MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE «ABBES LAGHROUR» DE KHENCHELA
FACULTE DES SCIENCES ET DE LA TECHNOLOGIE
DEPPARTEMENT SCIENCE DE LA MATIERE

N° de série : …

Mémoire de fin d'études

Pour l'obtention du diplôme de Master (L.M.D)

Spécialité : Chimie Analytique et Environnement

Intitulé :

Modélisation du potentiel de demi-vague de réduction d'une série des composés organiques

Réalisé par :

-HANOU Kenza

-GAAGAI Khouloud

Dirigé par* : *KERTIOU Noureddine

Membres de jury :

SAMAI Salima	*MCB*	*Présidente*
BOUAAKADIA Amel	*MCA*	*Examinatrice*

Année Universitaire 2021-2022

Remerciements

Avant tout, nous tenons à remercier **le DIEU** qui illumine notre chemin et qui nous a armés de courage et de volonté pour réaliser ce travail.

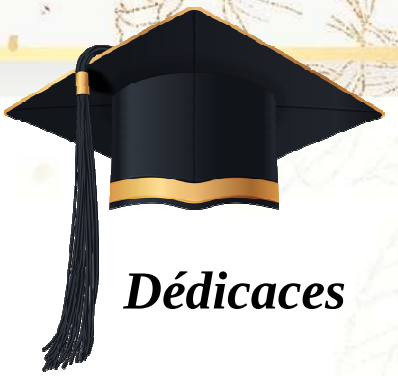
Nous tenons à remercier nos parents, qui nous procurent un soutien affectif, moral et émotionnel.

Nous exprimons nos sincères remerciements à notre Encadreur, le **Dr. KERTIOU Noureddine,** pour sa disponibilité, son aide, son suivi, son support, ses encouragements et ses précieux conseils durant notre travail malgré ses obligations et ses préoccupations, et surtout pour la rigueur intellectuelle avec laquelle elle a partagé tous ses savoirs. *Nous avons été heureux de travailler avec lui*
Nous tenons également à exprimer nos sincères remerciements aux membres de jury :

Dr. SAMAI Salima
et
Dr. BOUAAKADIA Amel

Pour avoir accepté d'examiner ce travail

Enfin nous tenons à remercier tous nos professeurs qui nous en enseignés durant notre cursus Et à toute personne qui nous nous aidons de prés ou de loin à achever notre mémoire.



Dédicaces

Avec l'expression de ma reconnaissance, je dédie ce modeste travail à ceux qui, quels que soient les termes embrassés, je n'arriverais jamais à leur exprimer mon amour sincère.

A ma chère mère, Fatiha
A mon cher père, Tayeb

Qui n'ont jamais cessé, de formuler des prières à mon égard, de me soutenir Et de m'épauler pour que je puisse atteindre mes objectifs.

A mes très chers frères, Salah Eddine et Ziad
Pour ses soutiens moral, Puisse Dieu vous donne santé, bonheur, courage et surtout réussite

A l'âme de mon grand-père Omar

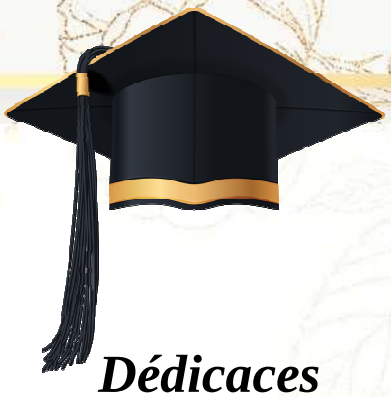
A ma chère grand-mère, Hafsia
Qui je souhaite une bonne santé.

A mes chers ami (e) s,
Hanane, Rania, Amel, Dounia, Wided, Islam, khouloud, Chihab, Malak
Pour leurs aides et supports dans les moments difficiles.

A toute ma famille,
A tous mes autres ami(e)s,
Aux étudiants de la promotion de Master Chimie analytique et environnement de l'année 2021/2022.

A tous ceux que j'aime et ceux qui m'aiment.

Kenza



Dédicaces

C'est avec une grande joie que je dédie ce modeste mémoire,

À tous mes proches particulièrement ***:***

A mon père Abd Elouassie

A ma très chère maman Fahima

à mes chers frères

Taki Eddine, Salah Eddine

A mes chères sœurs

Khawla, Anfel

A mes grands-pères et mes grands-mères

A mon mari ***Oussama***

A mes meilleurs amis

Mouna, Chaima, Donya, Amel, Zanouba ,Lina ,Hiba Wahiba ,Sara ,Abir, Nabila,Warda ,Nesrine ,Talin ,Kenza , Toulin ,Malek, Tawba , Tesnim, Salsabi, Jinan .

Khouloud

SOMMAIRE

Partie théorique

Application

LISTE DES TABLEAUX

LISTE DES FIGURES

SYMBOLES ET ABREVIATIONS

Symboles	Définitions
ACP	Analyse en composantes principales.
AG	Algorithme génétique (Genetic Algorithm).
Du	D total accessibility index / unweighted
ei	Différence entre les valeurs observées et estimées.
ei std	Résidu de prédiction standardisé.
$E_{1/2}$	Le potentiel de demi-vague
F	Statistique de Fisher.
FIV	Facteur d'inflation de la variance.
H	Matrice de projection, ou matrice chapeau.
h_{ii}	Eléments diagonaux de la matrice chapeau.
k	Nombre de descripteurs.
LMO	leave – many- out.
LOO	leave – one – out.
MCO	Les moindres carrés ordinaires
MCP	Les moindres carrés partiels.

MLR	Régression linéaire multiple.
MM^+	Mécanique Moléculaire.
n	Dimension de la population.
n-p	Nombre de degrés de liberté.
p	Nombre de descripteurs en comptant la constante (Nombre de paramètres).
PLS	Moindres carrés partiels.
PRESS	Somme des carrés des erreurs de prédiction.
Q^2	Coefficient de prédiction.
Q^2ext	Coefficient de prédiction externe
QSAR	Quantitative Structure/ Activity Relationships. Relations Structure/Activités Quantitatives).
QSPR	Quantitative Structure/ Propriety Relationships. Relations Structure/Propriétés Quantitatives).
R^2	Coefficient de détermination.
RMSE	Root Mean Squared Error.
RNA	Réseau de neurones artificiel.
S	Erreur standard.
SCE	Somme des carrés des écarts.
SCT	Somme des carrés totale.
SDEC	Standard Deviation Error in Calculation : Déviation standard de l'erreur calculée.

SDEP	Standard Deviation Error of Prediction : Déviation standard de l'erreur de prédiction.
$SDEP_{ext}$	External Standard Deviation Error of Prediction: Déviation standard de l'erreur de prédiction externe.
t	t de Student.
X	Matrice des valeurs observées des variables explicatives.
X'	Matrice transposée de X.
y	Vecteur de dimension n.
yi	Valeur observée.
ŷi	Valeur estimée.

Introduction générale

La croissance d'outil informatique fiable avec le développement de leur puissance a permis la mise en œuvre des techniques de modélisation moléculaire qui sont aujourd'hui des outils importants dans les domaines de prédiction des propriétés physiquo-chimiques et la conception des nouvelles molécules.

Au cours des dernières décennies, la modélisation moléculaire a connu des développements très importants dans des nombreuses branches, à savoir la structure électronique des atomes, des molécules et des complexes organométalliques, l'évaluation de leurs spectres et propriétés magnétiques ou encore la structure des molécules d'intérêt biologique. C'est un ensemble de techniques pour étudier et traiter des problèmes chimiques sur un ordinateur, sans avoir besoin d'aller au laboratoire pour des expériences.

L'utilisation de méthodes alternatives à l'expérimentation, parmi lesquelles les relations quantitatives structure propriété/activité QSPR/QSAR (en anglais : Quantitatifs Structure Propriety/Activity Relationships) sont devenues d'un grand intérêt et sont même recommandées dans les nouvelles règlementations [1, 2].

Le développement de modèles mathématiques QSPR/QSAR reliant les propriétés physico-chimiques et/ou les activités biologiques à la structure moléculaire, d'une part, peut expliquer l'origine de ces propriétés/activités, et d'autre part, prédire leurs propriétés expérimentales.

Les principales étapes de la construction d'un modèle QSPR/QSAR peuvent être décrites comme suit : Extraire les descripteurs des structures moléculaires, sélectionner les descripteurs adaptés à l'étude liée au problème exposé, et utiliser les descripteurs comme variables explicatives pour définir comment ils se rapportent aux propriétés/activités en question. Chaque modèle doit être validé sur un ensemble de données de test.

L'objectif de ce travail vise à présenter brièvement les différentes outils employées pour la mise en place des modèles QSPR/QSAR et leurs évaluations: base de données expérimentales, descripteurs moléculaires, sélection des descripteurs pertinentes, méthodes d'analyse de données, techniques de validation (interne et externe) et détermination des domaines d'applicabilité [3, 4].

Les descripteurs moléculaires typiques utilisés se répartissent généralement en trois catégories : physico-chimiques, topologiques ou électroniques, qui peuvent être déterminés de manière empirique ou par des expériences théoriques en chimie computationnelle. Ces descripteurs sont des caractéristiques de la structure bidimensionnelle (2D) ou tridimensionnelle (3D) de la molécule.

Les techniques les plus courantes pour établir des modèles QSPR utilisent l'analyse de régression (la régression multilinéaire utilisée dans cette étude : MLR; la projection des structures latentes par les moindres carrés partiels : PLS), les réseaux neuronaux RNA, et les méthodes de classification.

Le potentiel demi-vague ($E_{1/2}$) est une propriété électrochimique importante des composés organiques. Cette constante caractéristique d'un système d'oxydo-réduction réversible peut être utile pour prédire d'autres propriétés électrochimiques des composés organiques [5]. Il y a beaucoup de méthodes électrochimiques qui permettent la détermination de ce potentiel d'une large variété de composés organiques, inorganiques et organométalliques [6]. Une stratégie réussie pour la prédiction des potentiels de réduction est la construction des modèles QSPR. P. Tompe et al. ont rapporté une étude quantitative de la relation structure-électrochimie sur le potentiel demi-vague de Cétones α,β-insaturées en solution non gazeuse d'acétonitrile [7]. Ils ont trouvé une relation linéaire entre les constantes des substituants électroniques et $E_{1/2}$. H. Li et al. ont utilisé des indices topologiques pour corréler le potentiel demi-vague de différentes classes de composés organiques [8].

Nous avons appliqué une méthode hybride: algorithme génétique/régression multilinéaire (GA/MLR) pour modéliser, le potentiel de demi-vague de 68 composés organiques industriellement importantes dans la perspective du génie chimique.

Partie théorique

I Introduction

Les relations quantitatives structure-activité/propriété (QSAR/QSPR) sont de plus en plus utilisées, du fait de la croissance des moyens de calculs. Très récemment, la mise en place du nouveau règlement européen REACH, qui recommande leur utilisation pour limiter le recours à l'expérience, donne un nouvel essor au développement de tels modèles prédictifs. Dans les dernières années, l'utilisation des méthodes QSAR n'a cessé de progresser. Elle est même devenue indispensable en chimie pharmaceutique et pour la conception de médicaments. Leur développement dans une gamme plus large d'applications leur ouvre d'ailleurs de grandes perspectives (ex : solubilité, points d'ébullition, températures critiques, densité . . . etc.). Il s'agit de présenter ici le principe des modèles QSPR ainsi que ceux des différents outils employés pour leur mise en place et leur évaluation : bases de données expérimentales, descripteurs, outils d'analyse de données.

I.1 Historique de (QSAR)

Les relations quantitatives structure-activité/propriété (QSAR/QSPR) sont de plus en plus utilisées, du fait de la croissance des moyens de calculs. Elles ont été abondamment utilisées dans les industries pharmaceutiques, chimiques et cosmétiques, tout particulièrement pour la conception rationnelle de nouveaux principes actifs et de nouvelles entités chimiques.

Les premiers travaux utilisant la méthode (QSAR) ont commencé au début des années 60 avec Hansch [9] d'une part et de Free et Wilson [10] d'autre part, qui ont proposé un modèle mathématique pour corréler l'activité biologique et la structure chimique.

Les développements de cette étude (QSAR) sont très anciens, à partir de 1868, lorsqu'Alexander Crum-Brown et Thomas Fraser montrent l'existence d'une relation entre l'activité physiologiques et la structure chimique [11], suivis, avec Richet qui établit une relation entre la toxicité et les propriétés physicochimiques [11, 12], De manière indépendante, Meyer et Overton ont décrit une corrélation linéaire entre la lipophile (coefficient de partage huile-eau) et les effets biologiques (narcotiques) [13, 14].

Pendant les dernières décennies, ce domaine a largement été étudié et les données bibliographiques disponibles sur cette approche sont maintenant importantes [15].

I.2 Définition

Les méthodes QSAR sont basées sur l'hypothèse que l'activité ou la propriété d'un composé chimique est liée à sa structure, plus précisément cette approche affirme que l'activité et la structure d'un composé chimique sont liées d'un certain algorithme mathématique, cela est basé sur le postulat de base « les composés chimiques similaires ont des activités similaires». De plus, lorsque les paramètres moléculaires sont exprimés par des chiffres, on peut proposer une relation mathématique, ou relation quantitative structure activité, entre les deux.

Par définition, Une QSAR est un modèle mathématique qui associe un ou plusieurs paramètres quantitatifs dérivés de la structure chimique, à une mesure quantitative d'une activité **[16].**

I.3 QSAR/QSPR

La notion SAR (Structure-Activity Relationship) (relation entre la structure et l'activité) tente à établir une relation entre les différentes caractéristiques comportementales des composés chimiques (exemple : activité et toxicité) et les informations issues de leurs structures chimiques, en d'autres termes, l'étude SAR offre la possibilité d'établir une équation mettant en jeu l'activité/propriété/toxicité spécifique des produits chimiques en utilisant des informations sur leurs structures chimiques, du coup l'expression quantitative de l'activité d'un produit chimique définit l'étude QSAR. L'axiome central de la modélisation QSAR repose sur la présentation de la réponse chimique en termes de propriétés moléculaires, sachant que chaque propriété contenant une information chimique significative peut être employée comme descripteur. Une fois l'équation est établie, la méthode QSAR nous permet la prédiction de l'activité du produit chimique étudié, en outre, la méthode QSAR met l'accent sur la modification de la structure chimique pour obtenir les produits chimiques d'intérêt avec les valeurs de réponse désirées. L'appellation est influencée par le point final modélisé, par conséquence le processus peut être défini comme suit : QSAR/QSPR/QSTR pour activité/propriété/toxicité.

L'équation mathématique : Réponse = f (propriétés structurale/chimiques) [17].

L'analyse QSAR a été créé afin de remplir les objectifs suivants : (a) la prédiction de nouveaux analogues avec une meilleure activité, (b) améliorer la compréhension et l'investigation du mode d'action de produits chimiques et pharmaceutiques, (c) l'optimisation de la molécule type en congénères moins toxiques, (d) la rationalisation des expérimentations

humides (QSAR offre une alternative économique et rapide aux essais in vitro à débit moyen ainsi qu'aux essais in vivo à faible débit) [17].

I.4 Principe

Le principe d'une étude QSAR/QSPR **(Figure -01-),** consiste à trouver une relation mathématique reliant de manière quantitative une activité/propriété, mesurée pour une série de composés similaires dans les mêmes conditions expérimentales, avec des descripteurs moléculaires à l'aide des méthodes statistiques. L'objectif de ces études est d'analyser les données structurales afin de détecter les facteurs déterminants pour l'activité ou la propriété étudiée. Pour ce faire, différents types de méthodes statistiques peuvent être employées.

L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de l'activité/propriété étudiée pour de nouvelles molécules ou des molécules pour lesquels les données expérimentales ne sont pas disponibles.

