



الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي



Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة الشاذلي بن جديد - الطارف

Université Chadli Bendjedid – El Tarf

كلية العلوم والتكنولوجيا

Faculté des Sciences et de la Technologie

قسم الكيمياء

Département de Chimie

Mémoire de fin d'études

En vue de l'obtention du diplôme de Master

Domaine : Sciences de la matière

Filière : Chimie

Spécialité : Chimie Analytique

Thème

Modélisation de la toxicité des herbicides présents dans les sols agricoles

Présenté par :

SOLTANI Zakaria

Devant le Jury :

Dr. MOKRANI Karima	MCA	Univ Chadli Bendjedid El Tarf	Présidente
Dr. BOUGHERARA Boudjema	MCA	Univ Chadli Bendjedid El Tarf	Encadrant
Dr. LOURICI Leila	MCB	Univ Chadli Bendjedid El Tarf	Examineur

Année Universitaire : 2025/2026

Remerciements

Avant tout, je remercie Dieu le Tout-Puissant de m'avoir accordé la santé, la volonté et la persévérance nécessaires pour mener à bien ce travail.

J'adresse mes plus sincères remerciements à mon encadrant, ***Dr. BOUGHARA Boudjema***, pour son accompagnement, ses précieux conseils, sa disponibilité et son soutien tout au long de la réalisation de ce mémoire.

Je tiens également à remercier l'ensemble des enseignants du département de Chimie pour la qualité de leur enseignement et les connaissances qu'ils m'ont transmises durant mon parcours universitaire.

Mes remerciements s'adressent également aux membres du jury qui ont accepté d'évaluer ce travail et de lui consacrer leur temps.

Je remercie enfin tous ceux qui, de près ou de loin, ont contribué à l'aboutissement de ce mémoire par leurs encouragements, leur aide et leur soutien.

Dédicace

Je dédie ce travail :

À mes chers parents,

Pour leur amour inconditionnel, leurs sacrifices, leurs encouragements constants et leur confiance qui m'ont accompagné tout au long de mon parcours. Aucun mot ne saurait exprimer ma profonde gratitude envers eux.

À ma famille,

Pour son soutien, sa patience et ses encouragements qui m'ont permis de poursuivre mes études dans les meilleures conditions.

À mes amis,

Pour leur présence, leurs conseils et les moments de motivation partagés tout au long de cette aventure universitaire.

À tous ceux qui croient en la valeur du savoir, de la persévérance et du travail.

Je dédie humblement ce mémoire.

Résumé

L'utilisation intensive des herbicides en agriculture contribue à l'augmentation des risques de contamination des sols et peut engendrer des effets néfastes sur les écosystèmes et la santé humaine. Dans ce contexte, les approches QSAR (Quantitative Structure–Activity Relationship) constituent des outils performants permettant de prédire les propriétés toxicologiques des substances chimiques à partir de leur structure moléculaire.

L'objectif de ce travail est de développer des modèles QSAR capables de prédire la toxicité aiguë des herbicides présents dans les sols agricoles. Une base de données composée de 34 herbicides a été constituée à partir de données toxicologiques disponibles dans la littérature et les bases de données spécialisées. Les descripteurs moléculaires ont été calculés à l'aide du logiciel ChemMaster puis sélectionnés après différentes étapes de prétraitement et d'analyse statistique.

Quatre modèles QSAR ont été construits en utilisant la méthode de Régression Linéaire Multiple (MLR). Les performances des modèles ont été évaluées à l'aide des coefficients de détermination (R^2), de validation croisée (Q^2), ainsi que des indicateurs RMSE et MAE. Le modèle optimal a présenté un coefficient de détermination de 0,905 et les plus faibles erreurs de prédiction parmi les modèles étudiés.

Afin d'évaluer son applicabilité, le modèle retenu a été utilisé pour prédire les valeurs de toxicité (LD50) de dix herbicides externes. Les résultats obtenus démontrent l'intérêt des approches QSAR comme outil d'aide à l'évaluation préliminaire des risques toxicologiques liés aux herbicides et soulignent leur potentiel pour réduire le recours aux expérimentations animales.

Mots-Clés : QSAR, herbicides, LD50, sols agricoles, régression linéaire multiple (MLR), ChemMaster.

Abstract

The intensive use of herbicides in agriculture contributes to soil contamination and may cause adverse effects on ecosystems and human health. In this context, Quantitative Structure–Activity Relationship (QSAR) approaches represent valuable computational tools for predicting the toxicological properties of chemical compounds based on their molecular structure.

The objective of this study was to develop QSAR models capable of predicting the acute toxicity of herbicides commonly found in agricultural soils. A database containing 34 herbicides was compiled from toxicological information available in scientific literature and specialized databases. Molecular descriptors were calculated using ChemMaster software and subsequently selected through statistical preprocessing and variable selection procedures.

Four QSAR models were developed using Multiple Linear Regression (MLR). Model performance was assessed using the coefficient of determination (R^2), cross-validation coefficient (Q^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The best-performing model achieved an R^2 value of 0.905 and exhibited the lowest prediction errors among all developed models.

To evaluate its practical applicability, the selected model was applied to predict the LD50 values of ten external herbicides. The obtained results demonstrate the usefulness of QSAR approaches as effective tools for the preliminary assessment of herbicide toxicity and highlight their potential to reduce reliance on animal testing.

Keywords : QSAR, herbicides, LD50, agricultural soils, Multiple Linear Regression (MLR), ChemMaster

ملخص

يُعد الاستخدام المكثف لمبيدات الأعشاب في الزراعة من العوامل الرئيسية المسببة لتلوث التربة، مما قد يؤدي إلى آثار سلبية على البيئة وصحة الإنسان. لذلك أصبح من الضروري تطوير وسائل سريعة وفعالة لتقييم السمية دون الاعتماد الكامل على التجارب المخبرية المكلفة والمستهلكة للوقت.

يهدف هذا العمل إلى نمذجة سمية مبيدات الأعشاب الموجودة في التربة الزراعية باستخدام منهجية العلاقة الكمية بين البنية والنشاط (QSAR). تم إنشاء قاعدة بيانات تضم 34 مبيد أعشاب مع قيم السمية الحادة، (LD50) كما تم حساب مجموعة من الواصفات الجزيئية باستعمال برنامج ChemMaster.

بعد معالجة البيانات واختيار المتغيرات الأكثر أهمية، تم تطوير أربعة نماذج QSAR باستعمال الانحدار الخطي المتعدد (MLR). وقد تم تقييم أداء النماذج باستخدام معاملات إحصائية مختلفة مثل معامل التحديد، (R^2) ومعامل التحقق المتقاطع، (Q^2) ومتوسط الخطأ المطلق، (MAE) والجذر التربيعي لمتوسط مربع الخطأ. (RMSE)

أظهرت النتائج أن النموذج الرابع كان الأفضل من حيث الأداء، حيث حقق قيمة R^2 بلغت 0.905 مع أقل قيم للأخطاء الإحصائية. كما تم تطبيق هذا النموذج على عشرة مبيدات أعشاب جديدة من أجل التنبؤ بقيم السمية الخاصة بها، مما أكد فعالية منهجية QSAR في التقييم الأولي للمخاطر السمية.

تُبرز هذه الدراسة أهمية النمذجة الجزيئية وأساليب QSAR كأدوات فعالة لدعم تقييم المخاطر البيئية وتقليل الحاجة إلى التجارب الحيوانية، مع إمكانية تطوير نماذج أكثر دقة من خلال توسيع قاعدة البيانات واستخدام تقنيات تعلم آلي أكثر تقدماً.

الكلمات المفتاحية :

QSAR ، مبيدات الأعشاب، LD50 ، التربة الزراعية، الانحدار الخطي المتعدد، (MLR) ChemMaster

Sommaire

Résumé	B
Abstract	C
ملخص	D
Liste des abréviations	i
Liste des figures	ii
Liste des tableaux	iii
Introduction Générale	1
Chapitre I : SYNTHÈSE BIBLIOGRAPHIQUE	4
1. Les Herbicides et agriculture	5
1.1. Aperçu historique.....	5
1.1.1. Définition des herbicides.....	6
1.1.2. Classification des herbicides.....	7
1.1.2.1. Classification selon la structure chimique	7
1.2.2. Classification selon le mode d'action	8
1.1.2.3. Classification selon la sélectivité.....	8
1.1.3. Exemples d'herbicides courants	8
1.2. Comportement des herbicides dans les sols.....	9
1.2.1. Adsorption et désorption	10
1.2.2. Dégradation des herbicides	10
1.2.3. Mobilité et lixiviation.....	10
1.2.4. Persistance et bioaccumulation	11
1.2.5. Influence des propriétés du sol	12
1.3. Toxicité environnementale	12
1.3.1. Définition de la toxicité	13
1.3.2. Paramètres toxicologiques.....	13
1.3.3. Organismes bioindicateurs	14
1.3.4. Effets éco toxicologiques des herbicides	14
1.4. Approche QSAR	15
1.4.1. Principe de l'approche QSAR	15
1.4.2. Historique du QSAR	16
1.4.3. Descripteurs moléculaires.....	17
1.4.4. Construction des modèles QSAR	18

I.4.5. Validation des modèles QSAR.....	19
Chapitre II : Matériels et Méthodes	22
II.1. Constitution de la base de données.....	23
II.1.1. Sélection des herbicides	24
II.1.2. Sources des données	24
II.1.2.1. PubChem	24
II.1.2.2. United States Environmental Protection Agency	25
II.2. Calcul des descripteurs moléculaires.....	26
II.2.1. Types de descripteurs utilisés.....	26
II.2.2. Logiciel utilisé :	26
II.2.3. Prétraitement des données.....	27
II.3. Sélection des variables.....	28
II.3.1. Élimination des descripteurs redondants	28
II.3.2. Méthodes de sélection.....	28
II.3. Sélection des variables.....	29
II.3.1. Élimination des descripteurs redondants	29
II.3.2. Méthodes de sélection.....	29
II.3.2.1. Analyse de corrélation	29
II.3.2.2. Analyse en composantes principales (PCA)	30
II.3.2.3. Méthodes automatiques de sélection.....	30
II.4. Construction des modèles QSAR	31
II.4.1. Division des données.....	31
II.4.2. Méthodes utilisées.....	32
II.4.3. Paramétrage du modèle MLR.....	33
II.5. Validation des modèles.....	34
II.5.1. Validation interne.....	35
II.5.2. Validation externe.....	35
II.5.3. Indicateurs statistiques	36
CHAPITRE III : RÉSULTATS ET DISCUSSION	38
III.1. Analyse de la base de données.....	39
III.1.1. Description générale de la base de données	39
III.1.2. Statistiques descriptives	40
III.2. Analyse des descripteurs moléculaires.....	42
III.2.1. Analyse des corrélations.....	42
III.2.2. Importance des variables.....	42

