Université Mohamed Khider de Biskra
Faculté des Sciences Exactes et Sciences de la Nature et de la Vie
Département des Sciences de la Matière

MÉMOIRE DE MASTER

Domaine : Sciences de la matière
Filière : Chimie
Spécialité : Chimie pharmaceutique

Réf. : /

Présenté par :
KEBAIRI Maroua **OMRI Maroua**

Modélisation par les réseaux de neurones artificiels : application QSAR

Jury :

BOUMEDJANE Youcef	MCA	Université Med Khider de Biskra	Président
KENOUCHE Samir	MCA	Université Med Khider de Biskra	Rapporteur
LEHRAKI Faiza	MAA	Université Med Khider de Biskra	Examinatrice

Année universitaire : 2020-2021

Table des matières

Chapitre 2 Outils de la modélisation

Chapitre 3 Traitement, analyse et interprétation des données

Liste des figures

Chapitre 1

Chapitre 2

Chapitre 3

Liste des tableaux

Chapitre 1

Chapitre 2

Chapitre 3

Liste des abréviations

Les différentes abréviations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

ADME	Administration Distribution Métabolisme Elimination
ANOVA	Analysis Of Variation (analyse de la variance)
ATP	Adénosine TriphosPhate
B3LYP	Becke 3-Parameter Lee-Yang-Parr
BIC	Bayesian Information Criterion (Critère d'information Bayésian)
CME	Carré Moyen Expliqué
CMR	Carré Moyen Résiduel
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Indices Analysis
COV	Covariance
DCMU	DiChlorophényle diMéthylUrée
DDL	Degré De Liberté
DFT	Density Function Theory
DL_{50}/CL_{50}	Dose/Concentration létale médiane
DM	Dipole Moment (Moment Dipolaire)
IA	Intelligence Artificiel
IUPAC	International Union of Pure and Applied Chemistry
MCO	Moindres Carrés Ordinaires
MLP	Multilayer Perceptron
MSE	Mean Square Error (L'erreur quadratique moyenne)
NADPH	Nicotinamide Adénine Dinucléotide Phosphate-H
PI/PII	Photosystème I/II
PMC	Perceptron Multi-Couche
PPP	Produits phytopharmaceutiques
QSAR	Quantitative Structure-Activity Relationships
RLM	Régression linéaire multiple
RNA	Réseaux de Neurones Artificiels
SCE	Somme des Carrés Expliqués
SCR	Somme des Carrés Résiduelles
SCT	Somme des Carrés Totales
Se	Standard Error (écart-type)
VarCov	Variance-Covariance
XOR	Exclusif OR (OU exclusif)

Remerciements

Au nom de DIEU Le Plus Clément et Le Plus Miséricordieux.

Tout d'abord, nous remercions ALLAH le Tout Puissant qui nous a accordé la volonté et le courage pour réaliser ce mémoire.

Nos plus sincères remerciements vont à M. KENOUCHE Samir notre promoteur de mémoire, pour sa disponibilité, ses corrections, ses conseils judicieux et pour nous avoir guidés tout au long de ce travail.

Merci au Professeur BOUMEDJANE Youcef, pour avoir accepté de présider le jury de ce mémoire. On tient à remercier Mme. LEHRAKI Faiza, pour avoir bien voulu participer au jury de ce mémoire et accepter d'être examinatrice.

Un grand merci à ceux qui ont été à nos côtés et à ceux qui ont cru en nous.

Dédicace

To my family, especially to my parents.

MAROUA. K

To my parents and my beloved ones.

MAROUA. O

Introduction générale

Actuellement dans le domaine agricole, la majorité des cultures sont menacées par des mauvaises herbes. Nous comprenons aisément que la lutte contre ces mauvaises herbes demeure une étape cruciale pour de meilleures performances agricoles.

La chimie peut désormais traiter la plupart des problèmes de protection des cultures à l'aide des produits phytopharmaceutiques (PPP). Ces produits sont très importants et jouent un rôle très influent dans notre environnement. Parmi les PPP qui ont été étudiés dans cette étude sont les herbicides de type inhibitrice du Photosystème II (PS II). Ces herbicides sont abondamment utilisés dans les pratiques phytosanitaires.

Pour toutes ces raisons, la chimie moderne s'oriente vers de nouvelles méthodes de recherche, qui consistent à prédire les propriétés et les activités de ces molécules avant même que celles-ci ne soient synthétisées. Tout cela grâce à la Chimie Computationnelle, encore appelée Chimie Numérique. La relation quantitative structure-activité (QSAR) est l'un des moyens les plus utilisables en physico-chimie computationnelle. Elle est considérée comme une approche de « chimie verte » puisque les déchets chimiques ne sont pas générés lors de la réalisation de prévisions *in silico*. Par conséquent, les études QSAR sont incontestablement d'une grande importance en chimie de façon générale.

Dans ce travail de Master, il s'agit de développer une architecture neuronale robuste non linéaire et un modèle de régression multiple pertinent afin de modéliser et de prédire l'activité biologique de certaines molécules organiques herbicides dérivées d'anilide. La question qui se pose est : Quels sont les modèles QSAR qui donnent les bonnes performances, en terme du pouvoir prédictif ?

Ce thème s'articule autour de trois chapitres. Dans le premier chapitre, nous avons réalisé une enquête exploratoire sur les herbicides en question, leur mode action contre le photosystème II ainsi que leur toxicité. Par la suite, nous présenterons un aperçu sur l'étude QSAR, sa méthodologie ainsi que certains domaines d'application de cette approche.

Le deuxième chapitre sera consacré aux fondements théoriques des méthodes statistiques utilisées dans le cadre de ce travail.

Le troisième chapitre portera sur l'interprétation et l'analyse des différents résultats issus des recherches engagées dans le cadre de ce travail de Master. Nous dresserons une étude comparative entre les deux méthodes statistiques, pour déterminer les performances prédictives des modèles obtenus. Finalement, nous concluons ce travail avec quelques avantages et inconvénients pour les deux méthodes appliquées.

Chapitre 1

Généralités

Ce chapitre fournit une introduction à la nature des données que nous allons traiter, à la méthodologie générale de l'étude QSAR et résume ce qui est inclus dans la portée de cette étude.

Les végétaux sont des organismes autotrophes, ils synthétisent la matière organique à partir des composés inorganiques grâce à l'énergie de soleil, c'est la photosynthèse. Ce processus a lieu dans des organites spécialisées nommées chloroplastes (Fig 1.1). Chaque chloroplaste est constituée d'une membrane externe et d'une membrane interne repliées sur elles-mêmes en formant des thylakoïdes, les thylakoïdes ressemblent à des sacs à l'intérieur desquels se trouvent l'espace intrathylakoïdien, un empilement de thylakoïde se nomme granum ou grana au pluriel, ces grana baignent dans le stroma.

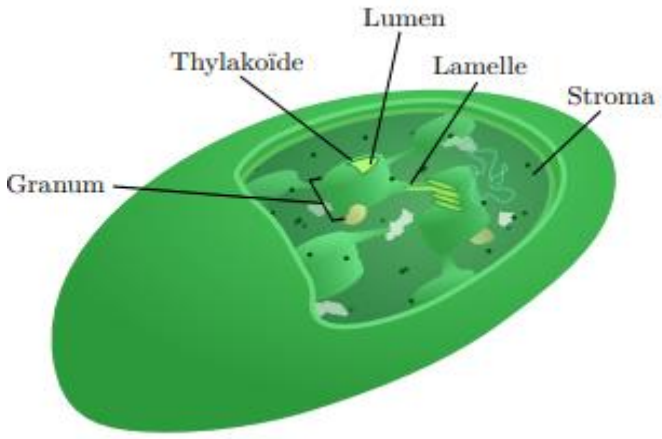


Figure 1.1 : Schéma de Chloroplaste [1]

Les membranes des thylakoïdes possèdent des pigments photosynthétiques groupés en photosystèmes, PI et PII, absorbent des longueurs d'ondes légèrement différentes, chaque photosystème comprend une antenne qui capte l'energie des photons et une centre réactionnelle qui transmet cette énergie en accepteur primaire des électrons. L'antenne de PII absorbe l'énergie des photons et la transmet au centre réactionnel, ce dernier céde deux électrons à l'accepteur primaire qui les transmet à une chaine de transporteurs d'électrons qui, à son tour, les dérige au PI. La présence de deux quinones successives Q_A et Q_B comme accepteurs d'électrons terminaux du centre réactionnel assure la processus de transfert d'électrons. Certains herbicides agissent comme inhibiteurs du PII en empêchant le transfert d'électrons entre les deux quinones [2]. On distingue plusieurs groupes d'herbicides ayant ce mode d'action, parmi ces inhibiteurs, dérivés à partir d'anilides, d'urées substituées, de triazines, d'uraciles, de pyridazinones, de triazinones et autres, sont de nombreux herbicides commerciaux. Ce groupe est mieux représenté par DCMU (le diuron, dichlorophényle diméthylurée) car ce composé est utilisé le plus souvent dans la recherche sur la photosynthèse [3].

Dans ce travail on propose une modélisation par QSAR et analyse d'un ensemble de données de 76 molécules dérivées de l'anilide meta et parasubstituées comme herbicides inhibiteurs du photosystème II.

Le choix de la base de données expérimentales de référence est décisif dans une étude QSAR. Elle doit être composée de données expérimentales fiables obtenues en suivant un protocole expérimental unique. En effet, la robustesse du modèle dépend fortement de la base sur lequel il se fonde [4]. La méthode QSAR inclut toutes les méthodes statistiques par lesquelles des activités biologiques sont reliées avec les éléments structuraux (analyse de Free Wilson), les propriétés physico-chimiques (analyse de Hansch) ou différents paramètres liés à la notion de champ aidant à la description de la structure (3D QSAR) (Fig 1.2) [5]. Le lien entre les descripteurs et l'activité se fait grâce à les outils de modélisation, comme la régression linéaire simple et multiple et les réseaux de neurones artificiels.

Enfin, les informations extraites à partir des résultats d'étude de QSAR peuvent être utilisées pour prévoir les propriétés physicochimiques et les activités biologiques de nouveaux composés ainsi que pour concevoir de nouvelles structures [5].

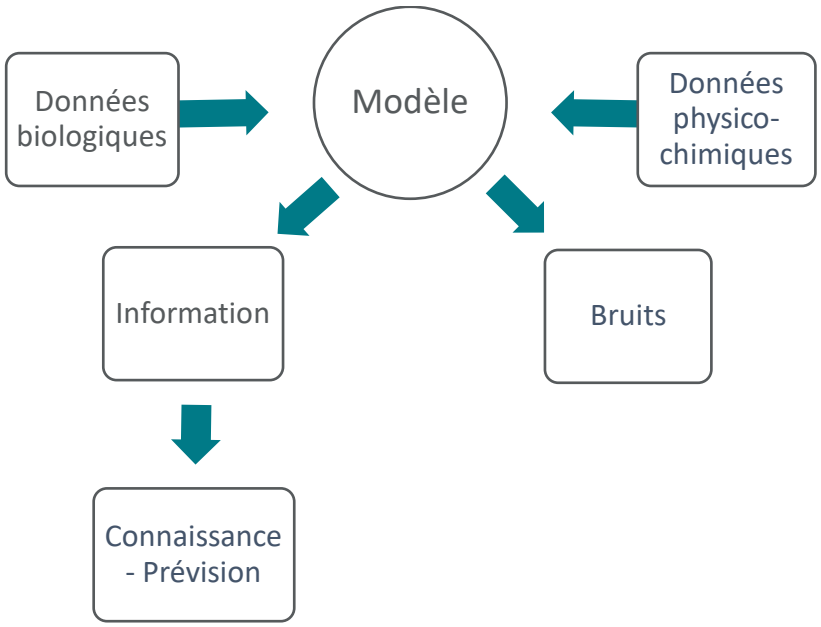


Figure 1.2 : Etude QSAR et son application [5]

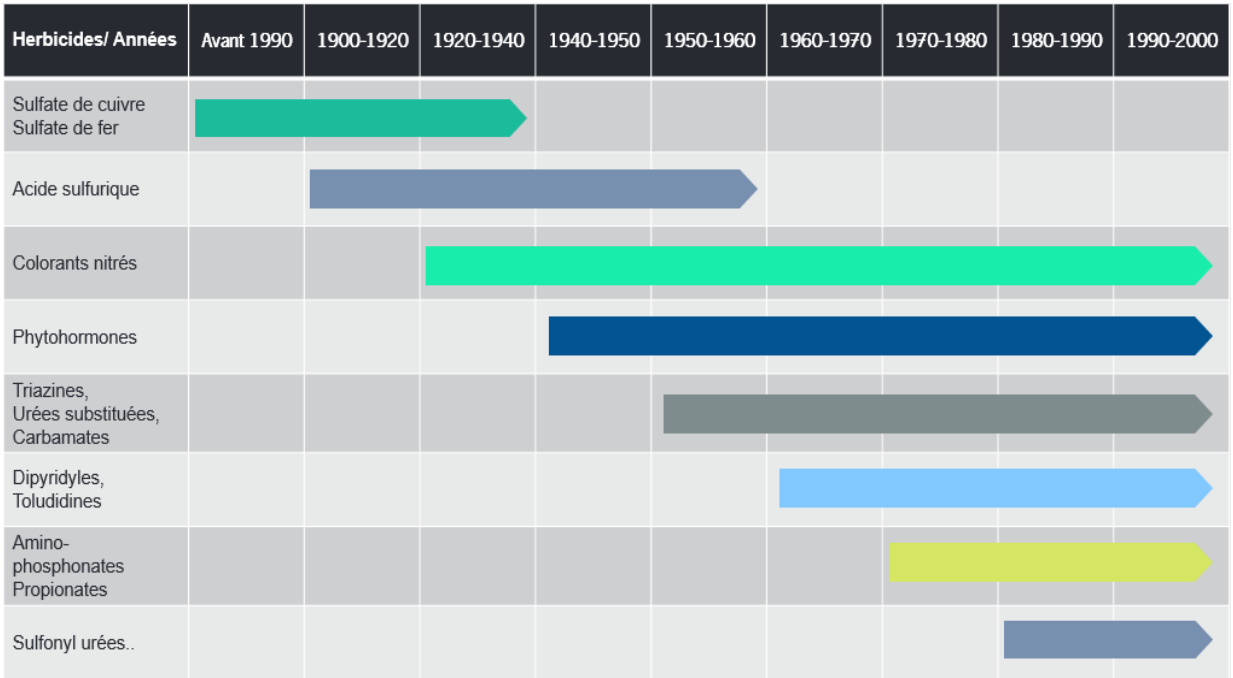
1. Herbicides

1.1 Aperçu historique

Les herbicides sont des substances ayant la capacité d'éliminer les adventices (ou mauvaises herbes). Leur emploi ne se limite pas au domaine agricole, ils sont utilisés aussi bien pour la protection des cultures que pour le confort (jardinage, entretien des villes, des voies ferrées…). Ce large spectre d'usage donne à ces molécules un caractère ubiquitaire.

Comme tous les pesticides, les herbicides ont connu un très grand développement depuis le XIXe siècle (Tab 1.1). A partir des années 1950, le marché des herbicides a reçu un grand nombre de nouvelles molécules (toluidines, aminosulphonates, triazines...). Les molécules synthétiques ont remplacé les substances naturelles telles que le soufre depuis les années 1990. La composition chimique des herbicides de synthèse est souvent très complexe ce qui rend difficile la compréhension de leur danger sur l'environnement et la prévision de leur devenir. Actuellement, les herbicides occupent la deuxième position après les fongicides, en matière de consommation dans le monde bien que les statistiques indiquent que les taux de consommation soient en baisse de 1,7% par rapport aux années précédentes.

Tableau 1.1 : Évolution chronologique de l'utilisation des herbicides [6]



Parmi l'arsenal des molécules actives à effet herbicide existant actuellement sur le marché, certaines exercent un désherbage total, d'autres un désherbage sélectif. Celles à effet total éliminent toute la végétation qui leur est exposée tandis que les molécules sélectives n'éliminent en général qu'une ou deux espèces définies de mauvaises herbes sans pour autant affecter grandement les cultures. Ce dernier type d'herbicides est le plus utilisé en agriculture [6].

1.2 Définition

Les herbicides sont des pesticides utilisés pour détruire la végétation indésirable, en particulier les divers types de mauvaises herbes, les graminées et les plantes ligneuses [7].

Ils sont des composés chimiques qui, après absorption par les plantes, par voie racinaire ou foliaire, sont responsables de la mort des végétaux. Très utilisés en agriculture intensive, distinés à la destruction des mauvaises herbes responsables d'une diminution importante de la productivité, qui peut atteindre 50% en l'absence totale de traitement.

Les herbicides sont en général des composés stables, non biodégradables. Leur action doit être sélective et ne s'exerce que sur les mauvaises herbes.

Le mode d'action des herbicides varie selon [2] :

- Inhibition de l'activité photosynthétique à la suite d'un blocage des photosystèmes I ou II ;
- Blocage des voies de biosynthèse de certains composés comme les acides aminés, les lipides, etc ;
- Inhibition de certaines enzymes comme la glutamine synthétase.

1.3 Composition et formulation

La formulation d'un herbicide se réfère au matériau dont il est transporté dans ou sur, et sa concentration dans ce support. Les formulations d'herbicides contiennent des ingrédients actifs (les composants de la formulation d'herbicide responsables d'être phytotoxique pour la mauvaise herbe [9]) et des ingrédients inertes. [8] Les ingrédients inertes comprennent autre chose que l'ingrédient actif et peuvent généralement être divisés en deux catégories, y compris les solvants et les adjuvants. Les solvants courants comprennent des substances telles que l'eau, les distillats de pétrole et l'argile.

Un adjuvant ou formulant est un additif (généralement en quantités relativement faibles par rapport au support) qui améliore l'application, les performances, la sécurité, le stockage ou la manipulation d'un ingrédient actif [9].

La formulation correspond à la forme physique sous laquelle le produit phytopharmaceutique est mis sur le marché. Les plus couramment répandues sont les suivantes [10] :

- Pour les formulations solides :
  - les granulés solubles,
  - les poudres mouillables.
- Pour les formulations liquides :
  - les concentrés solubles, composés de produits solubles dans l'eau,
  - les concentrés émulsionnables, composés de produits liquides en émulsion dans le produit,
  - les suspensions concentrées, composées de particules solides dispersées dans le produit.

1.4 Action des herbicides : La photosynthèse et l'inhibition photosynthétique

La photosynthèse est le processus photochimique qui permet aux organismes chlorophylliens de synthétiser de la matière organique à partir de dioxyde de carbone, d'eau et d'énergie lumineuse. Le bilan global de la photosynthèse dioxygénique est décrit par l'équation suivante [1] :

$$6\,CO_2 + 12\,H_2O \xrightarrow{h\nu} C_6H_{12}O_6 + 6\,O_2 + 6\,H_2O$$

La photosynthèse comporte deux étapes principales : la phase lumineuse (phase claire) et la phase métabolique (phase sombre). La phase lumineuse capte l'énergie de la lumière solaire et la stocke sous forme de molécules énergétiques ATP et NADPH. Dans la deuxième phase indépendante de la lumière (phase sombre), l'ATP et le NADPH produits au première phase sont utilisés pour convertir le dioxyde de carbone atmosphérique en molécules de sucre. Les réactions dépendant de la lumière impliquent deux complexes protéiques, à savoir le photosystème I et le photosystème II (Fig 1.3).

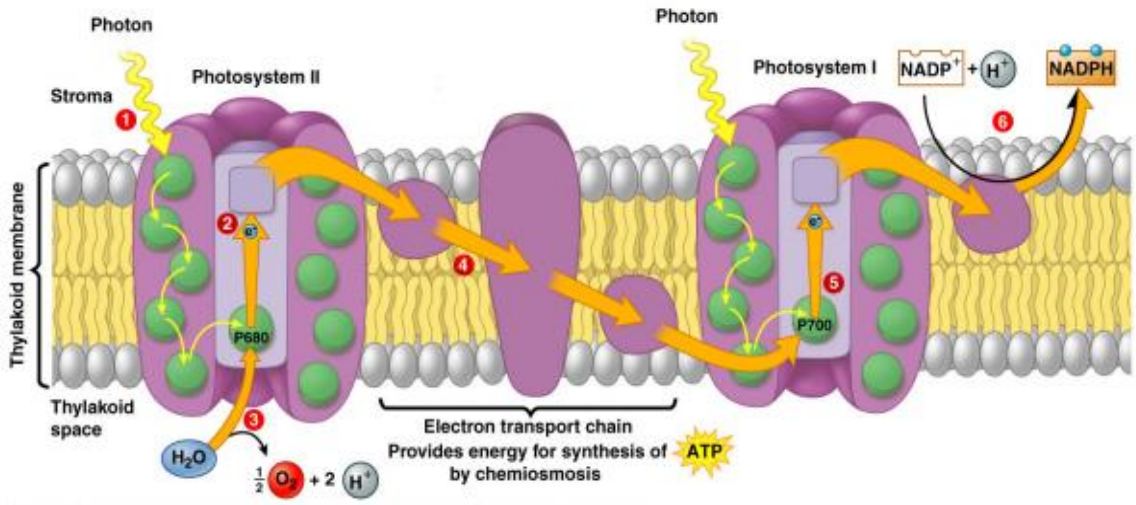
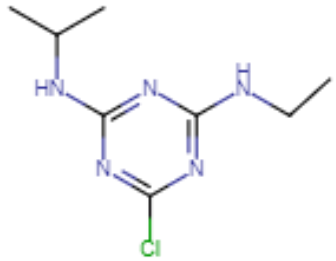
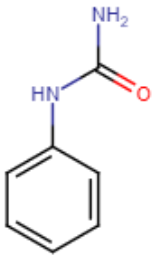
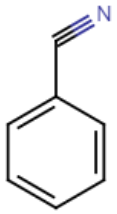
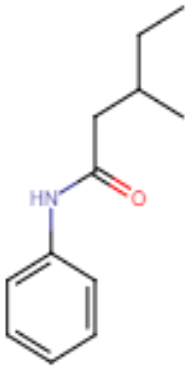
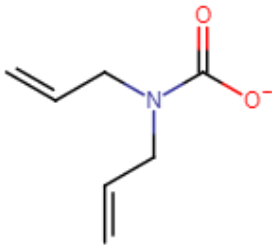


Figure 1.3 : Appareil photosynthétique et ses sous-unités : photosystème II et photosystème I [1]

Le PS II utilise des photons de la lumière du soleil et excite une paire d'électrons qui sont transférés à travers une chaîne de porteurs d'électrons. Une molécule d'eau est divisée dans le processus et les électrons perdus sont remplacés par les électrons de l'eau. L'énergie des électrons excités est utilisée pour créer un gradient de protons avec les ions hydrogène générés lors du fractionnement de l'eau. Ce gradient est utilisé pour générer de l'ATP. Dans le PS I, les électrons sont à nouveau excités par la lumière et ces électrons sont transférés via une chaîne de transport d'électrons vers le NADP $^{+}$ où des ions hydrogène sont également ajoutés pour former le NADPH. Les électrons transférés dans PS I sont remplacés par les électrons excités de PS II. L'ATP et le NADPH ainsi produits sont d'avantage utilisés dans la phase indépendante de la lumière pour synthétiser des molécules de sucre en utilisant du dioxyde de carbone atmosphérique.

Les herbicides qui inhibent la photosynthèse agissent en bloquant les réactions lumineuses, empêchant ainsi la conversion de cette énergie lumineuse en sa forme chimique. Plus d'herbicides agissent sur le processus photosynthétique que sur tout autre processus biochimique. Les herbicides qui inhibent le PS II bloquent le transport d'électrons en se liant à des sites adjacents sur des protéines de quinone qui aident à former la chaîne de transport. Ces inhibiteurs comprennent les triazines, les phénylurées, les uraciles, les hydrobenzo-nitriles, les acylanilides et les bis-carbamates [11].

		
Atrazine 6-chloro-1,3,5-triazine-2,4-diamine.	**Uracile** 2,4(1H,3H)-Pyrimidine-dione	**Phénylurée**
		
Benzonitrile	**Anilide** 2-butyl acetanilide	**Bis-carbamate** N,N-Bis(prop-enyl) carba-mate

Une structure chimique commune aux différents composés a été identifiée : X = C-N

Cette séquence assure l'interaction d'atomes des composés avec certains atomes de la structure de quinone Q_B du PSII ; ainsi, X peut être un oxygène établissant une liaison hydrogène d'un groupe OH appartenant à la quinone ou l'atome d'hydrogène d'un groupe NH comme présenté dans la figure suivante [2] :

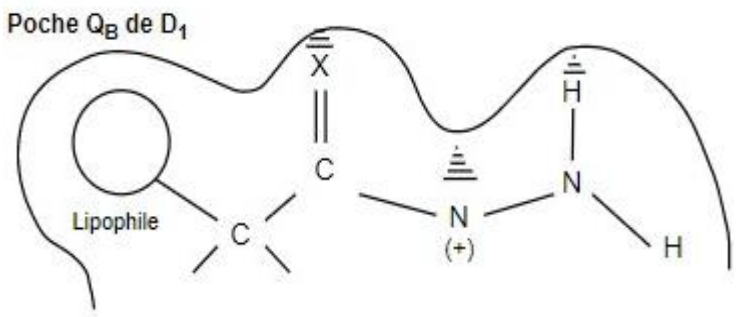


Figure 1.4 : Mode d'action des inhibiteurs de PS II [2]

1.5 Toxicité des herbicides

1.5.1 Phrases de risques

Il faut souligner que le risque toxicologique est résultant du danger présenté par le produit et de l'exposition de l'homme à ce produit. Ce danger résulte du fait qu'une molécule donnée, eu égard à sa structure, peut déclencher des effets néfastes connus à certaines concentrations sur la matière vivante, en particulier sur l'homme.

On remarquera que les phrases R dites de risque, portées obligatoirement sur les étiquettes des emballages, sont en réalité des phrases de danger.

En ce qui concerne les herbicides, le tableau (Tab 1.2) montre : 20 phrases de risque utilisés pour les substances actives, 22 pour les spécialités commerciales. En plus de ces phrases R, s'ajoutent quatre mentions particulières liées à l'environnement classification N (dangereux pour l'environnement) [12].