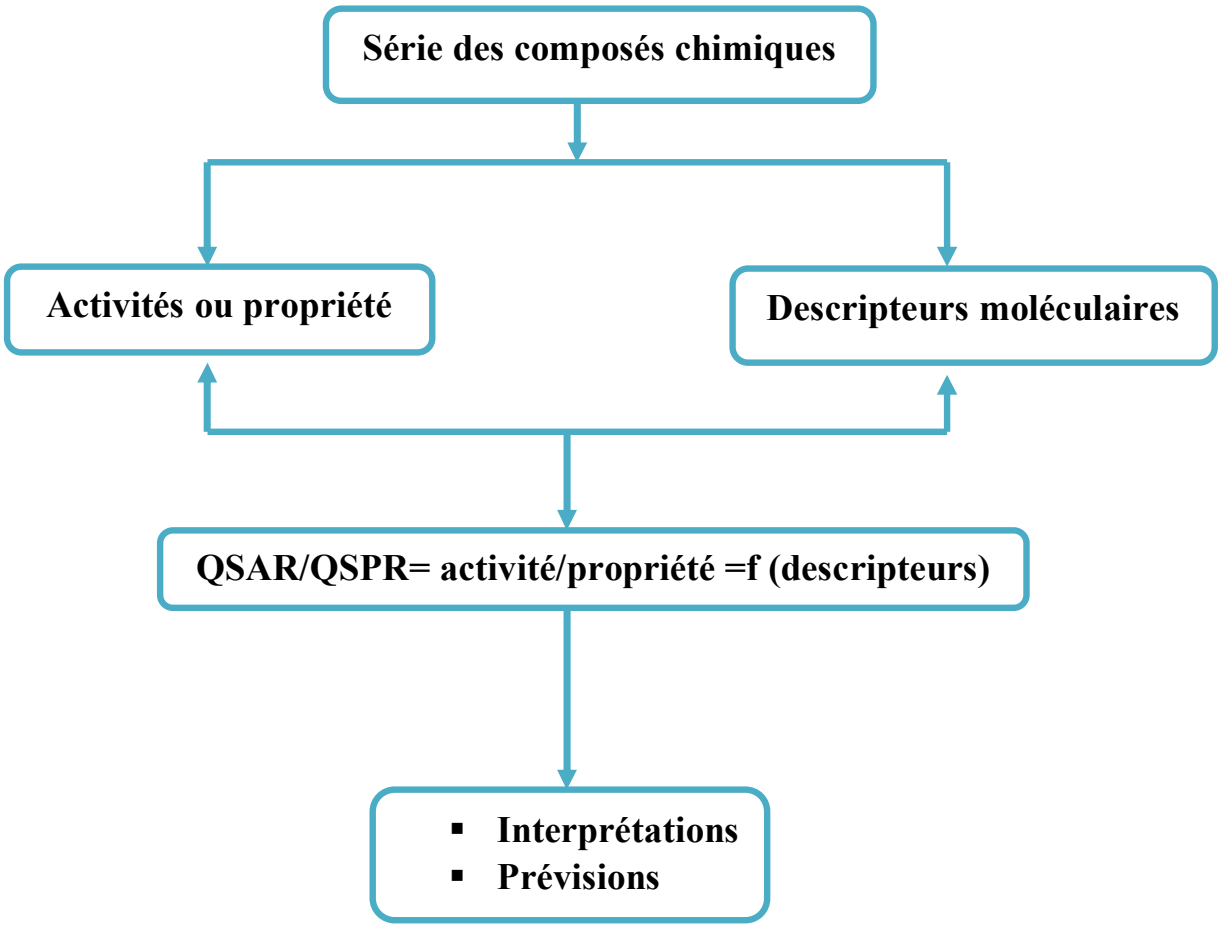


Figure -01- Modèle de l'étude de relation structure activité

I.5 Méthodologie générale d'une étude QSPR/QSAR

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

- Collecte d'une base de données.
- Recherche de descripteurs adéquats pour l'activité/proriété étudiée.
- Le choix d'une méthode d'analyse des données.
- Validation du modèle.
- Validation interne : on utilise la série d'apprentissage constituée de 2/3 de 4 la base de données. Ce type de validation a pour but de vérifier la stabilité et la robustesse du modèle retenu.
- Validation externe : on utilise la série de test constituée généralement de 1/3 de la base de données. Le but de cette validation est de vérifier le pouvoir prédictif du modèle élaboré [18].

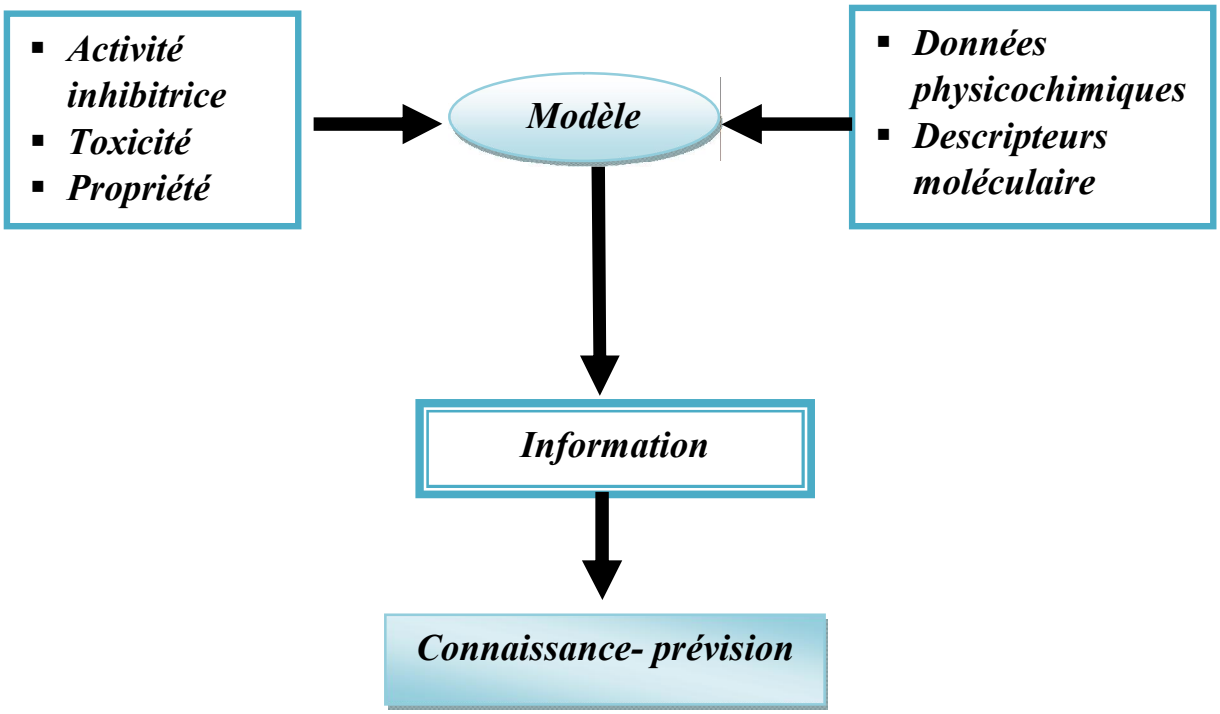


Figure -02- *Présentation de la méthodologie de QSAR*

I.6 Les applications de l'étude QSAR

Il existe un grand nombre d'application de ces modelés tels que :

• L'optimisation de l'activité pharmacologique.

• La conception rationnelle de nombreux autres produits tels que des agents tensioactifs, des parfums, des colorants et des produits chimiques fins.

• L'identification des composés dangereux dans les premiers stades de développement des produits ou la projection des stocks de composés existants.

• La prédiction de la toxicité et les effets secondaires de nouveaux composés.

• La prédiction de la toxicité pour les espèces environnementales.

• La sélection des composés ayant des propriétés pharmacocinétiques optimales, que ce soit la stabilité ou la disponibilité dans les systèmes biologiques.

• La prédiction d'une variété de propriétés physico-chimique des molécules.

• La prédiction du devenir des molécules qui sont libérées dans l'environnement.

• La prédiction des effets conjugués des molécules, que ce soit dans des mélanges ou des formulations **[19].**

I.7 Les méthodes mathématiques utilisés par le model QSPR

Pour les méthodes utilisées en QSPR sont deux types:

a) Linéaires

- Régression linéaire simple
- Régression linéaire multiple MLR.
- Régressions aux moindres carrées partielles (PLS).

b) Non linéaires

- Réseau de neurones artificiel RNA.
- SVM.Arbres de décision.

II COLLECTE DES DONNEES

Le potentiel de demi-vague est l'un des grandeurs physiques importantes. Les données utilisées dans ce travail, concernent 68 composés, sont réunies Dans le tableau -01-: On a choisi aléatoirement (17) composés pour validation ; et le reste (51) pour la calibration ou à la construction du modèle.

Tableau-1- Nomenclature et valeurs de la propriété étudiée

N	Composés	$E_{1/2}$(V)
1	Anthraquinone	-0,54
2	2,3-Dimethyl naphtoquinone	-0,22
3	Duroquinone	-0,09
4	Toluquinone	0,09
5	Methyl o-nitrobenzoate	-0,25
6	Methyl m-nitrobenzoate	-0,24
7	Methyl p-nitrobenzoate	-0,2
8	o-Nitroaniline	-0,29
9	p-Nitroanisole	-0,35
10	m-Nitrobenzaldehyde	-0,28
11	o-Nitrobenzoic acid	-0,23
12	m-Nitrobenzoic acid	-0,2
13	p-Nitrobenzoic acid	-0,17
14	o-Nitrophenol	-0,23
15	p-Nitrophenol	-0,35
16	m-Nitrotoluene	-0,22
17	p-Nitrotoluene	-0,24
18	Acetaldehyde	-1,89
19	Acrolein	-1,36
20	Benzaldehyde	-0,94
21	Formaldehyde	-1,59
22	Furfural	-1,06
23	Glyoxal	-1,41

Tableau -1- *(suite)*

N	Composées	$E_{1/2}$(V)
24	p-Hydroxybenzalhedyde	-1,16
25	o-Methoxybenzaldehyde	-1,03
26	p-Methoxybenzaldehyde	-1,07
27	Salicylaldehyde	-1,02
28	Acridine	-0,8
29	Pyridine	-1,49
30	Quinaldinic acid	-0,86
31	Quinoline	-1,23
32	Quinoline-8-carboxylic acid	-1,11
33	Saccharin	-1,77
34	Ascorbic acid	-0,17
35	Bromoacetic acid	-0,54
36	α-Bromopropionic acid	-0,39
37	Crotonic acid	-1,94
38	Dibromoacetic acid	-0,03
39	Diethyl fumarate	-0,84
40	Diethyl maleate	-0,95
41	Ethyl dichloroacetate	-0,86
42	Fumaric acid	-1,6
43	Trichloroacetic acid	-0,84
44	Allyl chloride	-1,91
45	Allyl bromide	-1,29
46	Benzotrichloride	-0,68
47	Benzyl chloride	-1,94
48	Bromobenzene	-2,32
49	n-Butyl bromide	-2,27
50	p-Dibromobenzene	-0,78

Tableau -1-(suite et fin)

N	Composés	$E_{1/2}$(V)
51	Nitromethane	-0,83
52	p-Hydroxybenzalhedyde	-1,39
53	o-Methoxybenzaldehyde	-1,41
54	p-Methoxybenzaldehyde	-1,43
55	Salicylaldehyde	-1,44
56	p-Nitroaniline	-0,36
57	m-Nitrophenol	-0,37
58	o-Nitrotoluene	-0,26
59	Crotonaldehyde	-0,92
60	Methyl glyoxal	-0,83
61	8-Hydroxyquinoline	-1,39
62	Nicotinamide	-1,56
63	Acrylonitrile	-1,94
64	Maleic acid	-1,36
65	Methylacrylonitrile	-2,07
66	Pyruvic acid	-0,86
67	Benzal chloride	-1,81
68	m-Dichlorobenzene	-0,3

II.1 Préparation de base des données

II.1.1 Calcul du modèle :

Les molécules sont dessinées par le logiciel ChemDraw (ChemDraw ultra 7.0) puis elles sont optimisées en utilisant le logiciel HyperChem [20] .Les descripteurs moléculaires ont été calculés à l'aide du logiciel informatique Dragon [21] plus de 1600 descripteurs sont calculés. L'ensemble des données a été décomposé en deux sous-ensembles aléatoirement, 75% de la totalité des composés pour la construction du modèle et 25% pour la validation externe.

D'après l'algorithme génétique dans la version MobyDigs **[22]** plusieurs modèles sont obtenues pour chaque jeu de groupe; le choix a été opté pour le modèle qui conduit aux meilleurs statistiques des 100 modèles générés par algorithmes génétiques.

II.1.2 Logiciels « ChemDraw»

ChemDraw est l'outil complet destiné aux chimistes et biologistes, intégrant toute une gamme d'outils intelligents permettant de faciliter les travaux des chercheurs au quotidien.

Utiliser pour créer des publications prêt, dessins scientifiquement significatifs de molécules et de réactions [23]. Essayant de concevoir et dessiner de nouvelles molécules en utilisant les différents outils disponibles dans ChemDraw, et de les visualiser, et plus important encore, les enregistrer dans différents formats. ChemDraw est très pratique pour les réactions d'écriture à l'aide de produits chimiques. On peut réaliser des structures dans différents formats disponibles [24]. Exemple : Anthraquinone

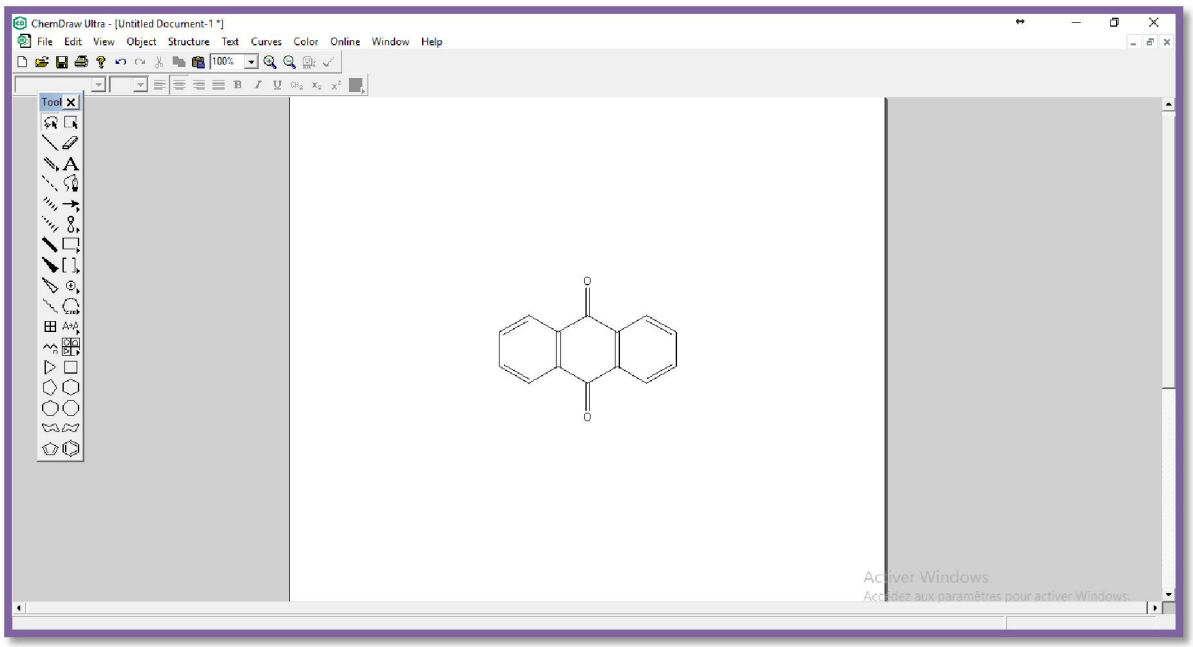


Figure-03- *Représentation des molécules par ChemDraw*

II.1.3 Le logiciel hyperchem professionnel :

HyperChem est un logiciel de modélisation moléculaire chimique développé par Hypercube Inc. Il fonctionne en unissant la visualisation et l'animation 3D avec des calculs de chimie quantique, la mécanique moléculaire et dynamique. Il est facile et flexible.

HyperChem est utilisé dans cette étude pour construire et optimiser les molécules, est enregistrée comme un fichier nommé "Hin" après l'optimisation. Nous avons utilisé la méthode semi empirique MM+ pour l'optimisation [25].

III Récupération et stabilisation les molécules de fichier Hin :

III.1 Stabilisation structure des molécules (minimisation de l'énergie) :

Pour stabilise forme de structure de chaque molécule ou minimisation de l'énergie on utilise le HyperChem, pouvez effectuer une minimisation de l'énergie (ou de la géométrie d'optimisation) d'une molécule en utilisant une variété de méthodes de calcul. Les deux mécanismes moléculaires et les méthodes semi-empiriques sont disponibles. Minimisation de l'énergie modifie la géométrie ou la forme d'une molécule d'abaisser l'énergie potentielle de la molécule et pour donner une conformation plus stable [26].