III.3. Modèles QSAR obtenus	43
III.3.1. Modèle QSAR 1.....	43
III.3.2. Modèle QSAR 2.....	44
III.3.3. Modèle QSAR 3.....	45
III.3.4. Modèle QSAR 4.....	46
III.3.1. Résultats des modèles QSAR.....	47
III.3.2. Comparaison des méthodes	48
III.4. Validation des modèles.....	49
III.4.1. Résultats de la validation interne	49
III.4.2. Résultats de la validation externe	50
III.5. Application du modèle optimal à de nouveaux herbicides	51
III.5.1. Sélection des herbicides externes	51
III.5.2. Prédiction des valeurs de LD50	52
III.5.3. Discussion des résultats obtenus.....	53
III.6. Discussion générale	54
III.6.1. Forces du modèle.....	54
III.6.2. Limites de l'étude.....	54
III.6.3. Comparaison avec la littérature	55
III.6.4. Perspectives d'amélioration	55
CONCLUSION GÉNÉRALE	57
REFERENCES BIBLIOGRAPHIQUES.....	59

Liste des abréviations

AD : Applicability Domain (Domaine d'applicabilité)
EPA : Environmental Protection Agency
LD50 : Dose Létale Médiane (Lethal Dose 50%)
MAE : Mean Absolute Error (Erreur Absolue Moyenne)
MLR : Multiple Linear Regression (Régression Linéaire Multiple)
OECD : Organisation for Economic Co-operation and Development (Organisation de Coopération et de Développement Économiques)
PCA : Principal Component Analysis (Analyse en Composantes Principales)
PubChem : Public Chemical Database
QSAR : Quantitative Structure–Activity Relationship (Relation Quantitative Structure–Activité)
Q² : Coefficient de Validation Croisée
R² : Coefficient de Détermination
RMSE : Root Mean Square Error (Erreur Quadratique Moyenne)
SMILES: Simplified Molecular Input Line Entry System
SVM : Support Vector Machine
AI : Artificial Intelligence (Intelligence Artificielle)
ANN : Artificial Neural Network (Réseau de Neurones Artificiel)
MW : Molecular Weight (Masse Moléculaire)
LogP : Coefficient de Partage Octanol/Eau
MR : Molar Refractivity (Réfractivité Molaire)
RF : Random Forest (Forêt Aléatoire)
kNN: k-Nearest Neighbors (k Plus Proches Voisins)
CV : Cross Validation (Validation Croisée)
ADMET: Absorption, Distribution, Metabolism, Excretion and Toxicity
h* : Levier Critique utilisé dans le diagramme de Williams

Liste des figures

Figure I-1 : La surutilisation des pesticides	5
Figure I-2 : Diagramme conceptuel des différentes sources et voies de contamination des eaux souterraines	11
Figure I-3: Les différents mécanismes biologiques de la phytoremédiation des sols contaminés	12
Figure I-4: Représentation graphique de la courbe dose-réponse et échelle d'évaluation de la Dose Létale 50 % (DL50)	13
Figure I-5: Le ver de terre (<i>Eisenia fetida</i>), un organisme bioindicateur clé utilisé pour l'évaluation de la toxicité des sols.	14
Figure I-6 : Étapes fondamentales de la construction d'un modèle prédictif basé sur la relation structure-activité/propriété	16
Figure I-7: Frise chronologique des étapes clés et jalons historiques du développement de la modélisation QSAR	17
Figure I-8 : Logigramme décisionnel du processus d'apprentissage (Training) et de test (Testing) pour le développement d'un modèle QSAR robuste.....	19
Figure I-9 : Diagramme synoptique des principaux domaines d'application des modèles QSAR dans la recherche et l'industrie	20
Figure II-1: Étapes méthodologiques allant de la collecte des données chimiques à l'interprétation structurale des modèles prédictifs	23
Figure II-2: Interface utilisateur de la base de données ouverte PubChem (NCBI) pour la collecte des données structurales et toxicologiques	25
Figure II-3 : Étude écotoxicologique des pesticides par modélisation moléculaire	27
Figure II-4 : Principe de la séparation aléatoire des données en sous-ensembles d'apprentissage et de test	31
Figure II-5: Les différentes méthodes de répartition d'un jeu de données en sous-ensembles d'apprentissage, de validation et de test	32
Figure II-6 : Étapes globales du développement, de la sélection de variables et de la validation d'un modèle QSAR	34
Figure II-7 : Logigramme de sélection des paramètres optimaux pour la construction d'un modèle QSAR robuste	36
Figure III-1: Diagramme en boîte (boxplot) illustrant la dispersion des valeurs de LD50 au sein de la base de données.	41
Figure III-2: Comparaison des performances statistiques des quatre modèles QSAR développés.	49
Figure III-3 : Interface logicielle présentant les structures 2D des herbicides et leurs valeurs de toxicité prédites par le modèle QSAR 4.....	52

Liste des tableaux

Tableau III-1: Distribution des valeurs de LD50 des 34 herbicides étudiés.	41
Tableau III-2 : Performances des modèles QSAR développés	47
Tableau III-3 : Résultats de la validation externe	50
Tableau III-4: Herbicides utilisés pour l'application du modèle QSAR.....	52
Tableau III-5 : Valeurs de LD50 prédites par le modèle QSAR 4.....	53

Introduction Générale

INTRODCUTION GENERALE

L'agriculture moderne repose largement sur l'utilisation des produits phytosanitaires afin d'augmenter la productivité agricole et de protéger les cultures contre les mauvaises herbes, les insectes et les maladies. Parmi ces produits, les herbicides occupent une place importante en raison de leur efficacité dans le contrôle des adventices et de leur utilisation intensive dans les systèmes agricoles. Cependant, l'emploi excessif et parfois incontrôlé de ces substances chimiques entraîne une contamination progressive des sols agricoles ainsi que des différents compartiments de l'environnement.

Après leur application, les herbicides peuvent subir plusieurs phénomènes tels que l'adsorption, la dégradation, la lixiviation ou encore la bioaccumulation. Une partie de ces composés persiste dans le sol pendant de longues périodes, ce qui peut provoquer des effets toxiques sur les organismes non ciblés, notamment les microorganismes du sol, les plantes, les vers de terre comme *Eisenia fetida*, ainsi que sur les ressources hydriques. Certains herbicides tels que Glyphosate, Atrazine et 2,4-D ont fait l'objet de nombreuses études en raison de leurs impacts éco-toxicologiques et de leur persistance environnementale.

L'évaluation expérimentale de la toxicité des herbicides nécessite généralement des essais biologiques longs, coûteux et parfois difficiles à réaliser. Face à ces contraintes, les méthodes de modélisation *in silico* se sont fortement développées au cours des dernières années. Parmi elles, l'approche QSAR (Quantitative Structure–Activity Relationship) représente un outil scientifique puissant permettant de relier les propriétés structurales et physicochimiques des molécules à leur activité biologique ou à leur toxicité.

Le principe du QSAR repose sur l'hypothèse selon laquelle des molécules présentant des structures chimiques similaires possèdent des comportements toxicologiques comparables. Cette approche utilise différents descripteurs moléculaires, tels que la masse moléculaire, la lipophilie, la polarité ou encore les propriétés électroniques, afin de construire des modèles mathématiques capables de prédire la toxicité de composés chimiques sans recourir systématiquement à l'expérimentation animale ou aux analyses biologiques classiques.

Dans ce contexte, la modélisation QSAR constitue une alternative prometteuse pour l'évaluation de la toxicité environnementale des herbicides présents dans les sols agricoles. Elle permet non seulement de réduire le temps et le coût des études toxicologiques, mais également d'identifier les facteurs moléculaires responsables des effets toxiques et de contribuer au développement de substances moins dangereuses pour l'environnement.

Le présent travail a pour objectif de développer et d'évaluer des modèles QSAR capables de prédire la toxicité des herbicides à partir de leurs caractéristiques structurales. Pour atteindre cet objectif, une base de données regroupant différents herbicides et leurs paramètres toxicologiques sera constituée à partir de bases de données scientifiques telles que PubChem et EPA. Des descripteurs moléculaires seront calculés à l'aide de logiciels spécialisés comme PaDEL-Descriptor et analysés par différentes méthodes statistiques et informatiques afin de construire des modèles prédictifs fiables.

Ce mémoire est structuré en trois chapitres principaux. Le premier chapitre présente une synthèse bibliographique sur les herbicides, leur comportement dans les sols agricoles, les notions de toxicité environnementale ainsi que les principes fondamentaux de l'approche QSAR. Le deuxième chapitre est consacré aux matériels et méthodes utilisés pour la

INTRODCUTION GENERALE

constitution de la base de données, le calcul des descripteurs moléculaires, la construction et la validation des modèles QSAR. Enfin, le troisième

chapitre expose les résultats obtenus, leur interprétation et la discussion scientifique des performances des modèles développés

Chapitre I : SYNTHÈSE

BIBLIOGRAPHIQUE

Ce chapitre sert à présenter toutes les bases théoriques et scientifiques nécessaires pour comprendre le sujet avant de passer à la modélisation QSAR

1. Les Herbicides et agriculture

L'agriculture moderne connaît une utilisation croissante des produits phytosanitaires afin d'assurer une production agricole suffisante pour répondre aux besoins alimentaires de la population mondiale. Parmi ces produits, les herbicides occupent une place importante grâce à leur efficacité dans la lutte contre les mauvaises herbes, appelées également adventices. Ces dernières concurrencent les cultures pour l'eau, la lumière, les éléments nutritifs et l'espace, entraînant ainsi une diminution importante des rendements agricoles.



Figure I-1 : La surutilisation des pesticides [1]

1.1. Aperçu historique

Depuis l'Antiquité, les pratiques agricoles reposent principalement sur le désherbage manuel et mécanique pour éliminer les mauvaises herbes et protéger les cultures. Cependant, avec l'augmentation de la population mondiale et le développement de l'agriculture intensive, ces méthodes sont devenues insuffisantes pour assurer des rendements agricoles élevés. L'apparition des herbicides chimiques au XXe siècle a donc constitué une étape majeure dans l'évolution des pratiques agricoles [2].

Le développement des herbicides modernes débute réellement dans les années 1940 avec la découverte de 2,4-D (acide 2,4-dichlorophénoxyacétique), considéré comme le premier herbicide sélectif utilisé à grande échelle. Ce composé permettait de détruire certaines mauvaises herbes à feuilles larges sans affecter significativement les cultures céréalières. Son introduction a marqué le début de l'agriculture chimique moderne [3].