La figure (Fig 1.5) précise les 19 phrases caractérisant les dangers toxicologiques. Le tableau (Tab 1.2) reprend l'ensemble de ces phrases concernant les substances actives et les spécialités commerciales :

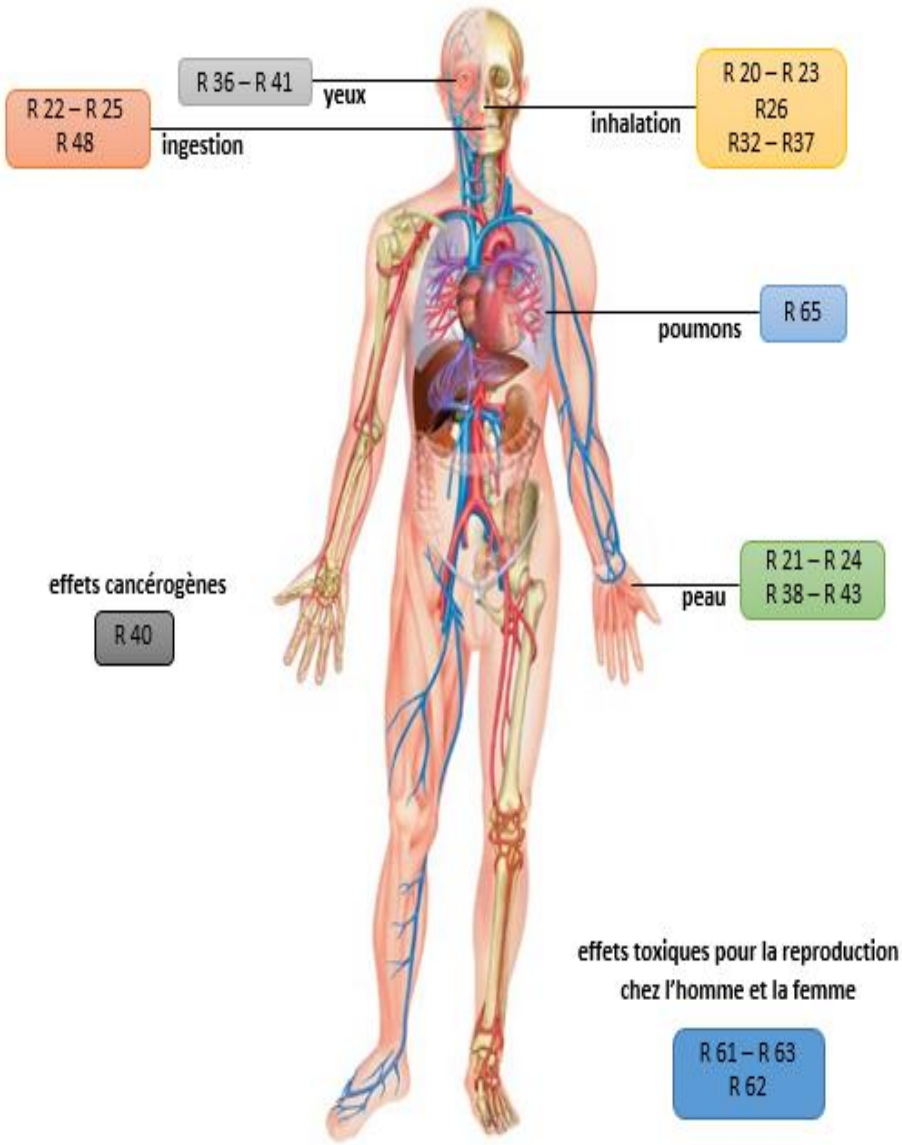


Figure 1.5 : Représentation des phrases de risque caractérisant les dangers toxicologiques [12]

Légende : phrases de risques et mentions concernant les dangers :

Physicochimiques Toxicologiques Ecotoxicologiques

Tableau 1.2 : Phrases de risques concernant les substances actives et débrouissaillantes [12]

Phrases R	Libellé	Substances actives	Spécialités commerciales
R 10	Inflammable		
R 11	Facilement inflammable		
R 20	Nocif par inhalation		
R 21	Nocif par contact avec la peau		
R 22	Nocif en cas d'ingestion		
R 23	Toxique en cas d'inhalation		
R 24	Toxique par contact avec la peau		
R 25	Toxique en cas d'ingestion		
R 26	Très toxique par inhalation		
R 32	Au contact d'un acide, dégage un gaz très toxique		
R 36	Irritant pour les yeux		
R 37	Irritant pour les voies respiratoires		
R 38	Irritant pour la peau		
R 40	Effet cancérogène suspecté : preuves insuffisantes		
R 41	Risque de lésions oculaires graves		
R 43	Peut entraîner la sensibilisation par contact avec la peau		
R 48	Risque d'effets graves pour la santé en cas d'exposition prolongée		
R 50	Très toxique pour les organismes aquatiques		
R 51	Toxique pour les organismes aquatiques		
R 52	Nocif pour les organismes aquatiques		
R 53	Peut entraîner des effets néfastes à long terme pour l'environnement aquatique		
R 54	Toxique pour la flore		
R 61	Risque pendant la grossesse d'effets néfastes pour l'enfant		
R 62	Risque possible d'altération de la fertilité		
R 63	Risque possible pendant la grossesse d'effets néfastes pour l'enfant		
R 65	Peut provoquer une atteinte des poumons en cas d'ingestion		
AQUA	Dangereux pour les organismes aquatiques		
DAFT	Dangereux pour la faune terrestre		
FAUN	Dangereux pour la faune aquatique		
POIS	Dangereux pour les poissons		

1.5.2 Différents niveaux de risques

On distingue, la toxicité pour l'homme (applicateur, public, personnels assurant la fabrication de la spécialité, voire le consommateur de produits récoltés dans la zone traitée), pour les animaux domestiques et l'écotoxicité pour le milieu. On discerne en outre [13] :

- La toxicité aiguë qui se traduit par des effets pathologiques après une absorption unique de produit ;
- La toxicité chronique qui résulte d'absorptions ou contacts répétés de petites quantités sans trouble immédiatement décelable : les effets nocifs ne se constatent qu'à long terme.

1.5.2.1 Toxicité aiguë

A. Toxicité aiguë orale

Elle s'exprime par la dose létale orale 50% (DL_{50}%). C'est la plus connue. Elle est généralement déterminée chez le rat. Elle correspond à la quantité de produit qui, ingérée en seule fois, entraîne la mort de 50% des individus en expérimentation. Elle s'exprime en mg/kg de poids corporel. Plus la DL_{50} est basse, plus le produit est toxique. Les anciens herbicides de la famille phénols nitrés présentaient notamment les DL_{50} les plus basses concernant les herbicides.

Pour les plus basses valeurs de DL_{50}, on retrouve les substances actives classées T (toxique) et T+ (très toxique) [12] :

- Ioxynil 110 mg/kg
- Paraquat 157 mg/kg
- Bromoxynil 190 mg/kg
- Diquat 231 mg/kg

B. Toxicité aiguë dermique

Elle résulte d'une absorption du produit toxique par la peau en un seul contact. Comme la DL_{50} orale, la DL_{50} cutanée qui en résulte s'exprime en mg/kg de poids corporel. Pour les herbicides cette donnée de toxicologie est peu souvent avancée mais il est vrai que, hormis les phénols nitrés désormais interdits, peu d'herbicides présentent un danger majeur par intoxication cutanée [12].

C. Toxicité aiguë respiratoire

Cette toxicité par voie respiratoire concerne les substances susceptibles de se trouver dans l'atmosphère ambiante soit sous forme de vapeur, soit sous forme de très fines gouttelettes en suspension dans l'air. La CL_{50} ou concentration létale 50% s'exprime en mg de substance par litre d'air. Seule quelque substances actives sont réputées dangereuses par inhalation [12] :

- Substances classés R26 : paraquat
- Substances classés R20 : acétochlore, cléthodime, clomazone.
- Substances classés R37 : acétochlore, paraquat, diquat.

1.5.2.2 Toxicité chronique

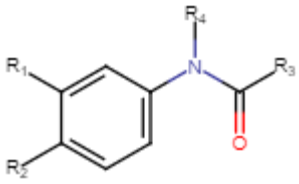
La classification concernant la toxicité chronique permet de définir des substances actives présentant des risques d'effets spécifiques sur la santé (cancérogènes, mutagènes et reprotoxiques) et comporte trois niveaux. Les substances et spécialités herbicides sont classées de niveau 3, le plus bas sur cette échelle. Ce niveau correspond à des « substances préoccupantes pour l'homme en raison d'effets possibles ».

Les scientifiques savent qu'un large éventail de substances dont des pesticides peut avoir des effets nocifs, les mécanismes par lequels les substances peuvent perturber ces systèmes sont trés complexes et subtils et ont la possibilités d'entraîner des changements sur la croissance, le developpement, la reproduction ou le comportement qui induisent des conséquences sur l'oragnisme lui-même ou la génération suivante [12].

1.6 Composés modèles : dérivés d'anilide

1.6.1 Définition

Les anilides (ou phénylamides) sont une classe de composés chimiques qui sont des dérivés acyliques de l'aniline, selon l'IUPAC [14] :

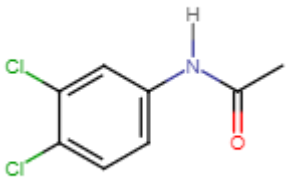


1- Composés dérivés des oxoacides en remplaçant un groupe OH par le groupe NHPh ou dérivé formé par substitution de cycle ; N-phényl amides, par exemple : Acétanilide.

2- Sels formés par remplacement d'un hydron aniline lié à l'azote par un métal ou un autre cation, par exemple : Anilide de sodium NaNHPh.

1.6.2 Propanil

1) Définition



Le propanil (3,4-dichloropropionanilide) est un anilide résultant de la condensation formelle du groupe carboxy de l'acide propanoïque avec le groupe amino de la 3,4-dichloroaniline. C'est un herbicide utilisé pour le traitement de nombreuses herbes et mauvaises herbes à feuilles larges dans le riz, les pommes de terre et le blé. Il est dérivé de 3,4-dichloroaniline [15].

Il est légèrement soluble dans l'eau, mais se dissout dans la plupart des solvants organiques. C'est un herbicide actif pour le contrôle des mauvaises herbes miliaires et autres dans les rizières. La DL_{50} pour les rats est d'environ 1300 mg / kg ; pour les lapins, elle est de 500 mg / kg [16].

2) Propriétés physico-chimiques

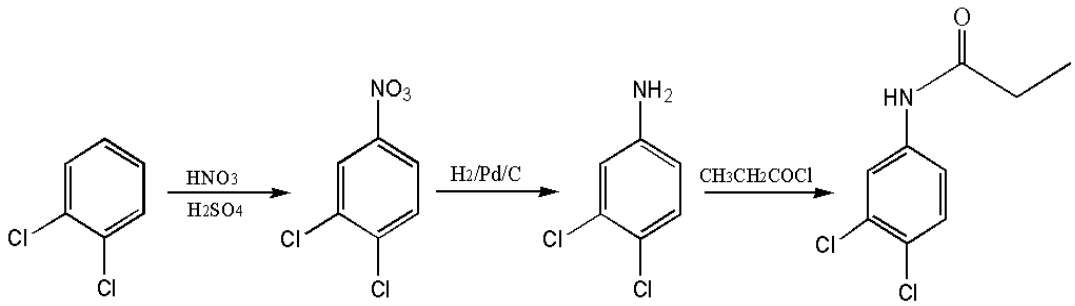
- Le propanil apparaît comme des cristaux incolores à bruns.
- Solubilité dans l'eau à température ambiante.
- Peu soluble dans les solvants aromatiques.
- Non corrosif dans des conditions normales d'utilisation.
- Corrosif pour le polyéthylène. [15]

3) Conditions de consérvation et décomposition

- L'activité biologique reste pratiquement inchangée pendant deux ans dans des conditions environnementales, à condition que le produit soit stocké dans ses conteneurs d'origine non ouverts et non endommagés, dans des endroits ombragés et éventuellement bien aérés.
- Lorsqu'il est chauffé jusqu'à décomposition, il émet des fumées très toxiques de chlorure d'hydrogène et d'oxydes d'azote [15].

4) Préparation

Le propanil peut être préparé par réaction de 3,4-dichloroaniline avec de l'acide propionique en présence de chlorure de thionyle [17] :



5) Utilisations

Le tableau (Tab 1.3) ci-dessous représente les utilisations de propanil dans des différentes catégories [15] :

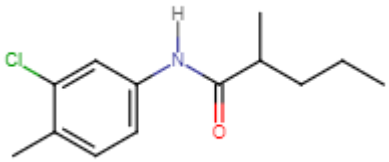
Tableau 1.3 : Utilisations de Propanil.

Catégorie	Description
Résidus alimentaires	Résidus trouvés dans les aliments, généralement des médicaments ou des pesticides.
Pesticide/ additif alimentaire	Comprend des épices, des extraits, des colorants, des saveurs, etc. ajoutés aux aliments pour la consommation humaine
Pesticide	Substances utilisées pour prévenir, détruire ou atténuer les ravageurs.

1.6.3 Pentanochlore

1) Définition

Le pentanochlore, $C_{13}H_{18}ClNO$, N-(3-chloro-4-methyl phenyl)-2-methylpentanamide [18], est un anilide herbicide. Aussi nommé solan, est un solide incolor, soluble dans le xylène (0.25kg/l). Il est stable dans les solutions aqueuses à pH 7-9 à température ambiante. Il produit également du chlorure toxique et de l'oxyde d'azote gazeux.



Le pentanochlore, un herbicide sélectif de contact et de pré- et post-levée, agit en inhibant la photosynthèse. il est utilisé pour contrôler les mauvaises herbes annuelles dans les carottes, le fenouil et le persil. Il est également appliqué au stade de la pré-plantation sur les tomates. Cet herbicide est irritant pour les yeux, la peau et les voies respiratoires. La DL_{50} orale aiguë chez le rat est > 1000 mg / kg. Le pentanochlore est modérément toxique [11].

2) Propriétés physico-chimiques

- Solide et incolor.
- Insoluble dans l'eau ;
- Soluble dans l'huile de pin, la diisobutyl cétone, l'isophorone et le xylène.
- Combustible. [18]

3) Utilisations

Le tableau (Tab 1.4) ci-dessous représente les utilisations de pentanochlore dans des différentes catégories [18] :

Tableau 1.4 : Utilisations de pentanochlore

Catégorie	Description
Soins personnels	Produits de soins personnels, y compris cosmétiques, shampooings, parfums, savons, lotions, dentifrices, etc.
Soins personnels/cosmétiques	Sous-catégorie des soins personnels, comprend les parfums, les shampooings, le maquillage, etc.
Pesticide	Substance utilisée pour prévenir, détruire ou atténuer les ravageurs.

2. Relation quantitative structure-activité (QSAR)

En raison de la demande de produits chimiques plus sûrs dans les disciplines médicales et agricoles. Au cours des 20 dernières années, les scientifiques et les ingénieurs ont travaillé pour concevoir des substances basées sur l'atténuation des effets toxiques sur l'environnement humain et l'écologique [19]. Ces derniers ont constaté que l'utilisation de méthodes prédictives *in silico*, basées sur des outils informatiques, offre une alternative rapide, rentable et éthique. Ces méthodes comprennent les modèles de QSAR [20].

2.1 Aperçu historique

En 1868, Crum-Brown et Fraser ont publié une équation qui est considérable pour être la première formulation d'une relation quantitative structure-activité, dans leurs recherches sur différents alcaloïdes [21]. Par conséquent, ils ont supposé que « l'activité physiologique » Φ devait être fonction de la structure chimique C (Eq. 1) [22]

$$\Phi = f\,(C) \tag{1.1}$$

Richet a découvert en 1893 que la toxicité des composés organiques suit inversement leur solubilité dans l'eau (il a conclu "*plus ils sont solubles, moins ils sont toxiques*") [23]. Une telle relation correspond à l'équation (Eq. 2) [22], où ΔΦ traduit les différences dans les valeurs d'activité biologique, causées par des changements correspondants dans les propriétés chimiques et surtout physicochimiques : ΔC.

$$\Delta\Phi = f\,(\Delta C) \tag{1.2}$$

Le développement de la technique QSAR a subi plusieurs transformation mais l'année 1964 est considérée comme le début des méthodes QSAR modernes où Hansch et Fujita ont établi les premières corrélations entre les propriétés physico-chimiques (log P, pKa, paramètres stériques et électroniques) et l'activité biologique (activité enzymatique, pharmacologique) [24].

La méthodologie QSAR s'est rapidement développée à partir du 1988, en raison de l'introduction du paramètre moléculaire tridimensionnel, qui expliquaient l'influence de différents conformères, stéréoisomères or énantiomères (modèles 3D QSAR).

Le premier modèle publié possédant ces caractéristiques était l'analyse comparative du champ moléculaire (CoMFA "Comparative Molecular Field Analysis") proposé par Cramer et al, qui est actuellement l'une des méthodologies QSAR les plus utilisées. [25]. La figure ci-dessous résume le développement de l'étude QSAR.

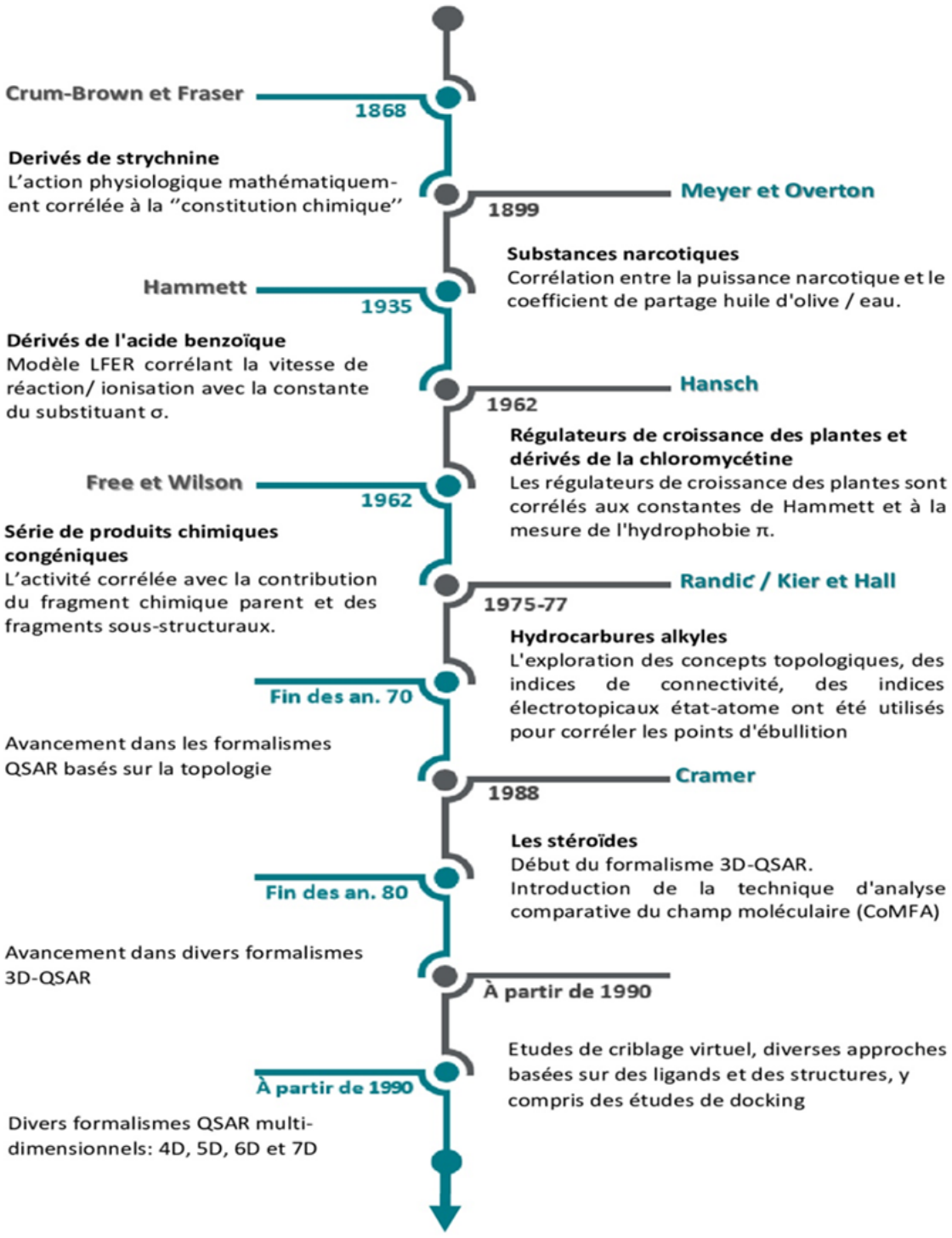


Figure 1.6 : Résumé des découvertes qui ont conduit à l'évolution progressive de l'étude QSAR [26]

2.2 Définition

Par définition, QSAR (relation quantitative structure-activité) est une corrélation mathématique entre une activité biologique ou une propriété moléculaire spécifiée et une ou plusieurs propriétés structurelles physicochimiques et/ou moléculaires, appelées descripteurs puisqu'elles "décrivent" l'activité ou la propriété examinée [27]. Le terme générique QSAR peut être exprimé comme :

$$\textbf{Activité biologique} = f\ \textbf{(paramètres)}$$

2.3 Principe

Le principe d'une étude QSAR (Fig 1.7), consiste à trouver une relation mathématique reliant de manière quantitative une activité biologique, mesurée pour une série de composés similaires dans les mêmes conditions expérimentales, avec des descripteurs moléculaires à l'aide des méthodes statistiques. Ces études permettent à analyser les données structurales afin de détecter les facteurs déterminants pour l'activité étudiée. Pour ce faire, différents types de méthodes statistiques peuvent être employées, les plus répondues sont : la méthode de moindres carrées partielles, la régression linéaire simple et multiple et les réseaux de neurones artificiels. (Pour plus de détails, voir : Chapitre 2).
La modélisation QSAR comporte généralement trois étapes [28] :

- Recueillir ou, si possible, concevoir un ensemble d'information sur la structure et les propriétés des produits chimiques.
- Choisir des descripteurs capables de relier correctement la structure chimique à l'activité biologique.
- Appliquer des méthodes statistiques qui corrèlent les changements de structure avec les changements dans l'activité biologique.

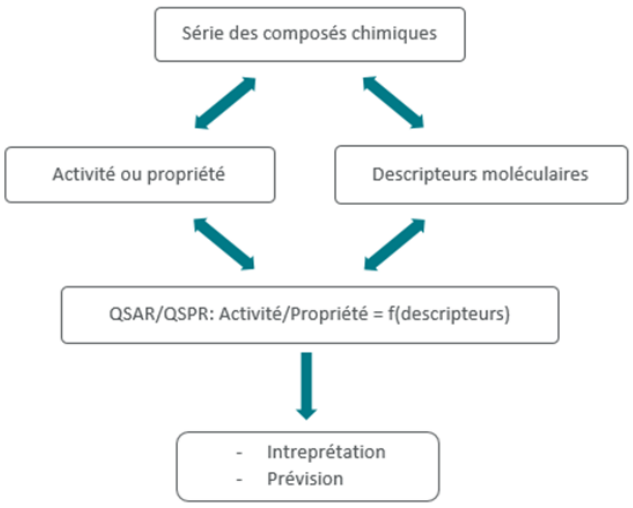


Figure 1.7 : Modèle de l'étude QSAR [29]

L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de l'activité étudiée pour de nouvelles molécules ou des molécules pour lesquelles les données expérimentales ne sont pas disponibles [24].

2.4 Objectifs

Les objectifs pratiques d'une étude QSAR sont nombreux et ces techniques sont largement utilisées dans de nombreuses situations. Le but des études in silico comprend donc ce qui suit [30] :

- Prédire l'activité biologique et les propriétés physico-chimiques par des moyens rationnels.
- Comprendre les mécanismes d'action au sein d'une série de produits chimiques.
- Économies sur les coûts de développement de produits.
- Les prévisions pourraient réduire la nécessité de tests sur les animaux longs et coûteux.

2.5 Applications QSAR

QSAR présente une option appropriée dans la surveillance rationnelle de l'activité / propriété / toxicité des produits chimiques et est donc utile dans une grande variété d'applications. Dans une perspective globale, les produits chimiques modélisés à l'aide de la méthode QSAR peuvent être présentés en trois grands types, qui sont [26] :

- Produits chimiques bénéfiques pour la santé : médicaments, ingrédients alimentaires...
- Les produits chimiques impliqués dans les processus industriels de laboratoire : solvants, réactifs, etc.
- Les produits chimiques présentant un résultat dangereux : polluants organiques persistants, toxines, xénobiotiques, cancérogènes, composés organiques volatils, etc.

La figure (Fig 1.8) suivante présente l'utilisation de l'application QSAR en trois domaines principaux :

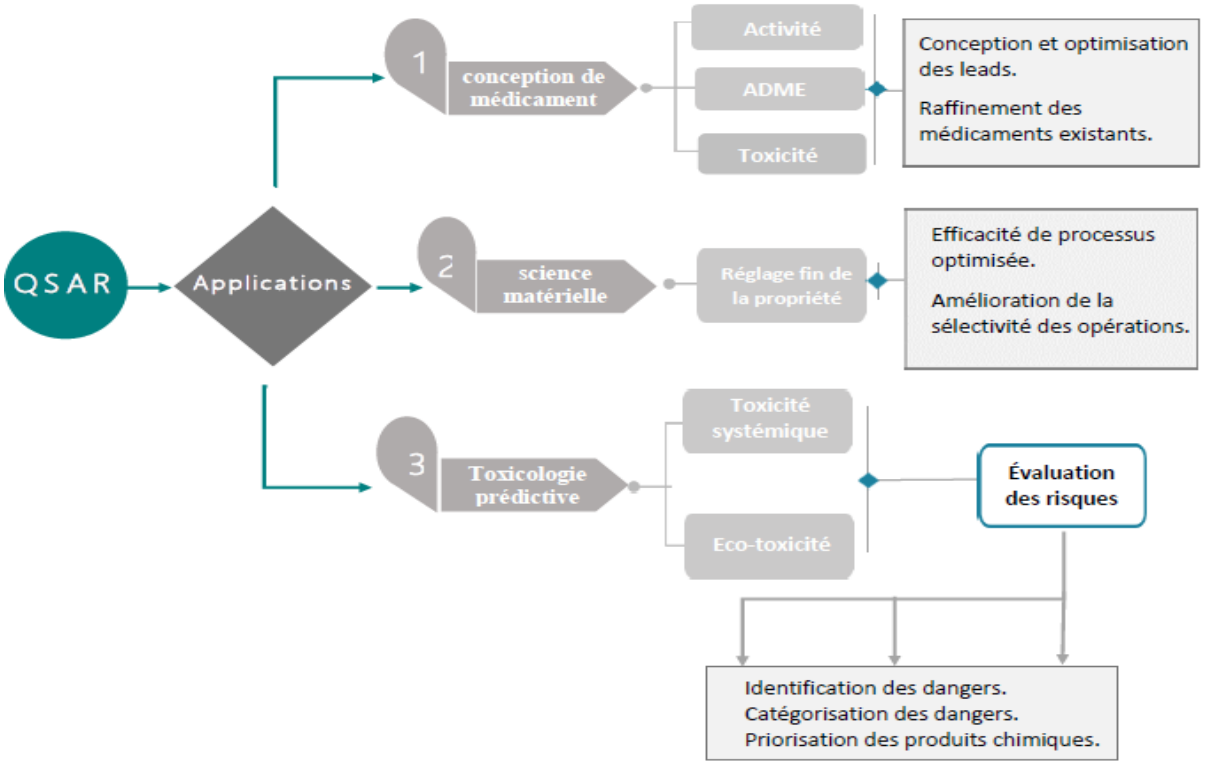


Figure 1.8 : Grands domaines d'application traités par les études QSAR [26]

Il existe alors un grand nombre d'applications de ces modèles au sein de l'industrie, du monde universitaire et des organismes réglementaires. Un petit nombre d'utilisations potentielles sont répertoriées ci-dessous [30] :

- L'identification de nouvelles pistes à activité pharmacologique, biocide ou pesticide.
- L'optimisation de l'activité pharmacologique, biocide ou pesticide.
- La conception des médicaments et de nombreux autres produits tels que les agents tensio-actifs, parfums, les colorants et les produits de la chimie fine.
- L'identification des composés dangereux aux premiers stades de développement.
- La prédiction de la toxicité et les effets secondaires de nouveaux composés.
- La prédiction de la toxicité pour l'homme par une exposition délibérée, occasionnelle et professionnelle.
- La prédiction de la toxicité pour l'écosystème.
- La prédiction d'une variété de propriétés physico-chimiques des molécules.
- La prédiction du sort des molécules libérées dans l'environnement.