III.2 Mécanique Moléculaire

Les champs de force mécaniques moléculaires utilisent les équations de la mécanique classique Décrire l'énergie potentielle de surface et les propriétés physiques des molécules. Une sorte d'une molécule est décrite comme un ensemble d'atomes en interaction par des fonctions analytiques simples. Cette description s'appelle un champ de force. Une sorte de

Les composantes du champ de force sont l'énergie produite par la compression et l'étirement de l'objet obligations [27]. HyperChem comprend quatre domaines de la mécanique des polymères Power : Nouvelles implémentations de technologies développées et publiées par des groupes Recherche respectée mais nous sommes cohérents avec l'approche MM^{+} dans ce travail Propriétés de la force de champ illustrées dans la **figure -04** :

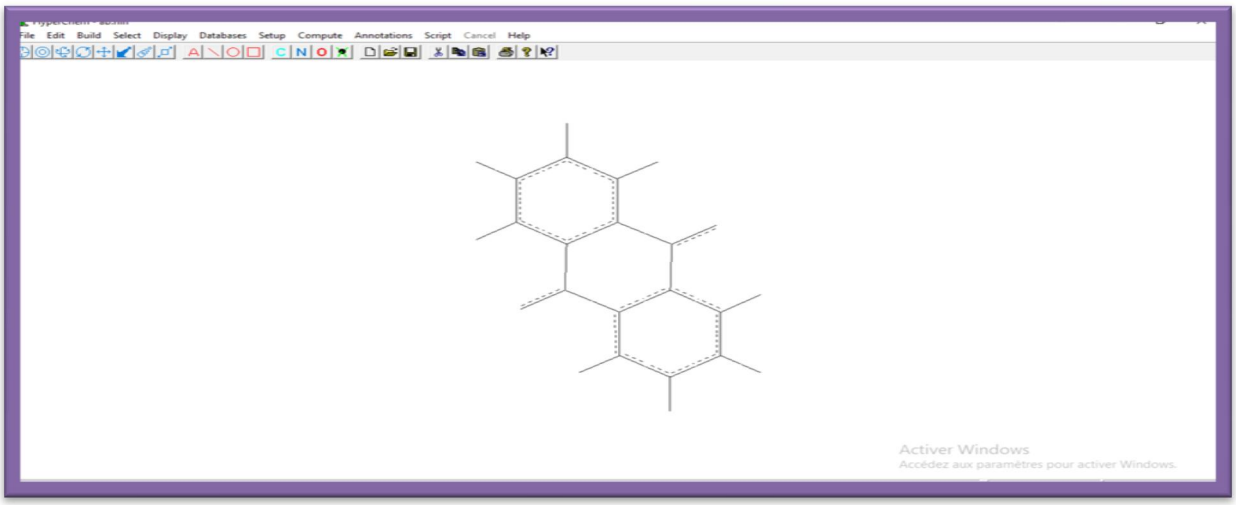


Figure-04- *Le Logiciel Hyperchem*

III.3 Récupération des fichiers HyperChem HIN

Après avoir construit la structure dans HyperChem, nous pouvons l'enregistrer pour une utilisation ultérieure. C'est une bonne idée car cela nous fait gagner du temps si nous voulons revoir notre structure plus tard. Pourquoi le construire deux fois ? ! On pout le faire en allant dans fichier et en enregistrant en le donnant un nom hin. Le fichier peut être rappelé à tout moment pour visualisation et manipulation.

Les structures chimiques des 68 composés de notre ensemble de données ont été établies dans le logiciel HyperChem et pré-optimisées à l'aide du champ MM^{+} ensuite la méthode semi-empirique AM1 [26].

III.4 Le Logiciel DRAGON

C'est une application pour le calcul des descripteurs moléculaires. Ces descripteurs peuvent être utilisés pour évaluer l'influence de la structure moléculaire ou les relations propriétés-structure, aussi pour l'analyse de symétrie et la projection des bases de données des molécules [28].

DRAGON fournis 1664 descripteurs moléculaires qui sont divisés en 20 blocs logiques figure -05- L'utilisateur peut calculer non seulement le descripteurs de type d'atome (Atom type), groupe fonctionnel, comptes de fragment, mais aussi des descripteurs topologiques et géométriques .Quelque propriétés moléculaires, comme logP, (LogKow), molar refractivity, et topological surface area (TPSA) sont calculés par l'utilisation des modèles communs.

Figure-05- *Le logiciel Dragon*

IV Descripteurs :

Le descripteur moléculaire est le résultat final d'une procédure logique et mathématique qui transforme l'information chimique chiffrée dans une représentation symbolique d'une molécule à un nombre utile ou le résultat de quelques expériences standard. Les descripteurs moléculaires sont les traits communs les plus considérables de structure moléculaire qui peut être utilisée pour développer la « Relation Structure – Activité » avec le but de prédire l'activité biologique et propriétés physico-chimique des molécules**. [29, 30].**

IV.1 Types de descripteurs :

Il y a beaucoup de descripteurs moléculaires qui ont été répertoriés, qu'ils soient dérivés de la théorie ou tirés des approches différentes.

IV.1.1 Descripteurs constitutionnels :

Les descripteurs constitutionnels sont directement liés à la formule brute de la molécule, à l'aide de la composition moléculaire, c'est-à-dire les atomes qui le constituent, Il s'agit de :

- La masse molaire.
- Les nombres absolus et relatifs d'atomes (C, H, O, S, N, F, Cl, Br, I, P.…).
- Les nombres absolus et relatifs de groupes fonctionnels (NH_2, COOH, OH. . .).
- Les nombres absolus et relatifs de liaisons (simples, doubles, aromatiques. . .).
- Les nombres absolus et relatifs de cycles (aromatiques ou non).

Ces descripteurs sont très utilisés du fait de leur extrême simplicité non seulement d'un point de vue conceptuel mais surtout calculatoire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution. Autrement dit, si on développe des modèles avec ce type de descripteurs seulement, ils peuvent poser un problème pour l'interprétation des mécanismes d'interaction mis en jeu pour la propriété étudiée [31].

IV.1.2 Descripteurs topologiques :

Les descripteurs topologiques "ou indices topologiques", décrivent la connectivité atomique dans la molécule, ils sont obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Ces descripteurs s'inspirent de la théorie des graphes appliquée à la table de connectivité qui n'est autre qu'une représentation compacte de la connectivité interatomique au sein de la molécule. Les indices topologiques les plus fréquemment utilisés sont l'indice de Wiener, l'indice de Randic, l'indice de connectivité de valence de Kier-Hall et l'indice de Balaban [31].

IV.1.3 Descripteurs géométriques :

Les descripteurs géométriques d'une molécule sont évalués à partir des positions relatives de ses atomes dans l'espace, et décrivent des caractéristiques plus complexes ; leurs calculs nécessitent de connaître, la géométrie 3D de la molécule, par modélisation moléculaire empirique ou ab initio, Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés qui dépendent de la structure 3D. On distingue plusieurs descripteurs importants, le volume moléculaire, la surface accessible au solvant, le moment d'inertie.

Le volume moléculaire est le volume occupé par la molécule en appliquant une grille 3D de cubes dans la boîte parallélépipédique dont les dimensions Xmax, Ymax et Zmax. La surface accessible au solvant SAS, ou la zone de surface accessible est la surface d'une molécule qui est accessible à un solvant, généralement mesuré en unités d'angströms carrés. Le moment d'inertie est une grandeur physique qui caractérise la distribution de masse dans la molécule [31].

IV.1.4 Descripteurs électrostatiques :

Ces descripteurs reflètent les caractéristiques de la distribution de charge de la molécule. Les charges partielles empiriques dans la molécule sont calculées en utilisant l'approche proposée par Zefirov. Cette méthode est basée sur l'échelle d'électronégativité. Sur la base de ces charges partielles les descripteurs électrostatiques suivantes sont calculés comme suivants :

• Les charges partielles minimales et maximales dans la molécule (qmin, qmax).

• Les charges partielles minimales et maximales pour l'atome (C, N, O.…).

• Les indices électroniques topologiques.

Ces descripteurs sont responsables sur des interactions entre les molécules polaires [31].

IV.1.5 Descripteurs thermodynamiques :

Les descripteurs thermodynamiques sont calculés sur la base de la fonction de partition totale Q de la molécule. La fonction de partition commode la façon avec laquelle l'énergie d'un système de molécules est répartie parmi les individus moléculaires. Sa valeur dépend du poids moléculaire, de la température, du volume moléculaire, des distances inter nucléaires, des mouvements moléculaires et des forces intermoléculaires. La fonction de partition est le point le plus commode entre les propriétés microscopiques des molécules individuelles (niveaux d'énergie, moments d'inertie) avec les propriétés macroscopiques (chaleur spécifique, entropie). La molécule peut accroître son énergie de translation, de vibration, de rotation de façon pratiquement indépendante [31].

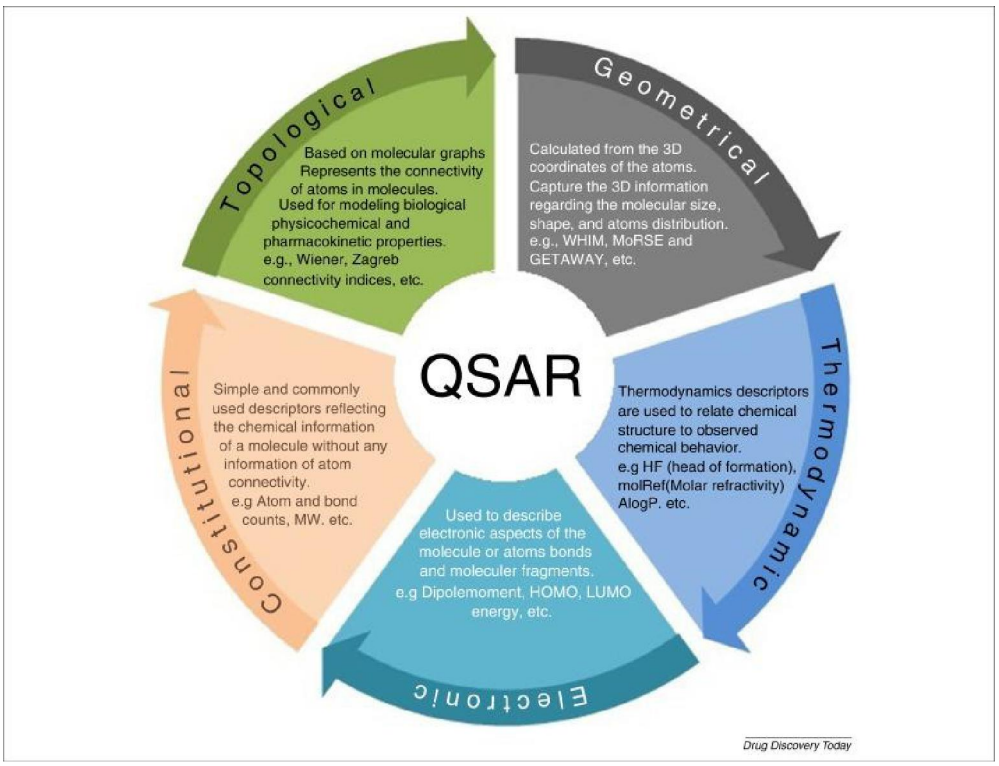


Figure -06- *Représentation des descripteurs moléculaires utilisés à la modélisation QSAR [30]*

IV.2 Travaux bioinformatique pour la sélection des descripteurs :

Dans cette section nous allons présenter les principales approches pour la sélection de descripteurs.

IV.2.1 Les méthodes de filtres :

Les méthodes filtres sont les méthodes les plus simples pour la sélection des caractéristiques, ces approches utilisent les données d'entrainement (Training Data) afin de sélectionner les caractéristiques sans appliquer des algorithmes ou les techniques de l'apprentissage automatique [32].

IV.2.2 Les méthodes de cohérence :

Ce sont des méthodes basées sur la robustesse des données d'entrainement et évaluent essentiellement la cohérence de l'ensemble des caractéristiques sélectionnées. Même si cette procédure est simple permettant de réaliser de petit sous-ensemble, elle a des inconvénients, en fait, elle s'applique uniquement avec des caractéristiques discrètes (discontinues), et si le sous-ensemble consiste de caractéristiques continues, elles doivent être discrétisées en 1er lieu [32].

IV.2.3 Méthodes d'information

Ceux-ci sont des méthodes qui comparent principalement l'information obtenue par la nouvelle caractéristique par rapport à la précédente [32].

IV.2.4 Les méthodes de dépendance :

Les méthodes de dépendance évaluent comment la valeur d'une variable peut être prédite utilisant une valeur d'une autre variable, dans ce cas, la méthode sélectionne la caractéristique qui corrèle le plus avec la classe cible sélectionnée [32].

IV.2.5 Les méthodes de la distance :

Elles constituent une grande classe des Méthodes FS, d'un point de vue général, elles utilisent des distances conventionnelles pour mesurer la similarité entre deux échantillons [32].

IV.2.6 La sélection en avant (Forward sélection) :

C'est une méthode très largement utilisée pour FS. C'est un type spécifique de la régression pas à pas qui commence par un sous ensemble de variables vides et ajoute des caractéristiques une par une à chaque étape. La caractéristique sélectionnée est celle qui permet une meilleure amélioration du model. Cette procédure continue jusqu'à ce qu'il n'y ait aucune caractéristique capable d'améliorer le model. L'inconvénient principal de cette méthode c'est qu'elle tend vers un sur-apprentissage grâce auquel il est important d'avoir un strict critère d'arrêt [32].

IV.2.7 Élimination en arrière :

Cette méthode fonctionne d'une manière opposée à la méthode de la sélection en avant (Forward selection), en fait, elle commence en incluant toutes les caractéristiques, puis elle élimine une par une à chaque étape tout en évaluant la contribution de la caractéristique à l'amélioration du model. Cette procédure n'est pas très utilisée en raison de la production de modèles surchargés [32].

IV.2.8 La sélection progressive :

Celle-ci est probablement la plus utilisée pour FS dans QSAR. C'est une méthode hybride basée sur les deux méthodes précédentes (sélection en avant et élimination en arrière). L'avantage majeur de cette méthode c'est que la variable qui rentre dans le model peut être supprimé après si elle s'est avérée impertinente. En vérité, le processus commence par l'ajout de la variable la plus en corrélation avec le point terminal. à chaque étape, la variable avec la plus de corrélation est ajouté jusqu'à ce qu'il n'y aura plus de variables significatives parmi l'ensemble des caractéristiques [32].

Les descripteurs moléculaires calculés par DRAGON sont divisés en 20 blocs logiques. Ces descripteurs moléculaires sont représentés dans le tableau et la figure suivante :

Tableau- 02- *Quelques blocks des descripteurs calculés par logiciel Dragon*

Classe	***Sous classe***
Descripteurs Constitutionnels	***- Dénombrement des atomes ou des liaisons.*** ***- Descripteurs basés sur les masses atomiques.***
Descripteurs Topologiques	***- Indices topologiques (connectivité).*** ***- Descripteurs théoriques d'information.*** ***- Descripteurs topo-chimiques.***
Descripteurs Géométriques	***- Descripteurs liés à la distance.*** ***- Descripteurs liés à l'aire de la surface.*** ***- Descripteurs liés au volume.*** ***- Descripteurs du champ stérique moléculaire.***
Descripteurs liés à la distribution de charge	***- Charges atomiques partielles.*** ***- Moments électriques moléculaires*** ***- Polarisabilités moléculaires.*** ***- Descripteurs du champ électrique moléculaire.***
Descripteurs liés aux orbitales Moléculaires	***- Energie des OM frontières*** ***- Ordres de liaison*** ***- Indices de réactivité de Fukui.***

Tableau- 02-*(suite et fin)*

Classe	*Sous classe*
Descripteurs température Dépendants	***- Fonctions thermodynamiques.*** ***- Descripteurs facteurs de Boltzmann pondérés.***
Descripteurs de Solvatation	***- Energie électrostatique de solvatation.*** ***- Energie de dispersion de solvatation.*** ***- Enthalpie libre de formation de cavité.*** ***- Descripteurs de liaison hydrogène.*** ***- Entropie de solvatation.*** ***- Descripteurs d'énergie de solvatation linéaire théorique.***
Descripteurs mixtes	***- Descripteurs topographiques.*** ***- Descripteurs électro-topologiques.*** ***- Descripteurs de la charge partielle de l'aire de la surface.***

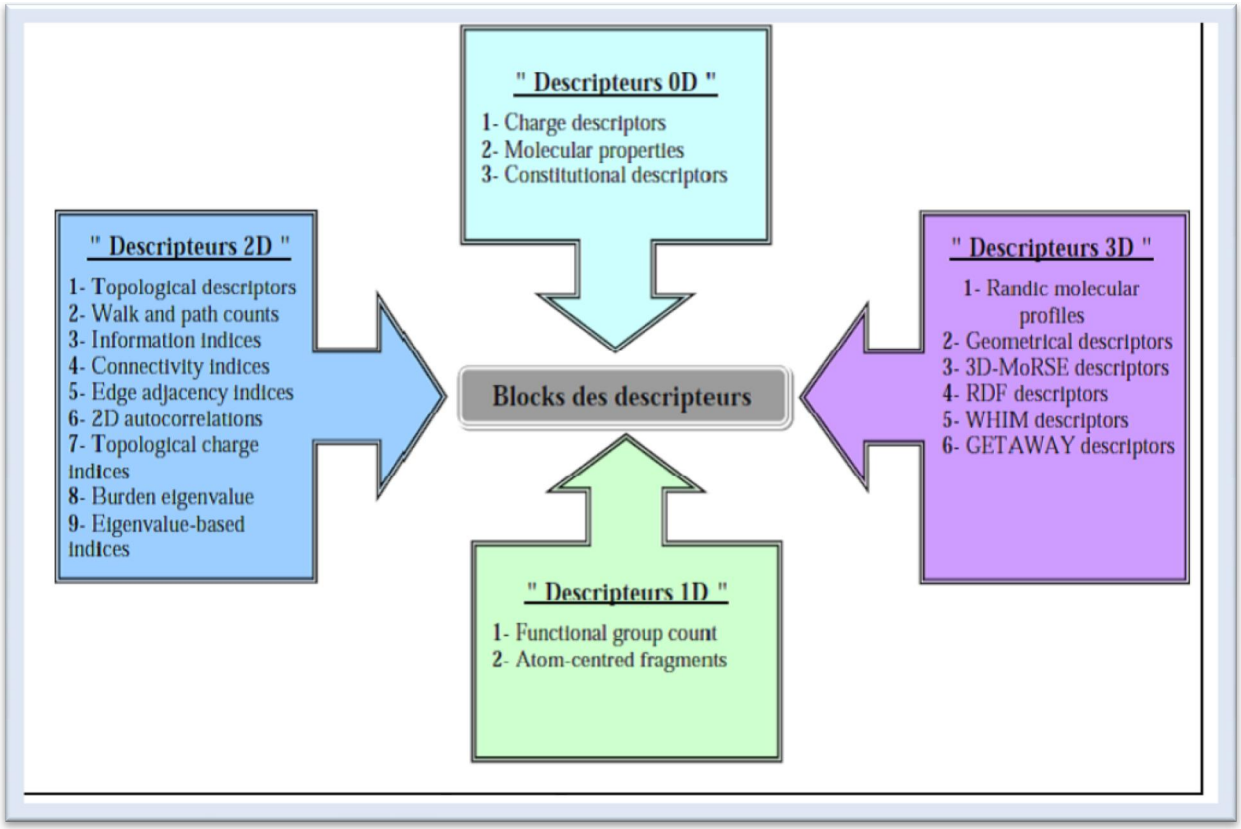


Figure -07- *Une autre* représentation des blocs des descripteurs moléculaires [33]

IV.3 Importance des descripteurs

Les descripteurs moléculaires jouent un rôle important en chimie, en science médicale, la protection de l'environnement, et la recherche.

