REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DE SAÏDA - Dr MOULAY TAHAR

FACULTE DES SCIENCES
Département de Chimie

MEMOIRE

Présenté par:
Hamdani Fatima Zohra

En vue de l'obtention du

Diplôme de Master en Chimie

Spécialité : Chimie Théorique et Computationnelle

Thème:

Étude QSPR du PKa et température de fusion de quelques composés organiques

Soutenu le 04/10/2020

Devant le jury composé de :

Présidente	Dr Djallila Missaoui	*MCB*	*Université de Saida*
Encadreur	Dr Rekia KADARI	*MCB*	*Université de Saida*
Co-Encadreur	Dr Houari BRAHIM	*MCA*	*Université de Saida*
Examinateur	Dr Ali Rahmouni	*Professeur*	*Université de Saida*
Examinateur	Dr Noureddine Doumi	*MCB*	*Université de Saida*

Année universitaire 2019/2020

ملخص

تم تطبيق العلاقات الكمية بين البنية وخاصية (QSPR) علي مدار العقد الماضي للحصول علي نماذج احصائية موثوقة للتنبؤ بأنشطة الكيانات الكيميائية الجديدة. ينصب اهتمام نموذج QSPR علي استخلاص المعلومات من مجموعة الواصفات العددية التي تميز التركيب الجزيئي و بالتالي التنبؤ بالأنشطة البيولوجية.

في اطار هده الاطروحة، اجريت دراسات QSPR علي 25 مركب من المركبات الكيميائية (الهيدروكربونات , مشتقات احماض البنزويك). تم استخدام تقنيات مختلفة مثل انحدار الخطي، و شبكة الخلايا العصبية متعددة الطبقات، و الجار الاقرب ل Kلإعداد نماذج للتنبؤ بالأنشطة البيولوجية .تشير نتائجنا الي نموذجQSPR بناءا علي واصفات مختلفة. تم تقدير التنبؤ بالنموذج من خلال التحقق الداخلي والخارجي. ثم التحقق من وجود علاقة ارتباط قوية بين قيم الانشطة التجريبية والانشطة المتوقعة، و التحقق من صحة نماذجQSPR الناتجة.

الكلمات المفتاحية:

QSPR ، الجزيئات ، التفاعلات البيولوجية ، الانحدار الخطي ، شبكة الخلايا العصبية متعددة الطبقات ، الجار الأقرب k ، الواصفات ، التنبؤ.

Résumé

Les relations quantitatives structure-propriété (QSPR) ont été appliquées au cours de la dernière décennie pour obtenir des modèles statistiques fiables pour la prédiction des activités de nouvelles entités chimiques. L'intérêt d'un modèle QSPR est de tirer des informations à partir de l'ensemble des descripteurs numériques caractérisant la structure moléculaire et prédire ainsi les propriétés chimiques.

Dans le cadre de ce mémoire des études QSAR ont été effectuées sur 25 composés d'entités chimiques (hydrocarbures, des dérivées d'acides benzoïques). Différentes techniques tel que la régression linéaire, les perceptron multi couche et le K-plus proche voisin ont été utilisées pour mettre en place des modèles pour la prédiction des activités biologiques. Nos résultats suggèrent des modèle QSAR en fonction de différents descripteurs. La prédictive du modèle a été estimé par la validation interne et externe. Une forte corrélation entre les valeurs des activités expérimentales et prédites a été observée, indiquant la validation et la bonne qualité des modèles QSAR issus.

Mots clés :

QSAR, molécules, Activités Biologiques, Régression linéaire, perceptron multicouche, k-plus proche voisin, Descripteurs, Prédiction.

Abstract

Quantitative structure-property relationships (QSPR) have been applied over the past decade to obtain reliable statistical models for the prediction of the activities of new chemical entities. The interest of QSPR models is to derive information from the set of numerical descriptors characterizing molecular structure and thus predict biological activities.

As part of this thesis, QSPR studies were carried out on 25 compounds of chemical entities (hydrocarbons, derivatives of benzoic acids). Different techniques such as linear regression, multilayer perceptron and K-nearest neighbors have been used to set up models for the prediction of biological activities. Our results suggest QSPR models based on different descriptors. The predictive of the model was estimated by internal and external validation. A strong correlation between the values of the experimental and predicted activities was observed, indicating the validation and the good quality of the resulting QSPR models.

Keywords:

QSPR, molecules, Biological Activities, Linear regression, multilayer perceptron, K-nearest Neighbors, Descriptors, Prediction.

Remercîment

Je remercie ALLAH le Tout-puissant de m'avoirdonné le courage, la volonté et la patience de mener à terme ce présent travail.

Nous exprimons nos sincères remerciements à Dr. ***REKIA KADARI*** *pour son acceptation de travailler avec moi. Elle peut trouver ici mon sentiment de gratitude et de respect.*

Nous remercions Dr. ***Houari Ibrahim*** *pour son acceptation en tant que Co-Encadreur de ce modeste travail.*

Nous remercions Dr. ***Djallila Missaoui****, d'avoir accepté de présider le jury de ce modeste travail.*

Un remerciment spécial Professeur ***Ali Rahmouni*** *et Dr.* ***Noureddine Doumi*** *qui ont bien voulu s'intéresser à ce modeste travail et il a accepté d'en d'être examinateur.*

Finalement, toute ma gratitude va à mes parents qui n'ont jamais douté de moi et qui m'ont aidé et encouragé tout le long de mes études, à mon frère et ma sœur pour leur patience et leur soutien pendant cette longue et pénible épreuve que représente ce mémoire.

Dédicace

Je dédie ce travail

A ma mère ;
En vous, je vois la maman parfaite, toujours prête à se sacrifier pour le bonheur de ses enfants.

A mon père ;
Qui m'a indiqué la bonne voie en me rappelant que la volonté fait toujours des grandes personnes.

A ma chère sœur Rekia pour ses encouragements permanents, et son soutien moral.

À mon cher frère Amine pour son soutien et ses encouragements.

À mes chères copines, je dois leur offrir un soutien moral afin de terminer le travail.

À toute ma famille pour leur soutien tout au long de mon parcours universitaire.

Table des matières

Chapitre II :*Relations Quantitatives Structures Activités/Propriété (QSAR/QSPR)*

Chapitre III : Résultats et Discussion

Conclusion générale

Liste Des Figures

Liste Des Tableaux

Liste Des Abréviations

OM	Orbitale Moléculaire
LCAO	Combinaison Linéaire d'Orbitales Atomique
DFT	Théorie de la Fonctionnelle de la Densité
B3LYP	Becke, trois paramètres, Lee-Yang-Parr
SAR	La Relation Structure-Activité
QSAR	Relation Quantitative Structure-Activité
QSPR	Relation Quantitative Structure-Propriété
LR	La Régression Linéaire
MLP	Perceptron multicouche
MAE	Erreur Absolue Moyenne
RMSE	Erreur quadratique moyenne
R	Coefficient de Corrélation
R^2	Coefficient de Détermination.
HOMO	Orbitale moléculaire occupée la plus élevée
LUMO	Orbite moléculaire inoccupée la plus basse
ΔE	GAP énergétique
KNN	k plus proches voisins
PCA	Analyse des composants principaux
SVM	Machine a Vecteur de Support

Introduction Générale

Introduction Générale

Au cours de ces dernières années, l'informatique a pris beaucoup de place dans la science. La chimie computationnelle est l'un des domaines d'application les plus importants des sciences informatiques.

La chimie computationnelle désigne l'application de compétences chimiques, mathématiques et informatiques à la recherche de la solution de problèmes chimiques intéressants en utilisant des ordinateurs, et de générer des informations telles que les propriétés des molécules ou des résultats expérimentaux simulés.

QSAR (Quantitative structure Activity Relationship) est un outil important en bio / chimio-informatique, qui peut être construit principalement à partir des données générées par la modélisation moléculaire et la chimie computationnelle pour faire des prédictions sur l'activité biologique de nouveaux produits chimiques.

Ce travail de recherche se place dans le contexte d'une étude QSPR sur deux types de molécules, des dérivés d'acides benzoïque et des hydrocarbures. Le principal objectif de ce travail est l'application de différentes méthodes de la modélisation moléculaire pour prédire des activités biologiques cibles de nouvelles molécules pour les deux types de molécules étudiées.

Le présent mémoire comporte trois chapitres. Le premier chapitre est scindé en deux parties : dans la première partie, nous allons présenter des généralités sur les méthodologies choisie dans la modélisation moléculaire qui comporte les différentes méthodes de calcul utilisées et engagées dans notre travail. La deuxième partie, comporte une étude bibliographique des molécules utilisés ainsi que les indices de la réactivité.

Le deuxième chapitre est consacré à la description de la méthodologie QSAR/QSPR. Dans ce chapitre nous présenterons le processus d'une étude QSAR/QSPR ainsi une description de certaines méthodes statistiques.

Dans le troisième chapitre, une étude quantitative des relations structure-activité des dérivés d'acide benzoïque et hydrocarbures. Nous présenterons le développement des modèles statistiques de deux activités biologiques: PKa et

température de fusion avec les méthodes statistiques (régression linéaire, perceptron multi couche, K plus proche voisin).

Chapitre I: *Méthodologie et Études bibliographiques*

Chapitre I: Méthodologie et Études bibliographiques

I.1 Introduction

La mécanique quantique est basée sur la résolution du système moléculaire de l'équation de Schrödinger en résolvant une équation indépendante du temps avec des valeurs propres et des vecteurs propres.

$$H\Psi = E\Psi$$

Où :

- H est l'hamiltonien non relativiste ;
- E l'énergie totale et ;
- Ψ la fonction d'onde de système.

La modélisation moléculaire qui est un ensemble technique pour modéliser ou simuler le comportement de molécules cela se fait à l'aide de différents logiciels, par exemple, le logiciel ***Gaussian*****[1]**.

I.2 Historique

Dans les années 1920, la mécanique quantique s'est développée de façon spectaculaire, et c'est aux scientifiques comme Bohr, Schrödinger, Born, Oppenheimer, Hartrèe ou encore Slater. En 1930, Hartrèe et Fock développèrent la méthode du champ auto cohérent qui permet d'effectuer les premiers calculs *ab initio* sur des systèmes diatomiques Il a fait son sur ordinateur dans les années 50. En 1964, la théorie de la fonctionnelle de la densité (DFT) a été introduite par Hohenberg et Kohn, en 1970 pople a fait le programme Gaussian et en 1970 et 1980 des méthodes semi-empiriques. Avec la forte augmentation de la puissance de calcul, la modélisation s'est invitée dans nos ordinateurs à partir des années 1990. En 1993 apparaît la fonctionnelleB3LYP, ehybride qui permet calculs DFT. en 1998, le prix Nobel de chimie a été décerné à John. A. Pople et Walter Kohn pour leurs travaux dans le domaine de la chimie informatique et la modélisation moléculaire (chimie quantique)**[2]**.

I.3 Equations de Schrödinger

Le comportement d'un système moléculaireest entièrement déterminé par une fonction d'onde multi-particules Ψ_0**,** solution de l'équation de Schrodinger dépendante du temps :

$$\mathbf{i\hbar\, d/dt\ \Psi_0 = H_0\ \Psi_0}$$

Dans le cadre d'un système non-relativiste et n'évoluant pas dans le temps, cette ´équation se simplifie en:

$$\mathbf{H_0\ \Psi_0 = E_0\ \Psi_0}$$

Où :

- H_0 représente l'opérateur Hamiltonien du système;
- E_0 son énergie ; et Ψ_0 (r_1,...,r_N), ou N est le nombre de particules, une fonction d'onde d´écrivant l'état du système énergétique E_0 ;
- Ψ_0 est un vecteur propre de H_0 associe à la valeur propre E_0**[3].**

I.4 Les Méthodes de Calculs

I.4.1 Méthode de Hartree-Fock-Roothaan

I.4.1.1 Approximation du champ moyen de Hartree

L'approximation du champ moyen, proposée par Hartree en 1927, consiste à remplacer l'interaction d'un électron avec les autres électrons par l'interaction de celui-ci avec un champ moyen créé par la totalité des autres électrons. Cela peut être remplacé par un potentiel électronique binaire$\sum_j \boldsymbol{e^2/rij}$, qui représente la résonance immédiate entre l'électron i et les autres électrons i differents de j . à partir d'une seule tension électronique moyenne. La forme de l'électron i est U (i). Par conséquent, sur la base de la théorie électronique, nous pouvons écrire la fonction d'onde totale comme un produit d'une seule fonction électronique **[4]**:

$$\mathbf{\Psi = \Psi_1\ (1)\ .\Psi_2(2).\ \Psi_3(3)\ldots\ldots.\Psi_n(n)}$$

I.4.1.2 Méthode de Hartree-Fock

Selon la méthode Hartree, la fonction d'onde totale est exprimée en fonction de la fonction d'onde de chaque électron, et chaque électron a sa propre orbite. Ce principe de la l'indiscernabilitédes électrons, mais Pauli l'a ignoré.

En 1928, Focka eu l'idée d'exprimer la fonction d'onde complète comme le déterminant de Slater. Chaque orbitale moléculaire(OM) peut contenir jusqu'à 2 électrons (spin α, spin β) **[5]**

$$\boldsymbol{\Psi_{(1,2,\ldots.n)}} = \frac{\mathbf{1}}{\sqrt{\boldsymbol{2n}!}} \begin{vmatrix} \emptyset_1\ x_1 & \emptyset_2 x_1 \ldots. & \emptyset_{2n} x_1 \\ \emptyset_1 x_2 & \emptyset_1 x_2 \ldots. & \emptyset_1 x_2 \\ \emptyset_1 x_{2n} & \emptyset_1 x_{2n} \ldots. & \emptyset_1 x_{2n} \end{vmatrix}$$

Utilise le déterminant et l'intégrale de Slater pour exprimer la fonction d'onde, qui sera l'énergie de Hartree Fook :

$$\mathbf{E_{Tot}} = 2\sum_{i=1}^{n} H_{ii}^{c} + \sum_{i=1}^{n}\sum_{j=1}^{n}(2J_{ij} - K_{ij})$$

Où :

- $H_{ii} = \int \emptyset_i^*\ (1)\widehat{h_\iota}\emptyset_i(1)d\overrightarrow{r_1}$
- $J_{ij} = \int \emptyset_i^*\ (1)\emptyset_j^*(2)\frac{1}{r_{12}}\emptyset_i(1)\emptyset_j(2)d\overrightarrow{r_1}\mathrm{d}\overrightarrow{r_2}$
- $K_{ij} = \int \emptyset_i^*\ (1)\emptyset_j^*(2)\frac{1}{r_{12}}\emptyset_j(1)\emptyset_i(2)d\overrightarrow{r_1}\mathrm{d}\overrightarrow{r_2}$

I.4.1.3 Méthode de Hartree-Fock-Roothaan

Selon la méthode Hartree-Fock, l'expression de l'analyse orbitale moléculaire ϕ_i n'est pas définie. Roothaan a utilisé OM-CLOA pour construire OM. La méthode consiste à utiliser un ensemble d'orbitales atomiquesΨ_i par une combinaison linéaired'orbitale moléculaire φ_u :

$$\Phi_i = \sum_{\mu=1}^{N} C_{i\mu}\emptyset_\mu$$

Où :

- C_{iu} est un coefficient variable. N est le nombre d'OA.

Les meilleures offres sont celles qui réduisent l'énergie. En procédant par la méthode des variations et après certaines manipulation algébriques, on aboutit aux équations de Roothan définies par le système séculaire suivant :

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

Avec :

$$\begin{cases} F_{rs} = h_{rs}^{c} + \sum_{p=1}^{n}\sum_{q=1}^{n} P_{pq}\{2\langle rs|pq\rangle - \langle rq|ps\rangle\} \\ S_{rs} = \langle \emptyset_r|\emptyset_s\rangle \\ h_{rs}^{c} = \int \emptyset_r^{*}(i) h^{c} \emptyset_s(i) d\tau_i \end{cases}$$

Où :

- r, s, p et q représentent OA. P
- pq est un élément de la matrice de densité.
- Les termes|rq>|ps> désignent respectivement l'intégration électronique coulombienne et le binaire**[6]**.

I.4.1.4 Méthode post-SCF

Le principal inconvénient de la méthode hartree-fock-roothaan est qu'elle ne tient pas compte des liens électroniques qui existent entre les mouvements électroniques. Cette méthode relativement limitée conduit à des calculs quantitatifs des propriétés thermodynamiques, telles que l'enthalpie d'activation.

Ces propriétés peuvent être efficacement calculées avec la méthode post-scf qui considère la corrélation électronique. Deux grandes familles de méthodes développées sont la théorie de la perturbation de Müller-Pleset (MPn) d'ordre n et les méthodes DFT.L'énergie du système est la différence entre l'énergie Hartree-Fock et l'énergie non proportionnelle précise du système**[6]**.