2.6 Méthodologie générale d'une étude QSAR

La méthodologie générale d'une étude QSAR est la suivante [28] :

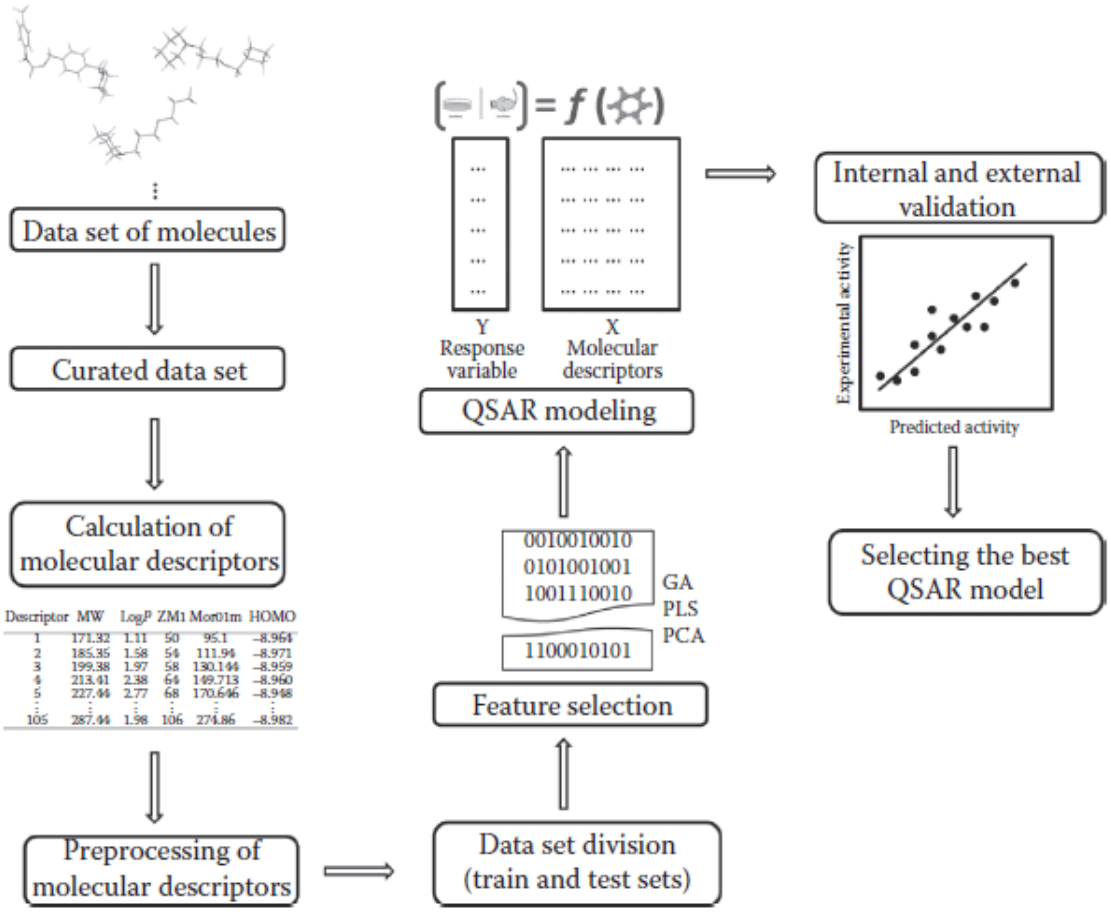


Figure 1.9 : Représentation schématique de travail QSAR [31]

- Constituer une base de données à partir des mesures expérimentales fiables de l'activité de chaque composé.
- Sélectionner les descripteurs en relation avec l'activité étudiée.
- Diviser cette base de données, aléatoirement, en une série d'apprentissage (Training set) qui contient généralement les 2/3 de la base de données et une série de test (Test set) constituée par le 1/3 restant.
- Etablir des modèles mathématiques en utilisant la série d'apprentissage.
- Valider les modèles élaborés en utilisant la série de test et calculer leurs paramètres statistiques de validation externe.
- Elaborer le domaine d'applicabilité du modèle retenu.

2.6.1 Base de données

Le choix de la base de données expérimentale initiale est une étape critique pour le développement des modèles QSAR. Généralement, les composés testés ont deux origines possibles, soit des produits de synthèse ou bien des produits d'extraction à partir de plantes [32]. Les données devraient provenir du même protocole d'analyse, et il faudrait veiller à éviter la variabilité inter-laboratoire. Tout mauvais point de données aura tendance à corrompre la corrélation correcte de la structure et de l'activité [28].

2.6.2 Descripteurs moléculaires

Les descripteurs moléculaires sont des représentations mathématiques formelles d'une molécule, obtenues par un algorithme bien défini et appliquées à une représentation moléculaire ou à une procédure expérimentale bien définie [33]. Ils permettent d'effectuer des prédictions sans avoir à synthétiser les molécules, ce qui est l'un des objectifs de la modélisation moléculaire [34].

La nature des paramètres moléculaires varie en fonction de leurs types et de leurs algorithmes. Les paramètres sont souvent classés comme descripteurs constitutionnels, topologiques, géométriques, thermodynamiques et électroniques. De plus, ils peuvent également être classés sur la base de la dimensionnalité comme 0D, 1D, 2D, 3D et autres [31].

Les descripteurs 1D : Ils sont accessibles à partir de la formule brute de la molécule, et décrivent des propriétés globales du composé [35]. Les descripteurs les plus courants sont le nombre d'atomes, le nombre de liaisons, le type d'atome, le nombre de cycles et le poids moléculaire [28].

Les descripteurs 2D : Ce sont les descripteurs ou des indices topologiques qui prennent en compte l'arrangement atomique interne des composés [28]. Ils sont calculés à partir de la formule développée de la molécule. On distingue :

- **Les indices constitutionnels** caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles…

- **Les indices topologiques** peuvent être obtenus à partir de la structure 2D de la molécule [36], et décrivent les connectivités atomiques dans la molécule. Ce sont des descripteurs plus "sophistiqués", ils contiennent en leur sein des informations sur la taille globale du système, sa forme globale et ses ramifications [24].

Ces descripteurs 2D permettent de prédire les propriétés physiques mais sont insuffisantes pour expliquer certaines propriétés et activités biologiques comme la toxicité [36].

Les descripteurs 3D : Ils 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 [37] le plus souvent par modélisation moléculaire empirique ou ab-initio. Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent de la structure 3D. On distingue plusieurs familles importantes de descripteurs 3D :

- **Les descripteurs géométriques** : Ils sont basés sur l'arrangement spatial des atomes constituant la molécule et sont définis par les coordonnées des noyaux atomiques et la grosseur de la molécule représentée. Dans ce mémoire nous avons choisi : le volume et la surface moléculaire et le nombre de liaisons rotatives.

- **Les descripteurs physico-chimiques :** certains d'entre eux reflètent la composition moléculaire du composé (le nombre et le type d'atomes et de liaisons présents dans la molécule, le nombre de cycle, les propriétés donneur/accepteur de liaison H, cation, anion, etc.…). D'autres représentent le caractère hydrophile ou lipophile de la molécule généralement évalué à partir du coefficient de partage octanol/eau. Dans ce mémoire nous avons choisi : le coefficient de partage et le nombre de liaisons hydrogène accepteurs.

- **Les descripteurs quantiques/électroniques :** qui caractérisent la distribution de charge des molécules (polarité des molécules). Par exemple, les énergies de la plus haute orbitale moléculaire occupée (HOMO) et de la plus basse vacante (LUMO) sont des descripteurs fréquemment sélectionnés. Le moment dipolaire, le potentiel d'ionisation, et différentes énergies relatives à la molécule sont d'autres paramètres importants [24]. Dans ce mémoire nous avons choisi : Le moment dipolaire, les charges des atomes.

Les descripteurs 4D : Ils correspondent à la mesure des propriétés 3D (potentiel électrostatique, d'hydrophobicité, de liaison hydrogène...) d'une molécule en tout point de l'espace. Ils permettent d'avoir l'information sur la structure de la cible (protéine). Ces descripteurs sont obtenus par le calcul des champs d'interactions moléculaires (CoMFA, CoMSIA) entre une molécule et une sonde représentée par une autre molécule (eau, amide, ...) [24].

A. Choix des descripteurs moléculaires

Pour avoir les bons choix de descripteurs moléculaires on doit respecter les points suivants [38] [39] :

- Les descripteurs doivent avoir une relation directe, et une description suffisante à la structure moléculaire des composés chimiques à étudier.
- Les descripteurs doivent avoir des liens avec les propriétés chimiques et physiques et l'activité des composés chimiques.
- La représentation descriptive de la totalité de la structure moléculaires des composées.
- La diversité (variétés) de ces descripteurs.
- La sélection exacte et la distinction entre les descripteurs décisifs (principaux), et d'écarts (secondaires).
- La simplicité et l'efficacité de ces descripteurs, puisque la complexité des provoque une perturbation sur les résultats d'estimation.
- Il n'est pas nécessaire d'avoir des relations parfaites, entre les variables (descripteurs) prédictives, parce qu'elles causent des difficultés sur la régression linéaire (confusion parfaite d'effets et perturbation dans l'interprétation des résultats).

2.6.3 Méthodes d'analyse des données

Une méthode d'analyse des données est nécessaire pour développer des modèles QSAR. Il existe plusieurs méthodes pour construire un modèle et d'analyser ses données statistiques. Ces méthodes sont disponibles dans des logiciels, tels que : Matlab, Excel, Minitab, Statistica, SPSS, R, etc. Les méthodes utilisées dans ce travail sont la régression linéaire multiple et l'apprentissage automatique via les réseaux de neurones artificiels implémentées sur Matlab, qui seront présentées en détails dans le deuxième chapitre.

2.6.4 Validation du modèle

Une fois développé, le modèle doit être validé afin de vérifier sa stabilité, son pouvoir explicatif et son pouvoir prédictif. Deux types d'approches de validation sont décrits ci-dessous :

A. Validation Interne

1) Validation croisée k-fold

La procédure « k-fold » correspond à un découpage du jeu d'apprentissage en k parties. On sélectionne un des k échantillons comme ensemble de validation (Test) et les (k-1) autres échantillons constitueront l'ensemble d'apprentissage (Training), puis on construit le modèle QSAR à partir de ce dernier, et on prédit les activités de l'ensemble de validation. L'opération est répétée k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La figure (Fig 1.10) peut aider:

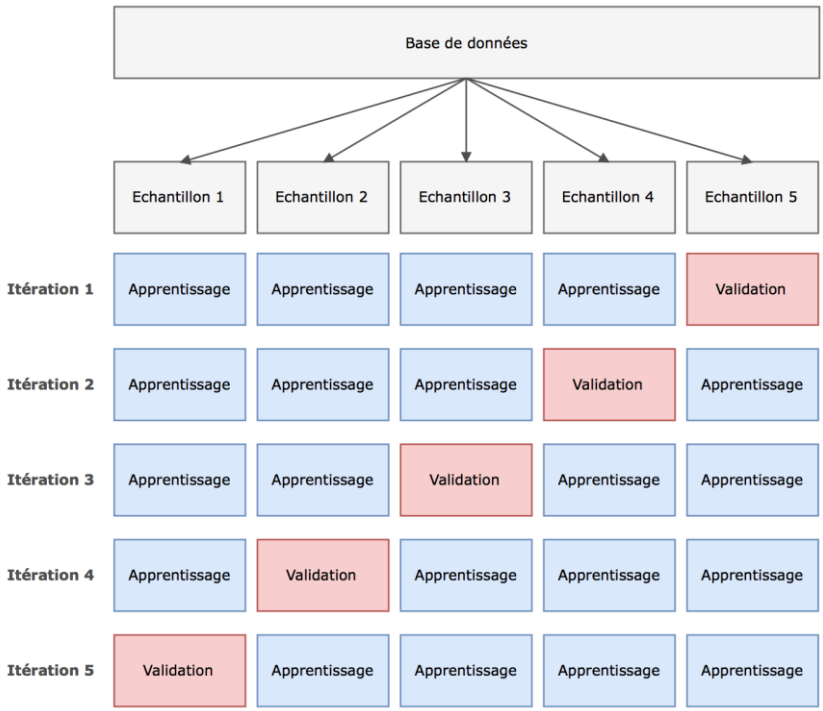


Figure 1.10: k-fold validation croisée avec k = 5 [40]

Le cas particulier où nous prenons $k = n$[1] s'appelle la validation croisée leave-one-out. C'est-à-dire que l'on apprend sur (n-1) observations pour construire le modèle QSAR puis on le valide sur la n^{ieme} observation et l'on répète cette opération n fois pour qu'en fin de compte chaque observation ait été utilisée exactement une fois comme ensemble de validation [24].

[1] n représente le nombre d'échantillon présents dans la base de construction.

B. Validation Externe

L'autre évaluation critique à faire sur le modèle QSAR si les valeurs des indices internes élevées est la validation externe [31].

La sélection d'un sous-ensemble de données comme ensemble de test est le point de départ de la validation externe. Pour ce faire, les activités biologiques des composés de l'ensemble Test sont prédites pour déterminer le pouvoir prédictif du modèle. Le paramètre le plus couramment utilisé pour évaluer les performances prédictives du modèle est le coefficient de corrélation prédictive R^2_{pred} entre les activités observées et les activités prédites pour l'ensemble de test qui est donné par l'équation suivante:

$$R^2_{pred} = 1 - \frac{\sum_{i=1}^{n_{ts}} \left(y^i_{test} - \hat{y}_{test}\right)^2}{\sum_{i=1}^{n_{ts}} \left(y^i_{test} - \bar{y}_{train}\right)^2} \quad (1.3)$$

Où y^i_{test}, $\hat{y}_{test}$ et $\bar{y}_{train}$ sont respectivement les valeurs observées, prédites et moyennes de la réponse biologique pour les ensembles de test et de train. La valeur R^2_{pred} varie de 0 à 1, et il est suggéré qu'elle ne devrait pas être inférieure à 0,6.

2.6.5 Domaine d'applicabilité du modèle

Le but ultime d'une analyse QSAR est de prédire les valeurs de la réponse pour des nouvelles entités chimiques. Dans ce contexte, le domaine d'applicabilité (DA) de l'ensemble d'apprentissage devrait être déterminé. DA est défini comme suit : « *Le domaine d'applicabilité QSAR est l'espace, les connaissances ou les informations physiques, chimiques, structurelles ou biologiques sur lequel l'ensemble d'apprentissage du modèle a été développé, et pour lequel il est applicable de faire des prédictions pour de nouveaux composés* [31]. *Ainsi, la prédiction d'une activité à l'aide du modèle QSAR n'est valable que si le composé à prédire se situe dans le DA.* » Bien qu'il existe plusieurs façons de présenter un DA, le graphe de William (méthode de leviers) est considéré comme l'une des méthodes les plus courantes basées sur la distance de levier h_i [41] (Cette méthode est détaillée dans Chapitre 2).

3. Conclusion

Dans ce chapitre, nous avons présenté en première partie la nature de la molécule étudiée et ses effets biologiques : sa composition, son mode d'action à l'égard des végétaux ainsi que ses effets toxiques. Ensuite, nous avons abordé la méthodologie QSAR en décrivant ses différentes caractéristiques. La base de données, les descripteurs moléculaires, la méthode de validation du modèle et le domaine d'applicabilité du modèle ont été décrits en détails.

Chapitre 2

Outils de la modélisation

Ce chapitre présente un aperçu des outils statistiques utilisés dans ce travail. Une introduction à l'apprentissage automatique est décrite, ainsi que l'analyse par la régression linéaire multiple.

L'objectif de ce travail est la recherche d'un modèle mathématique fiable et représentatif de la relation entre la structure des anilides et leur activité herbicide à partir d'une base de données de mesures expérimentales. Le modèle obtenu sera utilisé pour prédire l'activité de nouveaux composés et concevoir de nouvelles structures. Parmi les différentes méthodes statistiques de QSAR utilisées pour construire ce modèle : la régression linéaire. Les outils d'apprentissage automatique (Machine Learning Tools) tels que les réseaux de neurones artificiels sont également très efficaces pour développer des modèles prédictifs [42].

Ce chapitre est une introduction des deux méthodes statistiques, la régression linéaire multiple (RLM) et les réseaux de neurones artificiels (RNA), où on retrouve la plupart des éléments de base inhérents à ces deux méthodes.

1. Régression linéaire multiple

On cherche à décrire la relation existant entre une variable quantitative Y appelée variable à expliquer (ou encore, réponse, exogène, dépendante), ici l'activité biologique, et une série de n variables quantitatives $X_1, \dots, X_n$ dites variables explicatives (ou encore, endogènes, régresseurs) [43].

1.1 Modèle RLM

1.1.1 Définition

Un modéle de régression linéaire[2] qui contient plus qu'un variable prédicteur est applé un modéle de régression multiple, notée par RLM, il est défiué par :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad (2.1)$$

Où :

- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ sont appelés les paramètres ou les coefficients inconnus du modèle que l'on veut estimer à partir des données.
- $i = 1, \dots, n$ correspond au numéro d'obsérvation.
- y_i est la $i^{ème}$ obsérvation de variable Y.
- X_{ij} est la $i^{ème}$ obsérvation de la $j^{ième}$ variable.
- ε_i est l'erreur du modèle (bruit). Il représente la déviation entre ce que le modèle prédit et la réalité [43].

[2] Le modèle est linéaire par rapport aux coefficients et non aux variables. Lorsque le dérivé partiel par rapport aux coefficients est indépendent de ces derniers, le modèle est dite linéaire : $\partial y / \partial \beta \neq f(\beta)$

La Figure (Fig 2.1) donne une représentation graphique de la méthode pour deux variables dépendantes. Il s'agit en fait de définir le plan au plus proche de tous les points de l'espace (au sens des moindres carrés). Pour n variables, il s'agira d'un hyperplan d'ordre n.

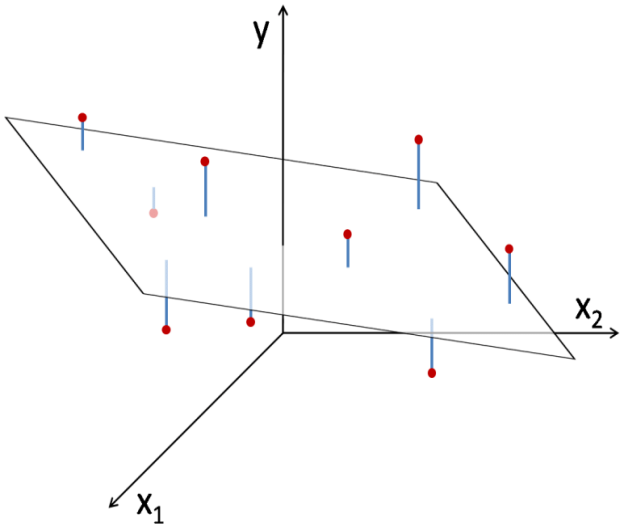


Figure 2.1 : Représentation graphique de la régression multi-linéaire pour deux variables indépendantes x_1 et x_2 et une variable dépendante y [44]

Remarques

1. Le coefficient β_0 est un paramètre appelé intercepte qui représente la moyenne des y_i lorsque la valeur de chaque variable explicative est égale à zéro.

2. Les coefficients β_j ($j = 1, \ldots, p$) représentent le changement subi par $E(y_i)$ correspondant à un changement unitaire dans la valeur de la $j^{ième}$ variable explicative, lorsque les autres variables explicatives demeurent inchangées [43].

3. Les résidus ε_i représentent l'erreur du modèle, constituée par l'incertitude sur la variable dépendante y_i d'une part, sur les variables indépendantes x_i d'autre part, mais aussi par les informations contenues dans les variables indépendantes mais non exprimées via les variables dépendantes [44].

1.2 Notation matricielle

Les observations $y_1, y_2, \dots, y_n$ enregistrées pour chacun de ces n niveaux peuvent être exprimées sous la forme d'équation comme suit :

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_2 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \quad \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases} \quad (2.2)$$

Ce système de n équations peut être représenté en notation matricielle comme suit :

$$y = X\beta + \varepsilon \quad (2.3)$$

Où:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \quad (2.4)$$

Avec :

- $\boldsymbol{X}$ est la matrice de taille $(n, p+1)$ contenant l'ensemble des variables explicatives observées avec la première colonne formée par la valeur 1 permettant d'intégrer la constante du modèle.
- $\boldsymbol{Y}$ est la matrice colonne de taille $(n, 1)$ contenant l'ensemble des variables à expliquées observées.
- $\boldsymbol{\varepsilon}$ représente le vecteur colonne de taille $(n, 1)$ contenant tous les résidus.
- $\boldsymbol{\beta}$ représente le vecteur colonne de taille $(p+1,1)$ contenant les coefficients de régression [45].

1.3 Estimation des paramètres par Moindres Carrés Ordinaires

1.3.1 Hypothèses relatives au modèle RLM

L'utilisation de la méthode des moindres carrés pour l'estimation des coefficients nécessite de faire certaines hypothèses pour qu'elle soit applicable. Les hypothèses sont les suivantes [45] :

$\mathcal{H}_1$ X est une donnée exogène dans le modèle, elle est supposée non aléatoire. $\hat{Y}$ est aléatoire par l'intermédiaire du résidu ε.

$\mathcal{H}_2$ Les résidus ε_i sont indépendants et identiquement distribués. Les résidus suivent la loi gaussienne d'espérance nulle et de variance qui est constante et indépendante des observations. On parle de l'hypothèse d'homoscédasticité:

$$\mathbb{E}(\varepsilon^i) = 0 \quad ; \quad \mathbb{V}(\varepsilon^i) = \sigma^2$$

$\mathcal{H}_3$ Les résidus sont indépendants de la variables exogène X et les erreurs relatives entre deux observations sont indépendantes. C'est l'hypothèse de non auto-corrélation des erreurs :

$$COV(\varepsilon^i, \varepsilon^j) = 0 \quad \text{pour} \quad i \neq j.$$

$\mathcal{H}_4$ La matrice $X'X$ de taille $(p+1, p+1)$ est inversible. Elle indique l'absence de colinéarité entre les variables exogènes.

1.3.2 Estimation des paramètres par MCO

La méthode des moindres carrés a pour but de trouver la droite de régression qui se rapproche le mieux de nuage de points. Cette droite est celle pour laquelle la somme des carrés des distances verticales des points à la droite est la plus petite possible (Fig 2.2). Cela consiste d'abord à déterminer les valeurs des coefficients de régression $\hat{\beta}_i$ qui minimisent la quantité de la somme quadratique des résidus.

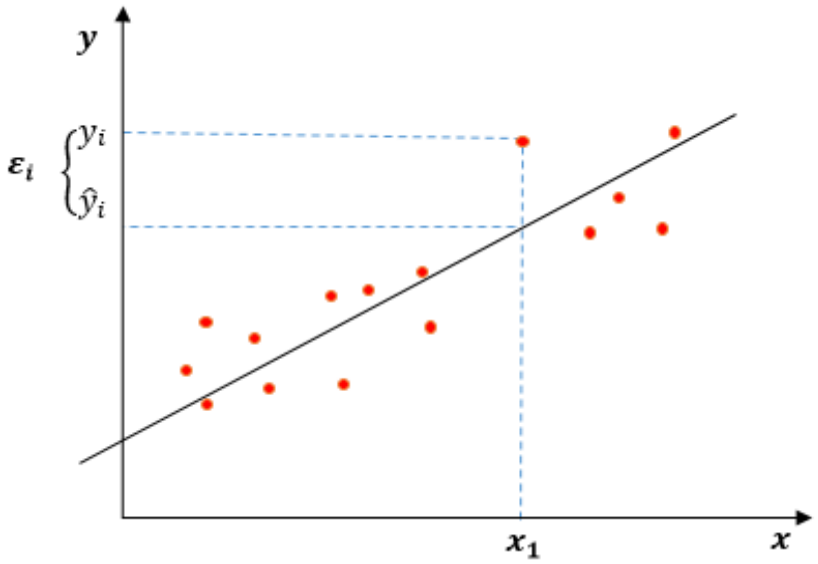


Figure 2.2 : Droite de la régression linéaire et résidus.

Notons que la distance minimisée avec les MCO est $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

- Soit $\hat{\varepsilon} = (\widehat{\varepsilon_1}, \dots, \widehat{\varepsilon_n})'$ le vecteur de dimension $n \times 1$ des résidus défini par :

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta} \tag{2.5}$$

Où $\hat{Y} - X\hat{\beta}$ représente les valeurs estimées par le modèle, on les appelle aussi valeurs ajustées [43].

On appelle estimateur des moindres carrés des coefficients β, l'estimateur $\hat{\beta}$ obtenu par minimisation de la quantité :

$$\hat{\beta} = argmin\, S(\beta) \tag{2.6}$$

Tel que :

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) \tag{2.7}$$

Notez que $S(\beta)$ peut être exprimé comme [46]:

$$\begin{aligned} S(\beta) &= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta \ ^{3} \\ &= y'y - \beta'X'y - \beta'X'y + \beta'X'X\beta \\ &= y'y - 2\beta'X'y + \beta'X'X\beta \end{aligned} \tag{2.8}$$

Pour que $S(\beta)$ soit minimum, il faut que les dérivées par rapport à $\hat{\beta}$ soient nulles :

$$\begin{aligned} \left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = 0 \ &\Leftrightarrow -2X'y + 2X'X\hat{\beta} = 0 \\ &\Leftrightarrow \hat{\beta} = (X'X)^{-1}X'y \end{aligned} \tag{2.9}$$

Pour obtenir une information sur la qualité de la régression, il est important d'avoir une estimation de l'erreur standard sur les coefficients estimés $\hat{\beta}$ qui nous donnera alors une estimation sur $\hat{y}_i$. L'estimation de l'écart type des coefficients $(\hat{\beta}_0, \dots, \hat{\beta}_p)$ s'obtient en utilisant la variance des résidus.

1.3.3 Propriétés des estimateurs MCO

On cherche maintenant à vérifier que les estimateurs des MCO que nous avons construit admettent de bonnes propriétés au sens statistique. Cela peut se résumer en deux parties: l'estimateur des MCO est-il sans biais et est-il de variance minimale dans sa classe d'estimateurs ?

En admettant, sous l'hypothèse $\mathcal{H}_2$, que : $\mathbb{E}(\varepsilon^i) = 0$ et $\mathbb{V}(\varepsilon^i) = \sigma^2$. Cette hypothèse nous permet de calculer [47] :

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}((X'X)^{-1}X'y) = (X'X)^{-1}X'\mathbb{E}(y) = (X'X)^{-1}X'X\beta = \beta \tag{2.10}$$

L'estimateur $\hat{\beta}$ est donc sans biais, sa variance vaut :

$$\mathbb{V}(\hat{\beta}) = \mathbb{V}((X'X)^{-1}X'y) = (X'X)^{-1}X'\,\mathbb{V}(y)\,X(X'X)^{-1} = \sigma^2(X'X)^{-1} \tag{2.11}$$

[3] D'après les équations (Eq. 7) et (Eq. 8) on voit que $(X\beta)' = \beta'X'$ parce que la transposée du produit de deux matrices est égale au produit des transposées de ces deux matrices, mais dans l'ordre inverse . Puisque $(\beta'X'y)$ est une matrice 1×1, ou un scalaire, sa transposée est le même scalaire $(\beta'X'y)' = \beta'X'y$.

Remarque 1 : Les estimateurs MCO sont à variance minimale, c'est-à-dire qu'il n'existe pas d'autres estimateurs linéaires sans biais présentant une plus petite variance. On parle d'estimateurs efficaces [45].

Remarque 2 : La matrice de variance-covariance $varcov(\hat{\beta})$ des coeffcients, de dimension $(p+1;\ p+\ 1)$; est donnée par :

$$varcov(\widehat{\beta}) = \begin{pmatrix} var(\widehat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) & \cdots & cov(\hat{\beta}_0, \hat{\beta}_p) \\ \cdots & var(\widehat{\beta}_1) & \cdots & cov(\hat{\beta}_1, \hat{\beta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & var(\widehat{\beta}_p) \end{pmatrix}$$

Cette matrice est symétrique, sur sa diagonale principale on observe les variances des coefficients estimés $(var(\widehat{\beta}_0), \ldots, var(\widehat{\beta}_p))$ [3].

