-coeur par une fonction de Voityuk [42] qui permet de prendre en compte la répulsion de deux atomes non chargés grâce à l'incorporation d'un terme diatomique.

De plus, les paramètres pour le traitement des orbitales *d* ont été ajoutés ce qui permet d'avoir, désormais, 80 atomes paramétrés pour cette méthode et de pouvoir ainsi traiter les métalloprotéines.

Malgré les améliorations apportées, la méthode PM6 échoue pour la description des interactions non-covalentes notamment en ce qui concerne la dispersion et la

représentation des liaisons hydrogènes [**43**] en sous estimant la force de ces interactions. Plusieurs méthodologies ont été présentées à partir des années 2009 afin de pallier ces problèmes dont PM6-DH[**43**], PM6-DH2 [**44**], PM6-DH+[**45**] et PM7[**46**].

Références Bibliographiques

[1] E. Schrödinger, Ann. Phys. Leipzig. 1926, 76, 361.

[2] M. Born et J. R. Oppenheimer, Ann. Phys. 1927, 84, 457.

[3] V. Minkine, B. Simkine, R. Minaev, « Théorie de la structure moléculaire » Edition Mir,Moscou, 1982.

[4] V. Fock, Z. Physik., 1930, 61, 126.

[5] J. C Slater, Phys. Rev., 1929, 34, 1293; 1931, 38, 38.

[6] C. C. Roothaan, Rev. Mod. Phys., 1951, 23, 69.

[7] I. Shavitt, « *Methods of Electronic Structure Theory* », H. F. Shaefer, Ed., Plenum Press, New-York, 1977, 189.

[8] A. Jugl, « Chimie Quantique Structurale et Eléments de Spectroscopie Théorique », 1978.

[9] C. Moller et M. S. Plesset, *Phys. Rev.*, 1934, 46, 618.

[10] J. L. Rivail, *Eléments de chimie quantique*, InterEditions/CNRS Editions, Paris, 1994.

[11] (a) R. G. Parr and W. Yang «Density Functional Theory», Oxford University, Press, 1989.
(b) L. J. Bartolotti and K. Flurchick, *Rev. Comput. Chem.*, 1996, 7, 187.
(c) St-Amant. Rev. Comput. Chem., 1996, 7, 217.
(d) T. Ziegler. Chem. Rev., 1991, 91, 651.
(e) E. J. Baerends et O. V. Gritsenko. J. Phys. Chem., 1997, 101, 5383.

[12] P. Hohenberg and W. Kohn, Phys. Rev., 1964, 136, B846.

[13] W. Khon and L. J. Sham, Phys. Rev., 1965, 140, A1133.

[14] S. J. Vosko, L.Wilk and M. Nusair, Can. J. Phys., 1980 , 58, 1200.

[15] F. Jensen « Introduction to Computational Chemistry », John Wiley & Sons, 1999.

[16] J. C. Slater, Phys. Rev., 1951, 81, 385.

[17] A. D. Becke, Phys. Rev., B, 1988, 38, 3098.

[18] C. Lee, W. Yang and R. G. Parr, Phys. Rev. B, 1988, 37, 785.

[19] A. D. Becke, J. Chem. Phys., 1993, 98, 5648.

[20] S. Dapprich, I. Komaromi, K. S. Byun, K. Morokuma and M. J. Frisch, "A New ONIOM Implementation for the Calculation of Energies, Gradients and Higher Derivatives Using Mechanical and Electronic Embedding I.," Theo.Chem. Act., in prep. 1998.

[21] I. Komaromi, S. Dapprich, K. S. Byun, K. Morokuma and M. J. Frisch, "A New ONIOM Implementation for the Calculation of Energies, Gradients and Higher Derivatives Using Mechanical and Electronic Embedding II.," Theo. Chem. Act., in prep. 1998.

[22] S. Humbel, S. Sieber and K. Morokuma, J. Chem. Phys. 1996, 105, 1959.

[23] F. Maseras and K. Morokuma, J. Comp. Chem. 1995, 16, 1170.

[24] T. Matsubara, S. Sieber and K. Morokuma, "A Test of the New "Integrated MO + MM" (IMOMM) Method for the Conformational Energy of Ethane and n-Butane," Journal of Quantum Chemistry, 1996, 60, 1101.

[25] M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber and K. Morokuma, J. Phys. Chem., 1996, 100, 19357.

[26] SPARTAN `04` , Wavefunction Inc., Irvine, 2003, CA 92612.

[27] J. C. Slater, J. Chem. Phys., 1930, 36, 57.

[28] S. F. Boys, Proc. Roy. Soc. 1950, A200, 542.

[29] E. Clementi, Ed., «Modern Techniques in Computationnal Chemistry», MOTECC™ 89, (ESCOM, Leiden), 1989.

[30] S. Huzinaga, J. Chem. Phys. 1965, 42, 1293.

[31] T. H. Dunning, J. Chem. Phys. 1971, 55, 716.

[32] S. M. Mekelleche , Thèse de doctorat d'état, Université de Tlemcen, 2000.

[33] M. Yoshimine, B. H. Lengsfield, P. S. Bagus, McLean, and B. Liu, Alchemy II (IBM, Inc., 1990) from MOTECC-90.

[34] A. Bouferguène, M. Fares, and p. E. Hoggan, Int. J. Quant. Chem., 1996, 57, 810.

[35] E. Van Lenthe, R. Van Leeuwen, E. J. Baerends, and J. G. Snijders, «in New hallenges in Computational Quantum Chemistry», (Ed Bagus, Groningen, 1994, 93.

[36] M. J. S. Dewar, W. Thiel, J. Am. Chem. Soc. **1977**, 99, 4899–4907.

[37] M. J. S. Dewar, W. Thiel, J. Am. Chem. Soc. **1977**, 99, 4907–4917.

[38] M. J. S. Dewar, E. G. Zoebisch, E. F.Healy, J. J. P. Stewart, J. Am. Chem. Soc. **1985**, 107, 3902–3909.

[39] J. J. P. Stewart, J. Comput. Chem. **1989**, 10, 209–220.

[40] J. J. P. Stewart, J. Mol. Model. **2007**, 13, 1173–1213.

[41] M. I.Bernal-Uruchurtu, M. F. Ruiz-López, Chem. Phys. Lett. **2000**, 330, 118–124.

[42] A. A.Voityuk, N. Rösch, J. Phys. Chem. A **2000**, 104, 4089–4094.

[43] J. Rezac , K. J. Fanfrl k, D. Salahub, P. Hobza, J. Chem. Theory. Comput. **2009**, 5, 1749–1760.

[44] M. Korth, K. M. Pito , J. Rezac , P.Hobza, J. Chem. Theory. Comput. **2010**, 6, 344–352.

[45] M. Korth, J. Chem. Theory. Comput. **2010**, 6, 3808–3816.

[46] J. P. Stewart, J. Mol. Model. **2013**, 19, 1–32.

CHAPITRE IV

GENERALITES SUR LA TOXICOLOGIE

MODELISATION QSTR

Introduction

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

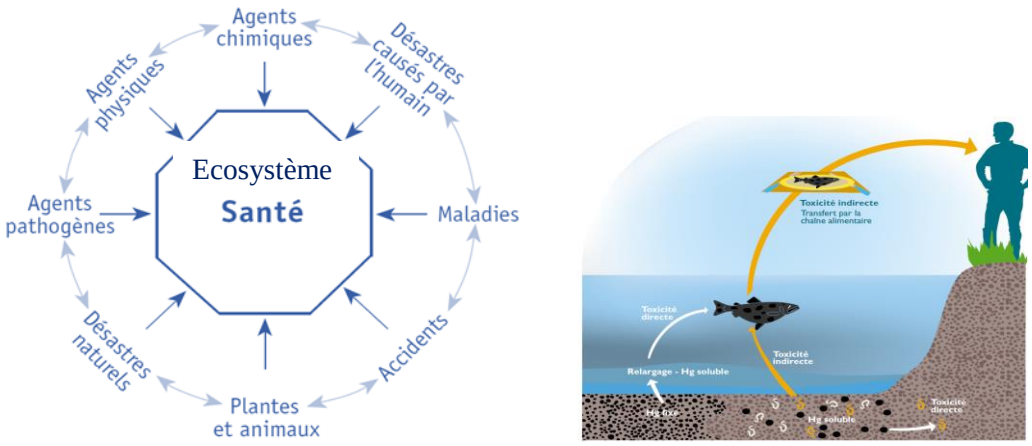


Figure. 1. Le milieu et les différents éléments pouvant affecter l'organisme vivant et l'écosystème

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

IV. 1. Notions et définitions [1]

IV.1. 1. Définition de la toxicologie

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

IV.1. 2. Définition de l'écotoxicologie :

L'écotoxiologie étudie les impacts des agents polluants sur la structure et le fonctionnement des écosystèmes.

Les effets d'un agent polluant dépendent de plusieurs facteurs, comme par exemple l'évolution du polluant dans le milieu, le mode et la voie d'administration du polluant.

IV. 1. 3. Définition d'un toxique (poison)

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

IV.1.4. Définition de la toxicité : il s'agit de la capacité inhérente à une substance chimique de produire des effets nocifs chez un organisme vivant et qui en font une substance dangereuse.

On parle alors d'un effet toxique, cet effet peut être local qui survient au point de contact, ou bien systémique qui survient à un endroit éloigné du point de contact initial.

IV.1. 5. Définition de la dose

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

IV.1. 6. Types de toxicité

- Toxicité aigüe: est définie comme celle qui résulte de l'exposition unique et

massive (ou de doses ramassées dans le temps) à un produit chimique entraînant des dommages corporels pouvant conduire à la mort.

Elle introduit la notion de dose « absorbée » (par ingestion, inhalation ou contact cutané) et se mesure par la **DL 50** (**dose létale**, ou dose provoquant la mort de 50% des animaux exposés à une dose unique du produit incriminé), exprimée en mg/kg de l'animal d'expérience retenu.

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

IV.1.7. Notions d'exposition

L'exposition est le couple « concentration en polluant/durée » auquel les organismes sont exposés.

- La biodisponibilité : se définit comme la propriété d'un élément d'atteindre les membranes cellulaires des organismes vivants.

Un polluant biodisponible est un polluant auquel les organismes sont exposés.

- La dégradation et la biodégradation : ce sont les principaux facteurs qui régissent le devenir des substances chimiques dans l'environnement.
- La bioaccumulation : est l'accumulation de substances toxiques dans les tissus des organismes vivants. Le paramètre utilisé pour mesurer la concentration du polluant dans l'organisme est BCF.

Cependant, très couramment on s'appuie sur le coefficient de partage octanol/eau (Kow) pour prédire la capacité d'un polluant à se bio-accumuler.

BCF= Concentration de polluant dans l'organisme/concentration de polluant dans le milieu.

BCF= Kow x concentration de la substance dans les lipides.

Il est considéré qu'une substance est bio-accumulable si

Log P≥3

BCF> 100

IV.1. 8. Voies d'absorption d'un toxique

Les principales façons d'absorption sont: l'inhalation (voie respiratoire), l'absorption par la peau (voie cutanée), l'ingestion (voie digestive)

IV.1. 9. Les phases du processus d'intoxication

a- La phase d'exposition (mise en contact avec le toxique suivie de sa résorption)

b-La phase toxico cinétique (elle commence après la résorption et aboutit à la présence du toxique dans le milieu intérieur)

c- La phase toxico dynamique (interaction avec le tissu cible).

La fraction de substance qui passe de la phase d'exposition à la phase toxico cinétique détermine sa disponibilité chimique.

IV.1.10. Les manifestations toxiques :

- **Selon les différents types d' effets toxiques**

Mortalité, Irrhitation et corrosion, la concerogénèse, la mutagénèse, l'allergie

Les effets sur la reproduction et le developpement.

- **Selon les systèmes biologiques et les organes cibles**

- L'hépato-toxicité, la néphro-toxicité, la neuro-toxicité, la dermato-toxicité, la toxicité de l'appareil respiratoire, la toxicité cardiovasculaire…

IV.1.11. Evaluation de l'effet toxique

On peut citer quatre catégories pour évaluer un effet toxique (toxicité) (**Figure. 2**) :

- ❖ Les études épidémiologiques, qui comparent plusieurs groupes d'individus ou les études de cas.
- ❖ Les études expérimentales in vivo, qui utilisent des animaux (ex. : lapin, rat et souris, *Tetrahyména-pyriformis*, poissons, algues…) ou les bio-essais sont utilisés
- ❖ Les études in vitro, effectuées sur des cultures de tissus ou des cellules.
- ❖ Les études théoriques par modélisation en l'occurrence les méthodes QSAR.

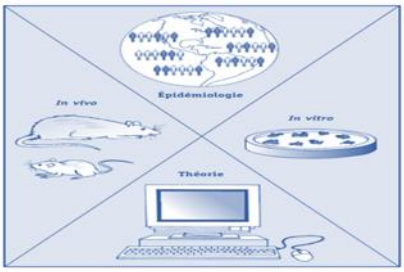


Figure.2. Méthodes d'évaluation d'un effet toxique

IV. 2. Toxicité des produits chimiques organiques

Sur 100 000 produits chimiques libérés dans l'environnement, moins de 1 - 5% ont des données de toxicité disponibles. Même pour les produits chimiques de volume de production élevé ou HPVCs (les substances chimiques produites en quantités>1000 tonnes par an dans l'UE ou > environ 442 tonnes par an aux Etats-Unis) il y'a un manque d'informations concernant leur toxicité, comme le montre le **Tableau 1** [2].

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

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

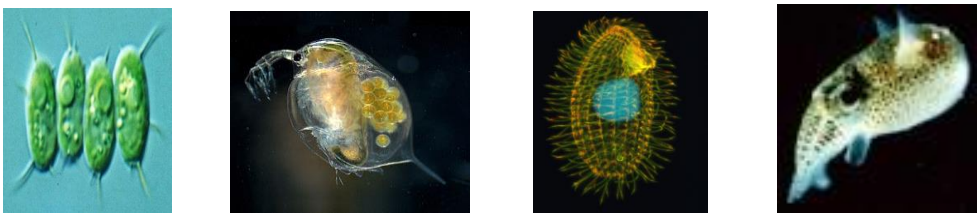
Avec une préoccupation croissante sur l'environnement et la santé humaine, les gouvernements et les organismes de réglementation à travers le monde cherchant à évaluer les risques éco-toxicologiques posés par la libération des substances chimiques. Ils ont proposé que **30000** produits chimiques existants soient testés sur les animaux pour une gamme des effets toxiques [3]. Ce serait évidemment une tache très coûteuse, impliquant l'utilisation des milliers d'animaux, par conséquent le recours aux méthodes alternatives est devenu d'une grande importance.

IV. 3. Les approches QSAR pour l'étude de la toxicité

QSAR pour la toxicité remonte au 19ème siècle. En 1863, A.F.A. Cros à l'Université de Strasbourg a observé que la toxicité des alcools à des mammifères augmente lorsque la solubilité des alcools dans l'eau diminue. Dans les années 1890, Hans Horst Meyer de l'Université de Marburg et Charles Ernest Overton de l'Université de Zurich, ont noté que la toxicité des composés organiques était tributaire de la lipophilie [4]. Au début des années 1960 Corwin Hansch a proposé un modèle mathématique pour corréler l'activité biologique et la structure chimique [5], cette date est considérée comme étant la naissance des méthodes QSAR. Depuis, l'utilisation des QSAR en toxicologie n'a pas cessé d'évoluer.

Des modèles QSAR sont maintenant mis au point en utilisant une variété d'approches, de méthodes d'analyse de données et de paramètres [6].

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



Vibrio fischeri *Tetrahymena.p.* Daphnie Algues

Figure.3. *Les différentes espèces aquatiques utilisées pour l'étude de la toxicité aquatique*

Cronin et Dearden [7, 8] ont examiné la littérature concernant la modélisation QSAR de la toxicité aquatique. Plusieurs modes d'action ont été identifiés chez les espèces aquatiques, à savoir narcose non polaire, narcose polaire, découplage de la phosphorylation oxydative, irritation de la membrane respiratoire, inhibition d'acétylcholinestérase, saisie de système nerveux central, inhibition de photosynthèse, et l'alkylation. Cependant, ceux-ci sont généralement plus largement regroupés comme: narcose non polaire, narcose polaire, réactifs sélectifs (électrophiles), et les molécules à mécanismes d'action spécifiques. Verhaar et al. [9] ont développé un système sur la base de la présence de groupes fonctionnels pour classer les produits chimiques dans ces quatre groupes **Tableau 2**.

Tableau.2. Les différents mécanismes possibles pour la toxicité (classification de Verhaar)

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

La dépendance de la toxicité des narcoses avec le coefficient de partition, en particulier le coefficient de partage octanol-eau, a été montrée par de nombreux auteurs (Dearden et al. [10], Freidig et Hermens, [11], Kapur et al. [12], Parkerton et Konkel [13]; Bundy et al.[14], Gramatica et al. [15], Ren et Frymer, [16], Sverdrup et al. [17], Worgan et al. [18]. Le QSAR représentant une toxicité de référence a été dérivé pour un groupe de narcotiques non polaires (alcools saturés, des cétones, des nitriles, des esters et des composés contenant du soufre). Ils ont conclu que le coefficient de partage octanol-eau est suffisant pour expliquer la toxicité des narcoses apolaires tandis que pour les narcoses polaires il faut la présence d'un descripteur supplémentaire qui explique le caractère électronique des molécules.

En outre, des modèles de combinaison des deux groupes de composés (narcoses apolaires et polaires) et les deux types de descripteurs ont été développés. Freidig et Hermens [11] ont conclu qu'en utilisant des modèles QSAR distincts pour les composés agissant par des mécanismes différents, y compris un descripteur qui caractérise le mécanisme de toxicité particulière, donne de meilleurs résultats que l'utilisation d'un modèle unique qui combine tous les composés avec les mêmes descripteurs.

Certains auteurs ont utilisé l'approche «réponse-surface» sur la base de l'hydrophobie et l'électrophilie des composés. Dans cette approche, les QSAR comprennent un descripteur qui caractérise la bio-absorption et la distribution (généralement partage octanol-eau ou coefficients de distribution (log *P* ou log*D*) et un descripteur de réactivité électrophile (habituellement LUMO ou la superdelocalisabilité maximale (Amax)). Cette approche a été appliquée à des différentes espèces aquatiques, y compris la bactérie *Vibrio fischeri* (Cronin et al. [19], les protozoaires *Tetrahymena pyriformis* (Cronin et Schultz [20], Cronin et al. [21], Schultz et al. [22], les algues de *Scenedesmus* (Wargan.) [23] et *Chlorella vulgaris* (Cronin et al.) [24]. L'avantage de l'approche réponse-surface est qu'elle est simple et a une interprétation mécanistique. Alors que certains auteurs (par exemple, Cronin et Schultz [20], Cronin et al. [19], Cronin et al. [21]) ont utilisé LUMO comme descripteur de réactivité électrophile entraînant des **interactions covalentes** dans les systèmes biologiques, Dimitrov et al. [25] et Dimitrov et al. [26] ont suggéré que LUMO peut être également utilisé pour décrire **l'interaction électrophile non covalente** des produits chimiques narcotiques avec le site d'action. Certains auteurs ont prolongé la démarche réponse-surface en ajoutant un indicateur supplémentaire est difficile et d'autres paramètres afin d'améliorer l'ajustement statistique des modèles (Schmitt et al. ; Wang et al. [27], Huang et al. [28], Cronin et al. [29], Netzeva et al., [30]). Toutefois, selon Schultz et al. [31] la modélisation QSAR des composés électrophiles en raison de la limitation des données et des descripteurs par rapport à la modélisation QSAR des composés agissants par d'autres mécanismes toxiques.

Cependant, des QSAR basés sur des indices topologiques pour l'étude de la toxicité ont fait l'objet de nombreux travaux (Burden [32], Gramatica et al. [15], Grodnitzky et Coats [33], Huuskonen [34], Rose et Hall [35]).

Récemment plusieurs programmes ont été développés pour le calcul des descripteurs pour l'étude QSAR de la toxicité, Katritzky et ses collaborateurs ont développé CODESSA [36], ils ont publié de nombreux articles sur les QSAR pour la prédiction de la toxicité des produits chimiques. DRAGON [37] est un autre programme pour le calcul des descripteurs moléculaires, ces derniers ont été utilisés pour élaborer des

modèles QSTR pour examiner la toxicité de nombreuses familles de molécules toxiques. Les modèles QSTR obtenus avec ces programmes sont des combinaisons de plusieurs types de descripteurs, quantiques, topologiques, thermodynamiques…en revanche ces modèles parfois n'ont aucune relation avec les mécanismes de toxicité et sont difficiles à interpréter. Plus récemment plusieurs modèles QSAR sont élaborés sur le programme CORAL [38-41] pour l'étude de la toxicité de différentes familles de molécules toxiques. Ces modèles sont de bonne qualité du coté statistique mais leur interprétation chimique est difficile.

Références Bibliographiques

[1] Notions de Toxicologie www.csst.qc.ca

[2] C. John. Dearden, Prediction of Environmental Toxicity and Fate Using Quantitative Structure-ActivityRelationships (QSARs) *J. Braz. Chem. Soc*. **2002**, 13, 754-762.

[3] Commission of the European Communities; White Paper on a Strategy for a Future Chemicals Policy, Brussels, Belgium, **2001**. http://europa.eu.int/comm/environment/chemicals/whitepaper.htm.

[4] R. L. Lipnick, *Trends Pharmacol. Sci*. **1986**, 7, 161–164.

[5] C. Hansch, P.P. Maloney, T. Fujita, R.M. Muir, *Nature* **1962**, 194, 178.

[6] I. Lessigiarska, A.P. Worth, T.I. Netzeva, Comparative review of QSARs for acute toxicity. EUR report No. 21559 EN. EC Joint Research Centre, Ispra, Italy (2005b).

[7] M.T.D.Cronin, J.C. Dearden, *Quant. Struct.-Act. Relat*. **1995**, 14, 1.

[8] M.T.D. Cronin, , Netzeva, T.I, Dearden, J.C., Edwards, R., Worgan, A.D.P. (). Assessment and Modeling of the Toxicity of Organic Chemicals to Chlorella vulgaris: Development of a Novel Database. *Chem. Res. Toxicol*, **2004**, 17, 545-554.

[9] H.J.M.Verhaar, C.J. Van Leeuwen, J.L.M. Hermens. Classifying environmental pollutants 1. Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere*, **1992**, 25, 471-491.

[10] J.Dearden, ECVAM Workshop on The Use of Computer Models as Alternatives to Animal Experiments in Chemical Risk Assessment, October 3-4, **2002**, Praha, Czech Republic.

[11] A.P.Freidig and J.L.M. Hermens. Narcosis and chemical reactivity QSARs for acute fish toxicity. *Quant. Struct. Act. Rel.*, **2000**, 19, 547-553.

[12] S. Kapur, A. Shusterman, R.P.Verma, C. Hansch, C.D.Selassie, Toxicology of benzylalcohols: a QSAR analysis. *Chemosphere*, **2000**, 41,1643-1649

[13] T.F.Parkerton and W.J.Konkel, Application of quantitative structure-activity relationships for assessing the aquatic toxicity of phthalate esters. *Ecotox. Environ. Saf*, **2000**, 45,61-78.

[14] J.G.Bundy, A.W.J.Morriss, D.G.Durham, C.D.Campbell, G.I.Paton.

Development of QSARs to investigate the bacterial toxicity and biotransformation Potential of aromatic heterocylic compounds. *Chemosphere*, **2001**, 42, 885-892.

[15] P.Gramatica, M.Vighi, F. Consolaro, R. Todeschini, A. Finizio, M. Faust. QSAR approach for the selection of congeneric compounds with a similar toxicological mode of action. *Chemosphere*, **2001**, 42, 873-883.

[16] S. Ren and P.D. Frymier. Estimating the toxicities of organic chemicals to bioluminescent bacteria and activated sludge. *Water. Res*, **2002**, 36, 4406-4414.

[17] L.E. Sverdrup, T. Nielsen, P.H. Krogh. Soil ecotoxicity of polycyclic aromatic hydrocarbons in relation to soil sorption,lipophilicity,and waters olubility. *Environ. Sci.Tech*, **2002**, 36, 2429-2435

[18] A.D.P.Worgan, J.C. Dearden, R. Edwards, T.I. Netzeva, M.T.D. Cronin. Evaluation of an ovelshort-termalgal toxicity assay by the development of QSARs and inter-species relationships for narcotic chemicals. *QSAR. Comb. Sci*, **2003**, 22, 204-209.