Au cours des années 1950 et 1960, plusieurs nouvelles familles d'herbicides ont été développées, notamment les triazines comme Atrazine, largement utilisées dans les cultures de maïs. Ces composés étaient appréciés pour leur efficacité et leur persistance dans les sols, permettant une protection prolongée contre les adventices. Cependant, leur forte stabilité chimique a également favorisé leur accumulation dans l'environnement.[4].

Dans les années 1970, l'introduction de Glyphosate a révolutionné les pratiques agricoles grâce à son large spectre d'action et à son efficacité élevée contre de nombreuses espèces

CHAPITRE I : SYNTHÈSE BIBLIOGRAPHIQUE

végétales. Son utilisation s'est encore intensifiée avec le développement des cultures génétiquement modifiées résistantes au glyphosate dans les années 1990[5]

Malgré leurs avantages agricoles, plusieurs études scientifiques ont montré que certains herbicides peuvent persister dans les sols, contaminer les eaux souterraines et provoquer des effets toxiques sur les organismes non ciblés. Ces préoccupations environnementales et sanitaires ont conduit à un intérêt croissant pour les méthodes alternatives d'évaluation de la toxicité, notamment les approches de modélisation QSAR permettant de prédire les effets toxiques à partir de la structure chimique des molécules [6].

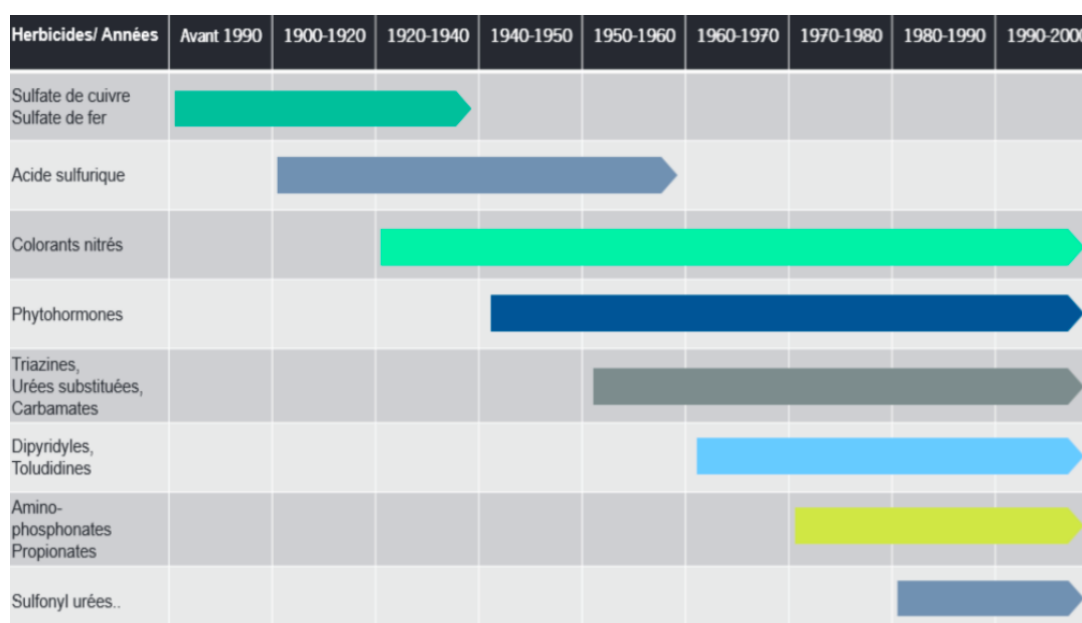


Tableau 1-1 : Évolution chronologique de l'utilisation des herbicides [7]

1.1.1. Définition des herbicides

Les herbicides sont des substances chimiques ou biologiques utilisées pour détruire, inhiber ou contrôler la croissance des mauvaises herbes, appelées également adventices, présentes dans les cultures agricoles [8]. Leur utilisation vise principalement à réduire la concurrence entre les adventices et les plantes cultivées pour les ressources essentielles telles que l'eau, la lumière, les éléments nutritifs et l'espace.

Les herbicides occupent une place importante parmi les produits phytosanitaires en raison de leur efficacité dans l'amélioration des rendements agricoles et la réduction des pertes de production [9]. Ils sont largement employés dans l'agriculture moderne, notamment dans les cultures céréalières, maraîchères et industrielles. Certains herbicides sont appliqués avant la levée des cultures (pré-émergence), tandis que d'autres sont utilisés après le développement des plantes (post-émergence).

Selon leur mode d'action, les herbicides peuvent agir de différentes manières sur les végétaux. Certains inhibent la photosynthèse, d'autres perturbent la synthèse des protéines, des lipides ou des acides aminés essentiels à la croissance des plantes [10]. Par exemple,

Glyphosate agit en bloquant une enzyme impliquée dans la biosynthèse des acides aminés aromatiques, alors qu'Atrazine inhibe le processus de photosynthèse

Les herbicides peuvent également être classés en deux grandes catégories : les herbicides sélectifs et les herbicides non sélectifs. Les herbicides sélectifs ciblent certaines espèces végétales sans affecter significativement la culture principale, tandis que les herbicides non sélectifs détruisent la majorité des plantes avec lesquelles ils entrent en contact [11].

Malgré leurs avantages agricoles, l'utilisation intensive des herbicides peut entraîner plusieurs conséquences environnementales, notamment la contamination des sols et des eaux, la persistance de résidus toxiques ainsi que des effets nocifs sur les organismes non ciblés [6]. Ces préoccupations ont conduit au développement de nouvelles approches scientifiques, telles que les modèles QSAR, permettant d'évaluer et de prédire la toxicité des herbicides à partir de leurs propriétés moléculaires.

1.1.2. Classification des herbicides

Les herbicides peuvent être classés selon plusieurs critères, notamment leur structure chimique, leur mode d'action, leur sélectivité ainsi que leur période d'application. Cette classification permet de mieux comprendre leurs mécanismes d'action, leurs propriétés environnementales et leurs effets toxiques [8].

1.1.2.1. Classification selon la structure chimique

Les herbicides sont regroupés en différentes familles chimiques possédant des propriétés et des mécanismes d'action spécifiques.

- a. **Les triazines** : sont des herbicides utilisés principalement pour le contrôle des mauvaises herbes dans les cultures céréalières. Elles agissent en inhibant la photosynthèse au niveau du photosystème II. Parmi les composés les plus connus figure Atrazine [12].
- b. **Les acides phénoxy** : Cette famille comprend notamment 2,4-D et MCPA. Ces herbicides agissent comme des hormones végétales synthétiques provoquant une croissance incontrôlée des plantes ciblées [3].
- c. **Les organophosphorés** : Le principal représentant de cette famille est Glyphosate. Il agit en inhibant une enzyme essentielle à la biosynthèse des acides aminés aromatiques chez les plantes [5].
- d. **Les sulfonilurées** : Les sulfonilurées sont des herbicides très actifs à faible dose. Elles inhibent la synthèse de certains acides aminés indispensables à la croissance végétale [13].
- e. **Les carbamates** : Les herbicides carbamates perturbent plusieurs processus métaboliques des végétaux et sont utilisés pour le contrôle de diverses adventices [8].

1.2.2. Classification selon le mode d'action

les herbicides peuvent également être classés selon leur mécanisme d'action physiologique sur les plantes.

- a. **Inhibiteurs de la photosynthèse** : Ces herbicides bloquent le transfert des électrons lors de la photosynthèse, empêchant ainsi la production d'énergie nécessaire à la plante. **Exemple** : Atrazine [12].
- b. **Inhibiteurs de la synthèse des acides aminés** : ils empêchent la formation des protéines essentielles au développement des végétaux. **Exemple** : Glyphosate [5].
- c. **Inhibiteurs de la division cellulaire** : Ces herbicides bloquent la multiplication des cellules végétales et perturbent la croissance des racines et des tiges [14].
- d. **Herbicides hormonaux** : Ils imitent les hormones naturelles des plantes et provoquent une croissance anormale conduisant à la mort de la plante. **Exemple** : 2,4-D [3].

1.1.2.3. Classification selon la sélectivité

- a. **Herbicides sélectifs** : Les herbicides sélectifs détruisent certaines espèces végétales sans affecter significativement la culture principale lorsqu'ils sont utilisés correctement [8].
- b. **Herbicides non sélectifs** : Les herbicides non sélectifs éliminent la majorité des plantes avec lesquelles ils entrent en contact. Ils sont généralement utilisés pour le désherbage total. **Exemple** : Glyphosate [5].

1.1.3. Exemples d'herbicides courants

De nombreux herbicides sont utilisés dans les pratiques agricoles modernes pour contrôler les mauvaises herbes et améliorer les rendements des cultures. Ces composés diffèrent par leur structure chimique, leur mode d'action, leur persistance dans l'environnement ainsi que leur niveau de toxicité. Parmi les herbicides les plus répandus figurent Atrazine, 2,4-D et MCPA.

A) Atrazine

L'atrazine appartient à la famille des triazines et est principalement utilisée pour le désherbage des cultures de maïs et de canne à sucre [12]. Cet herbicide agit en inhibant le processus de photosynthèse au niveau du photosystème II.

L'atrazine est connu pour sa persistance relativement élevée dans les sols ainsi que sa mobilité vers les eaux souterraines. En raison de ses effets éco toxicologiques, son utilisation a été interdite ou fortement limitée dans plusieurs pays [15].

B) 2,4-D

Le 2,4-D (acide 2,4-dichlorophénoxyacétique) est un herbicide sélectif appartenant à la famille des acides phénoxy. Il est considéré comme l'un des premiers herbicides synthétiques utilisés à grande échelle depuis les années 1940 [3].

Le 2,4-D agit comme une hormone végétale synthétique provoquant une croissance anormale des adventices à feuilles larges. Il est largement utilisé dans les cultures céréalières et les pelouses. Bien qu'il soit moins persistant que d'autres herbicides, des études ont signalé certains effets toxiques sur les organismes aquatiques et les sols [16].

C) MCPA

Le MCPA (acide 2-méthyl-4-chlorophénoxyacétique) appartient également à la famille des acides phénoxy. Il est principalement utilisé pour lutter contre les mauvaises herbes dans les cultures de blé, d'orge et de riz [17].

Son mécanisme d'action est similaire à celui du 2,4-D, perturbant la croissance normale des plantes ciblées. Le MCPA présente une toxicité modérée mais peut contaminer les sols et les eaux lorsqu'il est utilisé de manière excessive [6].

Ces herbicides représentent des modèles importants dans les études éco-toxicologiques en raison de leur large utilisation agricole et de leurs impacts potentiels sur l'environnement. Leur diversité structurale et toxicologique les rend particulièrement intéressants dans le cadre des études QSAR visant à prédire la toxicité à partir des propriétés moléculaires.