L'importance des descripteurs peut être résumée en deux points :

- Indication d'une description de la configuration moléculaire à étudier.
- Décrivons tous les paramètres descriptifs de la molécule.

Les descripteurs moléculaires sont utilisés pour, une connaissance de statistiques, chimiometriques, et les principes des approches QSAR/QSPR sont nécessaires en plus de la connaissance spécifique du problème [34].

IV.4 Les étapes de prédiction :

La première expérience de modélisation QSPR développée par Wiener, et depuis, de nouvelles techniques de modélisation d'apprentissage d'abord linéaires, puis non linéaires ont permis la mise en œuvre d'approches multiples, dont la plupart sont fondées sur la recherche d'une relation entre un ensemble de nombres réels, des descripteurs moléculaires et des propriétés ou activités que l'on souhaite prédire [35].

La prédiction des propriétés s'effectué par plusieurs étapes intéressantes illustrant dans le diagramme suivant:

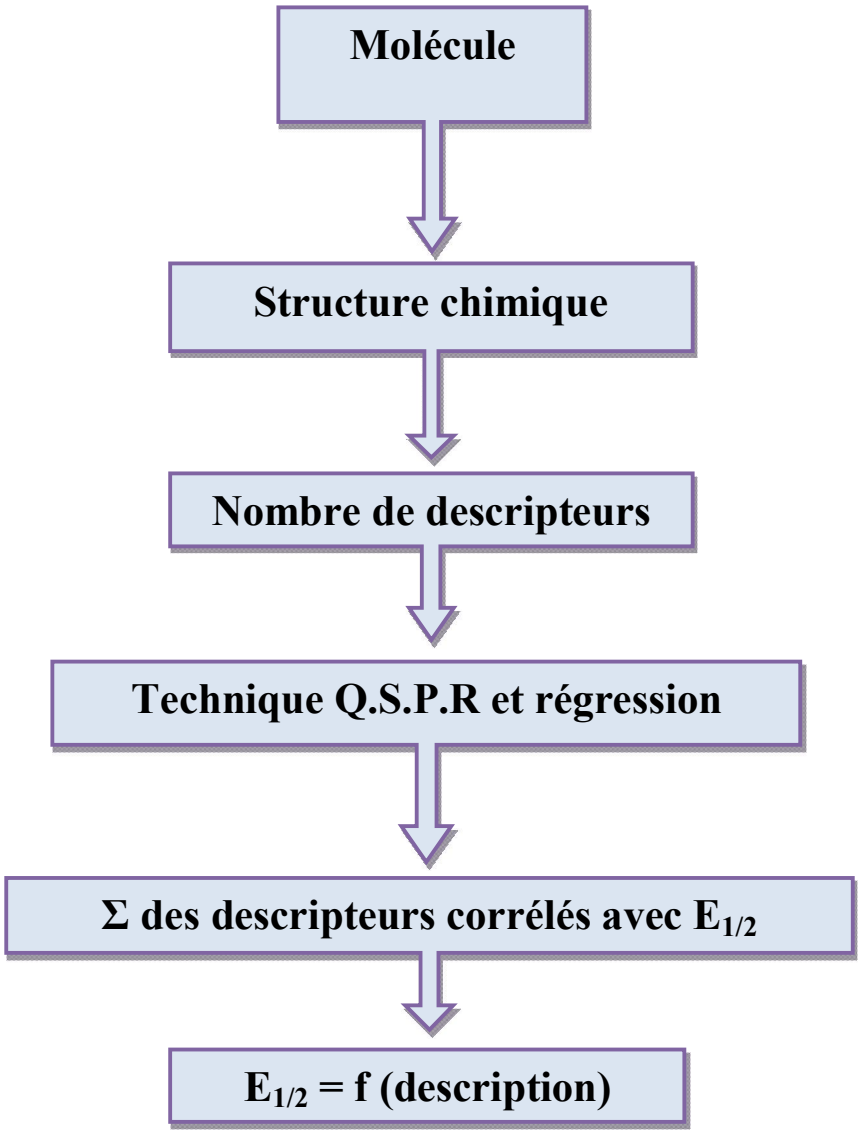


Figure -08- *Diagramme de prédiction par QSPR*

IV.5 L'objectif de la prédiction :

L'objectif principal est de construire un Structures moléculaires, plus précisément descripteurs moléculaires. Ce modèle permet de Suivre la classification du composé qui fait la prédiction

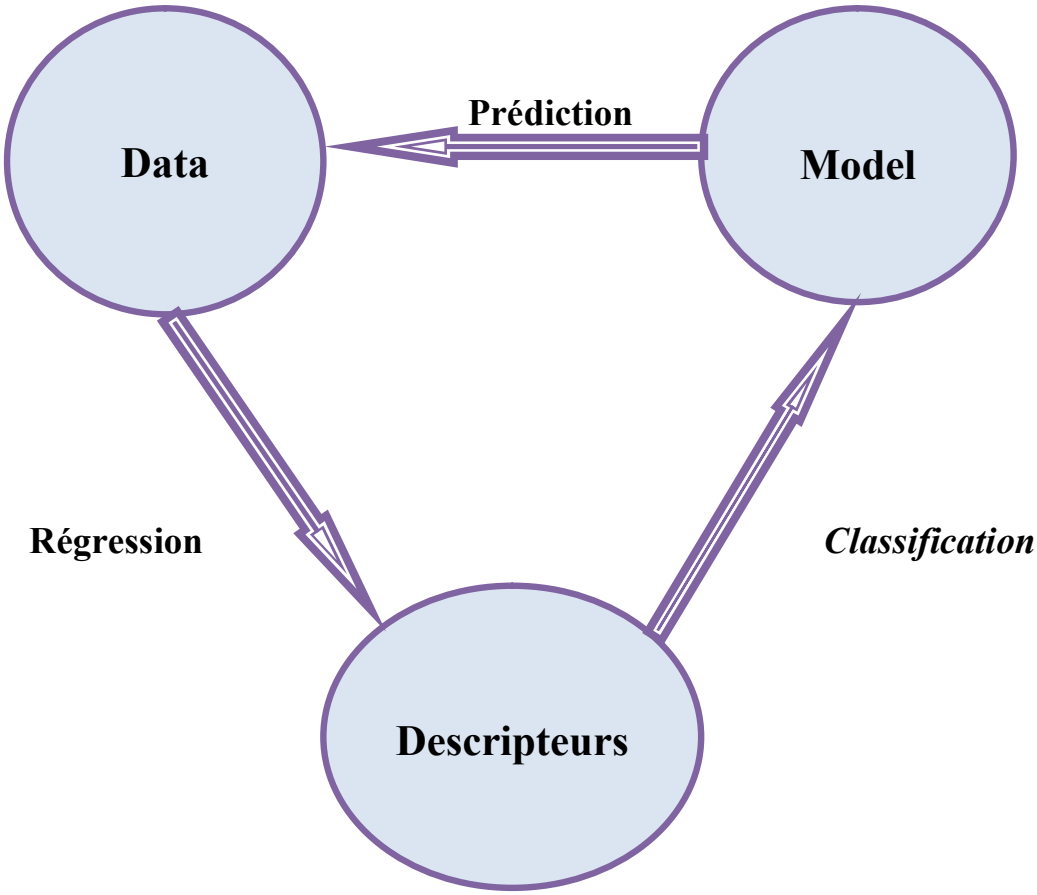


Figure -*09-* *Le cycle de prédiction*

IV.6 Les étapes de travail:

IV.6.1 Modélisation :

Le diagramme suivant illustré les étapes du travail et les techniques quant utilisé dans la Procédure de prédiction de la propriété étudiée (le potentiel de demi-vague).

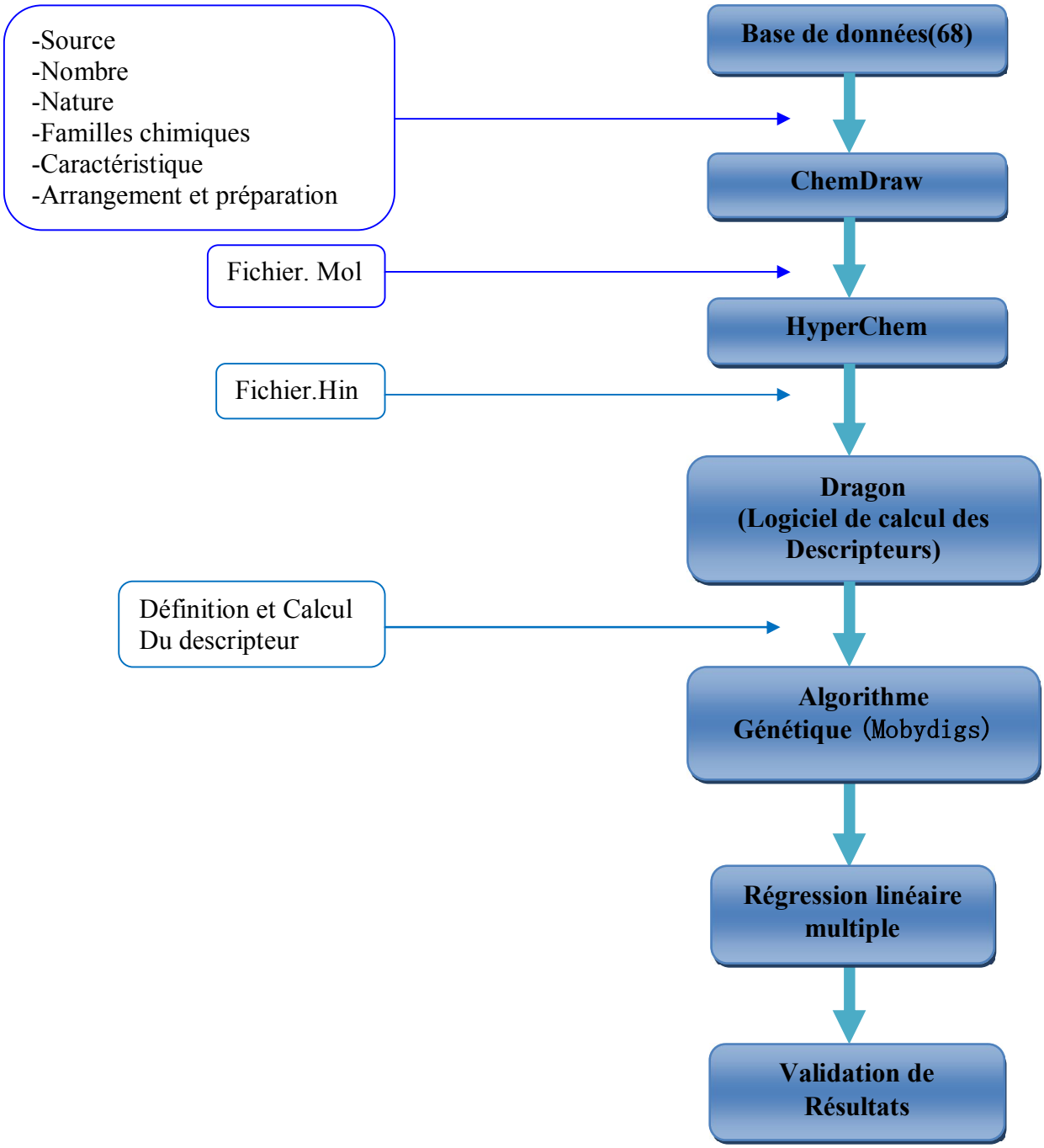


Figure-10- *Diagramme de notre travail*

V La régression linéaire multiple

L'objectif de la régression linéaire simple et multiple : est d'apprendre comment analyser un phénomène quelconque en utilisant des méthodes statistiques dites économétriques.

En effet, la régression linéaire est une relation stochastique entre une ou plusieurs variables. Elle est appliquée dans plusieurs domaines, tels que la physique, la biologie, la chimie, l'économie…etc.

Un modèle de régression linéaire multiple entre une variable expliquée Y et p variables explicatives X_1…, X_P, s'écrit pour tout i=1,…, n :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_i X_{ij} + \varepsilon \qquad (01)$$

Où les y_i, x_{i1}, x_{i2}…, x_{ip} : sont des données respectivement relatives aux variables Y, X_1...,X_p.

Les estimateurs βj sont calculés en utilisant la méthode des moindres carrés ordinaires. Les variables aléatoires ξ représentent les termes d'erreur non observables du modèle. On peut estimer ces erreurs par les résidus ordinaires ei, différence entre les valeurs observées yi et les valeurs estimées ŷi.

Pour construire le modèle et admettre que les coefficients de la régression sont sans biais et convergents, on montre qu'il faut poser comme hypothèses :

a) Les résidus (E) ont une espérance mathématique nulle **: E (e) = 0**
b) Le modèle choisi est correct (aucune variable explicative n'a été omise).
c) Les résidus sont indépendants entre eux :

$$\mathbf{E\ (ei, ej) = 0\ Si\ i \neq j} \qquad (02)$$

Leurs covariances sont nulles.

d) Les résidus ont tous même variance σ2 (propriété d'homoscédasticité).

Par ailleurs, l'emploi de tests statistiques pour analyser la variation expliquée par la régression conduit à admettre que :

Les résidus suivent une distribution normale (de Laplace-Gauss).

L'analyse des résidus présente un intérêt à plusieurs égards. Elle permet en effet de vérifier, a posteriori, la validité du modèle utilisé, en ce qui concerne, d'une part la forme de celui- ci (linéarité ou non linéarité de la relation, par exemple) et d'autre part, certaines hypothèses

plus spécifiques, telles que l'égalité des variances résiduelles, la normalité des résidus ou l'absence d'auto- corrélation.

Pour minimiser l'influence des erreurs de détermination des valeurs explicatives (ou régresseurs) sur la précision des résultats de la régression 5 données (variables dépendantes, ou encore observations) doivent, à la limite, être associées à chaque variable explicative. Le nombre de degrés de liberté final (n-p-1) doit être [36] tel que :

$$n - p - 1 \geq 10 \qquad (03)$$

N'étant la dimension de l'échantillon, et p le nombre de variables explicatives entrant dans la construction du modèle.

V.1 Méthodes de sélection des descripteurs

V.1.1 Algorithme génétique

La modélisation de processus génétiques a initié le développement des algorithmes génétiques, qui peuvent être exploités dans une grande variété de problèmes d'optimisation [37]. Dans un algorithme génétique adapté à l'optimisation, une solution potentielle est considérée comme un individu dans une population. La valeur de la fonction de coût associée à une solution mesure « l'adaptation » de l'individu associé à son environnement. Un algorithme génétique simule l'évolution, sur plusieurs générations, d'une population initiale dont les individus sont mal adaptés au moyen d'opérateurs génétiques de reproduction et de mutation. Après un certain nombre de générations, la population est constituée d'individus bien adaptés, autrement dit des solutions supposées « bonnes » au problème d'optimisation.

Dans ce travail les sélections des descripteurs en utilisant le logiciel de calcul statistique MINITAB version 16.2.0 [38]; et par algorithme génétique, dans la version MOBY DIGS de Todeschini [39].

V.1.2 Evaluation préliminaire des données

Avant d'entamer le développement effectif des équations de régression QSPR, il est hautement recommandé d'examiner la qualité statistique des données de départ, à la fois les

données à corréler (variable dépendante) et les descripteurs utilisés dans la corrélation (variables indépendantes).

On distingue habituellement dans un tel pré- traitement des données les analyses uni variées des analyses bi-variées [40,41].