[19] M.T.D. Cronin, G.S. Bowers, G. D. Sinks, T.W. Schultz, Structure-toxicity relationships for aliphatic compounds encompassing a variety of mechanisms of toxic action to V. fischeri. *SAR QSAR. Environ. Res*, **2000**, 11, 301-312.

[20] M.T.D.Cronin, and T.W. Schultz. Development of quantitative structure-activity relationships for the toxicity of aromatic compounds to T. pyriformis: comparative assessment of the methodologies. *Chem. Res. toxicol*, **2001**, 14, 1284-1295.

[21] M.T.D.Cronin, A.O. Aptula, J.C. Duffy, T.I. Netzeva, P.H. Rowe, I.V.Valkova, , T.W. Schultz. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymenapyriformis. *Chemosphere*, **2002**, 49, 1201–1221.

[22] T.W. Schultz, D.T. Lin, T.S. Wilke, L.M. Arnold, Quantitativestructure- activity relationships for theTetrahymena pyriformis population growth endpoint: a mechanism of action approach. In: Karcher,W and Devillers, J.(Eds), Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology.Kluwer Academic Publishers, **1990**, 61-82.

[23] A.D.P. Worgan. The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. *SAR QSAR Environ. Res.* **2002**, 13, 167-176.

[24] M.T.D. Cronin, J.C. Dearden, J.C. Duffy, R.Edwards, N.Manga, A.P.Worth, ,

A.D.P. Worgan. The importance of hydrophobicity and electrophilicity

descriptors in mechanistically-based QSARs for toxicological endpoints. *SAR .QSAR Environ. Res*, 2002, 13, 167-176.

[25] S.D. Dimitrov, O.G. Mekenyan, T.W. Schultz, (). Interspecies modeling of narcotic toxicity to aquatic animals. *Bull. Environ. Contam. Toxicol*, **2000**, 65, 399-406.

[26] S.D. Dimitrov, O.G. Mekenyan, G.D. Sinks, T.W. Schultz,). Global modeling of narcotic chemicals: ciliate and fish toxicity. *J. Mol. Struc. (Theochem)*, **2003**, 622, 63-70.

[27] H. Schmitt, R. Altenburger, B. Jastorff, G.Schüürmann, Quantitative structure-activity analysis of the algae toxicity of nitroaromatic compounds. *Chem. Res. Toxicol.* **2000**, 13,441-450.

[28] H. Huang, X. Wang, W. Ou, J. Zhao, Y. Shao, L. Wang. Acute toxicity of benzene derivatives to the tadpoles (Ranajaponica) and QSAR analyses. *Chemosphere*, **2003**, 53, 963-970.

[29] M.T.D. Cronin, T.I. Netzeva, J.C. Dearden, R.Edwards, A.D.P. Worgan Assessment and Modeling of the Toxicity of Organic Chemicals to Chlorella vulgaris: Development of a Novel Database. *Chem. Res. Toxicol*. **2004**, 17, 545-554.

[30] T.I. Netzeva, J.C. Dearden, R. Edwards, A.D.P. Worgan, M.T.D. Cronin, QSAR analysis of the toxicity of aromatic compounds to Chlorella vulgaris in a novel short-term assay. *J. Chem. Inf. Comp. Sci*, **2004**, 44, 258-265.

[31] T.W. Schultz, G.D. Sinks, A.P. Bearden, QSAR in aquatic toxicology: a mechanism of action approach comparing toxic potency to Pimephales promelas,T. pyriformis,andV.fischeri.In:Devillers,J.(Ed.),Comparative QSAR.Taylor & Francis New **1998**, York,51-109.

[32] F.R. Burden. Quantitativestructure-activity relationship studies using Gaussian processes. *J. Chem. Inf. Comp. Sci*. **2001**,41, 830-835.

[33] Grodnitzky,J.A.,andCoats,J.R. QSAR evaluation of monoterpenoids' insecticidal activity. *J. Agr. Food. Chem*, **2002**, 50,4576-4580.

[34] J. Huuskonen, QSARmodelingwiththeelectrotopologicalstateindices: predicting the toxicity of organic chemicals. *Chemosphere*, 2003, 20,949-953.

[35] K. Rose, and L.H. Hall, E-state modelling of fish toxicity independent of 3D structure information. *SAR. QSAR. Environ. Res.*, **2003**, 14, 113-129.

[36] CODESSA PRO, University of Florida, www.codessa-pro.com

[37] R.Todeschini and V.Consonni: "Molecular Descriptors for Chemoinformatics", (2 volumes), WILEY-VCH,Weinheim (Germany) **2009**, PP 1257.

[38] A.A. Toropov, and E.Benfenati, QSARmodelling of aldehyde toxicity by means Of optimization of correlation weights of nearest neighbouring codes. *J. Mol. Struct*. (Theochem), **2004**, 676,165-169.

[39] A.A.Toropov, and T.W. Schultz, Predictionofaquatictoxicity:useof optimization of correlation weights of local graph invariants. *J.Chem. Inf. Comput. Sci*, **2003**, 43,560-567.

[40] A.P. Toropova, A.A. Toropov, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: QSAR modeling of toxicity of organic chemicals towards Daphnia magna. *Chemom. Int.Lab. Syst* **2012,** 110, 177-181

[41] A. P Toropova, A. A Toropov, E. Benfenati, T. Puzyn, D. Leszczynska, J. Leszczynksy. Optimal descriptor as a translator of eclectic information into the prediction of membrane damage: the case of a group of ZnO and TiO_2 nanoparticles. *Ecotox. Environ. Saf*. **2014**, 108, 203–209.

CHAPITRE V

APPLICATIONS – RESULTATS ET DISCUSSION

Application I

Etude QSAR de la toxicité des nitrobenzènes vis-à-vis *Tetrahymena-pyriformis* à l'aide des descripteurs quantiques

Résumé

Les relations quantitatives structure-activité (QSAR) sont très utiles pour comprendre comment la structure chimique se corrèle à l'activité biologique et à la toxicité des produits chimiques naturels et synthétiques. La présente application montre que l'indice d'électrophilie de Parr « ω » en combinaison avec deux autres descripteurs, à savoir, l'énergie LUMO et l'indice de lipophilie log P, sont des indices de choix pour la prédiction de la toxicité d'une série constituée d'une cinquantaine de dérivés de nitrobenzène. Les modèles QSAR sont développés en utilisant la méthode de régression linéaire multiple MLR. Il s'avère que le meilleur modèle dont sa stabilité est confirmée par la validation « leave-1/3-of-set-out » est capable de décrire environ 87% de la variance de la toxicité expérimentale. Le modèle obtenu montre que les nitrobenzènes les plus toxiques sont caractérisés par de grandes lipophilies et un pouvoir électrophile élevé. Le modèle QSAR obtenu pourrait être efficacement appliqué pour l'estimation de la toxicité des nitrobenzènes pour lesquels les mesures expérimentales ne sont pas disponibles.

1.Introduction

L'amplification spectaculaire des nouveaux composés par l'industrie chimique en général et en particulier par l'agrochimie, la pétrochimie et la pharmaco-chimie est accompagnée d'une augmentation de la charge toxique dans l'environnement. Pour cette raison, le développement d'outils capables d'évaluer les effets dangereux sur les espèces vivantes devrait recevoir une attention particulière [1].

Les méthodes Quantitatives structure-propriété/activité (QSPR/QSAR) sont parmi les outils les plus pratiques en chimie physique. Ces méthodes sont basées sur l'axiome que la variance dans les propriétés physico-chimiques et les activités des composés chimiques est déterminée par la variance dans leurs structures moléculaires. Ainsi, si les données expérimentales sont disponibles pour seulement certains produits chimiques dans un groupe, on peut prédire le reste en utilisant des descripteurs moléculaires calculés pour l'ensemble du groupe et adapté un modèle mathématique [2-4]. La prédiction de la toxicité en utilisant des QSAR a été l'objectif de nombreux chercheurs qui utilisaient une variété d'approches. Cet objectif est séduisant, mais n'a pas encore été atteint de manière satisfaisante. Il y'a un certain nombre de raisons pour l'absence de succès [5]. Ce manque de réussite a été aggravé dans de nombreuses études par une mauvaise appréciation de l'hétérogénéité insuffisante, ou la diversité chimique, dans la base de données. En outre, certaines propriétés moléculaires (comme l'hydrophobie) sont bien décrites, d'autres, y compris la réactivité électrophile, l'ionisation, et la liaison hydrogène, sont mal paramétrisées. Enfin, les mécanismes d'action toxiques sont mal interprétés ou ne sont pas totalement compris, ou leur pertinence dans la modélisation de la toxicité est ignorée [6].

Les nitrobenzènes sont des produits chimiques dangereux qui affichent plusieurs manifestations de la toxicité, y compris la sensibilité de la peau, l'immunotoxicité, la dégénérescence des cellules germinales, l'inhibition des enzymes hépatiques. La modélisation de la toxicité des composés nitro-aromatiques a été compliquée à cause de la pénurie des données expérimentales. Les nitrobenzènes (NB) sont largement utilisés comme produits chimiques industriels, et par conséquent ont un potentiel élevé dans la pollution de l'environnement. Ils ont été rapportés [7] d'être présents dans les eaux de

surface. Se sont des produits chimiques réactifs, étant signalés comme découpleurs de la phosphorylation oxydative [8] et peuvent être considérés comme pro-électrophiles [9]. Ils peuvent subir un certain nombre de différentes réactions électrophiles. En raison de leur utilisation à grande échelle, la toxicité des nitro- produits a été assez largement étudiée, et ils ont fait l'objet d'un certain nombre d'études QSAR. La toxicité des NBs a été largement étudiée par plusieurs groupes de chercheurs avec l'utilisation de différentes méthodologies. Les tentatives visant à modéliser la toxicité aiguë des NBs ont été examinés par Dearden et al. [10]. En raison de la nature réactive électrophile des dérivés du nitrobenzene, il n'est pas surprenant que les efforts de modélisation précédentes s'étaient focalisées sur l'utilisation des descripteurs moléculaires électroniques [11-18].

Puisque l'expression de la toxicité chimique est une combinaison de pénétration dans ou à travers des membranes biologiques et de l'interaction de la substance toxique avec le site d'action. Ce principe est représenté mathématiquement par le QSAR générique suivant [19] :

$$(\text{toxicité})^{-1} = A(\log \text{ de pénétration}) + B\ (\log \text{ de l'interaction}) + C \qquad (1)$$

La pénétration sur le site d'action est généralement représentée par l'hydrophobie, le plus souvent quantifiée par le coefficient de partage octanol/eau (log *P*). L'interaction du produit chimique avec le site d'action est plus compliquée et décrit les propriétés électroniques et /ou stérique.

L'objectif de cette application est de développer des modèles QSAR fiables et prédictifs pour identifier les principaux facteurs moléculaires expliquant l'effet d'interaction (électronique / pénétration) et régissant la toxicité de ces dérivés de NB et d'examiner quels sont les facteurs autres que log P qui peuvent contrôler la toxicité de ces composés.

2. Méthodologie

2.1. Base de données:

Un total de 50 structures de NBs contenant chacune soit un halogène ou / et un groupe méthyle, un groupe nitro et d'autres substituants, sont prises de la base de données décrite par Cronin et al. [6]. la structure modèle du nitrobenzène est présentée sur le schéma 1.

La toxicité expérimentale exprimée par (1/IGC50) des 50 congénères sont numérotés dans le tableau 1.

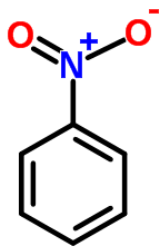


Schéma.1. Structure du nitrobenzène

Tableau. 1. Données expérimentales de la toxicité des NB vis-à-vis Tetrahymena-Pyriformis (Cronin et al., 2001) [6].

No	Composé	Toxicité Observée log (IGC_{50}^{-1})
1	2,6-dimethylnitrobenzene	0,30
2	2,3-dimethylnitrobezene	0,56
3	2-methyl-3-chloronitrobenzene	0,68
4	2-methylnitrobenzene	0,05
5	2-chloronitrobenzene	0,68
6	2-methyl-5-chloronitrobenzene	0,82
7	2,4,5-trichloronitrobenzene	1,53
8	2,5-dichloronitrobenzene	1,13
9	6-chloro-1,3-dinitrobenzene	1,98
10	nitrobenzene	0,14
11	3-methylnitrobenzene	0,05
12	1,3-dinitrobenzene	0,89
13	3,4-dichloronitrobenzene	1,16
14	4-methylnitrobenzene	0,17
15	1,4-dinitrobenzene	1,30
16	4-chlonitrobenzene	0,43
17	2,3,5,6tetrachloronitrobenzene	1,82
18	6-methyl-1,3-dinitrobenzene	0,87
19	3-chloronitrobenzene	0,73
20	1,2-dinitrobenzene	1,25
21	2-bromonitrobenzene	0,75

22	6-bromo-1,3-dinitrobenzene	2,31
23	3-bromonitrobenzene	1,03
24	4-bromonitrobenzene	0,38
25	2,4,6-trimethylnitrobenzene	0,86
26	5-methyl-1,2-dinitrobenzene	1,52
27	2,4-dichlorobenzene	0,99
28	3,5-dichlorobenzene	1,13
29	6-iodo-1,3-dinitrobenzene	2,12
30	2,3,4,5-tetrachloronitrobenzene	1,78
31	2,3-dichlorobenzene	1,07
32	2,5-dibromobenzene	1,37
33	1,2-dichloro-4,5-dinitrobenzene	2,21
34	3-methyl-4-bromonitrobenzene	1,16
35	2,3,4-trichloronitrobenzene	1,51
36	2,4,6-trichlonitrobenzene	1,43
37	4,6-dichloro-1,2-dinitrobenzene	2,42
38	3,5-dinitrobenzyl alcohol	0,53
39	3,4-dinitrobenzyl alcohol	1,09
40	2,4,6-trichloro-1,3-dinitrobenzene	2,19
41	2,3,5,6-tetrachloro-1,4-dinitrobenzene	**2,74**
42	2,4,5-trichloro-1,3-dinitrobenzene	2,59
43	4-fluoronitrobenzene	0,25
44	4-fluoro-2-nitrotoluene	0,25
45	1-fluoro-2-nitrobenzene	0,23
46	1-fluoro-3-nitrobenzene	0,20
47	4-nitrobenzaldehyde	0,20
48	2-nitrobenzaldehyde	0,17
49	3-nitrobenzaldehyde	0,14
50	3-nitroacetophenone	0,32

2.2. Calculs de la chimie quantique

Le programme MOPAC6 (Semi-empirical Molecular Orbital Package) [20] a été utilisé pour effectuer des calculs de chimie quantique et les géométries ont été optimisées en utilisant la méthode AM1.

Selon Cronin et al. [21], la toxicité des dérivés du nitrobenzène peut être expliquée en terme de pouvoir électrophile de ces composés, qui a été exprimé par le descripteur Amax qui exprime la super-délocalisabilté [22]. Cependant, le concept d'électrophilie est plus adéquatement défini dans le cadre de la DFT conceptuelle.

Dans le cadre de la DFT, le potentiel chimique μ et la dureté chimique η pour un système moléculaire à N électrons et d'une énergie totale E et un potentiel externe v sont définis comme les dérivées première et seconde de l'énergie par rapport à N, respectivement. Les valeurs approchées sont données par :

$$\mu = \frac{(E_{LUMO} + E_{HOMO})}{2} \quad (2)$$

$$\eta = E_{LUMO} - E_{HOMO} \quad (3)$$

Où E_{LUMO} est l'énergie de la plus basse orbitale moléculaire inoccupée et E_{HOMO} est l'énergie de la plus haute orbitale occupée.

Parr et al. [23] ont défini un nouveau descripteur, de la chimie quantique, connu sous le nom indice d'électrophilie " ω ", qui mesure le pouvoir à absorber les électrons, et se définit comme suit

$$\omega = \frac{\mu^2}{2\eta} \quad (4)$$

L'hydrophobie est exprimée comme le rapport des concentrations d'une substance en phase organique (octanol) et aqueuse (eau). Les valeurs de lipophilie ou coefficient de partition, log *P*, sont calculées en utilisant le programme ACD / Labs [24].

2.3. Analyse statistique

Des modèles Structure-toxicité ont été générés en utilisant la méthode de régression multilinéaire (MLR) de MINITAB version 15 [25], les valeurs log ($1/IGC_{50}$) indiquées en millimolaire, comme variable dépendante. ω, E_{LUMO} et log *P* sont utilisés comme variables indépendantes. Les modèles sont évalués par la valeur de R^2 (coefficient de détermination), le R^2-ajusté, la valeur SD (racine du carré moyen des erreurs) et la valeur F (statistique Fischer). Le nombre d'observations N est également à noter.

3. Résultats

Les valeurs d'électrophilie ω, E_{LUMO}, et le coefficient de partage estimé par log *P* sont donnés dans le **Tableau 2**.

Plusieurs modèles QSAR linéaires impliquant un, deux et trois descripteurs sont élaborés et les plus fortes corrélations multivariables sont identifiées par l'option "Best sub-set" du programme MINTAB.

Modèles QSAR à un seul paramètre*:*

Model#1

$$\log (1/ IGC_{50}) = 0{,}04 + 0{,}39 \log P \quad (5)$$

N =50 ; $R^2 = 0{,}16$; $R^2_{adj} = 0{,}14$; SD = 0,68

Model#2

$$\log (1/ IGC_{50}) = - 0{,}89 - 1{,}23\, E_{LUMO} \quad (6)$$

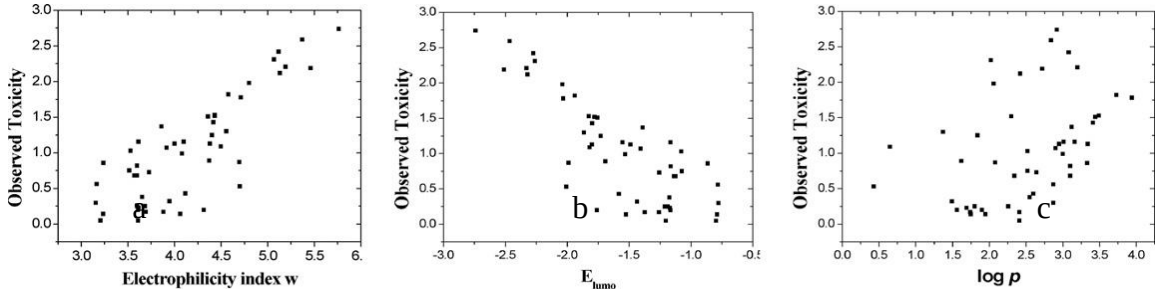
N =50 ; $R^2 = 0{,}66$; $R^2_{adj} = 0{,}65$; SD = 0,43

Model#3

$$\log (1/ IGC_{50}) = - 2{,}89 + 0{,}94\, \omega \quad (7)$$

N =50 ; $R^2 = 0{,}68$; $R^2_{adj} = 0{,}67$; SD = 0,42

La relation entre chaque descripteur et la toxicité expérimentale est présentée sur les figures 1(a-c).



Figures. 1a-c. Relation entre chaque descripteur et la toxicité mesurée

Il s'avère que le meilleur modèle QSAR à un seul paramètre est obtenue avec l'indice d'électrophilie de Parr ω (modèle # 3, R^2 = 0,68). Afin d'améliorer le pouvoir prédictif des modèles QSAR, il est nécessaire d'élaborer des modèles QSAR multilinéaires impliquant deux et trois paramètres.

Modèles QSAR à deux paramètres*:*

Model#4

$$\log (1/ IGC_{50}) = - 8,28 + 3,54\ \omega + 3,43\ E_{LUMO} \quad (8)$$

N =50 ; R^2 = 0,71 ; R^2_{adj} = 0,69 ; SD= 0,41 ; F= 56,44

Model#5

$$\log\ (1/ IGC_{50}) = - 1,78 + 0,36 \log P - 1,21\ E_{LUMO} \quad (9)$$

N =50 ; R^2 = 0,80 ; R^2_{adj}= 0,79 ; SD= 0,33 ; F= 94,75

Model#6

$$\log (1/ IGC50) = - 3,78 + 0,93\ \omega + 0,38 \log P \quad (10)$$

N = 50 ; R^2 =0,82 ; R^2_{adj}= 0,82; SD= 0,31 ; F= 112,34

Il s'avère que le meilleur modèle QSAR à deux paramètres est obtenu avec la combinaison du coefficient de partage log *P* et l'indice d'électrophilie ω (Modèle # 6, R^2 = 0,82).

Modèle QSAR à Trois paramètres

Model#7

$$\log(1/IGC_{50}) = -11,70 + 4,7\ \omega + 4,9\ E_{LUMO} + 0,42 \log P \quad (11)$$

N =50 R^2 = 0,87, R^2_{adj} = 0,86, SD = 0,27, F= 106,11

Une amélioration significative de la qualité du modèle QSAR est obtenue avec la combinaison des trois paramètres, à savoir, l'indice d'éléctrophilie de Parr ω le coefficient de partage log *P*, et l'énergie E_{LUMO}. La **Figure. 2** montre la corrélation linéaire entre les valeurs observées et les valeurs prédites de toxicité obtenue en utilisant l'équation. (11).

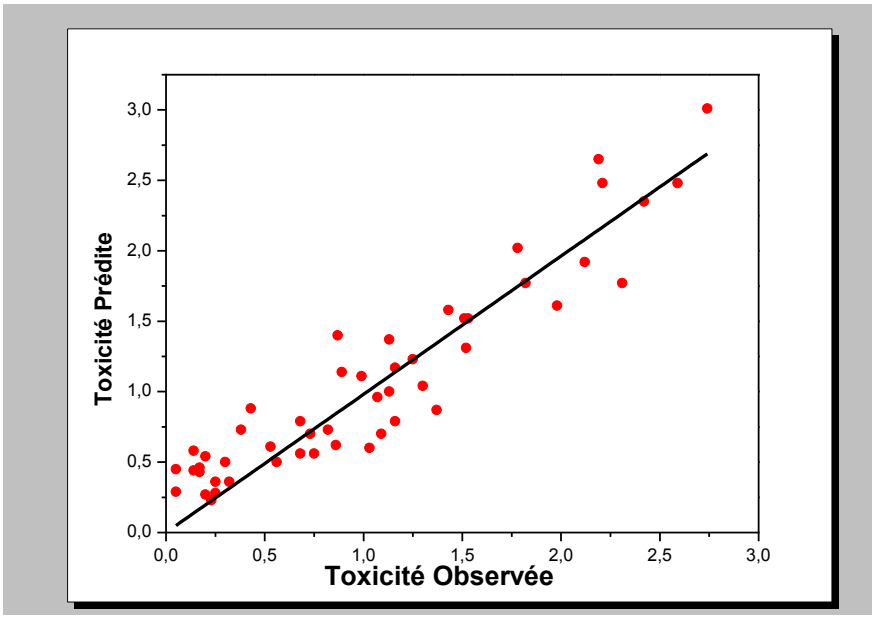


Figure.2. Corrélation entre la toxicité prédite et expérimentale

Tableau. 2. Valeurs des descripteurs et de toxicité observée/calculée en utilisant l'Eq.11

Composé	log *P*	E_{LUMO} (eV)	ω (eV)	Tox Obs	Tox Pred	Residual
1	2,87	-0,7769	3,1543	0,30	0,50	0,20
2	2,87	-0,7819	3,1642	0,56	0,50	0,06
3	3,10	-1,1248	3,5681	0,68	0,79	0,11
4	2,41	-0,7982	3,2040	0,05	0,45	0,40
5	2,34	-1,1346	3,5960	0,68	0,56	0,12
6	3,10	-1,1627	3,5966	0,82	0,73	0,09
7	3,49	-1,8247	4,4298	1,53	1,52	0,01
8	2,95	-1,4879	4,0007	1,13	1,00	0,13
9	2,06	-2,0373	4,7984	1,98	1,61	0,37
10	1,95	-0,7888	3,2322	0,14	0,44	0,30
11	2,41	-1,2031	3,6078	0,05	0,29	0,24
12	1,62	-1,6926	4,3708	0,89	1,14	0,25
13	3,16	-1,5547	4,0985	1,16	1,17	0,01
14	2,41	-1,2583	3,6922	0,17	0,46	0,29
15	1,37	-1,8640	4,5546	1,30	1,04	0,26
16	2,60	-1,5815	4,1154	0,43	0,88	0,45
17	3,73	-1,9382	4,5766	1,82	1,77	0,05
18	2,08	-1,9897	4,6950	0,87	1,40	0,53
19	2,64	-1,2550	3,7270	0,73	0,70	0,03
20	1,84	-1,7289	4,4023	1,25	1,23	0,02
21	2,52	-1,0742	3,5152	0,75	0,56	0,19
22	2,02	-2,2595	5,0662	2,31	1,77	0,54
23	2,52	-1,0775	3,5275	1,03	0,60	0,43
24	2,55	-1,1738	3,6505	0,38	0,73	0,35
25	3,33	-0,8642	3,2370	0,86	0,62	0,24
26	2,30	-1,7793	4,4306	1,52	1,31	0,21
27	3,00	-1,5317	4,0798	0,99	1,11	0,12
28	3,34	-1,7999	4,3797	1,13	1,37	0,24
29	2,42	-2,3197	5,1297	2,12	1,92	0,20
30	3,94	-2,0324	4,7113	1,78	2,02	0,24
31	2,90	-1,4078	3,9151	1,07	0,96	0,11
32	3,12	-1,3881	3,8587	1,37	0,87	0,50
33	3,20	-2,3302	5,1914	2,21	2,48	0,27
34	3,01	-1,1640	3,6124	1,16	0,79	0,37
35	3,44	-1,7567	4,3577	1,51	1,52	0,01
36	3,41	-1,7980	4,4151	1,43	1,58	0,15
37	3,08	-2,2735	5,1149	2,42	2,35	0,07
38	0,43	-2,0062	4,6984	0,53	0,61	0,08
39	0,65	-1,8173	4,4955	1,09	0,70	0,39
40	2,72	-2,5126	5,4605	2,19	2,65	0,46
41	2,92	-2,7402	**5,7611**	**2,74**	3,01	0,27
42	2,84	-2,4658	5,3695	2,59	2,48	0,11
43	1,80	-1,2118	3,6838	0,25	0,36	0,11
44	2,26	-1,1866	3,5977	0,25	0,28	0,03
45	1,69	-1,1698	3,6146	0,23	0,23	0,00
46	1,90	-1,1652	3,6092	0,20	0,27	0,07
47	1,56	-1,7598	4,3135	0,20	0,54	0,34
48	1,74	-1,3739	3,8796	0,17	0,43	0,26
49	1,75	-1,5254	4,0612	0,14	0,58	0,44
50	1,49	-1,4369	3,9445	0,32	0,36	0,04

Dans le **Tableau 3**, sont donnés les coefficients, les erreurs des coefficients et les valeurs t-test des trois paramètres moléculaires correspondants au meilleur modèle # 7.