- L'estimateur sans biais de la variance des erreurs est donné par :

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-(p+1)} = \frac{SCR}{n-p-1} \qquad (2.12)$$

Où : $\hat{\varepsilon}_i = y_i - X\hat{\beta}$, et SCR est la somme des carrés résiduelles [48].

- Les erreurs standards (standard error, se) des coefficients estimés sont alors calculés à partir la formule suivante [5] :

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 (X'X)^{-1}_{(p+1,p+1)}} = \sqrt{\hat{\sigma}^2\, C_{jj}} \qquad (2.13)$$

Où : C_{jj} est le $j^{ième}$élément diagonal de la matrice $(X'X)^{-1}$.

1.4 Analyse de variance et qualité de l'ajustement

1.4.1 Décomposition de la variance

L'intérêt d'un modèle de régression linéaire réside dans sa capacité à expliquer une partie des variations de la variable Y par les variations de la variable X. La variation d'une variable Y est obtenue en considérant les différences entre les valeurs observées y_i et leur moyenne $\bar{y}$.

On a:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \tag{2.14}$$

On peut obtenir la décomposition :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.15}$$

Où :

- $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$ avec $\bar{y}$ la moyenne des y_i. Ce terme correspond à la somme des carrés totaux (SCT). Elle indique la variabilité totale de Y. Elle représente l'information disponible dans les données.

- $SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ est la somme des carrés expliqués (SCE). Elle montre la variabilité expliquée par le modèle, c-à-d la variation de Y expliquée par X.

- $SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(\varepsilon_i)^2$ Correspondant à la somme des carrés des résidus. C'est le terme que l'on cherche à minimiser ($SCR = S(\beta)$). Elle représente la variabilité non-expliquée par le modèle. [45]

On peut produire le tableau d'analyse de variance (Tab 2.1) à partir de la décomposition de la variance, comme suit :

Tableau 2.1 : Tableau d'analyse de variance pour la régression multiple [49]

Source de variation	Somme des carrés	Ddl	Carrés moyens
Expliquée	$SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	P	$CME = \frac{SCE}{p}$
Résiduelle	$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	n – p – 1	$CMR = \frac{SCR}{n-p-1}$
Totale	$SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$	n – 1	/

Remarque *Deux situations extrêmes peuvent survenir :*

- Dans le meilleur des cas, $SCR = 0$ et donc $SCT = SCE$: les variations de Y sont complètement expliquées par celles de X. On a un modèle parfait, la droite de régression passe exactement par tous les points du nuage $(\hat{y}_i = y_i)$.
- Dans le pire des cas, $SCE = 0$: X n'apporte aucune information sur Y [43].

1.4.2 Coefficient de détermination R^2

D'une valeur comprise entre 0 et 1, le coefficient de corrélation multiple R^2 mesure la qualité d'adéquation entre le modèle et les données observées. Il permet de quantifier la force de la relation linéaire entre la variable dépendante et les variables indépendantes. Il est défini par le rapport [50]:

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^{n}(\varepsilon_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (2.16)$$

Bien évidemment $0 \leq R^2 \leq 1$, plus il tend vers 1 meilleur sera le modèle. Lorsqu'il est proche de 0, cela veut dire que les exogènes X_i n'expliquent en rien les valeurs prises par Y.

1.4.3 Coefficient de détermination ajusté R^2_{adj}

Le R^2 est un indicateur de qualité, mais il présente un défaut : plus nous augmentons le nombre des variables explicatives, même non pertinentes, n'ayant aucun rapport avec le problème que l'on cherche à résoudre, plus grande sera sa valeur, mécaniquement, et de même temps, nous diminuons le degré de liberté.

Il faudrait donc intégrer cette dernière notion pour opposer l'évolution du R^2. C'est exactement ce que fait le R^2_{adj} [49] :

$$R^2_{adj} = 1 - \frac{CMR}{CMT} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} \tag{2.17}$$

Il peut s'exprimer en fonction du R^2 selon:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2) \tag{2.18}$$

R^2_{adj} nous permet de comparer la qualité de deux modèles avec un nombre d'observations différent.

1.5 Tests de signification

Une fois qu'on a estimé les paramètres du modèle, on doit répondre à deux questions :

1. Quelle est l'adéquation globale du modèle ?
2. Quelles sont les variables individuelles qui semblent significatives ?

Pour répondre à ces deux questions critiques, on doit effectuer les tests d'hypothèses suivants.

1.5.1 Test de la significativité globale de la régression

Le test de significativité globale consiste à vérifier si le modèle, pris dans sa globalité, est pertinent. L'hypothèse nulle correspond à la situation où aucune des exogènes n'emmène de l'information utile dans l'explication de *Y, c*-à-d le modèle ne sert à rien. Le test s'écrit [49]:

$$\begin{cases} H_0 : \beta_j = 0, & j = 1, \dots, p \\ H_1 : \beta_j \neq 0 \end{cases}$$

La statistique de test est extraite du tableau d'analyse de variance (Tab 2.5), elle s'écrit :

$$F = \frac{CME}{CMR} = \frac{SCE/p}{SCR/(n-p-1)} \quad (2.19)$$

On peut l'exprimer à partir du coefficient de détermination :

$$F = \frac{R^2/p}{1-R^2/(n-p-1)} \quad (2.20)$$

Où F suit une loi de Fisher avec p et $(n-p-1)$ degrés de liberté. [43]
On rejette H_0 si :

$$F \geq F^{1-\alpha}_{(p,n-p-1)}$$

- $F^{1-\alpha}_{(p,n-p-1)}$ est le quantile d'ordre (1-α) d'une loi de Fisher à $(p, n-p-1)$ ddl.
- Le rejet de l'hypothèse H_0 implique qu'au moins un des régresseurs $x_1, x_2, \dots, x_p$ contribue de manière significative au modèle [46].

1.5.2 Test de significativité d'un coefficient

Après avoir déterminer qu'au moins un des régresseurs est important, la question qui se pose maintenant est lequel (ou lesquels) ?

Pour savoir si une variable joue un rôle explicatif dans un modèle, on effectue un test de Student. Ce test appelé aussi test de significativité partielle. Il permet d'évaluer l'influence de X_i sur Y. Pour faire un test de Student, il faut vérifier au préalable que les erreurs suivent une loi normale [51] :

$$\varepsilon_i \equiv N(0, \sigma_\varepsilon^2) \quad (2.21)$$

Le test s'écrit:

$$\begin{cases} H_0: \beta_j = 0 \Rightarrow le\ coefficient\ n'est\ pas\ significatif \\ H_1: \beta_j \neq 0 \Rightarrow le\ coefficient\ est\ significatif \end{cases}$$

La statistique de test s'écrit :

$$T_{\hat{\beta}} = \frac{\widehat{\beta_J}}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\widehat{\beta_J}}{se(\hat{\beta}_j)} \qquad (2.22)$$

Où $se(\hat{\beta}_j)$ est l'écart type de $\widehat{\beta_J}$. Cette statistique suit une loi de Student avec (n-p-1) degrés de liberté. On rejette H_0 si :

$$\left|T_{\hat{\beta}}\right| \geq t_{(n-p-1)}^{1-\frac{\alpha}{2}}$$

- $t_{(n-p-1)}^{1-\frac{\alpha}{2}}$ est la quantile de niveau $(1-\frac{\alpha}{2})$ d'une loi du Student à $(n-p-1)$ ddl.
- Si on rejette H_0 et on accepte H_1 : le coefficient est significativement différent de zéro donc la variable joue un rôle explicatif dans le modèle.
- Dans le cas où l'on ne rejette pas cette hypothèse nulle, on dit que X_j n'est pas significatif au sein du modèle. Ceci implique qu'un modèle plus simple, sans cette variable, peut être considéré [52].

1.6 Intervalle de confiance

On peut maintenant construire l'intervalle de confiance pour les coefficients de régression β_j ; Pour tout $j \in \{1, \dots, p\}$, un intervalle de confiance de niveau $(1-\alpha)$ pour β_j est :

$$\left[\hat{\beta}_j - t_{n-p-1}^{1-\frac{\alpha}{2}} se(\hat{\beta}_j), \hat{\beta}_j + t_{n-p-1}^{1-\frac{\alpha}{2}} se(\hat{\beta}_j)\right] \qquad (2.23)$$

Où $t_{n-p-1}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1-\frac{\alpha}{2})$ d'une loi Student T_{n-p-1} [53].

1.7 Analyse des résidus

Une fois le modèle mis en œuvre, on doit vérifier la présence des données aberrantes qui peuvent être influentes ou pas. Une observation influente est donc une observation qui, enlevée, conduit à une grande variation dans l'estimation des coefficients, et *vice versa*.

D'ailleurs on peut étudier :

1.7.1 Effet Levier

L'estimation des paramètres d'ajustement est très sensible à la présence de points extrêmes susceptibles d'affecter sérieusement les résultats. L'impact d'un point éloigné de la variable explicative, est donné par le levier défini pour chaque observation [54] :

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad avec \quad \sum_{i=1}^{n} h_{ii} = p + 1 \quad et \quad \frac{1}{n} \leq h_{ii} \leq 1 \qquad (2.24)$$

Un point x_i présente un effet levier important si $h_{ii} > 2(p + 1)/n$ ou $h_{ii} > 3(p + 1)/n$. L'effet levier apparait pour des observations ayant des valeurs de la variable explicative loin du centre. Autrement dit, la distance $x_i - \bar{x}$ est élevée. Une observation avec un h_{ii} proche de 1 est une observation avec un levier extrêmement important.

Un point ayant un effet levier important est susceptible de masquer ou bien de créer une fausse corrélation.

1.7.2 Cook's D

La distance de Cook mesure l'influence de l'observation i sur l'estimation du paramètre β. Pour établir une telle mesure, nous considérons la distance entre le coefficient estimé $\hat{\beta}$ et le coefficient $\hat{\beta}_i$ que l'on estime en enlevant l'observation i, mais en gardant le même modèle et toutes les autres observations bien évidemment. Si la distance est grande, alors l'observation i influence beaucoup l'estimation de β, puisque la laisser ou l'enlever conduit à des estimations éloignées [47].

Il est possible de montrer que la distance de Cook peut être réécrite comme une fonction des leviers et des et des résidus :

$$C_i = \frac{h_{ii}}{p(1 - h_{ii})^2} \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2} \qquad (2.25)$$

On considère en général que C_i est importante si elle est supérieure à 1.

2. Réseaux de neurones artificiels

2.1 Neurone formel

2.1.1 Présentation et historique

L'évolution technologique durant les dernières années a permis aux scientifiques et aux chercheurs d'élaborer et de perfectionner des méthodes pour différents domaines [55].
Les méthodes connexionnistes ont été initialisées à l'ère de la cybernétique. L'objectif des chercheurs était de construire une machine capable de reproduire le plus fidèlement possible certains aspects de l'intelligence humaine. Dès 1943, Mac Culloch et Pitts ont proposé des neurones formels mimant les neurones biologiques et capables de mémoriser des fonctions booléennes simples [56].
Les réseaux de neurones formels sont à l'origine d'une tentative de modélisation mathématique du cerveau humain. Ils présentent un modèle assez simple pour les neurones et explorent les possibilités de ce modèle [57]. Ils sont conçus pour reproduire certaines caractéristiques des mémoires biologiques par le fait qu'ils sont :

- Massivement parallèles;
- Capables d'apprentissage;
- Capables de mémoriser l'information dans les connexions interneurones ;
- Capables de traiter des informations incomplètes [56].

2.1.2 Neurone biologique

Le neurone biologique est une cellule vivante spécialisée dans le traitement des signaux électriques. Les neurones sont reliés entre eux par des liaisons appelées axones. Ces axones vont eux-mêmes jouer un rôle important dans le comportement logique de l'ensemble. Ils conduisent les signaux électriques de la sortie d'un neurone vers l'entrée (synapse) d'un autre neurone. Les neurones font une sommation des signaux reçus en entrée et en fonction du résultat obtenu vont fournir un courant en sortie [58].

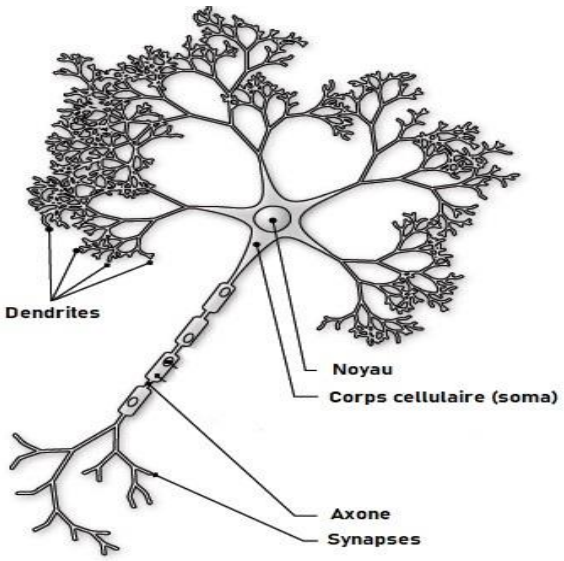


Figure 2.3 : Neurone biologique [59]

Le neurone biologique (Fig 2.3) comprend [56] :

- Les dendrites, qui sont les récepteurs principaux du neurone, captant les signaux qui lui parviennent ;
- Le corps cellulaire (ou soma), qui fait la somme des influx qui lui parviennent ; si cette somme dépasse un certain seuil, il envoie lui-même un influx par l'intermédiaire de l'axone ;
- L'axone, qui permet de transmettre les signaux émis par le corps cellulaire aux autres neurones ;
- Les synapses, qui permettent aux neurones de communiquer avec les autres via les axones et les dendrites.

Le tableau ci-dessous montre l'analogie entre le neurone biologique et le neurone formel :

Tableau 2.2 : Analogie entre le neurone biologique et le neurone formel [60]

Neurone biologique	Neurone formel
Dendrites	Signal d'entrée
Soma	Fonction d'activation
Axones	Signal de sortie
Synapses	Poids de connexion

2.1.3 Neurone artificiel (formel)

C'est l'élément de base d'un réseau de neurones [56]. Par définition, un neurone formel est une fonction algébrique non linéaire et bornée, dont la valeur dépend de paramètres appelés coefficients ou poids [61].

Un neurone lorsqu'il est activé, effectue la somme de ses entrées $(x_0, x_1, x_2, \dots, x_n)$, préalablement pondérées par leur coefficient synaptique w_i, et applique la fonction d'activation au résultat $f(y)$, pour déterminer sa valeur de sortie y [62] :

$$y = \sum_{i=0}^{n} w_i\, x_i = \sum_{i=1}^{n} w_i\, x_i + \theta \tag{2.26}$$

$$z = f(y) \tag{2.27}$$

Où $\theta = w_0$: le seuil (ou biais) propre au neurone qui est un nombre réel et qui représente la limite à partir de laquelle le neurone s'activera. Ce seuil peut jouer le rôle de poids de la connexion qui existe entre l'entrée fixée à "+1" et le neurone [63].

La figure ci-dessous résume la composition d'un neurone artificiel :

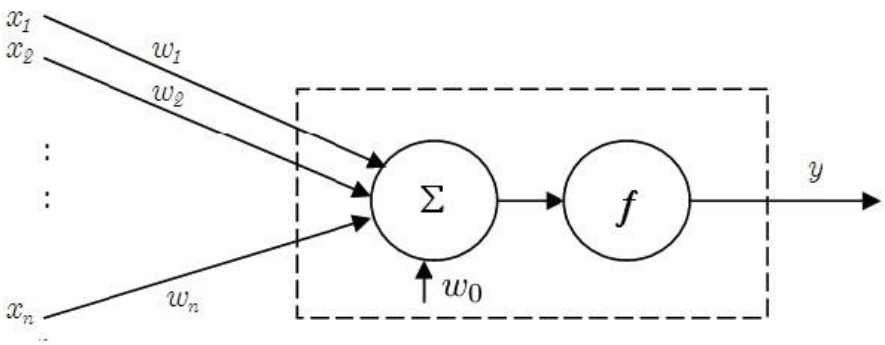
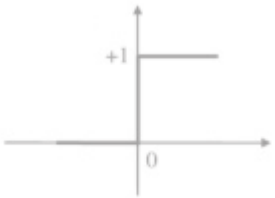
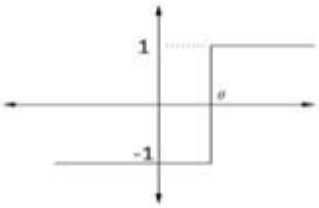
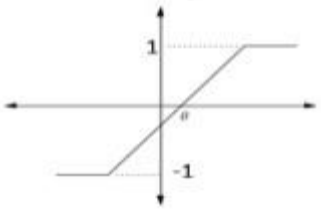
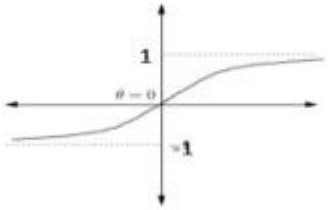
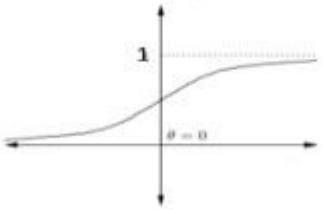


Figure 2.4 : Structure d'un neurone artificiel [64]

Dans le neurone de McCulloch et Pitts, la fonction d'activation f est du type fonction à seuil prenant les valeurs 0 ou 1. Le seuil de déclenchement est en général provoqué par une entrée inhibitrice x_0, parfois appelée biais [56]. Les fonctions d'activation les plus utilisées sont présentées dans le tableau suivant :

Tableau 2.3 : Fonctions de transfert $z = f(y)$ [63]

Nom de la fonction	Relation d'entrée/sortie	Graphe
Seuil	$z = 0 \quad si\ y < 0$ $z = 1 \quad si\ y \geq 0$	
Signe	$z = -1 \quad si\ y < 0$ $z = 1 \quad si\ y \geq 0$	

Linéaire saturée	$z = 0 \quad si\ y < 0$ $z = y \quad si\ 0 \leq y \leq 1$ $z = 1 \quad si\ y > 1$	
Tangente Hyperbolique	$z = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}$	
Sigmoïde	$z = \frac{1}{1 + exp^{-y}}$	

La fonction d'activation (la fonction de transfert) joue un rôle très important dans le comportement du neurone. Elle a comme paramètre la somme pondérée des entrées ainsi que le seuil d'activation. La nature de cette fonction diffère selon le réseau [57].

2.2 Apprentissage des réseaux de neurones

Le meilleur moteur d'intelligence artificielle n'est rien sans apprentissage [65]. L'apprentissage automatique (Machine Learning) est une application de l'intelligence artificielle (IA) qui offre aux systèmes la possibilité d'apprendre et de s'améliorer automatiquement à partir de l'expérience sans être explicitement programmé. Il se concentre sur le développement de programmes informatiques pouvant accéder aux données et les utiliser pour eux-mêmes.

On appelle « apprentissage » des réseaux de neurones la procédure qui consiste à estimer les paramètres (poids) des neurones du réseau, afin que celui-ci remplisse au mieux la tâche qui lui est affectée [66]. Les connaissances d'un réseau connexionniste sont mémorisées dans les poids qui seront déterminés lors de l'apprentissage. Le but est donc de trouver un ensemble de poids synaptiques qui minimisent l'erreur entre la sortie du réseau et le résultat désiré [67].

L'ensemble du cycle d'apprentissage automatique peut être résumé comme suit [68] :

- Acquisition des données : La première étape consiste à obtenir des données pertinentes pour l'application à développer.
- Préparation des données : Cette étape est également appelée nettoyage des données. Les données doivent être précises, propres et sécurisées.
- Sélection de l'algorithme : L'algorithme le plus approprié pour l'application à développer doit être choisi.
- Entraînement du modèle : L'algorithme retenu doit être entraîné sur les données pour créer le modèle. Le processus d'entraînement peut être supervisé ou non supervisé.
- Évaluation du modèle : Pour s'assurer que l'algorithme retenu est le mieux adapté.
- Déploiement : Il faut décider si le modèle doit être déployé dans le nuage informatique ou sur place.
- Test : Le modèle doit être testé avec des données nouvelles et pour faire des prédictions.
- Évaluation finale : La validité des prédictions établies par le modèle doit être évaluée, et le raffinement des données, du modèle et de l'algorithme doit être mis en œuvre selon qu'il convient.

Pour veiller à ce que les connexions au sein d'un réseau de neurones artificiels soient correctement établies, il faut au préalable procéder à son entraînement. On peut distinguer ici deux procédures de base : l'apprentissage supervisé et non supervisé.

2.2.1 Apprentissage supervisé

L'apprentissage supervisé est sûrement le plus courant. Le réseau reçoit/perçoit des exemples étiquetés (labeled data), à charge pour lui de les lier au mieux aux exemples de même étiquette (Label), et de les différencier des autres [69]. C'est à dire celui-ci va comparer la sortie obtenue par le réseau avec la sortie attendue, les poids w_i sont ensuite modifiés pour minimiser cette erreur, jusqu'à ce que les résultats soient satisfaisants pour tous les exemples fournis [70].

La figure ci-dessous résume ce processus :

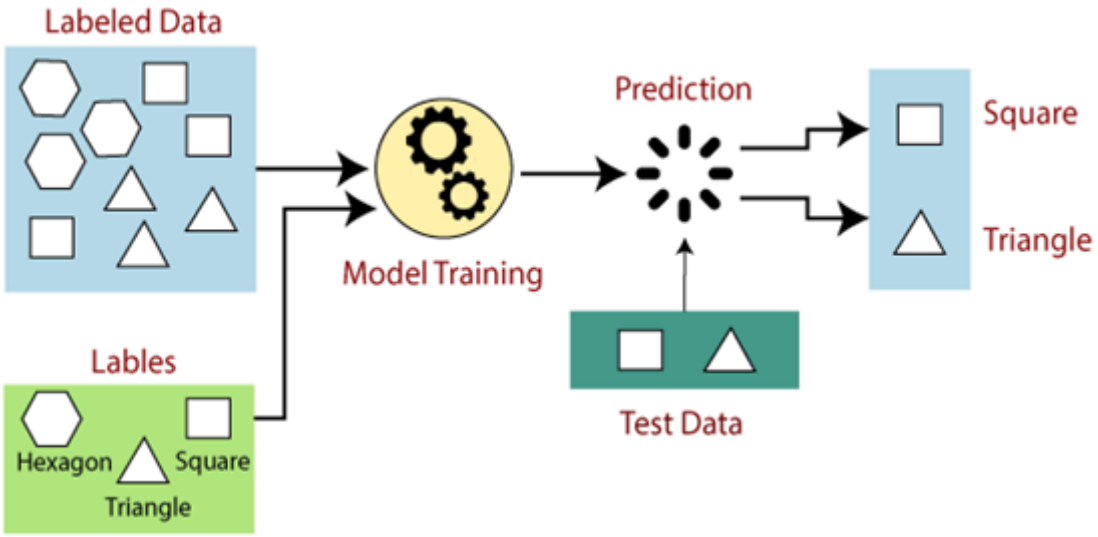
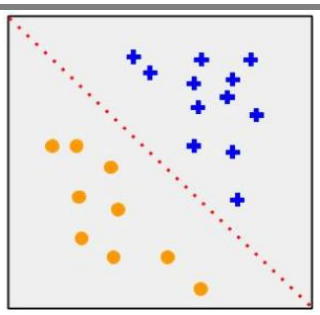
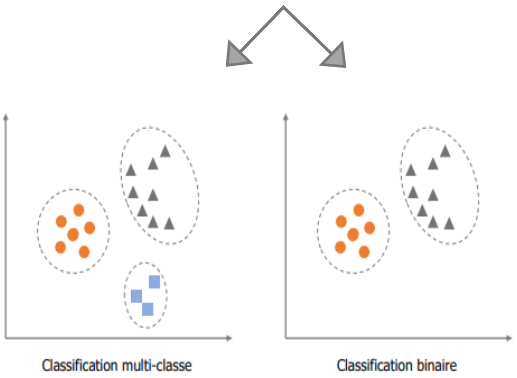
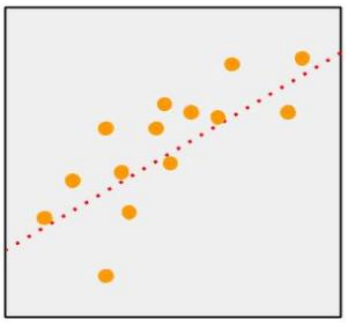


Figure 2.5 : Processus de l'apprentissage supervisé [71]

Les applications de l'Apprentissage Supervisé sont nombreuses, On peut les diviser en deux catégories de problèmes : *la régression*, et *la classification* ; qui sont résumés dans le tableau (Tab 2.4) suivant :

Tableau 2.4 : Catégories de l'apprentissage supervisé

Problèmes	Définitions	Exemples
Classification	La classification consiste à identifier les classes d'appartenance de nouveaux objets à partir d'exemples antérieurs connus, la variable à prédire peut donc prendre des valeurs discrètes de type qualitative appelées classes. [72] On distingue deux cas [73] : - Quand l'ensemble de sorties sont deux catégories, la machine doit classer ce qu'on lui donne dans deux classes, on parle de classification binaire. - Quand l'ensemble des valeurs dépasse deux éléments, on parle de classification multi-classes.	 Figure 2.6 : Représentation des observations de type classification [74]  Figure 2.7: Représentation de deux types de classification [72]
Régression	La régression permet de trouver un modèle (mathématique) en fonction des données d'entrainement. On utilise la régression lorsque la variable d'intérêt est quantitative (variable continue qui prendre une valeur réelle) [73]. La régression linéaire est le modèle le plus simple : Il consiste à trouver la meilleure droite qui s'approche le plus des données d'apprentissage [75].	 Figure 2.8 : Représentation des observations de type régression [74]

Selon le type de réseau choisi, les algorithmes d'apprentissage supervisés sont différents. Les trois principaux sont [70] :

- Descente de gradient.
- Algorithme de Widrow- Hoff.
- Rétropropagation.

Les deux premiers algorithmes ne s'appliquent qu'aux réseaux monocouches de type perceptron (avec une fonction seuil (Tab 2.3)). Les poids sont optimisés par plusieurs passes sur l'ensemble d'apprentissage dans l'algorithme de type Descente de gradient. Par contre l'algorithme de Widrow-Hoff applique la même modification, mais au lieu de la faire après avoir vu tous les exemples, celle-ci est appliquée après chaque test, il n'est pas très rapide.

Dans de nombreux cas, on utilisera des réseaux multicouche (MLP pour multilayer perceptron), avec ici une fonction d'activation sigmoïde (Tab 2.3) : l'algorithme adoptée par ce type de réseau est la rétro-propagation du gradient (Backpropagation en anglais) :

La phase d'apprentissage des MLP consiste à adapter les poids des connexions en fonction des erreurs de prédiction constatées à chaque classification d'une nouvelle instance. Cet algorithme permet de déterminer le gradient de l'erreur pour chaque neurone du réseau en partant de la dernière couche et en arrivant jusqu'à la première couche cachée. L'objectif est d'ajuster les poids des connexions dans le but de minimiser l'erreur quadratique [76]

$$E = \frac{1}{2}\sum_{i=1}^{N}(ref_i - hyp_i)^2 \qquad (2.28)$$

qui représente l'écart entre la sortie attendue (la référence) et la sortie produite par le réseau (hypothèse) correspondant à un vecteur d'entrée donné. *N* représente la taille des vecteurs en sortie.