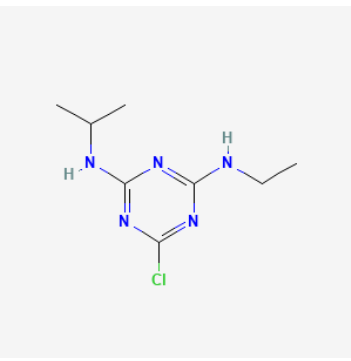
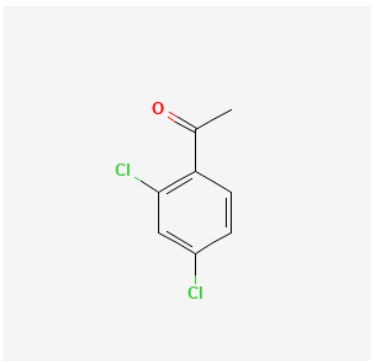
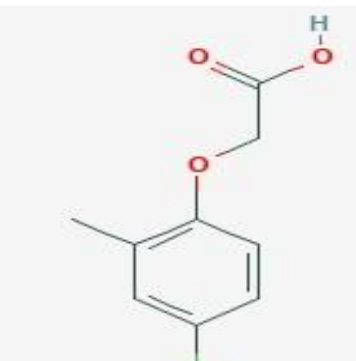
		
Atrazine	MCPA	2,4-D

Tableau 1-2 : Structures chimiques des principaux exemples herbicides étudiés (Atrazine, MCPA et 2,4-D)

I.2. Comportement des herbicides dans les sols

Le comportement des herbicides dans les sols agricoles est un aspect fondamental pour comprendre leur efficacité agronomique ainsi que leurs impacts environnementaux. Une fois appliqués, les herbicides ne restent pas statiques : ils subissent plusieurs processus physico-chimiques et biologiques qui déterminent leur devenir, leur persistance et leur toxicité potentielle dans l'environnement [18].

I.2.1. Adsorption et désorption

L'adsorption correspond à la fixation des molécules d'herbicides sur les particules du sol, notamment l'argile et la matière organique. Ce phénomène joue un rôle clé dans la disponibilité du produit dans le sol. L'intensité de l'adsorption dépend de plusieurs facteurs :

- la teneur en matière organique du sol
 - la texture (argile, limon, sable)
 - le pH du sol
- les propriétés physico-chimiques de l'herbicide (lipophilie, polarité)

Une forte adsorption limite la mobilité de l'herbicide mais peut prolonger sa persistance dans le sol, tandis qu'une faible adsorption favorise son lessivage vers les eaux souterraines [19].

I.2.2. Dégradation des herbicides

Les herbicides présents dans le sol peuvent être dégradés par deux principaux mécanismes:

a) Dégradation biologique

Elle est assurée par les microorganismes du sol (bactéries, champignons) qui transforment les herbicides en composés plus simples, parfois moins toxiques. Ce processus dépend fortement de l'activité biologique du sol [20].

b) Dégradation abiotique

Elle inclut :

- l'hydrolyse chimique
- la photodégradation (lumière solaire)
- les réactions d'oxydation-réduction
- La vitesse de dégradation est souvent exprimée par la demi-vie (DT50), qui représente le temps nécessaire pour que 50 % de la substance soit dégradée [21].

I.2.3. Mobilité et lixiviation

La mobilité des herbicides dans le sol dépend de leur solubilité dans l'eau et de leur capacité d'absorption. Les herbicides très mobiles peuvent migrer vers les couches profondes du sol et atteindre les nappes phréatiques, provoquant une contamination des ressources en eau

potable. Par exemple, certains herbicides hydrophiles présentent un risque élevé de lixiviation, surtout dans les sols sableux et fortement irrigués [22].



Figure I-2 : Diagramme conceptuel des différentes sources et voies de contamination des eaux souterraines [23].

I.2.4. Persistance et bioaccumulation

La persistance d'un herbicide dépend de sa résistance à la dégradation. Certains composés peuvent rester actifs dans le sol pendant plusieurs semaines, mois, voire années.

La bioaccumulation correspond à l'accumulation progressive de substances chimiques dans les organismes vivants. Elle peut affecter les microorganismes du sol, les plantes et les organismes supérieurs, perturbant ainsi les équilibres écologiques [24].

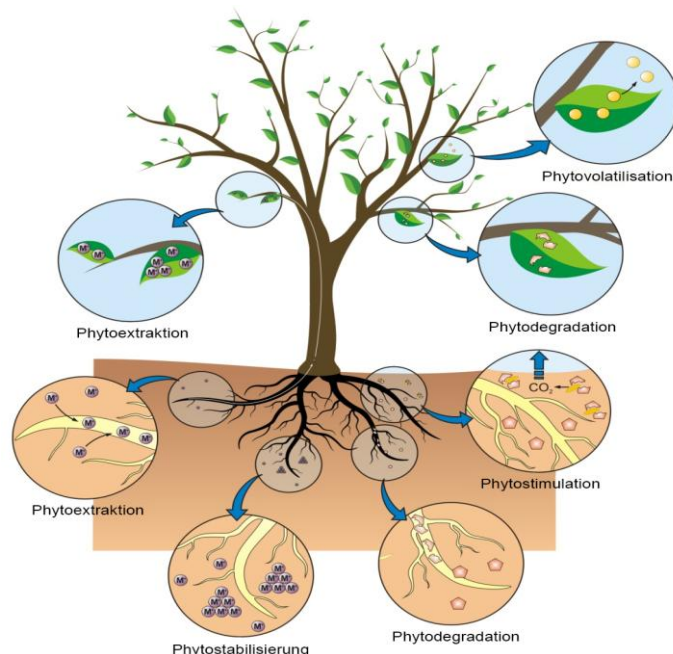


Figure I-3: Les différents mécanismes biologiques de la phytoremédiation des sols contaminés [25].

I.2.5. Influence des propriétés du sol

Le comportement des herbicides est fortement influencé par les caractéristiques du sol :

- pH : influence la charge des molécules
- matière organique : augmente l'adsorption
- humidité : favorise la mobilité et la dégradation biologique
- température : accélère les réactions chimiques et biologiques

Ainsi, un même herbicide peut avoir un comportement très différent selon le type de sol où il est appliqué [26].

I.3. Toxicité environnementale

La toxicité environnementale représente l'ensemble des effets nocifs qu'une substance chimique peut provoquer sur les organismes vivants et les différents compartiments de l'environnement. Dans le domaine agricole, les herbicides constituent une source importante de contamination environnementale en raison de leur utilisation intensive et répétée dans les cultures [6].

Après leur application, une partie des herbicides peut persister dans les sols, être transportée vers les eaux de surface ou les nappes souterraines, puis affecter différents organismes non

ciblés tels que les microorganismes, les plantes, les invertébrés et les animaux aquatiques [27]. Les effets toxiques dépendent de plusieurs facteurs, notamment la concentration du produit, la durée d'exposition, les propriétés physico-chimiques du composé ainsi que la sensibilité biologique des organismes exposés.

I.3.1. Définition de la toxicité

La toxicité peut être définie comme la capacité d'une substance chimique à provoquer des effets néfastes sur un organisme vivant, un organe ou un écosystème [28]. Ces effets peuvent être immédiats ou apparaître après une exposition prolongée.

On distingue généralement :

- **la toxicité aiguë** : effets rapides après une courte exposition
- **La toxicité chronique** : effets apparaissent après une exposition prolongée à faible dose.

I.3.2. Paramètres toxicologiques

L'évaluation de la toxicité environnementale repose sur plusieurs paramètres expérimentaux permettant de mesurer les effets des substances chimiques.

- LC50 (Lethal Concentration 50)** : représente la concentration d'une substance provoquant la mort de 50 % des organismes testés pendant une période déterminée [29]. Une faible valeur de LC50 indique généralement une forte toxicité.
- LD50 (Lethal Dose 50)** : La LD50 correspond à la dose nécessaire pour provoquer la mort de 50 % des organismes exposés [30]. Ce paramètre est souvent exprimé en mg/kg de masse corporelle.

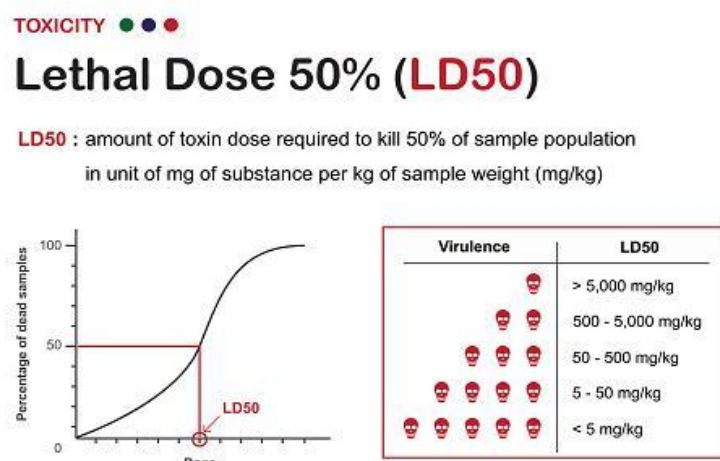


Figure I-4: Représentation graphique de la courbe dose-réponse et échelle d'évaluation de la Dose Létale 50 % (DL50) [31].

- c) **EC50 (Effective Concentration 50)** : La EC50 désigne la concentration entraînant un effet biologique chez 50 % des organismes étudiés, sans nécessairement provoquer leur mort [32].
- d) **NOEC (No Observed Effect Concentration)** : La NOEC correspond à la concentration maximale ne provoquant aucun effet observable sur les organismes testés [33].

I.3.3. Organismes bioindicateurs

Les études éco toxicologiques utilisent plusieurs organismes pour évaluer les effets des herbicides sur l'environnement.

- a) **Les vers de terre** : Les vers de terre tels que *Eisenia fetida* sont largement utilisés comme bioindicateurs de la qualité des sols en raison de leur sensibilité aux contaminants chimiques [34].



Figure I-5: Le ver de terre (*Eisenia fetida*), un organisme bioindicateur clé utilisé pour l'évaluation de la toxicité des sols[35].

- b) **Les microorganismes du sol** : Les bactéries et champignons jouent un rôle essentiel dans la fertilité des sols. Les herbicides peuvent perturber leur activité biologique et modifier les cycles biogéochimiques [36].
- c) **Les organismes aquatiques** : Les herbicides transportés vers les milieux aquatiques peuvent affecter les algues, les poissons et les invertébrés aquatiques [37]

I.3.4. Effets éco toxicologiques des herbicides

Les herbicides peuvent provoquer différents impacts environnementaux :

- diminution de la biodiversité
- perturbation des chaînes alimentaires
- contamination des eaux souterraines
- réduction de l'activité microbienne des sols

- effets toxiques chroniques sur les organismes non ciblés

Certains herbicides comme Atrazine et Glyphosate sont particulièrement étudiés en raison de leur persistance environnementale et de leurs effets éco toxicologiques potentiels [5].