Tableau. 3. Données du meilleur modèle # 7 Eq. 11

Numéro	X	DX	t-Test	Desripteur
0	-11,717	1,942	-6,03	Intercept
1	0,420	0,053	7,90	logP
2	4,705	0,916	5,14	ω
3	4,903	1,206	4,13	E_{LUMO}

4. Validation interne du meilleur modèle

Afin de vérifier la fiabilité et la stabilité du meilleur modèle QSAR (Eq.11), nous avons utilisé la validation interne « leave-1/3-of-set-out » de la manière suivante: les valeurs de données expérimentales parentes ont été divisées en fonction des valeurs expérimentales dans trois sous-ensembles (le 1er, 4ème , 7ème , …forment le premier sous-ensemble A, le 2nd , 5ème , 8ème, …. forment le sous-ensemble B, et le 3ème , 6ème , 9ème ,… forment le troisième sous-ensemble C). En combinant deux des sous-ensembles A, B , C, on obtient trois combinaisons et l'équation de corrélation est dérivée avec les mêmes descripteurs. L'équation obtenue a été utilisée pour prédire les données pour le sous-ensemble restant. Il s'avère que les valeurs prédites en utilisant R^2 pour les sous-ensembles (A + B), (B + C), (A + C) sont très proches de celle correspondant à l'ensemble complet de la série d'apprentissage (A + B + C) et les valeurs moyennes de R^2 (Fit) et R^2 (Prédites) (voir **Tableau 4**) sont également très proches. Notons que la valeur R^2_{adj} des modèles correspondant à des sous-ensembles A + B, A + C, et B + C sont beaucoup plus grandes que 0,80, ce qui indique que notre modèle est stable et peut être efficacement utilisée pour estimer la toxicité d'autres nitrobenzènes pour lesquels les données expérimentales ne sont pas disponibles.

Tableau. 4. Validation croisée du meilleur modèle présenté par l'Eq.11

Training set	N	R^2 (Fit)	R^2_{adj} (Fit)	SD (Fit)	Test set	N	R^2(pred.)	R^2_{adj} (pred.)
A + B	34	0,88	0,87	0,27	C	16	0,82	0,81
A + C	33	0,86	0,85	0,28	B	17	0,88	0,88
B + C	33	0,88	0,87	0,25	A	17	0,84	0,83
Average		0,87	0,86	0,26			0,85	0,84

5- Discussion statistique et mécanistique du modèle QSAR obtenu

Les modèles QSAR élaborés (Eqs.5-11) révèlent que la toxicité des nitrobenzènes pourrait s'expliquer par un certain nombre de facteurs électroniques et du transport. L'électrophilie défini par ω et E_{LUMO}, sont importantes dans la description de l'interaction électronique et la réactivité de ces toxines, tandis que l'hydrophobie, telle qu'elle est exprimée par log *P* est importante pour décrire le transport vers le site d'action. En effet, il s'avère que la combinaison des trois paramètres augmente considérablement le pouvoir prédictif du modèle QSAR donné par Eq.11 (R^2= 0,8 ; R^2_{adj} = 0,86 ; SD = 0,27 ; F = 106,11). Comme on le voit à partir des paramètres statistiques de l'équation ci-dessus, une amélioration considérable est obtenue en combinant les trois descripteurs (Eq. 11). Le QSAR modèle obtenu peut expliquer environ 87% de la variance expérimentale de la variable dépendante (IGC_{50}^{-1}) en plus il présente un F élevé de Fischer (F = 106,11) et une faible déviation standard (SD = 0,27) ce qui confirme que le modèle # 7 explique la toxicité (variable dépendante) d'une manière statistiquement significative satisfaisante.

Selon les valeurs du test t (| t |), l'importance des descripteurs impliqués dans ce modèle est dans l'ordre suivant: log P> ω > E_{LUMO}. Le descripteur le plus important selon le t- test (voir **Tableau 3**) est le coefficient de partage log *P*. Le second descripteur est l'indice de l'électrophile de Parr ω et le dernier est l'énergie LUMO.

E_{LUMO} est directement liée à l'affinité électronique d'une molécule et il caractérise la sensibilité de la molécule d'être attaqué par les nucléophiles, tandis que, l'indice d'électrophilie définie en terme ω de Parr, exprime l'énergie de stabilisation lorsque le système acquiert une charge additionnelle électronique de l'environnement.

Pour confirmer le comportement électrophile de ces nitrobenzènes, nous avons effectué une comparaison du pouvoir électrophile des cinquante toxines avec le pouvoir électrophile de certains acides nucléiques (AN). Les valeurs de potentiel électronique chimique, la dureté chimique, les indices d'électrophilie pour l'adénine A, guanine G, cytosine C, l'uracile U et T thymine sont données dans le **Tableau 5**.

Tableau. 5. Dureté, Potentiel chimique et électrophilie des acides nucléiques (NA)

Acide Nucléique	η (eV)	μ (eV)	ω (eV)
Adenine A	4,32	-4,44	**1,27**
Thymine T	4,66	-4,94	**2,62**
Guanine G	4,18	-4,49	**2,41**
Cytosine C	4,63	-4,77	**2,45**
Uracil U	4,82	-5,14	**2,74**

Il s'avère que les indices d'électrophilie des cinquante toxines (voir tableau 2) sont tous supérieurs à ceux des acides nucléiques (voir **Tableau 5**). Par conséquent, la molécule toxique va agir comme un électrophile (accepteur d'électrons), tandis que l'acide nucléique (NA) agira comme un nucléophile (donneur d'électrons) lors de l'interaction toxine-NA. Le coefficient positif obtenu pour l'indice de l'électrophilie de Parr ω comme descripteur de la chimie quantique dans le meilleur modèle, appuie le concept que la toxicité des dérivés de nitrobenzène augmente avec l'augmentation de leur capacité d'accepter des électrons, indiquant, que le transfert d'électrons a lieu de l'organisme aux toxines. Les mêmes constatations ont été obtenues pour la toxicité des polychloro-dibenzofuranes [**26**] et des composés aromatiques [**6**]. Par conséquent, les nitrobenzènes les plus toxiques sont prédits d'être caractérisés par un pouvoir électrophile fort (accepteurs forts d'électrons) et une lipophilie élevée.

Conclusion

La présente étude montre que les descripteurs de la chimie quantique, à savoir, l'énergie LUMO et l'indice d'életrophilie ω de Parr, en combinaison avec l'indice d'hydrophobicité « log *P* » facteur de transport, sont utiles pour la prédiction de la toxicité (log (IGC_{50}^{-1}) des nitrobenzènes vis-à-vis l'espèce aquatique *Tetrahymena pyriformis.* Le meilleur modèle QSAR (Eq. 11) est capable de décrire environ 87% de la variance de la toxicité expérimentale et pourrait être utilisé efficacement pour estimer la toxicité des dérivés du nitrobenzène pour lesquels les données expérimentales sont indisponibles. Notre étude montre que l'indice de l'électrophilie de Parr constitue le descripteur principal pour expliquer la toxicité de ces toxines, même si la contribution du descripteur E LUMO est également importante. En effet, les modèles QSAR élaborés révèlent que les nitrobenzènes les plus toxiques sont caractérisés par leur forte hydrophobie et un pouvoir électrophile élevé.

Références Bibliographiques

[1] M. Smiesko, E. Benfenati, Predictive Models for Aquatic Toxicity of Aldehydes Designed for various Model Chemistries. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 976.

[2] W. Karcher, and J. Devillers, SAR and QSAR in environmental chemistry and toxicology: Scientific tool or wishful thinking? In practical Applications of Quantitative Structure-Activity Relationships (QSAR) in environmental Chemistry and Toxicology (Karcher W and Devillers J, Eds). Kluwer Academic, Dordrecht, The Netherlands. **1990**, pp 1-12.

[3] T.W. Schultz, M.T..D.Cronin, J.D. Walker, Quantitative structure-activity relationships (QSARS) in toxicology: A historical perspective. *Theochem*. 2003, 622, 1.

[4] A.R. Katritzky, R. Petrukhin, D. Tatham, S. Basak, E. Benfenati, Interpretation of QSPR and QSAR relationships. *J. Chem. Inf. Comput. Sci*. **2001**, 41, 679.

[5] S. Arulmozhiraja, M. Morita, Structure-Activity Relationships for the Toxicity of Polychlorinated Dibenzofurans: Approach through Density Functional Theory-Based Descriptors. *Chem. Res. Toxicol*. **2004**, 17, 348.

[6] M.T.D. Cronin, N. Manga, J.R. Seward, G.D. Sinks, and T.W. Schultz, Parametrization of Electrophilicity for the prediction of the Toxicity of Aromatic Compounds. *Chem. Res. Toxicol.* **2001**, 14**,** 1498.

[7] B.C.J. Zoeteman, K. Harmsen, , J.B.H.J. Linders, C.F.H. Morra, and W. Slooff, Persistent organic polluants in river water and ground water of the Netherlands. *Chemosphere*, **1980**, 9, 231.

[8] R. Purdy, in QSAR, (Ed.Turner JE, Williams MW, Schultz TW and Kwaak NJ), U.S.Dept of Energy. Oak Rdge, **1988**, pp. 99.

[9] D.W. Roberts, An analysis of published data on fish toxicity of nitrobenzenes and aniline derivatives. In QSAR in Environmental Toxicology- II (Kaiser, KLE, Ed.), D. Reidel, Dordrecht, The Netherlands, **1987**, pp 295-308.

[10] J.C. Dearden, M.T.D. Cronin, T.W. Schultz, D.T. Lin, QSAR Study of the Toxicity of Nitrobenzenes to *Tetrahymena Pyriformis. Quant. Struct-Act. Relat*. **1995**, 14, 427.

[11] J.W. Deneer, T.L. Sinnige, W. Seinen, and J.L.M. Hermens, ,. Quantitative structure-activity relationships for the toxicity and bioconcentration factor of nitrobenzene derivatives to wards the guppy (Poecilia reticulate) *Aquat.Toxicol*. **1987**, 10, 115.

[12] J.W.V. Deneer, C.J. Leeuwen, W. Seinen, J.L. Diepeveen, and J.L.M. Hermens, QSAR study of the toxicity of nitrobenzene derivatives towards *Daphnia magna*, *Chlorella pyrenoidosa*, and Photobacterium phosphoreum. *Aquat. Toxicol*, **1989**.15, 83.

[13] G.D. Veith, and O. G. Mekenyan, A QSAR approach for estimating the aquatic toxicity of

soft electrophiles. *Quant Struct Act Relat*. **1993**, 12, 349.

[14] O. Mekenyan, D.W. Roberts, and W. Karcher, Molecular orbital parameters as predictors of skin sensitization potential of halo-and pseudohalobenzenes acting as SNAR electrophiles. *Chem Res Toxicol*. **1997**, 10, 994.

[15] X. Yuan, G. Lu, and P. Lang, QSAR study of the toxicity of nitrobenzenes to river bacteria and Photobacterium phosphireum. *Bull Eviron Contam Toxicol*. **1997**, 58,123.

[16] Y.H. Zhao, X. Yuan, G-D. Ji, L.X. Sheng and L.S. Wang, Quantitative structure-activity relationships of nitroaromatic compounds to four aquatic organisms. *Chemosphere*. **1997**, 34, 1837.

[17] A.K. Debnath, R.L.L. Compadre, G. Debnath, A.J. Shusterman, and C. Hansch, Structure-activity relationship of mutagenic aromatic and heteroaromatic nitrocompounds. Correlation with molecular orbital energies and hydrophobicity. *J. Med. Chem.* **1991**, 34, 786.

[18] P.Z. Lang, X.F. Ma, G.H. Lu, Y. Wang, and Y. Bian.,. QSAR for the acutr toxicity of nitroaromatics to the carp (*Cyprinus carpio*). *Chemosphere*, **1996**, 32, 1547.

[19] T.W. Schultz, Structure –Toxicity Relationships for Benzenes Evaluated with *Tetrahymena Pyriformis*. *Chem. Res. Toxicol*. **1999**, 12,1262.

[20] J.J.P. Stewart, MOPAC Program Package, Quantum Chemistry Program Exchange **1989**.

[21] M.T.D. Cronin, B.W. Gregory, and T.W. Schultz, Response surface-based analyses of nitrobenzene toxicity to *Tetrahymena Pyriformis*. *Chem. Res. Toxicol*. **1998,** 11, 902.

[22] K. Fukui, T. Yonezawa, and C. Nagata, Theory of substitution in conjugated molecules. *Bull. Chem. Soc. Jpn*. **1954**, 27, 423.

[23] R.G. Parr, L.V. Szentpaly, S. Liu, Electrophilicity index. *J. Am. Chem. Soc*. **1999**, 121,1922.

[24] ACD/Labs, Release 12**, 2009**, http://www.acdlabs.com

[25] MINITAB, State College, PA *Minitab*, Inc **2006**.

[26] U. Sarkar, J. Padmanabhan, R. Parthasarathi, V. Subramanian, P.K. Chattaraj, Toxicity analysis of polychlorinated dibenzofurans through global and local Electrophilicities. *J. Mol. Struct. Theochem*. **2006**, 758, 119.

Application II

Estimation de la toxicité des nitro-aromatiques :

Modèles QSAR pour les mécanismes « *Redox Cycling* » et « *Nucleophilic Attack* »

Résumé

Des modèles QSAR prédictifs ont été élaborés pour l'estimation de la toxicité d'une série constituée de ***72*** *composés nitro-aromatiques vis à vis de l'espèce aquatique Tetrahymena pyriformis en tenant compte du type de mécanisme. La série totale a été divisée en deux sous-séries. La sous-série A constituée de 42 composés dont leur toxicité est expliquée par le mécanisme « Redox Cycling » et la sous-série B formée par 30 composés dont leur toxicité est expliquée par l'attaque nucléophile. Les modèles QSAR obtenus montrent que l'énergie de l'orbitale moléculaire occupée par un seul électron* ***Esomo****, et la densité de spin de Mulliken* ***MSD*** *en combinaison avec le paramètre d'hydrophobie* ***log P*** *sont utiles pour expliquer la toxicité de la sous-série A; tandis que l'indice d' électrophilie de Parr,* ***ω****, en combinaison avec le paramètre de lipophilie* ***log P*** *et la polarisabilité sont efficients pour la rationalisation de la toxicité de la sous-série B. La présente étude montre l'importance de la prise en considération du type de mécanisme dans la modélisation QSAR de la toxicité des composés nitro-aromatiques.*

Introduction

L'environnement est régulièrement exposé à des substances chimiques issues de processus biologiques, industriels et naturels. Toutefois, la voie expérimentale, en utilisant des tests sur les animaux est généralement un processus long, coûteux et techniquement difficile et même éthiquement discutable **[1].** Pour ces raisons, les techniques de modélisation capables d'estimer les propriétés biologiques d'une manière plus économique, plus rapide et plus facile sont devenues d'intérêt potentiel. L'objectif majeur de l'analyse QSAR est de trouver une relation mathématique entre l'activité et les descripteurs liés à la structure de la molécule **[2, 3]**. Ces études peuvent éclairer et même élucider le mécanisme par lequel l'activité en question est liée à la structure chimique.

Les composés nitro-aromatiques et leurs nombreux dérivés sont utilisés comme des intermédiaires en synthèse organique, notamment pour la fabrication de l'aniline, la benzédine, d'explosifs et des colorants et dans d'autres applications industrielles. Les déchets de composés nitro-aromatiques sont facilement générés au cours de la fabrication, le stockage, et le transport, ce qui conduit à un danger potentiel pour l'homme et l'environnement. Un certain nombre d'études ont montré que les nitro-composés, leurs métabolites de transformation et les sous-produits de la synthèse, sont nocifs pour la biosphère en raison de leur toxicité **[4-6]**. La toxicité des nitro-aromatiques a été assez largement examinée par plusieurs groupes de chercheurs utilisant différentes méthodologies **[7-13]**. Il a été conclu que la toxicité des nitro-aromatiques peut être considérée à partir de deux points de vue:

(i) la réduction du groupement nitro NO_2.

(ii) la tendance à agir comme électrophile dans une substitution nucléophile aromatique (SNAr) **[13-15].**

La réduction d'un groupe nitro peut s'effectuer par la réduction en une seule étape avec une enzyme telle que la nitro-réductase c'est ce qu'on appelle « *Redox-Cycling* ». Pour les composés nitro-aromatiques qui peuvent exercer le stress oxydant (*oxydative stress*), responsable de plusieurs maladies (Alzeimer, cancer, vieillissement des cellules, maladies

cardiovasculaires …), le radical anion nitro-aromatique formé après réduction est oxydé en un composé parent tout en formant superoxyde ($O_2^{-\cdot}$), cela conduit à la génération de peroxyde d'hydrogène (H_2O_2) et les radicaux hydroxyles qui sont des oxydants très réactifs. A partir de ces considérations, Smith et al. [16] ont proposé un descripteur de la chimie quantique pour quantifier ce mécanisme, à savoir, l'énergie de l'orbitale moléculaire occupée par un seul électron (E_{SOMO}) de l'anion radical. Katritzky et al. [3] ont également montré l'utilité du descripteur E_{SOMO} en présence d'autres descripteurs générés par le programme CODESSA-PRO pour expliquer le mécanisme succession Redox (*Redox Cycling)* d'une série de nitro-aromatiques. L'utilisation d'autres descripteurs pour la modélisation de la toxicité de ces composés a été proposée par Agrawal et Khadikar [17] qui ont élaboré des modèles QSAR basés uniquement sur des descripteurs topologiques. Dearden et al. [18] ont utilisé les descripteurs A_{max} et E_{LUMO} pour expliquer le mécanisme attaque Nucléophile « *Nucleophilic Attack* » d'une série de nitrobenzènes. Récemment, en utilisant une technique hiérarchique (Hit QSAR), Artemenko et al. [19] ont constaté que les paramètres de substituants caractérisant l'hypophilie et les interactions électrostatiques sont les facteurs les plus importants qui déterminent la toxicité des composés nitro-aromatiques.

L'objectif de cette application est d'établir des modèles QSAR fiables pour la prédiction de la toxicité des composés nitro-aromatiques, d'éclairer et d'interpréter le mécanisme de la toxicité de ces produits : Succession-Redox et attaque nucléophile.

-La première sous-série est constituée de 42 composés dont leur toxicité est expliquée par le mécanisme *Redox Cycling* (mécanisme A) [19]. Les descripteurs utilisés pour expliquer ce mécanisme sont l'énergie de l'orbitale moléculaire occupée par un seul électron, E_{SOMO}, la densité de spin Mulliken, **MSD**, sur l'atome d'azote du radical anion, le moment dipolaire, **DM,** et le paramètre d'hydrophobie log *P*.

-La seconde sous-série est constituée de 30 composés dont leur toxicité est expliquée par le mécanisme *Nucleophilic Attack* (mécanisme B). Les descripteurs utilisés pour expliquer ce

mécanisme sont l'indice d'électrophilie de Parr, ω, l'énergie de l'orbitale moléculaire inoccupée la plus basse, E_{LUMO}, la polarisabilité, α, et le paramètre d'hydrophobie log *P*.

1. Base de données et méthodes de calcul

1.1. Base de données

Les données de toxicité des nitro-aromatiques ont été prises de la référence [19]. IGC_{50} signifie la concentration en mM provoquant 50% d'inhibition de la croissance de *Tetrahymena pyriformis*. La base de données expérimentale a été divisée en trois sous-séries :

- Sous-série **A** constituée de 42 molécules (33 composés pour la série d'apprentissage et 9 molécules pour la série de test), leur toxicité est régie par le mécanisme *Redox-Cycling* (**Tableau 1**).
- Sous-série **B** constituée de 30 composés (24 composés pour la série d'apprentissage et 6 molécules restantes pour la série de test), leur toxicité est expliquée par le mécanisme (*Nucleophilic Attack*) (**Tableau 2**).
- Sous-série **C** constituée de 7 molécules pour lesquelles le mécanisme est **inconnu** ou **incertain** (**Tableau 3**). Cette sous- série a été utilisée pour une validation externe du modèle de toxicité et la prédiction du mécanisme (A vs B).

1.2. Calculs de la chimie quantique

Tous les calculs ont été effectués à l'aide du programme informatique GAUSSIAN 03 [20]. La densité fonctionnelle hybride à trois paramètres, B3LYP/6-31G*, a été utilisée pour l'optimisation des structures moléculaires.

-L'énergie de l'orbitale occupée par un seul électron (E_{SOMO}), le moment dipolaire (DM) et la densité de spin de Mulliken (MSD) ont été calculés en utilisant l'anion radical de chaque molécule de la sous série A.

- Les valeurs d'hydrophobie (log *P*) et la polarisabilité ont été calculées en utilisant le programme ACD/Labs [21].

1.3. Analyse Statistique

Les modèles QSAR ont été générés en utilisant la procédure de régression multilinéaire du logiciel MINITAB (version 15) [**22**]. Les modèles sont évalués par la valeur R^2, le $R^2_{ajusté}$, la valeur de SD (racine de la moyenne des carrés des erreurs) et la valeur F (Fischer statistique), le nombre d'observations N est également noté. Le pouvoir prédictif est évalué par le coefficient de la validation croisée (R^2_{CV}) et par la validation externe.

2. Résultats

2. 1. Modèles QSAR pour le mécanisme A « *Redox Cycling* »

Plusieurs modèles QSAR linéaires impliquant un, deux, trois et quatre descripteurs ont été établis.