I.5 Théorie de la fonctionnelle de densité (DFT)

I.5.1 Fondement de la théorie DFT

Historiquement, il fut Thomas (1927), Fermi (1927, 1928) et Dirac (1930) le premier à exprimer l'énergie en fonction de la densité dans le modèle unifié de gaz d'électrons . Le but de la méthode DFT est de déterminer la fonction qui relie la densité électronique à l'énergie**[7]**. Cependant, la DFT fait en fait partie de la théorie fondamentale de Hohenberg et Kuhn en 1964 [8], qui établissait une relation fonctionnelle entre l'énergie de la densité électronique fondamentale de l'état [4].

I.5.2théorème de Hohenberg et Kohn

La théorie de Hohenberg et Cohen a souligné: "L'énergie moléculaire, la fonction d'onde et toutes les autres propriétés des électrons dans l'état fondamental sont déterminées par la densité électronique de l'état fundamental **[4]**.

Mentionnez l'expression hamiltonienne de systèmes électroniques multiples:

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

Avec :

$$V(r_i) = -\sum_a \frac{Z_a}{r_{ia}}$$

Où :

- $V(r_i)$le potentiel électrique externe de l'électron i.

Ce potentiel électrique correspond à la gravitation l'e- (i) de tous les noyaux externes du système électronique **[4]**.

I.5.3Approximation de la densité locale LDA

L'approximation de la densité locale LDA représente, en première approximation, la densité qui peut être réglée localement. Par conséquent, l'énergie du réseau est déterminée par la formule suivante:

$$E_{XC}^{LDA}\ [\rho] = \int \rho(r)\varepsilon_{XC}\,(\rho)dr$$

Où :

- ε_{XC}: représente l'énergie du réseau des électrons dans le modèle de gaz homogène de l'électron (Gelium) qui est représentée par la somme de deux contributions :

$$\varepsilon_{XC}(\rho) = \varepsilon_X\,(\rho) + \varepsilon_C(\rho)$$

Avec

- $\varepsilon_x(\rho)$: l'énergies d'échange proposée par Dirac comme approximation, selon la formule :

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

Donc:

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

Et le terme $\varepsilon c(\rho)$ de l'énergie d'échange est donné par la formule de Vosko, Wilk et Nusair (VWN) **[9]**.

I.5.4Approximation de la densite de spin locale LSDA

Dans la méthode LDA, les électrons avec des spins opposés ont les mêmes Kohn et ShamΘ^{KS}, ce qui rend la méthode adaptée aux systèmes à coque fermée. En ce qui concerne la géométrie des systèmes en couche ouverte et des particules de quasi-désintégration, la

méthode LSDA peut fournir de meilleurs résultats. LSDA distingue les pistes audio. Les électrons tournent dans le sens oppose**[5]**.

$$E_{XC}^{LSDA} = E_{xc}\left[\rho^{\alpha},\ \rho^{\beta}\right]$$

I.5.5Approximation du Gradient Generalise (GGA)

La première amélioration possible de la méthode LDA est de montrer que la fonction énergétique du réseau est fonction de la densité électronique et du gradient. Par conséquent, la solution consiste à réécrire l'expression liée dans LDA:

$$E_{xc} = \int \varepsilon_{XC}^{GGA}(\rho\ , \nabla\rho)dr$$

Où ε_{XC}^{GGA}la densité d'énergie d'échange est pertinente. La difficulté réside donc dans la recherche d'expressions analytiques.

Pour l'échange et l'association, plusieurs fonctions GGA ont été développées. Les fonctions d'échange de Becke (B88) et de Perdew et Wang (PW91) sont les plus connues et les plus fréquemment utilisées. Pour la corrélation, nous avons les fonctions de Perdew (P86), Lee, Yang et Parr (LYP) et Perdew et Wang (PW91) . Par rapport à l'approximation LDA locale, toutes ces fonctions peuvent améliorer l'évaluation de l'énergie de liaison dans la molécule et la barrière énergétique**[6]**.

I.5.6 Fonctionnelle hybride B3LYP

La fonction hybride B3LYP (Becke-3parametres-Lee-Young-Parr) est une fonction à trois paramètres associée à une fonction d'échange local. Becke et HF sont échangés via la fonction de liaison locale (VWN) et sont fixés au niveau de Lee. Bar Yang**[5]**:

$$E_{XC}^{B3LYP} = (1 - a_0 - a_x)E_X^{LDA} + a_0E_X^{HF} + a_XE_X^{B88} + \mathrm{a}_cE_C^{LYP}(1 - \mathrm{a}_c) + E_c^{VWN}$$

Les valeurs des 3 paramètres d'ajustement sont:

- a_0 = 0.20

- a_x = 0.72
- a_c = 0.81

I.6 Les fonctions de bases

La base 3-21G et la base 4-31G,6-31G : c'est ce qu'on appelle une base devalence partagée(« spiltvalence ») où seules les orbitales de valence seront triplées deux fois….les nombres 3,4 et 6 avant le correspondant au nombre de primitives gaussiennes qui seront utilisées pour décrire les orbitales centrales des atomes lourds (autre que l'hydrogène).les nombres 21 et 31 avant le lettre G correspondent au nombre de fonctions de base qui seront utilisées pour décrire les orbitales de valence. La base 311G indiquerait une triple base zeta. Et la base 6-31G* : ajout d'une orbitale d pour les éléments de la deuxième ligne et d'une orbitale f pour les métaux de transition.

- **6-31G**** : ajout d'une orbitale de type p pour les atomes d'hydrogène ;
- **6-31+G** : orbitales diffuses ajoutées sur les atomes d'hydrogène ;
- **6-31++G** : ajout supplémentaires sur les atomes d'hydrogène.

I.7 Les Descripteurs Quantiques

- **Highest Occupied Molecular Orbital (HOMO) :** Highest Occupied Molecular Orbital (HOMO) Traduit le caractère électro-donneur (nucléophile) de la molécule **[10]**. Plus l'énergie de cette OM est élevée, plus la molécule cédera facilement des électrons. notée E HOMO , mesurée en eV, est le niveau d'énergie le plus élevé dans la molécule qui contient des électrons, il est directement lié au potentiel d'ionisation. Lorsqu'une molécule agit comme une base de Lewis (un doublet d'électrons donneur) dans la formation d'une liaison, les électrons sont alimentés à partir de cette orbite. Il mesure la nucléophilie d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par des électrophiles**[7]**, Les énergies de l'HOMO sont de descripteur très populaires de produit chimique de **[8]**.
- **Lowest Unoccupied Molecular Orbital (LUMO) :**Lowest Unoccupied Molecular Orbital (LUMO)traduit le caractère électro-accepteur (électrophile) de

la molécule**[10]**. Plus l'énergie de cette OM est faible, plus la molécule acceptera facilement des électrons.notée E LUMO , mesurée en eV, est le niveau d'énergie le plus bas dans la molécule qui ne contient pas d'électrons, il est directement lié à l'affinité d'électron. Lorsqu'une molécule agit comme un acide de Lewis (un doublet d'électrons accepteur) dans la formation de liaisons, des doublets d'électrons entrants sont reçus dans cette orbite. Il mesure l'‘électrophilicité d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par les nucléophiles**[7]**.

Les énergies de LUMO sont de descripteur très populaires de produit chimique[8].

Le Gap énergétique (ΔE) :C'est une différence d'énergie entre une orbitale moléculaire occupée supérieure et une orbitale moléculaire vacante inférieure, et elle est mesurée en unités eV.**[7]**.

▪ **Moment dipolaire (M dipole) :** Le moment dipolaire électrique caractérise la distribution de charges dans une molécule. La connaissance de cette distribution est fondamentale pour comprendre les propriétés électroniques de la molécule, sa géométrie, les interactions avec d'autres particules, Cette grandeur physique peut également permettre d'obtenir des informations sur la dynamique et d'aborder des problèmes tels que la rigidité d'une molécule ; le couplage rotation-vibration L'unité couramment utilisée en physique et en chimie est le Debye qui est mieux adapté aux ordres de grandeur rencontrés dans les atomes et les molécules **[11]**.

▪ **No** : C'est le nombre d'atome d'oxygène dans une molécule.

▪ **Ncl :** C'est le nombre d'atome de chlore dans une molécule.

▪ **Nn :** C'est le nombre d'atome de l'azote dans une molécule.

▪ **Nf :** C'est le nombre d'atome de fluor dans une molécule.

▪ **Nbr :** C'est le nombre d'atome de brom dans une molécule.

▪ **Nc** : C'est le nombre d'atome de carbone dans une molécule.

▪ **Nh** :C'est le nombre d'atome d'hydrogène dans une molécule.

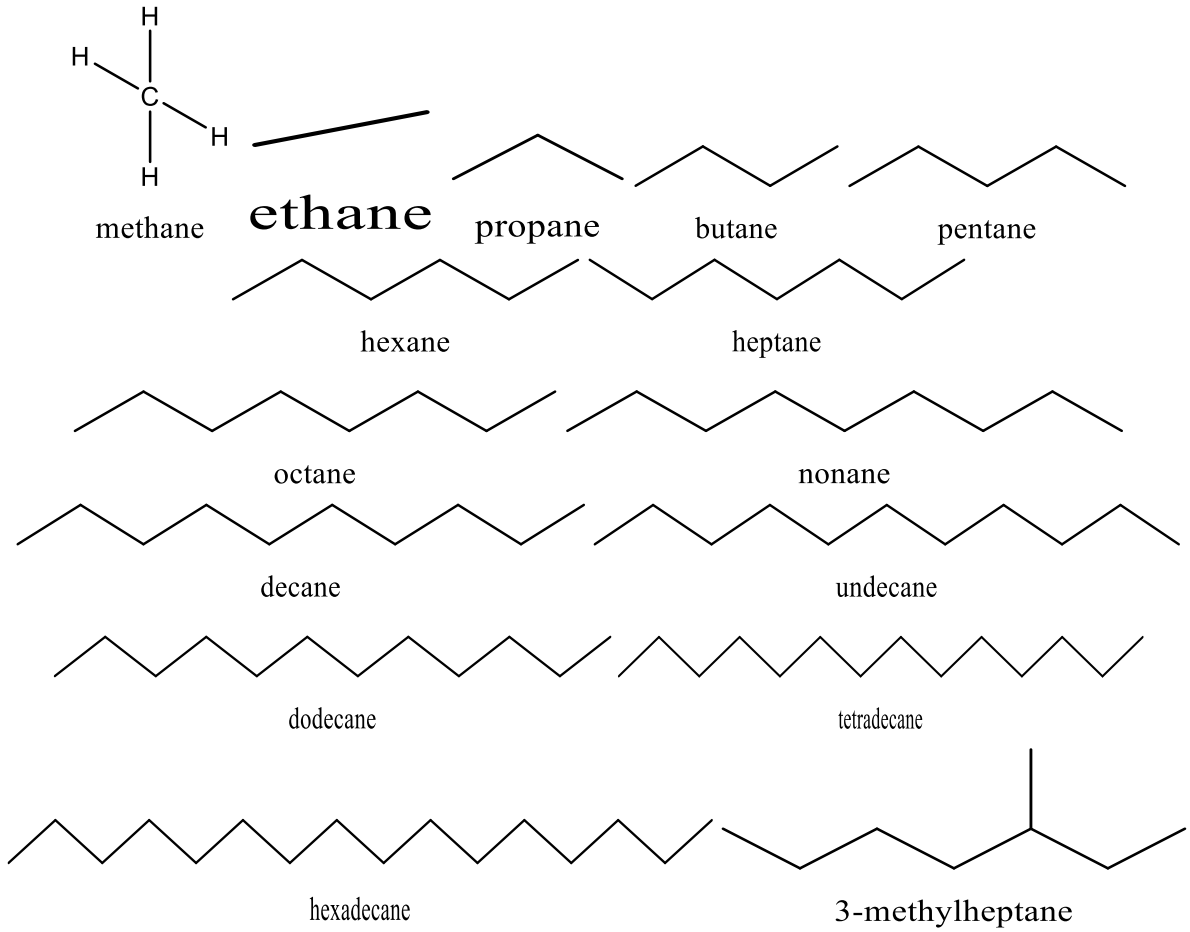
I.8 Études bibliographiques

I.8.1 Les Hydrocarbures

Les hydrocarbures sont des molécules organiques principalement composées de carbone (84%) et d'hydrogène (14%) mais aussi du Soufre (1-3%), de l'azote (-1%) de l'oxygène(-1%) des métaux (-1%) et de sel (-1%). Ils peuvent êtres saturés (alcanes) ou insaturés (alcènes, alcynes et composants aromatiques) tout en présentant une structure linéaire, ramifiée ou cyclique **[12]**.

Les alcanes : les alcanes sont des hydrocarbures saturés : ils sont constitués d'atomes de C et de H et ne possèdent pas de liaison multiple. Les alcanes non cycliques ont pour formule brute C_nH_{2n+2}**[13]**

I.8.2 Les Hydrocarbures utilisés



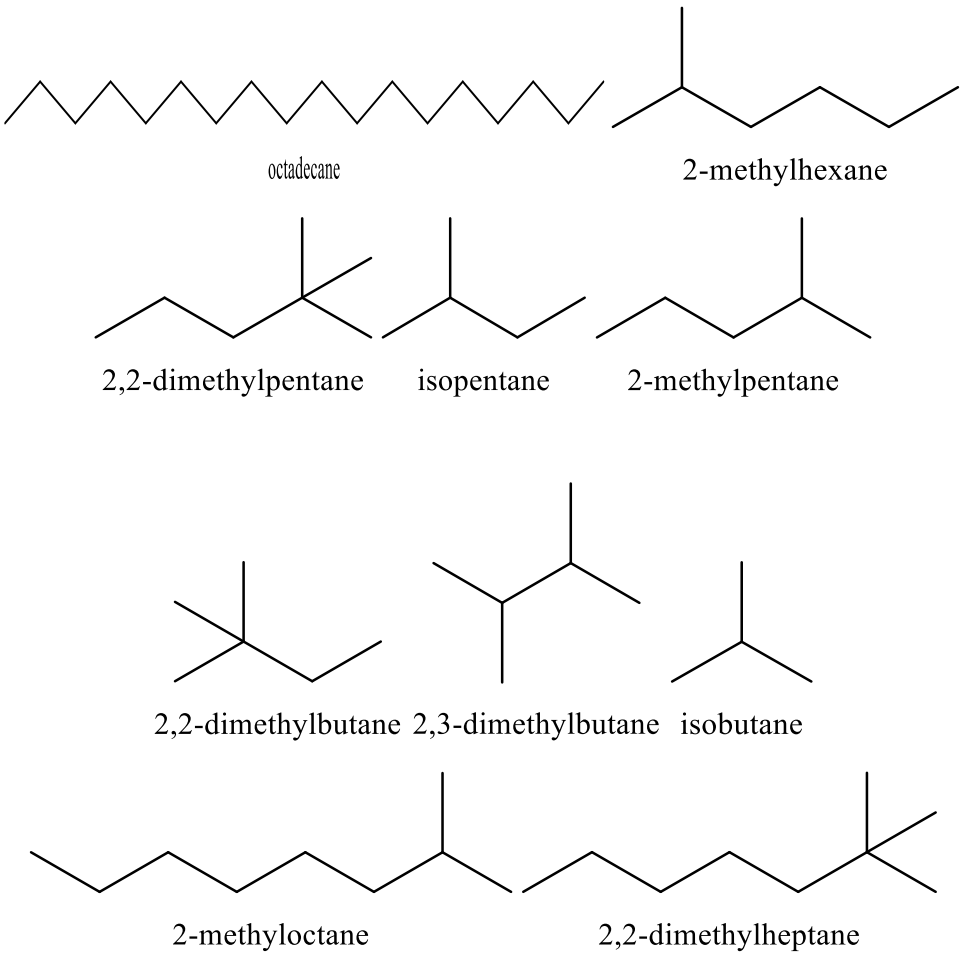


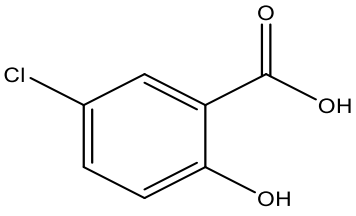
Figure 1. La liste des hydrocarbures utilisés.

I.8.3 Les acides benzoïques

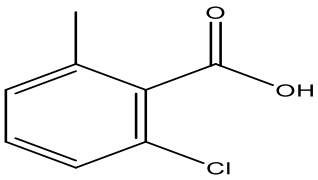
L'acide benzoïque est un acide organique de formule chimique C_6H_5COOH (ou $C_7H_6O_2$) est un acide carboxylique aromatique dérivé du benzène que l'on trouve naturellement dans certaines plantes.

L'acide benzoïque est un acide faible souvent utilisé comme conservateur.il est peu soluble dans l'eau à cause de son cycle aromatique apolaire. Cependant, l'acide salicylique et l'acide acétylsalicylique (aspirine) sont les principaux dérivés de l'acide benzoïque **[12]**.

