

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة مولاي الطاهر، سعيدة

Université MOULAY Tahar, Saida



كلية العلوم

Faculté des Sciences

قسم الكيمياء

Département de chimie

Mémoire pour l'obtention du diplôme de Master

En Sciences de la matière

Spécialité : Chimie Théorique et Computationnelle

Étude QSPR basée sur l'apprentissage automatique pour la prédiction des pKa des phénols dans les solvants organiques à l'aide de descripteurs quantiques

Présenté par :

Mlle : LAZRAGUI HADIL IMENE

Soutenu le : 20 – 06 – 2023

Devant le jury composé de :

Présidente

Dr. Djalila MISSAOUI

MCB

Université de Saida

Examinateur

Pr. Ali RAHMOUNI

Professeur

Université de Saida

Encadrante

Dr. Rekia KADARI

MCB

Université de Saida

Année universitaire 2022/2023



REMERCIEMENT

Je tiens à exprimer ma profonde reconnaissance et mes sincères remerciements à toutes les personnes qui ont joué un rôle essentiel dans mon parcours académique et professionnel.

Tout d'abord, je voudrais rendre hommage à mes parents, en particulier à mon père, qui, malheureusement, n'a pas pu être présent pour assister à ce jour de soutenance. Son rêve était de me voir réussir, et je suis déterminée à honorer sa mémoire en poursuivant mes objectifs avec détermination et persévérance.

Je souhaite également exprimer ma gratitude envers mes professeurs, en particulier mon encadrante, **Rekia Kadari**, pour son soutien inestimable tout au long de ce projet. Ses conseils éclairés et sa patience ont été d'une grande valeur pour moi.

Je tiens également à remercier le professeur **Rahmouni Ali** pour sa présence en tant que membre du jury et pour ses précieux commentaires et suggestions lors de ma soutenance. La présence de la professeure **Missaoui Djalila** en tant que présidente du jury a également été très appréciée.

Un grand merci à tous mes chers collègues qui m'ont soutenue et encouragée tout au long de ce parcours. Vos mots d'encouragement et votre soutien ont été une source de motivation constante.

Enfin, je souhaite exprimer ma gratitude envers toutes les personnes qui ont pris le temps de me féliciter et de m'encourager. Vos paroles aimables et vos encouragements ont renforcé ma détermination à réussir.

TABLE DE MATIÈRE

Introduction générale	(01)
------------------------------	-------------

CHAPITRE I: Etude bibliographique

I.1 Les phenols	
I.1.1 Introduction	(04)
I.1.2 Généralités sur les phénols	(04)
I.1.2.1 Historique	(04)
I.1.2.2 Définition.	(04)
I.1.2.3 Classification des phénols	(05)
I.1.2.4 Propriétés du phénol	(07)
I.2 L'acidité	
I.2.1 Introduction	(09)
I.2.2 Généralité sur les acides et les bases	(09)
I.2.3 La constante d'acidité	(10)
I.2.4 L'équation de Hammett pour calculer le pka	(12)
I.2.5 Les solvants	(13)
I.2.5.1 Le diméthylesulfoxyde (DMSO)	(14)
I.2.5.2 Le méthanol	(14)
I.2.5.3 Le N,N-diméthylformamide (DMF)	(15)

CHAPITRE II : Relations Quantitatives Structure-Propriétés

II.1 Introduction	(21)
--------------------------	-------------

II.2.	Définition de la modélisation QSPR	(21)
II.3	Historique de la modélisation QSPR	(21)
II.4	Principe	(22)
II.5	Étapes clés du développement des modèles QSPR	(22)
II.5.1	Collecte des données	(23)
II.5.2	Génération des descripteurs	(24)
II.5.2.1	Les descripteurs moléculaires	(24)
II.5.3	Sélection des descripteurs	(25)
II.5.4	Développement du modèle QSPR	(26)
II.5.4.1	Les méthodes statistiques	(26)
II.6	Evaluation du modèle QSPR	(30)
II.7	Les métriques d'évaluation	(31)
II.8	Application du QSPR	(33)
II.9	Conclusion	(34)

CHAPITRE III : Résultats et discussions

III.1	Introduction	(38)
III.2	Calcul et Sélection des descripteurs	(38)
III.3	Développement du modèle QSPR	(51)
III.4	Les résultats des méthodes statistiques développés	(52)
III.5	Discussion	(57)
III.6	Conclusion	(58)
	Conclusion Générale	(60)

LISTE DES FIGURES

Figure I.1 Structure moléculaire du phénol	(05)
Figure I.2 Effet oxydant de l'air sur le phénol	(08)
Figure I.3 Stabilité de résonance de l'ion phénolate	(08)
Figure I.4 Molécule de dimethylsulfoxyde	(14)
Figure I.5 Molécule de méthanol	(15)
Figure I.6 Molécule de DMF	(15)
Figure II.1 Les étapes de base de construction du modèle de QSPR	(23)
Figure II.2 Représentation d'un réseau de neurone	(29)
Figure III.1 Exemple de la représentation moléculaire sur Gaussview	(47)

LISTE DES TABLEAUX

Tableau I.1 Certaines propriétés du phénol	(07)
Tableau I.2 Certaines propriétés du DMSO	(14)
Tableau I.3 Certaines propriétés du méthanol	(15)
Tableau I.4 Certaines propriétés du DMF	(16)
Tableau III.1 Les molécules de DMSO	(39)
Tableau III.2 Les molécules de Me-OH	(44)
Tableau III.3 Les molécules deDMF	(46)
Tableau III.4 Résultats de calcul des descripteurs	(48)
Tableau III.5 Liste des descripteurs obtenus après le processus de sélection sous Weka	(51)
Tableau III.6 La répartition des données	(51)
Tableau III.7 Performance de la régression linéaire multiple sur les données d'apprentissage et les données de test	(52)
Tableau III.8 Performance de la regression linéaire multiples sur les données globales	(53)
Tableau III.9 Comparaison des données expérimentales avec les données prédites par le modèle basé sur la régression linéaire.	(54)
Tableau III.10 Performance de perceptron multicouche sur les données d'apprentissage et les données de test	(55)
Tableau III.11 Performance de perceptron multicouches sur les données globales	(55)
Tableau III.12 Comparaison des données expérimentales avec les données prédites par le modèle basé sur le perceptron multicouches	(57)

LISTE DES ABRÉVIATIONS

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 Corrélacion
R²	Coefficient de Détermination.
HOMO	Orbitale moléculaire occupée la plus élevée
LUMO	Orbite moléculaire inoccupée la plus basse
DFT	Théorie de la Fonctionnelle de la Densité
B3LYP	Becke, trois paramètres, Lee-Yang-Parr
K_a	La constante d'équilibre.
pK_a	La constante d'acidité.
DMSO	Le diméthylsulfoxyde
DMF	N,N-diméthylformamide.
Me-OH	méthanol .

Abstract

Multiple linear regression and Artificial neural networks based QSPR methodologies are applied to the prediction of pKa of phenols in different solvents. Our models are composed of 62 phenols and 3 solvents. The proposed models contains seven descriptors; six of them belong to the solutes and the other one to the solvents. The obtained models with 3 three selected descriptors show good predictive ability. We obtained a coefficient of determination R^2 of 0.9455 with the MLR model, while the MLP gains R^2 of 0.9228. The validation results through the test set indicate that the proposed models are robust and satisfactory.

Keywords: QSPR, pka, phenol, MLR, prediction.

Résumé

Ce Travail représente l'application de la régression linéaire multiple et des réseaux de neurones artificiels pour la prédiction pKa des phénols dans différents solvants. Notre ensemble de données comprend 62 phénols et 3 solvants. Nos modèles comprennent sept descripteurs, dont six appartiennent aux solutés et un est lié aux solvants. Les trois descripteurs sélectionnés présentent une bonne capacité prédictive. Nous avons obtenu un coefficient de détermination (R^2) de 0,9455 avec le modèle de régression linéaire multiple sur l'ensemble d'apprentissage, tandis que le perceptron multicouches obtient un R^2 de 0,9228. Les résultats de validation sur l'ensemble de test indiquent que les modèles proposés sont robustes et satisfaisants.

Mots-clés : QSPR, pKa, phénol, régression linéaire multiple, prédiction.

ملخص

في هذا البحث تم تطبيق منهجيات QSPR القائمة على الانحدار الخطي المتعدد والشبكات العصبية الاصطناعية للتنبؤ بثابت تفكك الحمض للفينولات في مذيبات مختلفة حيث يتكون النظام من 62 فينول و 3 مذيبات. النماذج المقترحة تتكون من 7 واصفات، 6 منها تنتمي إلى المذاب وواحدة إلى المذيب، لها قدرة تنبؤية جيدة. يضبط نموذج الانحدار الخطي المتعدد مجموعة التعلم بمعامل تحديد R^2 قدره 0.9455، بينما يحصل نموذج الشبكة العصبية الاصطناعية على قيمة R^2 قدرها 0.9228 حيث تشير نتائج التحقق من مجموعة الاختبارات إلى أن النماذج المقترحة قوية ومرضية.

الكلمات الرئيسية: QSPR, pKa, الفينول، الانحدار الخطي المتعدد، التنبؤ

Introduction Générale

Introduction Générale

Dans ce travail, nous nous sommes concentrés sur le développement de modèles QSPR pour prédire les valeurs de pKa de 62 molécules phénoliques. L'objectif principal de notre étude était de comprendre les relations complexes entre la structure moléculaire et les valeurs de pKa des phénols, et de développer des modèles prédictif fiable.

Pour atteindre cet objectif, nous avons utilisé une approche intégrant des descripteurs moléculaires quantiques, qui capturent les informations structurales et électroniques des molécules, ainsi que des techniques d'apprentissage automatique notamment la régression linéaire multiple et les perceptrons multicouches. Ces méthodes nous ont permis de construire des modèles capable de prédire le pKa des molécules phénoliques étudiées.

Les résultats obtenus dans cette étude contribuent à l'amélioration de la prédiction du pKa des phénols, ce qui peut avoir un impact significatif dans divers domaines de la chimie, de la pharmacologie et de la recherche biologique. De plus, notre approche méthodologique peut servir de base pour développer des modèles QSPR pour d'autres types de molécules et de propriétés, ouvrant ainsi la voie à de nouvelles applications dans le domaine de la chimie prédictive.

Le présent travail est divisé en trois chapitres :

-Le premier chapitre se divise en deux parties distinctes. La première partie couvre les notions générales concernant les phénols, tandis que la deuxième partie offre une revue de la littérature axée sur l'acidité.

-Dans le deuxième chapitre, nous présentons en détail les modèles QSPR, en exposant les différentes étapes de leur développement. Nous décrivons les méthodes et les techniques statistiques utilisées dans ces études, en mettant l'accent sur leur application dans le domaine de la prédiction des propriétés moléculaires.

-Dans le troisième chapitre, nous présentons nos résultats obtenus lors de cette étude. Nous exposons les modèles de prédiction du pKa que nous avons développés en utilisant la régression linéaire et le perceptron multicouche. Nous discutons également de l'efficacité et de la précision de ces modèles, en mettant en évidence les performances obtenues pour prédire les valeurs de pKa des phénols étudiés.

CHAPITRE I

Etude Bibliographique

I.1 Les phenols

I.1.1 Introduction

Le phénol est un composé organique aromatique soluble dans l'eau qui possède un groupement hydroxyle (OH) lié à un atome de carbone du cycle benzénique. Il est utilisé dans le commerce comme intermédiaires des colorants et des processus de synthèse organique. Ainsi que comme antiseptiques et désinfectants [1]. Il est largement utilisé en photographie, dans les industries du pétrole, des peintures, des explosifs, du caoutchouc, des matières plastiques et dans les industries pharmaceutiques et agroalimentaire.

I.1.2 Généralité sur les phenols

I.1.2.1 Historique

En 1650, le scientifique allemand Johann Rudolf Glauberat découvre le phénol dans l'état impure de la distillation du goudron de huile. Il l'appelait "l'huile vivante et Un rouge sang qui sèche et guérit tous les ulcères humides". Environ deux siècles après, en 1834 son concitoyen Friedrich Ferdinand Runge parvint à isoler pour la première fois le phénol et il le nomme "acide carbolique". Ensuite, En 1841 Auguste Laurent, un chimiste français fut le premier à préparer le phénol pur. Il le nomme acide phénolique. En 1843, le chimiste français, Charles Frédéric Gerhard invente le nom de «phénol», nom qui a peu à peu détrôné celui d'acide phénique. Le mot « phénol » est tout à fait conforme à la nomenclature officielle [2].

Le phénol a été produit, durant la première guerre mondiale, pour des applications militaires. Il est utilisé dans la synthèse des résines, plus tard il est devenu un objet de nombreuses études et plusieurs applications. Il intervient dans plusieurs procédés de fabrication de divers composés: sous produit de raffinage de pétrole, produits pharmaceutiques, colorants,...[3].

I.1.2.2 Définition

Le phénol est un composé organique aromatique composé de noyau benzénique relié avec un groupement hydroxyle (-OH), le phénol a des propriétés uniques et n'est pas classé comme un alcool. Sa structure est relativement simple, le groupement hydroxyle est lié à un atome de carbone du cycle benzénique (figure I-1). Il est présent dans de nombreux

végétaux. Dans les laboratoires, il est couramment utilisé comme un produit chimique pour diverses activités. est un produit de synthèse Pur, il se présente à la température ordinaire comme un solide blanc cristallisé. C'est un composé toxique Il peut être trouvé directement en tant que réactif commercial ou bien préparé dans le cadre d'un mélange dans le laboratoire. A l'heure actuelle le phénol est généralement préparé par le procédé Hock qui consiste à oxyder l'isopropylbenzène par le dioxygène de l'air. Le sous-produit de la réaction est la propanone qui est également un produit important utilisé notamment comme solvant. Par la suite, le phénol est séparé de l'acétone [4].