2.2.2 Apprentissage non supervisé

Au contraire en apprentissage non supervisé, l'apprenant reçoit/perçoit des exemples, sans autre indication, à charge pour lui de les lier à ce qu'il sait déjà [69]. On utilise cette forme d'apprentissage pour faire du *clustering* : on a un ensemble de données, et on cherche à déterminer des classes de faits [70].

Dans ce cas il n'y a pas de réponse correcte. Les algorithmes sont laissés à leurs propres mécanismes pour découvrir et présenter la structure intéressante des données.

La question qui se pose alors est de savoir comment on identifie les classes à partir des observations de données. La figure ci-dessous explique comment ça marche le processeur de ce type :

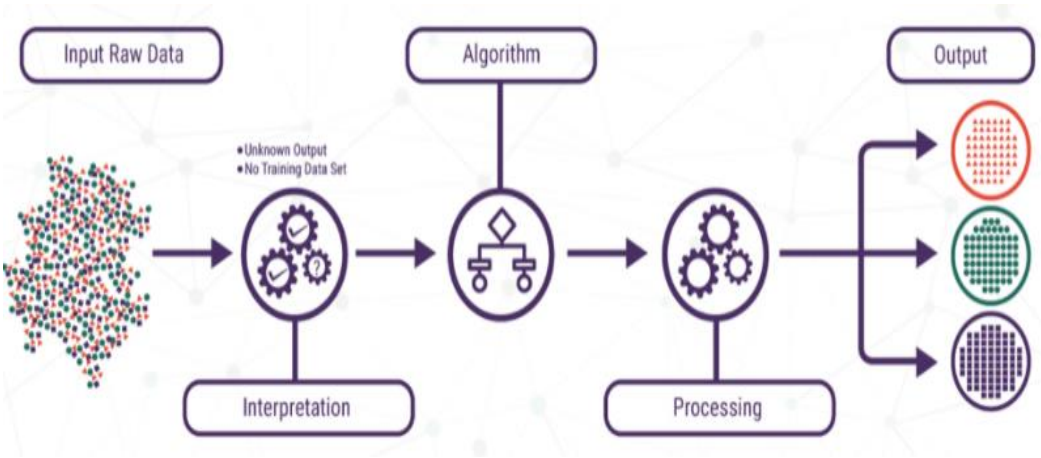
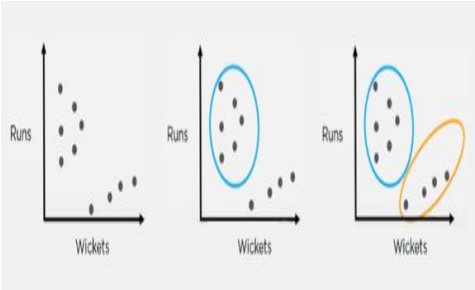


Figure 2.9 : Processus de l'apprentissage non supervisé [71]

L'apprentissage non supervisé comprend deux catégories d'algorithmes : Algorithmes *de regroupement* et *d'association* ; Sont présentées dans le tableau suivant :

Tableau 2.5 : Catégories de l'apprentissage non supervisé.

Problèmes	Définitions	Exemples
Regroupement (Clustering)	La mise en cluster consiste à séparer ou à diviser un ensemble de données en un certain nombre de groupes, c'est-à-dire partageant un certain nombre de caractéristiques identiques [77]. En outre, le clustering permet de mettre en évidence ces regroupements sans connaissance a priori sur les données traitées.	 Figure 2.10 : Représentation des observations de type clustering. [78]

Association	Dans un problème d'association qui permet a trouver des relations importantes entre les variables ou les caractéristiques d'un ensemble de données [79].	 Figure 2.11 : Représentation des observations de type association [78]

2.3 Réseaux de neurons

2.3.1 Définition

Haykin offre la définition suivante d'un réseau de neurones [80] :

Un réseau de neurones est un processeur distribué massivement parallèle qui a une propension naturelle pour emmagasiner la connaissance expérientielle et à la rendre disponible pour utilisation ultérieure. Il ressemble au cerveau selon deux aspects :

1. *La connaissance est acquise par le réseau à travers un processus d'apprentissage.*
2. *Les forces de connexion interneurones appelées poids synaptiques sont utilisées pour l'emmagasinage de l'information.*

2.3.2 Architecture

Un réseau de neurones est constitué de plusieurs couches de neurones connectées entres elles : la couche d'entrée (Input Layer) récolte d'abord ses informations à partir de capteurs disposés en dehors du réseau puis transmis l'information sur la couche suivante qui est elle-même connectée à une autre couche, et ainsi de suite jusqu'à la couche de sortie (Output Layer). Les couches trouvées entre celles d'entrée et de sortie sont appelées les couches cachées (Hidden Layers). Chaque couche i est ainsi composée d'une quantité n_i de neurones connectés en entrée sur les $n_i - 1$ neurones de la précédente couche. Ces connexions entre les neurones sont souvent appelées synapses. Un réseau de neurone se présente comme suit [62] :

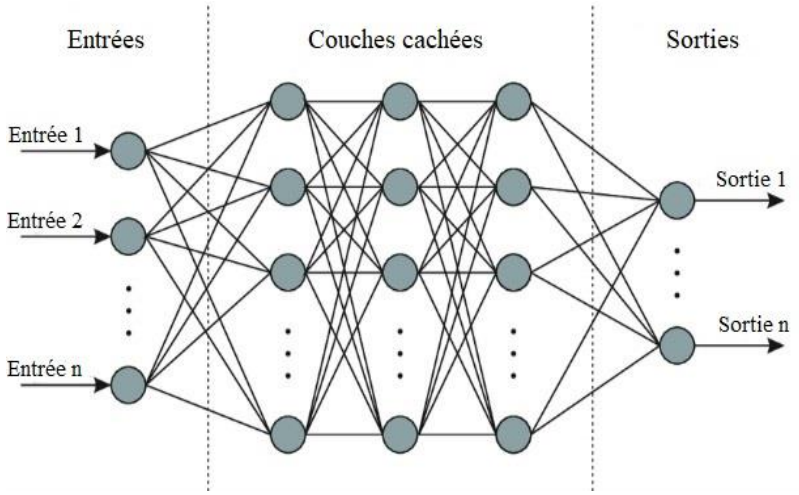


Figure 2.12 : Architecture d'un réseau de neurones [81]

Une couche est un ensemble de neurones n'ayant pas de connexion entre eux [82], le nombre de ses couches définit le type de réseau. On distingue :

A. Perceptron Monocouche

Le Perceptron, probablement le plus ancien modèle de calcul neuronal, est inventé par F. Rosenblatt et date de 1958 [83], il est considéré comme un des algorithmes d'apprentissage supervisé les plus simple pour la classification binaire. Ce réseau neuronal contient une seule couche d'entrée et un nœud de sortie. D'autre part, il est la première application reconnue du principe des réseaux neuronaux introduits par Pitts et McCulloch en 1943. Il existe plusieurs types de Perceptron, toutefois sous sa version la plus simple, il est conçu à partir d'une seule couche constituée d'un unique neurone connecté à n entrées [62] :

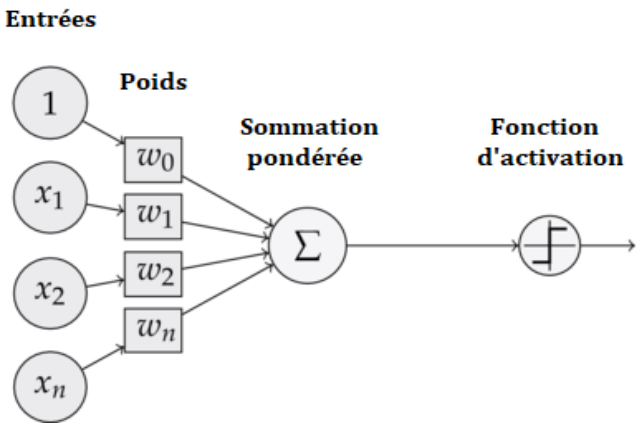


Figure 2.13 : Perceptron monocouche [84]

Notons la sortie de sommation du $i^{ème}$ neurone par y_i la relation de sommation du Perceptron est donnée par [83] [56]:

$$y_i = \sum_{j=1}^{n} w_{ij}x_j + w_{i0}x_0 \tag{2.29}$$

$$z_i = f(y_i) \tag{2.30}$$

Où w_{ij} étant le poids synaptique de la $j^{ième}$ entrée de la $i^{ème}$ neurone. L'équation (Eq. 29) peut être écrite sous forme vectorielle comme suit :

$$y_i = \mathbf{w}_i^T \mathbf{x}_i + w_{i0}x_0$$

Avec : $\mathbf{w}_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$
Et : $\mathbf{x}_i = [x_1, x_2, \dots, x_n]^T$

T désignant la transposée de w et x_0 est l'entrée inhibitrice constante $x_0 = 1$.

1) Fonction d'activation

La première fonction d'activation adoptée par les fondateurs McCulloch et Pitts était une fonction de seuillage brutale de type $sign(y)$, dès lors, la sortie peut prendre que des valeurs binaires (-1 *ou* 1) et cela est utile pour les tâches de classifications. Récemment, la fonction continue de forme sigmïdale est plus utilisée, on la retrouve sous deux formes, soit la fonction log-sigmoïde [80] :

$$f(y) = \frac{1}{1 + e^{-y}}$$

Soit la tangente hyperbolique:

$$f(y) = tanh = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}$$

Elles diffèrent principalement par leur intervalle de sortie, soient respectivement [0,1] et [-1,1]. Leurs formes graphiques sont illustrées dans le tableau (Tab 2.7).

2) Limites du Perceptron

Le perceptron permet de calculer les opérateurs logiques AND et OR. Rappelons que, si l'on note 0 le booléen « faux » et 1 le booléen « vrai », leurs tables de vérité sont les suivantes [85] :

x_1		x_2	y
0	AND	0	0
0	AND	1	0
1	AND	0	0
1	AND	1	1

x_1		x_2	y
0	OR	0	0
0	OR	1	1
1	OR	0	1
1	OR	1	1

Les calculs de **AND**, **OR**, ou **XOR** peuvent être vus comme des problèmes de classification :

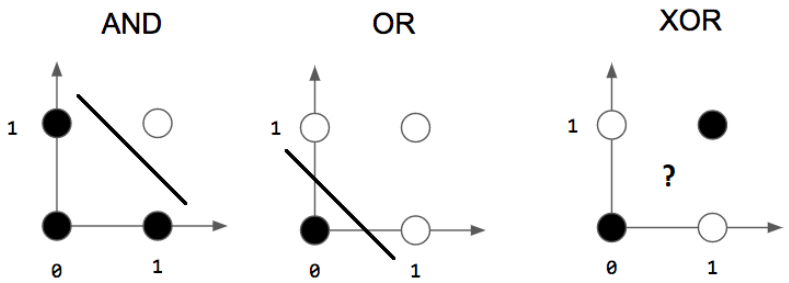


Figure 2.14 : Les opérateurs logiques AND, OR et XOR

Comme l'illustre la figure (Fig 2.14), le calcul de ces deux opérateurs revient à un problème de classification pour deux classes linéairement séparables. Néanmoins, le perceptron ne peut pas calculer l'opérateur logique XOR (« OU exclusif »), qui a pour table de vérité :

x_1		x_2	y
0	XOR	0	0
0	XOR	1	1
1	XOR	0	1
1	XOR	1	0

Le perceptron, comme étant un classifieur linéaire, ne permet pas donc de réaliser le OU exclusif (XOR). Ce problème a été souligné par Minsky et Papert. Les réseaux multicouches permettent de pallier à cette limitation.

B. Perceptron Multi-couche

Pour surmonter les limites soulignées par Minsky et Papert, qui à l'époque entraînaient une grande déception avec les RNA et une forte baisse (presque totale) des recherches sur ces derniers, il était nécessaire d'aller au-delà du réseau de neurones monocouche [83].

Les réseaux neuronaux multicouches MLP contiennent plus d'une couche de calcul. Contrairement au perceptron dont la couche de sortie est la seule couche effectuant les calculs qui sont complètement visibles pour l'utilisateur. Les réseaux multicouches contiennent plusieurs couches de calcul. Les couches intermédiaires supplémentaires (entre l'entrée et la sortie) sont appelées couches cachées car les calculs effectués ne sont pas visibles pour l'utilisateur [85] [86] :

- La couche d'entrée est composée de neurones qui lisent les composantes d'un vecteur et envoient l'information aux neurones de la première couche cachée.

- Chaque neurone d'une couche cachée fait une moyenne pondérée des informations reçues et réémet aux neurones de la couche suivante cette information moyenne modifiée par une fonction d'activation.

- Plusieurs couches cachées peuvent se succéder, chacune étant constituée d'un certain nombre de neurones connectés aux neurones de la couche précédente.

- Enfin, la dernière couche émet le vecteur ou le scalaire de sortie du réseau. Sa nature exacte dépend du problème traité (classification ou régression).

L'architecture spécifique des réseaux de neurones multicouches est un réseau à propagation directe (*feed-forward neural network*), car l'information se propage de la couche d'entrée à la couche de sortie :

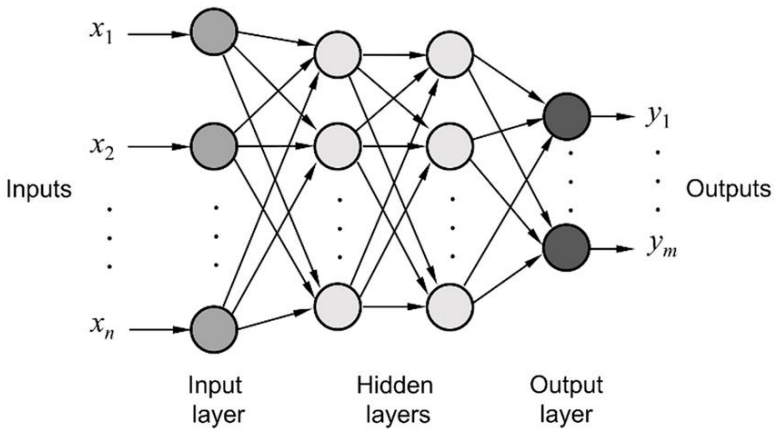


Figure 2.15 : Structure de PMC « *feed-forward* » [87]

Si nous considérons un PMC avec n neurones d'entrée, activés par un vecteur d'entrée x (de taille n), et par $w_{ij}^{0,1}$le poids correspondant à la connexion entre le neurone j de la couche 0 et le neurone i de la couche 1, la sortie z_i^1 de chacun des neurones de la première couche cachée sera exprimée par [88] :

$$z_i^1 = f_i(y_i^1) \tag{2.31}$$

$$y_i^1 = \sum_{j=1}^{n} w_{ij}^{0,1}\, x_j + \theta_i^1 \tag{2.32}$$

Où f_i est la fonction d'activation décrite précédemment, et θ_i^1 est le biais, dont le rôle est de rajouter un degré de liberté supplémentaire en agissant sur la position de la frontière de décision.

On peut généraliser l'équation (Eq. 32) à toutes les couches suivantes y compris la couche de sortie, tandis que chaque sortie d'une couche l joue le rôle de couche d'entrée pour la couche suivante $l+1$:

$$z_i^l = f_i(y_i^l) \tag{2.33}$$

$$y_i^{l,l+1} = \sum_{j=1}^{L} w_{ij}^{l,l+1}\, z_j^l + \theta_i^{l+1} \tag{2.34}$$

Où L est le nombre de neurones de la couche l.

3. Conclusion

Dans la deuxième partie de la présente étude, nous avons mis en évidence les méthodes statistiques les plus répandus. Dans un premier temps, on a expliqué les hypothèses nécessaires et les termes du modèle RLM, les notions d'estimation des paramètres du modèle par la méthode MCO, l'estimation par intervalle de confiance, les testes de signification des paramètres et la signification globale du modèle.

Par la suite nous avons donné une brève introduction (apprentissage et son architecture) à un outil d'apprentissage automatique qui est RNA.

Chapitre 3

Traitement, analyse et interprétation des données

Ce chapitre présente les principaux résultats de ce mémoire. Il y aura également une étude comparative des deux méthodes utilisées.

Dans ce travail on va essayer de trouver le meilleur modèle de QSAR qui permet d'expliquer la corrélation entre les paramètres physicochimiques et électroniques et les activités biologiques IC_{50} des dérivés d'anilide. Ce modèle sera développé par deux méthodes : la régression linéaire multiple et les réseaux de neurones artificiels. Une étude comparative sera réalisée pour déterminer leurs capacités prédictives ainsi que leurs avantages et inconvénients.

1. Régression linéaire multiple

1.1 Traitement des données

Les travaux de D. Zakarya et al. ont permis de rassembler les données d'activité inhibitrice de système PSII *in vitro* (IC_{50}) de soixante-seize dérivés d'anilide qui sont extraits de l'article « QSARs For A Series Of Inhibitory Anilides. (1997) ». On vise à construire un modèle mathématique capable d'expliquer et de prédire cette activité en fonction des descripteurs moléculaires.

Les 76 composés utilisés dans cette étude, ainsi que les valeurs de pIC_{50} sont présentées dans l'annexe (Annexe A). La structure modèle de ces molécules est représentée sur la figure suivante :

Figure 3.1 : Structure chimique des anilides

L'activité biologique inhibitrice est rapportée en terme IC_{50} : concentration micro molaire nécessaire pour inhiber 50% de l'activité photosynthétique de PS II. Pour notre cas, nous avons exprimé l'activité inhibitrice par le rapport logarithmique pIC_{50} [$\log(1/IC_{50})$].

Les calculs de modélisations moléculaires ont été exécutés en utilisant :

Logiciel HyperChem 7 Professional

HyperChem [89] est un environnement de modélisation moléculaire sophistiqué qui est connu pour sa qualité, sa flexibilité et sa facilité d'utilisation. Unissant la visualisation et l'animation 3D. Il possède plus de méthodes de calculs (mécanique moléculaire, semi-empirique et ab-initio) pour que vous puissiez calculer plus de propriétés.

Logiciel GaussView 5.0

GaussView [90] est l'interface graphique la plus avancée et la plus puissante disponible pour Gaussian. GaussView permet d'importer ou de construire les structures moléculaires, configurer, lancer, surveiller et contrôler les calculs gaussiens, et récupérer et visualiser les résultats. Il fournit des fonctionnalités pour chaque phase d'étude de grands systèmes moléculaires, de l'importation de molécules à partir de fichiers MDL, Puis le choix de la méthode de calcul DFT pour optimiser les molécules et calculer leur descripteur.

Molinspiration

Molinspiration [91] est un site de recherche en ligne indépendant axé sur le développement et l'application de techniques modernes de cheminformatique. Il propose une large gamme d'outils d'aide à la manipulation et au traitement des molécules, pour le calcul des propriétés physico-chimiques moléculaires pertinentes.

1.1.1 Calcul des descripteurs

Dans le but d'optimiser les molécules, nous avons dessiné ces dernières par le logiciel HyperChem. La géométrie des dérivés de l'anilide a été initialement optimisée entièrement par la mécanique moléculaire, avec le champ de force MM+.

Une fois les molécules sont stabilisées, nous avons calculé leur surface (surface area grid) et volume par le module 'QSAR propreties' intégré dans HyperChem.

Ensuite, nous avons ré-optimisé nos composés en utilisant GaussView par la théorie de fonctionnelle de la densité (DFT) en utilisant le niveau de théorie B3LYP/6-31G. Cette théorie a été utilisée pour calculer un certain nombre des descripteurs électroniques tel que : le moment dipolaire (MD) et les charges des atomes Q_4, Q_5 et $Q_{2'}$.

Enfin, quelques descripteurs physicochimiques ont été calculés par Molinspiration en ligne : le coefficient de partage (log P), le nombre des liaisons hydrogène accepteurs (HBA) et le nombre des liaisons rotatives (nrotb).

Les descripteurs calculés sont listés dans le tableau suivant.

Tableau 3.1 : Descripteurs physicochimiques et électroniques

Mol	log P	HBA	nrotb	Q_4	Q_5	$Q_{2'}$	MD	Volume	Surface
1	3.28	2	2	-0.069	-0.031	0.666	0.9915	609.00	390.45
2	3.52	2	2	-0.059	-0.040	0.668	1.7834	657.21	416.73
3	3.79	2	3	-0.059	-0.040	0.673	1.6615	691.94	427.34
4	4.34	2	4	-0.062	-0.038	0.669	2.0070	726.33	461.57
5	**4.05**	**2**	**3**	**-0.061**	**-0.039**	**0.669**	**1.8839**	**711.06**	**440.68**
6	4.58	2	4	-0.060	-0.040	0.672	2.0463	762.57	472.76
7	4.29	2	4	-0.061	-0.039	0.670	1.9558	761.79	470.48
8	**4.39**	**2**	**4**	**-0.041**	**-0.239**	**0.679**	**1.2885**	**747.83**	**472.07**
9	4.36	2	4	-0.048	-0.004	0.670	3.5569	767.93	474.57
10	**4.85**	**2**	**5**	**-0.064**	**-0.038**	**0.667**	**4.7798**	**782.16**	**494.31**
11	4.32	2	4	-0.062	-0.039	0.670	2.0066	759.34	477.35
12	4.69	2	2	-0.065	-0.038	0.669	4.6573	758.60	466.63
13	5.41	2	4	-0.062	-0.039	0.671	2.0001	868.22	529.67
14	4.72	2	4	-0.058	-0.041	0.661	1.5958	795.48	470.23
15	4.08	3	4	-0.060	-0.040	0.645	2.6763	742.56	431.03
16	1.94	2	2	-0.254	-0.230	0.664	3.4417	591.31	368.93
17	**2.60**	**2**	**2**	**-0.268**	**-0.011**	**0.666**	**3.0098**	**634.90**	**395.73**
18	2.73	2	2	-0.263	-0.103	0.666	2.9046	653.70	407.42
19	1.97	3	3	-0.335	0.314	0.665	4.9828	670.48	414.48
20	**2.81**	**2**	**3**	**-0.221**	**-0.168**	**0.666**	**3.6488**	**676.39**	**416.57**
21	3.63	3	6	-0.334	0.319	0.664	5.1186	881.92	536.20
22	**2.10**	**2**	**2**	**0.384**	**-0.291**	**0.664**	**1.5925**	**603.16**	**378.98**
23	**2.62**	**2**	**2**	**-0.036**	**-0.243**	**0.665**	**1.1766**	**634.22**	**395.82**
24	2.00	3	3	0.297	-0.263	0.662	4.3833	670.25	409.86
25	1.70	3	2	-0.187	-0.180	0.668	2.4981	651.83	405.59
26	3.23	2	2	-0.064	-0.038	0.667	2.2675	675.96	419.26
27	3.67	3	6	-0.331	0.318	0.655	5.3475	843.34	520.66
28	3.82	3	6	-0.334	0.313	0.655	4.8920	844.25	530.45
29	4.34	3	7	-0.331	0.316	0.655	5.0856	939.96	561.32
30	5.12	3	9	-0.331	0.318	0.655	5.3675	1049.87	633.82
31	3.97	4	8	-0.328	0.319	0.655	4.2874	986.53	599.59
32	4.24	4	9	-0.331	0.318	0.655	6.0331	1039.73	630.22
33	2.92	2	3	-0.242	-0.233	0.668	4.0624	651.40	394.86
34	3.06	2	3	-0.305	0.403	0.669	3.0168	660.73	400.61
35	4.61	3	7	-0.325	0.317	0.669	5.7812	928.77	517.45
36	3.60	2	3	-0.032	-0.243	0.670	1.0176	699.16	418.77
37	3.37	2	3	-0.032	-0.229	0.669	4.1712	696.80	409.63
38	**4.20**	**2**	**3**	**-0.055**	**-0.043**	**0.669**	**1.3540**	**725.91**	**424.74**

39	3.06	2	4	-0.251	-0.229	0.666	3.6158	633.20	429.06
40	**2.56**	**3**	**4**	**-0.329**	**0.319**	**0.667**	**4.8792**	**663.82**	**450.16**
41	3.09	3	5	-0.335	0.317	0.668	5.0768	710.76	524.88
42	3.97	3	7	-0.336	0.315	0.662	3.2759	832.40	520.85
43	**3.83**	**3**	**6**	**-0.338**	**0.315**	**0.662**	**3.2433**	**828.85**	**551.29**
44	4.53	3	8	-0.337	0.315	0.665	3.3203	884.24	551.19
45	4.21	3	7	-0.336	0.316	0.662	3.2048	880.75	536.82
46	4.33	3	7	-0.332	0.319	0.667	5.3963	871.01	580.08
47	**4.75**	**3**	**8**	**-0.336**	**0.315**	**0.663**	**3.3270**	**933.64**	**580.14**
48	5.04	3	9	-0.337	0315	0.665	3.3740	940.24	583.34
49	4.89	3	8	-0.339	0.316	0.665	3.3646	932.53	575.93
50	6.04	3	11	-0.337	0.315	0.665	3.4035	1048.46	645.43
51	5.90	3	10	-0.337	0.316	0.662	3.4162	1040.58	638.09
52	5.38	3	7	-0.336	0.316	0.662	3.2767	977.73	591.31
53	3.32	5	8	-0.339	0.310	0.663	4.0456	913.25	566.98
54	3.69	5	9	-0.279	0.316	0.666	4.4819	962.09	593.16
55	3.12	3	5	0.290	-0.314	0.659	4.8942	723.07	463.02
56	3.49	3	6	0.291	-0.256	0.665	4.3845	770.52	486.22
57	3.99	3	7	0.305	-0.264	0.665	5.1101	823.54	514.64
58	3.85	3	6	0.294	-0.264	0.659	2.0026	827.05	522.07
59	4.55	3	8	0.305	-0.264	0.665	5.1816	883.60	546.65
60	4.24	3	7	0.305	-0.263	0.665	4.8629	876.02	545.29
61	4.36	3	7	0.306	-0.264	0.665	5.1549	873.99	541.24
62	4.77	3	8	0.304	-0.263	0.665	5.0884	927.58	573.53
63	5.06	3	9	0.304	-0.263	0.665	5.1082	937.37	589.75
64	4.92	3	8	0.306	-0.264	0.665	5.2109	927.27	576.32
65	4.86	3	8	0.304	-0.263	0.665	5.1408	921.29	564.91
66	6.07	3	11	0.292	-0.256	0.665	4.6336	1042.54	639.48
67	5.93	3	10	0.306	-0.265	0.665	5.3504	1035.38	541.38
68	5.40	3	7	0.307	-0.265	0.665	5.1270	974.62	588.23
69	3.34	5	8	0.296	-0.260	0.665	4.7331	909.63	563.73
70	**3.72**	**5**	**9**	**0.290**	**-0.255**	**0.659**	**6.5069**	**971.79**	**607.58**
71	4.08	5	9	0.290	-0.257	0.660	3.5945	1024.38	626.05
72	**4.99**	**3**	**8**	**-0.288**	**0.316**	**0.665**	**4.9321**	**980.49**	**600.05**
73	3.92	2	4	-0.049	-0.045	0.672	1.8608	771.12	477.47
74	**4.32**	**3**	**8**	**-0.271**	**0.313**	**0.666**	**3.6396**	**959.71**	**585.25**
75	4.34	3	8	0.311	-0.267	0.669	5.4190	978.19	599.24
76	**4.59**	**3**	**7**	**-0.320**	**0.314**	**0.657**	**5.3620**	**959.97**	**544.68**

En gras : Test set

1.1.2 Normalisation et prétraitement des données

En général, la base de données contient des valeurs brutes avec des ordres de grandeurs très différents selon les variables. Pour que le facteur d'échelle n'influe pas nos résultats, la base de données doit subir un prétraitement pour uniformiser les échelles de mesures avec conversion des données en variables standardisées.