I.4. Approche QSAR

L'approche QSAR (Quantitative Structure–Activity Relationship), ou relation quantitative structure–activité, constitue une méthode de modélisation mathématique permettant d'établir une relation entre la structure chimique d'une molécule et son activité biologique ou sa toxicité [38]. Cette approche repose sur l'hypothèse selon laquelle des molécules possédant des structures similaires présentent généralement des propriétés biologiques comparables.

Dans le domaine de l'écotoxicologie, les modèles QSAR occupent aujourd'hui une place importante dans plusieurs domaines scientifiques tels que la chimie médicinale, la toxicologie, la pharmacologie, la chimie environnementale et l'écotoxicologie, ils sont utilisés pour prédire la toxicité des substances chimiques, notamment les herbicides, à partir de leurs propriétés moléculaires. Ces méthodes représentent une alternative rapide et économique aux expérimentations biologiques classiques souvent coûteuses et longues à réaliser [39].

I.4.1. Principe de l'approche QSAR

Le principe fondamental du QSAR consiste à relier quantitativement l'activité biologique ou la toxicité d'un composé à des paramètres appelés descripteurs moléculaires.

La relation générale peut être représentée comme suit :

$$\text{Activité biologique} = f(\text{descripteurs moléculaires})$$

Les descripteurs moléculaires traduisent différentes propriétés physico-chimiques des molécules telles que : la lipophilie ; la masse moléculaire ; la polarité ; les propriétés électroniques ; les caractéristiques géométriques.

l'activité peut correspondre à : LC50, EC50, LD50, biodégradabilité, bioaccumulation.

Dans une étude QSAR, les propriétés moléculaires sont transformées en données numériques afin d'établir des corrélations statistiques entre structure et toxicité [40].

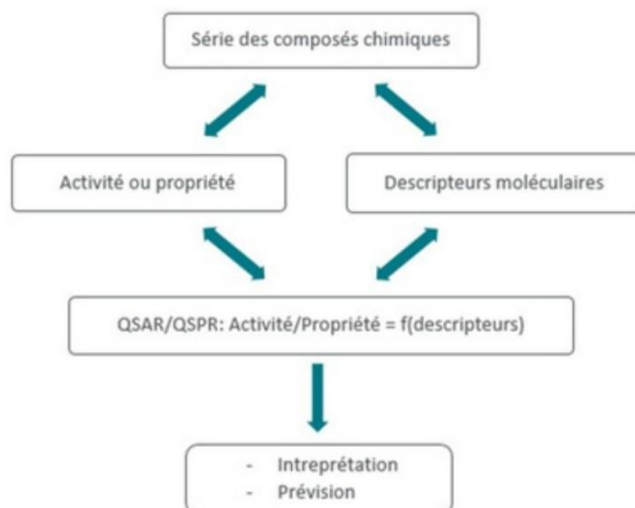


Figure I-6 : Étapes fondamentales de la construction d'un modèle prédictif basé sur la relation structure-activité/propriété [41].

I.4.2. Historique du QSAR

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. Par conséquent, ils ont supposé que « **l'activité physiologique** » dût être fonction de la structure chimique C.

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**") [1]. Une telle relation correspond à l'équation, où A traduit les différences dans les valeurs d'activité biologique, causées par des changements correspondants dans les propriétés chimiques et surtout physicochimiques : ΔC .

Le développement de la technique QSAR a subi plusieurs transformations 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).

La méthodologie QSAR s'est rapidement développée à partir du 1988, en raison de l'introduction du paramètre moléculaire tridimensionnel, qui expliquait l'influence de différents conformères, stéréoisomères ou 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. La figure ci-dessous résume le développement de l'étude QSAR [42].



Figure I-7: Frise chronologique des étapes clés et jalons historiques du développement de la modélisation QSAR [43].

Par la suite, les méthodes QSAR se sont considérablement développées grâce aux progrès de l'informatique, des statistiques et de la chimio-informatique. Aujourd'hui, les approches QSAR utilisent des techniques avancées telles que l'intelligence artificielle, l'apprentissage automatique (machine learning) et les réseaux de neurones artificiels pour améliorer la précision des prédictions [44].

I.4.3. Descripteurs moléculaires

Les descripteurs moléculaires sont des paramètres numériques décrivant les propriétés structurales et physico-chimiques des molécules.

a) Descripteurs physico-chimiques Ils incluent :

la masse moléculaire (MW): Elle influence : mobilité, diffusion, biodisponibilité.

le coefficient de partage octanol/eau ($\log P$): Le coefficient $\log P$ est particulièrement important dans les études QSAR car il reflète la lipophilie du composé :

$$\log P = \log \left(\frac{[\text{composé}]_{\text{octanol}}}{[\text{composé}]_{\text{eau}}} \right)$$

Un $\log P$ élevé indique : forte affinité pour les lipides et risque accru de bioaccumulation [45].

la surface polaire (TPSA): Elle influence : solubilité, transport membranaire, interactions biologiques.

- b) **Descripteurs topologiques** : Ils décrivent l'organisation des atomes et les connexions moléculaires.
- c) **Descripteurs électroniques** : Ils représentent les propriétés électroniques telles que les charges atomiques et la distribution des électrons.
- d) **Descripteurs géométriques** : Ils décrivent la forme tridimensionnelle et le volume des molécules [46].

I.4.4. Construction des modèles QSAR

La construction d'un modèle QSAR comprend plusieurs étapes :

1. sélection des molécules ;
2. collecte des données toxicologiques ;
3. calcul des descripteurs moléculaires ;
4. sélection des variables pertinentes ;
5. développement du modèle mathématique ;
6. validation statistique [47].

Les modèles peuvent être construits à l'aide de différentes méthodes statistiques ou informatiques :

- régression linéaire multiple (MLR) ;
- analyse PLS ;
- Random Forest ;
- Support Vector Machine (SVM) ;
- réseaux de neurones artificiels (ANN).

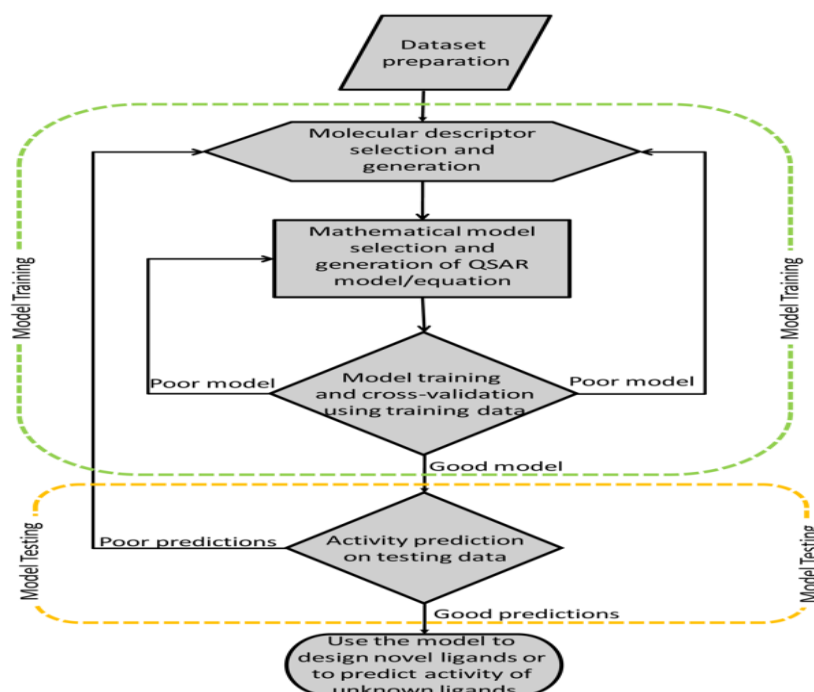


Figure I-8 : Logigramme décisionnel du processus d'apprentissage (Training) et de test (Testing) pour le développement d'un modèle QSAR robuste [48].

I.4.5. Validation des modèles QSAR

La validation permet d'évaluer la fiabilité et la capacité prédictive du modèle QSAR.

Les principaux paramètres statistiques utilisés sont :

a) Coefficient de détermination (R^2)

Une valeur élevée de R^2 indique une bonne corrélation entre les valeurs expérimentales et prédites.

b) Erreur quadratique moyenne (RMSE)

Une faible valeur de RMSE traduit une meilleure précision du modèle.

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 [48]:

- Produits chimiques bénéfiques pour la santé: médicaments, ingrédients alimentaires...etc
- Les produits chimiques impliqués dans les processus industriels de laboratoire: solvants, réactifs, etc.

CHAPITRE I : SYNTHÈSE BIBLIOGRAPHIQUE

- Les produits chimiques présentant un résultat dangereux: polluants organiques persistants, toxines, xénobiotiques, cancérogènes, composés organiques volatils,

La figure suivante présente l'utilisation de l'application QSAR en trois domaines principaux

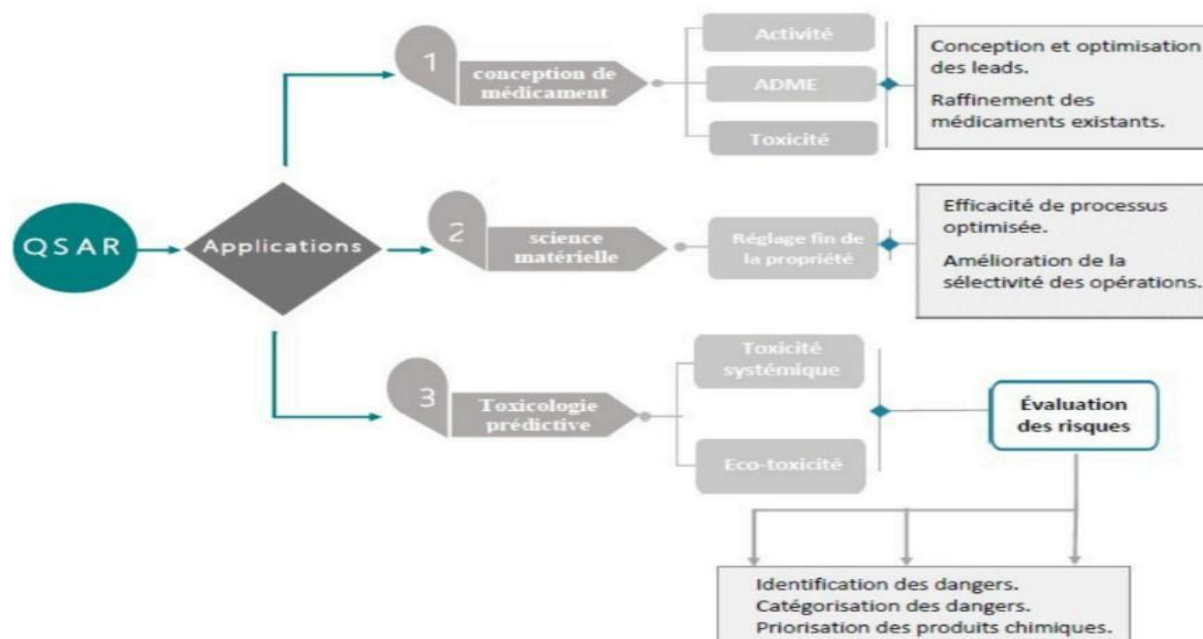


Figure I-9 : Diagramme synoptique des principaux domaines d'application des modèles QSAR dans la recherche et l'industrie [49].