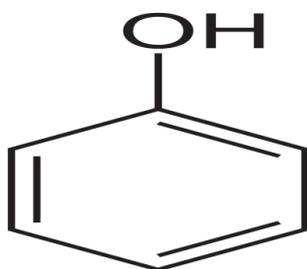


Figure I.1 Structure moléculaire du phénol.

I.1.2.3 Classification des phénols

Le terme composé phénolique est un ou plusieurs cycles aromatiques substitués par un ou plusieurs groupes hydroxyle[4]. Cette définition introduit une dizaine de familles de phénols.

I.1.2.3.1 Les phénols simples

Les phénols simples englobent toutes les molécules hydroxylées diversement substituées de noyau du phénol. Selon cette définition plusieurs phénols appartiennent à ce groupe, par exemple : le phénol, les chlorophénols, les alkylphénols, les acides hydroxybenzoïque, les nitrophénols... Dans la prochaine section nous expliquerons un exemple de ce type de phénols tels que les chlorophénols.

I.1.2.3.2 Les chlorophénols

Les chlorophénols sont des composés organiques dans lequel le noyau phénolique est substitué en différentes manières par un ou plusieurs atomes de chlore remplaçant l'hydrogène du cycle aromatique. Par conséquent nous comptons 19 congénères des

chlorophénols qui sont: (2 ; 3 ; 4)-chlorophénol, 2,(3 ; 4 ; 5 ; 6)-dichlorophénol, 3,(4 ; 5)-dichlorophénol, 2,3,(4 ; 5 ; 6)-trichlorophénol, 2,4(5 ; 6)-trichlorophénol, 3,4,5-trichlorophénol, 2,3,4 (5 ; 6)-tétrachlorophénol, 2,3,5,6-tétrachlorophénol et pentachlorophénol [5].

Les chlorophénols servent pour produire des intermédiaires dans la synthèse des phénols plus chlorés ou de dérivés des chlorophénols comme les herbicides, les colorants, les pigments et les résines phénoliques. Dans les stations d'épuration, les phénols peuvent se retrouver dans les eaux de surface destinées à la préparation d'eau potable. Après désinfection de cette eau par le chlore ou ses dérivés, les composés phénoliques peuvent générer des chlorophénols. Aussi, certains chlorophénols sont utilisés comme fongicides, antiseptiques, désinfectants et agents antigommage pour l'essence. D'une autre part, comme la plupart des composés phénoliques, les chlorophénols sont toxiques pour les milieux aquatiques. A certains seuils ils sont capables d'entraîner des changements histopathologiques mutagènes et cancérigènes [6].

I.1.2.3.3 Les polyphénols

Chimiquement, les polyphénols sont des composés phénoliques à haut poids moléculaires. Ils se composent d'un ou plusieurs cycles benzéniques portant un ou plusieurs groupements hydroxyles et autres constituants [7]. Les polyphénols regroupent un vaste ensemble de substances chimiques. Les polyphénols naturels sont des familles de substances chimiques largement distribués dans les végétaux. Ils ne sont pas directement impliqués dans un processus métabolique quelconque et sont donc considérés comme des métabolites secondaires. Ils sont présents dans diverses substances naturelles : sous forme d'anthocyanine dans les fruits rouges, de proanthocyanidines dans le chocolat, d'acides cafeoylquinique et feruloylquinique dans le café, de flavonoïdes dans les agrumes, de catéchines comme l'épigallo-catéchine gallate du thé vert, de quercitrine dans les pommes, ...etc [8]. Le polyphénol se présente comme un composé de base, constituant plusieurs produits synthétiques simples possédant une fonction alcool supplémentaire comme l'hydroquinone, catéchol et sous forme de polyphénols polymériques comme les colorants, les plastiques et les résines à base de bisphénol.

Au-delà de leur implication reconnue dans « l'effet-santé » de l'aliment, les polyphénols naturels participent aussi pleinement aux caractéristiques organoleptiques des denrées et jouent un rôle important sur la durée de vie des aliments. Ils peuvent générer des problèmes

de couleurs ou d'instabilité colloïdale au cours du temps. Concernant les polyphénols simples, dans une perspective écologique, le pyrocatechol et l'hydroquinone sont dangereux pour les écosystèmes et notamment pour les eaux car ils sont peu biodégradables et en partie toxiques pour les poissons. Le contact cutané avec le catéchol ou l'hydroquinone provoque une dermatite eczémateuse chez l'homme. Une respiration de fortes doses de catéchol peut entraîner une dépression du système nerveux et l'ingestion répétée de l'hydroquinone à faible quantité n'entraîne qu'une affection oculaire. Dans le cas de polyphénols polymériques, par exemple le bisphénol nuit beaucoup au système hormonal de l'organisme. Cette molécule peut, en effet, agir en tant qu'analogue d'œstrogène dans les systèmes biologiques. Les œstrogènes sont des molécules organiques dérivées du cholestérol. Chez la femme, ils jouent un rôle important dans la croissance et la fonction de reproduction, mais aussi dans la féminisation. Chez l'homme, ils sont présents à plus faibles doses et jouent un rôle dans la croissance et le développement [9,10].

I.1.2.4 Propriétés du phénol

I.1.2.4.1 Propriétés physicochimiques

Le phénol est un solide incolore cristallisé sous forme d'aiguilles dans les conditions ambiantes habituelles, sa formule chimique est la suivante : C_6H_6O .

Masse molaire (g/mol)	94,11
Densité liquide	1,049
Densité solide	1,071
pKa à 25°C	9.92
Point d'ébullition (°C)	182
Point de fusion (°C)	41
Tension de vapeur à 25°C (mmHg)	0,350

Tableau I.1 Certaines propriétés du phénol.

En plus, le phénol est hygroscopique et a une odeur âcre et douceâtre. Au contact de l'air ou sous l'influence de l'humidité le phénol s'oxyde légèrement pour donner des traces de quinone. Il prend alors une couleur rose, puis rouge. Il est miscible dans l'eau. Sa

solubilité est limitée entre 80- 100g/L à 25°C. Il est aussi très soluble dans plusieurs solvants organiques tels que l'acétone, l'éthanol, l'oxyde de diéthyle.

I.1.2.4.2 Propriétés oxydantes

L'oxydation du phénol peut avoir lieu sous O₂. Elle conduit à la formation de radicauxphénoxyles qui évoluent pour donner par couplage des produits complexes souvent colorés, dont la structure est mal définie (voir figure I.2). C'est la raison pour laquelle les récipients contenant du phénol doivent être soigneusement conservés à l'abri de l'air.

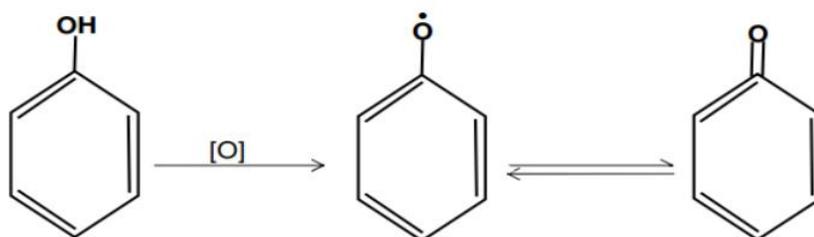


Figure I.2 Effet oxydant de l'air sur le phénol

Parfois, le phénol attaque certains métaux contenus en solution comme le plomb, le zinc, l'aluminium et aussi certains plastiques comme le polyéthylène.

I.1.2.4.3 Propriétés acido-basiques

En solution, le phénol forme un acide très faible. Par contre son acidité est plus forte que ceux des alcools (pK_a à 25°C du couple phénol/phénolate est de l'ordre de 9,9). Le phénol peut perdre un ion hydrogène et l'ion phénolate (phénoxyde) se stabilise dans la solution. En effet, lors de la prise du proton H⁺ du groupement hydroxyle (-OH), la charge négative est ainsi délocalisée autour du cycle en en position ortho ou para selon la stabilité de résonance de l'ion phénolate comme le présente la figure-3. L'une des paires d'électrons libres sur l'atome d'oxygène s'interfère avec les électrons délocalisés sur le noyau benzénique [11].

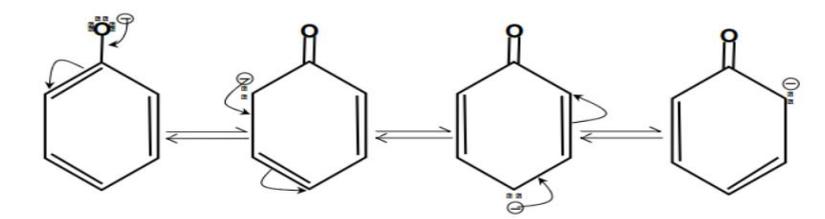


Figure I.1.3 Stabilité de résonance de l'ion phénolate.

I.2 L'acidité

I.2.1 Introduction

L'acidité joue un rôle essentiel dans de nombreuses réactions chimiques, notamment les réactions acido-basiques, les réactions d'oxydoréduction et les réactions de complexation. En chimie organique, l'acidité d'une molécule peut affecter sa réactivité et être utilisée pour contrôler l'issue des réactions chimiques. En biochimie, l'acidité des protéines et d'autres biomolécules peut affecter leur structure et leur fonction. En chimie environnementale, l'acidité de l'eau et du sol peut avoir des effets importants sur les écosystèmes et conduire à la pollution et à d'autres problèmes environnementaux[12].

Dans l'ensemble, le concept d'acidité est essentiel à notre compréhension des réactions chimiques et du comportement des molécules dans un large éventail de contextes. En comprenant les propriétés et le comportement des acides, nous pouvons mieux prédire et contrôler l'issue des réactions chimiques, concevoir de nouveaux matériaux et médicaments, et aborder d'importantes questions environnementales [13].

I.2.2 Généralité sur les acides et les bases

Les acides et les bases sont des produits chimiques importants et largement utilisés dans l'industrie. Ils peuvent être sous forme ionique ou non, solide ou liquide. Ces produits sont connus depuis l'Antiquité et leur histoire à la suite de l'évolution des connaissances et des théories chimiques[14]. Au cours des années allant de la fin du XIX^e siècle aux années 1960, quatre approches théoriques ont été développées pour comprendre les acides et les bases. La théorie la plus couramment utilisée en chimie des solutions est celle de Brønsted-Lowry, datant de 1923, bien qu'il existe d'autres approches, telles que celles proposées par Arrhenius et Lewis.

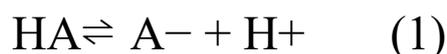
Depuis 1887, l'idée d'un acide est connue une évolution. Arrhenius a proposé que certains composés se dissocient en ions dans une solution [15] et a défini un acide comme un composé qui libère des ions hydrogène (H^+) en solution et une base comme un composé qui libère des ions hydroxyle (OH^-)[16]. Cette définition permet d'expliquer les propriétés des acides et des bases dans l'eau. Les travaux d'Arrhenius ont été très importants pour le développement de la chimie. Cependant, cette théorie est maintenant considérée comme insuffisante car elle ne s'applique pas à tous les solvants. L'ion H^+ , qui est le noyau de

l'atome d'hydrogène, une forte tendance à se lier à une molécule de solvant en raison de sa petite taille (cent mille fois plus petite que celle de la molécule H₂O)[17].

En 1923, deux scientifiques, Johannes Bronsted, un Danois, et Thomas Lowry, un chimiste-physicien anglais, ont proposé de manière indépendante que l'acide soit considéré comme un composé qui donne un proton (H⁺) et la base comme un composé qui accepte un proton, mettant ainsi l'accent sur les processus de libération et d'acceptation des protons. Cette théorie a été affinée et généralisée au fil du temps, grâce à un quart de siècle de travaux dans ce domaine.

Le modèle Brønsted-Lowry identifie les acides contenant de l'hydrogène (comme le HCl, HNO₃). Par conséquent, en l'absence d'hydrogène, plusieurs composés couramment appelés comme des acides, tels que SO₃ ou BCl₃, sont exclus de cette catégorisation. Ce modèle présente l'avantage de pouvoir être appliqué non seulement dans des milieux aqueux, mais également dans des solvants non aqueux et en phase gazeuse, ce qui le rend plus universel que le modèle d'Arrhenius [18]. Il fonctionne très bien dans tous les solvants polaires tels que l'eau, l'ammoniac, l'acide acétique, etc. Cependant, il ne peut pas expliquer le comportement acido-basique des solvants apolaires tels que le benzène et l'acétone.

La dissociation d'un acide de Brønsted donne un proton et la base conjuguée (un anion, si l'acide est un composé neutre) de l'acide:



$$ka = \frac{[A^-][H^+]}{[HA]} \quad (2)$$

Ou:

A⁻ est la base conjuguée d'AH. Inversement,

AH est l'acide conjugué de la base

A⁻. L'acide et sa base conjuguée forment un couple acido-basique AH/ A⁻.

K_a est une constante d'ionisation de l'acide, exprimée par les concentrations des réactifs et des produits, elle ne dépend que de la température. Le proton H⁺ ne se contente pas de se dissocier et de flotter en solution.

En 1923, une nouvelle définition de l'acide a été proposée par le chimiste-physicien américain GN Lewis. Selon sa théorie, un acide est un accepteur de doublet d'électrons, tandis qu'une base est un donneur de doublet. Cette définition est beaucoup plus générale que celle d'Arrhenius ou de Brønsted-Lowry, car elle n'est pas limitée aux composés hydrogénés. Selon cette théorie, tout ion positif est considéré comme un acide de Lewis, et les réactions de formation de complexes sont provoquées comme des réactions acide-base[19].