Z-score (un outil statistique sur Matlab [92]) permet de normaliser les variables en variables centrées réduites (adimensionnelles), cela nous permettra de comparer deux ou plusieurs ensembles de données avec des unités différentes. Pour les données d'échantillon avec une moyenne $\bar{x}$ et un écart-type S, z-score d'un point de données x est [93]:

$$z = \frac{x_i - \bar{x}}{s} \tag{3.1}$$

1.1.3 Sélection des descripteurs

Pour que les modèles QSAR soit simple et compréhensible, il faut que les descripteurs employés soient significatifs et interprétables [94]. La sélection des descripteurs candidats au modèle est une étape cruciale et la qualité du modèle dépendra de leur pertinence, ils doivent apporter l'information qui peut expliquer la réponse (l'activité biologique).

Afin de réduire le nombre des descripteurs, nous avons utilisé l'algorithme Pas à pas (Stepwise) par Matlab, son principe est :

1- Sélection *forward* : À chaque pas, une variable est ajoutée au modèle. C'est celle permettant de réduire au mieux le critère BIC (le critère de qualité choisi) du modèle obtenu. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque BIC ne décroît plus [82] :

$$BIC = n \ln\left(\frac{SCR}{n}\right) + \ln(n)\,(p+1) \tag{3.2}$$

Avec : n est le nombre d'échantillon, SCR la somme des carrées résiduelles (voir Chapitre 2) et p est le nombre des paramètres du modèle.

2- Élimination *backward* : L'algorithme démarre cette fois du modèle complet. À chaque étape, la variable dont l'élimination conduit à BIC le plus faible est supprimée. La procédure s'arrête lorsque BIC ne décroît plus.

3- Mixte *stepwise* : Combinaison des deux méthodes *forward* et *backward*. Cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection.

Cette méthode a été utilisée pour sélectionner les descripteurs les plus pertinents. Les quatre descripteurs qui contribuent à l'activité inhibitrice ayant été sélectionnés sont : log P (x_1), nombre des liaisons hydrogène accepteur (x_2), la charge Q_4 (x_4) et le volume (x_8).

1.1.4 Division de la base de données

Pour construire un modèle mathématique ruboste et capable de prédire l'activité biologique, on doit l'entraîner en premier puis le valider pour tester sa capacité à prédire l'activité des nouvelles structures. En fait, si le modèle est construit sur tout l'ensemble de données, on risque d'avoir le *sur-apprentissage* ou l'apprentissage par cœur, par conséquent, le modèle apprend très bien les données présentées sans pouvoir généraliser le modèle à des données nouvelles [94]. Pour ce faire, il faut découper la base de données en deux sous-ensembles :

- Training set : pour l'entraînement et la construction du modèle. Constitué par 75% de la base de données.
- Test set : pour évaluer la capacité prédictive du modèle. Constitué par le 25% de la base de données restante.

Cette division a été réalisée par la validation croisée.

1.2 Interprétation des résultats

1.2.1 Construction du modèle

La méthode Stepwise a été utilisée pour sélectionner les descripteurs les plus pertinents. Sur ceux qui ont été sélectionnés, une analyse de régression linéaire multiple a été réalisée sur l'ensemble d'apprentissage (Training set, 61 molécules), puis évaluée par l'ensemble Test (Test set, 15 molécules).

L'équation de régression obtenue est donnée par :

$$pIC_{50} = 4.8743 + 0.0199logP - 0.5038HBA - 0.5763Q_4 + 0.4509Volume$$

$$n_{training} = 61$$

1.2.2 Signification des coefficients

Les coefficients de régression représentent le changement moyen de la variable de réponse y pour une unité de changement de la variable de prédicteur x_i tout en maintenant les autres prédicteurs constants dans le modèle.

Le test de signification des coefficients des modèles mathématiques est mené au moyen du test de Student. Rappelons qu'un coefficient est dit significatif si la variable qui lui est associée à une influence sur la réponse. Dans ce cas de figure, la p-value est la probabilité qu'un coefficient soit négligeable. On calcule cette p-value à partir du rapport du coefficient $\hat{\beta}_i$ à son écart-type $se(\hat{\beta}_i)$. Les résultats de la p-value sont portés dans le tableau ci-dessous [95] :

Tableau 3.2 : Résultats du test de Student

	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$	T	$\boldsymbol{se}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})$	*pvalue*
Intercept	4.8743	52.6025	0.0927	< 0.0001
log P	0.0199	0.0775	0.2571	0.9385
HBA	-0.5038	-2.4010	0.2098	0.0197
Q_4	-0.5763	-5.9519	0.0968	< 0.0001
Volume	0.4509	1.3915	0.3240	0.1696

La question qui se pose ici c'est : est-ce que le coefficient de partage (log P), le nombre des liaisons hydrogène accepteurs (HBA), la charge Q_4 et le volume de la molécule prédisent significativement l'activité inhibitrice pIC_{50} ? Si oui, alors laquelle qui présente la contribution majeure à cette activité ?

La p-value nous permet de répondre à cette question. A partir du tableau, on remarque que les p-values de $\hat{\beta}_2$ et $\hat{\beta}_3$ sont inférieures au seuil 5%, cela signifie qu'ils sont significativement différents de zéro donc il faut les conserver dans le modèle prédictif.

On peut également interpréter ces résultats comme suit : le nombre des liaisons hydrogène accepteur a un pouvoir explicatif sur pIC_{50}. On peut dire qu'une augmentation d'une unité de nombre de ces liaisons diminuera la valeur de pIC_{50} de la molécule par 0,5 unités, même chose pour la charge Q_4, une augmentation d'une unité de la charge diminuera la valeur pIC_{50} de la molécule par 0,6 unités, tout en maintenant les autres variables constantes.

D'un autre côté, on remarque que les p-values $\hat{\beta}_1$ et $\hat{\beta}_4$ sont supérieures au seuil ce qui indique qu'ils ne sont pas statistiquement significatives. Cependant, nous avons fait le choix de les garder dans le modèle final car ces descripteurs véhiculent une information chimique pertinente.

Si l'on veut savoir laquelle de ces variables est celle qui contribue le plus à l'activité inhibitrice, on devrait voir les valeurs de $|t_{cal}|$, plus la valeur de ce dernier est élevée, plus

la variable est dite significative et contribue davantage à cette activité (les signes + et – désignent la nature de la relation entre les variables indépendantes et la réponse). A partir de là, on peut résumer la contribution de chaque variable :

$$\boldsymbol{Q_4 > HBA > volume > logP}$$

1.2.3 Evaluation globale de la régression

L'évaluation globale de la pertinence du modèle de prédiction, s'appuie sur l'équation fondamentale d'analyse de la variance qui est donnée par :

$$SCT = SCR + SCE$$

Pour donner le tableau d'analyse de variance. Il s'agit de calculer les quantités précédentes (voir Chapitre 2). Le tableau (Tab 3.3) est donné par :

Tableau 3.3 : Tableau d'analyse de variance ANOVA de notre modèle

Source de variation	Ddl	Somme des carrés	Carrés moyens	F	p-value
Expliquée	4	SCE = 29.3311	CME = 7.3327	14.6386	$3.0338\ 10^{-8}$
Résiduelle	56	SCR = 30.6689	CMR = 0.5283		
Totale	61	SCT = 60			

SCT est la variabilité totale de la variable dépendante Y. Le but de régression est d'expliquer une partie de cette variabilité, c'est donc le rôle SCE et SCR.

A partir de tableau ANOVA, on voit que le modèle explique 29.33 (SCE) sur 60 (SCT) de variabilité totale en utilisant nos quatre variables indépendantes, cette proportion est appelée le coefficient de détermination R^2 :

$$R^2 = \frac{SCE}{SCT} = 0.511$$

Ainsi, la valeur de R^2_{adj} est de 0.48. Ils nous disent que 50% de la variation en Y est expliquée par nos variables X.

Il semble donc que log P, HBA, Q_4 et le volume de la molécule expliquent la moitié de la variation de l'activité inhibitrice. Cela indique que le modèle peut être appliqué avec succès pour prédire l'activité inhibitrice des nouvelles structures.

On retrouve aussi la statistique dite de Fisher F qui permet de tester l'ajustement du modèle :

$$F_{obs} = \frac{CME}{CMR} = 14.64 \quad \text{avec} \quad pvalue = 3.0338\ 10^{-8}$$

La statistique de Fisher observée ($F_{obs} = 14.64$) est supérieure à ($F_{(4,56)}^{0.05} = 2.53$)[4], ce qui nous permet d'accepter l'hypothèse alternative H_1. On remarque que la pvalue est très inférieure au seuil 5%. Par conséquent, la régression est fortement significative.

Une autre mesure statistique qu'on peut tirer à partir de tableau ANOVA est l'erreur quadratique moyenne CMR (ou Mean Squared Error MSE), c'est juste une mesure pratique pour nous dire, à quel point notre droite de régression est proche de l'ensemble des points :

$$MSE = 0.53$$

Il n'y a pas de valeur correcte pour MSE. En termes simples, plus la valeur est faible (ou proche de 0), plus le modèle est adapté. MSE = 0 signifie que le modèle est parfait.

Donc, Il est évident que le modèle construit a montré de meilleurs résultats pour l'ensemble d'entraînement. Le graphe (Fig 3.2) ci-dessous montre la distribution des ensembles d'apprentissage et Test : en abscisse les activités prédites et en ordonnée les activités observées de pIC_{50}.

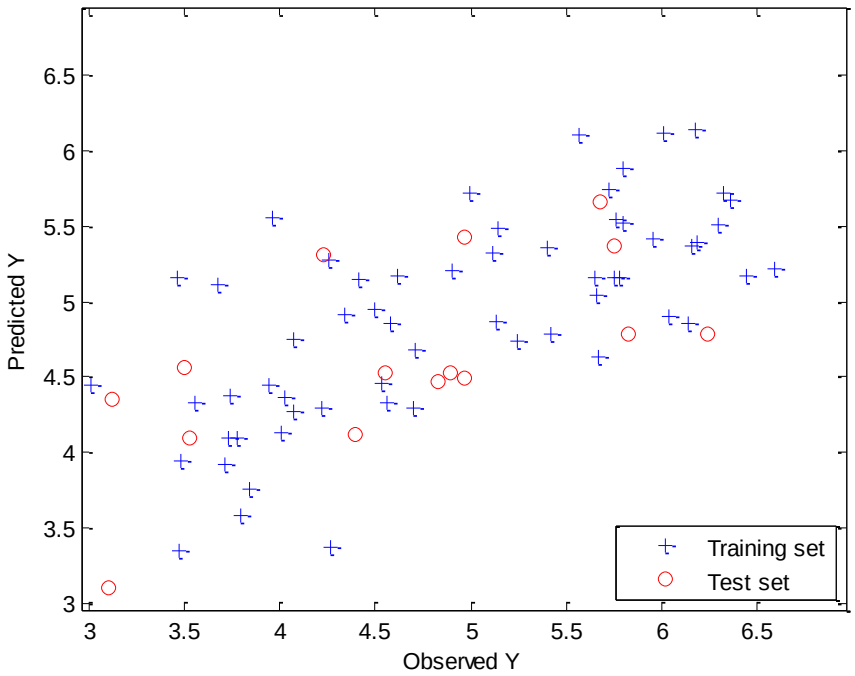


Figure 3.2: Activité observée et prévue du modèle par l'ensemble d'apprentissage et test.

[4] Quantile de la loi de Fisher $F_{(4,56)}^{0.05} = 2.53$, voir l'annexe B.

1.2.4 Diagnostic de la régression

Les points aberrants sont localisés loin des valeurs de l'activité pIC_{50} prédites (points aberrants sur l'axe des Y) ou loin des valeurs des descripteurs (points aberrants sur les axes des X).

On peut déterminer ainsi les points aberrants sur l'axe des descripteurs en utilisant la valeur de levier h_{ii}, dont les points possédant des valeurs supérieures aux seuils critiques

$$h_{ii}^{(1)} = \frac{2(p+1)}{n} = \frac{2(4+1)}{61} = 0.17$$

$$h_{ii}^{(2)} = \frac{3(p+1)}{n} = \frac{3(4+1)}{61} = 0.24$$

sont considérés comme points à grand levier, avec p et n respectivement nombre de descripteurs dans le modèle et le nombre d'observations. D'après le tableau de diagnostic (Tab 3.4) et le graphe (Fig 3.4) suivant on remarque qu'il existe cinq points qui sont supérieurs au limite inférieure $h_{ii}^{(1)} = 0.23, 0.19, 0.22, 0.17$ et 0.20.

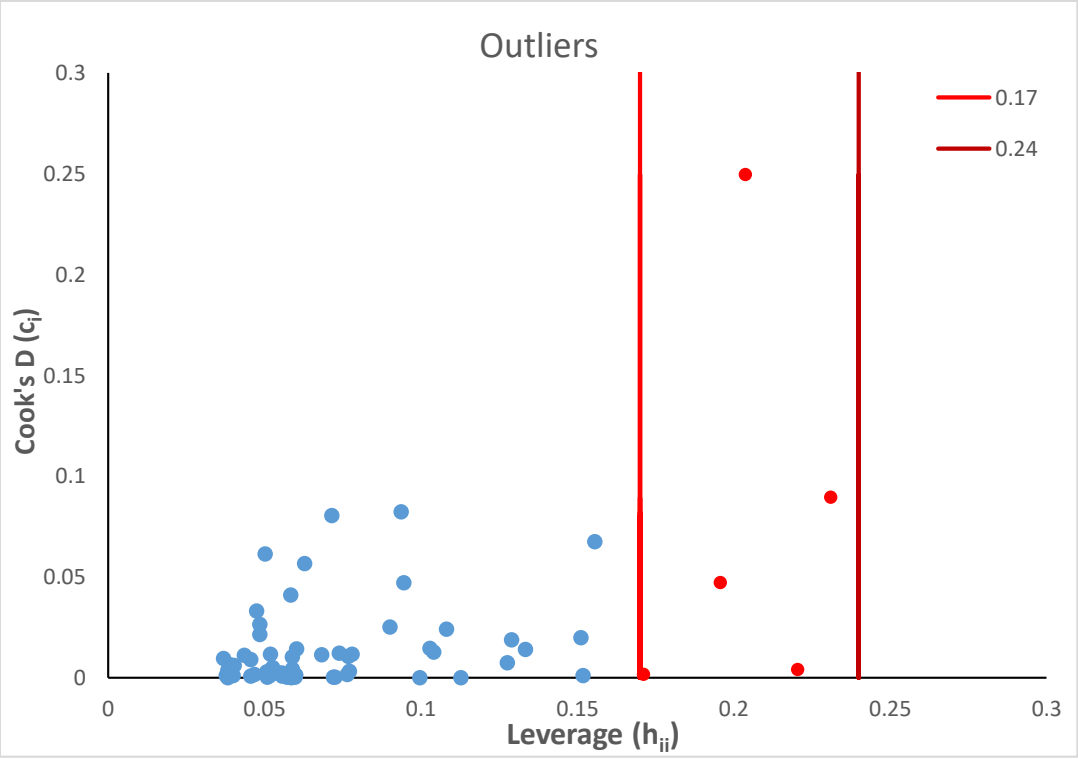


Figure 3.3: Graphe des points aberrants $c_i = f(h_{ii})$

Si les points à grand levier sont supérieurs aux limites supérieures $h_{ii} = 0.24$, on doit les supprimer et réconstruire le modèle. Dans notre cas, les points ne dépassent pas cette limite donc on va les conserver.

Tableau 3.4 : Diagnostic de régression

	Molécules	Y_{obs}	Y_{pred}	e_i	h_{ii}	c_i
Training	1	6.350	4.634	1.032	0.095	0.047
	2	6.070	4.783	0.634	0.059	0.010
	3	6.770	4.907	1.133	0.049	0.026
	4	6.340	5.043	0.614	0.068	0.011
	6	7.220	5.168	1.274	0.059	0.041
	7	6.330	5.162	0.486	0.038	0.004
	9	6.480	5.156	0.626	0.039	0.006
	11	6.450	5.156	0.599	0.040	0.006
	12	5.170	5.168	-0.555	0.077	0.011
	13	4.440	5.553	-1.591	0.072	0.080
	14	4.770	5.280	-1.024	0.049	0.021
	15	4.420	4.442	-0.498	0.151	0.020
	16	5.040	4.953	-0.456	0.134	0.014
	18	5.500	5.203	-0.296	0.077	0.003
	19	4.570	4.757	-0.679	0.108	0.024
	21	7.050	5.515	0.776	0.090	0.025
	24	4.780	3.366	0.899	0.156	0.067
	25	4.510	4.362	-0.337	0.128	0.007
	26	6.880	4.852	1.287	0.048	0.033
	27	6.900	5.377	0.780	0.044	0.011
	28	6.930	5.389	0.794	0.037	0.010
	29	7.090	5.722	0.604	0.074	0.012
	30	6.730	6.116	-0.111	0.152	0.001
	31	7.390	5.221	1.373	0.094	0.082
	32	6.670	5.416	0.536	0.129	0.019
	33	4.950	5.153	-0.736	0.060	0.014
	34	5.730	5.327	-0.214	0.060	0.001
	35	7.130	5.676	0.686	0.046	0.009
	36	5.750	4.869	0.262	0.040	0.001
	37	5.130	4.856	-0.279	0.047	0.002
	39	4.120	5.113	-1.437	0.063	0.057
	41	4.870	4.918	-0.572	0.078	0.012
	42	6.050	5.356	0.042	0.038	$2.84\ 10^{-5}$
	44	6.460	5.548	0.216	0.046	0.001
	45	6.500	5.527	0.273	0.038	0.001
	46	5.760	5.487	-0.348	0.039	0.002
	48	6.410	5.751	-0.031	0.059	0.000
	49	5.600	5.725	-0.729	0.052	0.012
	50	6.920	6.143	0.032	0.113	$5.58\ 10^{-5}$
	51	6.240	6.113	-0.545	0.103	0.015
	52	6.500	5.884	-0.084	0.073	0.000
	53	3.990	4.333	-0.773	0.231	0.089
	54	4.190	4.376	-0.638	0.196	0.047

55	4.260	3.586	0.215	0.077	0.002
56	4.310	3.754	0.091	0.060	0.000
57	4.160	3.916	-0.204	0.052	0.001
58	3.900	3.949	-0.470	0.053	0.005
59	4.490	4.134	-0.127	0.051	0.000
60	4.240	4.101	-0.318	0.056	0.002
61	4.180	4.095	-0.365	0.051	0.003
62	5.270	4.292	0.411	0.059	0.004
63	5.110	4.331	0.228	0.060	0.001
64	4.730	4.289	-0.069	0.057	0.000
65	4.570	4.272	-0.194	0.056	0.001
66	5.880	4.739	0.507	0.104	0.013
67	5.280	4.681	0.030	0.100	$4.25\ 10^{-5}$
68	5.080	4.459	0.073	0.072	0.000
69	3.460	2.924	0.163	0.221	0.004
71	3.890	3.347	0.124	0.171	0.001
73	3.880	5.160	-1.698	0.050	0.061
75	3.380	4.442	-1.426	0.204	0.249

En examinant les valeurs de distance de Cook pour notre régression, on remarque que toutes les valeurs de c_i sont inférieurs au seuil critique $c_i = 1$. Ainsi, nous n'avons pas de points atypiques dans la direction des ordonnées.

1.2.5 Validation du modèle

Une fois l'équation de modèle obtenue, il est important de déterminer sa fiabilité et son importance. Pour déterminer à quel point notre modèle prédit la réponse pour de nouvelles observations, nous utilisons le R^2 prédit. Les modèles qui ont des valeurs $R^2_{pred} \geq 0.6$ ont une meilleure capacité de prédiction :

$$R^2_{pred} = 1 - \frac{\sum_{i=1}^{n_{ts}} (y^i_{test} - \hat{y}_{test})^2}{\sum_{i=1}^{n_{ts}} (y^i_{test} - \bar{y}_{train})^2}$$

Où $\bar{y}_{train}$ est la valeur moyenne de la variable dépendante pour le Training set et n_{ts} est le nombre d'observation de Test set $n = 15$.

La capacité prédictive du modèle sélectionné a été confirmée par le R^2 du Test set $R^2_{pred} = 0.57 \cong 0.6$. Ce résultat indique que notre modèle a un pouvoir prédictif et capable de faire une généralisation sur de nouvelles données.

1.2.6 Modèle final

Après la confirmation par un ensemble des tests statistiques, notre modèle final est donné par :

$$\boldsymbol{pIC_{50}} = 4.8743 + 0.0199\boldsymbol{logP} - 0.5038\boldsymbol{HBA} - 0.5763\boldsymbol{Q_4} + 0.4509\boldsymbol{Volume}$$

$$n_{tr} = 61 \quad R^2 = 0.51 \quad R^2_{adj} = 0.48 \quad R^2_{pred} = 0.58 \quad F = 14.64 \quad MSE = 0.53$$

2. Réseaux de neurones artificiels

Dans la présente étude, nous avons choisi aussi d'établir une relation QSAR non linéaire avec les 76 molécules. En utilisant la méthode des réseaux de neurones artificiels. Dans une seconde étape nous procéderons à une comparaison entre les résultats obtenus par ce modèle et ceux obtenus par le modèle linéaire utilisant la régression linaire multiple.

2.1. Découpage de la base de données

Lors de l'entraînement de réseaux multicouches, la pratique générale consiste à diviser d'abord les données en trois sous-ensembles.

- Le premier sous-ensemble est l'ensemble d'apprentissage, qui est utilisé pour calculer le gradient et mettre à jour les poids et biais du réseau.

- Le deuxième sous-ensemble est l'ensemble de validation. L'erreur (MSE) sur le jeu de validation est surveillée pendant le processus de l'apprentissage. L'erreur de validation diminue normalement pendant la phase initiale de l'entraînement, tout comme l'erreur d'ensemble de l'apprentissage. Cependant, lorsque le réseau commence à surentraîner (en anglais Overfit [5]) les données, l'erreur sur l'ensemble de validation commence généralement à augmenter (Fig 3.4). Les poids et biais du réseau sont enregistrés au minimum de l'erreur de jeu de validation.

- L'erreur de l'ensemble Test n'est pas utilisée pendant la formation, mais elle est utilisée pour comparer différents modèles. Il est également utile de tracer l'erreur de l'ensemble de test pendant le processus de l'apprentissage. Si l'erreur sur l'ensemble de test atteint un minimum à un nombre d'itération significativement

[5] Overfitting : le modèle s'est tellement concentré sur l'ensemble d'entraînement particulier qu'il a raté le point. Overfitting se produit lorsque notre modèle devient vraiment efficace pour pouvoir classer ou prédire des données incluses dans l'ensemble d'apprentissage, mais n'est pas aussi efficace pour classer les données sur lesquelles il n'a pas été entraîné.

différent de l'erreur d'ensemble de validation, cela peut indiquer une mauvaise division de l'ensemble de données. [96]

Il est donc primordial de se munir d'un échantillon de validation qui permet de calibrer et de tester la performance du modèle pendant la phase d'apprentissage. La figure suivante résume le processus :

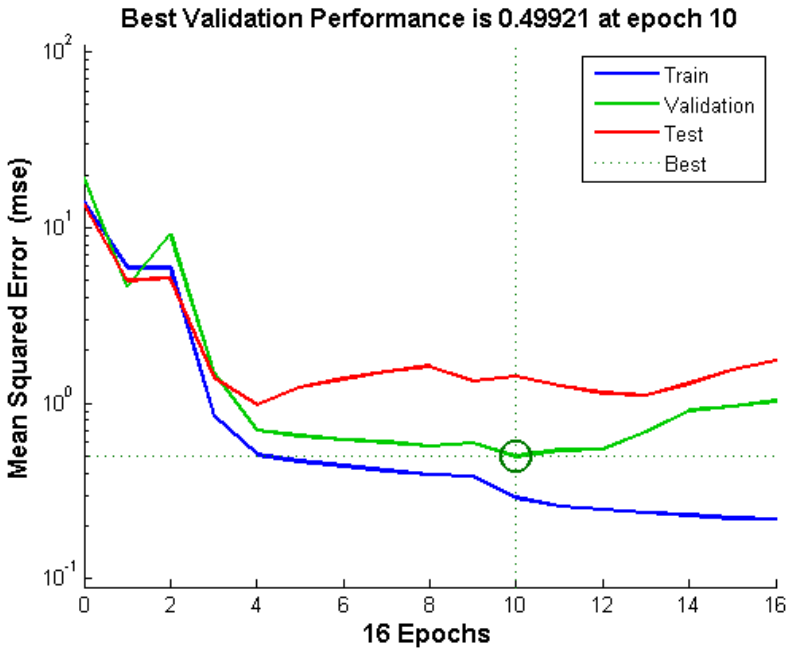


Figure 3.4: Résultats de l'erreur moyenne quadratique (MSE) du modèle RNA lors le processus de l'entraînement

Comme le montre la figure ci-dessus, l'erreur d'apprentissage diminue régulièrement au fil du temps tandis que l'erreur de validation commence à augmenter après un certain point.

Pour faire face à ce problème, une stratégie populaire de régularisation dans Deep Learning peut être utilisée, appelée *Early Stopping*. Dans cette stratégie, le meilleur réglage de paramètre générant l'erreur de validation la plus faible est réservé et lorsque l'entraînement est terminé (avec l'erreur d'apprentissage la plus faible), le modèle avec l'erreur de validation minimale est sélectionné (pas le dernier) [97].

2.2. Architecture du modèle RNA

Un réseau de neurones à trois couches de type feed-forward Backpropagation (un perceptron multicouche) a été utilisé dans ce travail, avec quatre entrées correspondant aux quatre variables obtenus par l'analyse RLM (log P, HBA, le volume et la charge Q_4), dix couches cachées et une sortie.

Les 76 données expérimentales sont réparties aléatoirement en trois sous-ensembles. Training (54 molécules), validation (11 molécules) et Test (les 11 molécules restantes) ces deux derniers ont été utilisés pour conquérir des *patterns* généraux entre les variables d'entrée et de sortie lors de la construction du modèle RNA.

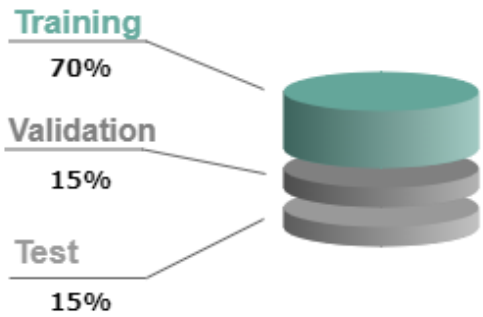


Figure 3.5 : Découpage de la base de données

L'ensemble d'apprentissage (Training) ajuste les poids de connexion et les paramètres du modèle. L'ensemble de validation vérifie les performances du modèle tout au long du processus de formation et arrête l'apprentissage pour éviter le surapprentissage, tandis que l'ensemble Test évalue les performances RNA entraînées et la puissance de généralisation [98]. La figure (Fig 3.6) illustre comment le modèle RNA est désigné :

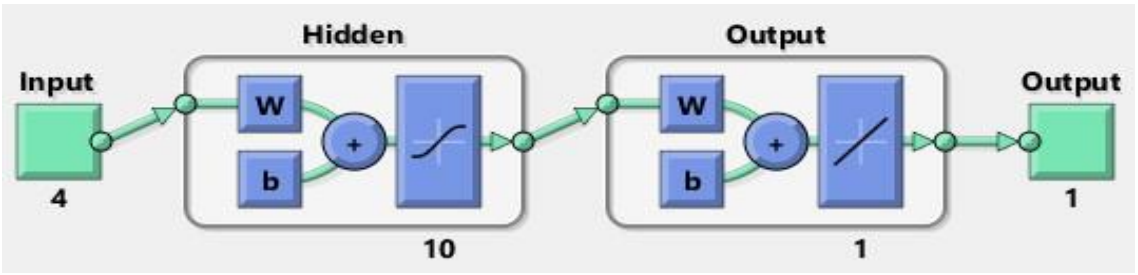


Figure 3.6 : Architecture du perceptron multicouche (4-10-1) que nous avons utilisé

Le réseau se compose de trois couches où la première couche est la couche d'entrée (quatre) qui est déclenchée à l'aide de la fonction d'activation tangente hyperbolique tandis que la deuxième couche est la couche cachée (dix) et la troisième couche est la couche de sortie (une) qui est déclenchée par une fonction d'activation linéaire. Le réseau est entraîné à l'aide de l'algorithme de Levenberg-Marquardt (LM)[6]. Dans le cas d'un apprentissage supervisé, le réseau est présenté à la fois avec les données d'entrée et les données cibles (Target) appelées ensemble d'apprentissage.