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 [50] :

- 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.

CHAPITRE I : SYNTHÈSE BIBLIOGRAPHIQUE

Dans le cas des herbicides, les approches QSAR permettent de prédire des paramètres tels que :

- LC50 ;
- EC50 ;
- biodégradabilité ;
- bioaccumulation ;
- persistance dans les sols.

Chapitre II : Matériels et Méthodes

Ce chapitre présente la méthodologie adoptée pour réaliser l'étude de modélisation QSAR de la toxicité des herbicides présents dans les sols agricoles. Il décrit les différentes étapes suivies, depuis la collecte des données toxicologiques et des structures moléculaires jusqu'à la construction et la validation des modèles prédictifs. Ce chapitre permet de détailler les outils, logiciels, descripteurs moléculaires et méthodes statistiques utilisés afin de garantir la fiabilité et la reproductibilité des résultats obtenus.

II.1. Constitution de la base de données

La constitution de la base de données représente la première étape essentielle dans le développement d'un modèle QSAR fiable. Elle consiste à sélectionner un ensemble de molécules d'herbicides pour lesquelles des données toxicologiques expérimentales sont disponibles. La qualité et la diversité des données collectées influencent directement la performance et la capacité prédictive du modèle développé.

Dans cette étude, les herbicides ont été sélectionnés à partir de bases de données chimiques et toxicologiques reconnues telles que PubChem, EPA et la littérature scientifique spécialisée. Pour chaque composé, les informations recueillies comprennent le nom de la molécule, sa structure chimique ainsi que les valeurs de toxicité utilisées comme variable dépendante dans la modélisation QSAR.

Afin d'assurer la cohérence des données, les molécules présentant des informations incomplètes ou des valeurs toxicologiques contradictoires ont été exclues de la base de données. Une attention particulière a été accordée à la diversité structurale des herbicides afin de couvrir différentes familles chimiques. Cette diversité est importante pour améliorer le domaine d'applicabilité du modèle QSAR.

Les structures moléculaires des herbicides sélectionnés ont été récupérées sous forme de représentations SMILES (Simplified Molecular Input Line Entry System), format largement utilisé en chimio-informatique pour le calcul automatique des descripteurs moléculaires.

La base de données finale obtenue constitue ainsi l'ensemble des composés utilisés pour les étapes ultérieures de calcul des descripteurs, de construction du modèle et de validation statistique.

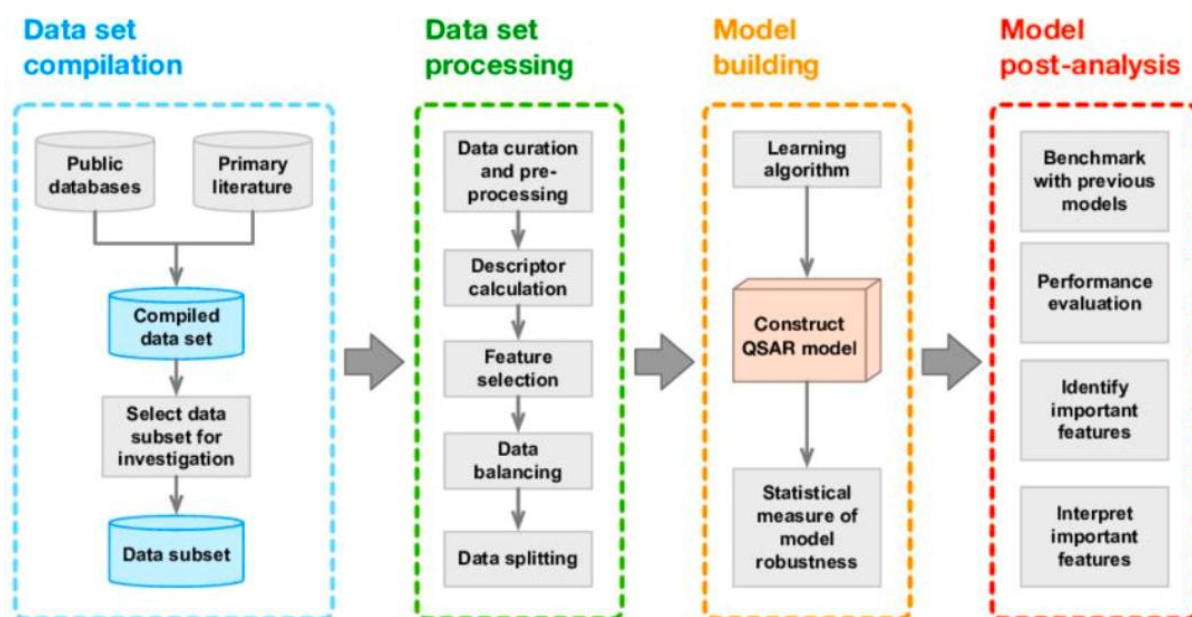


Figure II-1: Étapes méthodologiques allant de la collecte des données chimiques à l'interprétation structurale des modèles prédictifs [51].

II.1.1. Sélection des herbicides

La sélection des herbicides a été réalisée en tenant compte de plusieurs critères, notamment leur fréquence d'utilisation en agriculture, leur présence dans les sols agricoles et la disponibilité de données toxicologiques expérimentales. L'objectif était de constituer un ensemble de molécules représentatif de différentes familles chimiques afin de couvrir une diversité structurale suffisante pour la modélisation QSAR.

Les herbicides retenus appartiennent à plusieurs classes chimiques telles que les triazines, les organophosphorés, les acides phénoxy et les sulfonilurées. Pour chaque composé, les informations relatives à la structure moléculaire et aux paramètres toxicologiques ont été collectées puis vérifiées avant leur intégration dans la base de données.

Les molécules présentant des données incomplètes, ambiguës ou contradictoires ont été exclues afin de garantir la fiabilité de l'étude.

II.1.2. Sources des données

Les données utilisées dans cette étude ont été obtenues à partir de bases de données publiques reconnues dans le domaine de la chimie et de l'évaluation des risques environnementaux. Ces plateformes fournissent des informations détaillées sur les propriétés physico-chimiques, les structures moléculaires et les données toxicologiques des substances chimiques.

II.1.2.1. PubChem

PubChem est une base de données chimique développée par le National Center for Biotechnology Information (NCBI). Elle contient des millions de composés chimiques ainsi que des informations relatives à leurs propriétés moléculaires, leurs structures 2D et 3D et leurs représentations SMILES.

Dans cette étude, PubChem a été utilisé pour :

- récupérer les structures moléculaires des herbicides
- obtenir les représentations SMILES
- vérifier les formules moléculaires
- collecter certaines propriétés physico-chimiques utiles à la modélisation QSAR.

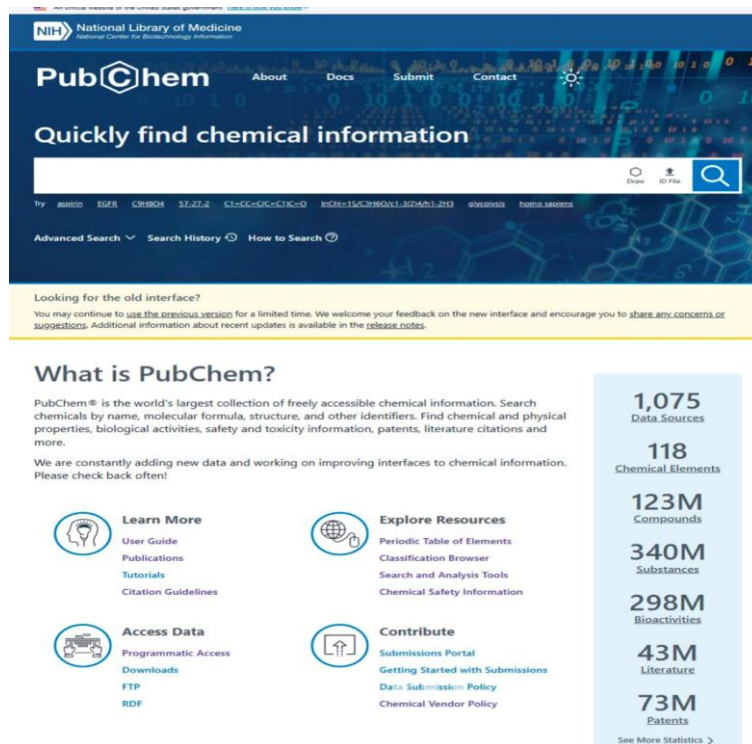


Figure II-2: Interface utilisateur de la base de données ouverte PubChem (NCBI) pour la collecte des données structurales et toxicologiques [52].

Grâce à son accès libre et à la qualité de ses données, PubChem constitue une référence importante dans les travaux de chimio-informatique et de modélisation moléculaire [53].

II.1.2.2. United States Environmental Protection Agency

L'EPA est une agence gouvernementale américaine chargée de la protection de l'environnement et de l'évaluation des risques liés aux substances chimiques. Elle met à disposition plusieurs bases de données toxicologiques contenant des informations sur les pesticides, les herbicides et leurs effets potentiels sur les organismes vivants [54].

Dans cette étude, les bases de données de l'EPA ont été utilisées pour :

- recueillir les valeurs toxicologiques expérimentales
- vérifier les données de toxicité environnementale
- compléter les informations nécessaires à la construction des modèles QSAR.

Les données fournies par l'EPA sont largement utilisées dans les études de toxicologie prédictive et d'évaluation des risques environnementaux.

II.2. Calcul des descripteurs moléculaires

Après la constitution de la base de données, les structures moléculaires des herbicides sélectionnés ont été utilisées pour calculer différents descripteurs moléculaires. Ces descripteurs représentent sous forme numérique les propriétés structurales, physico-chimiques et électroniques des molécules. Ils constituent les variables indépendantes utilisées dans la construction des modèles QSAR. Le choix de descripteurs pertinents est une étape essentielle car il influence directement la qualité des prédictions obtenues.