Les résultats obtenus pour la sous-série A ont donné les meilleurs modèles suivants :

- *Modèle à deux paramètres*

$$\log (1/ IGC_{50}) = = 0,06 + 0,54 \log P - 0,81 E_{SOMO} \quad (1)$$

$R^2 = 0,76$; $R^2_{adj} = 0,74$; $R^2_{CV} = 0,71$; $F = 47,76$; $SD = 0,39$; $N = 33$

- *Modèle à trois paramètres*

$$\log (1/ IGC_{50}) = - 0,11 + 0,50 \log P - 1,00 E_{SOMO} + 2,84 MSD \quad (2)$$

$R^2 = 0,77$; $R^2_{adj} = 0,74$; $R^2_{CV} = 0,69$; $F = 32,35$; $SD = 0,39$; $N = 33$

- *Modèle à quatre paramètres*

$$\log (1/ IGC_{50}) = - 0,30 + 0,41 \log P - 1,36 E_{SOMO} + 4,85 MSD + 0,05 DM \quad (3)$$

$R^2 = 0,78$; $R^2_{adj} = 0,74$; $R^2_{CV} = 0,69$; $F = 24,8$; $SD = 0,39$; $N = 33$

Une amélioration significative de R^2 est obtenue par la combinaison des quatre paramètres, à savoir l'énergie SOMO, le coefficient de partage log *P*, le moment dipolaire, DM, et la densité de spin de Mulliken, MSD. Néanmoins, les valeurs de R^2cv et SD restent les mêmes pour les modèles à trois et quatre paramètres, ceci dit que le moment dipolaire n'a pas une contribution significative dans le modèle QSAR présenté par Eq.3. Deux composés sont caractérisés par un résiduel élevé (> 0,7) sont, en l'occurrence, 3,4-

dinitrophénol, le 2,6-dinitrophénol et dans les corrélations données par Eqs.(2, 3). L'élimination de ces 2 molécules donne les modèles QSAR des équations. (4, 5).

$$\log (1/ IGC_{50}) = 0,18 + 0,51 \log P - 0,85 E_{SOMO} \quad (4)$$
$R^2 = 0,80$; $R^2adj = 0,79$; $R^2_{CV} = 0,75$; $F = 57,85$; $SD = 0,36$; $N = 31$

$$\log (1/ IGC_{50}) = - 0,05 + 0,46 \log P - 1,13 E_{SOMO} + 3,97 MSD \quad (5)$$
$R^2 = 0,82$; $R^2adj = 0,80$; $R^2_{CV} = 0,75$; $F = 40,71$; $SD = 0,35$; $N = 31$

Les valeurs des descripteurs et de la toxicité prédite sont données dans le **Tableau 1**.

Tableau. 1. Valeurs des descripteurs et des toxicités pour les composés de la sous-série A

Composé	log *P*	MSD	E_{SOMO} (ev)	Toxicité observée	Toxicité prédite	Résidus
2-Nitrobenzamide	0,170	0,078	0,842	-0,720	-0,614	-0,106
2-Nitrobenzaldehyde	1,740	0,092	0,539	0,170	0,507	-0,337
4-Nitrobenzaldehyde	1,560	0,091	0,208	0,200	0,794	-0,594
3,4-Dinitrophenol	2,170	0,060	0,176	0,270	0,988	-0,718
4-Bromonitrobenzene	2,550	0,159	0,897	0,380	0,741	-0,361
4-Chloronitrobenzene	2,600	0,163	0,944	0,430	0,726	-0,296
2,3-Dinitrophenol	2,310	0,063	0,196	0,460	1,041	-0,581
4-Nitroanisole	2,030	0,193	1,437	0,540	0,026	0,514
2,6-Dinitrophenol	1,870	0,143	-0,059	0,540	1,445	-0,905
4-Nitrobenzonitrile	1,190	0,127	0,272	0,570	0,694	-0,124
2-Chloronitrobenzene	2,340	0,240	1,008	0,680	0,840	-0,160
Ethyl-4-nitrobenzoate	2,330	0,103	0,373	0,710	1,009	-0,299
2-Bromonitrobenzene	2,520	0,166	0,958	0,750	0,686	0,064
1,3-Dinitrobenzene	1,620	0,101	0,256	0,890	0,807	0,083
2,5-Dinitrophenol	1,840	0,097	-0,172	0,950	1,376	-0,426
2-Nitrobenzonitrile	1,330	0,129	0,460	1,080	0,554	0,526
2,4-Dinitrophenol	1,740	0,099	0,477	1,080	0,604	0,476
3,4-Dichloronitrobenzene	3,160	0,153	0,605	1,160	1,327	-0,167
1,2-Dinitrobenzene	1,840	0,061	0,114	1,250	0,910	0,340
1,4-Dinitrobenzene	1,370	0,084	-0,217	1,300	1,159	0,141
2,5-Dibromonitrobenzene	3,120	0,151	0,526	1,370	1,390	-0,020
4-Butoxynitrobenzene	3,620	0,193	1,418	1,420	0,779	0,641
2,4,6-Trichloronitrobenzene	3,410	0,122	0,396	1,430	1,555	-0,125
2,3,4-Trichloronitrobenzene	3,440	0,134	0,402	1,510	1,610	-0,100
2,4,5-Trichloronitrobenzene	3,490	0,135	0,330	1,530	1,718	-0,188
2,3,4,5-Tetrachloronitrobenzene	3,940	0,127	0,104	1,780	2,149	-0,369
2,3,5,6-Tetrachloronitrobenzene	3,730	0,112	0,121	1,820	1,974	-0,154

2,4,6-Trichloro-1,3-dinitrobenzene	2,720	0,064	-0,373	2,190	1,877	0,313
1,2-Dinitro-4,5-dichlorobenzene	3,200	0,051	-0,485	2,210	2,173	0,037
6-Bromo-1,3-dinitrobenzene	2,020	0,085	-0,052	2,310	1,275	1,035
4,6-Dichloro-1,2-dinitrobenzene	3,080	0,049	-0,442	2,420	2,061	0,359
2,4,5-Trichloro-1,3-dinitrobenzene	2,840	0,077	-0,428	2,590	2,046	0,544
2,3,5,6-Tetrachloro-1,4-dinitrobenzene	2,920	0,051	-0,994	2,740	2,619	0,121
Série de test						
3-Nitrobenzamide	0,680	0,132	0,931	-0,190	-0,265	0,075
3-Nitrobenzaldehyde	1,750	0,147	0,724	0,140	0,520	-0,380
4-Nitrobenzamide	0,820	0,118	0,558	0,180	0,165	0,015
3-Nitroacetophenone	1,490	0,107	0,784	0,320	0,174	0,146
3-Nitrobenzonitrile	1,170	0,159	0,580	0,450	0,464	-0,014
3-Nitroanisole	2,170	0,180	1,281	0,670	0,215	0,455
3-Chloronitrobenzene	2,640	0,167	0,899	0,730	0,812	-0,082
4-Nitrophenetole	2,560	0,193	1,429	0,830	0,279	0,551
2,4-Dichloronitrobenzene	3,000	0,144	0,649	0,990	1,168	-0,178

La **Figure 1a** montre la corrélation linéaire entre les valeurs observées et les valeurs prédites de la toxicité, pour la série d'apprentissage et de test, obtenues en utilisant l'équation (5).

La **Figure 1b** montre les résidus en fonction des valeurs expérimentales de la toxicité obtenues avec ce modèle.

Les résultats de la validation externe pour le mécanisme « A »sont donnés dans le **Tableau 2.**

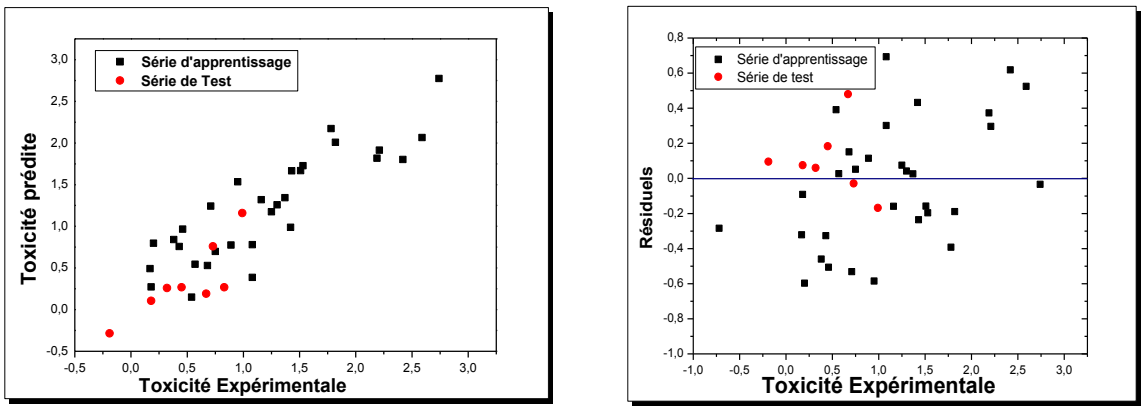


Figure.1a **–Toxicité** Prédite vs. Toxicité mesurée en utilisant le modèle Eq. (5)

b -Residus vs. toxicité observée

Tableau. 2. Résultats de la validation externe pour le mécanisme "A"

	Série d'apprentissage				Série de test			Critères de Tropsha			
	n	R^2	R^2_{cv}	SD	n	R^2	SD	R^2_0	k	$R^2 - R_0^2 / R^2$	$\lvert R^2 - R_0^2 \rvert$
Eq.(4)	33	0,80	0,75	0,36	9	0,63	0,14	0,66	0,79	-0,047	0,03
Eq.(5)	33	0,82	0,75	0,35	9	0,68	0,13	0,65	0,77	0,044	0,03

2. 2. Modèles QSAR pour le mécanisme B "*Nucleophilic Attack"*

Les résultats obtenus en utilisant la méthode de régression linéaire multiple (MLR) pour les composés de la sous-série B ont donné les meilleurs modèles suivants :

- **Modèle à deux paramètres**

$$\log (1/ IGC_{50}) = - 3,74 + 1,34\ \omega + 0,470 \log P \quad (6)$$

$R^2 = 0,75$; $R^2_{adj} = 0,73$; $R^2_{CV} = 0,71$; F= 35,40 ; SD = 0,42 ; N= 24

- **Modèle à trois paramètres**

$$\log (1/ IGC_{50}) = - 4,06 + 0,84\ \omega + 0,30 \log P + 0,11\ \alpha \quad (7)$$

$R^2 = 0,80$; $R^2_{adj} = 0,77$; $R^2_{CV} = 0,66$; F= 26,00 ; SD = 0,41 ; N= 24

L'analyse des résiduels a montré que les molécules : le 2-Nitrobenzamide et le 6-Bromo, 1,3-dinitrobenzène ont des résiduels élevés. L'élimination de ces 2 composés a donné le modèle QSAR suivant :

$$\log (1/ IGC_{50}) = - 4,04 + 0,61\ \omega + 0,12 \log P + 0,17\ \alpha \quad (8)$$

$R^2 = 0,82$; $R^2_{adj} = 0,79$; $R^2_{CV} = 0,72$; F= 28,00 ; SD= 0,33 ; N= 22

Il s'avère qu'une amélioration significative de la qualité du modèle QSAR a été obtenue par la combinaison des trois paramètres, à savoir, l'indice d'électrophilie, ω, de Parr, le coefficient de partage log P et la polarisabilité. Puisque E_{LUMO} et l'indice d'électrophilie sont corrélés, le meilleur modèle est seulement limité à trois descripteurs. Les valeurs des descripteurs et les valeurs de toxicité prédite en utilisant le meilleur modèle (Eq. (8)) sont données dans le **Tableau 3**.

Tableau. 3. Valeurs des descripteurs et des toxicités (observées et prédites) pour les composés de la sous-série B (*Nucleophilic-Attack*)

Nom du composé	E_{HOMO} (eV)	E_{LUMO} (eV)	μ (eV)	η (eV)	ω (eV)	log *P*	α	Toxicité Observée	Toxicité Prédite	Résiduel
2-Nitrobenzamide	-6,626	-2,453	-4,539	4,173	2,469	0,170	16,540	-0,720	0,298	-1,018
2-Nitrobenzaldehyde	-7,014	-2,855	-4,935	4,159	2,928	1,740	15,670	0,170	0,619	-0,449
5-Hydroxy-2-nitrobenzaldehyde	-7,033	-2,723	-4,878	4,310	2,761	1,750	16,420	0,330	0,646	-0,316
2,3-Dinitrophenol	-7,225	-3,044	-5,135	4,181	1,790	2,310	16,340	0,460	0,107	0,353
2,6-Dinitrophenol	-7,567	-3,393	-5,480	4,175	3,597	1,370	16,340	0,540	1,096	-0,556
2,6-Dichloro-4-nitrophenol	-7,258	-2,697	-4,978	4,561	2,716	2,940	17,620	0,660	0,965	-0,305
1,3-Dinitrobenzene	-8,412	-3,134	-5,773	5,278	3,157	1,490	15,590	0,890	0,715	0,175
2,5-Dinitrophenol	-7,553	-3,394	-5,474	4,159	3,602	1,750	16,340	0,950	1,145	-0,195
2,4-Dinitrophenol	-7,683	-2,830	-5,257	4,853	2,847	1,670	16,340	1,100	0,675	0,425
1,2-Dinitrobenzene	-7,589	-3,211	-5,400	4,379	3,330	1,690	15,590	1,250	0,844	0,406
1,4-Dinitrobenzene	-8,349	-3,493	-5,921	4,855	3,610	1,470	15,590	1,300	0,989	0,311
2,4,6-Trichloronitrobenzene	-7,648	-2,900	-5,274	4,748	2,929	3,690	18,820	1,430	1,389	0,041
2,3,4-Trichloronitrobenzene	-7,630	-2,932	-5,281	4,698	2,968	3,610	18,820	1,510	1,403	0,107
5-Methyl-1,2-dinitrobenzene	-7,493	-3,119	-5,306	4,374	3,218	2,300	17,500	1,520	1,174	0,346
2,4,5-Trichloronitrobenzene	-7,540	-2,991	-5,266	4,549	3,048	3,470	18,820	1,530	1,435	0,095
4,6-Dinitro-2-cresol	-7,413	-3,228	-5,321	4,185	3,383	2,200	18,250	1,720	1,390	0,330
2,3,4,5-Tetrachloronitrobenzene	-7,596	-3,108	-5,352	4,488	3,191	3,930	20,760	1,780	1,907	-0,127
2,3,5,6-Tetrachloronitrobenzene	-7,414	-3,048	-5,231	4,366	3,133	4,38	20,760	1,820	1,926	-0,106
2,4,6-Trichloro-1,3-dinitrobenzene	-7,643	-3,397	-5,520	4,246	3,589	2,970	21,410	2,190	2,145	0,045
1,2-Dinitro-4,5-dichlorobenzene	-7,933	-3,566	-5,749	4,367	3,785	3,200	19,470	2,210	1,963	0,247
6-Bromo-1,3-dinitrobenzene	-7,915	-3,261	-5,588	4,655	3,354	2,020	18,640	2,310	1,417	0,893
4,6-Dichloro-1,2-dinitrobenzene	-7,676	-3,495	-5,585	4,181	3,730	3,080	19,470	2,420	1,915	0,505
2,4,5-Trichloro-1,3-dinitrobenzene	-7,794	-3,465	-5,629	4,329	3,660	2,840	21,410	2,590	2,173	0,417
2,3,5,6-Tetrachloro-1,4-dinitrobenzene	-7,696	-3,830	-5,763	3,867	4,295	2,920	23,350	2,740	2,900	-0,160

Série de Test										
2-Nitrobenzoïc acid	-7,211	-2,665	-4,938	4,546	2,682	1,560	15,740	-1,640	0,459	-2,099
3,5-Dinitrobenzyl alcohol	-8,052	-3,051	-5,551	5,001	3,081	0,590	18,150	0,530	0,996	-0,466
6-Methyl-1,3-dinitrobenzene	-8,116	-2,986	-5,551	5,130	3,003	1,98	17,500	0,870	1,004	-0,134
2,6-Dinitro-4-cresol	-7,289	-3,304	-5,296	3,985	3,519	2,330	18,250	1,230	1,489	-0,259
2,6-Dibromo-4-nitrophenol	-7,129	-2,660	-4,894	4,469	2,680	3,570	19,840	1,360	1,396	-0,036
2,4-Chloro-6-nitrophenol	-6,976	-3,211	-5,093	3,765	3,445	3,410	17,620	1,750	1,466	0,284

La **Figure 2a** montre la corrélation linéaire entre la toxicté observée et les valeurs de la toxicité prédite obtenues en utilisant l'équation (8), de plus, les valeurs des résiduels de toxicité obtenues avec ce modèle en fonction des valeurs expérimentales sont présentées sur la **Figure 2b**.

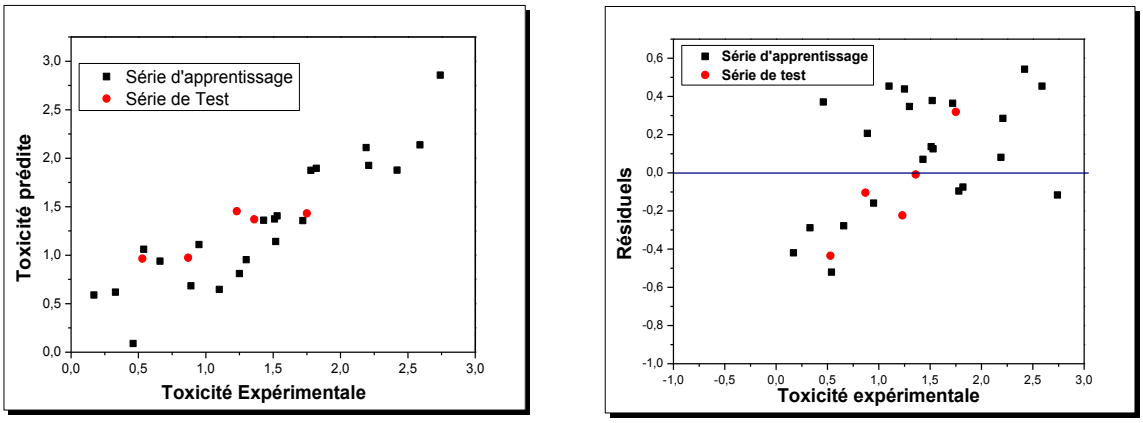


Figure. 2a - Toxicité Prédite vs Toxicité mesurée en utilisant le modèle (Eq.8)

b -Résiduels vs toxicité observée

les résultats de la validation externe pour les composés de la sous-série B sont présentés dans le **Tableau 4.**

Tableau. 4. Résultats de la validation externe du meilleur modèle pour le mécanisme B

Série d'apprentissage				Série de test			Critères deTropsha			
n	R^2	R^2_{cv}	SD	n	R^2	SD	R^2_0	k	$R^2 - R_0^2/R^2$	$\lvert R^2 - R_0^2 \rvert$
22	0,82	0,77	0,28	6	0,89	0,14	0,56	0,67	0,37	0,33

2. 3. Validation externe et prédiction du mécanisme

Afin de vérifier le pouvoir prédictif du meilleur modèle QSAR élaboré pour les deux sous-séries A et B et de prévoir le mécanisme de la toxicité, une validation externe a été effectuée pour 7 composés choisis aléatoirement. Les valeurs de tous les descripteurs et la toxicité prédite en utilisant les meilleurs modèles pour les mécanismes A et B sont présentées dans le **Tableau 5**.

Tableau. 5. Validation des modèles QSAR Eq. (5, 8) et prédiction du mécanismes A/B

			Mécanisme A				Mécanisme B				
Composé	Toxicité Observée	log *P*	E_{SOMO} (ev)	MSD	Toxicité Prédite		ω	α	Toxicité Prédite		Mécanisme Prédit
1,2,3-Trichloro-5-Nitrobenezene	1,55	3,74	0,29	0,14	1,899	(-0,349)	3,10	18,82	1,4992	(0,051)	B
1,2,4-Trichloro-6-Nitrobenezene	1,68	3,53	0,30	0,13	1,751	(-0,071)	3,09	18,82	1,4679	(0,212)	A/B
2,3-Dichloronitrobenezene	1,07	2,9	0,68	0,15	1,111	(-0,041)	2,80	16,88	0,8856	(0,184)	A
2,4,6-Trimethylnitrobenezene	0,86	3,33	1,26	0,19	0,812	(0,048)	2,14	18,74	0,8508	(0,009)	B
2,5-Dichloronitrobenzene	1,13	2,95	0,60	0,15	1,225	(-0,095)	2,88	16,88	0,9404	(0,190)	A/B
3-Bromonitrobenzene	1,03	2,52	0,86	0,16	0,773	(0,257)	2,71	16,05	0,644	(0,386)	A
3-Chloro-4-Fluoronitrobenzene	0,80	2,6	0,87	0,17	0,838	(-0,038)	2,76	14,94	0,4954	(0,305)	A

- Les valeurs entre parenthèses représentent la différence entre la toxicité observée et la toxicité prédite.

3. Discussion

Puisque la toxicité des composés nitro-aromatiques est expliquée par différents mécanismes (**Figure 3)** comme déjà mentionné, il est fortement recommandé de tenir compte du type de mécanisme dans l'élaboration des modèles QSAR de la toxicité.

Pour les composés de la sous-série A, les modèles QSAR élaborés (équations, (1-5)) révèlent que la toxicité pourrait s'expliquer par un certain nombre de facteurs électroniques et de transport. L'énergie SOMO est importante dans la description de la réduction du groupe nitro qui explique le mécanisme « *Redox-Cycling* », la densité de spin de Mulliken (MSD), décrit la densité de spin de l'électron après la formation de l'ion radical sur le groupe nitro anion, tandis que l'hydrophobie, décrit par log *P*, est importante pour décrire le transport vers le site d'action. Pour mettre en évidence la contribution de chaque paramètre E_{SOMO}, MSD et log *P* dans la prédiction de la toxicité, nous avons étudié la relation entre ces paramètres et la toxicité (IGC_{50}^{-1}). Aucune régression linéaire simple n'a été obtenue avec un seul descripteur (voir ANNEXE) suggérant que c'est la synergie et la combinaison de plusieurs descripteurs qui peuvent expliquer un bon pourcentage de la variance de la toxicité expérimentale. Le développement des QSAR à deux paramètres a révélé que la combinaison de log *P* avec E_{SOMO} donne un modèle, avec des paramètres statistiques acceptables, pour prédire la toxicité de cette sous-série de composés nitro-aromatiques, alors que les meilleurs modèles à deux paramètres sont obtenus avec E_{SOMO} et log *P*, cela montre l'importance de l'indice d'hydrophobie dans ces modèles multilinéaires. Les modèles avec trois descripteurs montrent que l'inclusion de la densité de spin de Mulliken MSD améliore la qualité de ces modèles QSAR. En effet, il s'avère que la combinaison des trois paramètres augmente modérément le pouvoir explicatif du modèle QSAR donné par l'équation (3) (R^2=0,78, R^2_{CV}=0,69). L'élimination des trois valeurs aberrantes (atypiques) (Eq.5) donne une amélioration significative de la qualité de ce modèle QSAR (R^2=0,82, R^2_{CV}=0,75) ce qui confirme que ce modèle explique la toxicité des composés nitro-aromatiques qui présentent un mécanisme A, de manière significative et statistiquement satisfaisante.

Pour la sous-série B, les modèles QSAR élaborés (Eqs. 6-8) révèlent que la toxicité des nitro-aromatiques pourrait s'expliquer par un certain nombre de facteurs électroniques, électriques et de transport. L'électrophilie, tel que définie par E_{LUMO}, ω, sont importants dans la description de l'interaction électronique et la réactivité de ces toxines; la polarisabilité est une mesure de la capacité d'une molécule à répondre à un champ électrique et d'acquérir un moment dipolaire électrique, tandis que l'hydrophobie, comme exprimée par log *P* est importante pour décrire le transport vers le site d'action.

Nous avons étudié la relation entre ces paramètres et la toxicité (IGC_{50}^{-1}), bien que la polarisabilité et les indices d'électrophilie ω et E_{LUMO} se trouvent à être plus importants dans les modèles QSAR à un paramètre en comparaison avec log *P,* l'inclusion de ce dernier dans les modèles QSAR à deux et trois paramètres, est d'une grande importance. En effet, la combinaison de log *P* soit avec ω, α et /ou E_{LUMO} conduit à des modèles plus fiables. Le développement des QSAR à deux paramètres a révélé que soit ω ou E_{LUMO} en combinaison avec log *P* donne des modèles acceptables pour prédire la toxicité des nitro-aromatiques de la deuxième sous-série (sous-série B). La relation entre l'électrophilie et la polarisabilité suggère qu'un toxique avec une grande valeur d'électrophilie et de polarisabilité pourrait permettre une déformation importante de son nuage électronique de sorte qu'il pourrait bien interagir avec les récepteurs tels que l'ADN et les protéines, ceci dit qu'un composé toxique avec une valeur élevée d'électrophilie peut être plus toxique que le composé d'une valeur plus petite. Par conséquent, les nitro-aromatiques les plus toxiques sont prévus pour être caractérisés par une forte électrophilie (des accepteurs d'électrons forts) et une hydrophobie (lipophilie) élevée.