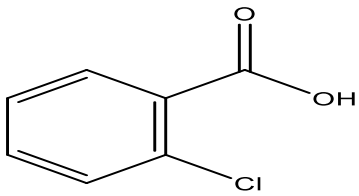
I.8.4Les dérivées des acides benzoïques utilisés



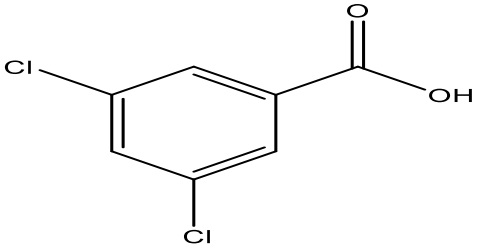
5-chloro-2-hydroxybenzoic acid



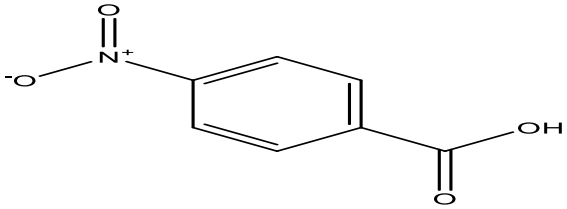
2-chloro-6-methylbenzoic acid



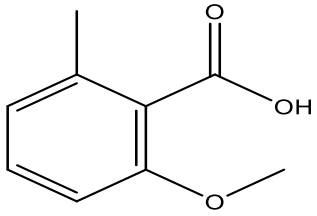
2-chlorobenzoic acid



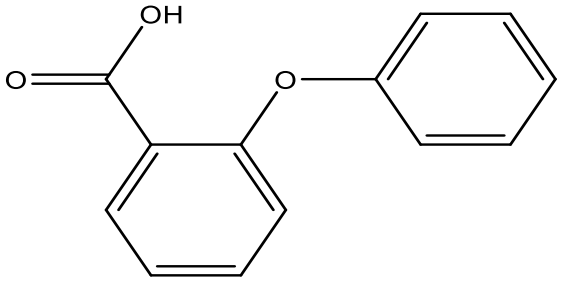
3,5-dichlorobenzoic acid



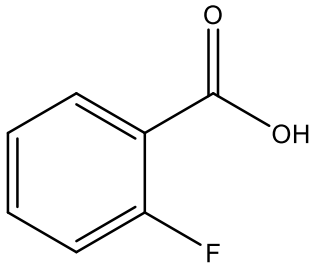
4-nitrobenzoic acid



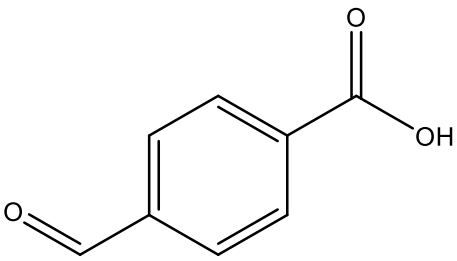
2-methoxy-6-methylbenzoic acid



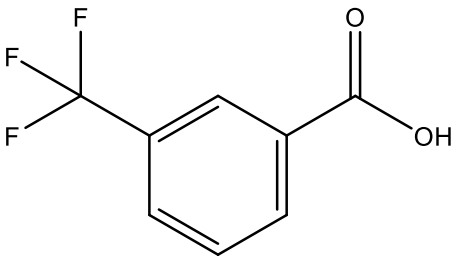
2-phenoxybenzoic acid



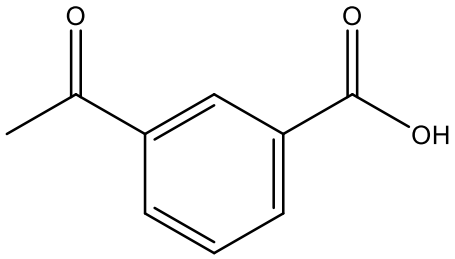
2-fluorobenzoic acid



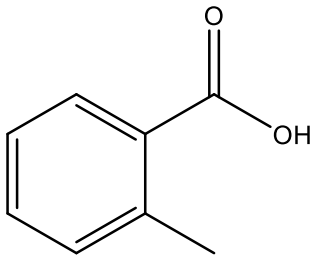
4-formylbenzoic acid



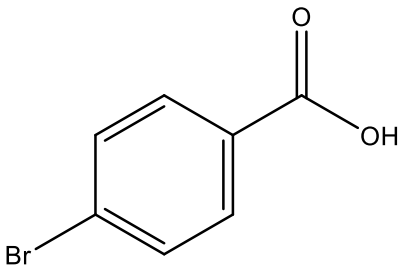
3-(trifluoromethyl)benzoic acid



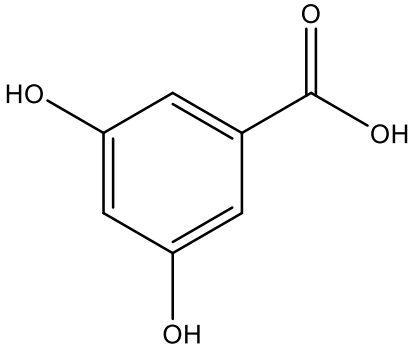
3-acetylbenzoic acid



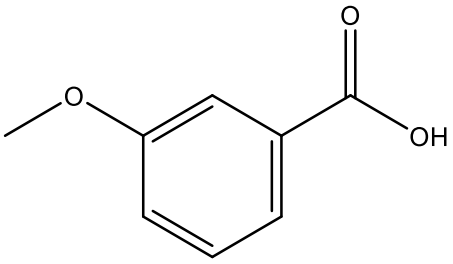
2-methylbenzoic acid



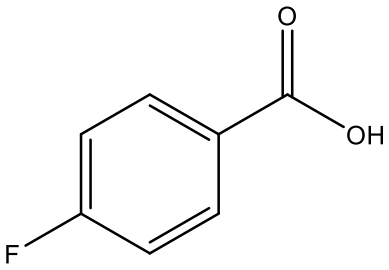
4-bromobenzoic acid



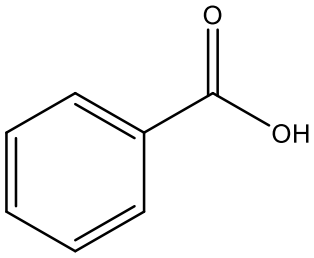
3,5-dihydroxybenzoic acid



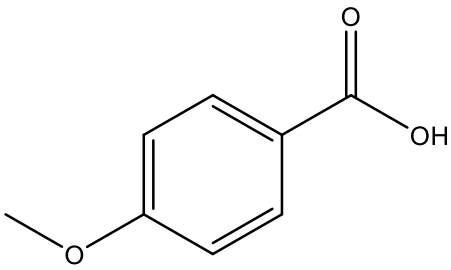
3-methoxybenzoic acid



4-fluorobenzoic acid



benzoic acid



4-methoxybenzoic acid

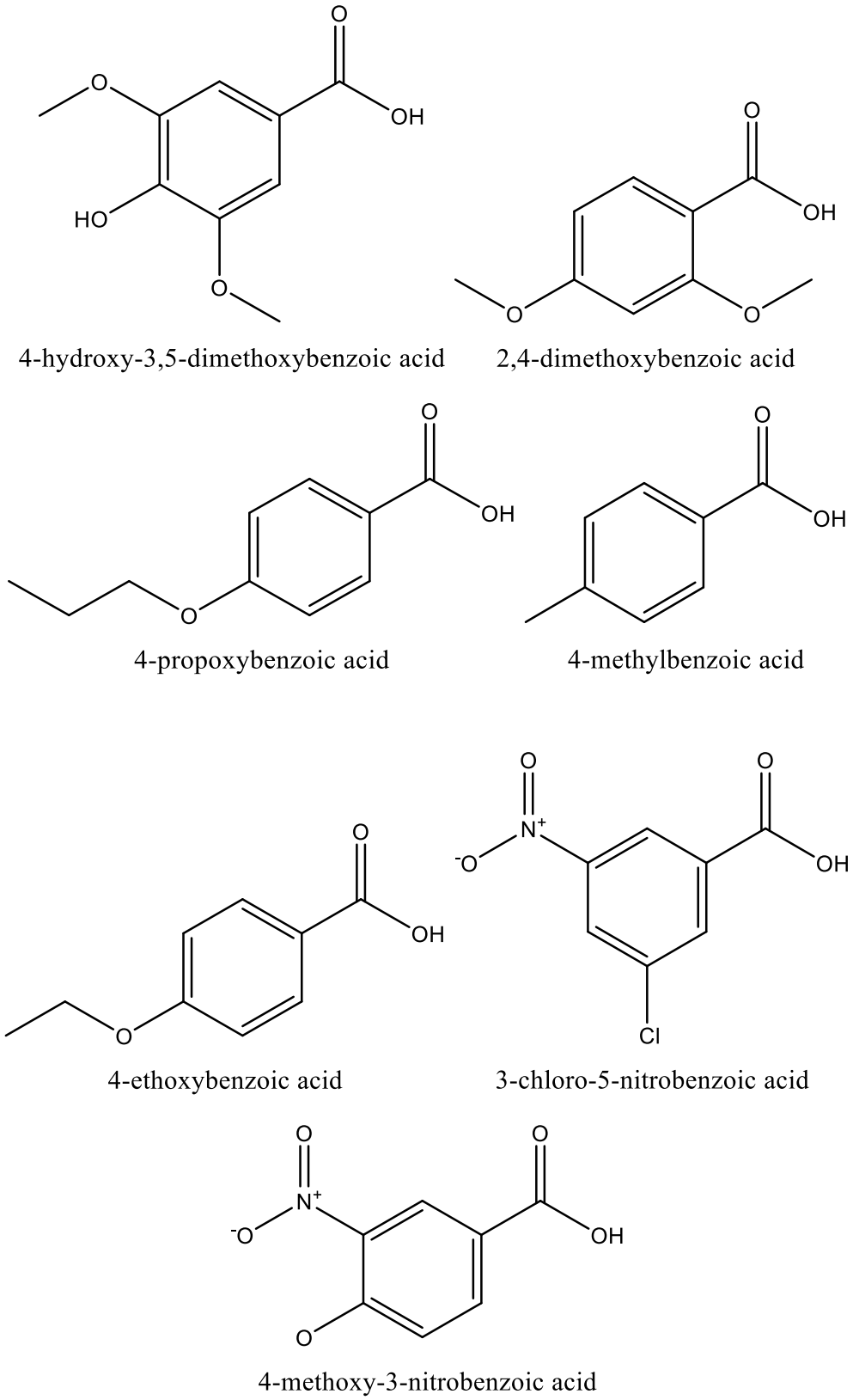


Figure 2. La liste des dérivées des acides benzoïques utilisés.

I.9 Programmes et matériels utilises

Les caractéristiques de l'ordinateur utilisé pour nos calculs sont les suivantes :

- Processeur : (Intel(R)Celeron (R) i3 CPU N3060 @ 1.60GHz 1.60GHz).
- Mémoire installée :2 GB.

- **Le logiciel ChemDraw :*ChemDraw*** est un éditeur de molécules développé pour la première fois en 1985 par David A. Evans et Stewart Rubenstein. ***ChemDraw*** est un programme simple à utiliser qui permet de dessiner de manière intuitive et efficace des représentations bidimensionnelles simples de molécules organiques**[14]**.***ChemDraw***, avec Chem3D et ChemFinder, fait partie de la suite logicielle ChemOffice et est disponible pour Macintosh et Microsoft Windows.

- **Le logiciel Gaussian :*Gaussian*** est un logiciel utilisé en chimie informatique initialement publié en 1970 par John Pople et son groupe de recherche à l'Université Carnegie Mellon. ***Gaussian*** est rapidement devenu un programme de structure électronique très populaire pour prédire les propriétés des molécules et des réactions, notamment : les énergies et structures moléculaires, les énergies et les structures des états de transition, les fréquences vibrationnelles, les propriétés thermochimiques, les énergies de liaison et de réaction, les voies de réaction, les orbitales moléculaires, les charges atomiques, les moments multipolaires, etc. Dans nos expériences, le logiciel ***Gaussian*** a été utilisé pour obtenir les descripteurs. Nous avons utilisé la version 5.0.8 du Gaussian.

- **Le logiciel WEKA :*Weka*** est un logiciel de data mining développé par l'Université de Waikato en Nouvelle-Zélande**[15]**.***Weka*** est une collection d'algorithmes d'apprentissage automatique pour les tâches de data mining. Le logiciel ***Weka*** contient des outils de prétraitement, de classification, de régression, de clustering, de règles d'association et de visualisation des données. Il est également bien adapté au développement de nouveaux programmes d'apprentissage automatique. Dans nos expériences, le logiciel ***WEKA*** a été utilisé pour développer le modèle QSAR.Nous avons utilisé la version 3.9.4 de Weka.

I.10 Conclusion

La chimie quantique est une branche de la chimie théorique qui applique la mécanique quantique aux systèmes moléculaires pour étudier les processus et les propriétés chimiques.

Dans ce chapitre, les différentes équations et estimations sont présentées. Parmi ces méthodes, nous avons insisté sur la théorie de la densité fonctionnelle, que nous avons utilisée pour calculer les descripteurs moléculaires des molécules utilisées dans ce travail.

Chapitre II : Relations Quantitatives Structures Activités/Propriétés (QSAR/QSPR)

Chapitre II : Relations Quantitatives Structures Activités/Propriétés (QSAR/QSPR)

II.1 Introduction

La modélisation quantitative des relations structure-activité/propriété (QSAR/QSPR) est un domaine de recherche très pertinent en chimie pharmaceutique. La modélisation quantitative des relations structure-activité/propriété (QSAR/QSPR) concerne la construction de modèles prédictifs d'activités biologiques en fonction des informations structurales d'une bibliothèque de composés.

Le concept du QSAR/QSPR a généralement été utilisé pour la découverte et le développement de médicaments et a acquis une large applicabilité pour corréler les informations moléculaires non seulement avec les activités biologiques, mais aussi avec d'autres propriétés physico-chimiques, qui a donc été appelée relation quantitative structure-propriété (QSPR).

Ce chapitre vise à couvrir les concepts et techniques essentiels qui sont pertinents pour la réalisation d'études QSAR / QSPR.

II.2 Historique

Le QSAR/QSPR a ses origines dans le domaine de la toxicologie. En 1863, Cros a proposé une relation qui existait entre la toxicité des alcools aliphatiques primaires et leur solubilité dans l'eau **[16]**.

En 1968, Crum-Brown et Fraser**[17]** ont proposé que l'activité biologique d'une molécule dépend de sa composition chimique en postulant le lien entre la constitution chimique et l'action physiologique dans leur enquête pionnière en 1868 comme suit:

> *"Effectuer sur une substance une opération chimique qui doit introduire un changement connu dans sa constitution, puis examiner et comparer l'action physiologique de la substance avant et après le changement"*

Peu de temps après, Richet (1893)**[18]**, Meyer (1899)**[19]** et Overton (1901)**[20]** ont découvert indépendamment une corrélation linéaire entre la lipophilicité (par exemple les coefficients de partage huile-eau) et les effets biologiques (par exemple les effets narcotiques et la toxicité).

En 1935, Hammett (1935, 1937)**[21-22]** a introduit une méthode pour rendre compte des effets des substituants sur les mécanismes de réaction grâce à l'utilisation d'une équation qui a pris deux paramètres en considération à savoir la constante de substituant et la constante de réaction.

En complément du modèle de Hammett, Taft a proposé en 1956 une approche pour séparer les effets polaires, stériques et de résonance des substituants dans les composés aliphatiques**[23]**.

L'année 1964 est considérée comme le début des méthodes QSAR/QSPR. En 1964, Hansch et Fujita **[24]**ont exposé la base mécanistique du développement QSAR / QSPR dans leur développement fondateur de l'équation linéaire de Hansch qui intégrait des paramètres hydrophobes aux constantes électroniques de Hammett. Cette corrélation analysera la relation entre ces méthodes en propriétés physiques et chimiques (log p, pka, paramètres spatiaux et électroniques) et en activité biologique (activité enzymatique, pharmacologie).

Depuis, QSAR / QSPR a reçu une forte impulsion avec le développement de descripteurs, de logiciels et d'ordinateurs plus récents et plus complexes. Cela a joué un rôle déterminant dans l'application des techniques de prédiction qui n'étaient pas réalisables ou qui prenaient auparavant trop de temps.

II.3 Définition

La relation structure-activité (SAR) est une approche pour trouver des relations qualitatives entre la structure chimique et leur activité biologique.

Les modèles QSAR/QSPR (Quantitative Structure Activity/ Property Relationship) sont des modèles théoriques qui relient une mesure quantitative de la structure chimique à une propriété physique ou à une activité biologique. Les modèlesQSAR/QSPR fournissent une solution statistique pour résoudre le problème du calcul des propriétés physique et biologiques directement à partir de la structure.

L'intérêt principal des modèles QSAR/QSPR est d'extraire des informations d'un ensemble de descripteurs numériques qui caractérisent la structure moléculaire, prédisant ainsi l'activité biologique de la nouvelle structure.