Les acides et les bases peuvent être classés en deux groupes en fonction de leur degré de dissociation : le premier groupe correspond aux acides et aux bases fortes, tandis que le deuxième groupe comprend les acides et les bases faibles. Les acides et les bases forts sont des électrolytes forts, avec un équilibre de dissociation nettement en faveur de la réaction directe. Les acides forts sont des espèces chimiques qui s'ionisent complètement dans l'eau pour donner des ions hydronium (H₃O⁺) et les bases fortes sont des espèces chimiques qui s'ionisent complètement dans l'eau pour donner des ions hydroxyde (OH⁻). En revanche, les acides et les bases faibles du deuxième groupe sont faiblement dissociés, avec un équilibre de dissociation nettement en faveur de la réaction inverse[20].

1.2.3 La constante d'acidité

La force d'un acide peut être activée par des facteurs locaux ou distaux, ce qui rend les termes "acide fort" et "acide faible" relatifs. Les facteurs locaux sont internes à la structure de la molécule d'acide, tels que la polarisation de la liaison AH résultant de l'effet inductif. La dissociation d'un acide de Brønsted forme une base conjuguée (anion A⁻), donc la stabilisation de cette dernière est importante pour classer les acides. Pour quantifier la force d'un acide, on peut utiliser les valeurs de pK_a plutôt que K_a[21].

Le pK_a est une mesure de l'acidité ou de la basicité d'une solution. ou constante d'acidité, mesure à quel point un proton est retenu par l'acide de Brønsted Il est défini comme le logarithme négatif (base 10) de la constante de la dissociation (K_a) d'un acide, il est exprimé par la relation suivante :

$$\text{pKa} = -\log (\text{Ka})$$

La valeur de pKa d'un composé est liée à sa structure. Elle peut être soit un petit nombre négatif, soit un nombre positif plus grand. Lorsqu'un acide de Brønsted a un pKa bas, cela signifie que le proton n'est pas fortement lié à l'acide et qu'il est donc facilement abandonné. Cela conduit à une dissociation complète de l'acide dans le solvant, et inversement.

A cause de signe moins de la définition, plus le pKa est petit, plus la constante d'équilibre est grande et plus l'acide est fort. Vous pouvez juger que la façon dont nous avons présenté le pKa est un concept plus utile pour visualiser le pKa : tout acide est à moitié dissocié dans une solution dont le PH est égale au pKa de l'acide. Aux pH supérieurs au pKa, l'acide est largement sous forme de sa base conjuguée (A⁻), mais aux pH inférieurs au pKa, l'acide est surtout sous la forme HA.

Le concept de pKa jou un rôle important dans de nombreux domaines de la chimie, notamment la biochimie, les produits pharmaceutiques et les sciences de l'environnement. Il est utilisé pour prédire le comportement des acides et des bases en solution et pour déterminer le pH d'une solution.

I.2.4 L'équation de Hammett pour calculer le pka

L'équation de Hammett est souvent utilisée pour corrélérer la réactivité chimique avec les substituants sur un composé organique, y compris les phénols. En utilisant cette équation, il est possible de prédire le pKa d'un phénol en fonction des effets inductifs et mésomères des substituants. L'équation de Hammett pour les acides de Brønsted (y compris les phénols) est la suivante :

$$\text{pKa} = \text{pKa}_0 + \sigma(\sigma_m + \sigma_p)$$

où pKa₀ est le pKa du composé de référence (généralement le phénol lui-même), σ est la constante de Hammett pour le substituant, et σ_m et σ_p sont les constantes de Hammett pour les effets mésomères et inductifs du substituant, respectivement[22].

Pour calculer le pKa d'un phénol en utilisant l'équation de Hammett, il faut tout d'abord déterminer les constantes de Hammett pour les substituants. Ces constantes peuvent être trouvées dans la littérature ou calculées à partir de données expérimentales ou

théoriques. Une fois que les constantes de Hammett pour les substituants ont été déterminées, il suffit de substituer ces valeurs dans l'équation de Hammett pour calculer le pKa du phénol correspondant. Il est important de noter que cette méthode de calcul du pKa ne prend pas en compte les effets de solvant ou d'autres facteurs qui peuvent influencer la valeur réelle du pKa d'un phénol. des solutions à une température et une pression spécifiques, et peuvent ne pas refléter avec précision le comportement d'une molécule dans un environnement biologique complexe.

En outre, les valeurs pKa ne tiennent pas compte d'autres facteurs susceptibles d'affecter l'acidité ou la basicité d'une solution, tels que la présence d'autres ions ou molécules. Il est donc important d'utiliser les valeurs de pKa en conjonction avec d'autres techniques d'analyse pour une compréhension complète d'un système.

I.2.5 Les solvants

Le solvant est un terme utilisé en chimie pour décrire un solvant capable de dissoudre un soluté particulier. Le solvant est généralement un liquide, mais il peut également s'agir d'un gaz ou d'un solide. La solvation décrit l'interaction entre le solvant et les molécules dissoutes, et la force et la nature de cette interaction ont un impact sur de nombreuses propriétés du soluté, notamment la solubilité et la réactivité. De plus, les propriétés physicochimiques du solvant, telles que la densité, la tension de vapeur, le point d'ébullition, la tension superficielle et surtout la constante diélectrique, sont également favorisées par cette interaction[23].

La constante diélectrique est une propriété physique importante d'un solvant en solution. Elle mesure la capacité d'un solvant à isoler les charges électriques et à réduire l'effet des forces électrostatiques entre les molécules dissoutes. Une constante diélectrique élevée indique que le solvant peut mieux stabiliser les charges électriques, tandis qu'une constante diélectrique faible indique que le solvant a moins d'effet sur les charges électriques et peut permettre une réactivité accrue entre les molécules dissoutes. La constante diélectrique est donc un facteur important dans de nombreux processus chimiques, tels que les réactions ioniques et les interactions électrostatiques.

I.2.5.1 Le diméthylesulfoxyde

Le diméthylesulfoxyde (DMSO) est un solvant organique polaire et aprotique, c'est-à-dire qu'il ne contient pas d'ions H^+ libres il est largement utilisé dans les laboratoires de recherche pour dissoudre des composés organiques et inorganiques. Il se présente sous la forme d'un liquide incolore, inodore et hygroscopique avec une formule chimique $(CH_3)_2SO$. Le DMSO a une polarité élevée en raison de la présence d'un atome de soufre oxydé et de deux groupes méthyles non polaires, ce qui lui permet de dissoudre une grande variété de composés polaires et apolaires. En raison de ses propriétés solvatantes uniques, le DMSO est couramment utilisé comme solvant dans les réactions chimiques, la chromatographie, la spectrométrie de masse, la cryoconservation et la thérapie médicale[24].

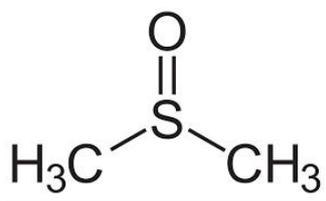


Figure I.4 Molécule de diméthylesulfoxyde.

I.2.5.1.1 Les principales constantes physiques du DMSO

La masse molaire	78,133
Point de fusion	18,45°C
Point d'ébullition (20°C, 1Atm)	189°C
Tension de vapeur saturante	0,417 mm Hg
densité	1,100
Constante Diélectrique	45
Moment Dipolaire	4,3D

Tableau I.2 Certaines propriétés du DMSO.

I.2.5.2 Le méthanol

Le méthanol (ou alcool méthylique, carbinol, alcool de bois, naphte de bois ou esprit de bois) est un composé organique de formule : CH_3OH (souvent abrégé en $MeOH$). C'est le plus simple des alcools. Il se présente sous la forme d'un liquid

léger, volatil, incolore, inflammable, toxique ayant une odeur caractéristique, plus douce et sucrée que celle de l'éthanol (alcool éthylique)[25].

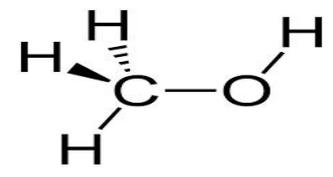


Figure I.5 Molécule de méthanol.

La masse molaire	32,04
Point de fusion	-97,8°C
Point d'ébullition (20°C, 1Atm)	64,5°C
Tension de vapeur saturante	12,3 mm Hg
densité	0,7915
Constante Diélectrique	32,7

Tableau I.3 Certaines propriétés du métnanol.

I.2.5.3 Le N,N-diméthylformamide

Le N,N-diméthylformamide ou DMF est un solvant courant en chimie organique de formule $\text{HCO-N}(\text{CH}_3)_2$. Ce liquide incolore est miscible à l'eau et à la majorité des liquides organiques. C'est un solvant aprotique polaire à haut point de fusion/ébullition utilisé couramment pour les réactions chimiques. Le DMF est un amide dérivé de l'acide formique et de la diméthylamine[26].

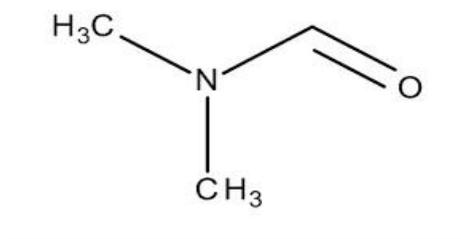


Figure I.6 Molécule de DMF .

La masse molaire	73,093
Point de fusion	-61°C
Point d'ébullition (20°C, 1Atm)	153°C
Tension de vapeur saturante	mm Hg
densité	944
Constante Diélectrique	36,7
Moment Dipolaire	3.86D

Tableau I.4 Certaines propriétés du DMF.

Références Bibliographiques

- [1] X. Wang, J. Yu, Y. Wang, L. Wang, *Chemosphere* 2002, 46, 241–250.
- [2] Z. Rappoport, “The Chemistry of Phenols”, John Wiley & Sons Ltd, Chichester, England (2003)
- [3] M. Ehtash, “Purification des eaux polluées par du phénol dans pertracteur à disques tournants”, Thèse de doctorat, Institut national des sciences appliquées de Rouen, France (2011)
- [4] B. Charriere, “Les composés phénoliques marqueurs de la matière organique terrestre dans deux écosystème marins : le delta du Rhône et le pro delta de la Têt”, Thèse de doctorat, Institut national polytechnique de Toulouse, Université de Perpignan, France (1991).
- [5] Organisation mondiale de la santé, “Environmental health criteria for chlorophenols”, (1987).
- [6] O. Gimeno, M. Carbajo, F. J. Beltrán and F. Javier Rivas, “Phenol and substituted phenols AOP remediation”, *Journal of Hazardous Materials B* 119 (2005) 99-108.
- [7] L.A. Pacheco Palencia, “Chemical characterization, bioactive properties and pigments stability of polyphenolics in ACAI (*Euterpe Oleracea Mart*)”, PhD Thesis, Texas A&M University, U.S.A (2009).
- [8] L. Layton, “Reversing itself, FDA expresses concerns over health risks from BPA”, *Washington Post*, U.S.A (2010).
- [9] D. Podlipna, M. Cichna-Marki, “Determination of bisphenol A in canned fish by sol–gel immunoaffinity chromatography, HPLC and fluorescence detection”, *European Food Research Technology* 224 (2007) 629-634.
- [10] J. Moreira del Rio, “Photocatalytic degradation of phenolic compounds in water: irradiation and kinetic modeling”, PhD thesis, University of Western Ontario, Canada (2011).

- [11] Phenol, <http://en.wikipedia.org/wiki/Phenol> mars (2015).
- [12] McMurry, J. (2016). Organic Chemistry. Cengage Learning
- [13] Brown, TL, LeMay Jr, HE, Bursten, BE, Murphy, CJ et Woodward, PM (2017). Chimie: La science centrale.
- [14] Bardez, É. (2012). L'acido-basicité. Des observations aux concepts, les étapes-clés du début du xix^e siècle. Bulletin de la Sabix. Société des amis de la Bibliothèque et de l'Histoire de l'École polytechnique, (50), 15-29.
- [15] Clayden. Greeves. Warren, Chimie Organique, 2eme édition, traduction d'André Pousse, de boeck
- [16] Dean's Office and Department of Chemistry & Physics, College of Science & Technology, Armstrong Atlantic State University, Savannah, GA, USA
- [17] R Ouahès et B Devallez, Chimie générale, 4e édition revue et corrigée, 2 septembre 1991, p192-197.
- [18] un b Miessler & Tarr . (1991). pp. 167-169 - Selon cette page, la définition originale était que "les acides ont tendance à perdre un proton.
- [19] Dean's Office and Department of Chemistry & Physics, College of Science & Technology, Armstrong Atlantic State University, Savannah, GA, USA
- [20] R Ouahès et B Devallez, Chimie générale, 4e édition revue et corrigée, 2 septembre 1991, p192-197
- [21] Chris P Schaller, Ph.D , (Collège Saint-Benoît / Université Saint John's)
- [22] C. Hansch, A. Leo, Exploring QSAR (American Chemical Society: Washington, DC), 1995.
- [23] D.Bégin, M.Gérin, Les grandes familles de solvant organiques, Utilisation et aspect physico-chimique.
- [23] Bégin, D., Lavoué, J., & Gérin, M. (2002). BILANS DE CONNAISSANCES.
- [25] Otera, J. and Nishikido, J. (2013). In Encyclopedia of Reagents for Organic Synthesis. Wiley. Methanol as a Solvent in Organic Synthesis and Catalysis" -

[26] Karimi, B. et al. (2016). Dans Réactions organiques. Wiley. N,N-diméthylformamide (DMF) en tant que réactif polyvalent dans la synthèse organique".

Chapitre II

Relations Quantitatives Structures Propriétés

II.1 Introduction

Le développement de modèles des Relations Quantitative Structures Propriétés (QSPR) joue un rôle crucial dans la conception et l'optimisation de médicaments et d'autres produits chimiques. Ces modèles permettent de prédire les propriétés de nouvelles molécules en se basant sur leurs caractéristiques structurales. Leur utilisation permet de réduire les coûts et le temps nécessaires à la synthèse et à l'expérimentation de nombreux composés.