D'après ces résultats, il s'avère que, pour la sous-série A dont la toxicité est gouvernée par le mécanisme *Redox-Cycling*, l'inclusion de la densité de spin avec la E_{SOMO} et le paramètre d'hydrophobie donne une prédiction acceptable pour le modèle QSAR. Pour la sous-série B, l'indice d'électrophile, ω, est important dans la description de la réactivité de ces xénobiotiques au niveau du site d'action.

Les résultats ont montré aussi une différence entre la toxicité des mono et des dinitro-aromatiques pour les deux sous-séries, les composés qui contiennent deux groupements nitro présentent une toxicité (observée et prédite) plus élevée que les composés avec un seul groupement nitro. Il a été montré que les mono-nitro-aromatique avec un substituant halogène ont le potentiel d'être métabolisés in vivo en groupement nitroso [23]. Ces groupements sont suffisamment électrophiles et peuvent réagir avec les sites riches en électrons sur les macromolécules tels que l'ADN et les protéines. Tandis que les dinitro-aromatiques avec un groupe halogène sont capables de subir un mécanisme d'action qui implique l'addition d'un nucléophile au cycle aromatique (c'est à dire, qu'ils agissent comme électrophiles SNAr) ceci a été déjà montré par Roberts [24] et Mekenyan et al. [25]. Cette toxicité dépend aussi de la position des deux groupements nitro, le 1,3-di nitrobenzène est le moins toxique en le comparant avec 1,2 et 1,4 dinitrobenzène (**Tableau 1**). La présence du chlore rend ces molécules plus toxiques, ceci est confirmé avec les valeurs prédites trouvées ce qui confirme que ces modèles sont explicatifs.

L'analyse des résultats de la validation externe et la prédiction du mécanisme a montré que la toxicité de 2 molécules sur les 7 de la série de test est régie par le mécanisme B. La toxicité de3 molécules sur les 7 est expliquée par le mécanisme A et pour 2 molécules, il est difficile de favoriser un mécanisme par rapport à l'autre. On note que les molécules correspondantes au mécanisme B contiennent les atomes Cl, Br et F et ont des valeurs de polarisabilité plus faibles par rapport aux autres molécules de la sous-série B, ceci peut s'expliquer par le principe HSAB défini par Pearson [26].

En se basant sur les résultats de la validation externe, nous pouvons conclure que les modèles QSAR élaborés pour les deux mécanismes méritent d'être améliorés pour qu'ils deviennent plus prédictifs pour l'étude de la toxicité des nitro-aromatiques qui peuvent subir un mécanisme « succession redox » (*Redox Cycling*) et /ou attaque-nucléophile (*Nucleophilic-Attack*) bien que la modélisation de la toxicité des nitro-aromatiques n'est pas une tâche facile et les deux mécanismes peuvent parfois se chevaucher.

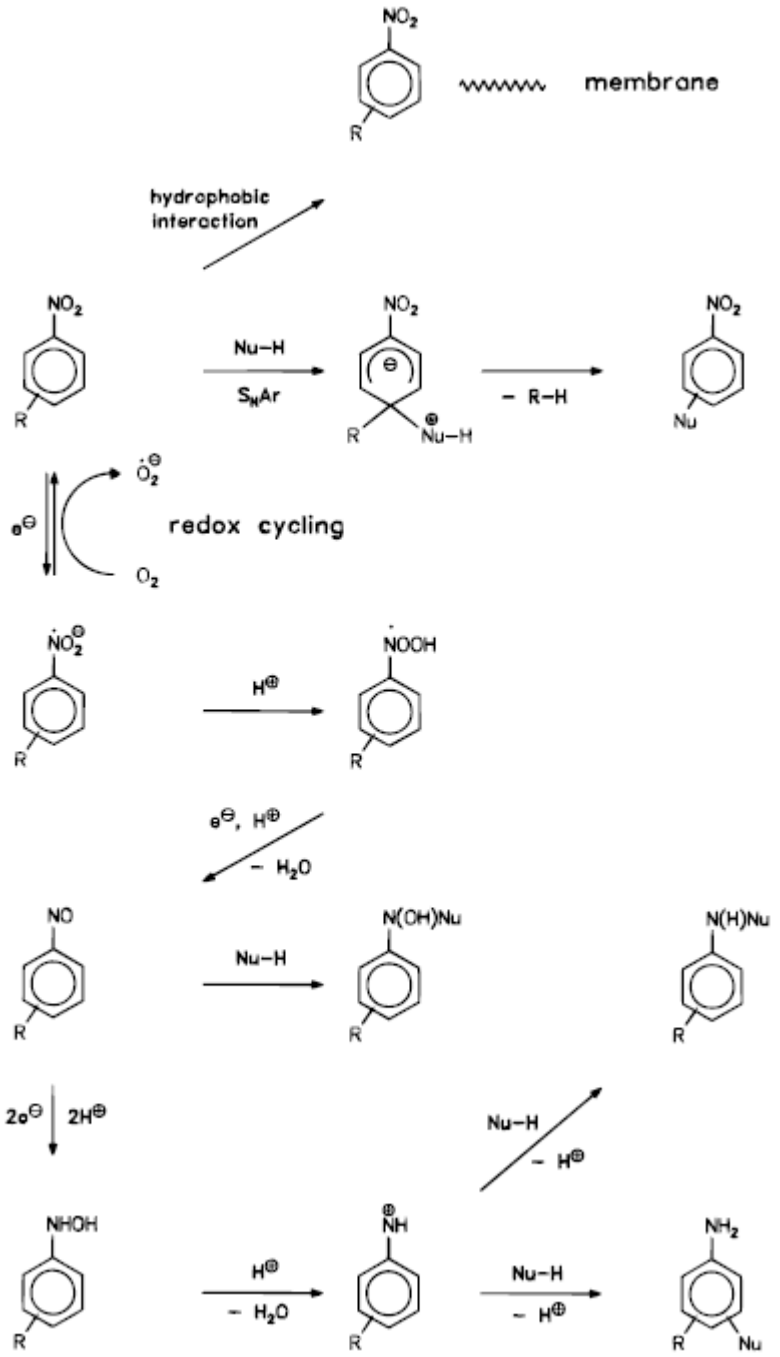


Figure. 3. Mécanismes possibles pour la toxicité des nitro-aromatiques

4. Conclusion

Dans la présente application, les deux mécanismes possibles (*Redox-Cycling* et Nucleophilic-Attack) de la toxicité des nitro-aromatiques sont examinés à l'aide des descripteurs quantiques en présence du coefficient de partage. Etant donné que la toxicité des composés nitro-aromatiques provient principalement de leur interactions avec les systèmes biologiques essentiellement par transfert d'électrons, l'utilisation de descripteurs qui expriment ce phénomène est d'une importance capitale dans l'élaboration des modèles QSAR. L'indice d'électrophilie de Parr a été utilisé pour présenter et quantifier ce transfert ainsi pour l'explication du mécanisme d'attaque nucléophile pour les nitro-aromatiques électrophiles. D'autre part, les résultats présentés ici suggèrent que la sensibilité des nitro-aromatiques pour le mécanisme "*Redox-Cycling*" peut être évaluée par l'énergie SOMO et l'indice de densité de spin de Mulliken, MSD, calculé par des méthodes de la chimie quantique. Enfin, cette étude représente une tentative d'élaborer des modèles QSAR pour la prédiction de la toxicité des composés nitro-aromatiques vis-à-vis *Tetrahymena pyriformis* en tenant compte du mécanisme d'action et en utilisant un nombre réduit de descripteurs simples et pertinents.

Références bibliographiques

[1] M. Hidalgo-Rodríguez, E. Fuguet, C. Ràfols, and M. Rosés, Modeling nonspecific toxicity of organic compounds to the Fathead Minnow fish by means of chromatographic systems, *Anal. Chem*, **2012**, 84, 3446-3452,

[2] A.R. Katritzky, V.S. Lobanov, and M. Karelson, QSPR: The correlation and quantitative prediction of chemical and physical properties from structure, *Chem. Soc. Rev*, **1995**, 24, 279-287.

[3] A.R. Katritzky, M. Karelson, and V.S. Lobanov, QSPR as a means of predicting and understanding chemical and physical properties in terms of structure, *Pur. App. Chem*, **1997**, 69, 245-248.

[4] A.H. Neilson, A.S. Allard, Environmental degradation and transformation of Organic chemicals, **2008**, CRC Press, Boca Raton, FL.

[5] S.S. Talmage, D.M. Opresko, C.J Maxwell, C.J. Welsh, F.M. Cretella, P.H. Reno, and F.B Daniel, Nitroaromatics munitions compounds: Environmental effects and screening values, *Rev. Environ. Contam. Toxicol*, **1999**, 161, 1-156.

[6] P.Y. Robidoux, C. Svendsen, J. Caumartin, J.Hawari, G. Ampleman, S. Thiboutot, J.M. Weeks, and G.I. Sunahara, Chronic toxicity of energetic compounds in soil determined using the earthworm (Eiseniaandrei) reproduction test, *Environ. Toxicol. Chem,* **2000**, *19*, 1764-1773.

[7] J.C. Dearden, M.T.D. Cronin, T.W Schultz, and D.T. Lin, QSAR study of the toxicity of nitrobenzenes to *Tetrahymena Pyriformis, Quant. Struct. Act. Relat,* **1995**, *14*, 427- 432.

[8] G.D. Veith, and O.G. Mekenyan, A QSAR approach for estimating the aquatic toxicity of soft electrophiles, *Quant. Struct.Act. Relat*, **1993**, *12*, 349-356.

[9] L. Pei-Zhen, M. Xun-Feng, L. Guang-Hua, W. Yi, and B. Yong, QSAR for the acute toxicity of nitroaromatics to the carp (*Cyprinuscarpio*), *Chemosphere,* **1996**, *32*, 1547-1552.

[10] Y.H Zhao, X. Yuan, G.D Ji, L.X. Sheng, and L.S. Wang, Quantitative structure-activity relationships of nitroaromatic compounds to four aquatic organisms, *Chemosphere,* **1997**, *34*, 1837-1844.

[11] A.K. Debnath, R.L.L. Compadre, G. Debnath, A. Shusterman, and C. Hansch, Structure-activity relationship of mutagenic aromatic and heteroaromatic nitrocompounds, Correlation with molecular orbital energies and

hydrophobicity, *J. Med. Chem,* **1991**, *34*, 786-797.

[12] O. Mekenyan, D,W. Roberts, and W. Karcher, Molecular orbital parameters as predictors of skin sensitization potential of halo-and pseudo halobenzenes acting as SNAR electrophiles, *Chem. Res. Toxicol,* **1997**,10, 994-1000.

[13] M.T.D. Cronin, B.W. Gregory, and T.W. Schultz, Quantitative structure-activity analysis of nitrobenzene toxicity to *Tetrahymena Pyriformis*, *Chem. Res. Toxicol,* **1998**, *11*, 902-908.

[14] M.T.D. Cronin, N. Manga, J.R. Seward, G.D. Sinks and T.W. Schultz, Parametrization of electrophilicity for the prediction of the toxicity of aromatic compounds, *Chem. Res. Toxicol,* **2001**, *14*, 1498-1505.

[15] A.R. Katritzky, P. Oliferenko, A. Oliferenko, A. Lomaka, and M. Karelson, , Nitrobenzene toxicity: QSAR correlations and mechanistic interpretations, *J. Phy. Org. Chem,* **2003**, *16*, 811-817.

[16] H. Schmitt, R. Altenburger, B. Jastorff, and G. Schuurmann, Quantitative structure-activity analysis of the Algae toxicity of nitroaromatic compounds, *Chem. Res. Toxicol,* **2000**,*13*, 441-450.

[17] W.K Agrawal, and P.V. Khadikar, QSAR prediction of toxicity of nitrobenzenes, *Bioorg. Med. Chem,* **2001**, *9*, 3035-3040.

[18] J.C. Dearden, M.T.D. Cronin, T.W. Schultz, D.T. Lin, QSAR Study of the Toxicity of Nitrobenzenes to *Tetrahymena Pyriformis. Quant. Struct. Act. Relat*, **1995**, 14, 427.

[19] A.G. Artemenko, E.N. Muratov, V.E. Kuz'min, N.N. Muratov, E.V. Varlamova, A.V. Kuz'mina, L.G Gorb, A. Golius, F.C. Hill, J. Leszczynski, and A. Tropsha, , QSAR analysis of the toxicity of nitroaromatics in *Tetrahymena pyriformis*: Structural factors and possible modes of action, SAR QSAR. *Environ. Res,* **2011**, *22*, 575-601.

[20] Gaussian 03, (**2003**) Revision B,01, Gaussian Inc, Pittsburgh.

[21] ACD/Labs, (**2009**) Release 12,http://www,acdlabs,com.

[22] MINITAB, (**2006**) State College, PA Minitab, Inc, http//www,minitab,com.

[23] D. W. Roberts, (1987) An analysis of published data on fish toxicity of nitrobenzenes and aniline derivatives. In *QSAR in Environmental Toxicology - II* (Kaiser, K. L. E., Ed.) pp 295-308, D.Reidel, Dordrecht, The Netherlands.

[24] D.W. Roberts, Linear free energy relationships for reactions of electrophilic halo- and pseudohalobenzenes, and their application in prediction of skin sensitization potential for SNAr electrophiles. *Chem. Res. Toxicol,* 1995, 8, 545-551

[25] O. Mekenyan, D.W. Roberts, and W. Karcher,) Molecular orbital parameters as predictors of skin sensitization potential of halo- and pseudohalobenzenes acting as SNAr electrophiles. *Chem. Res. Toxicol,*1997, 10, 994-1000.

[26] R.G. Pearson, Hard and soft acids and bases*, J. Am. Chem. Soc,* 1963, *85*, 3533-353.

Application III

Estimation de la toxicité aiguë des phénols halogénés en utilisant les paramètres d'hydrophobie et d'électrophilie

Résumé

Les phénols et en particulier les phénols halogénés représentent une partie importante des produits chimiques et sont connus comme des polluants aquatiques. Les relations quantitatives structure-toxicité (QSTR) sont utiles pour comprendre la corrélation existante entre la structure chimique et la toxicité des produits chimiques. Dans la présente étude, les toxicités aiguës de 45 phénols halogénés ont été estimées à l'aide des méthodes semi-empiriques AM1, PM3 et PM6. Des modèles QSTR ont été élaborés en utilisant la technique de régression linéaire multiple (MLR). Le pouvoir prédictif des modèles a été évalué par la validation croisée interne, la Y-randomisation et la validation externe. Leur domaine d'applicabilité a été également défini. Les résultats montrent que le meilleur modèle QSTR est obtenu avec la méthode AM1 (R^2 =0,91, R^2_{CV}=0,90, SD= 0,20 pour la série d'apprentissage et R^2=0.96, SD=0,13 pour la série de test). En outre, les 5 critères de« Tropsha» pour un modèle QSTR prédictif sont vérifiés. Les modèles QSTR obtenus ont été développés avec un nombre réduit de descripteurs pertinents et mis en évidence l'importance du facteur de transport exprimé par le paramètre de lipophilie log P et l'effet électronique exprimé par l'indice d'électrophilie de Parr dans l'interprétation et la prédiction de la toxicité des phénols halogénés.

Introduction

Une variété de composés organiques peut être des polluants environnementaux et des substances toxiques. Cependant, il est essentiel de protéger l'environnement et prévenir les intoxications professionnelles par l'étude de la toxicité de ces polluants. L'impact du danger potentiel des produits chimiques, un défi qui conforte les organismes de réglementation internationaux [1-4], peut être mesuré par des études expérimentales, mais cette approche est à la fois très coûteuse en argent et en temps [5]. Pour cette raison, une grande partie de l'effort a été mise sur l'utilisation des méthodes théoriques et numériques pour surmonter les contraintes de l'expérience. Une alternative est de s'appuyer sur les méthodes QSAR (structure-toxicité relation quantitative) pour établir des modèles décrivant une relation mathématique entre les caractéristiques structurelles des molécules chimiques et la toxicité particulière associée [6, 7]. Avec le développement rapide des techniques de calcul et des moyens informatiques, les descripteurs moléculaires peuvent être obtenus rapidement et avec précision pour un grand nombre de composés.

Les phénols représentent une partie importante des produits chimiques dans le monde. Ils ont été largement utilisés comme matériaux de base en pharmaco-chimie, en industrie et en agriculture [8]. Ils peuvent transpercer par l'air et de l'eau, avec une forte cancérogénèse et mutagenèse [9-10], ce qui provoque de grands dommages à l'environnement. Les risques pour l'environnement des composés phénoliques ont conduit à un large intérêt des chercheurs, et de nombreux travaux ont été réalisés pour élaborer des modèles QSTR ces dernières années [11-15]. Cronin et.al [14] ont obtenu des modèles QSAR pour une série de phénols en utilisant la méthode de régression linéaire multiple (MLR) et les réseaux des neurones (NN), les méthodes et les résultats obtenus montrent la capacité des modèles élaborés à prévoir deux mécanismes non-covalentes (narcoses polaires et découplants de la phosphorylation oxydative) et leur incapacité à estimer le mécanisme électrophile. Pacha et al. [16] ont étudié la toxicité d'une série de dérivés du phénol en utilisant des méthodes semi-empiriques et DFT. Cependant, les modèles QSAR élaborés impliquent plusieurs descripteurs moléculaires corrélés et les valeurs calculées de

l'électrophilie (voir les tableaux 2-5 de la Réf. [16]) sont erronées et insensées. Récemment, Ertürk et al. [17] ont étudié la toxicité d'une série de phénols vis-à-vis d'une algue marine, *Dunaliellatertiolecta,* utilisant les techniques MLR et NN. Leurs modèles QSTR, élaborés sur la base de descripteurs moléculaires calculés en utilisant les logiciels CODESSA [18] et DRAGON [19], ont donné des prédictions acceptables même si la signification physique des descripteurs concernés et leur corrélation avec la toxicité ne sont pas toujours clairement et rationnellement expliquée. Ertürk et al. [20] ont également modélisé la toxicité d'une série de phénols vis-à-vis *vulgaris Chlorella* en utilisant l'approche MLR, les résultats ont révélé que les modèles QSTR établis fournissent des prévisions acceptables (R^2= 0,84 ; SD =0,20) pour les narcotiques polaires et les découplants de la phosphorylation oxydative, mais ils sont incapables de prédire la toxicité des phénols réactifs présentant un mécanisme électrophile.

Les Phénols halogénés et particulièrement les chlorophénols, sont les plus répandus et la majorité des phénols halogénés sont généralement des narcotiques polaires [21]. Selon Schultz [14], il est difficile de modéliser l'ensemble des phénols à cause de l'existence de nombreux modes d'action. Pour cette raison, des modèles QSTRs ont été généralement développées en utilisant des composés d'une seule classe chimique (par exemple, les phénols halogénés) en supposant que les composés constituant cette classe agissent avec le même mode d'action.

Plusieurs études théoriques sur la prédiction de la toxicité des phénols halogènes peuvent être trouvées dans la littérature [22-24]. Cependant, plusieurs modèles QSTR élaborés ne répondent pas entièrement aux critères de l'OCDE (Organisation de coopération et de développement économique) concernant les principes de validation d'un modèle QSAR [25]. Par exemple, la validation externe n'est pas systématiquement effectuée ou les descripteurs du modèle sont fortement corrélés ce qui rend difficile de savoir le pouvoir prédictif externe. En outre, la « Y-Randomisation » et le domaine d'applicabilité du modèle ne sont pas toujours évalués et discutés. D'autre part, les phénols halogénés sont généralement des narcoses polaires, de sorte qu'il existe un transfert d'électrons entre la molécule toxique et l'organisme. Cet effet a été exprimé dans la

modélisation QSTR par différents descripteurs tels que l'énergie HOMO [24], E_{LUMO} [14,26], et la super-délocalisabilité électrophile A_{max} [26,28]. Etant donnée que E_{HOMO} exprime la tendance du système à fournir des électrons, à savoir le caractère nucléophile, ce descripteur ne peut pas être utilisé pour exprimer le comportement du caractère électrophile. D'autre part, E_{LUMO} et A_{max} sont des définitions approximatives et obsolètes de la notion d'électrophile. Récemment, Parr et al. [29] a proposé une définition précise et rigoureuse de l'électrophilie, notée ω, exprimant le gain d'énergie lorsqu'une molécule capte des électrons jusqu'à la saturation. L'indice d'électrophilie de Parr est d'un grand intérêt dans l'analyse de plusieurs et divers domaines de la chimie. En effet, il a été démontré que l'électrophilie peut donner des informations sur la structure, la stabilité, la réactivité, la toxicité [30]. Le descripteur ω a été utilisé pour l'étude de la toxicité des phénols chlorés par Chattaraj et al. [31]. Cependant, il a été utilisé seul et le facteur de pénétration, à savoir le paramètre de lipophilie log *P*, n'a pas été pris en compte. Dans le présent travail, les deux facteurs, électronique et de transport, seraient pris en compte.

Deux objectifs ont été ciblés pour la présente application :

i) Elaborer des modèles prédictifs pour la toxicité d'une série de phénols halogénés vis-à-vis *Tetrahymena pyriformis* en utilisant un nombre réduit de descripteurs pertinents et en respectant tout le protocole de la méthodologie QSAR (série d'apprentisasge + validation interne croisée + Y-randomization, série de test + validation externe + vérification des 5 critères de Tropsha, définition du domaine d'applicabilité)

ii) L'utilisation des méthodes semi-empiriques (AM1, PM3 et PM6) non couteuses en temps de calcul et analyse de l'influence de ces méthodes sur la qualité des modèles élaborés.

iii) Utilisation de la régression linéaire multiple donnant des modèles clairs et faciles à analyser et à interpréter.

1. Base de données et méthodes de calcul

1.1 Base de données :

La base de données se compose de 45 phénols halogénés prises de la référence [11] et numérotés dans le **Tableau 1**. Les données biologiques sont considérées comme de haute qualité, car elles se réfèrent aux mêmes grandeurs mesurées dans les mêmes conditions expérimentales (même protocole).

Tableau. 1. Numéro du composé (CAS), valeurs des descripteurs, des toxicités (observée et prédite) et des résiduels.