II.2.1. Types de descripteurs utilisés

Les descripteurs moléculaires calculés dans cette étude appartiennent à différentes catégories permettant de caractériser les molécules selon plusieurs aspects.

a) Descripteurs constitutionnels

Ils décrivent la composition élémentaire de la molécule, notamment : masse moléculaire ; nombre d'atomes ; nombre de liaisons ; nombre d'hétéroatomes.

b) Descripteurs topologiques

Ils traduisent l'organisation des atomes et des liaisons au sein de la molécule sans tenir compte de sa géométrie tridimensionnelle.

c) Descripteurs physico-chimiques

Ils incluent : $\log P$; polarité ; surface polaire ; réfractivité molaire.

d) Descripteurs électroniques

Ils renseignent sur la distribution électronique et les interactions potentielles des molécules avec leur environnement biologique.

L'ensemble de ces descripteurs permet de représenter quantitativement les caractéristiques moléculaires susceptibles d'influencer la toxicité des herbicides [38].

II.2.2. Logiciel utilisé :

Le calcul des descripteurs moléculaires a été réalisé à l'aide du logiciel ChemMaster. Ce logiciel permet l'analyse des structures chimiques et la génération automatique d'un grand nombre de descripteurs moléculaires à partir des représentations moléculaires des composés étudiés.

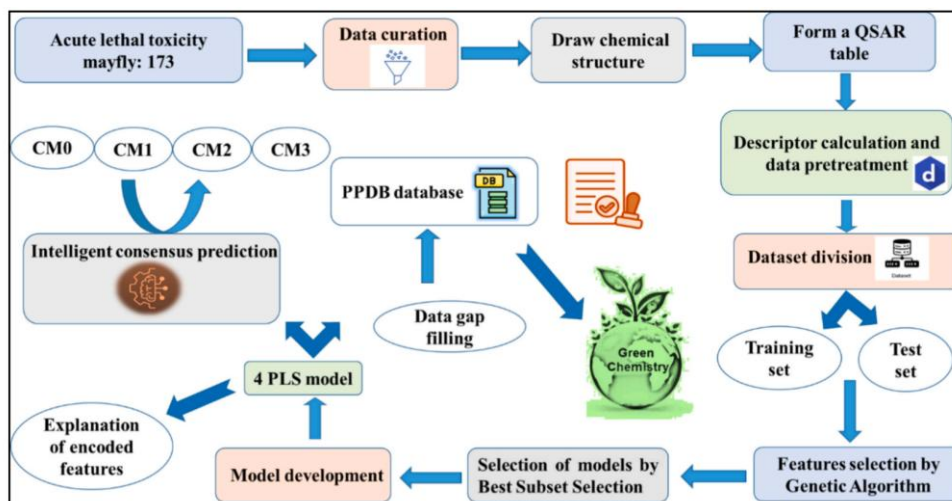


Figure II-3 : Étude écotoxicologique des pesticides par modélisation moléculaire [39].

Les structures des herbicides ont été importées dans ChemMaster puis soumises à une procédure de calcul automatique des descripteurs. Le logiciel a généré différentes catégories de paramètres décrivant les propriétés structurales et physico-chimiques des molécules.

L'utilisation de ChemMaster présente plusieurs avantages :

- Rapidité de calcul.
- Traitement simultané de nombreuses molécules.
- Génération automatisée des descripteurs.
- Exportation des résultats sous forme exploitable pour les analyses statistiques ultérieures.

II.2.3. Prétraitement des données

Avant la construction des modèles QSAR, les données obtenues ont été soumises à plusieurs étapes de prétraitement afin d'améliorer leur qualité et leur cohérence.

Les principales opérations réalisées sont :

- Suppression des descripteurs contenant des valeurs manquantes.
- Élimination des variables constantes ou quasi constantes.
- Vérification des erreurs de calcul.
- Normalisation des données lorsque cela est nécessaire.
- Contrôle des valeurs aberrantes.

Ces opérations permettent de réduire le bruit statistique et d'améliorer la stabilité des modèles développés [56].

II.3. Sélection des variables

Le calcul des descripteurs moléculaires génère généralement un nombre important de variables. Cependant, tous les descripteurs ne contribuent pas de manière significative à la prédiction de la toxicité. Une étape de sélection est donc nécessaire afin d'identifier les variables les plus pertinentes et de réduire la complexité du modèle QSAR.

II.3.1. Élimination des descripteurs redondants

Les descripteurs fortement corrélés entre eux apportent souvent des informations similaires. Leur présence simultanée peut entraîner des problèmes de multi colinéarité et diminuer les performances du modèle.

Pour éviter ce problème, les descripteurs présentant une forte corrélation mutuelle ont été identifiés puis éliminés. Seules les variables apportant une information indépendante ont été conservées pour les étapes suivantes de la modélisation [57].

II.3.2. Méthodes de sélection

a) Analyse de corrélation

L'analyse de corrélation a été utilisée pour évaluer les relations entre les descripteurs moléculaires et la variable de toxicité étudiée.

Les variables présentant une faible contribution ou une forte redondance ont été supprimées afin de simplifier le modèle et d'améliorer sa capacité prédictive.

b) Analyse en composantes principales (PCA)

L'analyse en composantes principales (PCA) est une méthode statistique permettant de réduire la dimension des données tout en conservant l'essentiel de l'information.

Cette technique facilite :

- L'identification des variables importantes.
- La détection des corrélations.
- La Visualisation de la distribution des molécules dans l'espace descriptif [58].

c) Méthodes automatiques de sélection

Des méthodes automatiques ont également été utilisées pour identifier les descripteurs les plus pertinents pour la modélisation QSAR.

Ces approches permettent de sélectionner automatiquement les variables ayant la plus grande influence sur la toxicité étudiée tout en éliminant les variables peu informatives.

L'objectif est d'obtenir un ensemble réduit de descripteurs capable d'expliquer efficacement les variations observées dans les données toxicologiques.

II.3. Sélection des variables

Après le calcul des descripteurs moléculaires, un grand nombre de variables est généralement obtenu. Cependant, tous les descripteurs ne contribuent pas de manière significative à l'explication de la toxicité des herbicides. La présence d'un nombre excessif de variables peut entraîner une augmentation de la complexité du modèle, un risque de surapprentissage (overfitting) et une diminution des performances prédictives. Une étape de sélection des variables est donc nécessaire afin d'identifier les descripteurs les plus pertinents pour la construction des modèles QSAR.

II.3.1. Élimination des descripteurs redondants

Avant l'étape de modélisation, les descripteurs redondants ont été identifiés et supprimés. Cette opération vise à réduire la multi colinéarité entre les variables explicatives et à améliorer la stabilité statistique des modèles QSAR.

Les descripteurs présentant une variance nulle ou très faible ont été éliminés puisqu'ils n'apportent aucune information discriminante entre les molécules étudiées. De même, les variables fortement corrélées entre elles ont été examinées afin de conserver uniquement les descripteurs les plus représentatifs.

Cette étape permet de diminuer le nombre de variables tout en préservant l'information essentielle contenue dans la base de données [56].

II.3.2. Méthodes de sélection

II.3.2.1. Analyse de corrélation

L'analyse de corrélation constitue l'une des méthodes les plus utilisées pour sélectionner les variables pertinentes dans les études QSAR. Elle permet d'évaluer la relation existant entre chaque descripteur moléculaire et la variable de toxicité étudiée.

Les coefficients de corrélation ont été calculés afin d'identifier les variables les plus influentes. Les descripteurs présentant une faible corrélation avec la propriété étudiée ou une forte corrélation avec d'autres descripteurs ont été exclus du processus de modélisation.

Cette méthode permet de réduire efficacement le nombre de variables tout en conservant celles qui contribuent le plus à l'explication de la toxicité [40].

II.3.2.2. Analyse en composantes principales (PCA)

L'analyse en composantes principales (Principal Component Analysis, PCA) est une méthode statistique multivariée utilisée pour réduire la dimension des données. Elle transforme les variables initiales en un nombre limité de composantes principales expliquant la plus grande partie de la variance observée dans le jeu de données.

La PCA a été utilisée pour :

- visualiser la distribution des herbicides.
- détecter les corrélations entre les descripteurs.
- Identifier les variables les plus informatives.
- Repérer d'éventuelles valeurs aberrantes.

Cette approche facilite la compréhension de la structure des données et contribue à la sélection des descripteurs les plus pertinents pour la modélisation QSAR [58].

II.3.2.3. Méthodes automatiques de sélection

En complément des méthodes statistiques classiques, des procédures automatiques de sélection ont été utilisées afin d'optimiser le choix des variables explicatives.

Ces méthodes permettent de rechercher automatiquement les combinaisons de descripteurs offrant les meilleures performances prédictives. Elles reposent sur des algorithmes capables d'évaluer l'importance de chaque variable et de sélectionner les plus pertinentes pour le modèle final.

L'utilisation de méthodes automatiques présente plusieurs avantages :

- Réduction du nombre de variables.
- Amélioration de la robustesse du modèle.
- Limitation du risque de surapprentissage.
- Optimisation des performances prédictives.

La combinaison des analyses de corrélation, de la PCA et des méthodes automatiques permet ainsi d'obtenir un ensemble de descripteurs représentatifs et adaptés au développement de modèles QSAR fiables [57].

II.4. Construction des modèles QSAR

Après la sélection des descripteurs moléculaires les plus pertinents, l'étape suivante consiste à construire les modèles QSAR permettant d'établir une relation quantitative entre les caractéristiques structurales des herbicides et leur toxicité. Cette étape repose sur l'utilisation de méthodes statistiques et d'apprentissage automatique capables d'extraire les relations existantes entre les variables explicatives (descripteurs moléculaires) et la variable réponse (toxicité). La qualité du modèle dépend à la fois de la pertinence des données utilisées et du choix de la méthode de modélisation.

II.4.1. Division des données

Afin d'évaluer objectivement les performances des modèles QSAR, la base de données a été divisée en deux sous-ensembles distincts : un ensemble d'apprentissage (training set) utilisé pour construire le modèle et un ensemble de test (test set) utilisé pour vérifier sa capacité prédictive sur des données non utilisées lors de l'entraînement.