II.4 Principe général du QSPR

Les modèles QSPR reposent sur le principe que « produits chimiques structurellement similaires sont susceptibles d'avoir des propriétés physico-chimiques et biologiques similaires ». Les méthodes QSAR/QSPR permet d'établir une la relation mathématique qui relie quantitativement les structure molécules, en caractéristiques moléculaires avec activité biologique ou propriété physicochimique appelées descripteurs. La forme mathématique générale de QSAR/QSPR est représentée par l'équation suivante :

Activité biologique = *f* (propriété physicochimique)

Les modèles QSAR/QSPR ont la forme suivante :

$$\mathbf{A_{pred} = f(D_1, D_2, \ldots D_n)}$$

Où :

- A_{pred} : activité biologique à prédire.
- D1,D2,...Dn: propriétés chimiques ou structurelles (descripteurs moléculaires)

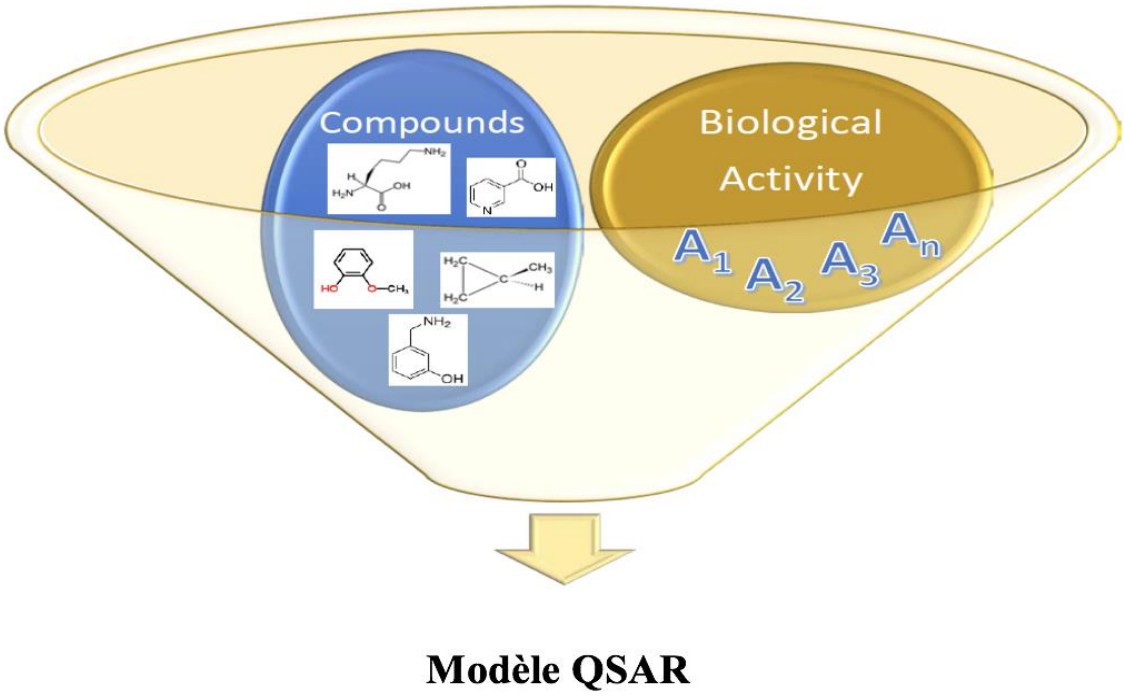


Figure 3. Schéma général d'un modèle QSAR/QSPR

Où :

- A1,A2,...An: activité biologique des produits chimiques d'entraînement.

II.5 Le processus du QSAR/QSPR

Les méthodologies QSAR/QSPR ont le potentiel de réduire considérablement le temps et les efforts requis pour la découverte de nouveaux composées. Une étape majeure dans la construction des modèles QSAR/QSPR est de trouver un ensemble ayant une activité cible descripteurs moléculaires qui représentent les variations des propriétés structurales de la molécule. L'analyse QSAR/QSPR utilise différents types de méthodes pour dériver une relation mathématique quantitative entre la structure chimique et l'activité biologique.

La représentation schématique de la construction du modèle QSAR/QSPR est donnée dans la figure 4. Une brève description des étapes impliquées dans la génération du modèle QSAR/QSPR est présentée dans la section suivante.

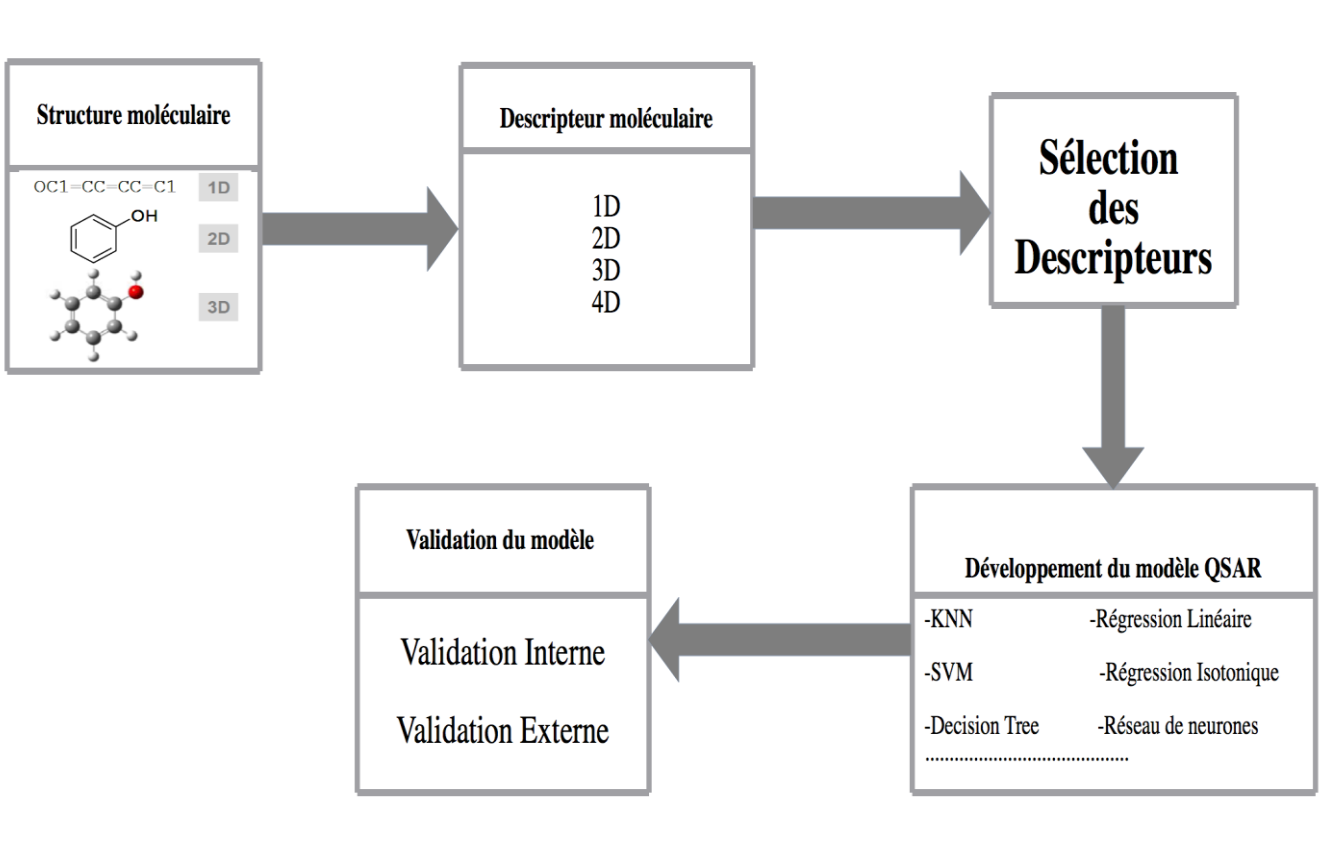


Figure 4. Le processus du QSAR/QSPR.

II.5.1 La collection et la compréhension des données

La collection et la compréhension des données est une étape cruciale qu'il ne faut pas négliger car elle aide le chercheur à se familiariser avec la nature des données avant la construction réelle du modèle QSAR / QSPR.

La collection et la compréhension des données s'effectue par une analyse exploratoire des données qui commence souvent par une simple observation de la matrice de données des échantillons.

En QSAR/QSPR, les échantillons de données représentent chaque composé unique ; les types de données se réfèrent aux caractéristiques ou aux types de données comme la valeur particulière est représentée, qui sont essentiellement de nature qualitative ou quantitative.

II.5.2 La Génération des descripteurs moléculaires à partir de la structure chimique

Les descripteurs moléculaires peuvent être définis comme les informations essentielles d'une molécule en termes de ses propriétés physico-chimiques telles que les descripteurs constitutionnels, électroniques, géométriques, hydrophobes, lipophiles, stériques, chimiques quantiques et topologiques.

Les descripteurs sont des valeurs numériques qui expriment les propriétés physiques et chimiques des molécules en fonction de leur représentation structurelle. Les descripteurs moléculaires peuvent être calculés à l'aide d'un logiciel de chimie quantique général tel que **Gaussian[25]**.

À l'heure actuelle, il existe un grand nombre de descripteurs moléculaires qui peuvent être utilisés dans les études QSAR/QSPR. Il existe différents niveaux de représentation chimique allant de 1D à 4D. Dans ce qui suit, nous allons présenter la classification des descripteurs moléculaires :

- **Les descripteurs 1 D :** des descripteurs unidimensionnels sont accessibles à partir de la formule générale de la molécule, et il faut noter que ces descripteurs ne permettent pas la distinction des isomères structuraux, ni le développement de modèles plus complexes.

- **Les descripteurs 2 D :** les descripteurs bidimensionnels sont des attributs numériques qui peuvent être calculés à partir à partir de la table de connectivité de la molécule ou à partir de la représentation plane (bidimensionnelle) de la structure.
- **Les descripteurs 3 D :** évaluer les descripteurs 3D des molécules en fonction de leurs positions relatives dans l'espace et décrire des propriétés plus complexes ; par conséquent, leurs calculs nécessitent généralement une modélisation moléculaire pour comprendre la géométrie tridimensionnelle des molécules. Par conséquent, ces descripteurs s'avèrent être relativement coûteux en calcul, mais ils fournissent plus d'informations et sont nécessaires pourla modélisation d'entités ou d'activités qui dépendent de structures 3D.
- **Les descripteurs 4 D:**il correspond à la mesure des propriétés tridimensionnelles (potentiel électrostatique, hydrophobicité, liaison hydrogène, etc.) de molécules n'importe où dans l'espace. Fournit des informations sur l'architecture cible. On distingue donc les descripteurs 4D, et ces descripteurs sont obtenus en calculant le champ d'interaction moléculaire (CoMFA, CoMSIA) entre une molécule et une sonde représentée par une autre molécule (eau, amide, etc.).

Une fois que les descripteurs moléculaires ont été calculés, ils serviront de variables indépendantes pour la poursuite de la construction du modèle QSAR/QSPR.

II.5.3 La Sélection des descripteurs moléculaires les plus pertinents

Les descripteurs utilisés au développement des modèles QSAR/QSRP doivent être utiles, faciles et interprétables d'un point de vue chimique et phénoménologique, c'est pourquoi il faut identifier les descripteurs les plus pertinents pour le sujet.

Pour trouver le meilleur modèle QSAR/QSPR, une sélection préalable de descripteurs (avant le calcul des modèles QSAR/QSPR et une sélection d'ensembles de descripteurs est une étape obligatoire.

La sélection des descripteurs doit supprimer les descripteurs "non significatifs" ayant une très faible corrélation avec la propriété dépendante.

Il existe différentes méthodes pour la sélection des descripteurs, telles que Ant Colony Optimization **[26],** Forward / Backward Stepwise (Avant / Arrière Pas à pas) **[27]**, les algorithmes génétiques**[28]**, la recherche séquentielle ...etc.

La sélection des descripteurs peut être effectuée à l'aide de plusieurs logiciels tel que ***WEKA*[15]**, etc.…

II.5.4 Le développement du modèle QSAR/QSPR

De nombreuses méthodes ont été utilisées pour mapper les descripteurs moléculaires aux propriétés. Les modèles QSAR/QSPR sont généralement divisés en deux classes selon le mode de fonctionnement : basé sur des règles (système expert) et basé sur des statistiques (système QSAR/QSPR).

Dans les sections suivantes, nous allons présenter la classification des méthodes de développement des modèles QSAR/QSPR:

II.5.4.1 Les modèles QSAR/QSPR basés sur des règles

La relation structure-activité (SAR) est une étude visant à comprendre comment les changements dans la structure moléculaire provoquent des changements dans les propriétés moléculaires, dont certains sont liés à des changements dans l'activité biologique. La recherche SAR suit des règles et des procédures strictes pour remplir des formulaires valides et fiables.

Les systèmes basés sur des règles également appelés systèmes experts sont des systèmes de prise de décision informatisés interactif et fiable qui utilisent à la fois des faits et des heuristiques pour résoudre des problèmes de prise de décision complexes.

Un système expert vise à imiter un expert humain pour raisonner sur un problème et faire des prédictions ou des recommandations. La connaissance d'une forme généralisée, stockée dans une base de règles, est utilisée par un programme, communément appelé moteur d'inférence, pour émettre des jugements.

Dans les systèmes QSAR/QSPR basés sur des règles, la prédiction qualitative est basée sur la présence de caractéristiques structurelles des produits chimiques d'essai. Les premiers systèmes QSAR/QSPR basés sur des règles étaient ceux de James et Elisabeth Miller **[30]**et les travaux ultérieurs de John Ashby et Raymond Tennant, qui ont systématisé la relation entre les structures chimiques et les résultats toxiques observés **[31]**.

II.5.4.2 Les modèles QSAR/QSPR basés sur des modèles statistiques

Les modèles QSAR/QSPR basés sur des méthodes statistiques sont des modèles basés sur des propriétés physico-chimiques exprimées en termes de descripteurs moléculaires ou de fragments structuraux connus pour être corrélés aux activités biologiques. La relation quantitative entre l'activité biologique et les descripteurs moléculaires est calculée par un algorithme d'apprentissage automatique**[32]**.

L'analyse statistique peut également identifier des descripteurs qui sont liés les uns aux autres, de sorte que seuls les descripteurs originaux sont préservés, ce qui réduit la duplication des informations **[33]**.

L'analyse statistique peut identifier la corrélation entre le descripteur et la variable cible. Il met également l'accent sur la contribution relative de chaque descripteur à l'interprétation globale de la performance.

Ce qui suit une représentation dequelques méthodes statistiques pour obtenir un modèle QSAR/QSPR.

- **La Régression linéaire : (LR : Linear Regression)** la régression linéaire est une technique de modélisation statistique. La régression linéaire cherche à modéliser la relation entre une variable dépendante Y et une ou plusieurs variables indépendantes X.Un modèle de régression linéaire simple est de la forme suivante **[34]**.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

  Où :

  - **y :** la variable à prédire ;
  - **x :** la variable exogène (indépendante, explicative) ;
  - β_0 β_1**:** les paramètres a estimé par le modèle ;
  - **ε:**l'erreur du modèle.

- **Le perceptron multicouche: (MLP : Multi Layer Perceptron)**le Perceptron multicouche est un classificateur linéaire de type réseau neuronal formel organisé en plusieurs couches au sein desquelles une information circule de la couche d'entrée vers la couche de sortie. Chaque couche est constituée d'un nombre variable de neurones, les neurones de la couche de sortie correspondant toujours aux sorties du système**[35]**.

  Comparés aux autres méthodes de classification supervisée, les réseaux de neurones sont rapides et permettent de régler le taux de mauvaise classification mais nécessitent également un entraînement long et laborieux, car demandant une certaine expertise pour

optimiser les différents paramètres.

Le perceptron multicouche de base se compose d'au moins trois nœuds disposés en trois couches fonctionnelles (figure 5) :

1. Couche d'entrée - où les informations entrent ;

2. Couche cachée - celle où se trouve toute l'action ;

3. Couche de sortie - les résultats de l'opération.

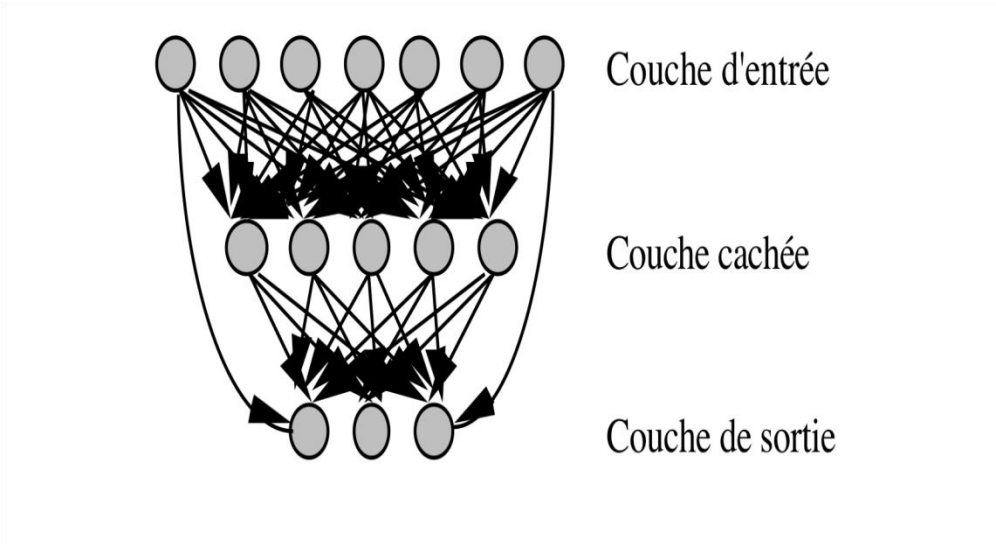


Figure 5. Les couches d'un réseau multicouche.