[6] Cet algorithme semble à être la méthode la plus rapide pour entraîner des réseaux de neurones feed-forward de taille moyenne (jusqu'à plusieurs centaines de poids). Il est également une implémentation efficace dans le logiciel MATLAB®.

Le réseau est ajusté en fonction de la comparaison des valeurs de sortie Y_{pred} et cible Y_{obs} jusqu'à ce que les sorties correspondent aux cibles. Les valeurs de Y_{pred} et Y_{obs} ainsi que l'erreur correspondant sont présentées dans le tableau suivant :

Tableau 3.5 : les valeurs des activités observées et prédites et ses résidus.

	Molécules	Y_{obs}	Y_{pred}	e_i
Training	2	6.070	6.118	-0.048
	6	7.220	6.048	1.172
	7	6.330	6.037	0.293
	8	5.800	6.108	-0.308
	9	6.480	6.009	0.471
	11	6.450	6.066	0.384
	12	5.170	5.898	-0.728
	13	4.440	4.498	-0.058
	14	4.770	5.866	-1.096
	15	4.420	4.384	0.036
	16	5.040	5.350	-0.310
	17	5.720	5.298	0.422
	18	5.500	5.593	-0.093
	20	5.320	5.726	-0.406
	22	5.140	5.308	-0.168
	23	5.800	6.178	-0.378
	24	4.780	4.416	0.364
	25	4.510	4.522	-0.012
	26	6.880	6.027	0.853
	27	6.900	6.150	0.750
	29	7.090	6.122	0.968
	30	6.730	6.390	0.340
	31	7.390	6.797	0.593
	33	4.950	5.303	-0.353
	36	5.750	5.864	-0.114
	37	5.130	5.752	-0.622
	39	4.120	4.075	0.045
	40	4.120	4.288	-0.168
	41	4.870	4.842	0.028
	42	6.050	5.706	0.344
	43	5.640	6.009	-0.369
	45	6.500	6.129	0.371
	46	5.760	5.729	0.031
	47	7.290	6.371	0.919
	48	6.410	6.126	0.284
	49	5.600	6.231	-0.631
	50	6.920	6.675	0.245
	51	6.240	6.729	-0.489

	54	4.190	4.210	-0.020
	55	4.260	4.399	-0.139
	56	4.310	4.105	0.205
	58	3.900	4.010	-0.110
	59	4.490	4.646	-0.156
	60	4.240	4.297	-0.057
	61	4.180	4.446	-0.266
	63	5.110	4.823	0.287
	64	4.730	4.796	-0.066
	65	4.570	4.784	-0.214
	68	5.080	4.714	0.366
	70	3.630	3.737	-0.107
	72	6.800	5.403	1.397
	74	3.650	3.643	0.007
	75	3.380	3.590	-0.210
	76	4.09	5.965	-1.875
Validation	1	6.350	5.827	0.523
	3	6.770	5.955	0.815
	4	6.340	5.948	0.392
	10	6.630	5.844	0.786
	35	7.130	6.137	0.993
	44	6.460	5.703	0.757
	52	6.500	6.370	0.130
	53	3.990	5.290	-1.300
	57	4.160	4.157	0.003
	62	5.270	4.663	0.607
	69	3.460	3.846	-0.386
Test	5	4.940	5.965	-1.025
	19	4.570	5.930	-1.360
	21	7.050	5.919	1.131
	28	6.930	6.122	0.808
	32	6.670	7.363	-0.693
	34	5.730	5.265	0.465
	38	6.720	6.044	0.676
	66	5.880	3.656	2.224
	67	5.280	3.980	1.300
	71	3.890	3.366	0.524
	73	3.880	5.529	-1.649

Les résultats obtenus montrent que les valeurs prédites sont proches des valeurs observées car les valeurs des résidus $e_i = y_{obs} - y_{pred}$ sont minimaux. Ce qui confirme que le modèle neuronal est capable de faire la prédiction d'une façon adéquate.

La figure ci-dessous, nous permet de représenter graphiquement les résultats présentés dans le tableau précédant (Y_{pred} et Y_{obs}).

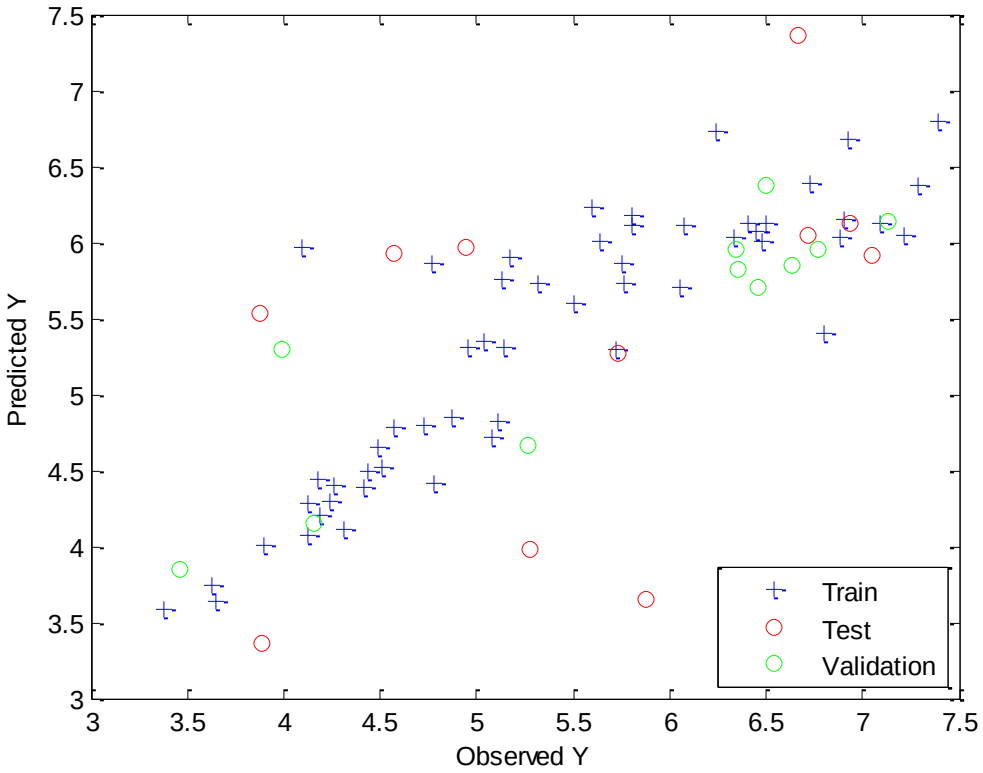


Figure 3.7: Activité observée et prévue du modèle par les trois sous-ensembles (Train, Test, Validation)

2.3. Mesure de la performance du modèle RNA

La performance du modèle développé a été évaluée par un ensemble des indicateurs statistiques : l'erreur moyenne quadratique (MSE), le coefficient de détermination R^2 et R^2_{adj}. Tandis que la prédictivité du modèle a été testée par le coefficient R^2_{pred}. Les résultats de ces derniers sont présentés dans le tableau suivant :

Tableau 3.6 : Résultats des indicateurs statistiques du modèle RNA

Data set	N	MSE	R^2	R^2_{adj}	R^2_{pred}
Training	(54)	0.289	0.716	0.680	0.607

En examinant ces résultats, l'ensemble d'apprentissage donne une valeur d'erreur quadratique moyenne très faible ce qui signifie que le modèle généré présente une meilleure régression avec des minimum erreurs entre les activités biologiques observées et prédites. La courbe de MSE (Fig 3.4) en fonction de nombre d'itérations dans la phase d'apprentissage montre que le réseau est arrivé au meilleur apprentissage et à l'erreur de validation la plus faible $MSE = 0.49921$ après dix itérations.

Les valeurs élevées de R^2 et R^2_{adj} indiquent qu'environs 70% de la variabilité totale est expliquée par ce modèle. On dit que les résultats sont satisfaisants si la valeur R^2 est supérieure à 0,60. Cela confirme que le modèle est bien entraîné, et que la tâche de l'apprentissage a réussi.

De plus, la valeur élevée de R^2_{pred} montre que le modèle obtenu peut prédire l'activité inhibitrice de photosystème II.

La figure (Fig 3.8) montre les graphes des coefficients de détermination R pour les ensembles d'apprentissage, de test, de validation et l'ensemble de prédiction globale sous la forme de sortie du réseau par rapport à l'expérimentale.

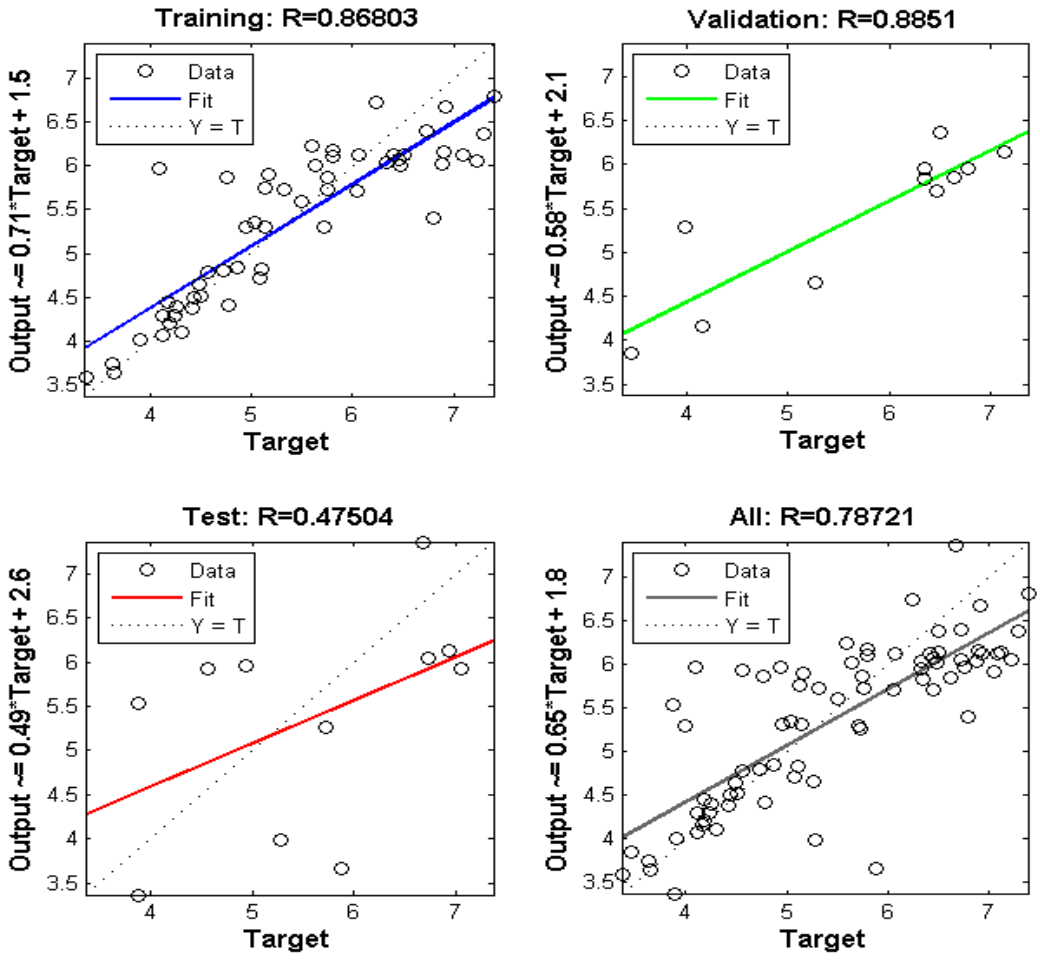


Figure 3.8 : Training, validation et test graphes de régression du modèle RNA

Les coefficients de corrélation pour l'apprentissage, la validation et les tests étaient respectivement de 0.86803, 0.8851 et 0,47504, tandis que l'ensemble de prédiction global était de 0,78721, ce qui confirme que le modèle RNA est satisfaisant pour ajuster les données expérimentales.

3. Comparaison des performances des modèles RLM et RNA

La performance des deux méthodes étudiées, à savoir la régression linéaire multiple et le réseau de neurones artificiels dans la prédiction de l'activité biologique d'inhibition de l'activité du photosystème II, est mesurée par un ensemble des indicateurs descriptifs et prédictifs utilisés pour quantifier l'exactitude de la valeur prédite. Le tableau suivant présente les valeurs des indicateurs de performance :

Tableau 3.7 : Résultats des indicateurs de performance des modèles RLM et RNA

	Indicateurs descriptifs			Indicateurs prédictifs
Modèle	R^2	R^2_{adj}	MSE	$R^2{}_{pred}$
RLM	0.51	0.48	0.524	0.57
RNA	0.72	0.68	0.289	0.61

Les valeurs du coefficient de détermination et du coefficient de détermination ajusté pour le modèle RNA ont des valeurs supérieures à 0,6 indiquant que les valeurs prédites sont très précises. Cependant, ces mesures de précision sont meilleures pour la RLM. L'erreur quadratique moyenne est plus petit pour le modèle RNA que pour la RLM. Cela montre que RNA donne le meilleur résultat basé sur les mesures d'erreur. Ainsi, RNA devrait fournir une meilleure prédiction que la RLM.

On remarque aussi que le modèle RNA donne une valeur de $R^2{}_{pred}$ un peu plus grande que celle de RLM ce qui confirme que RNA est plus prédictif.
Les figures (Fig 3.9 et 3.10) ci-dessous montrent les valeurs observées et prédites par les deux modèles :

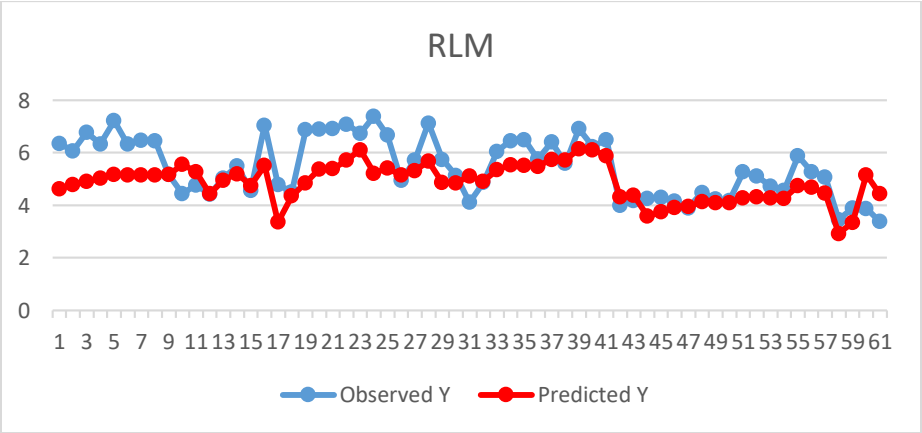


Figure 3.9 : Comparaison des valeurs observées et prédites par le modèle RLM.

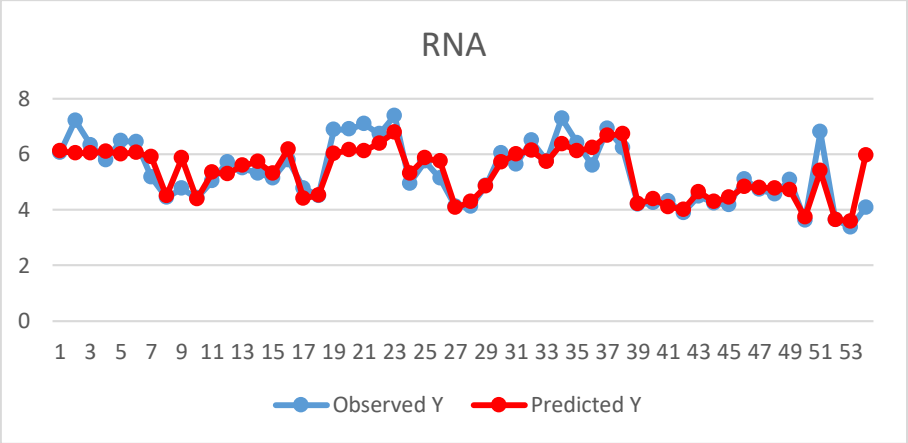


Figure 3.10 : Comparaison des valeurs observées et prédites par le modèle RNA

Ces figures suggèrent que le modèle RNA surpasse le modèle RLM puisque les valeurs prédites par RNA sont plus proches des valeurs réelles par rapport au modèle RLM.

4. Conclusion

La régression linéaire multiple donne des résultats assez acceptables pour la prévision de l'activité inhibitrice mais elle ne surpasse pas les RNA.

Des études cumulées ont montré que les réseaux de neurones artificiels sont meilleur en prédictions que la régression linéaire multiple [99]. Dans ce chapitre on a pu confirmer cette théorie. RNA est l'outil le plus puissant pour prédire l'activité inhibitrice en comparant les résultats des deux modèles. On trouve que les valeurs de R^2 le plus élevé et MSE le plus faible sont pour le modèle RNA. Cette puissance est dû à la capacité du réseau à reconnaitre les *patterns* entre les entrées et les sorties ou de la généralisation effectuée.

Conclusion générale

Notre travail a été consacré à la modélisation de la relation quantitative structure activité inhibitrice de photosystème II, pour construire des modèles QSAR fiables, robustes, stables et précis, capables de prédire efficacement cette activité. Pour cet objectif, nous avons mené notre étude à partir d'une bibliothèque de soixante-seize molécules, dérivées d'anilide.

La première partie de ce mémoire a été consacré sur les effets biologiques de la molécule étudiée ainsi que la méthodologie générale de l'étude QSAR et les deux méthodes statistiques utilisées pour la construction des modèles. Cette partie nous a permis d'avoir une idée sur les deux premiers chapitres.

La seconde partie était dédiée, dans un premier temps, à l'utilisation de la méthode de régression linéaire multiple dans le développement du modèle QSAR en tant que méthode d'apprentissage et de sélection, pour modéliser l'activité inhibitrice de photosystème II exprimée par la grandeur pIC_{50}. Quatre descripteurs moléculaires pertinents ont été sélectionnés par la méthode Stepwise afin de construire le modèle RLM, exprimés par : log P, HBA, la charge Q_4 et le volume. Puis nous avons fait appel aux réseaux de neurones artificiels en tant que méthode d'apprentissage non linéaire avec une grande capacité prédictive. La comparaison des résultats du modèle RLM avec ceux obtenus par le modèle non linéaire de RNA a montré que ce dernier est une méthode capable de modéliser efficacement l'activité inhibitrice de PSII, et que ses résultats sont meilleurs et plus fiables que ceux obtenu par la méthode RLM.

Les performances des deux méthodes étudiées, à savoir la régression linéaire multiple et le réseau de neurones artificiels, dans la prédiction de l'activité inhibitrice de PSII, ont été mesurées en utilisant l'erreur quadratique moyenne (MSE) et le coefficient de détermination R^2 qui indique la proportion de variation expliquée par le modèle de validation. On a trouvé que R^2 le plus élevé et la valeur MSE la plus faible pour pIC_{50} ont été respectivement obtenus via le modèle RNA comme 0,716 et 0,2896.

Malgré la capacité prédictive du réseau de neurones artificiels, elle manque en quelque sorte d'explication sur les paramètres utilisés. La régression linéaire multiple donne une explication simple et facile sur les paramètres estimés. Cela rend la méthode toujours très utile. D'un autre côté, la régression linéaire multiple nécessite des hypothèses de base telles que la linéarité et la normalité. Ils peuvent être violés lors de l'utilisation de données réelles. Si ces hypothèses ne sont pas respectées, des mesures correctives doivent être prises. Les réseaux de neurones artificiels n'ont pas besoin de ces hypothèses. Une grande quantité de données est suffisante pour que le réseau reconnaisse les modèles formés par les données. Une valeur aberrante ou influente peut affecter la fonction de régression estimée, mais le réseau neuronal artificiel est très puissant et peut gérer ce type de données.

Bien que tous les résultats montrent que RNA est meilleure que RLM du côté prédiction, on ne peut pas favoriser l'une sur l'autre. Le choix de la méthode à adapter est basé sur la nature du problème que nous cherchons à traiter. Si notre but est d'avoir un modèle à utiliser dans des fins futures, alors nous devrions choisir la régression linéaire et si notre but est de faire la prédiction alors on doit choisir les réseaux de neurones.

Glossaire

CHAPITRE 1 : EFFETS BIOLOGIQUES DE LA MOLECULE

- **Antenne :** Ou complexes antennaires, sont des complexes constitués de pigments liés à des protéines qui assurent essentiellement la capture de l'énergie lumineuse.

- **ATP :** Un nucléotide de la famille des purines servant à emmagasiner et à transporter de l'énergie dans la cellule végétale.

- **Chlorophylle :** De couleur verte, est le pigment photosynthètique essentiel des feuilles des plantes supérieures. C'est une molécule constituée d'un cycle tetrapyrrolique possédant un atome de magnésium et deux résidus alcools liés à des fonctions carboxyles sièges de l'absorption d'énergie faisant passer la molécule de l'état fondamentale à l'état excité.

- **Chloroplaste :** Organite cellulaire rencontrée chez les eucaryotes qui est le site de la photosynthèse oxygénique.

- **Glutamine synthétase :** Chez les plantes, cette enzyme intervient au sein des chloroplastes dans le cadre de la photorespiration.

- **DL_{50}/CL_{50} :** Indicateurs quantitatifs de la toxicité d'une substance. Dose/Concentration provoquant 50% de mortalité dans la population d'organismes étudiée, par administration unique.

- **Granum :** Grana au pluriel, en botanique c'est une structure fine des chloroplastes où se fait la captation de l'énergie lumineuse par la plante.

- **Hydron :** Le nom général du cation H^+.

- **NADPH :** Porteur de pouvoir réducteur, car il est capable de fournir de l'énergie lors du transfert de son atome d'hydrogène : il rend possible des réactions de réduction nécessaires à la cellule.

- **Oxoacides :** Nom traditionnel pour tout acide ayant un oxygène dans le groupe acide. Le terme oxoacide fait référence à un composé qui contient de l'oxygène, au moins un autre élément et au moins un hydrogène lié à l'oxygène, et qui produit une base conjuguée par perte d'ion (s) d'hydrogène positif.

- **Pesticide :** Produit chimique utilisé pour la protection ou le traitement des végétaux. Les pesticides regroupent principalement les fongicides, les insecticides et les herbicides, utilisés respectivement pour lutter contre les champignons, les insectes et les mauvaises herbes (adventices).

- **Photosystème :** Sous-unité de l'appareil photosynthétique. Les mécanismes de la photosynthèse font intervenir des photosystèmes, chargés d'absorber les photons et d'effectuer la conversion photochimique.

- **Stroma :** Liquide intra-chloroplastique. Le fluide incolore qui entoure les thylakoïdes dans les chloroplastes des cellules végétales.

- **Thylakoïde :** Saccule aplatis, prolongement de la membrane interne du chloroplaste logeant les photosystèmes.

- **Voie foliaire :** Type de pénétration des herbicides : par les feuilles, tissus chlorophylliens ou préchlorophylliens et par les tiges avant formation de l'écorce.

- **Voie racinaire :** Type de pénétration des herbicides : par les racines, les poils absorbants, par les radicules ou les tigelles sortant de graines en cours de germination.

CHAPITRE 1 : RELATION QUANTITATIVE STRUCTURE-ACTIVITE

- **Ab initio :** Méthode de chimie computationnelle basée sur la chimie quantique utilisée pour le calcul de structure électronique des atomes et des molécules.

- **Base de données :** Ensemble de données organisées en vue de son utilisation par des programmes correspondant à des applications distinctes et de manière à faciliter l'évolution indépendante des données et des programmes.

- **Coefficient de partage :** Généralement noté K ou P, est le rapport des activités chimiques d'un soluté entre deux phases.

- **Hydrophobicité :** Une substance est dite hydrophobe quand elle repousse l'eau ou est repoussée par l'eau.

- **Modélisation moléculaire :** Ensemble de techniques pour modéliser ou simuler le comportement de molécules. Elle est utilisée pour reconstruire la structure tridimensionnelle de molécules, en particulier en biologie structurale, à partir de données expérimentales.

- **Relation quantitative structure-activité (QSAR) :** Procédé par lequel une structure chimique est corrélée avec un effet bien déterminé comme l'activité biologique ou la réactivité chimique.

- **Potentiel d'ionisation :** Ou énergie d'ionisation d'un atome ou d'une molécule est l'énergie qu'il faut fournir à un atome pour arracher un électron (le moins lié) à l'état gazeux et former un ion positif.

- **Potentiel électrostatique :** L'énergie potentielle d'un système de charges électriques séparées.

- **Van der Waals :** En physique et en chimie, une force de van der Waals, interaction de van der Waals ou liaison de van der Waals est un potentiel interatomique dû à une interaction électrique de faible intensité entre deux atomes ou molécules.

CHAPITRE 2 : REGRESSION LINEAIRE MULTIPLE

- **Distance de Cook :** Un outil de diagnostic qui mesure l'effet de la suppression d'une donnée aberrante (Outlier).

- **Effet de Levier (Leverage) :** Un outil de diagnostic qui répond à la question suivante : "est-ce qu'un ou plusieurs points sont tellement influents qu'ils tirent la régression vers eux de manière abusive ?".

- **Espérance :** L'espérance mathématique est une valeur statistique que l'on s'attend à trouver, en moyenne, si l'on répète un grand nombre de fois la même expérience aléatoire.

- **Homoscédasticité :** Uniformité de la variance de l'erreur dans un ensemble de valeurs observées.

- **Loi gaussienne :** Loi de probabilité célèbre qui permet de décrire une grande part des distributions statistiques observées, aussi dite Loi normale.

- **Moindres carrés :** Cette méthode est une notion mathématique permettant d'apporter à un nombre d'éléments susceptibles de comporter des erreurs un ajustement afin d'obtenir des données proches de la réalité.

- **Régression linéaire :** Un modèle qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

- **Variance :** La variance d'une variable aléatoire est la mesure de la dispersion des échantillons autour de la moyenne.

CHAPITRE 2 : RESEAUX DE NEURONES ARTIFICIELS

- **Apprentissage non supervisé :** Une tâche d'apprentissage automatique consistant à apprendre à un algorithme des informations qui ne sont ni classées, ni annotées, et à permettre à cet algorithme de réagir à ces informations sans supervision.

- **Apprentissage supervisé :** Une tâche d'apprentissage automatique consistant à apprendre à un algorithme une fonction de prédiction à partir d'exemples annotés.

- **Axone :** Prolongement du neurone conduisant l'influx nerveux.

- **Booléen (ne) :** Le terme booléen s'applique à un langage informatique binaire, inventé par le mathématicien George Boole, qui consiste à programmer des variables qui peuvent être le chiffre 0, qui correspond à non et faux, ou 1 qui correspond à oui ou vrai.

- **Coefficient synaptique :** Nombre qui, en multipliant les différentes valeurs des signaux reçus à l'entrée d'un neurone artificiel, sert à calculer la valeur du signal émis à la sortie. Il fait référence à la force ou à l'amplitude d'une connexion entre deux nœuds.