Composé				AM1			PM3			PM6		
	CAS. N	Tox.Exp	log*P*	ω	Tox.Préd	Résid	ω	Tox.Préd	Résid	ω	Tox .Préd.	Résid
4-fluorophenol	371-41-5	0,017	1,915	1,114	0,222	-0,205	1,165	0,269	-0,252	1,274	0,325	-0,308
2-chlorophenol	95-57-8	0,183	2,155	1,124	0,375	-0,192	1,119	0,350	-0,167	1,233	0,396	-0,213
2-bromophenol	95-56-7	0,33	2,355	1,160	0,549	-0,219	1,155	0,526	-0,196	1,261	0,549	-0,219
3-fluorophenol	372-20-3	0,381	1,915	1,162	0,301	0,080	1,205	0,327	0,054	1,245	0,280	0,101
2-chloro-5-methylphenol	615-74-7	0,393	2,654	1,113	0,640	-0,247	1,106	0,640	-0,247	1,153	0,549	-0,156
4-chlorophenol	106-48-9	0,545	2,485	1,105	0,532	0,013	1,107	0,537	0,008	1,255	0,612	-0,067
2-bromo-4-methylphenol	6627-55-0	0,599	2,854	1,141	0,800	-0,201	1,266	0,999	-0,400	1,223	0,768	-0,169
2,4-difluorophenol	367-27-1	0,604	1,947	1,286	0,524	0,080	1,346	0,553	0,051	1,419	0,565	0,039
2-chloro-4,5-dimethylphenol	1124-04-5	0,688	3,103	1,102	0,878	-0,190	1,116	0,933	-0,245	1,132	0,766	-0,078
4-chloro-2-methylphenol	1570-64-5	0,701	2,984	1,098	0,804	-0,103	1,103	0,840	-0,139	1,191	0,790	-0,089
2,6-dichlorophenol	87-65-0	0,735	2,627	1,272	0,889	-0,154	1,245	0,827	-0,092	1,412	0,932	-0,197
2,6-dichloro-4-fluorophenol	392-71-*2*	0,804	2,797	1,404	1,205	-0,401	1,394	1,150	-0,346	1,604	1,322	-0,518
3-chlorophenol	108-43-0	0,871	2,485	1,159	0,622	0,249	1,167	0,626	0,245	1,293	0,671	0,200

2,4-dichlorophenol	120-83-2	1,036	2,957	1,254	1,047	-0,011	1,231	1,010	0,026	1,426	1,138	-0,102
2,5-dichlorophenol	583-78-8	1,125	2,957	1,275	1,083	0,042	1,231	1,012	0,113	1,418	1,125	0,000
3-chloro-4-fluorophenol	2613-23-2	1,131	2,717	1,272	0,942	0,189	1,294	0,955	0,176	1,444	1,032	0,099
2,4,6-trichlorophenol	88-06-2	1,41	3,367	1,376	1,398	0,012	1,317	1,391	0,019	1,573	1,592	-0,182
4-bromo-2,6-dimethylphenol	2374-05-2	1,167	3,633	1,093	1,165	0,002	1,138	1,294	-0,127	1,143	1,077	0,090
2,3,5,6-tetrafluorophenol	769-39-1	1,167	2,068	1,547	1,219	-0,052	1,716	1,168	-0,001	1,687	1,045	0,122
4-chloro-3,5-dimethylphenol	88-04-0	1,201	3,483	1,139	1,156	0,045	1,078	1,113	0,088	1,087	0,908	0,293
2,3-dichlorophenol	576-24-9	1,276	2,837	1,270	1,006	0,270	1,228	0,933	0,343	1,402	1,034	0,242
4-bromo-6-chloro-2-methylphenol	7530-27-0	1,276	3,606	1,241	1,397	-0,121	1,261	1,457	-0,181	1,375	1,419	-0,143
2,4-dibromophenol	615-58-7	1,398	3,307	1,307	1,336	0,062	1,425	1,511	-0,113	1,444	1,359	0,039
Pentafluorophenol	771-61-9	1,638	2,213	1,825	1,572	0,066	1,882	1,499	0,139	1,836	1,356	0,282
3,4-dichlorophenol	95-77-2	1,745	3,167	1,251	1,162	0,583	1,215	1,118	0,627	1,407	1,224	0,521
4-bromo-2,6-dichlorophenol	3217-15-0	1,778	3,517	1,388	1,589	0,189	1,365	1,554	0,224	1,577	1,681	0,097
2,4,6-tribromophenol	118-79-6	2,03	3,917	1,442	1,907	0,123	1,530	2,042	-0,012	1,600	1,938	0,092
Pentachlorophenol	87-86-5	2,049	4,323	1,618	2,431	-0,382	1,475	2,214	-0,165	1,806	2,481	-0,432
2,4,5-trichlorophenol	95-95-4	2,097	3,577	1,399	1,641	0,456	1,333	1,544	0,553	1,572	1,706	0,391
2,3,5-trichlorophenol	933-78-8	2,373	3,577	1,410	1,661	0,712	1,328	1,537	0,836	1,572	1,707	0,666
3,4,5,6-tetrabromo-2-methylphenol	576-55-*6*	2,574	4,967	1,563	2,706	-0,132	1,613	2,814	-0,240	1,683	2,649	-0,075

Pentabromophenol	608-71-9	2,664	4,853	1,741	2,937	-0,273	1,710	2,886	-0,222	1,889	2,903	-0,239
3-iodophenol	626-02-8	1,119	2,895	1,196	0,885	-0,382	1,308	1,037	-0,165	1,586	1,280	-0,432
4-iodophenol	540-38-5	0,854	2,895	1,146	0,803	0,456	1,286	0,975	0,553	1,499	1,153	0,391
Série de test												
2-fluorophenol	367-12-4	0,185	1,715	1,132	0,139	0,046	1,190	0,176	0,009	1,232	0,150	0,035
2,6-difluorophenol	28177-48-*2*	0,471	1,747	1,308	0,450	0,021	1,360	0,456	0,015	1,372	0,378	0,093
4-bromophenol	106-41-2	0,680	2,630	1,141	0,676	0,004	1,174	0,687	-0,007	1,266	0,691	-0,011
4-chloro-3-methylphenol	59-50-7	0,796	2,984	1,097	0,805	-0,009	1,098	0,777	0,019	1,164	0,727	0,069
4-chloro-3-ethylphenol	14143-32-9	1,081	3,513	1,094	1,102	-0,021	1,101	1,090	-0,009	1,171	1,021	0,060
3-bromophenol	591-20-8	1,145	2,635	1,198	0,774	0,371	1,206	0,738	0,407	1,321	0,777	0,368
4-bromo-3,5-dimethylphenol	7463-51-6	1,268	3,633	1,092	1,166	0,102	1,144	1,227	0,041	1,102	0,981	0,287
3,5-dichlorophenol	591-35-5	1,569	3,287	1,303	1,320	0,249	1,283	1,239	0,33	1,483	1,371	0,198
4-chloro--2-isopropyl-5-methylphenol	89-68-9	1,854	4,411	1,064	1,565	0,289	1,077	1,580	0,274	1,113	1,415	0,439
2,3,5,6-tetrachlorophenol	935-95-*5*	2,222	3,848	1,547	2,046	0,176	1,432	1,796	0,426	1,718	2,027	0,195
2,3,4,5-tetrachlorophenol	4901-51-3	2,712	4,058	1,484	2,060	0,652	1,371	1,825	0,887	1,664	2,058	0,654

1. 2.Calcul

1.2.1. Optimisation de la géométrie et calcul des descripteurs moléculaires

Les structures ont été établies avec le programme Chem-Office [32]. Tout d'abord, nous avons réalisé une optimisation préliminaire de la géométrie des molécules étudiées en utilisant la mécanique moléculaire, puis les méthodes semi-empiriques AM1 [33], PM3 [34], et PM6 [35] implémentées dans le programme Gaussian 09 [36] sont utilisées pour l'optimisation finale de la géométrie. Plusieurs descripteurs ont été calculés pour chaque composé. Le coefficient de partage octanol/eau (log *P*) et d'autres descripteurs (par exemple, la polarisabilité, la charge positive totale, la charge totale absolue, la surface moléculaire, la charge totale d'atomes halogénés, la charge totale d'atomes de carbone du cycle aromatique, ...) ont été calculés à l'aide de différents logiciels ACD /Labs [37], Hyperchem [38] et Molinspiration [39]. Les valeurs de log *P*, ont également été prises de la référence [11].

1.3 . Analyse statistique

La régression linéaire multiple (MLR) a été utilisée pour développer les modèles QSAR en utilisant MINITAB (version 15) [40].

2. Résultats et discussion

L'ensemble de données constitué par 45 phénols halogénés a été divisé en une série d'apprentissage formée de 34 composés et une série de test formée de 11 composés choisis d'une manière aléatoire. En utilisant l'option «Best subset» dans MINITAB, nous avons construit plusieurs modèles QSAR pour les 3 méthodes semi-empiriques AM1, PM3 et PM6. Ensuite, nous avons vérifié la non-colinéarité des descripteurs qui apparaissent dans chaque équation. Si les descripteurs dans l'équation MLR sont fortement corrélés, le modèle QSAR est systématiquement rejeté.

Nous avons d'abord testé la corrélation entre la toxicité mesurée et le coefficient de partage octanol-eau log *P*, ce qui traduit la pénétration de la substance toxique dans les lipides membranaires. Le paramètre log *P* a été calculé à l'aide de plusieurs logiciels. Cependant, les valeurs tirées de la référence [11] sont trouvées celles qui donnent le

meilleur modèle de régression linéaire simple (R^2= 0,68, SD=0,39). Ce résultat montre que la toxicité de cette série de phénols halogénés peut être expliquée en partie par le descripteur log *P*. Toutefois, afin d'améliorer la qualité des modèles QSAR, l'inclusion d'autres descripteurs est nécessaire. Selon de nombreuses études de la littérature, la majorité des dérivés du phénol agit comme narcoses polaires [11-14]. Pour cette raison, l'indice d'électrophile de Parr, ω, a été calculé et utilisé comme un potentiel descripteur de chimie quantique.

Les modèles MLR à un et à deux paramètres, log *P* et ω (obtenus à l'aide des méthodes semi-empiriques, AM1, PM3, et PM6), sont donnés dans les équations (1-7); où n est le nombre de composés inclus dans le modèle, SD est la déviation standard, R^2 est le coefficient de corrélation au carré, F est le rapport de Fischer, R^2cv est le carré du coefficient de corrélation croisée validé et P est le P -value.

Méthode AM1

Modèles à un seul paramètre

$pIGC_{50}$= - 1,06 + 0,74 log *P* (1)

n= 34 ; R^2 = 0,68 ; SD = 0,39 ; F=68,13 ; P=0,000

$pIGC_{50}$= - 2,11 + 2,54 ω (2)

n= 34 ; R^2 = 0,52 ; SD = 0,47 ; F=35,00 ; P=0,000

Modèle à deux paramètres

$pIGC_{50}$ = - 2,69 + 1,65 ω+ 0,57 log *P* (3)

n= 34, R^2 = 0,87 ; R^2_{adj} = 0,86 ; R^2_{cv}= 0,84 ; SD = 0,26 ; F= 100,43 ; P=0,000

Méthode PM3

Modèles à un seul paramètre

$pIGC_{50}$= - 1,74 + 2,23 ω (4)

n= 34 ; R^2 = 0,40 ; SD = 0,27 ; F=21,55 ; P=0,000

Modèle à deux paramètres

$pIGC_{50}$= - 2,63 + 1,47 ω+ 0,62 log *P* (5)

n=34 ; R^2= 0,84 ; R^2_{adj}= 0,83 ; R^2_{cv}= 0,81 ; SD = 0,28 ; F= 80,05 ; P =0,000

Méthode PM6

Modèles à un seul paramètre

$pIGC_{50}$= - 2,18 + 2,34 ω (6)

n= 34 ; R^2 = 0,52 ; SD = 0,47 ; F=35,16 ; P=0,000

Modèle à deux paramètres

$pIGC_{50}$= - 2,67 + 1,50 ω+ 0,57 log *P* (7)

n= 34 ; R^2 = 0,85 ; R^2_{adj}= 0,84 ; R^2_{cv}= 0,82 ; SD = 0,27 ; F=90,20 ; P=0,000

Il s'avère qu'une amélioration considérable des modèles QSAR est obtenue en combinant le paramètre log *P* avec l'indice ω. Les trois méthodes semi-empiriques (AM1, PM3, PM6) ont donné des modèles MLR satisfaisants, bien que la méthode AM1 semble donner les meilleurs paramètres statistiques. Dans le **Tableau 2**, sont reportés, le coefficient (Coef.), l'erreur-type de coefficients (SE Coef.), T-test, facteur de variance de l'inflation (VIF) et le coefficient de corrélation (R^2). L'analyse des valeurs de VIF et R_{cor} montre qu'il n'y a pas de corrélation (colinéarité) entre les deux descripteurs ω et log *P*.

Tableau. 2. Coefficients des variables, VIF et coefficient de correlation des descripteurs du meilleur modèle.

	Coef.	SE Coef.	T-test	VIF	R^2
Constant	-2.69	0.31	-8.73		
ω	1.65	0.25	6.57	1.19	0.39
log *P*	0.57	0.06	8.91	1.19	0.39

L'analyse des résidus standardisés montre l'existence de deux valeurs incohérentes (*outliers*) dans la série d'apprentissage avec une valeur résiduelle standardisé supérieure à 2,2 unités de la toxicité pour les trois modèles donnés dans les équations. (3, 5, 7). Ces composés sont : 2, 3, 5-trichlorophénol (No 25) et le 3,4-dichlorophénol (No 30). La toxicité prédite de ces molécules est inférieure à la toxicité mesurée. Après élimination de

ces deux molécules de la série d'apprentissage, les qualités des modèles MLR, donnés dans les équations. (8-10) sont remarquablement améliorées.

Méthode AM1

$pIGC_{50} = -2,64 + 1,64\ \omega + 0,54 \log P$ (8)

n= 32 ; $R^2 = 0,91$; $R^2_{adj} = 0,91$; $R^2_{cv} = 0,90$; SD = 0,20 ; F=151,22 ; P=0,000

Méthode PM3

$pIGC_{50} = -2,67 + 1,55\ \omega + 0,58 \log P$ (9)

n= 32 ; $R^2 = 0,90$; $R^2_{adj} = 0,89$; $R^2_{cv} = 0,89$; SD = 0,21 ; F=132,92 ; P=0,000

Méthode PM6

$pIGC_{50} = -2,60 + 1,46\ \omega + 0,54 \log P$ (10)

n= 32 ; $R^2 = 0,89$; $R^2_{adj} = 0,89$; $R^2_{cv} = 0,86$; SD = 0,22 ; F=121,20 ; P=0,000

Les valeurs élevées des coefficients de régression R^2_{cv} de *leave-one-out* et les faibles valeurs des déviations standards SD (voir les équations. 8-10) justifient le pouvoir prédictif et la stabilité interne des modèles élaborés.

Les résultats de la Y-randomisation (voir **chapitre I**, page 19) pour les dix premières itérations sont présentés dans le **Tableau 3**. On a constaté que toutes les valeurs de R^2_r des modèles aléatoires sont plus faibles que le R^2 correspondant au modèle d'origine. Ce résultat indique que les modèles obtenus ne sont pas dues à une chance.

Tableau. 3. Valeurs de R_r^2 et R^2_{cv} après les différentes Y-randomisations

Iteration	R_r^2(Eq.8)	R^2_{cv}(Eq.8)	R_r^2(Eq.9)	R^2_{cv}(Eq.9)	R_r^2(Eq.10)	R^2_{cv}(Eq.10)
	AM1		PM3		PM6	
1	0.040	0.000	0.031	0.000	0.020	0.000
2	0.078	0.000	0.016	0.000	0.027	0.000
3	0.082	0.000	0.008	0.000	0.037	0.000
4	0.034	0.000	0.124	0.000	0.003	0.000
5	0.138	0.000	0.041	0.000	0.200	0.077
6	0.213	0.000	0.000	0.000	0.010	0.000
7	0.073	0.000	0.012	0.000	0.044	0.000
8	0.132	0.000	0.038	0.000	0.230	0.000
9	0.079	0.000	0.005	0.000	0.006	0.000
10	0.019	0.000	0.139	0.000	0.104	0.000

2.1. Validation externe

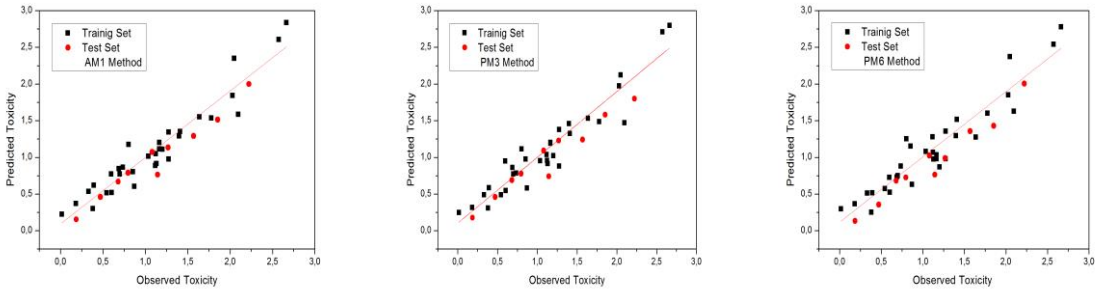
La validation interne n'est pas suffisante pour vérifier le pouvoir prédictif d'un modèle QSAR et une validation externe est nécessaire. Tous les modèles MLR donnés dans les équations (3, 5, 7) ont été utilisés pour prédire la toxicité d'une série de test constituée de 11 phénols halogénés choisis au hasard. L'analyse des résidus montre que le 2,3,4,5-tétrachlorophénol est un composé 'incohérent' et il est éliminé de manière systématique. Les résultats de la validation externe en utilisant les dix molécules restantes de la série de test sont présentés dans le **Tableau 4**.

Tableau. 4. Validation interne et externe des modèles QSAR

	Série d'apprentissage				Série de test			CritèresTropsha			
	n	R^2	R^2_{cv}	SD	n	R^2	SD	R^2_0	k	$R^2 - R_0^2 / R^2$	$\lvert R^2 - R_0^2 \rvert$
AM1											
Eq.(3)	34	0.86	0.85	0.26	10	0.96	0.14	0.96	0.90	0.000	0.000
Eq.(8)	32	0.91	0.90	0.20	10	0.96	0.13	0.96	0.86	0.001	0.001
PM3											
Eq.(5)	34	0.84	0.81	0.28	10	0.93	0.13	0.92	0.89	0.011	0.011
Eq.(9)	32	0.90	0.89	0.22	10	0.94	0.13	0.93	0.85	0.010	0.010
PM6											
Eq.(7)	34	0.86	0.83	0.26	10	0.96	0.13	0.96	0.86	0.000	0.000
Eq.(10)	32	0.89	0.87	0.22	10	0.96	0.13	0.96	0.84	0.009	0.009

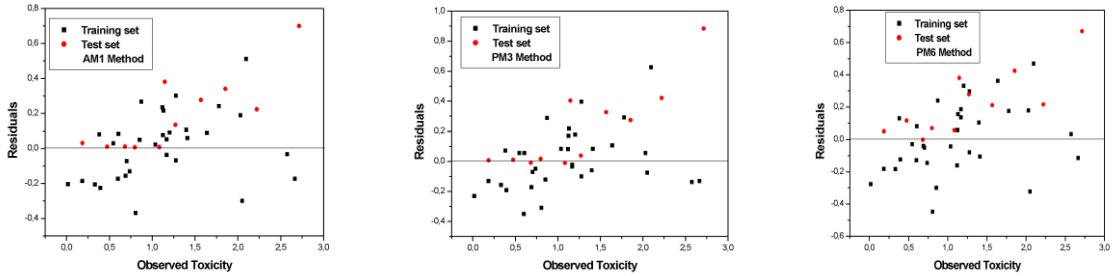
Comme mentionné dans le **chapitre I**, le pouvoir prédictif d'un modèle QSAR peut être vérifié à l'aide des critères de Tropsha (voir **chapitre I,** page 20). Ces résultats montrent le grand pouvoir de prédiction des modèles MLR particulièrement avec les méthodes AM1 et PM3. Pour le modèle obtenu avec la méthode PM6, les critères de Tropsha ne sont pas tous vérifiés, depuis la valeur de k est inférieur à 0,85. A la fin de processus rigoureux de validation (interne et externe), nous pouvons conclure que les modèles présentés par les équations ci-dessus sont stables et prédictifs. De plus, les deux descripteurs, à savoir, log *P* et ω sont pertinents et ne sont pas corrélés.

Les corrélations entre la toxicité prédite et la toxicité expérimentale pour les deux séries (apprentissage et de test) des modèles (8-10) sont représentées sur les **Figures (1a-1c)** respectivement.



Figures. (1a-1c). Toxicité prédite vs toxicité observée en utilisant Eqs.(8-10)

Les **Figures (2a-2c)** présentent la variation des résidus par rapport aux valeurs expérimentales pour les trois modèles#8-10. Comme la plupart des résidus calculés sont répartis sur deux côtés de la ligne zéro, nous pouvons conclure qu'il n'ya pas d'erreur systématique dans le développement des modèles actuels.



Figs. (2a-2c). Résiduels vs Toxicité observée en utilisant Eqs.(8-10)

2. 2. Domaine d'applicabilité

Le domaine d'applicabilité du meilleur modèle QSAR obtenu (Eq. 8), est présenté sur la **Figure 3**. Ce domaine est une représentation des valeurs des résidus de prédiction standardisés des composés pour les deux sous-groupes (apprentissage et test)
en fonction de leurs leviers (*leverage*) respectifs. Ce domaine nous permet une détection graphique à la fois des valeurs aberrantes et les produits chimiques influents dans un modèle. La **Figure 3** montre qu'il y'a, deux composés (N° 1 et 10) avec un h légèrement supérieur à la valeur critique ($h > h^* = 0,28$). L'applicabilité du modèle peut être évaluée avec les gammes de descripteurs, valeurs minimales et maximales pour la série de composés modélisés, donnés dans le **Tableau 1**, ces gammes peuvent être utilisées pour la prédiction de la toxicité de nouveaux composés.

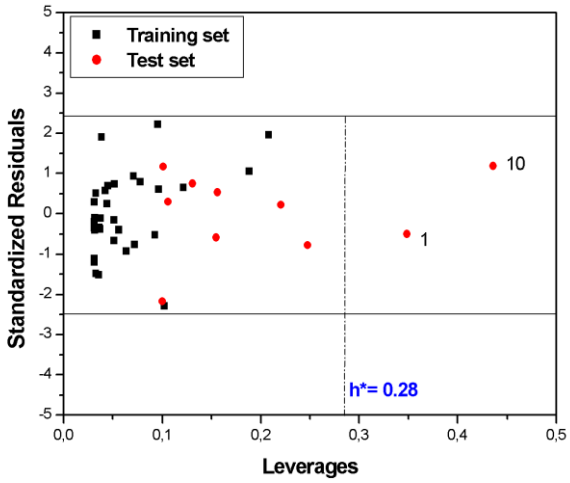


Figure.3. Domaine d'applicabilité du modèle présenté par Eq.8

2. 3. Discussion du mécanisme de toxicité

Dans le meilleur modèle MLR donné par l'équation (8), Les principaux facteurs qui peuvent influencer sur la toxicité sont le paramètre de lipophilie, présenté par log *P*, et l'indice d' électrophilie de Parr, ω, calculé selon la méthode AM1. L'analyse des valeurs T-test (**Tableau 3**) et le coefficient normalisé de chaque descripteur montre que le descripteur log *P* a la plus grande valeur de T-test. Par conséquent, le facteur d'hydrophobie, tel qu'il est exprimé par log *P* avec un coefficient positif, est utile pour décrire le transport vers le site d'action. Donc, si le composé a une valeur élevée de log *P*, on aura une bonne solubilité dans les lipides, et il peut diffuser facilement la membrane cellulaire et de se concentrer sur les organismes, ce qui conduit à une augmentation de la toxicité de la molécule. La contribution des effets électroniques, exprimée par le paramètre d'électrophilie de Parr, ω, est également importante pour prévoir la toxicité des phénols halogénés. Le coefficient positif de ω indique que l'augmentation du pouvoir d'électrophilie d'un composé conduit à l'augmentation de sa toxicité. Par conséquent, nous pouvons conclure que ce descripteur est plus approprié pour décrire la capacité électrophile des phénols halogénés comparant avec les descripteurs précédemment utilisés (E_{LUMO}, E_{HOMO} ou Amax). D'autre part, les phénols halogénés présentent un mécanisme narcose polaire de sorte qu'il y'a une interaction non covalente avec le composant

lipidique. Classiquement, les produits chimiques narcotiques polaires ont été modélisés dans le cadre de l'approche réponse-surface sous la forme d'un modèle à deux paramètres incluant à la fois le transport et les effets électroniques [**41**]. Dans cette approche, une variable indépendante, caractérisant l'absorption de la substance chimique dans la bio-phase (pénétration de la structure moléculaire). Une autre variable indépendante qui explique l'interaction avec le site d'action à savoir, les effets électroniques. Les modèles QSAR simples, décrits dans la présente application, combinent à la fois le transport et les facteurs électroniques et expliquent de façon adéquate le mécanisme narcose- polaire des phénols halogénés.

3. Conclusion

La toxicité de 45 phénols halogénés est modélisée avec succès par le paramètre de lipophilie et l'indice d'électrophilie de Parr, calculé par différentes méthodes semi-empiriques et en utilisant l'analyse de régression linéaire multiple (MLR). Les modèles QSAR développés sont simples, interprétables et transparents en utilisant un nombre réduit de descripteurs. En outre, ils ont une bonne stabilité, la robustesse et le pouvoir prédictif élevé, vérifié par la validation interne qui est clair à partir de son coefficient de corrélation R^2et le coefficient de validation croisée R^2_{cv} et plus précisément de sa validation externe. Ainsi, les modèles sont considérés comme validés et applicables pour l'exploitation de la base de données. Le domaine d'applicabilité du meilleur modèle étudié obtenue avec l'indice oméga calculé selon la méthode AM1 peut être servi comme un outil précieux pour filtrer les dissemblables et les valeurs aberrantes. Ainsi, il est applicable à faire des prédictions pour les nouveaux composés. L'indice d'électrophilie de Parr calculé en utilisant des méthodes économiques avec la combinaison du paramètre de lipophilie explique et rationalise le mécanisme de narcose polaire de cette série de phénols halogénés.

Références Bibliographiques

[1] E. Papa, F. Villa, P.Gramatica, Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephalespromelas (Fathead Minnow), *J. Chem. Inf. Model*. 45(**2005**), 1256-1266.

[2] J.D Wallker, Applications of QSARs in toxicology: a US Government perspective, *J. Mol. Struct.-Theochem*. 622 (**2003**), pp. 167- 184.