Dans l'analyse uni-variée, il est recommandé de vérifier la conformité des données à la distribution normale. Une précaution particulière doit être prise lors de la procédure de régression subséquente si les valeurs de la propriété étudiée, ou d'un descripteur, ne suivent pas la loi de Laplace-Gauss. Pour un ensemble de descripteurs différents, il est nécessaire d'effectuer une analyse des données bi-variée, c'est-à-dire de calculer le coefficient de corrélation linéaire R entre chacune des paires de l'ensemble des descripteurs. Si R est statistiquement significatif (R>0,9), ces deux descripteurs ne peuvent être utilisés simultanément lors de l'analyse par MLR.

VI Paramètres statistique

VI.1 Paramètres d'évaluation de la qualité de l'ajustement :

Deux paramètres sont couramment utilisés :

VI.1.1 Le coefficient de détermination multiple :

Pour comprendre la qualité de l'ajustement obtenu, nous avons calculé le coefficient de détermination R², qui représente la fraction de la variation de régression Y (= potentiel demi-vague)″ expliquée ou "raisonnable". Ce paramètre correspond au carré du coefficient de corrélation , entre 0 et 1, exprimé en Toujours un pourcentage.

Si la valeur de R² est proche de 1 ou 100% ; on a donc un excellent ajustement

qualité ; en revanche, si la valeur du R² est faible et proche de 0 ou 0 %, elle est mal ajustée.

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \qquad (04)$$

Où ŷi est la valeur estimée du paramètre physique, et ȳ la moyenne des valeurs expérimentales.

VI.1.2 La racine de l'erreur quadratique moyenne de prédiction (désignée également par SDEP) :

$$SDEP = \sqrt{\frac{1}{n}\sum_{1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2} = \sqrt{\frac{PRESS}{n}} \qquad (05)$$

VI.2 Validation externe :

Il est intéressant, pour juger de la qualité du modèle, de considérer la racine de l'écart Quadratique moyen (RMSE, pour Root Mean Squared Error), calculée sur différents ensembles :

- Ensemble d'estimation (appelée SDEC)
- Ensemble de validation croisée (appelée également SDEP)
- Ensemble de prédiction externe (désignée par SDEPext).

Ces valeurs RMSE sont mieux adaptées, pour juger de la qualité d'un modèle que les valeurs

De R^2 et Q^2 seules, qui constituent de bons tests uniquement pour des données réparties régulièrement.

$$SDEC = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (06)$$

$$SDEP = \sqrt{\frac{\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_{(i)})^2}{n_{ext}}} \qquad (07)$$

La validation croisée du « leave-one-out » (LOO) [45] consiste à recalculer les modèles (n-1) observations et utiliser le modèle résultant pour calculer la quantité d'intérêt Composés rejetés, notés ŷ(i). Répétez le processus pour chaque quantité d'intérêt.

La somme des erreurs de prédiction au carré, désignée par le symbole PRESS, est une mesure

Dispersion estimée. Il est utilisé pour définir des coefficients de prédiction :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \qquad (08)$$

Contrairement à R^2 au coefficient qui augmente avec le nombre de paramètres du modèle, le Facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain

Nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande Signification du coefficient Q^2_{LOO}. Les valeurs Q^2_{LOO} >0,5 sont considérées comme satisfaisantes, et les valeurs supérieures à 0,9 sont excellentes [46]. Si une valeur plus petite de Q^2_{LOO} indique un modèle plus faible, caractérisé par un pouvoir prédictif interne plus faible, pas nécessairement l'inverse. En effet, si une valeur élevée de Q^2_{LOO} est une condition nécessaire à la robustesse et éventuellement au pouvoir prédictif élevé du modèle, alors cette condition seule n'est pas suffisante et peut conduire à une surestimation du pouvoir prédictif du modèle.

Dans le cas de vrais composés externes, 2, 3 éléments ou plus peuvent devoir être jetés en même temps, ce qui conduit à une procédure LMO (Leave-More-Out). Cependant, ces procédures sont rarement signalées comme des résultats QSPR communs et sous-utilisées dans les travaux actuels. Dans le cas où on a suffisamment de données qui n'ont pas servi dans la création du modèle ou après collecte de nouvelles, on peut ou on doit procéder à la validation de ce dernier, c'est la validation externe. La statistique se rapportant à ce procédé, notée Q^2ext, est calculée comme suit :

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2 / n_{ext}}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / n} \qquad (09)$$

Pour une grande valeur de Q^2_{LOO}, une valeur élevée de Q^2_{ext} permet de présager d'une bonne Capacité prédictive du modèle.

VI.3 Facteur d'inflation de la variance [FIV]

Le facteur d'inflation de variance est utilisé pour détecter si le descripteur a une forte association linéaire avec les prédicteurs restants (multi colinéarité entre les prédicteurs). S'il existe une corrélation entre les prédicteurs, le facteur d'inflation de la variance mesure l'augmentation de la variance des coefficients de régression estimés (multi colinéarité).

FIV = 1 signifie aucune relation, s'il n'y a pas de FIV supérieur à 1, le plus grand facteur FIV parmi tous les prédicteurs est souvent utilisé comme indicateur de l'importance de la multi colinéarité, et si FIV> 5-10, le coefficient de régression de la qualité estimée est faible [47].

VI.4 Test de randomisation

Ce test peut mettre en évidence des corrélations dues au hasard. Elle consiste à générer un vecteur "propriété considérée" en randomisant les composantes d'un vecteur réel.

Le modèle QSPR est ensuite calculé sur les vecteurs obtenus (considérés comme des vecteurs expérimentaux réels) selon les méthodes usuelles. Ce processus est répété plusieurs fois (100 fois dans notre cas).

Application

Nous traiterons dans ce travail la propriété physique envisagée ($E_{1/2}$). Rappelant que les 68 composés ont été éclatés aléatoirement en deux sous ensemble comportant (51) pour la calibration et (17) pour la Validation. Les résultats ainsi obtenus seront discutés.

Conditions expérimentales

Nous avons commencés à utiliser les algorithmes génétiques en utilisant le logiciel Mobydigs [42], ce dernier nous permet de calculer une centaine de modèles de différentes dimensions.
Parmi ces modèles nous avons choisi celui porte le moins de descripteurs et obéi aux conditions expérimentales telles que :
La corrélation entre les descripteurs, le FIV et les grandes valeurs de R^2, Q^2 et Q^2ext.

Choix de la taille du modèle :

En traçant la courbe R^2 en fonction de nombre de descripteurs k **(figure-11-)** on peut dire que notre modèle optimal est celui à 6 descripteurs parce qu'après cela l'amélioration est faible lorsqu'on ajoute d'autre descripteurs.

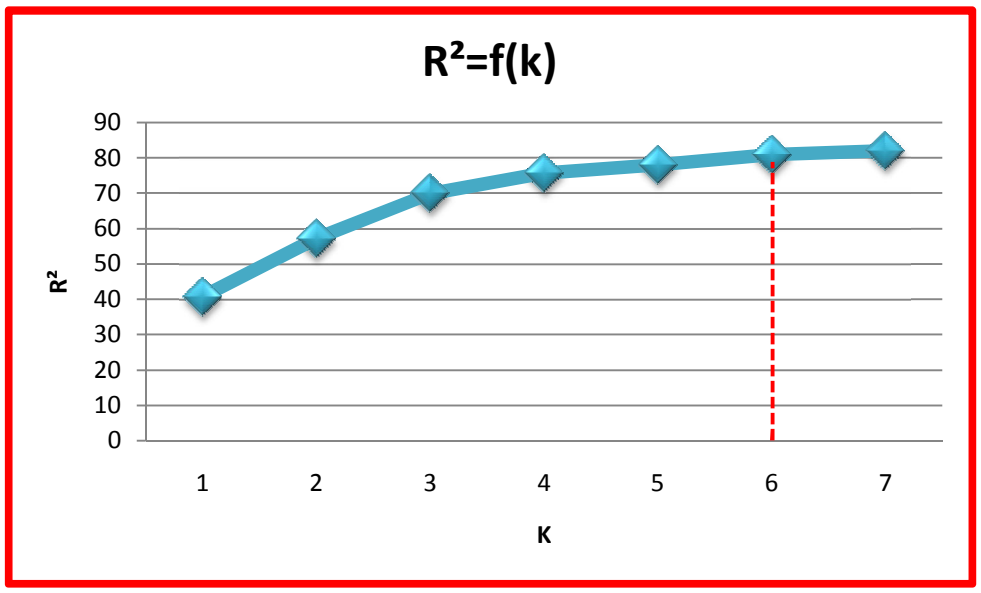


Figure -11- Variation de R^2 en fonction du nombre de descripteur

Sélection des descripteurs

Plusieurs classes de descripteurs moléculaires ont été calculées pour la molécule entière :
Descripteurs relatifs à la construction moléculaire, et descripteurs topologiques.
Parmi les modèles obtenus pour la taille optée, on a sélectionné les 6 descripteurs qui donnent les meilleurs résultats.

Nous rapporterons leurs valeurs, classes et quelques significations dans les tableaux suivants :

Tableau -03-Valeurs des descripteurs moléculaires sélectionnés.

N	Composé	E1/2(V)	MATS6v	MATS4e	Mor02m	Mor15m	Mor16m	Du
1	Anthraquinone	-0,54	-0,228	-0,156	12,785	1,121	-0,318	0,361
2	2,3-Dimethyl naphtoquinone	-0,22	-0,546	-0,059	11,776	0,431	-0,133	0,436
3	Duroquinone	-0,09	0,081	-0,35	9,974	0,166	-0,348	0,548
4	Toluquinone	0,09	1,16	-0,335	8,941	0,205	-0,34	0,39
5	Methyl o-nitrobenzoate	-0,25	-0,234	-0,321	11,51	0,207	-0,089	0,358
6	Methyl m-nitrobenzoate	-0,24	0,108	-0,504	12,34	0,56	-0,202	0,374
7	Methyl p-nitrobenzoate	-0,2	-0,14	-0,521	14,234	0,6	-0,243	0,37
8	o-Nitroaniline	-0,29	0,833	-0,094	10,505	0,358	-0,117	0,371
9	p-Nitroanisole	-0,35	0,287	-0,327	11,754	0,644	-0,241	0,359
10	m-Nitrobenzaldehyde	-0,28	0,632	-0,336	10,568	0,381	-0,248	0,338
11	o-Nitrobenzoic acid	-0,23	0,565	-0,269	11,115	0,166	-0,069	0,333
12	m-Nitrobenzoic acid	-0,2	0,3	-0,458	11,128	0,478	-0,199	0,345
13	p-Nitrobenzoic acid	-0,17	0,044	-0,477	12,874	0,493	-0,23	0,336
14	o-Nitrophenol	-0,23	0,771	-0,035	10,918	0,485	-0,265	0,332
15	p-Nitrophenol	-0,35	0,074	-0,398	10,703	0,498	-0,209	0,33
16	m-Nitrotoluene	-0,22	0,594	-0,133	9,938	0,392	-0,179	0,381
17	p-Nitrotoluene	-0,24	-0,29	-0,159	10,489	0,399	-0,174	0,375
18	Acetaldehyde	-1,89	0	0	2,553	-0,003	-0,093	0,553
19	Acrolein	-1,36	0	-0,575	3,397	-0,014	-0,046	0,324
20	Benzaldehyde	-0,94	0,823	-0,098	7,995	0,385	-0,036	0,344
21	Formaldehyde	-1,59	0	0	1,259	-0,078	-0,044	0,332
22	Furfural	-1,06	0	-0,096	8,107	0,374	0,143	0,363
23	Glyoxal	-1,41	0	0	3,631	-0,09	-0,126	0,476

Tableau-03- (Suite).

N	composé	E1/2.(V)	MATS6v	MATS4e	Mor02m	Mor15m	Mor16m	Du
24	p-Hydroxybenzalhedyde	-1,16	-0,158	-0,169	8,948	0,566	-0,07	0,33
25	o-Methoxybenzaldehyde	-1,03	0,399	0,072	9,056	0,716	-0,171	0,354
26	p-Methoxybenzaldehyde	-1,07	0,372	-0,166	9,832	0,676	-0,107	0,359
27	Salicylaldehyde	-1,02	0,928	0,059	8,806	0,578	-0,142	0,333
28	Acridine	-0,8	-0,244	-0,318	13,85	1,205	-0,229	0,369
29	Pyridine	-1,49	0	-0,047	7,247	0,308	0,1	0,384
30	Quinaldinic acid	-0,86	-0,117	-0,17	12,918	0,855	-0,101	0,334
31	Quinoline	-1,23	0,287	-0,283	10,486	0,755	-0,073	0,365
32	Quinoline-8-carboxylic acid	-1,11	0,049	0,044	11,201	0,733	-0,019	0,357
33	Saccharin	-1,77	1,068	0,006	15,139	0,192	0,511	0,409
34	Ascorbic acid	-0,17	0,07	-0,147	11,463	0,414	-0,029	0,391
35	Bromoacetic acid	-0,54	0	0,352	6,61	0,217	-0,086	0,472
36	α -Bromopropionic acid	-0,39	0	-0,913	8,027	0,378	-0,446	0,504
37	Crotonic acid	-1,94	0,732	-0,129	5,817	0,085	0,004	0,434
38	Dibromoacetic acid	-0,03	0	0,114	8,067	1,583	-1,118	0,503
39	Diethyl fumarate	-0,84	-0,142	-0,388	12,783	0,282	-0,025	0,496
40	Diethyl maleate	-0,95	-0,142	-0,388	10,718	0,29	-0,063	0,439
41	Ethyl dichloroacetate	-0,86	-0,42	-0,207	7,856	0,035	0,01	0,5
42	Fumaric acid	-1,6	0,325	-0,45	7,541	0,189	0,045	0,391
43	Trichloroacetic acid	-0,84	0	-0,91	11,408	0,264	-0,208	0,422
44	Allyl chloride	-1,91	0	-0,3	3,888	-0,063	0,139	0,414
45	Allyl bromide	-1,29	0	-0,286	4,844	0,51	-0,353	0,41
46	Benzotrichloride	-0,68	-1	-0,349	9,406	0,425	0,014	0,405

Tableau-03-(Suite et fin).

N	Composé	E1/2.(V)	MATS6v	MATS4e	Mor02m	Mor15m	Mor16m	Du
47	Benzyl chloride	-1,94	0,429	-0,051	7,558	0,253	0,113	0,396
48	Bromobenzene	-2,32	0	-0,195	7,203	1,145	-0,265	0,313
49	n-Butyl bromide	-2,27	0	-0,019	4,744	0,403	-0,331	0,589
50	p-Dibromobenzene	-0,78	0	-0,493	9,345	1,734	-0,852	0,299
51	Nitromethane	-0,83	0	0	5,149	-0,113	-0,174	0,553
52	Benzoquinone	0,15	0	-0,465	8,46	0,186	-0,307	0,338
53	2-Methyl-1,4-naphtoquinone	-0,17	-0,038	-0,076	11,541	0,801	-0,326	0,385
54	Azobenzene	-0,33	-0,077	-0,148	15,034	0,713	0,039	0,383
55	m-Dinitrobenzene	-0,26	0,485	-0,556	11,676	0,344	-0,369	0,34
56	p-Nitroaniline	-0,36	-0,14	-0,277	10,617	0,483	-0,192	0,336
57	m-Nitrophenol	-0,37	0,665	-0,231	10,105	0,508	-0,223	0,349
58	o-Nitrotoluene	-0,26	0,821	-0,198	10,149	0,265	-0,103	0,399
59	Crotonaldehyde	-0,92	0	-0,049	4,053	0,015	-0,062	0,411
60	Methyl glyoxal	-0,83	0	-0,399	5,169	-0,023	-0,195	0,392
61	8-Hydroxyquinoline	-1,39	0,346	-0,167	10,819	0,922	-0,109	0,345
62	Nicotinamide	-1,56	0,043	0,017	9,162	0,42	0,067	0,358
63	Acrylonitrile	-1,94	0	-1,782	3,051	-0,047	0,019	0,28
64	Maleic acid	-1,36	0,325	-0,45	8,549	0,032	-0,004	0,349
65	Methylacrylonitrile	-2,07	0	-0,598	4,277	-0,022	-0,008	0,4
66	Pyruvic acid	-0,86	0	-1,01	7,068	0,037	-0,156	0,363
67	Benzal chloride	-1,81	-0,167	-0,181	7,86	0,435	-0,029	0,392
68	m-Dichlorobenzene	-0,3	0	0,237	7,208	-0,141	0,099	0,343

-Les 17 derniers composés (52-68) sont destinés à la validation externe.

Tableau- 04- Classes et significations des descripteurs.