Dans ce chapitre, nous présenterons les étapes de développement d'un modèle QSPR et de son application dans la prédiction des propriétés chimiques. Nous examinerons les méthodes utilisées pour la collecte et la préparation des données, la sélection des descripteurs moléculaires pertinents, ainsi que la construction et l'évaluation des modèles QSPR [1].

II.2 Définition

Les modèles QSPR sont des approches de modélisation utilisées en chimie pour établir des relations quantitatives entre la structure des molécules et leurs propriétés physico-chimiques. Ces modèles reposent sur des données expérimentales ou calculées afin de prédire de manière quantitative les propriétés de nouvelles molécules en se basant sur des paramètres spécifiques appelés descripteurs.

L'objectif principal des modèles QSPR est de fournir des outils prédictifs et économiques pour évaluer les propriétés des composés chimiques avant leur synthèse ou leur expérimentation, ce qui permet d'accélérer le processus de découverte et de conception. Ces modèles sont basés sur des relations mathématiques et statistiques entre les caractéristiques structurales et les propriétés cibles, en utilisant des algorithmes d'apprentissage automatique ou des approches statistiques [2].

II.3 Historique

Les relations quantitatives structure-Propriétés (QSPR) sont de plus en plus utilisées, du fait de la croissance des moyens de calculs. Elles ont été abondamment utilisées dans les industries [3].

Les premières études utilisant la méthode QSPR ont commencé dans les années 60 avec Hansch et d'autres chercheurs, qui ont proposé un modèle mathématique pour corréler l'activité biologique et la structure chimique [4]. Les développements de cette méthode remontent même plus loin, à partir de 1868, lorsque des chercheurs ont montré l'existence d'une relation entre l'activité physiologique et la structure chimique [5]. Au fil des décennies, ce domaine a été largement étendu et il existe maintenant une grande quantité de données bibliographiques à ce sujet.

La méthode QSPR permet également d'obtenir des informations utiles sur les propriétés structurelles des composés qui sont responsables de leur activité biologique [6-7]. Cependant, établir une relation directe avec la structure moléculaire n'est pas facile. Bien que l'activité ne puisse pas être directement à partir de la structure moléculaire, il est possible d'identifier plusieurs facteurs connus sous le nom de descripteurs moléculaires qui influencent les propriétés moléculaires sélectionnées[8].

II.4 Principe

Les modèles QSPR reposent sur l'établissement de relations mathématiques entre les caractéristiques structurales d'une molécule (représentées par des descripteurs moléculaires) et leurs propriétés physico-chimiques [9]. L'objectif est de trouver une fonction mathématique f qui relie les descripteurs (D) à la variable cible (Y) :

$$Y = f(D_1, D_2, \dots, D_n)$$

Avec: D_1, D_2, \dots, D_n sont des descripteurs des structures moléculaires.

La fonction f peut être un modèle linéaire, non linéaire ou une combinaison de plusieurs fonctions. Les coefficients ou paramètres du modèle sont déterminés lors de la phase d'apprentissage, où des données expérimentales ou calculées sont utilisées pour ajuster le modèle aux valeurs réelles de la variable cible. En général, les modèles QSPR utilisent des techniques de régression pour estimer les paramètres du modèle[10].

II.5 Étapes clés du développement des modèles QSPR

Dans cette section, nous allons examiner les différentes étapes clés de construction des modèles QSPR, en mettant l'accent sur la collecte des données, la sélection des descripteurs appropriés, le développement du modèle prédictif et l'évaluation de sa performance. En général, les modèles QSPR pour prédire les propriétés physico-chimique implique la construction d'un modèle mathématique en cinq étapes de base (Figure II.1):

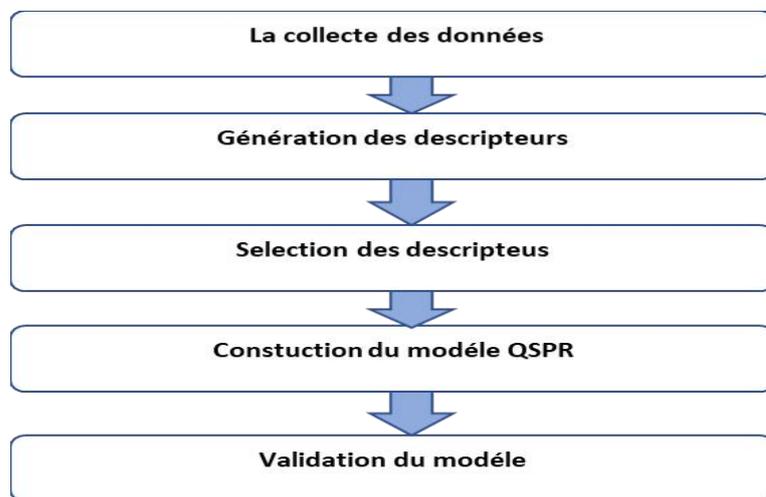


Figure II.1 Les étapes clés du développement des modèles QSPR

II.5.1 La collecte des données

La construction d'un modèle QSPR commence par la collecte de données. Toutes les données (toutes les molécules) utilisées dans les études QSPR sont collectées directement à partir de mesures expérimentales ou extraites de la littérature ou de bases de données. Dans un premier temps, la représentation structurale est réalisée sous forme d'esquisse en deux dimensions. Pour cela, on peut utiliser plusieurs programmes spécifiques tels que Hyperchem [10] (Hypercube Inc., Waterloo, Ontario) ou ChemDraw[11] (Cambridgesoft Corp., Cambridge, mA). Ces programmes permettent de dessiner les structures moléculaires et de stocker les informations structurales sur les composés organiques dans une table de connectivité, qui résume les types d'atomes, les liaisons ainsi que leurs multiplicités.

Une fois la représentation 2D obtenue, une structure tridimensionnelle (3D) est générée par l'optimisation quantique [12]. Cette étape permet de déterminer la conformation spatiale la plus stable et réaliste de la molécule étudiée. La structure 3D optimisée constitue une étape cruciale pour la construction d'un modèle QSPR précis, car elle capture les interactions spatiales et les propriétés moléculaires en 3D.

Ainsi, la collecte de données dans le cadre des modèles QSPR implique la représentation des

structures moléculaires en 2D, suivie de la génération de structures tridimensionnelles optimisées. Ces données structurales serviront de base pour l'extraction des descripteurs et la construction du modèle prédictif.

II.5.2 Génération des descripteurs

Les descripteurs sont des valeurs numériques qui encodent les caractéristiques physico-chimiques des molécules à partir de leurs représentations structurales. La génération de descripteurs constitue une étape importante dans l'analyse de QSPR, car la qualité du modèle dépend essentiellement du type de descripteurs générés. Actuellement, il existe plus de 3000 descripteurs différents qui peuvent être calculés théoriquement à l'aide de divers logiciels tels que OASIS[13], ADAPT[14], CODESSA[15] et DRAGON[16].

Ces logiciels permettent de calculer une large gamme de descripteurs qui capturent différentes caractéristiques des molécules, telles que leur taille, leur forme, leur polarité, leur hydrophobicité, etc. Les descripteurs peuvent également inclure des informations sur les groupes fonctionnels présents dans la molécule, les interactions entre les atomes, les propriétés électroniques, et bien d'autres[17].

Il existe différents types de descripteurs qui permettent de représenter diverses propriétés physico-chimiques des molécules. Les quatre principaux types de descripteurs couramment utilisés sont :

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

leur extrême simplicité, ces descripteurs sont couramment utilisés. Cependant, ils peuvent poser des problèmes d'interprétation des mécanismes, et d'interaction du fait qu'ils ne permettent pas de tenir en compte des effets stériques et d'isomérisation [18].

Les descripteurs 2D: les descripteurs 2D, sont des propriétés numériques qui peuvent être calculées à partir de la table de connectivité d'une molécule ou d'une représentation planaire (2D) de la structure. Ils sont basés sur les éléments présents, les charges partielles, la nature des liaisons, le nombre de cycles [19].

Ils sont calculés à partir de la formule développée de la molécule. On distingue :

- ◆ Les indices 2D constitutionnels : qui caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles ,...etc.
- ◆ Les indices 2D topologiques : peuvent être obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications[20].

Les descripteurs 3D: les descripteurs 3D sont évalués à partir des positions relatives des atomes dans l'espace, et décrivent des caractéristiques plus complexes. Leurs calculs nécessitent donc de connaître la géométrie 3D de la molécule[21].

Les descripteurs 3D géométriques : les plus importants sont le volume moléculaire, la surface accessible au solvant [22].

Les descripteurs 3D électroniques : permettent de quantifier les différents types d'interactions inter- et intramoléculaires.

Les descripteurs 3D ont une grande influence sur l'activité biologique des molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie est minimale, et fait souvent appel à la chimie quantique[23].

II.5.3 Sélection des descripteurs

La sélection des descripteurs les plus pertinents est une étape essentielle dans le développement des modèles QSPR. L'objectif est de choisir les caractéristiques structurales qui ont une influence significative sur la propriété cible, tout en éliminant les descripteurs redondants ou non informatifs.

En effet, il existe deux approches principales pour la sélection des descripteurs : la sélection subjective et la sélection objective[24] :

II.5.3.1 Sélection objective

La sélection objective des descripteurs repose sur des méthodes statistiques et d'apprentissage automatique pour évaluer l'importance des descripteurs de manière plus objective. Ces méthodes utilisent des critères et des algorithmes spécifiques pour évaluer la contribution de chaque descripteur à la prédiction de la variable cible. Cela peut inclure des techniques telles que l'analyse de corrélation, l'analyse en composantes principales, les méthodes de régression pénalisée, les méthodes basées sur l'information mutuelle, les arbres de décision, etc. Ces approches permettent de réduire les biais subjectifs et d'identifier les descripteurs les plus informatifs de manière plus rigoureuse.

En résumé, la sélection subjective repose sur l'expertise et l'intuition des experts, tandis que la sélection objective utilise des méthodes statistiques et d'apprentissage automatique pour évaluer l'importance des descripteurs de manière plus objective. Les deux approches peuvent être combinées pour obtenir une sélection des descripteurs pertinente et fiable [25].

II.5.3.2 Sélection subjective

La sélection subjective des descripteurs repose sur l'expertise et les connaissances préalables dans le domaine d'étude. Les experts choisissent les descripteurs en fonction de leur intuition, de leur expérience et de leur compréhension des relations entre les caractéristiques structurales et les propriétés cibles [26-27]. Cette approche peut être rapide et efficace, surtout lorsque les experts possèdent une expertise approfondie dans le domaine. Cependant, elle peut être influencée par des biais et des préjugés, et il est important de valider les choix effectués [28].

II.5.4 Développement du modèle QSPR

L'étape de développement du modèle QSPR est cruciale pour construire un modèle prédictif fiable. Cette étape comprend plusieurs étapes clés, notamment la division des ensembles d'apprentissage et de test, la sélection du modèle et l'optimisation des paramètres.

En premier temps, les ensembles de données sont divisés en ensembles d'apprentissage et de test. L'ensemble d'apprentissage est utilisé pour entraîner le modèle, tandis que l'ensemble de test est utilisé pour évaluer sa performance et sa capacité de prédiction.

Une fois les ensembles de données établis, différents modèles QSPR sont explorés. Il existe de nombreuses méthodes de modélisation, telles que la régression linéaire, les réseaux de neurones, les méthodes basées sur les arbres de décision, etc. Le choix du modèle dépendra de la nature des données et des objectifs de prédiction [29].

II.5.4.1 Les méthodes statistiques

La régression des moindres carrés partiels

La méthode de régression des moindres carrés partiels (PLS) est une technique d'analyse multivariée largement utilisée dans le domaine de la modélisation QSPR. Elle vise à établir une relation linéaire entre les descripteurs moléculaires et les propriétés cibles en utilisant une approche de régression partielle. Contrairement à la régression linéaire traditionnelle, la PLS tient compte à la fois de la variance des descripteurs moléculaires et de la covariance avec les propriétés cibles.

La méthode PLS est particulièrement adaptée lorsque les données présentent une haute

dimensionnalité et une corrélation complexe entre les variables. Elle est capable de réduire la dimensionnalité de l'espace des descripteurs tout en préservant les informations essentielles pour la prédiction.

La PLS est souvent utilisée lorsque les relations entre les descripteurs et les propriétés cibles sont non linéaires ou lorsque les interactions entre les variables sont complexes. Elle permet de capturer efficacement ces relations complexes en utilisant des combinaisons linéaires adaptatives des descripteurs.

Après avoir sélectionné le modèle, l'optimisation des paramètres est réalisée pour améliorer les performances du modèle. Cela implique l'ajustement des paramètres du modèle afin de minimiser l'erreur de prédiction et d'optimiser sa capacité à généraliser de nouvelles données. [30-31].

La régression linéaire multiple

La régression linéaire multiple (MLR) est l'une des méthodes les plus couramment utilisées et les plus simples pour construire des modèles QSPR. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre un ensemble de descripteurs moléculaires (représentés par un vecteur de caractéristiques x , chaque descripteur étant une composante) et une activité spécifique (représentée par une quantité y) d'un composé donné. Un modèle MLR peut être décrit à l'aide de l'équation suivante :

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (1)$$

Avec :

- ◆ $\{X_1, \dots, X_k\}$ sont des descripteurs moléculaires,
- ◆ β_0 est la constante du modèle de régression,
- ◆ β_1 à k sont les coefficients des descripteurs individuels. Coefficient pour chaque descripteur X_1 à X_k . Les valeurs pour 0 à k sont choisies en minimisant la somme des carrés des résidus entre les valeurs observées et prédites définies par l'équation de manière à donner la meilleure prédiction de y partir de x [32].