- **Dendrites :** Des prolongements cytoplasmiques qui entourent le corps cellulaire des neurones.

- **Descente de gradient :** Est un algorithme d'optimisation qui permet de trouver le minimum de n'importe quelle fonction convexe en convergeant progressivement vers celui-ci.

- **Fonction d'activation :** Il s'agit d'une fonction qui permet de transformer le signal entrant dans une unité (neurone) en signal de sortie (réponse).

- **Intelligence artificielle IA :** Permet à des machines, et plus particulièrement à des systèmes informatiques, de simuler les processus cognitifs humains (apprentissage, raisonnement...).

- **Neurone formel :** Une représentation mathématique et informatique d'un neurone biologique.

- **Perceptron :** L'un des tout premiers algorithmes de Machine Learning, et le réseau de neurones artificiels le plus simple.

- **Réseau de neurones artificiels :** Un système dont la conception est à l'origine schématiquement inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques.

- **Rétropropagation :** Une méthode pour calculer le gradient de l'erreur pour chaque neurone d'un réseau de neurones, de la dernière couche vers la première.

- **Soma :** Corps cellulaire qui forme la partie centrale du noyau d'un neurone.

- **Synapse :** Région spécialisée où un signal nerveux saute d'une cellule nerveuse à une autre. C'est le site de communication entre deux cellules nerveuses.

CHAPITRE 3 : INTERPRETATION DES RESULTAS

- **6-31G :** Une base de calcul (ensemble de fonctions) en chimie quantique utilisée afin de modéliser des orbitales moléculaires.

- **B3LYP** (fonctionnelle hybride) **:** Est un élément d'une classe d'approximations à la fonctionnelle d'échange-corrélation, utilisé au sein de la DFT. La caractéristique de ces fonctionnelles est d'avoir une partie d'échange basée sur la méthode de Hartree-Fock dépendante des orbitales alors que la partie de corrélation est basée sur une autre approche (soit issue de méthodes ab-initio, soit semi-empirique).

- **Backpropagation :** Ou la rétropropagation du gradient, est une méthode pour calculer le gradient de l'erreur pour chaque neurone d'un réseau de neurones, de la dernière couche vers la première.

- **BIC :** Ou Critère d'information Bayésian. Est un critère de sélection des modèles qui dépend de la taille de l'échantillon et du nombre des paramètres.

- **Champ de force :** Ensemble de potentiels et de paramètres permettant de décrire la structure de l'énergie potentielle d'un système de particules (typiquement des atomes).

- **Deep Learning :** Ou apprentissage profond, est un type d'intelligence artificielle dérivé du machine learning (apprentissage automatique) où la machine est capable d'apprendre par elle-même.

- **DFT :** La théorie de la fonctionnelle de la densité est une méthode de calcul quantique permettant notamment l'étude de la structure électronique.

- **Early stopping :** Forme de régularisation dans l'apprentissage automatique conçu pour surveiller l'erreur de généralisation d'un modèle et arrêter l'entraînement lorsque l'erreur de généralisation commence à se dégrader.

- **FeedForward :** Le premier et le plus simple type de réseau neuronal artificiel conçu. Dans ce réseau, l'information ne se déplace que dans une seule direction, vers l'avant, à partir des nœuds d'entrée, en passant par les couches cachées et vers les noeuds de sortie.

- **Mécanique moléculaire :** Méthode empirique de calcul des propriétés de molécules telles que la géométrie moléculaire, la chaleur de formation, l'énergie de déformation, le moment dipolaire et les fréquences de vibration.

- **Z score :** Une mesure numérique qui décrit la relation d'une valeur avec la moyenne d'un groupe de valeurs.

Références

[1] **J. Farineau, J. –F. Morot-Gaudry**. La photosynthèse, Processus physiques, moléculaires et physiologiques. Pages: 364, 365.

[2] **A. Trebst and S. Reimer, W. Draber and H. J. Knops**. The Effect of Analogues of Dibromothymoquinone and of Bromonitrothymol on Photosynthetic Electron Flow. Received June 1st, 1979.

[3] **Thi Ngoc Phuong Huynh**. Synthèse et études des relations structure/activité quantitatives (QSAR/2D) d'analyse benzo[c]phénanthridiniques. Sciences du Vivant. Université d'Angers, 2007.

[4] **Mohamed Amine Bedrane**. Agronomie Info. Les herbicides. Du réseau https://agronomie.info/fr/les-herbicides/

[5] National Center for Biotechnology Information. PubChem Database. Propanil, CID=4933, https://pubchem.ncbi.nlm.nih.gov/compound/Propanil (accessed on Feb. 16, 2020)

[6] Penn State: Herbicide Formulations and Adjuvants-Herbicide Applicator Training.

[7] **Celestine Duncan.** TechLine Invasive Plant News 2018. Introduction to Herbicide Formulations.

[8] Cirad. La recherche agronomique pour le developpement. du réseau http://agroecologie.cirad.fr

[9] **Camill Rhoul**. Simulation de la fluorescence de la végétation mesurée depuis une orbite géostationnaire. Interfaces continentales, environnement. Université Paris-Saclay, 2016.

[10] **Kalyani Paranjape, Vasant Gowariker, V N Krishnamurthy, Sugha Gowariker**. The Pesticide Encyclopedia. Pages : 248, 351.

[11] **Michel Tissut, Philippe Delval, Jean Mamarot et Patrick Ravanel**. Plantes, herbicides et désherbage. 2e édition. Pages : 390, 391, 395, 398, 399, 400.

[12] **A. Gama, coord**. Utilisation des herbicides en forêt et gestion durable. Pages: 51, 52.

[13] **S. J. Chalk.** IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by Blackwell Scientific Publications, Oxford (1997). Online version (2019-). https://doi.org/10.1351/goldbook.

[14] National Center for Biotechnology Information. PubChem Database. Propanil, CID=4933, https://pubchem.ncbi.nlm.nih.gov/compound/Propanil (accessed on Feb. 16, 2020).

[15] **N.N. MELNIKOV**, edited by **FRANCES A. GUNTHER and JANE DAVIES GUNTHER**. Chemistry of Pesticides. Page: 124.

[16] **Dr. A G Nikalje** Synthon Approach. Chavan College Of Pharmacy. Dr. Rafiq Zakaria Campus Aurangabad.

[17] National Center for Biotechnology Information. PubChem Database. Pentanochlor, CID=16826, https://pubchem.ncbi.nlm.nih.gov/compound/Pentanochlor (accessed on Mar. 11, 2020).

[18] ChemicalBook, CAS DataBase List. Pentanochlor. https://www.chemicalbook.com/ChemicalProductProperty_EN_CB9783960.htm

[19] **Angela Robyn Kana**. Quantitative Structure-Activity Relationship: Prediction of Anaerobic Transformation of Chloroacetanilide Herbicides. Faculty of The Graduate College of The Oklahoma State University. July, 2007.

[20] **Mabrouk Hamadache**, **Othmane Benkortbi**, **Salah Hanini**, **Abdeltif Amrane**, **Latifa Khaouane**, **Cherif Si Moussa**. A Quantitative Structure Activity Relationship for acute oral toxicity of pesticides on rats: Validation, Domain of Application and Prediction. Journal of Hazardous Materials, Elsevier, 2016. Pages: 28-40.

[21] **Umma Muhammad**, **Adamu Uzairu**, **David Ebuka Arthur.** Review on: quantitative structure activity relationship (QSAR) modelling. Nigeria. April 27, 2018.

[22] **Hugo Kubinyi**. QSAR: Hansch analysis and related approaches. Page: 4.

[23] **Hugo Kubinyi**. 3D QSAR in Drug Design: Volume 1: Theory Methods and Applications, Ludwigshafen, Germany.

[24] **Samir Chtita**. Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments. Université Moulay Ismail – Présidence. Maroc,2017.

[25] **Ramon Carbó-Dorca** ; **David Robert** ; **Lluís Amat** ; **Xavier Gironés** ; **Emili Besalú**. Molecular Quantum Similarity in QSAR and Drug Design.

[26] **Roy Kunal**, **Kar Supratik**, **Rudra Narayan Das.** A Primer on QSAR/QSPR Modeling Fundamental Concepts.

[27] **Kunal Roy**. Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental sciences. Pages: 57-58.

[28] **Benazzouz Hicham, Khebiza Ayoub**. Relation Structure Activité : Etude Qualitative et Quantitative et Développement de Recherche sur les Coumarines. Universite Abou Bekr Belkaîd Tlemcen.

[29] **Free SM**, **Wilson JW**. A mathematical contribution to structure-activity studies. Journal of Medicinal Chemistry. 1964.

[30] **Mark T. D. Cronin**. **Tomasz Puzyn**, **Jerzy Leszczynski**, **Mark T. Cronin**. Recent Advances in QSAR Studies: Methods and Applications. Page: 4.

[31] **Siavoush Dastmalchi**, **Maryam Hamzeh-Mivehroud**, **Babak Sokouti**. Quantitative Structure–Activity Relationship. Pages : 2, 11, 56-57,59.

[32] **A. Fortuné**, Techniques de Modélisation Moléculaire appliquées à l'étude et à l'optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésist-ance, Université Joseph Fourier – Grenoble I, France, 2006.

[33] **Khan AU**. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug discovery today. 2016.

[34] **Mehellou Mohammed Nadjib**. Etude des relations structure/activité quantitatives (QSAR/2D) d'une série de dérivés de Triazolothiadiazoles. Universite Echahide Hamma Lakhder D'el-Oued.

[35] **Bouabid Amel**, **Chibani Warda**, **Yahi Amina**. L'évaluation du potentiel génotoxique des nanoparticules (cas SiO2) par une approche prédictive (Relation Quantitative Structure-Activité-QSAR), Universite 8 Mai 1945 Guelma. Juin 2014.

[36] **Aurélie Goulon-Sigwalt-Abram**. Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments. Université Paris 6 Pierre et Marie Curie,2008.

[37] **Bellifa khadidja**. Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires. « Modélisation QSAR », Universite Abou Bekr Belkaïd De Tlemcen, 2015.

[38] **Amrani Hayat**, **Lebssisse Chaima**. Développement de modèles QSPR pour la prédiction d'enthalpie de décomposition des composés organiques à l'aide des descripteurs moléculaires. Université Echahid Hamma Lakhdar. El Oued 2018.

[39] **Melle Hamad B** et **Khahla S**. Contribution à la Prédiction de Coefficient de partage octanol /eau par la technique QSPR (Quantitive Structure Proprety Relationship) Université El-Oued (2013/2014).

[40] **Mathias Barbaste.** Une méthode de provisionnement individuel par apprentissage automatique

[41] **Ravichandran Veerasamy, Shalini Sivadasan, Christapher P. Varghese. Harish Rajak. Abhishek Jain, Ram Kishore Agrawal**. Validation of QSAR Models - Strategies and Importance. July – September 2011.

[42] **Kunal Roy. Supratik Kar. Rudra Narayan Das**. « A Primer on QSAR/QSPR Modeling, Fundamental Concepts ». SpringerBriefs in Molecular Science.

[43] **Berkani Fatiha**. « Application de la Régression Linéaire Multiples sur la Balance Commerciale Algériènne ». Universite de Kasdi Merbah Ouargla.

[44] **Guillaume Fayet**. « Développement de modèles QSPR pour la prédiction des propriétés d'explosibilité des composés nitroaromatiques. Chimie ». Chimie ParisTech. 2010 Français.

[45] **Mourad Bougrine**. « Analyse et prévision du Best Estimate dans le cadre de l'ORSA en assurance vie par régression linéaire multiple ».

[46] **Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining**. « Introduction to Linear Regression Analysis ». Pages : 72-84.

[47] **Éric Matzner-Løber**.« Régression Théorie et applications (Statistique et probabilités appliquées) ». Pages : 41-91.

[48] **Corinne Perraudin (2004-2005).** « Le modèle de régression linéaire économétrie ». Université Paris I Panthéon Sorbonne DESS Conseil en Organisation et Stratégie.

[49] **Ricco Rakotomalala.** Econométrie. « La régression linéaire simple et multiple ». Université Lumière Lyon2.

[50] **Nissay Lim.** « Élaboration d'un modèle de profitabilité ». Sciences de la gestion (Intelligence d'affaires).

[51] **Hélène Hamisultane**. « ECONOMETRIE ». Licence. France. 2002.

[52] **Dris Leila, Hachemi Warda épouse Hider**. « Régression linéaire multiple et modèle linéaire général ». Université Abderrahmane Mira – Bejaia 2015-2016.

[53] **Youssra Fekiyer**. « Comparaison de la régression linéaire multiple et des réseaux de neurones artificiels pour l'évaluation de la qualité chimique des eaux d'irrigation dans la région de Skhirat ».

[54] **Mouna Mouda**. « Analyse statistique avancée des modèles d'adsorption : dépollution des effluents industriels ». Université de Biskra.

[55] **Ali Djaidja**. « Etude de la classification supervisée des données environnementales à l'aide de réseaux de neurones de fonctions à base radiales ». Université Mohamed Boudiaf de M'sila.

[56] **Pierre Borne**, **Mohamed Benrejeb**, **Joseph Haggege**. « Les réseaux de neurones présentation et applications ». Pages : 1-2-3-4.

[57] **Idiou Ghania**. « Régression et modélisation par les réseaux de neurones ». Université de Constantine.

[58] **Fabrice Sorin**, **Lionel Broussard**, **Pierre Roblin** (2001). « Régulation d'un processus industriel par réseaux de neurones », Techniques de l'Ingénieur, traité Informatique industrielle.

[59] **Yann Boniface, Nicolas P. Rougier**. « Apprentissage et Mémoires Introduction aux réseaux de neurones artificiels ». 2012

[60] **El Mahdi Brakni**. « Réseaux De Neurones Artificiels Appliqués A La Méthode Electromagnétique Transitoire InfiniTEM ». Université de Québec en Abitibi-Témiscamingue.

[61] **G. DREYFUS**. « Les Réseaux De Neurones ». École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI), Laboratoire d'Électronique.1998

[62] **F. Tschirhart**. « Réseaux De Neurones Formels Appliques A L'intelligence Artificielle Et Au Jeu ». Ecole Superieure De Genie Informatique.

[63] **Mefenza N. Michael**. « Analyse Par Réseau De Neurones En Vue De La Caracterisation Des Antennes Intelligentes ». Master Recherche En Genie Des Telecommunications Ensp 2009.

[64] **Dr A. Djeffal**. « Classification Réseaux de neurones ». Université de Mohamed Khider Biskra.

[65] **Laurence Parisot**. « L'intelligence Artificielle : L'expertise Partout Accessible À Tous ». Février 2018. Page: 24.

[66] **Gérard Dreyfus**, **Jean-Marc Martinez**, **Manuel Samuelides**, **Mirta B. Gordon**, **Fouad Badran**, **Sylvie Thiria**, **Laurent Hérault**. « Réseaux de neurones-Méthodologie et applications ». Page : 9.

[67] **Hicham Chaoui**. « Conception Et Comparaison De Lois De Commande Adaptative À Base De Réseaux De Neurones Pour Une Articulation Flexible Avec Non-Linéarité Dure ». Université Du Québec À Trois-Rivières. Décembre 2002.

[68] **Kamalesh Gosalia**. « Introduction à l'apprentissage automatique Monographie De CPA Nouveau-Brunswick ». Page : 18

[69] **Louis Frécon** & **Okba Kaza**. « Manuel d'intelligence artificielle ». Page : 96

[70] **Virginie Mathivet**. « L'Intelligence Artificielle pour les développeurs - Concepts et implémentations en C# ». Pages: 440-442-443-446.

[71] **Ronald van Loon**. « Machine Learning Explained: Understanding Supervised, Unsupervised, and Reinforcement Learning ». https://datafloq.com/read/machine-learning-explained-understanding-learning/4478

[72] **Morgane Laur**. « Anticipation des changements de notes des obligations du portefeuille d'un assureur par méthode de machine Learning ». Université Paris Dauphine.

[73] **Chloé-Agathe Azencott**. « Introduction au Machine Learning ». Dunod 2018. Page: 6.

[74] **JR Oakes**. « An experiment in trying to predict Google rankings » in late 2015.

[75] **Younes Benzaki**. « Régression linéaire en Python par la pratique ». 18 avril 2017 : https://mrmint.fr/apprentissage-supervise-machine-learning.

[76] **Mohamed Bouaziz**. « Réseaux de neurones récurrents pour la classification de séquences dans des flux audiovisuels parallèles ». Académie D'aix-Marseille Université D'avignon Et Des Pays De Vaucluse.

[77] **Pascal Cuxac** et **Maha Ghribi** ,**Jean-Charles Lamirel**. « Les méthodes de classification non supervisées appliquées aux textes : mesure de la performance des résultats de clustering de documents ». Vandœuvre-lès-Nancy, France.

[78] **Eshna Verma**. « Machine Learning Interview Questions and Answers Lesson 12 of 13» (Last updated on Jul 13, 2020) : https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-interview-questions

[79] **Manil wagle**. «Association Rules: Unsupervised Learning in Retail ». (Mar 25)

[80] **Eric Larouche**. « Exploration De Différentes Architectures De Réseaux De Neurones Pour La Prédiction De La Glace Atmosphérique Sur Les Conducteurs Des Réseaux Électriques ».

[81] **Bre**, **Facundo** & **Gimenez, Juan** & **Fachinotti**, **Víctor** (2017). « Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks ». Energy and Buildings. 158.

[82] **Philippe Besse**. « Apprentissage Statistique ». INSA de Toulouse - Mathématiques Appliquées.

[83] **Daniel Graupe**. « Principles of artificial neural networks ». Pages: 17-18-24.

[84] « Learning machine-learning eBook ». Page: 45 https://riptutorial.com/Download/machine-learning.pdf

[85] **Frédéric Sur**. « Introduction à l'apprentissage automatique. Département Génie Industriel et Mathématiques Appliquées ». École des Mines de Nancy

[86] **Charu C. Aggarwal**. « Neural Networks and Deep Learning A Textbook ». Page: 17

[87] **Pouliakis**, **Abraham** & **Karakitsou**, **Effrosyni** & **Margari**, **Niki** & **Bountris**, **Panagiotis** & **Haritou**, **Maria** & **Panayiotides**, **John** & **Koutsouris**, **Dimitrios** & **Karakitsos**, **Petros**. (2016). «Artificial Neural Networks as Decision Support Tools in Cytopathology: Past Present and Future ».

[88] **Moez Baccouche**. « Apprentissage neuronal de caractéristiques spatio-temporelles pour la classification automatique de séquences vidéo ». École Doctorale Informatique et Mathématiques de Lyon.

[89] HyperChem (Molecular Modeling System) Hypercube, Inc., 1115 NW, 4th Street, Gainesville, FL 32601, USA (2007). http://www.hyper.com/News/PressRelease/Release7Oct2001/tabid/412/Default.aspx

[90] Gaussian 07 https://gaussview.software.informer.com/5.0/

[91] Molinspiration https://www.molinspiration.com/cgi-bin/properties

[92] Matlab R2012b. The MathWorks, Inc.

[93] MathWorks : https://www.mathworks.com/help/stats/zscore.html#btg5k75

[94] **Bouarra Nabil**. Etudes QSPR des propriétés contrôlant l'évolution de quelques HAP dans l'environnement. Université de Badji Mokhtar Annaba.

[95] **Samir Kenouche**. Méthodes d'analyse quantitatives 2018. Université Mohamed Khider Biskra.

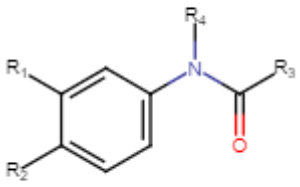
[96] Mathworks. Divide Data for Optimal Neural Network Training https://www.mathworks.com/help/releases/R2015b/nnet/ug/divide-data-for-optimal-neural-network-training.html

[97] **Ehsan Fathi. Babak Maleki Shoja**. Deep Neural Networks for Natural Language Processing..East Carolina University, Greenville, NC, United States. Page : 269

[98] **ÇALIŞKAN, E. SEVİM, Y**. A Comparative Study Of Artificial Neural Networks And Multiple Regression Analysis For Modeling Skidding Time. 2018.

[99]**Kenji Watanabe. Saiful Anwar.** Performance Comparison of Multiple Linear Regression and Artificial Neural Networks in Predicting Depositor Return of Islamic Bank.

Annexe A. Les structures chimiques des anilides étudiés



Mol	R_1	R_2	R_3	R_4	pIC_{50}
1	Cl	Cl	Et	H	6.35
2	Cl	Cl	i-Pr	H	6.07
3	Cl	Cl	1-Me-allyl	H	6.77
4	Cl	Cl	n-Bu	H	6.34
5	Cl	Cl	i-Bu	H	4.94
6	Cl	Cl	1-Me-n-Bu	H	7.22
7	Cl	Cl	2-Me-n-Bu	H	6.33
8	H	Cl	1,1-Me_2-n-Bu	H	5.80
9	Cl	Me	1-Me-n-Bu	H	6.48
10	Cl	Cl	n-pentyl	H	6.63
11	Cl	Cl	i-pentyl	H	6.45
12	Cl	Cl	c-Hx	H	5.17
13	Cl	Cl	$(CH_2)_2$-c-HX	H	4.44
14	Cl	Cl	$(CH_2)_2$-Ph	H	4.77
15	Cl	Cl	CH_2-OPh	H	4.42
16	H	H	1-Me-c-Pr	H	5.04
17	Cl	H	1-Me-c-Pr	H	5.72
18	Br	H	1-Me-c-Pr	H	5.50
19	OMe	H	1-Me-c-Pr	H	4.57
20	CF_3	H	1-Me-c-Pr	H	5.32
21	O-i-pentyl	H	1-Me-c-Pr	H	7.05
22	H	F	1-Me-c-Pr	H	5.14
23	H	Cl	1-Me-c-Pr	H	5.80
24	H	OMe	1-Me-c-Pr	H	4.78
25	H	CN	1-Me-c-Pr	H	4.51
26	Cl	Cl	1-Me-c-Pr	H	6.88
27	O-i-pentyl	H	c-Pr	H	6.90
28	$O(CH_2)_2Ph$	H	c-Pr	H	6.93
29	$O(CH_2)_3Ph$	H	c-Pr	H	7.09
30	$O(CH_2)_5Ph$	H	c-Pr	H	6.73
31	$O(CH_2)_3OPh$	H	c-Pr	H	7.39
32	$O(CH_2)_4OPh$	H	c-Pr	H	6.67
33	H	H	CH_2Ph	H	4.95
34	F	H	CH_2Ph	H	5.73
35	O-i-pentyl	H	CH_2Ph	H	7.13
36	H	Cl	CH_2Ph	H	5.75
37	H	Me	CH_2Ph	H	5.13

38	Cl	Cl	CH_2Ph	H	6.72
39	H	H	n-Bu	H	4.12
40	OH	H	n-Bu	H	4.12
41	OMe	H	n-Bu	H	4.87
42	O-n-Pr	H	n-Bu	H	6.05
43	O-i-Pr	H	n-Bu	H	5.64
44	O-n-Bu	H	n-Bu	H	6.46
45	O-i-Bu	H	n-Bu	H	6.50
46	O-sec-Bu	H	n-Bu	H	5.76
47	O-i-pentyl	H	n-Bu	H	7.29
48	O-n-pentyl	H	n-Bu	H	6.41
49	O-1-Me-n-Bu	H	n-Bu	H	5.60
50	O-n-heptyl	H	n-Bu	H	6.92
51	O-1-Me-n-Hx	H	n-Bu	H	6.24
52	OCH_2-c-Hx	H	n-Bu	H	6.50
53	O(CH)MeCOOMe	H	n-Bu	H	3.99
54	O(CH)MeCOOEt	H	n-Bu	H	4.19
55	H	OMe	n-Bu	H	4.26
56	H	OEt	n-Bu	H	4.31
57	H	O-n-Pr	n-Bu	H	4.16
58	H	O-i-Pr	n-Bu	H	3.90
59	H	O-n-Bu	n-Bu	H	4.49
60	H	O-i-Bu	n-Bu	H	4.24
61	H	O-sec-Bu	n-Bu	H	4.18
62	H	O-i-pentyl	n-Bu	H	5.27
63	H	O-n-pentyl	n-Bu	H	5.11
64	H	O-1-Me-n-Bu	n-Bu	H	4.73
65	H	$OCH(Et)_2$	n-Bu	H	4.57
66	H	O-n-heptyl	n-Bu	H	5.88
67	H	O-1-Me-n-Hx	n-Bu	H	5.28
68	H	OCH_2-c-Hx	n-Bu	H	5.08
69	H	O(CH)MeCOOMe	n-Bu	H	3.46
70	H	O(CH)MeCOOEt	n-Bu	H	3.63
71	H	O(CH)MeCOO-i-Pr	n-Bu	H	3.89
72	O-i-pentyl	H	1-Me-n-Bu	H	6.80
73	Cl	Cl	n-Bu	Me	3.88
74	O-i-pentyl	H	n-Bu	Me	3.65
75	H	O-i-pentyl	n-Bu	Me	3.38
76	$O(CH_2)_3Ph$	H	c-Pr	Me	4.09

Me : Methyl. **Et :** Ethyl. **Pr :** Propyl. **Bu :** Butyl. **Ph :** Phenyl. **Hx :** Hexyl. **i :** iso. **n :** normal. **c :** cyclo.

Annexe B. Quantile de la loi de Fisher

		TABLE du F de FISHER Degré de liberté du numérateur									F limite à p 0.05
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Degré de liberté du dénominateur	**1**	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
	4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
	5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
	6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
	7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
	8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
	9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
	10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
	11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
	12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
	13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
	14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
	15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
	16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
	17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
	18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
	19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
	20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
	21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
	22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
	23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
	24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
	25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
	26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
	27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
	28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
	29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
	30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
	40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
	50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
	60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
	70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
	80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
	90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
	100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927

Résumé

Ce mémoire est basé sur l'application des deux méthodes statistiques : la régression linéaire multiple et les réseaux de neurones artificiels en utilisant le logiciel Matlab. Lors de cette étude, nous avons tenté d'étudier la relation entre l'activité inhibitrice du photosystème II et quelques caractéristiques structurales d'une série, composée de soixante-seize (76) molécules dérivées d'anilide.

Un réseau neuronal artificiel a été construit pour modéliser et prédire l'activité biologique de cette série. Les résultats du modèle établi en utilisant l'algorithme de rétropropagation de gradient (R^2= 0.72 ; R^2_{pred}= 0.61 ; n = 76) étaient supérieurs à ceux obtenus par la régression linéaire multiple (R^2= 0.51 ; R^2_{pred}= 0.57 ; n = 76). L'étude comparative des performances de ces deux méthodes nous a permet de conclure que la RNA est plus puissante pour la prédiction et la généralisation.

MOTS CLÉS : QSAR ; Herbicides ; Anilide ; Activité inhibitrice ; Régression linéaire multiple; Réseau de neurones artificiels.

Abstract

The present work is based on the application of two statistical methods: multiple linear regression and artificial neural networks using Matlab software. Indeed, we tried to study the relationship between the inhibitory activity of photosystem II, and some structural features of a series composed of seventy-six (76) molecules derived from anilide.

A neural network was built to modelize and predict the biological activity of this series. The results of the model established using the Backpropagation algorithm (R^2= 0.72 ; R^2_{pred}= 0.61 ; n = 76) were superior to those obtained by multiple linear regression (R^2= 0.51 ; R^2_{pred}= 0.57 ; n = 76). The comparative study of the performance of these two methods allows us to conclude that RNA is more accurate for prediction and generaliz-ation.

KEY WORDS: QSAR; Herbicides; Anilide; Inhibitory activity; Multiple linear regression; Artificial neural network.