Descripteur	classe	Signification
MATS6v	**2D autocorrelations**	Moran autocorrelation - lag 6 /weighted by aomic van der Waals volumes
MATS4e		Moran autocorrelation of lag 4 weighted by Sanderson electronegativity
Mor02m	**3D-MoRSE descriptors**	signal 02 / weighted by mass
Mor15m		signal 15 / weighted by mass
Mor16m		signal 16 / weighted by mass
Du	**WHIM descriptors**	D total accessibility index / unweighted WHIM descriptors (Weighted Holistic Invariant Molecular descriptors) are geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes

Calcul des corrélations entre les différents descripteurs :

Le coefficient de corrélation, r, de Bravais-Pearson a servi pour mettre en évidence les relations possibles entre les différents descripteurs des 51 composés, la matrice de corrélation obtenue à l'aide de la commande "corrélation" du logiciel MINITAB, montre que les descripteurs sont entre eux plus ou moins corrélés.

Les couples des descripteurs qui présentent des valeurs de $r > 0.90$, sont très fortement corrélés et apportent la même information.

Tableau-05- Corrélation entre $E_{1/2}$ et les descripteurs.

Correlations: E1/2.(V); MATS6v; MATS4e; Mor02m; Mor15m; Mor16m; Du

	$E_{1/2}$	MATS6v	MATS4e	Mor02m	Mor15m	Mor16m
MATS6v	-0,011 0,938					
MATS4e	-0,241 0,089	0,173 0,225				
Mor02m	0,569 0,000	0,045 0,754	-0,264 0,062			
Mor15m	0,187 0,189	-0,117 0,413	-0,018 0,899	0,407 0,003		
Mor16m	-0,409 0,003	0,111 0,437	0,167 0,242	0,002 0,988	-0,637 0,000	
Du	-0,139 0,331	-0,231 0,103	0,132 0,357	-0,327 0,019	-0,346 0,013	-0,095 0,505

Cell Contents: Pearson correlation
P-Value

Cette matrice nous permette de voir que les descripteurs Mor02m (57%) et Mor16m (40%) ont une corrélation acceptable avec la propriété, par contre les autres sont moins corrélés mais ils portent un complément pour le modèle. Egalement tous les descripteurs ne sont pas corrélés entre eux et déjà confirmer par les valeurs de p> 0.05, à l'exception de Mor15m avec Mor16m (64%) ont un p<0.05.

Equation de régression :

L'équation de régression du modèle calculé est la suivante :

$$E_{1/2}(V) = - 0,666 - 0,309\ MATS6v + 0,773\ MATS4e + 0,201\ Mor02m - 1,89\ Mor15m - 3,29\ Mor16m - 3,73\ Du \quad (10)$$

Tableau-06- Paramètres de régression

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0,6662	0,3520	-1,89	0,065	
MATS6v	-0,3088	0,1071	-2,88	0,006	1,248
MATS4e	0,7727	0,2003	3,86	0,000	1,455
Mor02m	0,20104	0,01650	12,18	0,000	1,711
Mor15m	-1,8860	0,2010	-9,38	0,000	3,865
Mor16m	-3,2882	0,2898	-11,35	0,000	2,933
Du	-3,7310	0,7428	-5,02	0,000	1,741

Analyse de régression

Les valeurs de T des descripteurs (mor02m, mor16m et mor15m) sont proches et avec des grandes valeurs qui nous montre une grande contribution, mais les autres leur contribution est faible dans le modèle.

Les valeurs des VIF (< 5) suggèrent que ces descripteurs sont faiblement corrélés les uns avec les autres. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

Tableau -07- *Les Valeurs expérimentales, calculées, prédites et leurs erreurs pour l'ensemble de calibration*

N	Composé	$E_{1/2}$ Exp	$E_{1/2}$ Calc	$E_{1/2}$ Pred	Hat	Err. Calc.	Err. Pred.	Std.Err. Calc.	Std.Err. Pred.
1	Anthraquinone	-0,54	-0,56	-0,56	0,104	-0,02	-0,02	-0,08	-0,09
2	2,3-Dimethyl naphtoquinone	-0,22	-0,18	-0,17	0,154	0,04	0,05	0,16	0,19
3	Duroquinone	-0,09	-0,17	-0,19	0,168	-0,08	-0,1	-0,31	-0,37
4	Toluquinone	0,09	-0,21	-0,29	0,21	-0,3	-0,38	-1,2	-1,52
5	Methyl o-nitrobenzoate	-0,25	0,04	0,08	0,13	0,29	0,33	1,1	1,27
6	Methyl m-nitrobenzoate	-0,24	-0,4	-0,41	0,058	-0,16	-0,17	-0,57	-0,61
7	Methyl p-nitrobenzoate	-0,2	0,12	0,16	0,106	0,32	0,36	1,22	1,36
8	o-Nitroaniline	-0,29	-0,56	-0,58	0,086	-0,27	-0,29	-1	-1,1
9	p-Nitroanisole	-0,35	-0,41	-0,41	0,041	-0,06	-0,06	-0,2	-0,21
10	m-Nitrobenzaldehyde	-0,28	-0,16	-0,15	0,092	0,12	0,13	0,45	0,49
11	o-Nitrobenzoic acid	-0,23	-0,14	-0,13	0,103	0,09	0,1	0,33	0,37
12	m-Nitrobenzoic acid	-0,2	-0,41	-0,42	0,055	-0,21	-0,22	-0,77	-0,81
13	p-Nitrobenzoic acid	-0,17	0,11	0,14	0,096	0,28	0,31	1,06	1,17
14	o-Nitrophenol	-0,23	-0,02	0,02	0,139	0,21	0,25	0,81	0,94
15	p-Nitrophenol	-0,35	-0,33	-0,33	0,06	0,02	0,02	0,08	0,09
16	m-Nitrotoluene	-0,22	-0,53	-0,54	0,054	-0,31	-0,32	-1,12	-1,19
17	p-Nitrotoluene	-0,24	-0,17	-0,16	0,1	0,07	0,08	0,26	0,29
18	Acetaldehyde	-1,89	-1,9	-1,91	0,172	-0,01	-0,02	-0,06	-0,07
19	Acrolein	-1,36	-1,46	-1,49	0,237	-0,1	-0,13	-0,4	-0,53
20	Benzaldehyde	-0,94	-1,28	-1,31	0,089	-0,34	-0,37	-1,27	-1,39
21	Formaldehyde	-1,59	-1,36	-1,27	0,277	0,23	0,32	0,96	1,33
22	Furfural	-1,06	-1,64	-1,69	0,083	-0,58	-0,63	-2,16	-2,35
23	Glyoxal	-1,41	-1,13	-1,09	0,119	0,28	0,32	1,07	1,21
24	p-Hydroxybenzalhedye	-1,16	-1,02	-1,01	0,067	0,14	0,15	0,53	0,56
25	o-Methoxybenzaldehyde	-1,03	-1,02	-1,02	0,07	0,01	0,01	0,03	0,03

Tableau-07-(suite et fin)

N	Composé	$E_{1/2}$ Exp	$E_{1/2}$ Calc	$E_{1/2}$ Pred	Hat	Err. Calc.	Err. Pred.	Std.Err. Calc.	Std.Err. Pred.
26	p-Methoxybenzaldehyde	-1,07	-1,2	-1,2	0,049	-0,13	-0,13	-0,46	-0,48
27	Salicylaldehyde	-1,02	-1	-1	0,119	0,02	0,02	0,07	0,08
28	Acridine	-0,8	-0,95	-0,98	0,158	-0,15	-0,18	-0,58	-0,68
29	Pyridine	-1,49	-1,59	-1,6	0,07	-0,1	-0,11	-0,36	-0,39
30	Quinaldinic acid	-0,86	-0,69	-0,68	0,084	0,17	0,18	0,63	0,69
31	Quinoline	-1,23	-1,41	-1,43	0,084	-0,18	-0,2	-0,67	-0,74
32	Quinoline-8-carboxylic acid	-1,11	-1,05	-1,04	0,087	0,06	0,07	0,23	0,26
33	Saccharin	-1,77	-1,52	-1,33	0,428	0,25	0,44	1,2	2,09
34	Ascorbic acid	-0,17	-0,64	-0,66	0,044	-0,47	-0,49	-1,72	-1,79
35	Bromoacetic acid	-0,54	-0,95	-1,04	0,169	-0,41	-0,5	-1,61	-1,94
36	α -Bromopropionic acid	-0,39	-0,88	-1,09	0,294	-0,49	-0,7	-2,1	-2,97
37	Crotonic acid	-1,94	-1,62	-1,58	0,106	0,32	0,36	1,22	1,37
38	Dibromoacetic acid	-0,03	-0,14	-0,24	0,474	-0,11	-0,21	-0,55	-1,05
39	Diethyl fumarate	-0,84	-0,65	-0,62	0,137	0,19	0,22	0,72	0,83
40	Diethyl maleate	-0,95	-0,75	-0,73	0,056	0,2	0,22	0,75	0,8
41	Ethyl dichloroacetate	-0,86	-1,08	-1,11	0,104	-0,22	-0,25	-0,83	-0,93
42	Fumaric acid	-1,6	-1,56	-1,56	0,094	0,04	0,04	0,14	0,16
43	Trichloroacetic acid	-0,84	-0,46	-0,37	0,2	0,38	0,47	1,5	1,87
44	Allyl chloride	-1,91	-2	-2,01	0,136	-0,09	-0,1	-0,34	-0,4
45	Allyl bromide	-1,29	-1,24	-1,24	0,082	0,05	0,05	0,17	0,19
46	Benzotrichloride	-0,68	-1,09	-1,2	0,198	-0,41	-0,52	-1,65	-2,05
47	Benzyl chloride	-1,94	-1,64	-1,62	0,072	0,3	0,32	1,09	1,18
48	Bromobenzene	-2,32	-1,82	-1,71	0,191	0,5	0,61	1,96	2,42
49	n-Butyl bromide	-2,27	-1,6	-1,42	0,211	0,67	0,85	2,7	3,43
50	p-Dibromobenzene	-0,78	-0,75	-0,74	0,336	0,03	0,04	0,12	0,18
51	Nitromethane	-0,83	-0,91	-0,92	0,152	-0,08	-0,09	-0,31	-0,36

La valeur critique pour déterminer les points leviers correspond à :

$$h^* = \frac{3p}{n} = \frac{3(k+1)}{n} = \frac{3 \times 7}{51} = 0.41 \qquad (11)$$

Toutes les valeurs des paramètres statistiques de calibration sont regroupées dans le **tableau-08-**

Tableau -08- *Valeurs des paramètres statistiques pour l'ensemble de calibration*

N	**51**
R^2	**83,1 %**
Q^2	**75,58 %**
F	**36,05**
S	**0,2807**
SDEC	**0,2607**
SDEP	**0,3134**

Pour la robustesse du modèle est assurés par la valeur de Q^2_{LOO}> 75% alors que les valeurs de L'erreur quadratique moyenne de prédiction et de calcul sont petites et proches ; en plus ce modèle est significatif avec une valeur du paramètre de Fisher : (F=36 ,05).

Diagramme de williams

Le domaine d'application a été discuté à l'aide du diagramme de Williams (**figure-12-).**

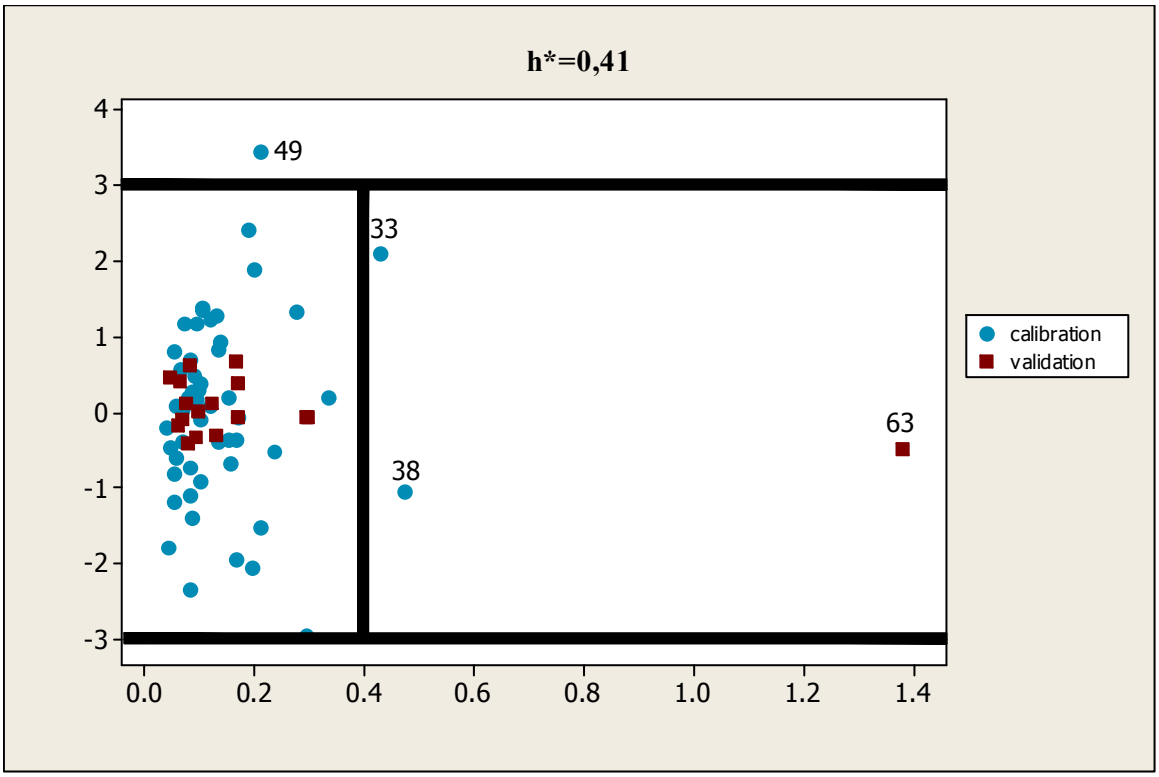


Figure-12- *Diagramme de Williams*

Toutes les erreurs standardisées sont comprises entre les limites ±3 à l'exception du composé 49 (n-Butyl bromide) le diagramme de williams fait ressortir trois points influents ce sont les composés (33) et (38) de l'ensemble de calibration et le composé (63) de l'ensemble de validation.

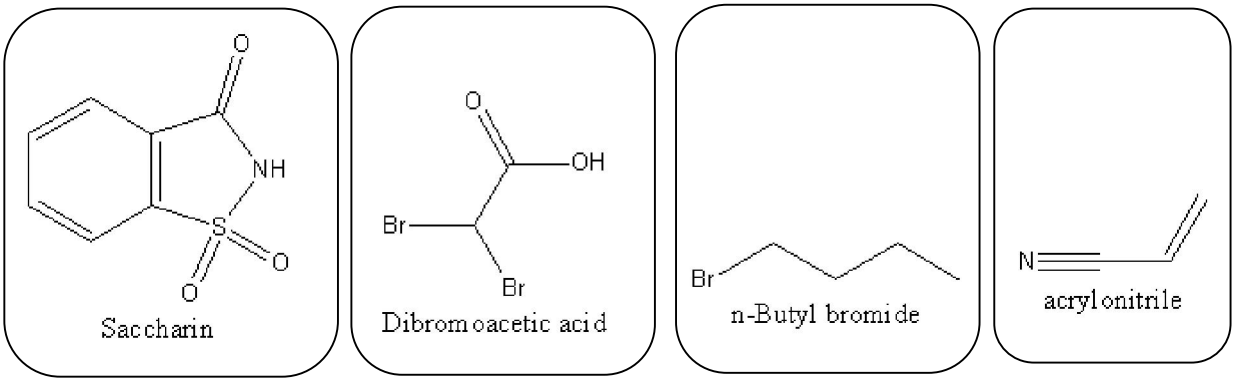


Figure -13- Les 04 composés influents

Vérification de la qualité de l'ajustement

La qualité de l'ajustement a été vérifiée en représentant les valeurs calculées de $E_{1/2}$ avec notre modèle en fonction des celles observées ou expérimentales. La (figure-14-) montre un bon ajustement traduit par une faible dispersion autour de la droite de régression.