L'avantage de la régression linéaire multiple (MLR) réside dans sa simplicité et son expression mathématique facilement interprétable. Les coefficients β_1 à β_k dans l'équation de régression indiquent si chaque descripteur moléculaire contribue positivement ou négativement à une activité spécifique, et leur valeur absolue reflète l'importance relative de chaque descripteur

pour cette activité[33].Cependant, il est important de noter que la MLR fonctionne de manière optimale dans les cas où la relation structure-activité est linéaire et lorsque les descripteurs moléculaires sont mathématiquement indépendants (orthogonaux) les uns des autres [34] .

II.5.4.1.3 Les réseaux de neurones artificiels

Les réseaux neuronaux artificiels (RNA) sont devenus des outils précieux dans de nombreux domaines en résolvant des problèmes de classification, de prédiction, de reconnaissance de formes, de catégorisation, et d'optimisation [35]. Les RNA sont conçus pour imiter le fonctionnement interne du cerveau, qui est composé d'un réseau interconnecté de neurones. Dans la modélisation, chaque neurone possède une couche d'entrée par laquelle les données sont introduites. Ensuite, le neurone produit une valeur +1 ou -1 en fonction de si la somme pondérée dépasse un certain seuil. Les poids de chaque neurone sont ajustés au cours de l'apprentissage pour optimiser la performance du réseau. Cette architecture interconnectée des unités neuronales dans les RNA permet de capturer des relations complexes entre les descripteurs et les propriétés cibles [36].

- ◆ **La couche d'entrée** : elle est constituée d'autant d'entrées que de descripteurs pour l'ensemble de données entier. Chaque neurone est connecté aux neurones de la couche cachée.
- ◆ **La couche cachée** : elle est constituée d'un nombre variable de neurones. Pour chaque neurone, le réseau effectue une opération de somme pondérée avec les différents poids de chaque neurone d'entrée. Un réseau de neurones peut avoir plusieurs couches cachées. Un poids est associé à chaque neurone de la couche cachée.
- ◆ **La couche de sortie** : le nombre de neurones est égale au nombre de propriétés modélisées. Pendant la phase d'apprentissage du modèle par un réseau de neurones, les molécules sont présentées une par une aux neurones de la couche d'entrée.

Le système change itérativement les poids associés aux neurones d'entrées et de la couche cachée de façon à minimiser l'erreur entre la propriété calculée et la propriété expérimentale. A cette fin, il ajuste les poids de chaque neurone[36].

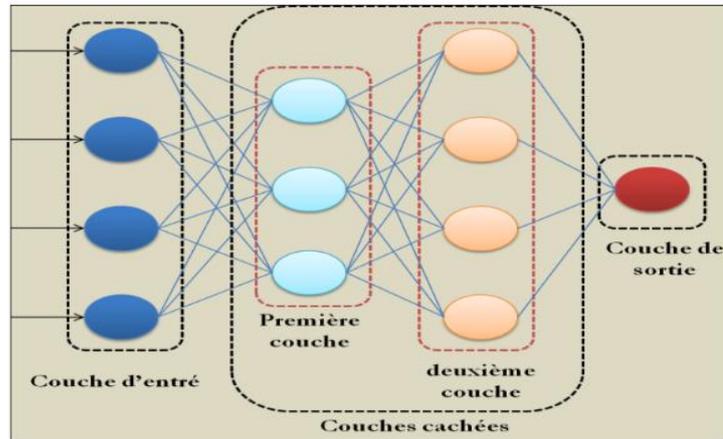


Figure II.2 Représentation d'un réseau de neurone

Les informations provenant de la couche d'entrée sont transmises à la couche cachée pour la configuration de traitement seront ensuite transmises de la couche cachée à la couche de sortie. Dans un algorithme de rétropropagation, l'erreur calculée est dérivée de la différence entre la valeur prédite et la valeur réelle, et si elle est acceptable, le processus d'apprentissage ne cesse autrement signaux seront envoyés vers l'arrière de la couche cachée a un traitement ultérieur et réajustements de poids. Ceci est réalisé de manière itérative jusqu'à ce qu'une solution soit trouvée et l'apprentissage est terminé[37].

II.5.4.1.4 Les machine à vecteurs de support

Les machines à vecteurs de support (SVM) sont des algorithmes d'apprentissage automatique puissants. Ils sont couramment utilisés pour des tâches de classification et de régression. L'idée centrale des SVM est de trouver une frontière de décision optimale qui sépare les différentes classes de données avec la plus grande marge possible. Cette frontière est déterminée en utilisant un sous-ensemble d'échantillons d'entraînement, appelés vecteurs de support, qui se trouvent le plus proche de la frontière de décision [38].

Les SVM offrent également une bonne capacité de généralisation, ce qui signifie qu'ils peuvent prédire avec précision les classes des nouvelles données qu'ils n'ont pas rencontrées lors de l'entraînement.

II.6 Evaluation du modèle

L'évaluation des modèles QSPR est une étape cruciale pour évaluer leur performance et leur capacité à prédire avec précision les propriétés cibles des composés chimiques. L'évaluation peut être effectuée à la fois en interne et en externe[40].

II.6.1 Evaluation interne

L'évaluation interne consiste à évaluer les performances du modèle sur les données d'apprentissage utilisées pour le construire. Elle permet de vérifier la capacité du modèle à ajuster les données d'entraînement et à capturer les relations entre les descripteurs moléculaires et les propriétés cibles. Les mesures d'évaluation interne couramment utilisées comprennent l'erreur quadratique moyenne (RMSE), le coefficient de détermination (R^2) et l'erreur absolue moyenne (MAE). Une évaluation interne rigoureuse implique souvent l'utilisation de méthodes de validation croisée telles que la validation croisée k-fold ou la validation croisée leave-one-out [41].

II.6.2 Evaluation externe

L'évaluation externe, en revanche, teste les performances du modèle sur un ensemble de données indépendant, distinct de l'ensemble d'apprentissage. Cela permet de mesurer la capacité du modèle à généraliser et à prédire les propriétés des nouveaux composés qui n'ont pas été utilisés pendant l'entraînement. L'évaluation externe fournit une estimation plus réaliste de la capacité prédictive du modèle dans des situations réelles. Les mêmes mesures d'évaluation que pour l'évaluation interne peuvent être utilisées pour évaluer les performances du modèle en externe.

Idéalement, un bon modèle QSPR devrait obtenir de bons résultats à la fois en évaluation interne et en évaluation externe. Cependant, si le modèle montre une performance élevée en évaluation interne mais une performance médiocre en évaluation externe, cela peut indiquer un surajustement (overfitting) aux données d'apprentissage et une faible capacité de généralisation [42].

II.7 Les métriques d'évaluation

II.7.1 Coefficient de corrélation (R)

Le coefficient de corrélation est une mesure statistique utilisée pour évaluer la relation linéaire entre deux variables. Il indique à quel point les variations d'une variable sont associées aux variations de l'autre variable. Le coefficient de corrélation est noté R, sa valeur varie entre 0 et 1:

$$R = \sqrt{1 - \frac{\sum (y - \hat{y})^2}{\sum (\hat{y} - y)^2}}$$

Avec : Y et \hat{y} sont, respectivement, les valeurs observées et calculées de la variable dépendante.

Lors de l'évaluation des modèles QSPR, le coefficient de corrélation est souvent utilisé pour mesurer la qualité de la prédiction. Un coefficient de corrélation élevé entre les valeurs prédites et les valeurs réelles indique une bonne capacité du modèle à capturer la relation entre la structure et la propriété étudiée.

II.7.2 Coefficient de détermination (R^2)

Le coefficient de détermination, également appelé R^2 , est une mesure statistique utilisée pour évaluer la qualité d'ajustement d'un modèle de régression.

Lors de l'utilisation de modèles QSPR, le coefficient de détermination est souvent utilisé pour évaluer la performance du modèle dans la prédiction de la propriété étudiée. L'équation du coefficient de détermination (R^2) peut être exprimée comme suit

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

Avec:

- ◆ Y : La valeur réelle.
- ◆ \hat{y} : La valeur prédite.

- ◆ \bar{y} : La moyenne de la somme des valeurs réels.

II.7.3 L'erreur absolue moyenne

L'erreur absolue moyenne (MAE - Mean Absolute Error) est une mesure statistique utilisée pour évaluer la précision d'un modèle de prédiction. Elle quantifie la différence moyenne absolue entre les valeurs prédites par le modèle et les valeurs réelles de la variable cible.

Pour calculer le MAE, on prend la différence absolue entre chaque prédiction et la valeur réelle correspondante, puis on fait la moyenne de ces différences.

L'erreur absolue moyenne est souvent utilisée dans divers domaines, y compris les sciences, l'économie, la finance et l'apprentissage automatique. Elle fournit une mesure facilement interprétable de la performance du modèle et est particulièrement utile lorsque des erreurs de prédiction significatives peuvent avoir un impact important.

L'équation de l'erreur absolue moyenne peut être exprimée comme suit :

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

Avec:

- ◆ \hat{y} : valeur prédite de y

II.7.4 L'erreur quadratique moyenne

La MSE (Mean Squared Error) est une autre mesure statistique utilisée pour évaluer la précision d'un modèle de prédiction. Elle mesure la moyenne des carrés des différences entre les valeurs prédites par le modèle et les valeurs réelles de la variable cible.

Pour calculer la MSE, on calcule d'abord l'erreur quadratique pour chaque prédiction en soustrayant la valeur prédite de la valeur réelle, puis on élève cette différence au carré. Ensuite, on fait la moyenne de ces erreurs quadratiques pour obtenir la MSE comme suit:

$$MSE = 1/N(y_i - \hat{y})^2$$

II.7.5 L'erreur quadratique moyenne

Le RMSE (Root Mean Squared Error) est une mesure statistique largement utilisée pour évaluer la précision d'un modèle de prédiction. Il est dérivé de la MSE (Mean Squared Error) en prenant la racine carrée de la moyenne des carrés des différences entre les valeurs prédites par le modèle et les valeurs réelles de la variable cible.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_i^N (y_i - \hat{y})^2}$$

II.8 Application du QSPR

Les applications des modèles QSPR sont vastes et couvrent de nombreux domaines scientifiques et industriels. Voici quelques-unes des principales applications des QSPR :

Conception de médicaments : Les modèles QSPR sont utilisés pour prédire les propriétés pharmacocinétiques et les interactions médicamenteuses, ce qui aide à la conception et à l'optimisation de nouveaux médicaments [43]. Ils permettent également d'évaluer la toxicité potentielle des composés avant leur synthèse et leur test *in vitro* ou *in vivo*, ce qui peut accélérer le processus de découverte de médicaments.

Chimie environnementale : Les modèles QSPR sont utilisés pour prédire les propriétés écotoxicologiques des produits chimiques et leur impact sur l'environnement.

Propriétés physicochimiques : Les modèles QSPR permettent de prédire diverses propriétés physicochimiques des composés chimiques, telles que la solubilité, la volatilité, la polarité, la stabilité et la réactivité chimique. Ces prédictions sont utiles dans de nombreux domaines, tels que la formulation de produits chimiques, la conception de catalyseurs et la modélisation des processus chimiques [44-45].

Toxicologie et sécurité chimique : Les modèles QSPR sont utilisés pour évaluer la toxicité et les propriétés de sécurité des produits chimiques. Ces informations sont essentielles pour garantir la sécurité des produits chimiques et protéger la santé humaine et l'environnement.

II.9 Conclusion

Ce chapitre a fourni une compréhension approfondie des modèles QSPR, en commençant par la définition de ces concepts et en explorant leur principe de base. Nous avons également

examiné l'histoire des modèles QSAR/QSPR, remontant aux premières contributions de chercheurs.

De plus, nous avons passé en revue les étapes clés du développement des modèles QSPR. Cela comprend la sélection appropriée des descripteurs moléculaires, qui sont des caractéristiques structurelles utilisées pour représenter les molécules, ainsi que les méthodes statistiques et d'apprentissage automatique utilisées pour créer des relations quantitatives entre la structure et les propriétés.

L'étape de développement des modèles QSAR/QSPR nécessite une collecte de données fiables et représentatives, ainsi que la validation rigoureuse des modèles pour s'assurer de leur précision et de leur généralisabilité. Nous avons également exploré les domaines d'application des modèles QSPR, qui s'étendent à de nombreux domaines de la chimie, de la pharmacologie et de la toxicologie.

Ces connaissances sont cruciales pour comprendre et appliquer efficacement les modèles QSPR dans la prédiction des propriétés moléculaires, ouvrant ainsi la voie à de futures avancées dans le domaine de la modélisation moléculaire et de la conception de composés.

Références Bibliographiques

- [1] Roy, K., Kar, S., & Das, R. N. (2015). A primer on QSAR/QSPR modeling: fundamental concepts. Springer.
- [2] BELLIFA, K. (2015). Etude des relations quantitatives structure–toxicité des composés chimiques à l’aide des descripteurs moléculaires.«Modélisation QSAR (Doctoral dissertation).
- [3] C. Hansch, T. Fujita, J. Am. Chem. Soc., 1964, 86, 1616-1626.S.M. Free, J.W. Wilson, J. Med. Chem., 1964, 7, 395-399
- [4] A. Crum Brown, T.R. Fraser, Trans. Roy.Soc. Edinburgh, 1868, 25, 151-203.
- [5] A. Crum-Brown, T.R. Fraser, Trans. R. Soc. Edinburgh, 25, 151 (1868).
- [6] C. Richet, C. R. Séances Soc. Biol. Ses. Fil., 9, 775 (1893).
- [7] KRAIM, K. (2009). ETUDES DE QSAR SUR DES ACTIVITES BIOLOGIQUES UTILISANT DES PRODUITS D'ORIGINES NATURELS (Doctoral dissertation).
- [8] S Chtita, Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments, 2017
- [9] K. H. Park; Y. D. Park; J. M. Han; K. R. Im; B. W. Lee; I. Y. Jeong; T. S. Jeong; W.S. Lee Bioorg. Med. Chem. Lett. 16 (2006) 5580
- [10] HyperChem Release 7, Hyper Cube, Inc., <http://www.hyper.com>.
- [11] Chem Draw (Cambridgesoft Corp., Cambridge), www.cambridgesoft.com.
- [12] M. J. S. Dewar; E. G. Zoebisch; E. F. Healy; J.J. P. Stewart; J. Am. Chem. Soc. 107 (1985) 3902- 3909.
- [13] O. Mekenyan; D. Bonchev Acta Pharm. Jugosl. 36 (1986) 225-237.
- [14] R. Guha; J. R. Serra; P. C. Jurs J. Mol. Graph. Mod. 23 (2004) 1–14.
- [15] A. R. Katritzky; V. S. Lobanov; M. Karelson. Manuel du logiciel CODESSA. Gainesville, FL. University of Florida. 1994. (www.semichem.com/codessraefs.html)
- [16] R. Todeschini, Milano Chemometrics and QSPR Group, (<http://www.disat.unimib.it/vhml>).
- [17-19] Hmamouchi, R., Bouachrine, M., & Lakhlifi, T. (2016). Tentative Pratique du Relation Quantitatives Structure-Activité/Propriété (QSAR/QSPR). Revue Interdisciplinaire, 1(1).