- **Le K plus proche voisin : (KNN : K-Nearest Neighbors) :**est un algorithme simple, qui stocke tous les cas et classifie les nouveaux cas en fonction de la mesure de similarité**[36]**. Le principe général du KNN est que des instances similaires ont une classification similaire. Du point d'apprentissage au point d'échantillonnage, la distance est évaluée et le point avec la distance la plus faible est appelé le plus proche voisin (Figure 6).

  L'algorithme de K plus proche voisin simple est effectué en 2 étapes :

  Étape 1 : devons choisir la valeur de K, (les points de données les plus proches).

  Étape 2 : Pour chaque point des données de test, procédez comme suit -

- Calculer la distance entre les données de test et chaque ligne de données d'apprentissage à l'aide de l'une des méthodes à savoir: distance euclidienne, Manhattan ou Hamming. La méthode la plus couramment utilisée pour calculer la distance est euclidienne.
- En fonction de la valeur de la distance, triez-les par ordre croissant.
- Choisir les K premières lignes du tableau trié.
- Attribuer une classe au point de test en fonction de la classe la plus fréquente de ces lignes.

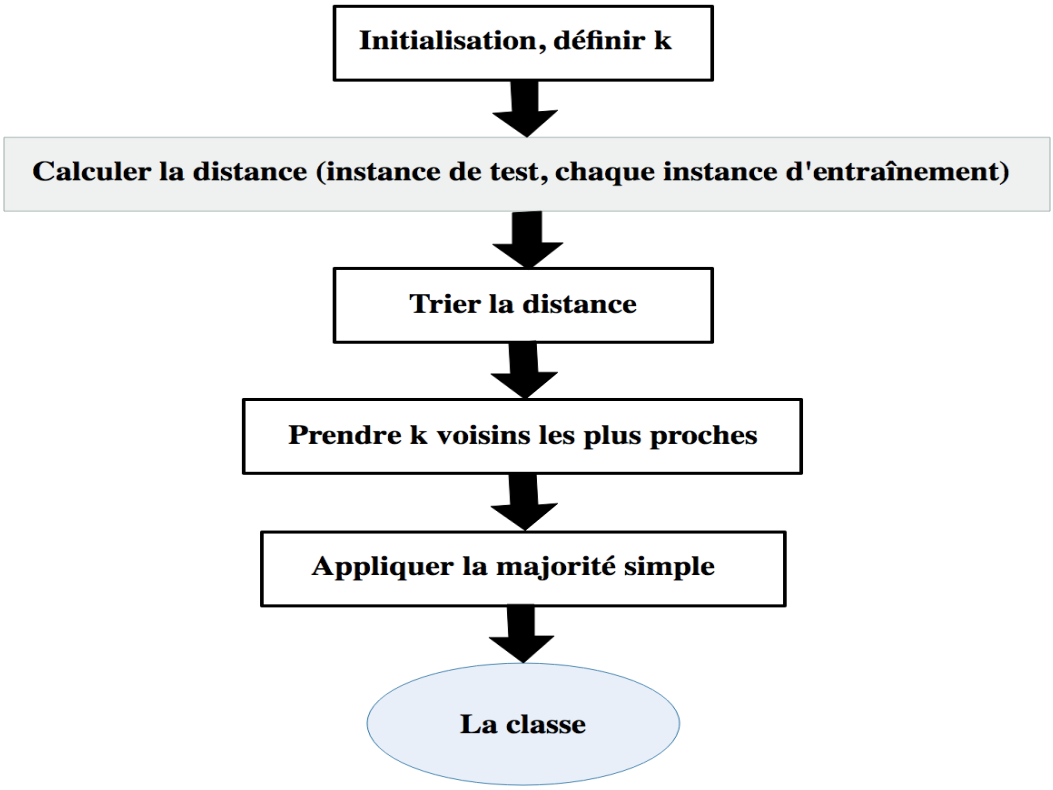


Figure 6. Les étapes de l'algorithme KNN.

- **L'analyse en composantes principales (PCA : Principal Components Analysis)** est une technique statistique qui permet d'identifier des modèles linéaires sous-jacents dans un ensemble de données afin qu'elle puisse être exprimée en termes d'autres ensembles de données d'une dimension significativement inférieure sans trop de perte d'information.

L'idée principale de le PCA est de déterminer les modèles et les corrélations entre les différentes caractéristiques de l'ensemble de données**[37]**. En trouvant une forte corrélation entre différentes variables, une décision finale est prise concernant la réduction des dimensions des données de telle sorte que les données significatives soient toujours conservées.

Un tel processus est très essentiel pour résoudre des problèmes complexes axés sur les données qui impliquent l'utilisation d'ensembles de données de grande dimension.

- **Les Arbres de Décision :** les algorithmes basés sur des arbres sont une famille populaire de méthodes d'apprentissage supervisées pour la classification et la régression**[37]**.

  Les arbres de décision ressemblent à un arbre avec une règle de décision à la racine, à partir de laquelle les règles de décision ultérieures s'étalent en dessous. Dans les arbres de décision les données sont continuellement divisées selon un certain paramètre. L'arbre peut être expliqué par deux entités, à savoir les nœuds de décision et les feuilles.
- **Machine a Vecteur de Support (SVM : Support Vector Machine):** Les SVM sont des algorithmes qui utilisent une transformation non linéaire des données d'apprentissage**[37]**.

  Ils projettent les données d'apprentissage dans un espace de plus grande dimension que leur espace d'origine. Dans ce nouvel espace, ils cherchent l'hyperplan qui permet une séparation linéaire optimale des données d'apprentissage en utilisant les vecteurs de support et les marges définies par ces vecteurs.

  La technique SVM fait partie des techniques classiques de fouille de données. Elle fait partie des méthodes d'apprentissage qui ont réalisé des performances meilleures que les méthodes statistiques traditionnelles en matière de classification.

II.5.5 Evaluation du modèle

II.5.5.1 Evaluation Interne

La performance interne est généralement évaluée à partir de la performance prédictive de l'ensemble d'apprentissage lors du développement des modèles QSAR.

La performance interne des modèles QSAR peut être évaluée en divisant les données d'apprentissage en deux ensembles : un ensemble d'apprentissage et un ensemble de validation. L'ensemble d'apprentissage est utilisé pour construire le modèle QSAR dont les performances prédictives sont évaluées sur l'ensemble de validation **[38]**.

Une autre technique d'évaluation interne est appelé validation croisée ou un ou

plusieurs composés sont éliminé de l'ensemble de données au hasard dans chaque cycle et le modèle est construit en utilisant le reste des composés. Le processus est répété jusqu'à ce que tous les composés soient éliminés une fois.

II.5.5.2 Evaluation Externe

Après le développement des modèles QSAR/QSPR, la manière dont le modèle se généralise sur des données invisibles est un aspect tout aussi important qui doit être pris en compte. L'évaluation du modèle développé permet de le tester par rapport à des données qui n'ont jamais été utilisées lors de l'apprentissage et de voir sa capacité de faire de bonnes prédictions sur des échantillons futurs, ou des échantillons qu'il n'a jamais vus auparavant **[38]**.

L'utilisation correcte des techniques d'évaluation du modèle, de sélection de meilleurs modèles et de sélection des méthodes de développement du modèle est vitale dans le développement de bon modèles QSAR/QSPR.

II.5.5.3 Les métriques d'évaluation

Dans cette section on passe en revue de différentes métriques pouvant être utilisées pour évaluer les modèles QSAR/QSPR.

- **La Corrélation (R) :** Le coefficient de corrélation de Pearson (r) est un paramètre couramment utilisé pour décrire le degré d'association entre deux variables d'intérêt. La valeur r calculée de deux variables d'intérêt peut prendre une valeur allant de –1 à +1, la première indiquant une corrélation indirecte (négative) tandis que la seconde suggère une corrélation directe (positive). Pour décrire la performance prédictive relative d'un modèle QSAR/QSPR, r est utilisé pour mesurer la corrélation entre les valeurs d'intérêt expérimentales (x) et prédites (y) afin d'observer la variabilité qui existe entre les variables**[39]**.

- **L'erreur absolue moyenne (MAE) :** est une indication de l'écart moyen des valeurs prédites par rapport aux valeurs observées correspondantes et peut présenter des informations sur les performances à long terme des modèles ; plus la MAE est faible, meilleure est la prédiction du modèle à long terme **[40]**.

- **L'erreur quadratique moyenne (RMSE) :**RMSE présente des informations sur

l'efficacité à court terme qui est une référence de la différence des valeurs prédites par rapport aux valeurs observées. Plus le RMSE est bas, plus l'évaluation est précise**[40]**.

- **Coefficient de Détermination (R^2) :** Le coefficient de détermination (également appelé carré R) mesure la variance interprétée par le modèle, qui est la réduction de la variance lors de l'utilisation du modèle **[40]**.
- **La Précision :** La précision représente le nombre de données classées correctement, comme positives par exemple, par rapport au nombre de données totales reconnues comme positives :

$$precision = \frac{\text{VP}}{VP + FP}$$

Où:

  - **VP :**varie positive
  - **FP :** Faux positive

II.6 Les applications du QSAR/QSPR

Les modèles QSAR/QSPR ont une grande importance en termes de prédiction des activités biologiques. Certaines applications des modèles QSAR/QSPR sont mentionnées ci-dessous :

- Application de certains modèles SAR dans les domaines de l'industrie.
- Optimisation de l'activité pharmacologique, biocide ou pesticide et identification rationnelle de nouveaux composé chimique pour leur activité.
- Rationalisation et prédiction des effets combinés des molécules, que ce soit dans des mélanges ou des formulations.
- Prédire une variété de propriétés physiques et chimiques des molécules.
- Prédire la toxicité de molécules pour les humains et les espèces environnementales.
- Le concept de toxicité et d'effets secondaires des nouveaux composés et la connaissance des composés dangereux au début du développement du produit.

II.7 Les limites et défis du QSAR/QSPR

Certaines limites des modèles QSAR/QSPR sont citées ci-dessous**[41]**:

- Un modèle QSAR/QSPR ne peut pas être considéré comme un modèle universel, parce qu'il est développé sur un nombre limité de composés qui ne couvrent pas tout l'espace chimique.
- L'activité/propriété prédite d'un composé, chimiquement dissimilaire au jeu d'apprentissage, ne pourra pas être considérée fiable.
- Effet de la qualité et de la quantité des données d'entraînement sous-jacentes.

II.8 Conclusion

Actuellement les méthodes QSAR/QSPR sont employées pour prédire l'activité d'une molécule avant de faire sa synthèse. C'est pour cette raison que les relations quantitatives structures-activités sont devenues de plus en plus utilisées dans le cadre de la conception de nouvelles entités chimiques. Dans ce chapitre nous avons présenté la méthode de QSAR/QSPR en décrivant ses différentes étapes ; la génération et la sélection des descripteurs moléculaires, les différents types des modèles et des méthodes d'apprentissage et la validation interne et externe des modèles de QSAR/QSPR.

Chapitre III: Résultats et Discussion

Chapitre III : Résultats et Discussion

III.1 Introduction

Un modèle QSPR est un modèle mathématique qui établit une relation entre les caractéristiques dérivées de la structure d'un composé et son activité biologique sous la forme d'un modèle mathématique.

Le but de ce chapitre est de présenter des modèles QSPR (Quantitative Structure Property Relationship) et d'évaluer leurs performances dans la prédiction d'activités biologiques. Pour ce faire, un ensemble de données a été collecté et a été soumis à des processus tels que le calcul des descripteurs, la sélection de descripteurs pertinents, la modélisation QSPR suivie de l'évaluation des modèles développés à l'aide de divers paramètres statistiques.

III.2. Calcul et Sélection des descripteurs

III.2.1 Représentation des molécules

Le travail a été effectué avec une version d'essai du logiciel ***ChemDraw***. Un exemple de représentation d'une molécule est représenté dans la figure suivante :

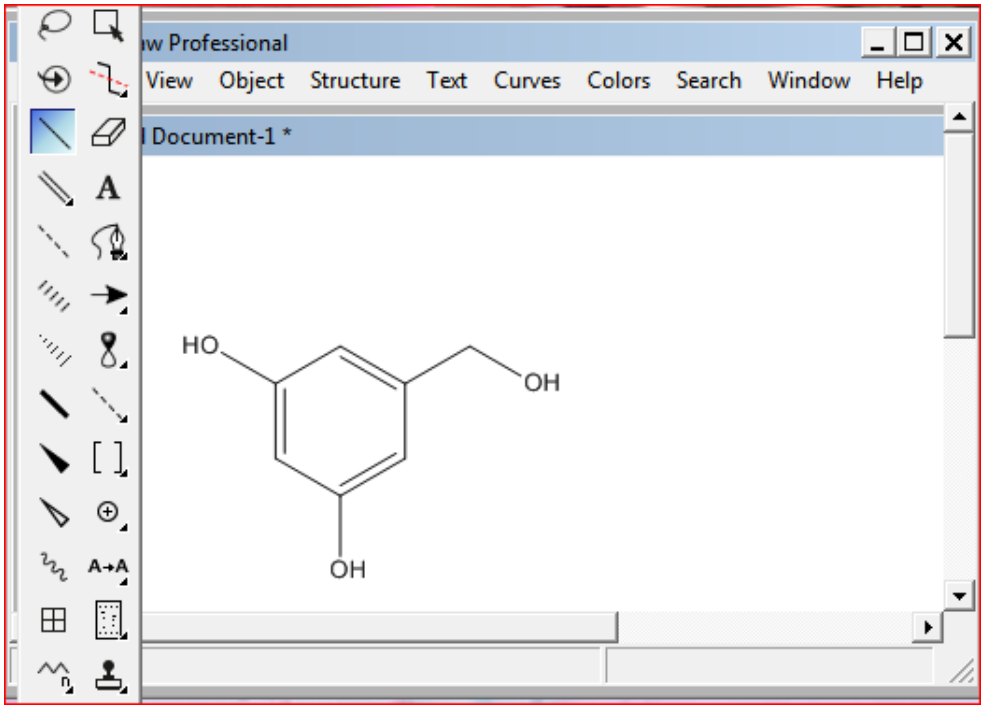
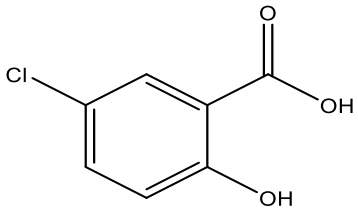
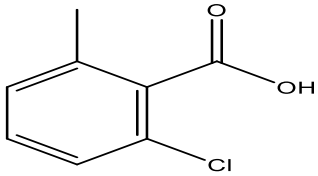
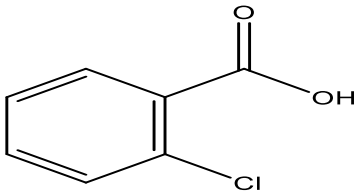
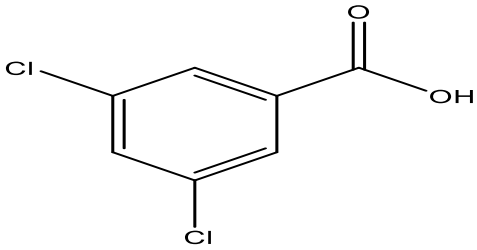
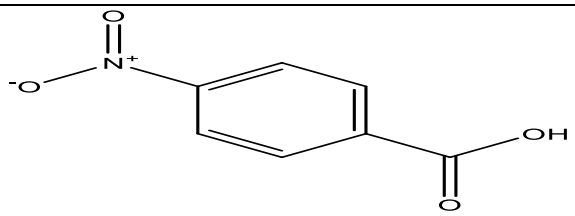
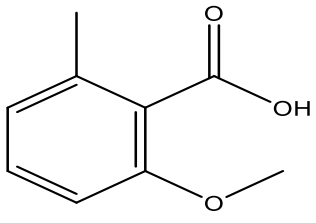
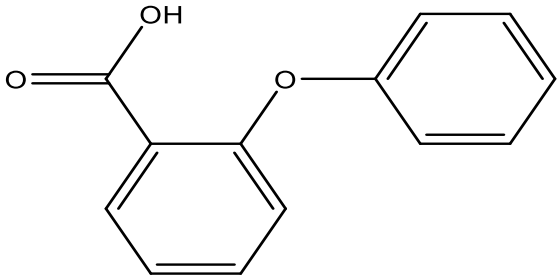
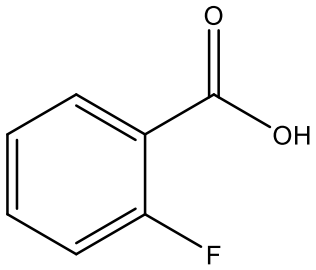
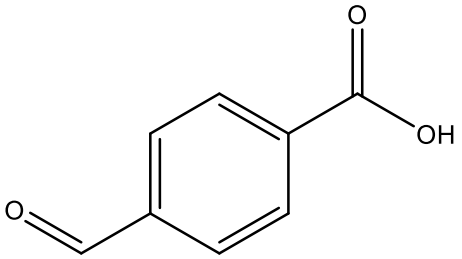
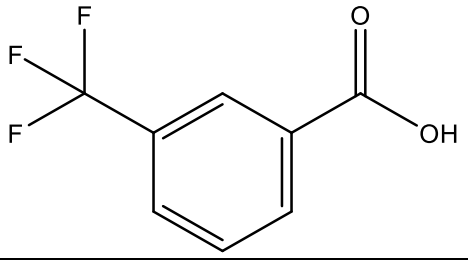