[3] S.P. Bradbury, C.L. Russom, G.T.Ankley, T.W. Schultz, J.D.Walker, Overview of data and conceptual approaches for derivation of quantitative structure–activity relationships for ecotoxicological effects of organic chemicals, *Environ. Toxicol. Chem*. 228 (**2003**), pp.1789-1798.

[4] European Commission. White Paper on a strategy for a future Community Policy for Chemicals, http://europa.eu.int:/comm/enterprise/reach/.

[5] M.W.Toussaint, T.R.Shedd, W.H.Van der Shalie, G.R.Leather, A comparison of standard acute toxicity tests with rapid-screening toxicity tests. *Environ. Toxicol.Chem*, 14 (**1995**), pp.907-915.

[6] H.Kubiniy, From narcosis to hyperspace: The history of QSAR. Quant.strut-Act.Rel, 21(**2002**), pp. 348-356.

[7] http://www.epa.gov/nrmrl/std/qsar/qsar.html.

[8] J.Michałowicz, W.Duda, Phenols – Sources and Toxicity, Polish. J.Environ.Stud. 16 (**2007**), pp. 347-362.

[9] G.M. Della, P.Monaco, G.Pinto, A.Pollio, L.Previtera and F.Temussi, Phytotoxicity of low-molecular-weight phenols from olive mill waste waters, *Bull.Environ.Contam.Toxicol*. 67 (**2001**), pp.352-359.

[10] R.Garg, S.Kapur and C.Hansch, Radical toxicity of phenols: a reference point for obtaining perspective in the formulation of QSAR, *Med Res Rev* 21(**2001**), 73-82.

[11] T.W. Schultz, A.P.Bearden, J.S.Jaworska, A novel QSAR approach for estimating toxicity of phenols, *SAR QSAR. Environ.Res* . 5 (**1996**), pp. 99–112,

[12] A.O. Aptula, T.I.Netzeva, I.V.Valkova, M.T.D.Cronin, T.W.Schultz, R.Kühne, G. Schüürmann, Multivariate Discrimination between Modes of Toxic Action of Phenols, *Quant Struct.-Act. Relat* . 21(**2002**), pp.12–22.

[13] M.T.D. Cronin, A.O. Aptula, J.C. Duffy, T.I .Netzeva, P.H. Rowe,I.V. Valkova, T.W. Schultz, Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymenapyriformis, *Chemosphere* . 49 (**2002**), pp.1201–1221.

[14] S.J. Enoch, M.T.D. Cronin, T.W. Schultz, J.C. Madden, An evaluation of global QSAR

models for the prediction of the toxicity of phenols to Tetrahymenapyriformis, *Chemosphere*. 71(**2008**), pp.1225–1232.

[**15**] C.Maria, E.G.Guimarães, D.G.Mota Silva, P.Matheus, Freitas, Aug-MIA-QSPR modelling of the toxicities of anilines and phenols to Vibrio fischeri and Pseudokirchneriellasubcapitata, *Chemo.Int. Lab. Syst*. 134 (**2014**), pp.53–57.

[**16**] F.A. Pasha, H.K. Srivestava, P.P. Singh, Comparative QSAR study of phenol derivatives with the help of density functional theory, *Bioorg. Med. Chem*. 13 (**2005**), pp.6823-6829.

[**17**] M.D.Ertürka, M.S. Türker, A.M. Novicb, N.Minovskib, Quantitative structure–activity relationships (QSARs) using the novel marinealgal toxicity data of phenols, *J.Mol.Graphics.Model* . 38(**2012**), pp.90–100.

[**18**] CODESSA PRO, University of Florida, www.codessa-pro.com

[**19**] Dragon, TALETE, Italy, www.talete.mi.it/products/dragon_description.htm.

[**20**] M.Erturkn, S.MelekTurker, Assessment and modelling of the novel toxicity dataset ofphenols to Chlorellavulgaris. *Ecotoxicol. Environ. Saf* . 90 (**2013**), pp.61–68.

[**21**] T.W.Schultz, D.T. LinS.K.Wesley, QSARs for monosubstituted phenols and the polar narcosis mechanism of toxicity. Quality *Assur. Good Pract. Regul. Law*, 2 (**1992**), pp.132–143.

[**22**] J.H. Xing, Y.T Zhang, A QSAR Study of Halogen Phenols Toxicity to the TetrahymenaPyriformis, *Comp.App. Chem*. 24 (**2007**), pp.87-90.

[**23**] Y.F.Peng, T.B.Liu, QSAR Study of Halogen Phenols Toxicity to TetrahymenaPyriformis, *Chin. J. Struct. Chem*. 28 (**2009**), pp.218-222.

[**24**] G.Hea, L.Fenga, H. Chen, A QSAR Study of the Acute Toxicity of Halogenated Phenols, *Procedia Engineering*. 43 (**2012**), pp. 204 – 209.

[**25**] http://www.oecd.org/document/23/0,2340,en_2649_201185_33957015_1_1_1_1,00.html.

[**26**] Cronin,M.T.D., Manga, N., Seward, J.R., Sinks G.D and Schultz, T.W. Parametrization of electrophilicity for the prediction of the toxicity of aromatic compounds. *Chem. Res. Toxicol*. 14 (**2001**), pp. 1498-1505

[**27**] M.T.D. Cronin, B. W .Gregory and T.W. Schultz, Quantitative structure-activity analysis of nitrobenzene toxicity to TetrahymenaPyriformis. *Chem. Res. Toxicol*. 11 **(1998)**, pp. 902-908.

[**28**] S.Ren, Determining the mechanisms of toxic action of phenols to *Tetrahymena pyriformis*. *Environ. Toxicol*. 17 **(2002)**, pp. 119-127.

[**29**] R.G. Parr, L.V. Szentpaly and S. Liu, Electrophilicity index. *J.Am.Chem.Soc*. 121 (**1999**),

pp. 1922-1924.

[30] P.K.Chattaraj, S.Giri, and S.Duley, Update 2 of: Electrophilicity Index. *Chem.Rev.*111 (**2011**), pp. PR43–PR75.

[31] J. Padmanabhan, R. Parthasarathi, V.Subramanian, P.K. Chattaraj, Group philicity and electrophilicity as possible descriptors for modelling ecotoxicity applied to chlorophenols. *Chem. Res. Toxicol.*19 (**2006**), pp.356-64.

[32] Chem-Office **2004**.

[33] M.J.S Dewar, E.G Zoebisch, E.F Healy, J.J.P. Stewart, Development and use of quantum molecular models. 75. Comparative tests of theoretical procedures for studying chemical reactions, *J. Am. Chem Soc*. 107(**1985**), pp.3902–3909.

[34] J.J.P.Stewart, Optimization of parameters for semiempirical methods I. Method. *J. Comput Chem.*10 (**1989**), pp. 209–220.

[35] J.J.P. Stewart, Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model*. 13 (**2007**), pp.1173–1213.

[36] Gaussian 09, Revision D.01, Gaussian, INC, Wallingford CT, **2009**.

[37] ACD/Labs, Release 12. **2009**, http://www.acdlabs.com

[38] HyperChem Release 7. HyperCube.Inc..http://www.hyper.com

[39] MolinspirationCheminformatics, http: //www.molinspiration.com

[40] MINITAB, State College, PA Minitab, Inc. **2006**.

[41] S.D. Dimitrov, O.G. Mekenyan, G.D. Sinks, T.W. Schultz Global modelling of narcoticchemicals:ciliate and fishtoxicity. J Mol.Struct. 622 (**2003**), pp. 63-70.

CONCLUSION GENERALE

L'objectif de cette thèse était de développer des modèles QSAR fiables pour la prédiction de la toxicité de certaines familles de molécules organiques, en l'occurrence, les nitrobenzènes, les nitro-aromatiques et les phénols halogénés vis-à-vis l'espèce aquatique *Tetrahymena-pyriformis*.

Une méthodologie basée sur la régression linéaire multiple (MLR), implémentée dans le logiciel MINITAB, a été utilisée. Cette méthode permet d'extraire de manière efficace des modèles QSAR transparents. Ces modèles sont à la fois fiables, c'est-à-dire explicatifs, prédictifs et interprétables en choisissant des descripteurs pertinents pour expliquer et interpréter la toxicité des composés étudiés du point de vue statistique et chimique.

La première application a montré que les descripteurs de la chimie quantique, à savoir, l'énergie LUMO et l'indice d'électrophilie ω de Parr, en combinaison avec l'indice d'hydrophobie « log *P* », facteur de transport, sont utiles pour la prédiction de la toxicité (log (IGC_{50}^{-1})) des nitrobenzènes. Le modèle QSAR obtenu est capable de décrire environ 87% de la variance de la toxicité expérimentale et pourrait être utilisé efficacement pour estimer la toxicité des dérivés du nitrobenzène pour lesquels les données expérimentales sont indisponibles. En effet, le modèle QSAR élaboré révèle que les nitrobenzènes les plus toxiques sont caractérisés par leur forte lipophilie et un pouvoir électrophile élevé.

Dans la deuxième application, les deux mécanismes possibles (*Redox-Cycling* et *Nucleophilic-Attack*) de la toxicité des nitro-aromatiques sont examinés à l'aide des descripteurs de la chimie quantique en présence du coefficient de partage. Etant donné que la toxicité des composés nitro-aromatiques provient principalement de leurs interactions avec les systèmes biologiques essentiellement par transfert d'électrons, l'utilisation de descripteurs qui expriment ce phénomène est d'une importance majeur dans l'élaboration des modèles QSAR. L'indice d'électrophilie de Parr a été utilisé pour présenter et quantifier ce transfert ainsi pour l'explication du mécanisme d'attaque nucléophile. D'autre part les résultats présentés ici suggèrent que la sensibilité des nitro-aromatiques pour le mécanisme « *Redox -Cycling* » peut être évaluée par l'énergie SOMO et

l'indice de densité de spin de Mulliken calculé par des méthodes de la chimie quantique.

Les deux applications montrent que l'électrophilie exprimée par l'indice de Parr est très utile pour l'étude de la toxicité des nitrobenzènes et des dérivés nitro-aromatiques, et ceci quelque soit la méthode utilisée pour le calcul de cet indice B3LYP/DFT ou bien les méthodes semi-empiriques (AM1 par exemple).

Dans la 3ème application, la toxicité d'une série de phénols halogénés est modélisée avec succès par l'hydrophobie et l'indice d'électrophilie de Parr calculé par différentes méthodes semi-empiriques (AM1, MP3, PM6). Les modèles QSAR développés sont simples, interprétables et transparents en utilisant un nombre réduit de descripteurs. En outre, ils sont caractérisés par la stabilité, la robustesse et le pouvoir prédictif élevé vérifié par la validation interne et externe. Ainsi, les modèles sont considérés comme validés et applicables pour l'exploitation de la base de données. L'indice d'électrophilie de Parr calculé en utilisant la méthode de faible coût avec la combinaison du paramètre d'hydrophobie peuvent expliquer et rationaliser le mécanisme « narcose polaire » de cette série de phénols halogénés.

Il est difficile d'obtenir des modèles QSTR parfaits vue la complexité du phénomène de toxicité qui est gouverné par de multiples facteurs.

Les perspectives de ce travail nous semblent diverses. D'une part, nous avons l'intention de reprendre les mêmes bases de données et élaborer des modèles en utilisant d'autres méthodes d'analyse de données telle que PLS, PCA, GA, NN, SVM,...D'autre part, nous projetons d'élaborer des modèles QSAR pour d'autres bases de données de molécules toxiques vis-à vis d'autres espèces aquatiques et vis-à-vis de l'être humain (cytotoxicité, mutagénèse…) en utilisant toujours les descripteurs quantiques et non-quantiques.

ANNEXE

ANNEXE (Application II)

Mécanisme A

Paramètre	R^2
log *P*	0,37
E_{SOMO}	0,39
MSD	0,09
DM	0,06

Mécanisme B

Paramètre	R^2
log *P*	0,40
α	0,60
ω	0,50

Arabian Journal of Chemistry (2012) **xxx**, xxx–xxx

King Saud University

Arabian Journal of Chemistry

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis* using quantum chemical descriptors

Khadidja Bellifa, Sidi Mohamed Mekelleche *

Laboratory of Applied Thermodynamics and Molecular Modelling, Department of Chemistry, Faculty of Science, Abou-Bekr Belkaïd University, BP 119, Tlemcen 13000, Algeria

Received 27 September 2011; accepted 25 April 2012

KEYWORDS

Nitrobenzenes;
Toxicity;
QSAR;
Electrophilicity index;
log *P*;
Quantum chemistry calculations

Abstract Quantitative Structure–Activity Relationship (QSAR) models are useful in understanding how chemical structure relates to the biological activity and the toxicity of natural and synthetic chemicals. The present study shows that Parr's electrophilicity index ω in combination of two other descriptors, namely, the LUMO energy and the hydrophobicity index log *P*, prove their utility for the prediction of the toxicity of a series constituted by 50 nitrobenzene derivatives. The QSAR models are developed using the Multiple Linear Regression (MLR) method. It turns out that the best model, which its stability is confirmed using the leave-1/3-of-set-out validation, is able to describe about 87% of the variance of the experimental toxicity. The satisfactory obtained results show that Parr's electrophilicity index is a useful quantum chemical descriptor for the toxicity modeling of nitrobenzene derivatives. Finally, the elaborated model shows that the most toxic nitrobenzenes are characterized by large hydrophobicities and high electrophilicity powers and could be efficiently applied for the estimation of the toxicity of nitrobenzenes for which the experimental measures are unavailable.

1. Introduction

The rapid development of new compounds by the chemical industry in general and in particular by the agrochemical, petrochemical and pharmaceutical industries is accompanied by an increasing toxic burden in the environment. Because of this, the development of tools able to assess hazardous effects on living species should receive high attention (Smiesko and Benfenati, 2004).

Quantitative Structure–Property/Activity Relationship (QSPR/QSAR) methods are among the most practical tools in computational physical chemistry. These methods are based on the axiom that the variance in the physicochemical

* Corresponding author. Tel./fax: +213 43286308.
E-mail addresses: sidi_mekelleche@yahoo.fr, sm_mekelleche@mail.univ-tlemcen.dz (S.M. Mekelleche).

Peer review under responsibility of King Saud University.
http://dx.doi.org/10.1016/j.arabjc.2012.04.031

ELSEVIER **Production and hosting by Elsevier**

properties and activities of chemical compounds is determined by the variance in their molecular structures. Thus, if experimental data are available for only some chemicals in a group, one can predict the missing from molecular descriptors calculated for the whole group and suitable mathematical model (Karcher and Devillers, 1990; Schultz et al., 2003; Katritzky et al., 2001). The global prediction of toxicity using QSARs has been the goal of many workers who utilized a variety of approaches. This goal is alluring, but has yet to be achieved satisfactorily. There are a number of reasons for the absence of success (Arulmozhiraja and Morita, 2004). The deficiency of available toxicity data has clearly held back progress. This lack of success has been compounded in many studies by a poor appreciation of the insufficient heterogeneity, or chemical diversity, in the dataset. Further, while some molecular properties (such as hydrophobicity) are well described, others, including electrophilic reactivity, ionization, and hydrogen bonding, are poorly parameterized. Last, mechanisms of toxic action are not fully understood or misinterpreted, or their relevance in the modeling of toxicity is ignored (Cronin et al., 2001).

Nitroaromatics are hazardous chemicals that display several manifestations of toxicity, including skin sensitization, immunotoxicity, germ cell degeneration, inhibition of liver enzymes and also a conjectured carcinogenicity. The modeling of toxicity of nitroaromatic compounds was complicated by the paucity of experimental data. Nitrobenzenes (NBs) are widely used as industrial chemicals, and consequently have high potential for environmental pollution, they have been reported (Zoeteman et al., 1980) to be present in surface waters. They are reactive chemicals, being reported to be uncouplers of oxidative phosphorylation (Purdy, 1988) and may be regarded as pro-electrophiles (Roberts, 1987) yielding the corresponding potentially highly toxic C-nitroso compounds. Because of their widespread use, the toxicity of NBs has been quite extensively examined, and they have been the subject of a number of QSAR studies. The NBs are representatives of electrophilic toxicants in that, depending on substitution pattern, they may undergo a number of different electrophilic reactions. NBs toxicity has been extensively studied by several groups of workers with the use of different methodologies. Attempts to model the acute toxicity of NBs have been reviewed by Dearden et al. (1995). Due to the reactive electrophilic nature of the NBs, it is not surprising that previous modeling efforts focused on the use of electronic molecular descriptor (Deneer et al., 1987, 1989; Roberts, 1987; Veith and Mekenyan, 1993; Lang et al., 1996; Yuan et al., 1997; Zhao et al., 1997; Debnath et al., 1991; Mekenyan et al., 1997).

Since the expression of chemical toxicity is a combination of penetration into, or through, biological membranes and the interaction of the toxicant with the site of action, this principle is represented mathematically as the following generic QSAR (Schultz, 1999):

$$\log(\text{toxicity})^{-1} = \text{A}(\log \text{ of penetration}) + \text{B}(\log \text{ of interaction}) + \text{C} \quad (1)$$

Penetration to the site of action is generally represented by hydrophobicity, most often quantified by the 1-octanol/water partition coefficient ($\log P$) (Schultz, 1999). Interaction of the chemical with the site of action is more complicated and describes electronic and/or steric properties.

Our aim of this study is to develop reliable and predictive QSAR models for identifying the primary molecular factors explaining the interaction effect (electronic/penetration) and governing the toxicity of these NB derivatives and to examine what factors other than $\log P$ are controlling the toxicity of compounds.

2. Materials and methods

2.1. Data set

A total of 50 NBs each containing either a halogenated or/and a methyl, nitro and other substituents is created from the data set described by Cronin et al. (2001). Scheme 1 shows the template structure of nitrobenzene.

The experimental toxicity data $\log(1/\text{IGC}_{50})$ of the 50 congeners are listed in Table 1.

2.2. Quantum chemistry calculations

MOPAC6 semi-empirical molecular orbital package (Stewart, 1989) was used for doing quantum chemistry calculations and the equilibrium geometries were optimized using the AM1 method. According to previous works (Cronin et al., 1998) the toxicity of NBs can be explained in terms of the electrophilicity power of these compounds which has been expressed by the $A_{\max}$ descriptor (Fukui et al., 1954). However, the electrophilicity concept is more adequately defined within the conceptual DFT (density functional theory). According to DFT, the chemical potential and chemical hardness for the n-electron molecular system with total energy E and external potential are defined as the first and second derivatives of the energy with respect to n, respectively:

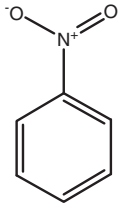
$$\mu = \frac{\varepsilon_{\text{lumo}} + \varepsilon_{\text{homo}}}{2} \quad (2)$$

$$\eta = \varepsilon_{\text{lumo}} - \varepsilon_{\text{homo}} \quad (3)$$

where $\varepsilon_{\text{lumo}}$ is the lowest unoccupied molecular orbital's energy and $\varepsilon_{\text{homo}}$ is the highest occupied orbital's energy. Using μ and η, Parr et al. (1999) have defined a new quantum chemical descriptor, known as electrophilicity index, ω, which measures the propensity to absorb electrons and is defined as

$$\omega = \frac{\mu^2}{2\eta} \quad (4)$$

The value for the hydrophobicity term, expressed as the logarithm of the octanol/water partition coefficient $\log P$, for each chemical is computed using ACD/Labs program (ACD/Labs, 2009)



Scheme 1 The structural template of nitrobenzene (NB).

Please cite this article in press as: Bellifa, K., Mekelleche, S.M. QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis* using quantum chemical descriptors. Arabian Journal of Chemistry (2012), http://dx.doi.org/10.1016/j.arabjc.2012.04.031

Table 1 NBs and their toxicity against *Tetrahymena pyriformis* (Cronin et al., 2001).

No.	Compound	Observed toxicity $\log(IGC_{50}^{-1})$
1	2,6-Dimethylnitrobenzene	0.30
2	2,3-Dimethylnitrobezene	0.56
3	2-Methyl-3-chloronitrobenzene	0.68
4	2-Methylnitrobenzene	0.05
5	2-Chloronitrobenzene	0.68
6	2-Methyl-5-chloronitrobenzene	0.82
7	2,4,5-Trichloronitrobenzene	1.53
8	2,5-Dichloronitrobenzene	1.13
9	6-Chloro-1,3-dinitrobenzene	1.98
10	Nitrobenzene	0.14
11	3-Methylnitrobenzene	0.05
12	1,3-Dinitrobenzene	0.89
13	3,4-Dichloronitrobenzene	1.16
14	4-Methylnitrobenzene	0.17
15	1,4-Dinitrobenzene	1.30
16	4-Chlonitrobenzene	0.43
17	2,3,5,6-Tetrachloronitrobenzene	1.82
18	6-Methyl-1,3-dinitrobenzene	0.87
19	3-Chloronitrobenzene	0.73
20	1,2-Dinitrobenzene	1.25
21	2-Bromonitrobenzene	0.75
22	6-Bromo-1,3-dinitrobenzene	2.31
23	3-Bromonitrobenzene	1.03
24	4-Bromonitrobenzene	0.38
25	2,4,6-Trimethylnitrobenzene	0.86
26	5-Methyl-1,2-dinitrobenzene	1.52
27	2,4-Dichlorobenzene	0.99
28	3,5-Dichlorobenzene	1.13
29	6-Iodo-1,3-dinitrobenzene	2.12
30	2,3,4,5-Tetrachloronitrobenzene	1.78
31	2,3-Dichlorobenzene	1.07
32	2,5-Dibromobenzene	1.37
33	1,2-Dichloro-4,5-dinitrobenzene	2.21
34	3-Methyl-4-bromonitrobenzene	1.16
35	2,3,4-Trichloronitrobenzene	1.51
36	2,4,6-Trichlonitrobenzene	1.43
37	4,6-dichloro-1,2-Dinitrobenzene	2.42
38	3,5-Dinitrobenzyl alcohol	0.53
39	3,4-Dinitrobenzyl alcohol	1.09
40	2,4,6-Trichloro-1,3-dinitrobenzene	2.19
41	2,3,5,6-Tetrachloro-1,4-dinitrobenzene	2.74
42	2,4,5-Trichloro-1,3-dinitrobenzene	2.59
43	4-Fluoronitrobenzene	0.25
44	4-Fluoro-2-nitrotoluene	0.25
45	1-Fluoro-2-nitrobenzene	0.23
46	1-Fluoro-3-nitrobenzene	0.20
47	4-Nitrobenzaldehyde	0.20
48	2-Nitrobenzaldehyde	0.17
49	3-Nitrobenzaldehyde	0.14
50	3-Nitroacetophenone	0.32

2.3. Statistical analysis

Structure–toxicity models are generated using the multilinear regression procedure of MINITAB version 15 (MINITAB, 2006). $\log(1/IGC_{50})$ values reported in millimolar are used as the dependent variable and $\log P$, ε_{lumo}, ω, as the independent variables. The models are assessed with the R^2 value (coefficient of determination), the R^2-adjusted, the SD value (root of the mean square of errors) and the F value (Fischer statistic). The number of observations N is also noted.

3. Results

The calculated quantum chemical descriptors, namely, ε_{lumo}, the electrophilicity power ω, and the estimated partition coefficient $\log P$ are given in Table 2.

Several linear QSAR models involving one, two, and three descriptors are established and strongest multivariable correlations are identified by the "Best Subsets" regression analysis of the MINTAB program.

3.1. One-parameter QSAR models

Model #1

$$\log(1/IGC_{50}) = 0.04 + 0.39\log P$$
$$N = 50, \quad R^2 = 0.16, \quad R^2_{adj} = 0.14, \quad SD = 0.68 \tag{5}$$

Model #2

$$\log(1/IGC_{50}) = -0.89 - 1.23\varepsilon_{lumo}$$
$$N = 50, \quad R^2 = 0.66, \quad R^2_{adj} = 0.65, \quad SD = 0.43 \tag{6}$$

Model #3

$$\log(1/IGC_{50}) = -2.89 + 0.94\omega$$
$$N = 50, \quad R^2 = 0.68, \quad R^2_{adj} = 0.67, \quad SD = 0.42 \tag{7}$$

The plots of $\log(1/IGC_{50})$ versus the three descriptors, $\log P$, ω, and ε_{lumo} are given in Fig. 1a–c, respectively.

It turns out that the best one-parameter QSAR model is obtained with Parr's electrophilicity index ω (model #3, $R^2 = 0.68$). In order to improve the predictivity of the QSAR models, it is necessary to investigate multilinear QSAR models involving two and three parameters.