- [20] V.Y. Nalimov, the Application of Mathematical Statistics to Chemical Analysis, AddisonWesley, Reading, MA, 1962.
- [21] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-Chemical Descriptors in QSAR/QSPR Studies, Chem. Rev.. 96, 1027-1044, 1996
- [22] F.Luan, X. Xu, H. Liu, Cordeiro, M.N.D.S. Prediction of the baseline toxicity of non-polar narcotic chemical mixtures by QSAR approach. Chemosphere, 90, 1980–1986, 2013
- [23] M. Srivastava, H.Singh, P.K. Naik, Quantitative structure–activity relationship (QSAR) of artemisinin: the development of predictive in vivo antimalarial activity models. J. Chemometr., 23, 618–635, 2009
- [24] J. H. Kalivas “Adaption of simulated annealing to chemical optimization problems” Ed. Elsevier Science B.V. Pays Bas, 1995.
- [25] KRAIM, K. (2009). ETUDES DE QSAR SUR DES ACTIVITES BIOLOGIQUES UTILISANT DES PRODUITS D'ORIGINES NATURELS (Doctoral dissertation).
- [26] P. Gramatica, P. Pilutti, E. Papa, J. Chem. Inf. Comp. Sci. 44, 2004, 1794–1802.
- [27] B. Bhatarai, P. Gramatica, Chem Res Toxicol. 23, 2010, 528–539
- [28] Daniel Borcard, « Régression multiple - corrélation multiple et partielle », Département de sciences biologiques Université de Montréal, 2009
- [29] R. Benigni “Quantitative Structure-Activity Relationship (QSAR): Models of Mutagens and Carcinogens” Ed. CRC Press LLC. Etats Unies, 2003.
- [30] S. Wold, M. Sjöström, L. Eriksson PLS-regression: a basic tool of chemometrics, Chemom Intell Lab Syst, 58, 109–130, 2010.
- [31] N. Chirico, P. Gramatica, Real External predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model, 51,2320–2335, 2011.
- [32] Allison, P. Multiple Regression. Pine Forge Press: Thousand Oaks, CA, 1999
- [33] Topliss, J. G.; Edwards, R. P. J. Med. Chem., 1979, 22, 1238.
- [34] Eriksson, L.; Jaworska, J.; Cronin, M.; Worth, A.; Gramatica, P.; McDowell, R. Environ.

Health Perspect., 2003, 111, 1361.

[35] P.J. DREW, et J.R.T. MONSON, (2000). Artificial neural networks, *Surgery* 127: 3-11

[36] Hmamouchi, R., Bouachrine, M., & Lakhlifi, T. (2016). Tentative Pratique du Relation Quantitatives Structure-Activité/Propriété (QSAR/QSPR). *Revue Interdisciplinaire*, 1(1).

[37] Hmamouchi, R., Bouachrine, M., & Lakhlifi, T. (2016). Tentative Pratique du Relation Quantitatives Structure-Activité/Propriété (QSAR/QSPR). *Revue Interdisciplinaire*, 1(1).

[38] COMPUTATIONNELLE, C. (2019). TOUHAMI Moufida (Doctoral dissertation, Université de Saida).

[39] MARSLAND, Stephen. Machine learning: an algorithmic perspective.

[40] A. Varnek; A. Tropsha "Chemoinformatics Approaches to Virtual Screening" Ed. Royal Society of Chemistry. Royaume- Uni, 2008.

[41] Hmamouchi, R., Bouachrine, M., & Lakhlifi, T. (2016). Tentative Pratique du Relation Quantitatives Structure-Activité/Propriété (QSAR/QSPR). *Revue Interdisciplinaire*, 1(1).

[42] G. Schuurmann, R.U. Ebert, J. Chen, B. Wang, R. Kuhne, External validation and prediction employing the predictive squared correlation coefficient-Test-set activity mean vs training set activity mean. *J Chem Inf Model*, 48, 2140–2145, 2008.

[43] T. Puzyn, J. Leszczynski, M. T. D. Cronin, « Recent advances in QSAR studies Methods and applications » Springer Dordrecht Heidelberg London New York, 2010

[44] M. Tichý, *Science.Tot. Env.* 109, 1991, 407-410.

[45] O. M. Bautista-Aguilera, G. Esteban, I. Bolea, K. Nikolic, D. Agbaba, I. Moraleda, I. Iriepa, A. Samadi, E. Soriano, M. Unzeta, J. Marco-Contelles, *Europ. J. Med. Chem.* 75, 2014, 82-95.

Chapitre III

Résultats et Discussion

III.1 Introduction

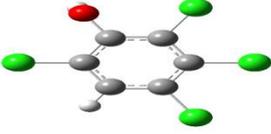
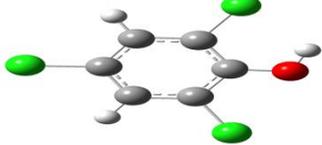
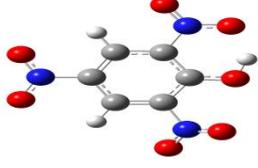
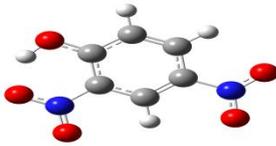
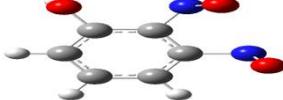
Les modèles QSPR (Quantitative Structure-Property Relationship) ont été largement utilisés pour la prédiction de diverses propriétés physico-chimiques des composés, y compris le pKa. Le pKa est un paramètre important qui détermine le comportement ionique des composés ionisables en solution. Dans le cas des phénols, le pKa détermine leur capacité à donner un ion phénolate (-O^-) en solution. Dans cette étude, nous avons développé un modèle QSPR pour la prédiction du pKa des phénols dans différents solvants. Ce modèle est basé sur une analyse de la structure moléculaire des phénols et de leur environnement solvant, ainsi que sur une base de données de pKa expérimentaux pour un ensemble de composés. Nous avons utilisé une approche statistique pour développer ce modèle, en utilisant des techniques d'apprentissage automatique pour prédire le pKa en fonction des descripteurs moléculaires et des propriétés des solvants.

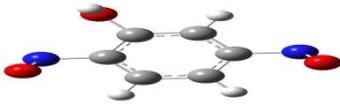
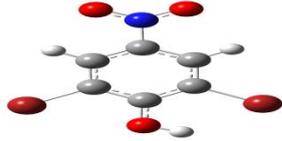
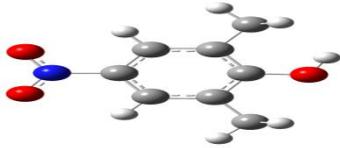
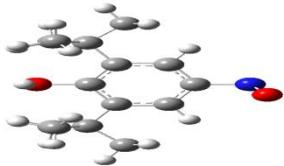
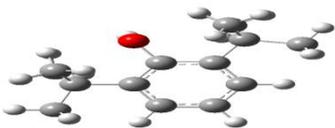
L'objectif de ce chapitre est de présenter des modèles QSPR (Quantitative Structure Property Relationship) et d'évaluer leur capacité à prédire les activités biologiques. Pour ce faire, nous avons collecté un ensemble de données et avons appliqué des processus tels que le calcul des descripteurs, la sélection des descripteurs pertinents, la modélisation QSPR, ainsi que l'évaluation des modèles développés à l'aide de divers paramètres statistiques. Cette approche nous a permis de déterminer l'efficacité de ces modèles dans la prédiction des activités biologiques, ce qui peut être utile pour la conception de nouveaux composés et la compréhension de leur activité biologique.

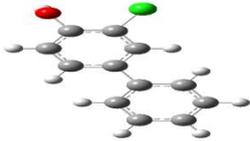
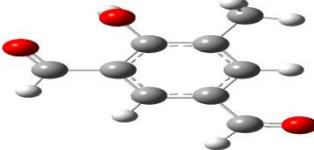
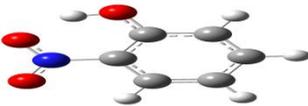
III.2 Calcul et Sélection des descripteurs

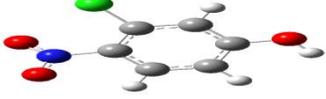
III.2.1 La collecte des données

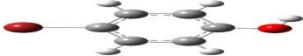
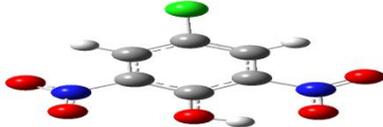
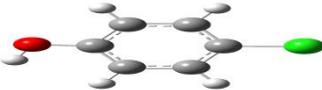
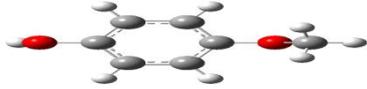
Avant de procéder au calcul des descripteurs moléculaires, il est important de définir une liste des molécules qui sont utilisées dans notre étude. Cette liste est constituée de 62 phénols. La liste des molécules utilisées dans notre étude est présentée dans les Tableaux III.1, III.2 et III.3. Ces tableaux fournissent une vue d'ensemble des 62 phénols sélectionnés, en présentant leurs noms et leurs structures moléculaires.

N°	Nom	Structure
1	2,3,4,6-Tetrachlorophenola	
2	2,4,5-Trichlorophenola	
3	2,4,6-Trichlorophenola	
4	2,4,6-Trinitrophenola	
5	2,4-Dichlorophenol	
6	2,4-Dinitrophenola	
7	2,3-dinitrophenol	

N°	Nom	Structure
8	2,5-Dinitrophenolb	
9	2,6-Dibromo-4-nitrophenola	
10	2,6-Dichlorophenola	
11	2,6-Dimethyl-4-nitrophenola	
12	2,6-Dinitrophenolc	
13	2,6-Di-tert-Butyl-4-nitrophenolb	
14	2,6-Ditertbutylphenol	

N°	Nom	Structure
15	2-Chloro-4-phenylphenola	
16	2-Methyl-4,6-dinitrophenol	
17	2-Nitrophenola	
18	3-(Methylsulfonyl)phenola	
19	3,4,5-Trichlorophenola	
20	3,4-Dichlorophenolc	
21	3,4-Dinitrophenola	
22	3,5-Dichlorophenolb	

N°	Nom	Structure
23	3,5-Dinitrophenolc	
24	3-Acetylphenola	
25	3-chloro4-nitrophenol	
26	3-chlorophenol	
27	3-Cyanophenol	
28	3-Fluorophenola	
29	3-Hydroxyphenola	
30	3-Methoxyphenolc	
31	3-Methylphenola	

N°	Nom	Structure
32	3-trifluoromethylpenol	
33	3-Nitrophenola	
34	3-Trifluoromethyl4-nitrophenola	
35	4-Acetylphenolc	
36	4-Bromophenolc	
37	4-Chloro-2,6-dinitrophenola	
38	4-Chlorophenola	
39	4-Cyanophenola	
40	4-Methoxyphenolb	

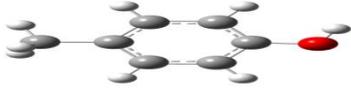
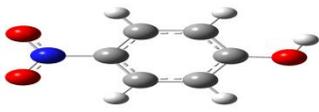
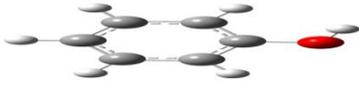
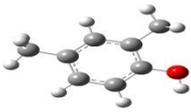
N°	Nom	Structure
41	4-Methylphenola	
42	4-Nitrophenola	
43	Pentachlorophenola	
44	Phenola	

Tableau III.1 Les molécules de DMSO

N°	Nom	Structure
45	2,3-Dimethylphenola	
46	2,4,6-Trimethylphenola	
47	2,4-Dimethylphenola	
48	2,4-Di-tert-Butylphenola	

N°	Nom	Structure
49	2,5-Dimethylphenola	
50	2,6-Dimethylphenolc	
51	2-Chloro-4-bromophenola	
52	2-Chlorophenolc	
53	2-Methylphenola	
54	2-tert-Butylphenola	
55	3,5-Dimethylphenola	
56	3,5-Di-tert-Butylphenola	
57	3-Bromophenola	
58	4-tert-Butylphenola	

Tableau III.2 Les molécules de Me-OH.

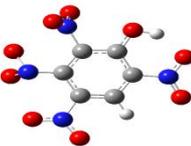
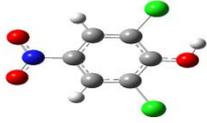
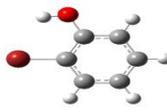
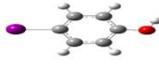
N°	Nom	Structure
59	2,3,4,6-Tetranitrophenolc	
60	2,6-Dichloro-4-nitrophenola	
61	2-Bromophenolc	
62	4-Iodophenola	

Tableau III.3 Les molécules DMF

Afin de préparer les fichiers d'entrée nécessaires pour le calcul des descripteurs, une étape préliminaire importante consiste à visualiser et manipuler les structures moléculaires des 62 phénols sélectionnés pour notre étude. Pour réaliser cette tâche, nous avons utilisé le logiciel GaussView, qui offre des fonctionnalités avancées de visualisation et de manipulation moléculaire. La Figure III.1, ci-dessous, illustre un exemple de la représentation moléculaire obtenue grâce à l'utilisation du logiciel **GaussView**[1].