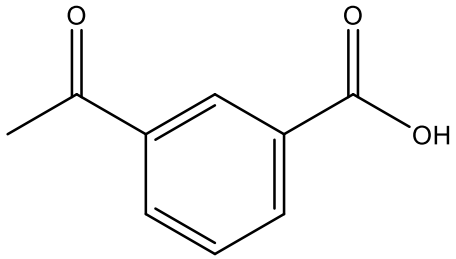
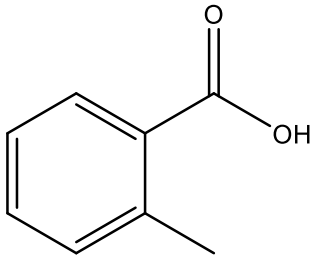
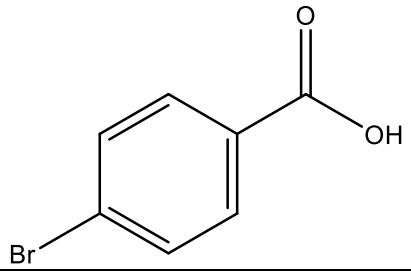
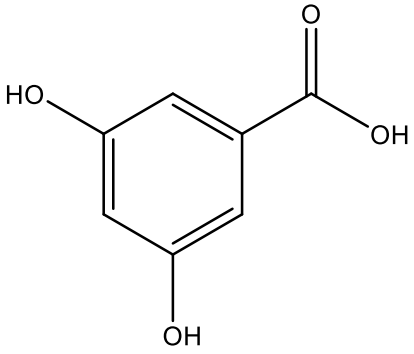
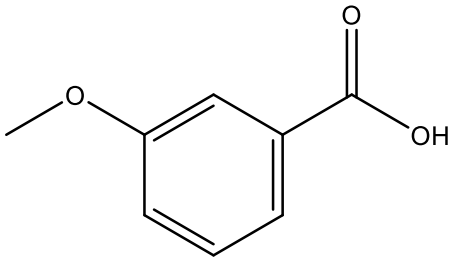
Figure 7. Exemple de représentation d'une molécule avec le logiciel ***ChemDraw***.

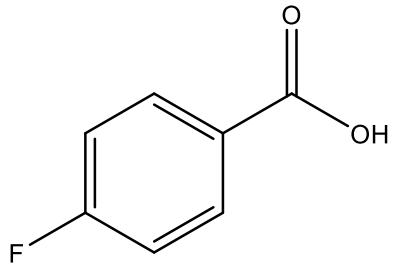
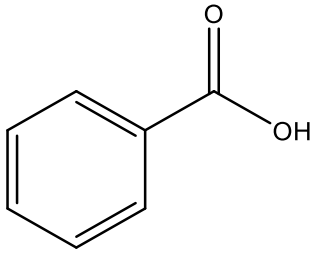
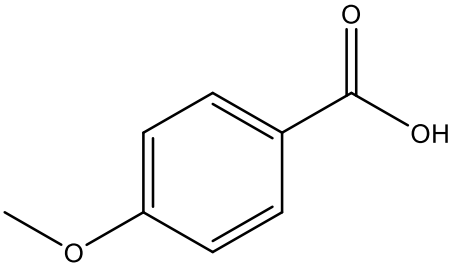
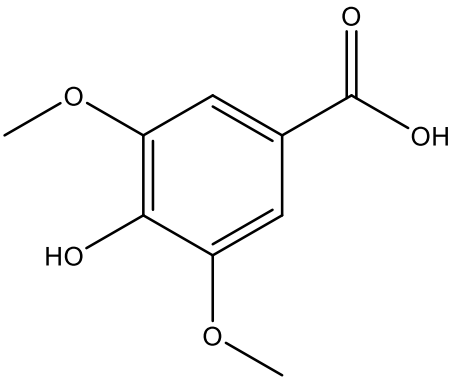
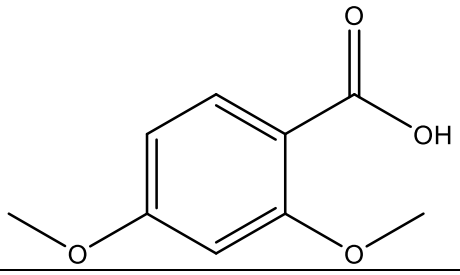
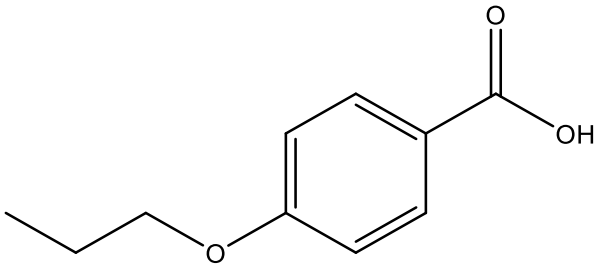
Dans notre travail nous avons utilisé 50 molécules dont 25 hydrocarbures et 25 dérivés d'acide benzoïque, Les résultats de représentations de ces molécules par le Logiciel ***ChemDraw*** sont résumés dans les ***tableaux 1*** et ***2*** pour les dérivés des acides benzoïques et les hydrocarbures comme suit :

Tableau 1. Les structures des dérivés d'acide benzoïque.

N°	***Nom***	***Structure***
1	Acide 2-hydroxy-5-chlorobenzoïque	
2	Acide 2-méthyl-6-chlorobenzoïque	
3	Acide 2-chlorobenzoïque	
4	Acide 3-5-dichlorobenzoïque	
5	Acide 4-nitrobenzoïque	
6	Acide 2-méthyl-6-méthoxybenzoïque	

		
7	Acide 2-phénoxybenzoïque	
8	Acide 2-fluorobenzoïque	
9	Acide 4-formylbenzoïque	
10	Acide 3-trifluorométhylbenzoïque	

11	Acide 3-acétylbenzoïque	
12	Acide 2-méthylbenzoïque	
13	Acide 4-bromobenzoïque	
14	Acide 3-5-dihydroxybenzoïque	
15	Acide 3-méthoxybenzoïque	

16	Acide 4-fluorobenzoïque	
17	Acide benzoique	
18	Acide 4-méthoxybenzoïque	
19	Acide 4-hydroxy-3-5-diméthoxybenzoïque	
20	Acide 2-4-diméthoxybenzoïque	
21	Acide 4-propoxybenzoïque	

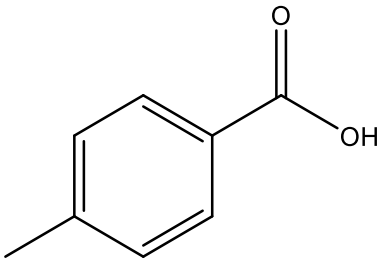
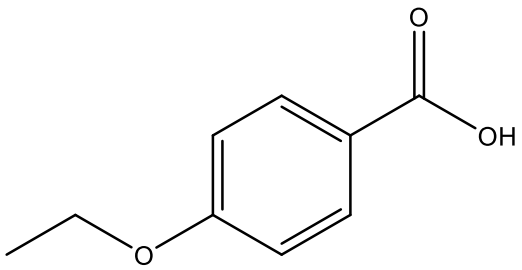
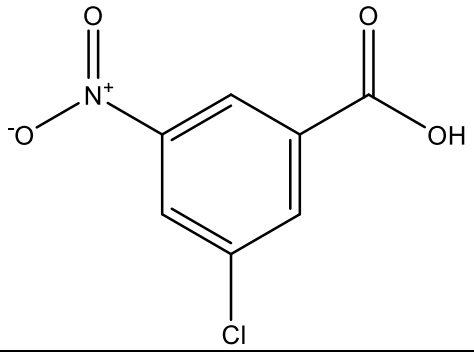
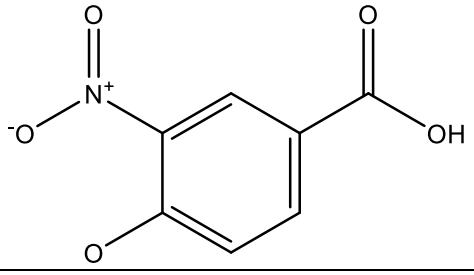
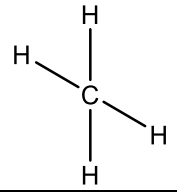
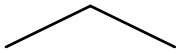
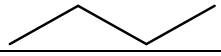
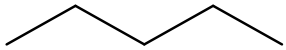
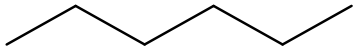
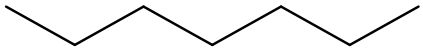
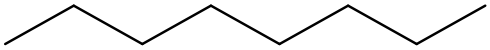
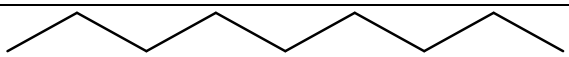
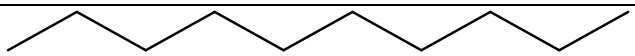
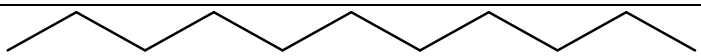
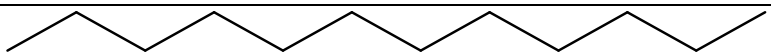
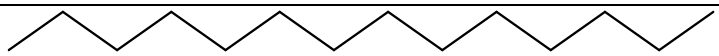
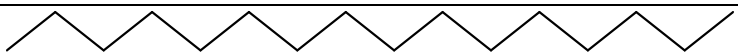
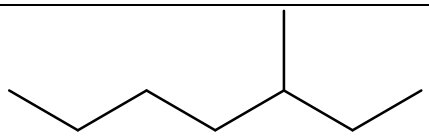
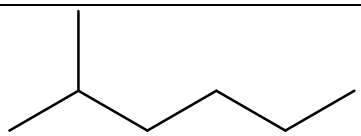
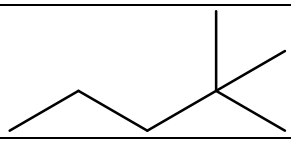
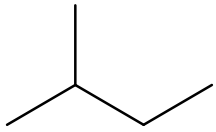
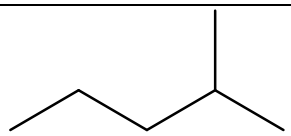
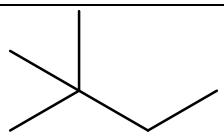
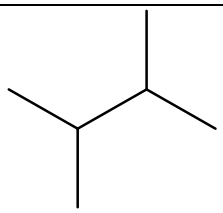
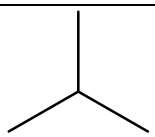
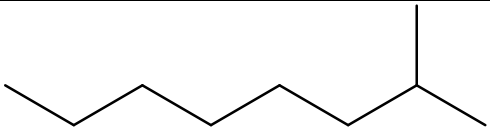
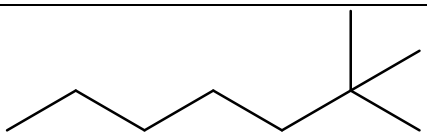
22	Acide 4-méthylbenzoïque	
23	Acide 4-éthoxybenzoïque	
24	Acide 3-chloro-5-nitrobenzoïque	
25	Acide 3-nitro-4-méthoxybenzoïque	

Tableau 2. Les structures des hydrocarbures.

N°	Nom	Structure
1	Methane	
2	Ethane	
3	Propane	
4	Butane	
5	Pentane	
6	Hexane	
7	Heptane	
8	Octane	
9	Nonane	
10	Decane	
11	Undecane	
12	Dodecane	
13	Tetradecane	
14	Hexadecane	
15	Octadecane	
16	3-Methylheptane	
17	2-Methylhexane	
18	2,2-Dimethylpentane	
19	Isopentane (2-methylbutane)	
20	2-Methylpentane	

21	2,2-Dimethylbutane	
22	2.3-dimethylbutane	
23	Isobutane (2-methylpropane)	
24	2-Methyloctane	
25	2,2-Dimethylheptane	

III.2.2 Calculs sur le logiciel Gaussian

Les molécules dessinées sur ***ChemDraw***, sont introduites dans le logiciel ***Gaussian***, pour optimisation par la méthode '***DFT***', pour pouvoir ensuite calculer les différents descripteurs moléculaires.

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

Gaussian a été utilisé pour calculer les paramètres suivants :Highest Occupied Molecular Orbital (**HOMO**), Lowest Unoccupied Molecular Orbital (**LUMO**), Le Gap énergétique (**ΔE**) , au niveau B3LYP/6-31G .Le nombre d'atome de carbone (**NC**), le nombre d'atome d'hydrogène (**NH**), Le moment dipolaire ($\mathbf{M_{Dipole}}$), le nombre d'atome d'oxygène (**NO**), le nombre d'atome de chlore (**NCL**), le nombre d'atome de azote (**Nn**), le nombre d'atome de fluor (**Nf**), le nombre d'atome de brom (**Nbr**). Les résultats des calculs des paramètres physico-chimiques sont représentés dans les ***tableaux 3*** et ***4***.

Tableau 3: Résultats de calcul des descripteurs pour les dérivées des acides benzoïques.

N°	HOMO(u.a)	LUMO(u.a)	ΔE(u.a)	M_{Dipole}(u.a	NC	NH	NO	NCL	Nn	Nf	Nbr
1	-0.24308	-0.07605	-0.16703	2.72	7	5	3	1	0	0	0
2	-0.26875	-0.05871	-0.21004	4.74	8	7	2	1	0	0	0
3	-0.28298	-0.0735	-0.20948	5.56	7	5	2	1	0	0	0
4	-0.27493	-0.08391	-0.19102	1.67	7	4	2	2	0	0	0
5	-0.29672	0.04552	-0.34224	4.02	7	5	4	0	1	0	0
6	-0.2536	-0.05307	-0.20053	2.91	13	10	3	0	0	0	0
7	-0.24077	-0.04523	-0.19554	2.88	9	10	3	0	0	0	0
8	0.70328	-0.06543	0.76871	3.49	7	5	2	0	0	1	0
9	-0.27677	-0.1042	-0.17257	4.58	8	6	3	0	0	0	0
10	-0.28918	-0.07806	-0.21112	4.33	8	5	2	0	0	3	0
11	-0.25756	-0.07543	-0.18213	4.99	9	8	3	0	0	0	0
12	-0.26481	-0.05466	-0.21015	5.46	8	8	2	0	0	0	0
13	-0.2605	-0.0679	-0.1926	1.37	7	5	2	0	0	0	1
14	-0.21139	0.00614	-0.21753	1.5	7	8	3	0	0	0	0
15	-0.23137	-0.05465	-0.17672	3.56	8	8	3	0	0	0	0
16	-0.26871	-0.06346	-0.20525	1.3821	7	5	2	0	0	1	0
17	-0.2646	-0.05711	-0.20749	2.1008	7	6	2	0	0	0	0
18	-0.23526	-0.04779	-0.18747	2.5729	8	8	3	0	0	0	0
19	-0.2147	-0.04929	-0.16541	0.5627	9	10	5	0	0	0	0
20	-0.22751	-0.0411	-0.18641	3.7736	9	10	4	0	0	0	0
21	-0.25459	-0.05281	-0.20178	2.5924	8	8	2	0	0	0	0
22	-0.23316	-0.04629	-0.18687	2.9901	10	12	3	0	0	0	0
23	-0.23355	-0.04659	-0.18696	2.8571	9	10	3	0	0	0	0
24	-0.29368	-0.12675	-0.16693	4.4632	7	4	4	1	1	0	0
25	-0.26359	-0.10051	-0.16308	7.5666	8	7	5	0	1	0	0

Tableau 4: Résultats de calcul des descripteurs pour les Hydrocarbures.