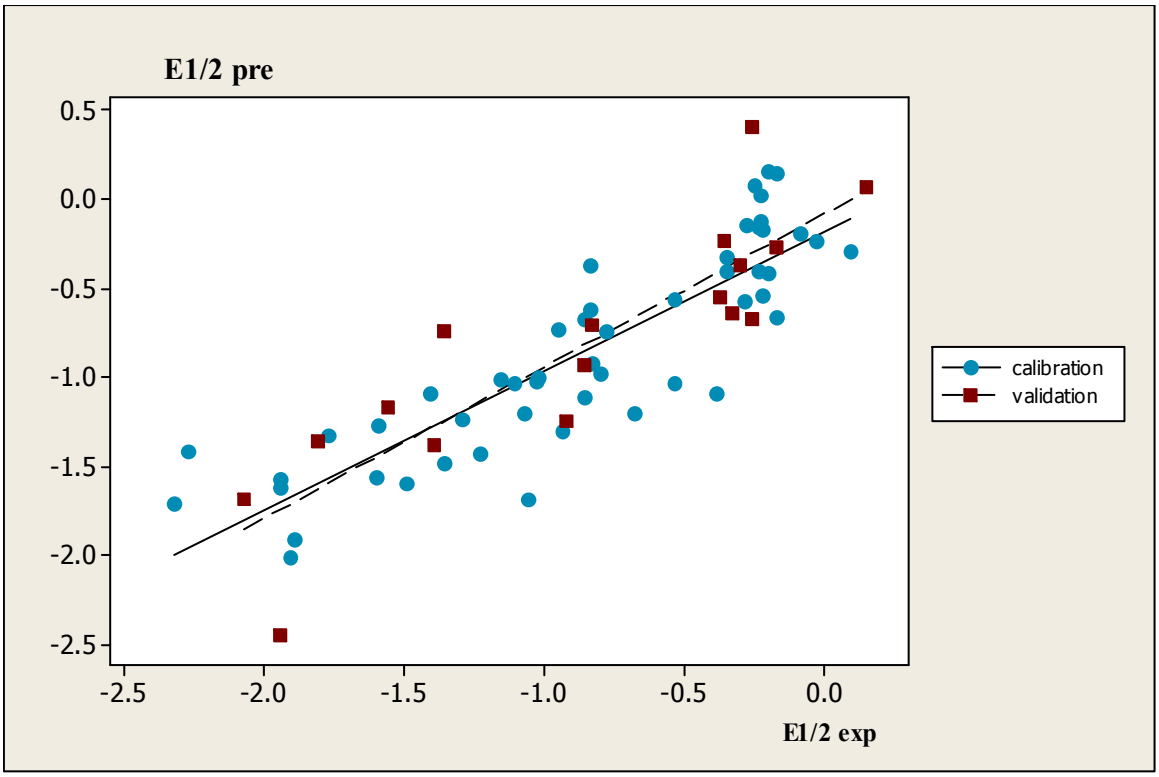


Figure-14-*Graphe des valeurs $E_{1/2}$ calculées en fonction des valeurs expérimentales*

Test de randomisation :

Dans le but d'établir que le modèle obtenu n'est pas dû au hasard ou à une sur spécification, nous avons appliqué le test de randomisation. Ainsi 100 nouveaux vecteurs de potentiel de demi-vague ont été générés par permutation des positions des composantes du vecteur réel.la figure -15- qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (carrée bleu) au modèle réel de départ (carré orange).

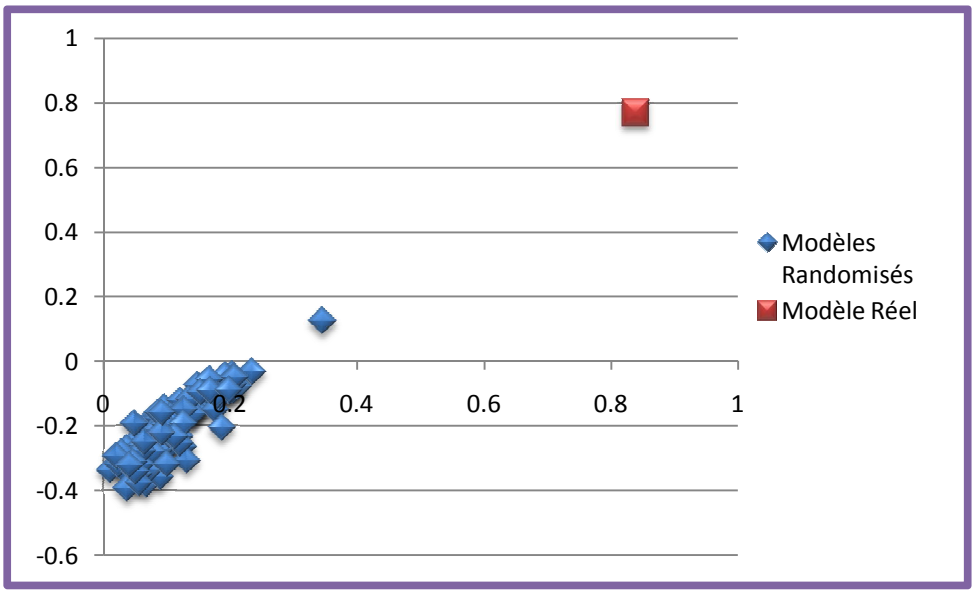


Figure-15- *Test de randomisation associé au modèle QSPR.*

Les carrés bleus représentent le potentiel de demi-vague ordonnée de façon aléatoire, et le carré orange correspond au modèle réel. Il est clair que les statistiques obtenues pour les vecteurs modifiés sont plus petites que celles du modèle QSPR réel, et pour la majeure partie on obtient même un $Q^2 < 0$. Ceci permet d'assurer qu'une relation structure/potentiel de demi-vague réelle a été très bien établie.

Validation externe :

Pour généraliser le modèle choisi on procède à une validation externe sur les (17) composés choisis aléatoirement et qui ne font pas partie de l'ensemble d'essai. Les résultats obtenus **Tableau -09-** montrent que les valeurs prédites sont proches des valeurs observées, ce qui confirme que le modèle choisi décrit de bonne façon la relation de potentiel de demi-vague prédite et celles observée avec une dispersion autre de la droite d'ajustement **(figure-14-).**

Tableau -09- *Valeurs expérimentales, prédites et leurs erreurs pour l'ensemble de validation*

N	composées	$E_{1/2}$	$E_{1/2}$Pred	Hat	Err.Pred.	Std.Err.Pred.
52	Benzoquinone	0,15	0,07	0,17	-0,08	-0.53
53	2-Methyl-1,4-naphtoquinone	-0,17	-0,27	0,071	-0,1	-0.48
54	Azobenzene	-0,33	-0,64	0,133	-0,31	-1.24
55	m-Dinitrobenzene	-0,26	0,4	0,168	0,66	2.47
56	p-Nitroaniline	-0,36	-0,24	0,078	0,12	0.29
57	m-Nitrophenol	-0,37	-0,55	0,061	-0,18	-0.68
58	o-Nitrotoluene	-0,26	-0,68	0,081	-0,42	-1.6
59	Crotonaldehyde	-0,92	-1,25	0,097	-0,33	-1.33
60	Methyl glyoxal	-0,83	-0,71	0,124	0,12	0.31
61	8-Hydroxyquinoline	-1,39	-1,39	0,1	0	0.09
62	Nicotinamide	-1,56	-1,17	0,067	0,39	1.33
63	acrylonitrile	-1,94	-2,45	1,378	-0,51	-0.55
64	Maleic acid	-1,36	-0,74	0,083	0,62	2.15
65	Methylacrylonitrile	-2,07	-1,69	0,173	0,38	1.51
66	Pyruvic acid	-0,86	-0,94	0,296	-0,08	-0.23
67	Benzal chloride	-1,81	-1,36	0,049	0,45	1.59
68	m-Dichlorobenzene	-0,3	-0,37	0,299	-0,07	-0.76

Toutes les valeurs des paramètres statistiques de validation sont regroupées dans le tableau ci dessous

Tableau -10- *Valeurs des Q^2_{ext} et $SDEP_{ext.}$*

N	**17**
Q^2_{Ext}	**74,19 %**
$SDEP_{ext}$	**0,345**

La valeur de Q^2ext est proche ou même supérieure à celle de la prédiction interne ce qui confirme que le modèle a une bonne capacité prédictive, également pour le SDEPext qui une Valeur proche à celle de SDEP.

Conclusion

Conclusion

La méthode QSPR a été utilisée pour relier les potentiels de demi-vague de mélanges de composés organiques à des descripteurs moléculaires théoriques calculés à l'aide d'un logiciel spécialisé.

Les 68 composés ont été divisés en deux groupes, dont 51 ont été utilisés pour la calibration et le reste pour la validation externe.

La taille du modèle a été fixée à 6 descripteurs par la valeur optimale de R^2, et la sélection des variables explicatives a été faite en maximisant la valeur du coefficient de prédiction Q^2_{LOO} par l'algorithme génétique du logiciel MOBY DIGS.

Les statistiques obtenues permettent de s'assurer de la qualité d'ajustement, de la robustesse interne et externe, du pouvoir prédictif et de la probabilité d'un dimensionnement adéquat du modèle choisi.
Ainsi, le potentiel demi-vague peut être prédit à partir de la structure ou de la géométrie de la molécule par un modèle linéaire.

Enfin on doit investiguer les causes possibles des aberrations relevées pour notre modèle, si c'est un problème d'optimisation alors il faut chercher d'autres méthodes ou ce sont des erreurs d'expérimentateur.

Ce travail peut être étendu à un nombre plus important de composés et le choix ou l'éclatement des données en deux ensembles disjoints (calibration et validation) pourrait se faire d'une manière plus réfléchie, en utilisant d'autres méthodes non-linéaires à savoir les réseaux de neurones artificiels ou le SVM.

Références bibliographiques

Références bibliographiques

N°	*Références*
[01]	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.
[02]	N. Margossian, Le règlement REACH - La règlementation européenne sur les produits chimiques, Dunod / L'Usine Nouvelle, Paris, (2008).
[03]	M. Ghamali, S. Chtita, R. Hmammouchi, A. Adad, M. Bouachrine, T. Lakhlifi, The inhibitory activity of aldose reductase of flavonoids compounds. Combining DFT and QSAR calculations, J. of Taibah Univ. for Sci. (2016). In Press, http://dx.doi.org/doi:10.1016/j.jtusci.2015.09.006.
[04]	S. Chtita, R. Hmamouchi, M. Larif, M. Ghamali, M. Bouachrine and T. Lakhlifi,, QSPR studies of 9-aniliioacridine derivatives for their DNA drug binding properties based on density functional theory using statistical methods: Model, validation and influencing factors, J. of Taibah Univ. for Sci. (2016), In Press, http://dx.doi.org/10.1016/j.jtusci.2015.04.007.
[05]	R. L. McNaughton, A. A. Tipton, N. D. Rubie, R. R. Conry and M. L. Kirk, Electronic Structure Studies of Oxomolybdenum Tetrathiolate Complexes: Origin of Reduction Potential Differences and Relationship to Cysteine-Molybdenum Bonding in Sulfite Oxidase, Inorg. Chem. 39, 5697-5706,(2000).
[06]	S. Niu, X. B. Wang, J. A. Nichols, L. S. Wang and T. Ichiye, Combined Quantum Chemistry and Photoelectron Spectroscopy Study of the Electronic Structure and Reduction Potentials of Rubredoxin Redox Site Analogues, J. Phys. Chem. A. 107, 2898-2907,(2003).
[07]	P. Tompe, Gy. Clementis, I. Petnehazy, Zs. M. Jaszay, L. Toke, Quantitative Structure-Electrochemistry Relationships of α, β-Unsaturated Ketones, Anal. Chim. Acta, 305, 295-303,(1995).
[08]	H. Li, L. Xu, Q. Su, Structure-Property Relationship Between Half-Wave Potentials of Organic Compounds and Their Topology, Anal. Chim. Acta. 316, 39-45,(1995).
[09]	C. Hansch, T. Fujita, *J. Am. Chem. Soc.*, *86*, 1616-1626, (1964).

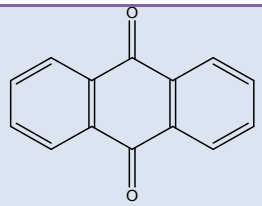
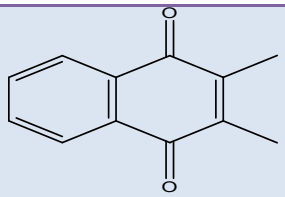
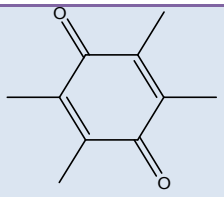
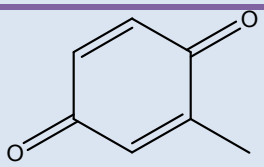
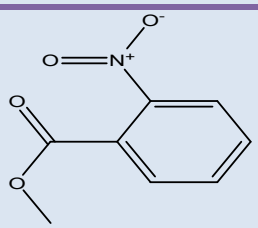
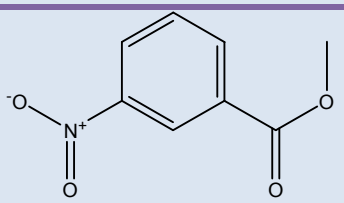
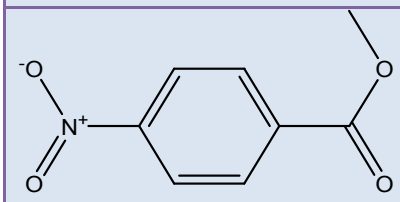
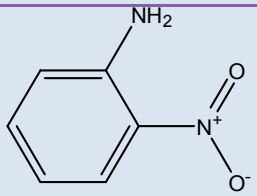
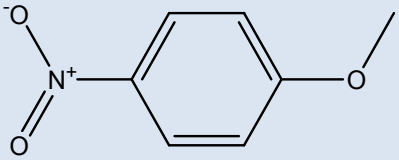
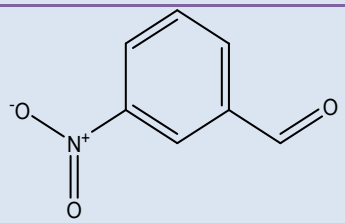
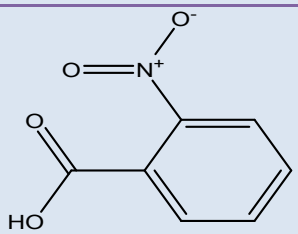
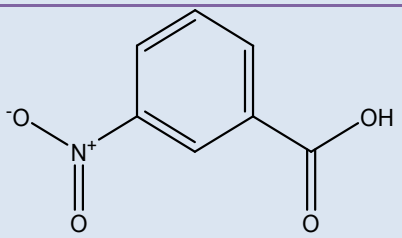
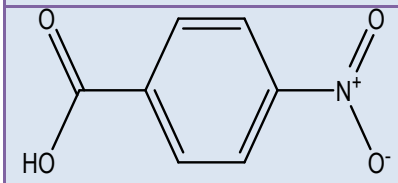
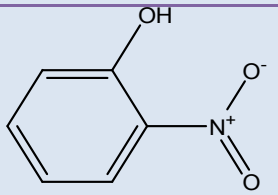
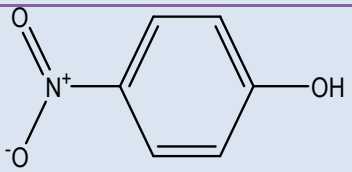
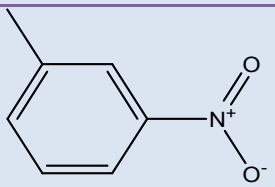
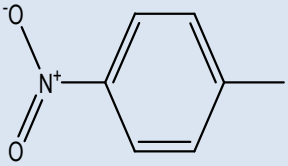
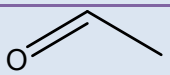
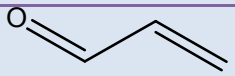
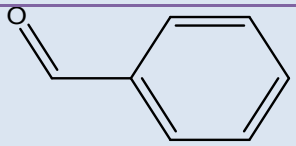
[10]	S.M. Free, J.W. Wilson, *J. Med. Chem.*, *7*, 395-399, (1964).
[11]	A. Crum-Brown, T.R. Fraser, *Trans. R. Soc. Edinburgh*, 25, 151 (1868).
[12]	C. Richet, C. R. Séances Soc. Biol. Ses. Fil., 9, 775 (1893).
[13]	H. Meyer, Zur Theorie der Alkoholnarkose, *Arch. Exp. Pathol. Pharmakol.*, 42, 109 (1899).
[14]	E. Overton, Studien über die Narkose. Zugleichnein Beitrag zur allgemeinen Pharmakologie. Jena: Gustav Fischer, Germany, (1901).
[15]	A. K. Debnath,"Quantitative Structure-Activity Relationship (QSAR) Paradigm Hansch Era to New Millenium". Mini Reviews in Medicinal Chemistry, I: 187-195. (2001).
[16]	C.Hansch, Lien E. J. Structure-activity relationships in antifungial agents. Surveyl, Journal of Medicinal Chemistry. 14(8). P 653-670, (1971).
[17]	K. Roy, S. Kar, and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, (2015).
[18]	Y, YOUSFI. Etude QSAR de l'activité anti oxydante d'une série de composes phénoliques. Master. Université Tlemcen. P21, (2017).
[19]	K. Dermiche. Etude in sillico de la thalidomide : Apport de la modélisation moléculaire. Thèse en vue de l'obtention du diplôme de doctorat en science. Université Mohamed Boudiaf d'Ouran. P 40, (2016).
[20]	HyperchemTM Release 6.03 for windows, Molecular Modeling system, (2000).
[21]	R. Todeschini, V. Consonni, Et M. Pavan, DRAGON, Software for the calculation of Molecular Descriptors. Release 5.3 for windows, Milano. 2005.
[22]	Todeschini. R, Ballabio. D, Consonni. V, Mauri. A, Pavan. M. MOBYDIGS Software for Multilinear Regression Analysis and variable Subset Selection by Genetic Algorithm. Release I.1 for Windows. Milano. (2009).
[23]	Searching scientific literature directly with Chem Draw v14 Chemistry World .29 July (2014).
[24]	M. P. Allen, D.J.Tildesley, Computer stimulation of liquids .Oxford .(1987)
[25]	R. Todeschini, V. Consonni, M. Pawan, DRAGON, Software for the calculation of Molecular Descriptors. Release 5.3 for Windows, Milano, (2005).
[26]	B .Hoggas, C.Amamiche , Modélisation De La Température D'ébullition Des Alcanes En Utilisant Une Approche QSPR, Mémoire de Master (L.M.D), chimie analytique et environnement :p8 ,(2016).