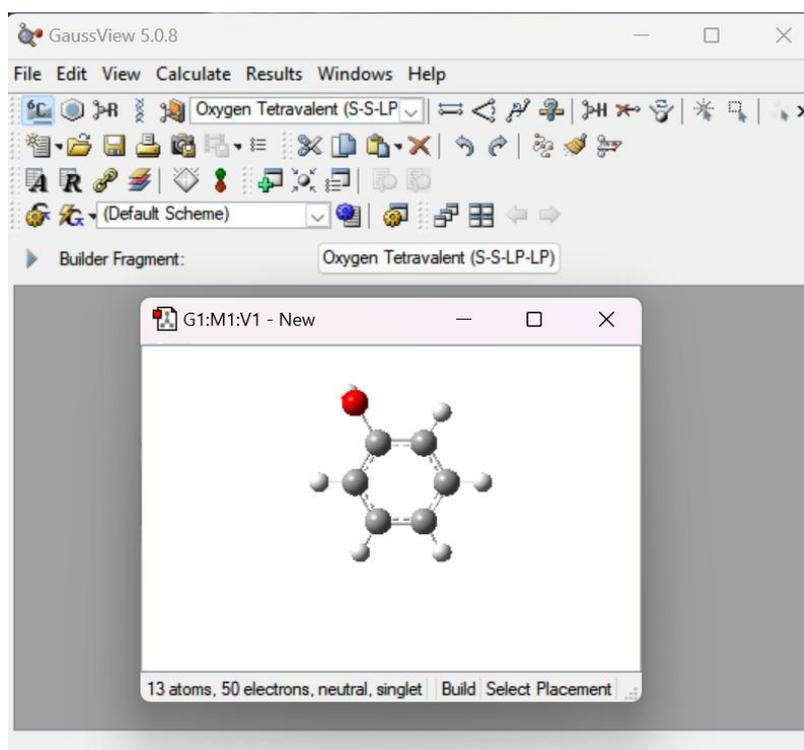


Figure III.1 Exemple de la représentation moléculaire sur Gaussview.

III.2.2 Calculs des descripteurs

La détermination des descripteurs moléculaires, qui représentent numériquement la structure chimique, joue un rôle essentiel dans le développement des modèles QSPR.

Dans notre étude, nous avons utilisé la méthode DFT (Density Functional Theory) pour calculer les descripteurs moléculaires. La méthode DFT est une approche théorique couramment utilisée pour calculer les propriétés électroniques et structurales des molécules. Contrairement aux orbitales atomiques, elle se base sur la densité électronique. Dans nos calculs, nous avons utilisé la base 6-31G+(d,p), qui est un ensemble de fonctions d'onde permettant de représenter les orbitales électroniques. Cette base est particulièrement adaptée pour les calculs de chimie organique, offrant une description précise des molécules contenant des atomes d'hydrogène, d'oxygène et de carbone, ce qui est le cas pour le groupement phénol.

Pour ce faire, nous avons effectué une optimisation de géométrie en utilisant le logiciel **Gaussian 09** [2]. Cette approche nous a permis de déterminer les descripteurs clés tels que l'énergie de la HOMO (plus haute orbitale moléculaire occupée), l'énergie de la LUMO (plus basse orbitale moléculaire inoccupée), ainsi que les charges des atomes d'hydrogène (H) et d'oxygène (O) présents dans le groupement phénol. De plus, nous avons également

obtenu des informations sur la distance de liaison du groupement phénol et l'angle HOC. Ces résultats sont fondés sur l'utilisation directe de la méthode DFT dans notre étude pour calculer les descripteurs moléculaires. En ce qui concerne les solvants, nous avons utilisé la constante diélectrique comme descripteur.

Le tableau présenté ci-dessous illustre les résultats des descripteurs calculés ainsi que les valeurs expérimentales des pKa correspondantes.

N°	HOMO(u.a)	LUMO(u.a)	La charge de H	La charge de O	La dist O-H	L'angleHOC	ϵ	Pka exp
1	-0.25376	-0.05743	0.384452	-0.529787	0.97058	109.59677	45	7.55
2	-0.25000	-0.05188	0.379716	-0.548173	0.97008	109.76881	45	10.97
3	-0.25053	-0.05218	0.382433	-0.540449	0.97026	109.48433	45	10.19
4	-0.31587	-0.15950	0.408203	-0.443383	0.99168	107.27975	45	-0.30
5	-0.24362	-0.04323	0.376800	-0.558252	0.966999	109.70052	45	13.25
6	-0.29422	-0.13910	0.386574	-0.548250	0.98821	107.49093	45	5.32
7	-0.28533	-0.13794	0.386770	-0.541929	0.98674	107.77324	45	4.09
8	-0.29559	-0.20581	0.387315	-0.548110	0.96000	109.4999	45	7.32
9	-0.27346	-0.11479	0.383938	-0.462032	0.97309	109.63477	45	5.17
10	-0.24890	-0.04025	0.380137	-0.538521	0.97026	109.29574	45	11.54
11	-0.25600	-0.09499	0.384226	-0.570908	0.96612	110.69933	45	9
12	-0.28971	-0.14010	0.400146	-0.453352	0.99010	106.98282	45	4.82
13	-0.25108	-0.9229	0.375877	-0.503368	0.96296	110.56920	45	7.60
14	-0.24101	-0.01201	0.370238	-0.522709	0.96616	108.89218	45	16.85
15	-0.23733	-0.05193	0.375398	-0.557791	0.96000	109.49995	45	14.90
16	-0.26364	-0.09366	0.388069	-0.524873	0.98575	104.35407	45	4.59
17	-0.26335	-0.11705	0.387183	-0.528128	0.98632	107.14708	45	11
18	-0.25665	-0.05467	0.351288	-0.520132	0.96637	110.42379	45	13.56
19	-0.25071	-0.05029	0.377314	-0.558562	0.96634	110.52231	45	12.58
20	-0.24285	-0.04191	0.374677	-0.570146	0.96629	110.51128	45	14.22
21	-0.38203	0.02882	0.399255	-0.596245	0.94417	112.72649	45	7.97
22	-0.24414	-0.04190	0.376312	-0.559809	0.96638	110.44334	45	13.09
23	-0.28990	-0.13334	0.382951	-0.556741	0.96693	110.98351	45	10.60
24	-0.24555	-0.07187	0.374089	-0.564700	0.96656	109.99186	45	15.14

N°	HOMO(u.a)	LUMO(u.a)	La charge de H	La charge de O	La dist O-H	L'angleHOC	ϵ	Pka exp
25	-0.27236	-0.10236	0.378447	-0.561579	0.96690	110.66000	45	1.16
26	-0.24271	-0.03129	0.373043	-0.571737	0.96622	110.38341	45	9.80
27	-0.25832	-0.06735	0.375761	-0.563341	0.96644	110.55955	45	14.76
28	-0.24192	-0.02483	0.372411	-0.580870	0.96620	110.29762	45	15.88
29	-0.22671	-0.01401	0.371279	-0.581870	0.96620	110.04085	45	15.30
30	-0.22356	-0.01213	0.371336	-0.581378	0.96617	109.95815	45	15.72
31	-0.22887	-0.011711	0.370091	-0.573765	0.96603	109.85572	45	16.86
32	-0.25328	-0.04815	0.374772	-0.563937	0.96640	110.42772	45	14.30
33	-0.26336	-0.10920	0.376173	-0.566541	0.96663	110.54412	45	13.75
34	-0.27941	-0.11002	0.381440	-0.551879	0.96711	110.99983	45	9.30
35	-0.24668	-0.06182	0.373413	-0.565963	0.96665	110.24544	45	13.68
36	-0.23345	-0.03293	0.371802	0.581117	0.96615	110.17879	45	15.50
37	-0.28606	-0.14779	0.402461	-0.452930	0.98944	107.08817	45	3.51
38	-0.23512	-0.03262	0.371885	-0.578263	0.96610	110.20641	45	16.10
39	-0.25522	-0.05526	0.375036	-0.566991	0.96668	110.54825	45	13.01
40	-0.24195	-0.02236	0.369831	-0.585770	0.96622	110.01019	45	17.58
41	-0.31587	-0.15950	0.408203	-0.443383	0.96599	109.86602	45	16.96
42	-0.26811	-0.10089	0.376771	-0.570078	0.96688	110.66398	45	11
43	-0.25839	-0.06463	0.384706	-0.532965	0.97081	109.52142	45	7.05
44	-0.23253	-0.01869	0.369577	-0.585162	0.96612	109.97973	45	16.47
45	-0.22356	-0.00966	0.370932	-0.579643	0.96584	109.61851	32.7	15.08
46	-0.21442	-0.00709	0.376856	-0.580279	0.96511	110.09379	32.7	15.04
47	-0.21859	-0.01264	0.372114	-0.575893	0.96587	109.67068	32.7	15.04
48	-0.21984	-0.01071	0.369961	-0.523141	0.96594	109.15207	32.7	16.77
49	-0.22333	-0.01216	0.373485	-0.577788	0.96537	110.26700	32.7	14.91
50	-0.22253	-0.00816	0.376439	-0.584163	0.96516	110.18160	32.7	15.26
51	-0.24195	-0.04343	0.376837	-0.559506	0.96992	109.68077	32.7	12.70
52	-0.24172	-0.03030	0.374707	-0.559251	0.96989	109.55166	32.7	12.14
53	-0.22734	-0.01320	0.372707	-0.575487	0.96535	110.37099	32.7	14.90
54	-0.22747	-0.01361	0.370934	-0.536848	0.96595	109.27222	32.7	16.50
55	-0.22462	-0.01284	0.370491	-0.569671	0.96609	109.77936	32.7	14.62
56	-0.22245	-0.01025	0.372489	-0.578371	0.96596	109.74581	32.7	14.89

N°	HOMO(u.a)	LUMO(u.a)	La charge de H	La charge de O	La dist O-H	L'angleHOC	ϵ	Pka exp
57	-0.24189	-0.03254	0.373040	-0.574823	0.96627	109.27493	32.7	13.30
58	-0.22366	-0.01520	0.368602	-0.574469	0.96605	109.83334	32.7	14.52
59	-0.32118	-0.16601	0.410203	-0.506273	0.99069	107.41945	37.8	1.1
60	-0.24077	-0.03138	0.373631	-0.526714	0.97159	109.34832	37.8	13.85
61	-0.24077	-0.03138	0.373631	-0.526714	0.97159	109.34832	37.8	13.85
62	-0.28956	-0.11631	0.387293	-0.535796	0.97071	109.50518	37.8	5.72

Tableau III.4 Résultats de calcul des descripteurs

III.2.4 La Sélection des Descripteurs

Pour la construction du modèle QSPR de pKa des phénols dans différents solvants, la sélection des descripteurs a été réalisée en utilisant le logiciel Weka. Weka est un logiciel de fouille de données développé par l'Université de Waikato en Nouvelle-Zélande. Il propose une collection d'algorithmes d'apprentissage automatique pour les tâches de fouille de données, comprenant des outils de prétraitement, de classification, de régression, de clustering et de visualisation des données [3]. Dans notre étude, les méthodes CfsSubsetEval et BestFirst ont été utilisées pour identifier les descripteurs les plus pertinents pour prédire le pKa des phénols. Cette approche vise à sélectionner les descripteurs les plus pertinents afin d'obtenir des prédictions précises.

-La méthode CfsSubsetEval évalue la corrélation entre les descripteurs et la variable cible (le pKa), en tenant compte des interactions entre les descripteurs. Elle permet de déterminer quels descripteurs sont les plus informatifs et contribuent le plus à la prédiction du pKa [4].

-Ensuite, la méthode BestFirst est employée pour explorer de manière efficace l'espace des combinaisons possibles de descripteurs et sélectionner le sous-ensemble optimal. Elle effectue une recherche heuristique en évaluant les différentes combinaisons de descripteurs selon leur performance prédictive [5].

Le tableau ci-dessous présente la liste des descripteurs les plus pertinents obtenue après l'application de ces deux méthodes sur nos données.

Descripteur	Description
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
Dist O-H	La distance de la liaison O-H du groupement de phenol

Tableau III.5 Liste des descripteurs obtenus après le processus de sélection sous Weka

III.3 Développement du modèle QSPR

III.3.1 La répartition des données

L'ensemble des données utilisées dans cette étude a été divisé en deux parties distinctes : les données d'apprentissage et les données de test. Cette division vise à évaluer les performances et la capacité de généralisation des modèles développés. Les données d'apprentissage sont utilisées pour entraîner les modèles et ajuster leurs paramètres afin qu'ils puissent apprendre les relations entre les descripteurs moléculaires et les valeurs de pKa. Les données de test, quant à elles, servent à évaluer la performance des modèles sur des données indépendantes, c'est-à-dire des données qu'ils n'ont pas utilisées lors de l'apprentissage. Cette séparation des données en ensembles d'apprentissage et de test permet de fournir une estimation réaliste de la capacité des modèles à prédire le pKa des phénols dans les solvants organiques.

Dans notre travail, nous avons utilisé le logiciel Dataset Division 1.2 pour diviser l'ensemble des données en utilisant la méthode 'Activity/Property based'. Ce logiciel nous a permis de séparer les données en ensembles d'apprentissage et de test de manière efficace, en prenant en compte les propriétés ou les activités associées à chaque molécule [6].