N°	HOMO(u.a)	LUMO(u.a)	ΔE(u.a)	M_{Dipole}	NC	NH
1	-0.39383	0.00529	-0.39912	0	1	4
2	-0.34452	0.00546	-0.34998	0	2	4
3	-0.32885	0.00635	-0.3352	0.0865	3	8
4	-0.32346	0.00573	-0.32919	0	4	10
5	-0.31754	0.00624	-0.32378	0.0813	5	12
6	-0.31132	0.00644	-0.31776	0.0003	6	14
7	-0.30656	0.00675	-0.31331	0.0815	7	16
8	-0.30294	0.00695	-0.30989	0	8	18
9	-0.30002	0.00715	-0.30717	0.0816	9	20
10	-0.29769	0.00728	-0.30497	0	10	22
11	-0.28924	0.08913	-0.37837	0.0417	11	24
12	-0.28763	0.08889	-0.37652	0	12	26
13	-0.28513	0.08853	-0.37366	0	14	30
14	-0.2833	0.08829	-0.37159	0	16	34
15	-0.28195	0.08811	-0.37006	0	18	38
16	-0.29506	0.08588	-0.38094	0.0336	8	18
17	-0.30023	0.0866	-0.38683	0.0471	7	16
18	-0.2898	0.08267	-0.37247	0.0664	7	16
19	-0.30865	0.08558	-0.39423	0.052	5	12
20	-0.31721	0.08516	-0.40237	0.0719	6	14
21	-0.31007	0.03506	-0.34513	0.0585	6	14
22	0.30045	0.07853	-0.37898	0.058	6	14
23	-0.31721	0.08516	-0.40237	0.0719	4	10
24	-0.29363	0.08639	-0.38002	0.0476	9	20
25	-0.293	0.07715	-0.37015	0.0546	9	20

Tableau 4: Résultats de calcul des descripteurs pour les Hydrocarbures.

III.2.3 Propriété chimique

La construction d'un modèle QSPR est très dépendante des données expérimentales de référence. Le choix de la base de données est un point critique de son développement. Dans la plupart des cas, les données expérimentales sont issues de la littérature.

Les données de l'activité à modéliser peuvent être considéré comme étant un effet biologique (activité toxique, …) ou d'une propriété physio-chimique (point de fusion, …). Dans cette section nous allons citer les activités biologiques utilisées dans l'analyse QSPR qui sont utilisées dans nos expériences.

- **PKA :** détermine la force d'un acide. Les données utilisées dans ce travail présentées dans le ***Tableau 5*** ont été prélevées à partir du CRC Handbook of Chemistry and Physics **[42].**

Tableau 5. La Propriété chimique **PKa** de la série des dérivées des acides benzoïques étudiées.

N°	***Nom***	**Pka**
1	Acide 2-hydroxy-5-chlorobenzoïque	2.59
2	Acide 2-méthyl-6-chlorobenzoïque	2.75
3	Acide 2-chlorobenzoïque	2.92
4	Acide 3-5-dichlorobenzoïque	3.1
5	Acide 4-nitrobenzoïque	3.43
6	Acide 2-méthyl-6-méthoxybenzoïque	3.46
7	Acide 2-phénoxybenzoïque	3.53
8	Acide 2-fluorobenzoïque	3.57
9	Acide 4-formylbenzoïque	3.69
10	Acide 3-trifluorométhylbenzoïque	3.75
11	Acide 3-acétylbenzoïque	3.83
12	Acide 2-méthylbenzoïque	3.91
13	Acide 4-bromobenzoïque	3.99
14	Acide 3-5-dihydroxybenzoïque	4.04
15	Acide 3-méthoxybenzoïque	4.12
16	Acide 4-fluorobenzoïque	4.15
17	Acide benzoique	4.21
18	Acide 4-méthoxybenzoïque	4.25
19	Acide 4-hydroxy-3-5-diméthoxybenzoïque	4.34
20	Acide 2-4-diméthoxybenzoïque	4.36
21	Acide 4-propoxybenzoïque	4.46
22	Acide 4-méthylbenzoïque	4.51
23	Acide 4-éthoxybenzoïque	4.8
24	Acide 3-chloro-5-nitrobenzoïque	3.13
25	Acide 3-nitro-4-méthoxybenzoïque	3.72

- **Température de fusion :** ils'agit de la température à laquelle un corps pur passe de l'état solide à l'état liquide. Cette température se note en général T_{fus} ou θ_{fus} , se mesure en kelvin (°k).Les données utilisées dans ce travail présentées dans le ***Tableau 6***ont été prélevées à partir des valeurs indiquées dans le travail de Beghou.M**[8]**.

Tableau 6. La Propriété chimique température de fusion de la série des hydrocarbures étudiée.

N°	***Nom***	**T_{fusion}**
1	Methane	90.69
2	Ethane	90.35
3	Propane	91.45
4	Butane	134.79
5	Pentane	143.43
6	Hexane	177.84
7	Heptane	182.59
8	Octane	216.39
9	Nonane	219.66
10	Decane	243.49
11	Undecane	247.57
12	Dodecane	264
13	Tetradecane	279
14	Hexadecane	291
15	Octadecane	301
16	3-Methylheptane	152.63
17	2-Methylhexane	154.89
18	2,2-Dimethylpentane	149.37
19	Isopentane (2-methylbutane)	113
20	2-Methylpentane	119.48
21	2,2-Dimethylbutane	173.33
22	2.3-dimethylbutane	144.35
23	Isobutane (2-methylpropane)	113.54
24	2-Methyloctane	192.79
25	2,2-Dimethylheptane	160.16

III.2.4 La Sélection des Descripteurs

La sélection des descripteurs a été performée avant le développement des modèles QSAR. Le logiciel ***WEKA*****[15]**a été utilisé pour effectuer cette étape. Dans ***WEKA***, De nombreuses techniques de sélection de fonctionnalités sont prises en charge. Afin de sélectionner des descripteurs, il faut spécifier une méthode de recherche et un évaluateur d'attributs. La méthode de recherche représente un algorithme de recherche (tel que

Exhaustive Search, *Genetic Search*, *BestFirst*, etc.), tandis que l'évaluateur d'attribut spécifie un moyen de calculer la valeur optimisée au cours de la sélection des descripteurs.

- **La méthode de recherche** par défaut dans ***WEKA*** est **BestFirst**. Cette méthode recherche l'espace des sous-ensembles de descripteurs par une escalade augmentée d'une fonction de retour en arrière. La méthode BestFirst peut commencer par l'ensemble vide de descripteurs et effectuer des recherches vers l'avant (comportement par défaut), ou commence par l'ensemble complet des attributs et des recherches vers l'arrière, ou commence à rechercher dans les deux directions (en considérant tous les ajouts de descripteurs uniques et suppressions à un moment donné).
- **L'évaluateur d'attributs** par défaut dans ***WEKA*** est **CfsSubsetEval**. Cette méthode évalue la valeur d'un sous-ensemble de descripteurs en considérant la capacité prédictive individuelle de chacun ainsi que le degré de redondance entre les descripteurs. Les sous-ensembles de descripteurs qui sont fortement corrélés avec les valeurs de propriété / activité et ayant une faible inter-corrélation sont préférés.

Dans notre travail, la combinaison par défaut **CfsSubsetEval** (évaluateur d'attributs) et **BestFirst** (méthode de recherche) est appliquée pour sélectionner les descripteurs.

Les Tableaux 7 et ***8*** représentent la liste des descripteurs sélectionnées pour chaque famille de molécules après le processus de sélection sous ***WEKA***.

Tableau 7 Liste des descripteurs obtenus après le processus de sélection sous Weka pour les dérivées des acides benzoïques.

Descripteur	**Description**
HOMO	Highest Occupied Molecular Orbital
ΔE	Le Gap énergétique
$\mathbf{M_{Dipole}}$	Le moment dipolaire
NCL	Le nombre d'atome de chlore

Tableau 8 Liste des descripteurs obtenus après le processus de sélection sous Weka pour les Hydrocarbures.

Descripteur	Description
ΔE	Le Gap énergétique
M_{Dipole}	Le moment dipolaire
NC	Le nombre d'atome de carbone

III.3 Développement du modèle QSPR

III.3.1 La répartition des données

Dans ce travail, l'ensemble des données (molécules) est divisé en deux parties : des données d'apprentissage et des données de test qui permettent de tester l'efficacité des modèles développés comme indiqué dans le ***Tableau 9***.

Tableau 9 La répartition des données

	Nombre de molécules
Données d'apprentissage	20
Données de Test	5

Dans la phase d'apprentissage, les données ont aussi été divisés en deux parties dont 80% des données sont utilisées pour développer le modèle (l'apprentissage) et 20% est utilisé pour valider le modèle pour chaque famille de molécules.

Pour le test (validation), nous avons pris cinq (les 5 derniers) composés qui n'ont pas participé au développement du modèle et nous avons calculé avec le modèle obtenu les ***Pka*** et $\boldsymbol{T_{fusion}}$ (estimée) afin de valider nos modèles.

III.3.2 Métriques d'évaluation

L'évaluation des modèles développés dans cette étude a été évaluée à l'aide des paramètres statistiques suivants :

- Coefficient de corrélation

$$\mathbf{R= \sum x_i y_{i-}(\sum x_i \sum y_i/N) \sqrt{(\sum x_i^2- (\sum x_i)^2/N) (\sum y_i^2- (\sum y_i)^2/N)}}$$

- Mean Absolute Error (MAE**) [43]**

$$\mathbf{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{(\boldsymbol{i}=\mathbf{1})}^{\boldsymbol{n}} |\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}}|$$

- Mean Absolute Square Error (MASE**) [43]**

$$\mathbf{RMSE} = \sqrt{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \frac{(\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})^{\mathbf{2}}}{\boldsymbol{n}}}$$

Où : x_i et $y_i$et $\hat{y}_i$désignent la valeur des activités réelle et prévue pour le ième composé, et «N» est le nombre de composés.

III.3.3 Méthodes statistiques pour former la relation Structure-Activité

Trois techniques statistiques ont été utilisées pour réaliser une étude QSPR qui sont : la régression linéaire (LR), perceptron multicouche (MLP) et le K plus proches voisins (KNN).

III.3.3.1 Les résultats des méthodes statistiques développés pour dérivés de l'acide benzoïque

a. La régression linéaire

L'analyse de la régression de l'activité étudiée nous a donné le modèle statistique à quatre descripteurs comme suit**(équation 1)**:

Pka = (-2.1868 * HOMO) + (1.8854 * AE) + (-0.1262 * $\mathbf{M_{Dipole}}$) + (-0.7445*NCL) + 4.1245**… (équation 1)**

Les résultats obtenus sont indiqués dans le ***Tableau 10*** suivant:

Tableau 10 : Performance de la régression linéaire sur les données d'apprentissage et les données de test pour lesdérivés de l'acide benzoïque.

	R	**MAE**	**RMSE**
Apprentissage	0.92	0.28	0.31
Test	0.97	0.45	0.53

Le tableau suivant représente une comparaison des données expérimentales avec les données prédites par le modèle basé sur la régression linéaire pour les dérivés de l'acide benzoïque.

Tableau 11 : Comparaison des données réelles avec les données prédites par le modèle basé sur la régression linéaire pour les dérivés de l'acide benzoïque.

N°	Valeurs Réelles	Valeurs Prédites	Erreur
1	4.46	3.974	-0.486
2	4.51	3.905	-0.605
3	4.8	3.922	-0.878
4	3.13	3.144	0.014
5	3.72	3.438	-0.282

b. Perceptron multicouche

Comme le montre le ***Tableau 12***, le modèle MLP donne de bons résultats avec des coefficients de corrélation élevés (R), à la fois dans l'ensemble d'apprentissage et dans l'ensemble de test, ce qui indique que le modèle MLP est non seulement bien performé dans le développement de modèles, mais aussi donne une excellente prédiction.La Figure 8 montre les couches de perceptron obtenues par lemodèle MLP.

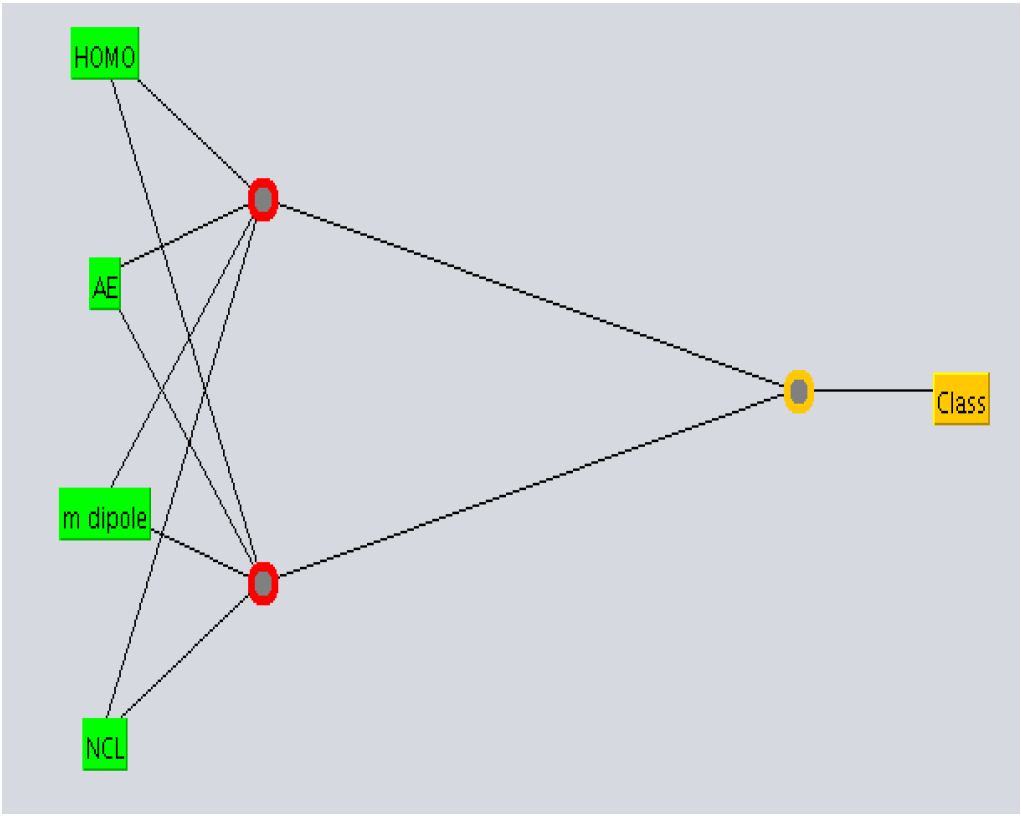


Figure 8. Perceptron multicouche du modèle développé pour les dérivés des acides benzoïques.

Tableau 12 : Performance du MLP sur les données d'apprentissage et les données de test pour les dérivés de l'acide benzoïque.

	R	MAE	RMSE
Apprentissage	0.88	0.3	0.33
Test	0.98	0.8	0.84

Le tableau suivant (***Tableau 13***) représente une comparaison des données expérimentales avec les données prédites avec par le modèle basé sur le MLP pour les dérivés de l'acide benzoïque.

Tableau 13 : Comparaison des données réelles avec les données prédites par le modèle basé sur le MLP pour les dérivés de l'acide benzoïque.

N°	Valeurs Réelles	Valeurs Prédites	Erreur
1	4.46	3.56	-0.9
2	4.51	3.596	-0.914
3	4.8	3.607	-1.193
4	3.13	2.69	-0.44
5	3.72	3.165	-0.555

c. K-plus proche voisin

K-NN prédit une classification des échantillons de test sur la base de leur similarité avec les exemples de la base d'apprentissage. La proximité est mesurée par les métriques de distance euclidienne. Si plusieurs instances ont la même (la plus petite) distance par rapport à l'instance de test, la première trouvée sera utilisée. Les résultats obtenus sont représentés dans le ***Tableau 14***.

Tableau 14 : Performance du KNN sur les données d'apprentissage et les données de test pour les dérivés de l'acide benzoïque.

	R	MAE	RMSE
Apprentissage	0.91	0.19	0.25
Test	0.54	0.66	0.79

Le ***Tableau 15*** représente une comparaison des données expérimentales avec les données prédites par le modèle basé sur la méthode KNN pour les dérivés de l'acide benzoïque.

Tableau 15 : Comparaison des données réelles avec les données prédites par le modèle basé sur la méthode KNN pour les dérivés de l'acide benzoïque.

N°	Valeurs Réelles	Valeurs Prédites	Erreur
1	4.46	3.89	-0.57
2	4.51	3.495	-1.015
3	4.8	3.495	-1.305
4	3.13	2.835	-0.295
5	3.72	3.87	0.15

III.3.3.2 Les résultats desméthodes statistiques développés pour les Hydrocarbures

a. La régression linéaire

L'analyse de la régression de l'activité étudiée nous a donné le modèle statistique à deux descripteurs suivant **(équation 2)**:

T_{fusion}= (423.3858 * AE) + (-228.0317 * $\mathbf{M_{Dipole}}$) + (14.4956 * NC) + 225.4794..**(équation 2)**

La performance du modèle QSPR basé sur la régression linéaire est montrée dans le ***Tableau 16***.