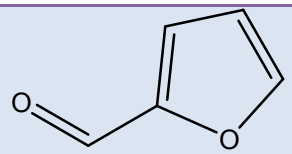
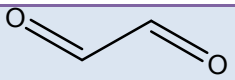
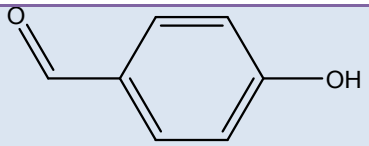
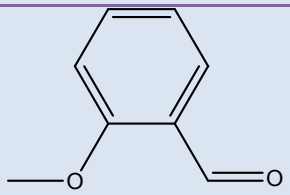
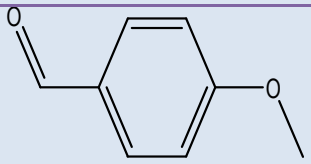
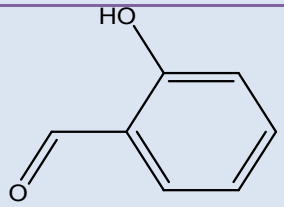
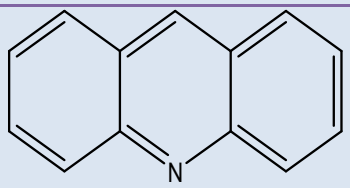
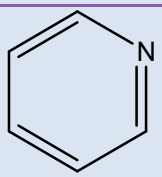
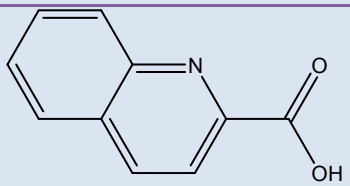
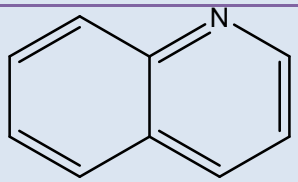
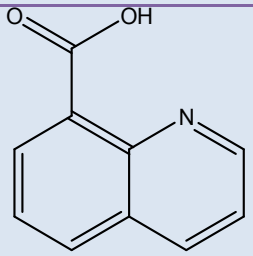
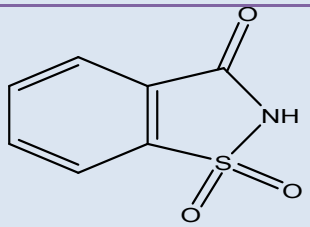
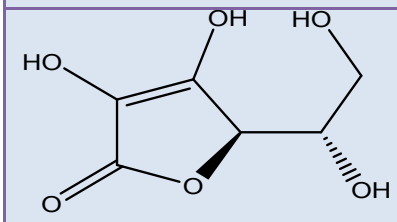
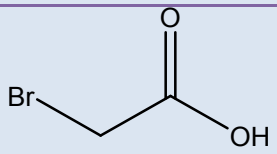
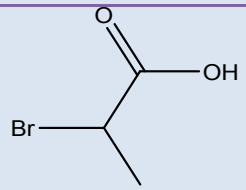
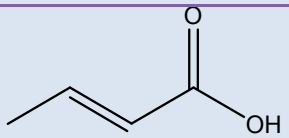
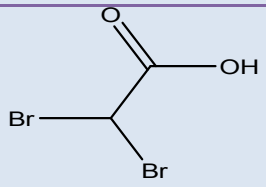
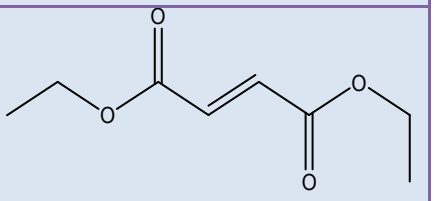
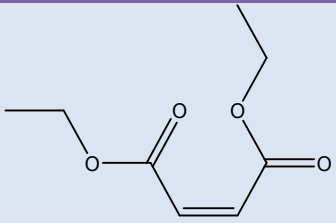
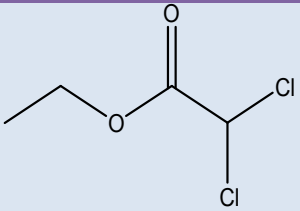
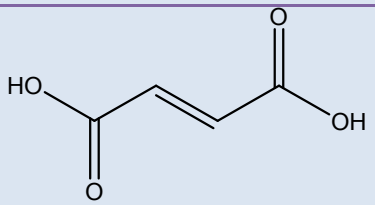
[27]	T.Thomas-Danguin, Intensité olfactive des composés purs et des mélanges: application au masquage des odeurs, Université Claude Bernard, Lyon, p224, (1997).
[28]	G.Martin, P.Laffort, Odeurs et désodorisations dans l'environnement, Lavoisier, Tec&Doc, Paris, (1991).
[29]	K. Roy, S. Kar, and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. (2015).
[30]	Danishuddin and A. U. Khan, "Descriptors and their selection methods in QSAR analysis: paradigm for drug design," *Drug Discov. Today*, vol. 21, no. 8, pp. 1291–1302, 2016, doi: 10.1016/j.drudis. 06.013, (2016).
[31]	E. T. D. E. La and R. Scientifique, "Élaboration des modèles QSPR prédictifs des propriétés physico- chimiques à l'aide des descripteurs moléculaires" (2015).
[32]	K. Roy, *Ecotoxicological QSARs*. (2020).
[33]	K. Saadi, Contribution á l'étude de la Relation structure chimique- odeur Utilisation de la technique Random Forest (Application à la famille des pyrazines), p36, (2009).
[34]	K. Saadi, Contribution l'étude de la Relation structure chimique- odeur Utilisation de la technique Random Forest (Application à la famille des pyrazines), Mémoire de Magister. UNIVERSITE KASDI Merbah Ourgla.p30, (2009).
[35]	AI ACCESS, 91940, Les Ulis, Franonce.
[36]	R. Tomassone, E. Lesquoy, C. Miller, La régression : nouveaux regards sur une ancienne méthode statistique. Masson, INRA .variables. Ecology 89(9): 2623-2632, (1983).
[37]	L. Chambers. Practical Handbook of Genetic Algorithms. Lewis Publishing, (1995).
[38]	MINITAB Release 16.2.0.0 for Microsoft langage pack 2.
[39]	R. Todeschini. MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm.Release for Windows. Milano Srl.
[40]	A. R. Katritzky, V. S. Lobanov, M. Karelson, CODESSA Reference Manual. University of Florida, Gainesville, (1994).
[41]	P. Dagnélie, Statistique Théorique Et Appliquée. Tomes 1 et 2. De Boeck &Larcier s. a, (1998).
[42]	L. Chambers. Practical Handbook of Genetic Algorithms. Lewis Publishing,

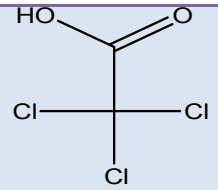
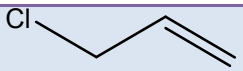
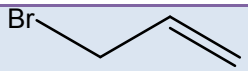
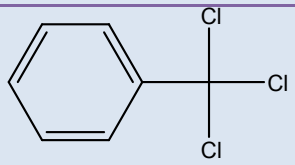
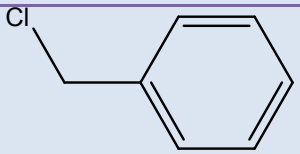
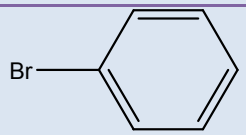
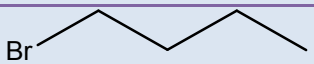
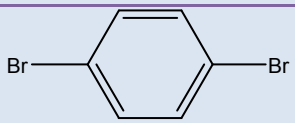
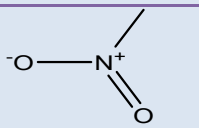
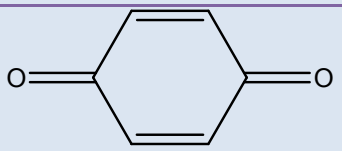
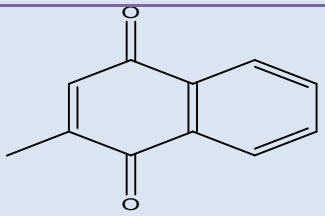
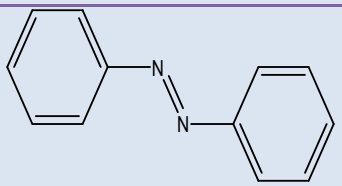
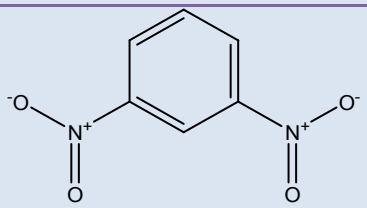
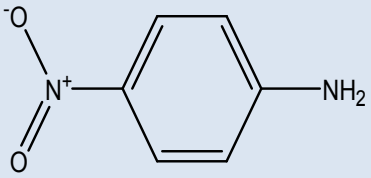
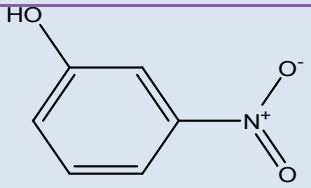
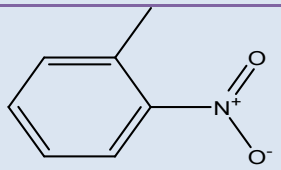
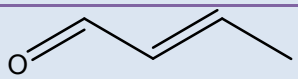
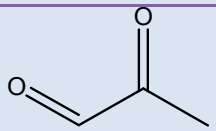
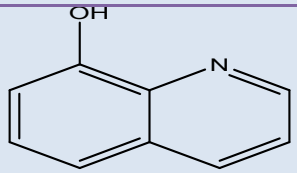
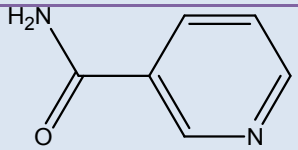
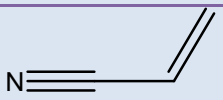
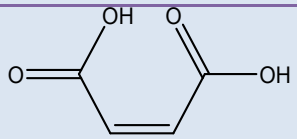
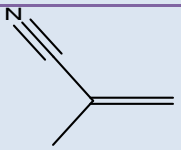
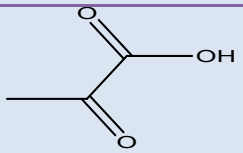
	(1995).
[43]	MINITAB Release 16.2.0.0 for Microsoft langage pack 2.
[44]	R. Todeschini. MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release for Windows. Milano Srl.
[45]	N.R Draper, H. Smith, Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New York, (1998).
[46]	L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Perspective, 111(10), 1361-1375, (2003).
[47]	S .Remache ,W.Redah , Etude QSRR de la rétention chromatographique des HAP, Mémoire de Master (L.M.D), chimie analytique et environnement , :p 20, (2018).

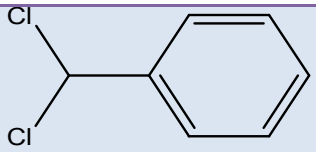
Annexes

Annexes

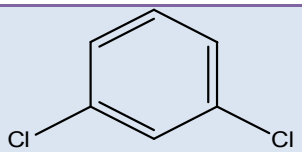
 Anthraquinone	 2,3-Dimethylnaphthoquinone	 Duroquinone
 Toluquinone	 Methyl o-nitrobenzoate	 Methyl m-nitrobenzoate
 Methyl p-nitrobenzoate	 o-Nitroaniline	 p-Nitroanisole
 m-Nitrobenzaldehyde	 o-Nitrobenzoic acid	 m-Nitrobenzoic acid
 p-Nitrobenzoic acid	 o-Nitrophenol	 p-Nitrophenol
 m-Nitrotoluene	 p-Nitrotoluene	 Acetaldehyde
 Acrolein	 Benzaldehyde	 Formaldehyde

 Furfural	 Glyoxal	 p-Hydroxybenzaldehyde
 o-Methoxybenzaldehyde	 p-Methoxybenzaldehyde	 Salicylaldehyde
 Acridine	 Pyridine	 Quinaldinic acid
 Quinoline	 Quinoline-8-carboxylic acid	 Saccharin
 Ascorbic acid	 Bromoacetic acid	 alpha-Bromopropionic acid
 Crotonic acid	 Dibromoacetic acid	 Diethyl fumarate
 Diethyl maleate	 Ethyl dichloroacetate	 Fumaric acid

 Trichloroacetic acid	 Allyl chloride	 Allyl bromide
 Benzotrichloride	 Benzyl chloride	 Bromobenzene
 n-Butyl bromide	 p-Dibromobenzene	 Nitromethane
 Benzoquinone	 2-Methyl-1,4-Naphthoquinone	 Azobenzene
 m-Dinitrobenzene	 p-Nitroaniline	 m-Nitrophenol
 o-Nitrotoluene	 Crotonaldehyde	 Methyl glyoxal
 8-Hydroxyquinoline	 Nicotinamide	 acrylonitrile
 Maleic acid	 Methylacrylonitrile	 Pyruvic acid



Benzal chloride



m-Dichlorobenzene

Résumés

Résumé:

Un modèle QSPR a été développé pour la prédiction du potentiel de demi-vague. D'une série de 68 composées. La série des données a été divisé aléatoirement en deux sous-ensembles, un ensemble de 51 composés pour la construction du modèle et un ensemble de 17 composés pour la validation. Nous avons calculé les descripteurs moléculaires en utilisant des logiciels de modélisation moléculaires bien spécifiques (HyperChem, ChemDraw, dragon, minitab).

Les descripteurs moléculaires jouent un rôle clé dans le développement des modèles QSAR/RQSP car ils représentent quantitativement les informations chimiques codées. Non seulement ils aident à dériver des corrélations mathématiques entre les structures chimiques et les réponses d'intérêt. Les modèles QSAR obtenus sont élaborés avec la méthode de régression RLM. La taille du modèle à été déterminée en optimisant la valeur de R^2, et la sélection des descripteurs réalisée par algorithme génétique. Les valeurs des paramètres statistiques (R^2, Q^2, SDEC, SDEP, SDEPext) obtenues attestent de la pertinence du modèle développé. Le domaine d'application a été discuté à l'aide de diagramme de Williams.

Mots-clés: QSAR/QSPR – potentiel de demi-vague – Descripteurs moléculaires – Algorithme génétique.

ABSTRACT:

A QSPR model was developed for the prediction of the half-wave potential. From a series of 68 composed. The data set was randomly divided into two subsets, a set of 51 compounds for model building and a set of 17 compounds for validation. We calculated the molecular descriptors using very specific molecular modeling software (HyperChem, ChemDraw, dragon, minitab).

Molecular descriptors play a key role in the development of QSAR/RQSP models because they quantitatively represent the encoded chemical information. Not only do they help derive mathematical correlations between chemical structures and responses of interest. the QSAR models obtained are elaborated with the RLM regression method. The size of the model was determined by optimizing the value of R^2, and the selection of descriptors carried out by genetic algorithm. The values of the statistical parameters (R^2, Q^2, SDEC, SDEP, SDEPext) obtained attest to the relevance of the model developed. The domain of application was discussed using Williams diagram.

Keywords: QSAR/QSPR – half-wave potential – Molecular descriptors – Genetic algorithm.

ملخص

تم تطوير نموذج QSPR للتنبؤ بإمكانية نصف الموجة. من سلسلة تتألف من 68. تم تقسيم مجموعة البيانات بشكل عشوائي إلى مجموعتين فرعيتين ، مجموعة من 51 مركبًا لبناء النموذج ومجموعة من 17 مركبًا للتحقق من صحتها. حسبنا الواصفات الجزيئية باستخدام برنامج نمذجة جزيئية محدد جدًا (chem draw ,hyperchem, dragon, minitab)

تلعب الواصفات الجزيئية دورًا رئيسيًا في تطوير نماذج QSPR/QSAR لأنها تمثل كميًا المعلومات الكيميائية المشفرة. فهي لا تساعد فقط في اشتقاق الارتباطات الرياضية بين الهياكل الكيميائية والاستجابات ذات الأهمية. تم تطوير نماذج التي تم الحصول عليها باستخدام طريقة الانحدار RLM .تم تحديد حجم النموذج من خلال تحسين قيمةR^2 ، واختيار الواصفات التي تم إجراؤها بواسطة الخوارزمية الجينية. قيم المعلمات الإحصائيةR^2 ، Q^2، SDEC، SDEP، SDEPextالتي تم الحصول عليها تشهد على ملاءمة النموذج الذي تم تطويره. تمت مناقشة مجال التطبيق باستخدام مخطط ويليامز.

الكلمات المفتاحية :

جهد نصف الموجة ـ الواصفات الجزيئية ـ الخوارزمية الجينية ـ QSAR/QSPR