	Nombre de molécules
Données d'apprentissage	50
Données de Test	12

Tableau III.6 La répartition des données.

De plus, dans la phase d'apprentissage les données ont également été divisées en deux parties, dont 90% des données ont été utilisées pour développer le modèle (l'apprentissage) et 10% ont été utilisées pour valider le modèle.

III.4 Les résultats des méthodes statistiques développés

Dans cette section, nous présentons les résultats obtenus en utilisant deux approches de modélisation différentes : MLR (Multiple Linear Regression) et RNA (Réseau de Neurones Artificiels). Nous avons appliqué ces méthodes sur les ensembles d'apprentissage et de test, tels que définis précédemment, afin de prédire le pKa des phénols dans les solvants organiques. Les performances des modèles ont été évaluées en comparant les prédictions avec les valeurs expérimentales de pKa. Nous discuterons des résultats obtenus et analyserons la précision et la fiabilité des deux approches de modélisation dans la prédiction du pKa des phénols.

a. La régression linéaire multiple MLR

Les résultats obtenus à l'aide de la méthode de régression linéaire multiple (MLR) pour prédire le pKa des phénols dans les solvants organiques sont prometteurs. Le modèle MLR développé a montré une bonne corrélation entre les descripteurs moléculaires sélectionnés et les valeurs expérimentales de pKa. L'équation 1 présente le modèle statistique développé:

$$\text{Pka} = 76.8995 * \text{HOMO} + 7.33 * \text{LUMO} - 208.3945 * \text{dis(O-H)} + 234.048 \quad (\text{eq1})$$

comme indiqué dans le **tableau III. 7**, les performances du modèle ont été évaluées à l'aide de métriques telles que l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2) et l'erreur absolue moyenne MAE. Ces résultats nous donnent confiance dans la capacité du modèle MLR à prédire de manière fiable le pKa des phénols.

	R^2	MAE	RMSE
Apprentissage	0,9455	1,52	2,1967
Test	0,9476	1,1286	1,5578

Tableau III. 7 Performance de la régression linéaire multiple sur les données d'apprentissage et les données de test.

Le tableau III.8 présente une évaluation globale du modèle de régression linéaire multiple sur l'ensemble des données d'apprentissage et de test. Ces résultats globaux démontrent la capacité du modèle MLR à prédire de manière satisfaisante le pKa des phénols dans les solvants organiques sur l'ensemble des données utilisées.

R²	MAE	RMSE
0.8298	1.7184	2.5266

Tableau III.8 Performance de la régression linéaire multiple sur les données globales.

Le tableau suivant (tableau III.9) présente une comparaison des valeurs prédites par le modèle MLR avec les valeurs expérimentales du pKa des phénols. Cette comparaison est essentielle pour évaluer la performance et l'exactitude du modèle dans sa capacité à prédire le pKa des phénols dans des conditions réelles.

N°	La valeur de pka exp	La valeur prédite	L'erreur
1	7.55	11.85	4.3
2	10.19	12.203	2.013
3	13.25	13.48	0.23
4	5.32	4.466	-0.854
5	4.09	5.464	1.374
6	7.32	9.75	2.43
7	5.17	9.391	4.221
8	11.54	12.416	0.876
9	4.82	4.411	-0.409
10	7.6	7.3	-0.3
11	16.85	14.084	-2.766
12	14.9	15.358	0.458
13	11	7.395	-3.605
14	13.56	12.525	-1.035
15	7.97	8.122	0.152
16	13.09	13.578	0.488
17	10.6	9.275	-1.325
18	15.14	13.213	-1.927
19	1.16	10.857	9.697
20	9.8	13.799	3.999
21	15.88	13.912	-1.968
22	15.72	15.423	-0.297
23	16.86	15.047	-1.813
24	14.3	12.826	-1.474

N°	La valeur de pka exp	La valeur prédite	L'erreur
25	13.75	11.555	-2.195
26	9.3	10.215	0.915
27	13.68	13.181	-0.499
28	15.5	14.514	-0.986
29	3.51	4.773	1.263
30	16.1	14.398	-1.702
31	13.01	12.566	-0.444
32	17.58	13.923	-3.657
33	16.96	7.282	-9.678
34	11	11.199	0.199
35	15.08	15.51	0.43
36	15.04	0.824	0.824
37	16.77	15.767	-1.003
38	14.91	15.607	0.697
39	15.26	15.742	0.482
40	12.7	12.998	0.298
41	12.14	13.118	0.978
42	14.9	15.295	0.395
43	16.5	15.157	-1.343
44	14.62	15.353	0.733
45	14.89	15.566	0.676
46	14.52	15.418	0.898
47	1.11	1.678	0.568
48	13.85	12.829	-1.021
49	14.31	12.103	-2.207
50	5.72	8.638	2.918
51	10.97	12.284	1.314
52	-0.3	1.928	2.228
53	9	12.331	3.331
54	4.59	7.663	3.073
55	12.58	13.02	0.44
56	14.22	13.696	-0.524
57	14.76	12.289	-2.471
58	15.3	15.161	-0.139
59	7.05	11.393	4.343
60	16.47	14.695	-1.775
61	15.04	16.384	1.344
62	13.3	13.843	0.543

Tableau III.9 Comparaison des données expérimentales avec les données prédites par le modèle basé sur la régression linéaire.

b. Perceptron multicouche

Selon les résultats présentés dans le **Tableau III.10**, le modèle MLP démontre de bonnes performances avec des coefficients de corrélation élevés (R) à la fois dans l'ensemble d'apprentissage et dans l'ensemble de test. Ces résultats indiquent que le modèle

MLP est non seulement performant lors du développement des modèles, mais il offre également une excellente capacité de prédiction.

	R²	MAE	RMSE
Apprentissage	0,9228	3,677	4,2514
Test	0,9537	1,4846	2,0021

Tableau III.10 Performance de perceptron multicouche sur les données d'apprentissage et les données de test.

Le tableau III.11 présente les résultats de l'évaluation globale du modèle MLP.

R²	MAE	RMSE
0.849	2.1583	2.8231

Tableau III.11 Performance de perceptron multicouche sur les données globales.

Ces résultats globaux confirment que le modèle MLP présente de bonnes performances dans la prédiction du pKa, avec des valeurs de RMSE, R² et MAE indiquant une précision et une adéquation satisfaisantes.

N°	Valeur expérimentl	Valeur prédite	L'erreur
1	7.55	9.322	1.772
2	10.19	9.85	-0.34
3	13.25	11.726	-1.524
4	5.32	1.815	-3.505
5	4.09	2.639	-1.451
6	7.32	8.495	1.175
7	5.17	6.077	0.907
8	11.54	10.143	-1.397
9	4.82	2.158	-2.662
10	7.6	5.597	-2.003
11	16.85	12.518	-4.332
12	14.9	14.094	-0.806
13	11	5.38	-5.62
14	13.56	10.24	-3.32
15	7.97	9.661	1.691
16	13.09	11.846	-1.244
17	10.6	7.08	-3.52
18	15.14	11.394	-3.746
19	1.16	8.236	7.076

N°	Valeur expérimentl	Valeur prédite	L'erreur
20	9.8	12.148	2.348
21	15.88	12.299	-3.581
22	15.72	14.459	-1.261
23	16.86	13.974	-2.886
24	14.3	10.672	-3.628
25	13.75	9.054	-4.696
26	9.3	7.647	-1.653
27	13.68	11.304	-2.376
28	15.5	13.274	-2.226
29	3.51	2.499	-1.011
30	16.1	13.096	-3.004
31	13.01	10.313	-2.697
32	17.58	12.308	-5.272
33	16.96	6.914	10.046
34	11	8.599	-2.401
35	15.08	14.578	-0.502
36	15.04	14.982	-0.058
37	16.77	14.871	-1.899
38	14.91	14.723	-0.187
39	15.26	14.889	-0.371
40	12.7	11.074	-1.626
41	12.14	11.212	-0.928
42	14.9	14.321	-0.579
43	16.5	14.13	-2.37
44	14.62	14.378	-0.242
45	14.89	14.64	-0.25
46	14.52	14.464	-0.056
47	1.11	0.316	-0.794
48	13.85	10.824	-3.026
49	14.31	10.344	-3.966
50	5.72	5.883	0.163
51	10.97	9.97	-1
52	-0.3	0.445	0.745
53	9	10.089	1.089
54	4.59	5.409	0.819
55	12.58	10.982	-1.598
56	14.22	12.031	-2.189
57	14.76	9.941	-4.819
58	15.3	14.13	-1.17
59	7.05	8.653	1.603
60	16.47	13.494	-2.976
61	15.04	15.581	0.541
62	13.3	12.226	-1.074

Tableau III.12 Comparaison des données expérimentales avec les données prédites par le modèle basé sur le perceptron multichouches.

Le **tableau III.12** présente une comparaison entre les valeurs réelles du pKa des phénols et les valeurs prédites par le modèle. On peut observer que les valeurs prédites du pKa sont très proches des valeurs réelles pour chaque molécule. Ces résultats renforcent la confiance dans l'exactitude et la performance du modèle dans la prédiction du pKa des phénols dans différents solvants.

III.5 Discussion

Les résultats obtenus à partir des évaluations du modèle MLR et MLP montrent des performances prometteuses dans la prédiction du pKa des phénols dans les solvants organiques. Les valeurs élevées du coefficient de détermination (R^2) indiquent une bonne adéquation du modèle aux données expérimentales, avec une capacité d'expliquer respectivement 94.76% pour le modèle MLR et 95.37% pour le modèle MLP sur les données de test.

Le modèle MLR, basé sur une approche statistique, présente une précision raisonnable avec une adéquation satisfaisante aux données. En revanche, le modèle MLP, basé sur un réseau de neurones artificiels, offre des performances encore meilleures, avec une plus grande capacité prédictive et une meilleure adéquation aux données.

Cependant, il est important de continuer à améliorer et à valider ces modèles en utilisant des ensembles de données plus vastes et diversifiés, ainsi que d'explorer d'autres méthodes de modélisation pour une meilleure prédiction des propriétés chimiques.

En conclusion, les résultats obtenus avec les modèles MLR et MLP démontrent leur potentiel dans la prédiction du pKa des phénols dans les solvants organiques. Ces modèles offrent une approche précise et prometteuse pour l'estimation de cette propriété chimique essentielle, ouvrant ainsi la voie à des applications dans la conception et la recherche de nouvelles molécules.

III.6 Conclusion

En conclusion, le chapitre a présenté les résultats de l'application des méthodes statistiques MLR et MLP pour la prédiction du pKa des phénols dans les solvants organiques. Les performances des deux modèles ont été évaluées à l'aide de différentes métriques telles que le coefficient de détermination (R^2), l'erreur moyenne absolue (MAE) et le Root Mean Square Error (RMSE).

Les résultats obtenus ont démontré la capacité des modèles à prédire avec précision le pKa des phénols. Le modèle MLP, basé sur un réseau de neurones artificiels, a affiché des performances supérieures, avec des valeurs de R^2 élevées et des erreurs moyennes faibles. Cela souligne l'importance des méthodes d'apprentissage automatique pour la modélisation de relations complexes entre les descripteurs moléculaires et les propriétés chimiques.

Cette étude a démontré l'efficacité des méthodes statistiques MLR et MLP ouvrant ainsi la voie à de futures recherches et applications dans le domaine de la chimie computationnelle et de la modélisation moléculaire.

Les références bibliographiques

- [1] LI, W., Xie, H., Huang, Y., Song, L., Shao, Y., & Qiu, K. (2016). Application of Gaussian 09/GaussView 5.0 in analytical chemistry teaching. *Journal of Kunming Medical University*, 134-136.
- [2] Frisch,A.(2009).gaussian09W Referenc.wallingford,USA,25p,470.
- [3] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, IH et Trigg, L. (2010). Weka-un atelier d'apprentissage automatique pour l'exploration de données. *Manuel d'exploration de données et de découverte de connaissances* , 1269-1277.
- [4] Abdi, M. K., & Lounis, H. Analyse de l'impact du changement: approche et étude de cas.
- [5] Cung, V. D., Hifi, M., & Le Cun, B. (2000). Constrained two-dimensional cutting stock problems a best-first branch-and-bound algorithm. *International Transactions in Operational Research*, 7(3), 185-210.
- [6] http://teqip.jdvu.ac.in/QSAR_Tools/

Conclusion Générale

Conclusion générale

Conclusion Générale

Dans ce travail de développement de modèles QSPR pour la prédiction du pKa de 62 molécules phénoliques en utilisant des descripteurs quantiques et des techniques d'apprentissage automatique a abouti à des résultats prometteurs. Les modèles construits à l'aide de la régression linéaire multiple (MLR) et des perceptrons multichouches (MLP) ont démontré de bonnes performances prédictives.

Les résultats obtenus ont montré que les descripteurs quantiques sélectionnés étaient pertinents pour capturer les informations structurelles des molécules, permettant ainsi d'établir des relations significatives avec les valeurs de pKa. Les modèles MLR et MLP ont été capables de prédire les valeurs de pKa des molécules étudiées, démontrant leur utilité dans le domaine de la prédiction des propriétés moléculaires.

Cependant, il est important de noter que des améliorations et des extensions peuvent être envisagées pour les travaux futurs. Tout d'abord, l'ajout de nouveaux descripteurs moléculaires ou l'utilisation d'autres techniques d'apprentissage automatique plus avancées pourrait améliorer encore les performances prédictives des modèles. De plus, l'extension de l'étude à un plus grand nombre de molécules phénoliques ou à d'autres types de molécules pourrait permettre de généraliser davantage les résultats.

En outre, il serait intéressant d'explorer d'autres propriétés moléculaires et d'étendre l'application des modèles QSPR développés à d'autres domaines de la chimie et de la pharmacologie. Par exemple, la prédiction de la toxicité ou de l'activité biologique des molécules pourrait être abordée à l'aide de la même approche.