Tableau 16 : Performance de la régression linéaire sur les données d'apprentissage et les données de test pour les hydrocarbures.

	R	MAE	RMSE
Apprentissage	0.98	9.12	12.35
Test	0.86	15.59	17.38

Le ***Tableau 17*** représente une comparaison des données expérimentales avec les données prédites par le modèle basé sur la régression linéaire pour les hydrocarbures.

Tableau 17 : Comparaison des données réelles avec les données prédites par le modèle basé sur la régression linéaire pour les hydrocarbures.

N°	Valeurs Réelles	Valeurs Prédites	Erreur
1	173.33	152.99	-20.34
2	144.35	138.773	-5.577
3	113.54	96.709	-16.831
4	192.79	184.191	-8.599
5	160.16	186.773	26.613

b. Perceptron multicouche

La Figure 9 montre les couches du perceptron obtenues par le MLP pour les hydrocarbures.

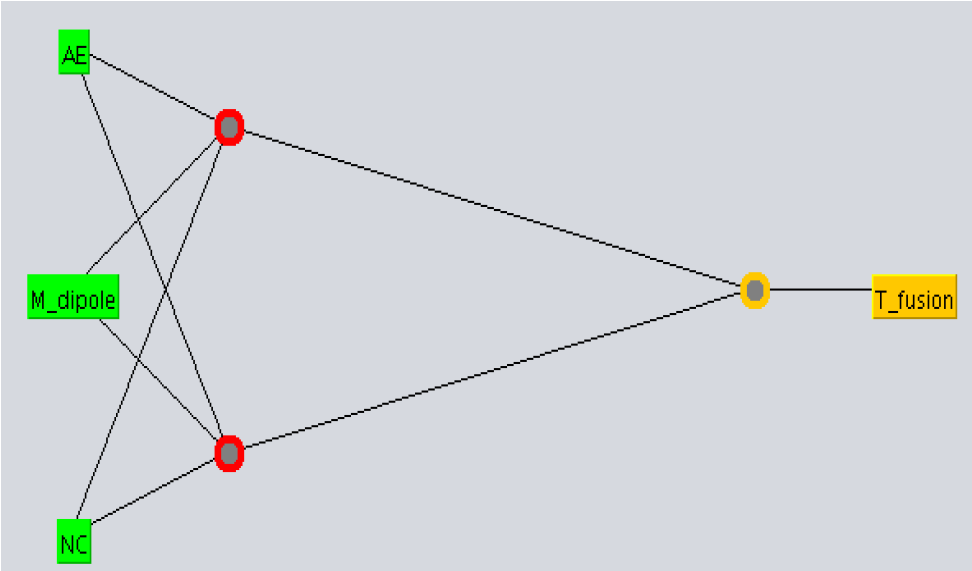


Figure 9. Perceptron multicouche du modèle développé pour les hydrocarbures.

Le ***Tableau 18*** montre les résultats du modèle basé sur MLP comme suit :

Tableau 18 : Performance du MLP sur les données d'apprentissage et les données de test pour les hydrocarbures.

	R	MAE	RMSE
Apprentissage	0.96	8.32	11.89
Test	0.82	20.76	23.60

Le tableau suivant représente une comparaison des données expérimentales avec les données prédites par le modèle basé sur le MLP pour les hydrocarbures.

Tableau 19 : Comparaison des données réelles avec les données prédites par le modèle basé sur le MLP pour les hydrocarbures.

N°	Valeurs Réelles	Valeurs Prédites	Erreur
1	173.33	147.053	-26.277
2	144.35	125.658	-18.692
3	113.54	93.759	-19.781
4	192.79	190.426	-2.364
5	160.16	196.888	36.728

c. K-plus proche voisin

Les résultats d'évaluation du modèle QSPR sont présentés dans le tableau suivant :

Tableau 20 : Performance du KNN sur les données d'apprentissage et les données de test pour les hydrocarbures.

	R	MAE	RMSE
Apprentissage	0.87	22.81	27.42
Test	0.77	16.83	21.59

Le tableau suivant représente une comparaison des données expérimentales avec les données prédites par le modèle basé sur le KNN pour les hydrocarbures.

Tableau 21 : Comparaison des données réelles avec les données prédites par le modèle basé sur le KNN pour les hydrocarbures.

N°	Valeurs Réelles	Valeurs Prédites	Erreur
1	173.33	146.4	-26.93
2	144.35	152.13	7.78
3	113.54	116.24	2.7
4	192.79	201.23	8.44
5	160.16	198.47	38.31

III.4 Discussion

La relation d'activité de la structure quantitative joue un rôle impératif dans la découverte de nouvelles entités chimiques puissantes. Dans cette étude, des modèles QSPR ont été construits pour la prédiction les valeurs de***pKa*** et ***Température de fusion*** pour une série de dérivés d'acide benzoïque et d'hydrocarbures respectivement.

Un nombre de descripteurs a été calculé pour chaque famille de composées ensuite limité à un sous ensemble de descripteurs les plus performant sur le développement des modèles à l'aide de ***WEKA***. Cette procédure a abouti à 4 descripteurs pour les dérivés de l'acide benzoïque(***Tableau 7***) et 3 pour les hydrocarbures (***Tableau 8***).

Ensuite, des modèles basés sur les méthodes statistiques : régression linéaire, perceptron multicouche et kNN ont été développés. Dans tous les cas, les modèle QSPR ont été construit sur un ensemble d'apprentissage et évalué par rapport à un ensemble de test. L'activité prédite a ensuite été comparée à la valeur réelle (expérimentale).

Les équations obtenues après la régression linéaire sont donnée dans les équation «(1) et (2)». Le coefficient de corrélation de 0,92 pour le modèle basé sur la MLR pour les dérivés de l'acide benzoïque et 0.98 pour les où on peut déduire que les descripteurs sélectionnés contribuent positivement à la prédiction de l'activité pka.

D'après les ***Tableaux*** [***10-21***], il est clair que le modèle k plus proche voisin (kNN) fonctionne moins que les deux autres techniques d'apprentissage en termes de corrélation, MAE et RMSE. La valeur R= 0.91 pour les acides benzoïques et R=0.87 pour les hydrocarbures spécifie que les variations de 91% de pka sont bien expliquées par les descripteurs pour les acides benzoïques ainsi que 87% de T_{fusion} sont bien expliquées par les descripteurs pour les hydrocarbures. RMSE est la mesure de la précision et est considérée comme idéale si elle est petite.

III.5 Conclusion

QSPR joue un rôle clé dans la découverte et le développement de nouveaux composés chimiques. Le choix d'une méthode de modélisation QSPR appropriée joue un rôle majeur dans la prédiction. L'utilisation de diverses techniques d'apprentissage automatique et l'utilisation des métriques de validation associées permet le développement de modèles QSPR hautement prédictifs. La création d'un modèle QSPR efficace serait très bénéfique.

Dans ce chapitre, l'activité PKa des acides benzoïques et température de fusion des hydrocarbures ont été modélisée par trois techniques d'apprentissage automatique différentes, à savoir la régression linéaire, le perceptron multicouche et k le plus proche voisin. D'après

les résultats, il était clair que le modèle de régression linéaire fonctionnait mieux que les deux autres modèles.

Par conséquent, le modèle QSPR basé sur la régression linéaire peut être considéré comme une approche prometteuse pour la prédiction des deux activités. En outre, le modèle doit être testé par rapport à un grand ensemble de données pour authentifier sa précision prédictive.

Conclusion Générale

Conclusion Générale

Les modèles quantitatifs de QSAR/QSPR présentent une solution statistique du problème de la difficulté du calcul direct des propriétés physiques et biologiques à partir de la structure.

L'intérêt d'un modèle de QSAR/QSPR est de tirer des informations à partir de l'ensemble des descripteurs numériques caractérisant la structure moléculaire et prédire ainsi les activités biologiques de nouvelles structures.

Dans ce travail, une étude quantitative de la relation structure-prpriété (QSPR) a été effectuée sur deux types de molécules : 25 molécules d'hydrocarbures et 25 dérivées d'acide benzoïques. Trois modèles QSPR ont été établis pour chaque famille de molécules en utilisant les méthodes de régression linéaire (LR), Le perceptron multi couche (MLP) et le K plus proche voisin. L'objectif de notre travail est de modéliser les activités pka des dérivées des acides benzoïques et la température de fusion des hydrocarbures pour former des modèles de QSAR robustes, stables, et précis capables de prédire efficacement ces activités.

Les modèles QSPR développés ont été obtenus avec quatre descripteurs (HOMO, **ΔE**, $\mathbf{M_{Dipole}}$,NCL) pour les dérivées des acides benzoïques qui ont une influence significative sur l'activité biologique Et trois descripteurs significatifs pour les hydrocarbures (**ΔE**, $\mathbf{M_{Dipole}}$, NC).

Le pouvoir prédictif des modèles obtenus a étéconfirmé par les résultats de la validation externe et interne des modèles. Une forte corrélation a été observée entre les valeurs expérimentales et prédites des activités biologiques pka et température de fusion, ce qui indique la validité et la qualité des modèles QSPR obtenus.

A travers les différents résultats obtenus au cours de ce travail, nous pouvons dire que les ensembles de molécules, les traitements statistiques et les techniques informatiques utilisés, lors du développement et l'analyse des modèles de QSPR, ont donné de bons résultats, ce qui nous permet d'entrevoir des perspectives assez prometteuses dans ce domaine par l'amélioration des traitements statistiques et par l'utilisation d'autres méthodes de calculs et de sélection des descripteurs.

Ce travail peut aussi être étendu à un nombre plus important de composés pour chaque famille chimique. Un prolongement évident dans la modélisation qualitative et précisément avec les méthodes de régression et de classification comme l'analyse par composantes principales ACP, machine à support vecteur SVM, ...etc.

Références Bibliographiques

Références Bibliographiques

[1] : Boumédiène bounaceur étude par spectometre et calculs quanto chimiques de la photo tranformation des cinnamatesdecholecteryles et de leurs dérivés halogénés. 2007.Thèse de doctorat.

[2] : MESSAADIA Lyamine. Chimie Quantique. 2015. Université Mohamed Seddik Ben Yahia Jijel.

[3]: Gérald Monard. Méthode Hartree-Fock. 2013. Université de Toulouse. Cours Modélisation Moléculaire.

[4] : BELLIFA, Khadidja. Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires.«Modélisation QSAR. 2015. Thèse de doctorat.

[5] : YOUSFI, YOUCEF. *ETUDE QSAR DE l'ACTIVITE ANTI-OXYDANTE D'UNE SERIE DE COMPOSES PHENOLIQUES*. 2018. Thèse de doctorat.

[6] : Charif Imad Eddine,élaboration des correlations quantitatives structure-activite des acides carbones.etude theorique des effets de solvants sur les equilibres ceto-enoliques des composes β-dicarbonyles cycliques et de leurs analogues a chaines ouvertes'. 2012. Thèse de doctorat.

[7] : Mehellou Mohammed Nadjib, Etude des relations structure/activité quantitatives (QSAR/2D) d'une série de dérivés de Triazolothiadiazoles. 2018.Mémoire de master.

[8] : BEGHOU Mahrez, Estimation statistique des températures de fusion pour quelques hydrocarbures aliphatiques. 2016. Mémoire de master.

[9] : Maddi Housny, Étude du mode de liaison et de la dynamique en solution de complexes binucléaires dissymétriques du pentalène. 2014. Mémoire de magister.

[10] : Mostefaoui, L.Contribution à la description et à la compréhension de la solvatation des biomolécules. 2011. Universite Abou-Bakr Belkaid de Tlemcen.

[11] : Aber achour ,Mekki Yakoub, Calcul théorique du moment dipolaire et polarisabilité pour des phosphazénes cycliques . 2013. Mémoire de licence.

[12] : kpenglame Kpassèmon. Situation des carburants au togo. 2017. Ministere Des Mines Et De L'energie.

[13] : Bourgeais.A. Alcanes. Cours de Chimie Organique. Ressource Nationales de Chimie.2009.

[14]: Cousins, Kimberley R. "Computer review of ChemDraw ultra 12.0." 2011: 8388-8388.

[15]: Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. 2016. WEKA manual for version 3-9-1. *University of Waikato, Hamilton, New Zealand*.

[16] : Cros AFA. Action de l'alcohol amylique sur l'organisme. 1863. Strasbourg, University of Strasbourg, Thesis.

[17]: Crum-Brown A, Fraser TR. On the connection between chemical constitution and physiological action. Pt 1. On the physio- logical action of the salts of the ammonium bases, derived from Strychnia, Brucia, The- bia, Codeia, Morphia, and Nicotia. T Roy Soc Edin 1868-1869; 25:151-203.

[18] : Richet MC. Note sur le rapport entre la toxicité et les propriétes physiques des corps. Compt Rend Soc Biol (Paris) 1893;45:775-6.

[19]: Meyer H. Zur Theorie der AIkoholnarkose. Arch Exp Path Pharm 1899;42:109-18.

[20] : Overton CE. Studien über die Narkose. Je- na: Fischer, 1901.

[21]: Hammett LP . Some relations between reaction rates and equilibrium constants. Chem Rev 1935;17:125-36.

[22]: Hammett LP. The effect of structure upon the reactions of organic compounds. Ben- zene derivatives. J Am Chem Soc 1937;59: 96-103.

[23]: Taft RW. Separation of polar, steric and resonance effects in reactivity.1956. In: Newman MS (ed.): Steric effects in organic chemis- try (pp 556-675). New York: Wiley.

[24]: Hansch C, Fujita T. *p-σ-π* analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 1964;86:1616-26.

[25]: Frisch MJ, Trucks GW, Schlegel HB, et al. Gaussian 03W, Revision C.02. Walling-ford: Gaussian Inc., 2004. .

[26]: Izrailev, S.; Agrafiotis, D. A Novel Method for Building Regression Tree Models for QSAR Based on Artificial Ant Colony Systems. J. Chem. Inf. Comput. Sci. 2001,41, 176–180.

[27]: Sutter, J.; Kalivas, J. Comparison of Forward Selection, Backward Elimination and Generalized Simulated Annealing for Variable Selection. Microchemical J. 1993, 47, 60–66.

[28]: Goldberg, D. Genetic Algorithms in Search, Optimization & Machine Learning; Addison-Wesley: Reading, MA, 2000.

[29] : Benazzouz,H.; Khebiza,A. "Relation Structure Activité : Etude Qualitative et Quantitative et Développement de Recherche sur les Coumarines". Thèse de doctorat. 2018.

[30]: Miller, A. and Miller, E. C. (1977) Ultimate chemical carcinogen as reactive mutagenic electorophiles. In Hiatt, H. H., Watson, J. D. and Winsten, J. A. (eds.), *Origin of Human Cancer*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 605–627.

[31]: Ashby, J. and Tennant, R. W. (1988) Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic car- cinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat. Res.*, 204, 17–115.

[32]: Honma, Masamitsu, et al. "Improvement of quantitative structure–activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project." *Mutagenesis* 34.1. 2019: 3-16.

[33] : Fortuné, Antoine. *Techniques de Modélisation Moléculaire appliquées à l'Etude et à l'Optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance*. Diss. 2006.

[34]: Schneider A, Hommel G, Blettner M: Linear regression analysis—part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(44): 776–82. DOI: 10.3238/arztebl.2010.0776

[35] : Haccoun, A. Comparaison de méthodes de classifications. From. 2012. https://www.lri.fr/~antoine/Courses/Master-ISI/ISI-10/Projets_2012/Projet_DM.pdf

[36] : Izabela Moise, Evangelos Pournaras, Dirk Helbing, K- Nearest Neighbour Classifier. 2015. ETH Zurich.

[37] : AIT MAHAMMED, Fatima. Approches d'apprentissage automatique pour la détection du Spam Web: exploration de diverses caractéristiques. 2018.

[38] : KADARI R. Introduction à l'intélligence artificielle. Cours Master 2 Chimie Théorique et Computationnelle. 2019.

[39]: Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., & Prachayasittikul, V. (2009). A practical overview of quantitative structure-activity relationship.

[40]: Vastrad, Chanabasayya. "Performance analysis of neural network models for oxazolines and oxazoles derivatives descriptor dataset." *arXiv preprint arXiv:1312.2853* (2013).

[41]: Pradeep, Prachi.: Quantitative Structure Activity Relationships: An overview. The United States Environmental Protection Agency's Center for Computational Toxicology and Exposure. (2018) Presentation. https://doi.org/10.23645/epacomptox.6856775.v1

[42]: David R. Lide, CRC Handbook of Chemistry and Physics 90th Edition, (2010), Edition CRC Press.

[43]: Tien Dat pham , et al."Estimating aboveground biomass of a mangrove plantation on the northern coast of Vietnam using machine learning techniques with an integration of alos-2 palsar-2 and sentinel-2a data".2018.