3.2. Two-parameters QSAR models

Model #4

$$\log(1/IGC_{50}) = -8.28 + 3.54\omega + 3.43\varepsilon_{lumo}$$
$$N = 50, \quad R^2 = 0.71, \quad R^2_{adj} = 0.69, \quad SD = 0.41, \quad F = 56.44 \tag{8}$$

Model #5

$$\log(1/IGC_{50}) = -1.78 + 0.36\log P - 1.21\varepsilon_{lumo}$$
$$N = 50, \quad R^2 = 0.80, \quad R^2_{adj} = 0.79, \quad SD = 0.33, \quad F = 94.75 \tag{9}$$

Model #6

$$\log(1/IGC_{50}) = -3.78 + 0.93\omega + 0.38\log P$$
$$N = 50, \quad R^2 = 0.82, \quad R^2_{adj} = 0.82, \quad SD = 0.31, \quad F = 112.34 \tag{10}$$

It turns out that the best two-parameters QSAR model is obtained with a combination of the electrophilicity index ω with the partition coefficient $\log P$ (model #6, $R^2 = 0.82$).

Please cite this article in press as: Bellifa, K., Mekelleche, S.M. QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis* using quantum chemical descriptors. Arabian Journal of Chemistry (2012), http://dx.doi.org/10.1016/j.arabjc.2012.04.031

Table 2 Descriptor values and predicted toxicity of nitrobenzene derivatives by Eq. (11).

Compound	log *P*	ε_{lumo} (eV)	ω (eV)	Observed toxicity	Predicted toxicity	Residual
1	2.87	−0.7769	3.1543	0.30	0.50	0.20
2	2.87	−0.7819	3.1642	0.56	0.50	−0.06
3	3.10	−1.1248	3.5681	0.68	0.79	0.11
4	2.41	−0.7982	3.2040	0.05	0.45	0.40
5	2.34	−1.1346	3.5960	0.68	0.56	−0.12
6	3.10	−1.1627	3.5966	0.82	0.73	−0.09
7	3.49	−1.8247	4.4298	1.53	1.52	−0.01
8	2.95	−1.4879	4.0007	1.13	1.00	−0.13
9	2.06	−2.0373	4.7984	1.98	1.61	−0.37
10	1.95	−0.7888	3.2322	0.14	0.44	0.30
11	2.41	−1.2031	3.6078	0.05	0.29	0.24
12	1.62	−1.6926	4.3708	0.89	1.14	0.25
13	3.16	−1.5547	4.0985	1.16	1.17	0.01
14	2.41	−1.2583	3.6922	0.17	0.46	0.29
15	1.37	−1.8640	4.5546	1.30	1.04	−0.26
16	2.60	−1.5815	4.1154	0.43	0.88	0.45
17	3.73	−1.9382	4.5766	1.82	1.77	−0.05
18	2.08	−1.9897	4.6950	0.87	1.40	0.53
19	2.64	−1.2550	3.7270	0.73	0.70	−0.03
20	1.84	−1.7289	4.4023	1.25	1.23	−0.02
21	2.52	−1.0742	3.5152	0.75	0.56	−0.19
22	2.02	−2.2595	5.0662	2.31	1.77	−0.54
23	2.52	−1.0775	3.5275	1.03	0.60	−0.43
24	2.55	−1.1738	3.6505	0.38	0.73	0.35
25	3.33	−0.8642	3.2370	0.86	0.62	−0.24
26	2.30	−1.7793	4.4306	1.52	1.31	−0.21
27	3.00	−1.5317	4.0798	0.99	1.11	0.12
28	3.34	−1.7999	4.3797	1.13	1.37	0.24
29	2.42	−2.3197	5.1297	2.12	1.92	−0.20
30	3.94	−2.0324	4.7113	1.78	2.02	0.24
31	2.90	−1.4078	3.9151	1.07	0.96	−0.11
32	3.12	−1.3881	3.8587	1.37	0.87	−0.50
33	3.20	−2.3302	5.1914	2.21	2.48	0.27
34	3.01	−1.1640	3.6124	1.16	0.79	−0.37
35	3.44	−1.7567	4.3577	1.51	1.52	0.01
36	3.41	−1.7980	4.4151	1.43	1.58	0.15
37	3.08	−2.2735	5.1149	2.42	2.35	−0.07
38	0.43	−2.0062	4.6984	0.53	0.61	0.08
39	0.65	−1.8173	4.4955	1.09	0.70	−0.39
40	2.72	−2.5126	5.4605	2.19	2.65	0.46
41	2.92	−2.7402	5.7611	2.74	3.01	0.27
42	2.84	−2.4658	5.3695	2.59	2.48	−0.11
43	1.80	−1.2118	3.6838	0.25	0.36	0.11
44	2.26	−1.1866	3.5977	0.25	0.28	0.03
45	1.69	−1.1698	3.6146	0.23	0.23	0.00
46	1.90	−1.1652	3.6092	0.20	0.27	0.07
47	1.56	−1.7598	4.31357	0.20	0.54	0.34
48	1.74	−1.3739	3.8796	0.17	0.43	0.26
49	1.75	−1.5254	4.0612	0.14	0.58	0.44
50	1.49	−1.4369	3.9445	0.32	0.36	0.04

3.3. Three-parameters QSAR model

Model #7

$$\log(1/IGC_{50}) = -11.70 + 4.7\omega + 4.9\varepsilon_{lumo} + 0.42\log P$$
$$N = 50 \quad R^2 = 0.87, \quad R^2_{adj} = 0.86, \quad SD = 0.27, \quad F = 106.11 \tag{11}$$

A significant improvement of the quality of the QSAR model is obtained with a combination of the three parameters, namely, Parr's electrophilicity index ω, the partition coefficient log *P*, and the LUMO energy ε_{lumo}. Fig. 2 shows the linear correlation between the observed and the predicted toxicity values obtained using Eq. (11).

In Table 3 are given the coefficients, the coefficient errors and *t*-test values of the three molecular parameters corresponding to the best model #7.

4. Cross-validation

In order to check the reliability and the stability of the best elaborated QSAR model (Eq. (11)), we have used the

Please cite this article in press as: Bellifa, K., Mekelleche, S.M. QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis* using quantum chemical descriptors. Arabian Journal of Chemistry (2012), http://dx.doi.org/10.1016/j.arabjc.2012.04.031

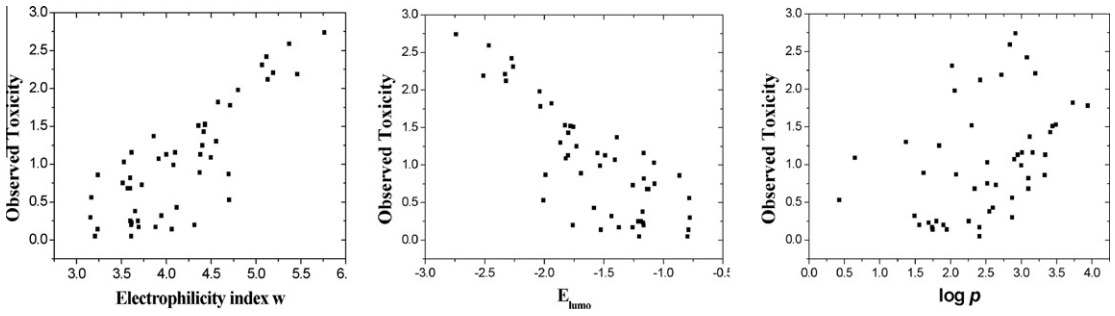


Figure 1 (a–c) Plot showing the relationship between each parameter and the observed toxicity.

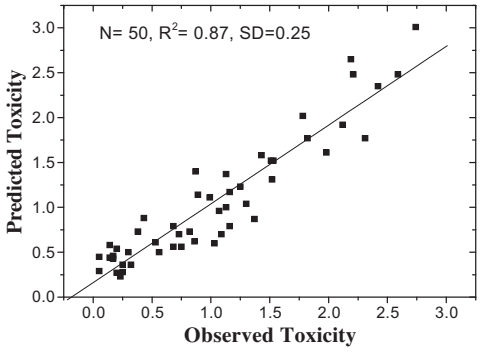


Figure 2 Predicted vs. observed toxicity using Eq. (11).

Table 3 Coefficients, coefficient errors and *t*-test values of the descriptors of the best QSAR model (Eq. (11)).

Number	X	DX	*t*-Test	Descriptor
0	−11.717	1.942	−6.03	Intercept
1	0.4204	0.0532	7.90	$\log P$
2	4.7055	0.9157	5.14	ω
3	4.983	1.208	4.13	ε_{lumo}

leave-1/3-of-set-out validation in the following manner: the parent data points were divided according to the experimental values into three subsets (the 1st, 4th, 7th, etc. entries go into the first subset A, the 2nd, 5th, 8th, etc. into the second subset B, and the 3rd, 6th, 9th, etc. into the third subset C). In each of three combinations, two of the subsets were combined into one and the correlation equation was derived with the same descriptors. The obtained equation was used to predict data for the remaining subset. It turns out that the predicted R^2 values using subsets (A + B), (B + C), (A + C) are close to that corresponding to the full training set (A + B + C) and the average values of R^2 (Fit) and R^2 (Predicted) (see Table 4) are also close. Note that the R^2_{adj} value of the models corresponding to the subsets A + B, A + C, and B + C is much larger than 0.80, indicating that our model is stable and can be efficiently used for estimating the toxicity of other nitrobenzenes for which no experimental data are available.

5. Discussion

The elaborated QSAR models (Eqs. (5)–(11)) reveal that the toxicity of the nitrobenzenes could be explained by a number of electronic and transport factors. Electrophilicity, as defined by ε_{lumo} and ω is important in describing the electronic interaction and the reactivity of these toxins; whereas hydrophobicity, as expressed by $\log P$ is important to describe the transport to the site of action. To put in evidence the contribution of each parameter (ω, ε_{lumo}, $\log P$) to the toxicity, we studied the relationship between these parameters and the toxicity $\log(IGC_{50}^{-1})$. Although the electrophilicity indexes ω and ε_{lumo} are found to be more significant in the one-parameter QSAR models (Eqs. (2) and (3)) in comparison with $\log P$ (Eq. (5)), the inclusion of the latter in the two- and three-parameter QSAR models is of great importance. Indeed, the combination of $\log P$ with either ω and/or ε_{lumo} provides more reliable models. The two-parameter model involving ω and ε_{lumo} explains only 71% of the variance of the toxicity; whereas the two-parameter models involving the combinations of ω with $\log P$ and ε_{lumo} with $\log P$ explain around 80–82% of the variance of the toxicity. This indicates the importance of the hydrophobicity index in the multilinear regression models. Indeed, it turns out that the combination of the three parameters increases remarkably the predictive power of the QSAR model given by Eq. (11) ($R^2 = 0.87$, $R^2_{adj} = 0.86$, SD = 0.27, $F = 106.11$). As can be seen from the statistical parameters of the above equation, a considerable improvement is achieved by combining the three descriptors. (Eq. (11)) can explain about 87% of the experimental variance of the dependent variable $\log(IGC_{50}^{-1})$ besides it presents a high F of Fischer

Table 4 Cross-validation of the best QSAR model.

Training set	*N*	R^2 (Fit)	R^2_{adj} (Fit)	S^2 (Fit)	Test set	*N*	R^2 (pred.)	R^2_{adj} (pred.)
A + B	34	0.88	0.87	0.27	C	16	0.82	0.81
A + C	33	0.86	0.85	0.28	B	17	0.88	0.88
B + C	33	0.88	0.87	0.25	A	17	0.84	0.83
Average		0.87	0.86	0.26			0.85	0.84

Please cite this article in press as: Bellifa, K., Mekelleche, S.M. QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis* using quantum chemical descriptors. Arabian Journal of Chemistry (2012), http://dx.doi.org/10.1016/j.arabjc.2012.04.031

Table 5 Chemical hardness (η), chemical potential (μ), electrophilicity (ω) of nucleic acid (NA) bases.

	η (eV)	μ (eV)	ω (eV)
Adenine A	4.32	−4.44	1.27
Thymine T	4.66	−4.94	2.62
Guanine G	4.18	−4.49	2.41
Cytosine C	4.63	−4.77	2.45
Uracil U	4.82	−5.14	2.74

($F = 106.11$) and a low standard deviation (SD = 0.27) which confirms that the model #7 predicts the toxicity (dependent variable) in a statistically satisfactory significant manner. According to the *t*-test values ($|t|$), the importance of the descriptors involved in the model decreases in the following order: $\log P > \omega > \varepsilon_{lumo}$. The most significant descriptor according to the *t*-test (see Table 4) is the partition coefficient $\log P$. The second significant descriptor is Parr's electrophilicity index ω and the last one is the LUMO energy ε_{lumo}. The ε_{lumo} is related directly to the electron affinity of a molecule and as such characterizes the susceptibility of the molecule to be attacked by nucleophiles; whereas, Parr's electrophilicity index, ω, defined in terms of the electronic chemical potential and the chemical hardness, expresses the stabilization energy when the system acquires an additional electronic charge from the environment. Our QSAR models reveal that Parr's electrophilicity index ω constitutes the main among the three descriptors in explaining the toxicity of the nitrobenzene derivatives. Indeed, ω alone explains about 68% of the variance of the toxicity (see Eq. (7)).

To confirm the electrophilic behavior of these nitrobenzenes, we have performed a comparison of the electrophilicity power of the 50 toxins with electrophilicity power of some nucleic acids (NA) bases. The values of electronic chemical potential, chemical hardness and electrophilicity indexes for adenine A, guanine G, cytosine C, uracil U and thymine T are given in Table 5.

It turns out that the electrophilicity indexes of the 50 toxins (see Table 2) are all greater than those of the NA bases (see Table 5). Consequently, the toxin will act as an electrophile (electron acceptor); whereas the NA will act as a nucleophile (electron donor) during the toxin–NA interaction. The positive coefficient obtained for Parr's electrophilicity index ω as a quantum chemistry descriptor in all QSAR models #1–7, supports the earlier concept that the toxicity of nitrobenzene derivatives increases with the increase in their electron accepting capability, indicating, that the electron flow takes place from the organism to the toxins. The same statements were obtained for the toxicity of polychlorinated dibenzofurans (Sarkar et al., 2006) and aromatic compounds (Cronin et al., 2001). Consequently, the most toxic nitrobenzenes are predicted to be characterized by high electrophilicity power (strong electron acceptors) and high hydrophobicity (lipophilicity) values.

6. Conclusion

The present study shows that quantum chemistry descriptors expressing the electrophilicity power, namely, the LUMO energy and Parr's electrophilicity index ω in combination with the hydrophobicity index, $\log P$, expressing the transport factor, are useful for the prediction of the toxicity of a nitrobenzenes to *Tetrahymena pyriformis* ($\log(IGC_{50}^{-1})$). The best QSAR model (Eq. (11)) is able to describe about 87% of the variance in the experimental toxicity and could be efficiently used for estimating the toxicity of nitrobenzene derivatives for which the experimental data are unavailable. Our study shows that Parr's electrophilicity index constitutes the main descriptor in explaining the toxicity of these toxins although the contributions of the $\log P$ and ε_{lumo} descriptors are also important. Indeed, the elaborated QSAR model reveals that the most toxic nitrobenzenes are characterized by large hydrophobicities and high electrophilicity powers.

References

ACD/Labs, 2009. Release 12. <http://www.acdlabs.com>.

Arulmozhiraja, S., Morita, M., 2004. Structure–activity relationships for the toxicity of polychlorinated dibenzofurans: approach through density functional theory-based descriptors. Chem. Res. Toxicol. 17, 348.

Cronin, M.T.D., Gregory, B.W., Schultz, T.W., 1998. Response surface-based analyses of nitrobenzene toxicity to *Tetrahymena pyriformis*. Chem. Res. Toxicol. 11, 902.

Cronin, M.T.D., Manga, N., Seward, J.R., Sinks, G.D., Schultz, T.W., 2001. Parametrization of electrophilicity for the prediction of the toxicity of aromatic compounds. Chem. Res. Toxicol. 14, 1498.

Dearden, J.C., Cronin, M.T.D., Schultz, T.W., Lin, D.T., 1995. QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis*. Quant. Struct.-Act. Relat. 14, 427.

Debnath, A.K., Compadre, R.L.L., Debnath, G., Shusterman, A.J., Hansch, C., 1991. Structure–activity relationship of mutagenic aromatic and heteroaromatic nitrocompounds. Correlation with molecular orbital energies and hydrophobicity. J. Med. Chem. 34, 786.

Deneer, J.W., Sinnige, T.L., Seinen, W., Hermens, J.L.M., 1987. Quantitative structure–activity relationships for the toxicity and bioconcentration factor of nitrobenzene derivatives towards the guppy (*Poecilia reticulate*). Aquat. Toxicol. 10, 115.

Deneer, J.W.V., Leeuwen, C.J., Seinen, W., Maas-Diepeveen, J.L., Hermens, J.L.M., 1989. QSAR study of the toxicity of nitrobenzene derivatives towards *Daphnia magna*, *Chlorella pyrenoidosa*, and *Photobacterium phosphoreum*. Aquat. Toxicol. 15, 83.

Fukui, K., Yonezawa, T., Nagata, C., 1954. Theory of substitution in conjugated molecules. Bull. Chem. Soc. Jpn. 27, 423.

Karcher, W., Devillers, J., 1990. SAR and QSAR in environmental chemistry and toxicology: scientific tool or wishful thinking? In: Karcher, W., Devillers, J. (Eds.), Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology. Kluwer Academic, Dordrecht, The Netherlands, pp. 1–12.

Katritzky, A.R., Petrukhin, R., Tatham, D., Basak, S., Benfenati, E., 2001. Interpretation of QSPR and QSAR relationships. J. Chem. Inf. Comput. Sci. 41, 679.

Lang, P.Z., Ma, X.F., Lu, G.H., Wang, Y., Bian, Y., 1996. QSAR for the acute toxicity of nitroaromatics to the carp (*Cyprinus carpio*). Chemosphere 32, 1547.

MINITAB, 2006. State College, PA Minitab, Inc.

Mekenyan, O., Roberts, D.W., Karcher, W., 1997. Molecular orbital parameters as predictors of skin sensitization potential of halo- and pseudohalobenzenes acting as S_NAR electrophiles. Chem. Res. Toxicol. 10, 994.

Parr, R.G., Szentpaly, L.V., Liu, S., 1999. Electrophilicity index. J. Am. Chem. Soc. 121, 1922.

Purdy, R., 1988. In: Turner, J.E., Williams, M.W., Schultz, T.W., Kwaak, N.J. (Eds.), QSAR. U.S. Dept of Energy, Oak Ridge, p. 99.

Please cite this article in press as: Bellifa, K., Mekelleche, S.M. QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis* using quantum chemical descriptors. Arabian Journal of Chemistry (2012), http://dx.doi.org/10.1016/j.arabjc.2012.04.031

Roberts, D.W., 1987. An analysis of published data on fish toxicity of nitrobenzenes and aniline derivatives. In: Kaiser, K.L.E. (Ed.), QSAR in Environmental Toxicology – II. D. Reidel, Dordrecht, The Netherlands, pp. 295–308.

Sarkar, U., Padmanabhan, J., Parthasarathi, R., Subramanian, V., Chattaraj, P.K., 2006. Toxicity analysis of polychlorinated dibenzofurans through global and local electrophilicities. J. Mol. Struct. Theochem. 758, 119.

Schultz, T.W., 1999. Structure–toxicity relationships for benzenes evaluated with *Tetrahymena pyriformis*. Chem. Res. Toxicol. 12, 1262.

Schultz, T.W., Cronin, M.T.D., Walker, J.D., 2003. Quantitative structure–activity relationships (QSARS) in toxicology: a historical perspective. THEOCHEM 622, 1.

Smiesko, M., Benfenati, E., 2004. Predictive models for aquatic toxicity of aldehydes designed for various model chemistries. J. Chem. Inf. Comput. Sci. 44, 976.

Stewart, J.J.P., 1989. MOPAC program package. Quantum chemistry program exchange.

Veith, G.D., Mekenyan, O.G., 1993. A QSAR approach for estimating the aquatic toxicity of soft electrophiles. Quant. Struct.-Act. Relat. 12, 349.

Yuan, X., Lu, G., Lang, P., 1997. QSAR study of the toxicity of nitrobenzenes to river bacteria and *Photobacterium phosphireum*. Bull. Eviron. Contam. Toxicol. 58, 123.

Zhao, Y.H., Yuan, X., Ji, G.-D., Sheng, L.X., Wang, L.S., 1997. Quantitative structure–activity relationships of nitroaromatic compounds to four aquatic organisms. Chemosphere 34, 1837.

Zoeteman, B.C.J., Harmsen, K., Linders, J.B.H.J., Morra, C.F.H., Slooff, W., 1980. Persistent organic pollutants in river water and ground water of the Netherlands. Chemosphere 9, 231.

Please cite this article in press as: Bellifa, K., Mekelleche, S.M. QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis* using quantum chemical descriptors. Arabian Journal of Chemistry (2012), http://dx.doi.org/10.1016/j.arabjc.2012.04.031

ملخص

يهدف العمل المقدم في هذه الأطروحة إلى

1- تطوير نماذج QSAR موثوقة وتنبؤية لتحديد العوامل الجزيئية الرئيسية شرح تأثير التفاعل (الإلكترونية / اختراق) والتي تنظم سمية مشتقات نترات البنزين.
2- إنشاء نماذج QSAR موثوقة تفسيرية لسمية المركبات نيترو العطرية، وتنوير وتفسير آلية سمية هذه المنتجات: دورة الأكسدة والاختزال وهجوم أليف النواة باستخدام نظرية DFT.
3- تطويرنماذج تنبؤية لسمية سلسلة من الفينولات الهالوجينية وجها لوجه رباعية الغشاء yriformisp باستخدام عدد قليل من الواصفات ذات الصلة واحترام جميع بروتوكول لمنهجية QSAR.
وأجري العمل باستخدام النهج الإحصائية، حالة الانحدار الخطي بسيطة ومتعددة. أجريت الحسابات مع برامج Gaussian , MOPAC وذلك باستخدام أساليب 6PM، 3PM، 1AM وG31-6 /3B YPL / DFT *

الكلمات المفتاحية: سمية ; نترات البنزين; الفينولات الهالوجينية; واصفات نظرية.

RESUME

Le travail présenté dans cette thèse a pour objectifs :

1- D évelopper des modèles QSAR fiables et prédictifs pour identifier les principaux facteurs moléculaires expliquant l'effet d'interaction (électronique /pénétration) et régissant la toxicité des dérivés de nitrobenzène.
2- Etablir des modèles QSAR explicatifs fiables pour la toxicité des composés nitro-aromatiques, éclairer et interpréter le mécanisme de la toxicité de ces produits : Succession-Redox et Attaque Nucléophile en utilisant la théorie de DFT.
3- Elaborer des modèles prédictifs pour la toxicité d'une série de phénols halogénés vis-à-vis *Tetrahymena pyriformis* en utilisant un nombre réduit de descripteurs pertinents et en respectant tout le protocole de la méthodologie QSAR.

Le travail a été mené à l'aide des approches statistiques, en l'occurrence la régression linéaire simple et multiple et des approches quantiques. Les calculs ont été effectués avec les programmes Mopac, G*aussian ,* en utilisant les méthodes AM1, PM3, PM6 et DFT/B3LYP/6-31G*

Mots-Clés :Toxicité; Nitrobenzènes ; phénols halogénés ; Descripteurs théoriques

ABSTRACT

The work presented in this thesis aims to:

1- Develop reliable and predictive QSAR models to identify key molecular factors explaining the interaction effect (electronic / penetration) and governing the toxicity of nitrobenzene derivatives.
2- Establish reliable explanatory QSAR models for toxicity of nitro-aromatic compounds, enlighten and interpret the mechanism of the toxicity of these products: Redox Cycling and Nucleophilic Attack by using the theory of DFT.
3- Develop predictive models for the toxicity of a series of halogenated phenols against *Tetrahymena pyriformis* using a small number of relevant descriptors and respecting all the protocol of the QSAR methodology.

The work was conducted using statistical approaches, the simple and multiple linear regression case and quantum approaches. The calculations were performed with the programs Mopac, Gaussian, using methods AM1, PM3, PM6 and DFT / B3LYP / 6-31G *

Key-Words: Toxicity; Tetrahymena; Nitrobenzenes; halogenated phenols; Theoretical Descriptors.