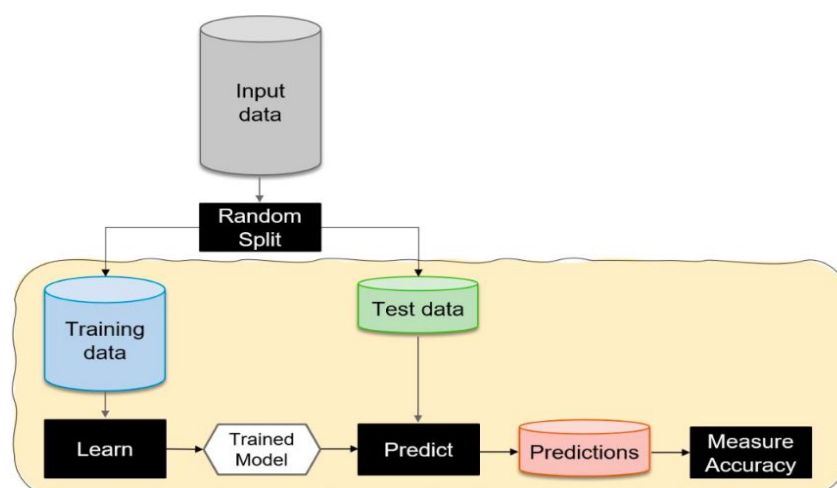


Figure II-4 : Principe de la séparation aléatoire des données en sous-ensembles d'apprentissage et de test [59].

a) Ensemble d'apprentissage (training set)

L'ensemble d'apprentissage regroupe la majorité des molécules de la base de données. Il est utilisé pour établir la relation entre les descripteurs moléculaires et la toxicité étudiée.

Durant cette phase, les paramètres du modèle sont ajustés afin d'obtenir la meilleure correspondance possible entre les valeurs expérimentales et les valeurs prédites. Généralement, environ 70 à 80 % des données sont affectées à l'ensemble d'apprentissage [60].

b) Ensemble de test

L'ensemble de test est constitué des molécules restantes qui ne participent pas à la construction du modèle.

Son rôle est de mesurer la capacité du modèle à effectuer des prédictions fiables sur de nouvelles molécules. Cette étape est essentielle pour vérifier que le modèle possède un réel pouvoir prédictif et qu'il n'est pas simplement adapté aux données d'apprentissage [61].

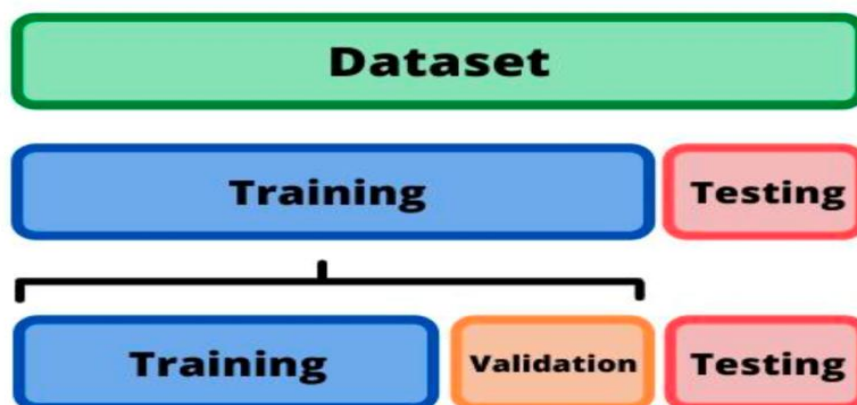


Figure II-5: Les différentes méthodes de répartition d'un jeu de données en sous-ensembles d'apprentissage, de validation et de test [62].

II.4.2. Méthodes utilisées

Plusieurs approches peuvent être employées pour développer des modèles QSAR. Dans cette étude, les méthodes de modélisation ont été sélectionnées en fonction de leur capacité à décrire les relations entre les descripteurs moléculaires et la toxicité des herbicides.

Les méthodes couramment utilisées en QSAR comprennent.

- la régression linéaire multiple (MLR).
- la régression PLS (Partial Least Squares).
- les arbres de décision.
- les forêts aléatoires (Random Forest).
- les machines à vecteurs de support (Support Vector Machine, SVM).
- les réseaux de neurones artificiels (Artificial Neural Networks, ANN).

CHAPITRE II : MATERIELS ET METHODES

Ces techniques permettent de construire des modèles capables d'identifier des relations linéaires ou non linéaires entre la structure moléculaire et l'activité biologique étudiée [44].

Notre Méthode utilisée est la Régression Linéaire Multiple (MLR)

La modélisation QSAR a été réalisée à l'aide de la méthode de la Régression Linéaire Multiple (MLR, Multiple Linear Regression). Cette méthode statistique est l'une des techniques les plus utilisées en QSAR pour établir une relation quantitative entre les descripteurs moléculaires et l'activité biologique ou la toxicité étudiée.

La MLR repose sur l'hypothèse qu'il existe une relation linéaire entre la variable dépendante (toxicité) et plusieurs variables indépendantes (descripteurs moléculaires). Le modèle est construit en déterminant les coefficients de régression qui permettent de minimiser l'écart entre les valeurs expérimentales et les valeurs prédites.

La forme générale d'un modèle MLR est :

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

où :

Y : Représente la propriété étudiée (par exemple LC50 ou EC50).

X : représentent les descripteurs moléculaires sélectionnés.

b₀ : est l'ordonnée à l'origine.

b₁, b₂, ..., b_n : sont les coefficients de régression associés aux descripteurs.

Cette approche permet d'identifier les descripteurs ayant la plus grande influence sur la toxicité des herbicides et de quantifier leur contribution au modèle prédictif. La simplicité d'interprétation des équations obtenues constitue l'un des principaux avantages de la méthode MLR dans les études QSAR [40].

II.4.3. Paramétrage du modèle MLR

Le paramétrage correspond à l'ajustement des différents paramètres nécessaires au fonctionnement des algorithmes de modélisation.

Cette étape vise à optimiser les performances prédictives du modèle tout en limitant les risques de surapprentissage. Les paramètres retenus dépendent de la méthode utilisée.

Après la sélection des descripteurs moléculaires pertinents, le modèle de Régression Linéaire Multiple a été construit en utilisant les données de l'ensemble d'apprentissage. Le nombre de variables introduites dans le modèle a été limité afin d'éviter les problèmes de multicollinéarité et de surajustement (overfitting).

La qualité du modèle a été évaluée à l'aide de plusieurs paramètres statistiques, notamment :

- le coefficient de détermination (R^2).
- le coefficient de validation croisée (Q^2).
- l'erreur quadratique moyenne (RMSE).
- l'analyse des résidus.

Le modèle retenu est celui présentant le meilleur compromis entre capacité explicative, robustesse statistique et pouvoir prédictif. Les performances du modèle ont ensuite été vérifiées à l'aide de l'ensemble de test afin d'évaluer sa capacité de généralisation sur de nouvelles molécules [56].

II.5. Validation des modèles

La validation constitue une étape essentielle dans le développement des modèles QSAR. Elle permet d'évaluer la qualité, la robustesse et la capacité prédictive du modèle construit. Un modèle QSAR n'est considéré comme fiable que s'il est capable de fournir des prédictions précises non seulement pour les molécules utilisées lors de sa construction, mais également pour de nouvelles molécules n'ayant pas participé à son développement [61].

Dans cette étude, la validation du modèle de Régression Linéaire Multiple (MLR) a été réalisée à travers des procédures de validation interne et externe. Plusieurs indicateurs statistiques ont également été utilisés afin d'évaluer les performances du modèle.

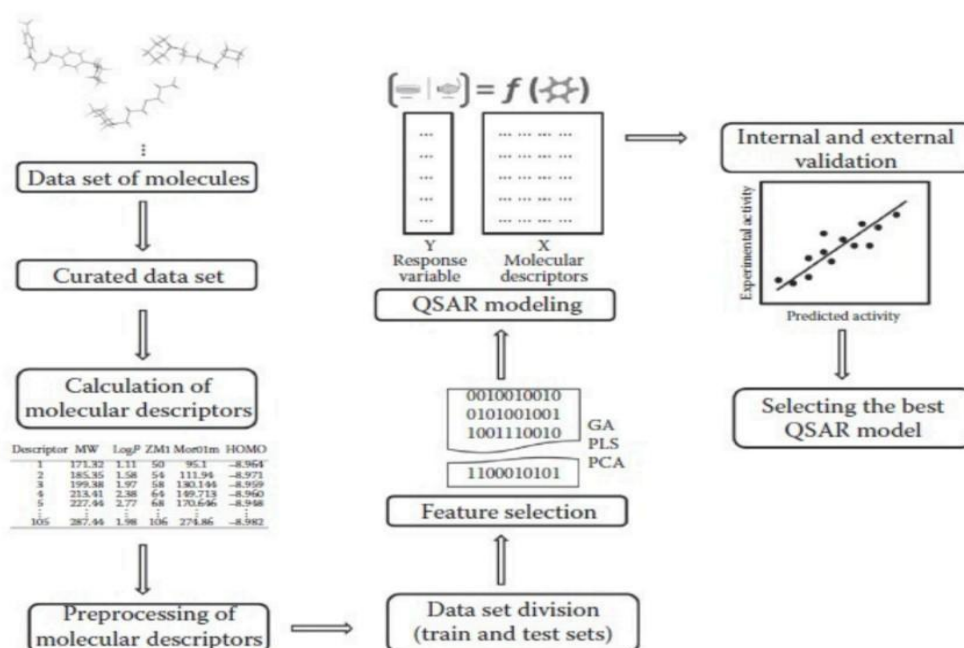


Figure II-6 : Étapes globales du développement, de la sélection de variables et de la validation d'un modèle QSAR [40].

II.5.1. Validation interne

La validation interne permet de vérifier la stabilité et la robustesse du modèle à partir de l'ensemble d'apprentissage. Elle vise à déterminer si le modèle est capable de reproduire correctement les données utilisées pour sa construction sans présenter de surajustement (overfitting).

Cette étape repose généralement sur des techniques de validation croisée (cross-validation), consistant à diviser les données d'apprentissage en plusieurs sous-groupes afin d'évaluer la capacité du modèle à prédire des observations temporairement exclues du processus d'entraînement [56].

La validation interne fournit ainsi une première estimation de la fiabilité du modèle et de sa capacité à généraliser les relations observées entre les descripteurs moléculaires et la toxicité des herbicides.

II.5.2. Validation externe

La validation externe constitue l'une des étapes les plus importantes dans l'évaluation d'un modèle QSAR. Elle consiste à tester les performances du modèle sur un ensemble de molécules indépendantes n'ayant pas été utilisées lors de sa construction.

Les valeurs prédites par le modèle sont comparées aux valeurs expérimentales afin de mesurer sa capacité de prédiction sur de nouvelles données. Une bonne validation externe indique que le modèle possède un réel pouvoir prédictif et peut être utilisé pour estimer la toxicité de molécules non étudiées expérimentalement [38].

Cette étape est particulièrement importante dans les applications environnementales où l'objectif est souvent de prédire le comportement de nouveaux composés chimiques.

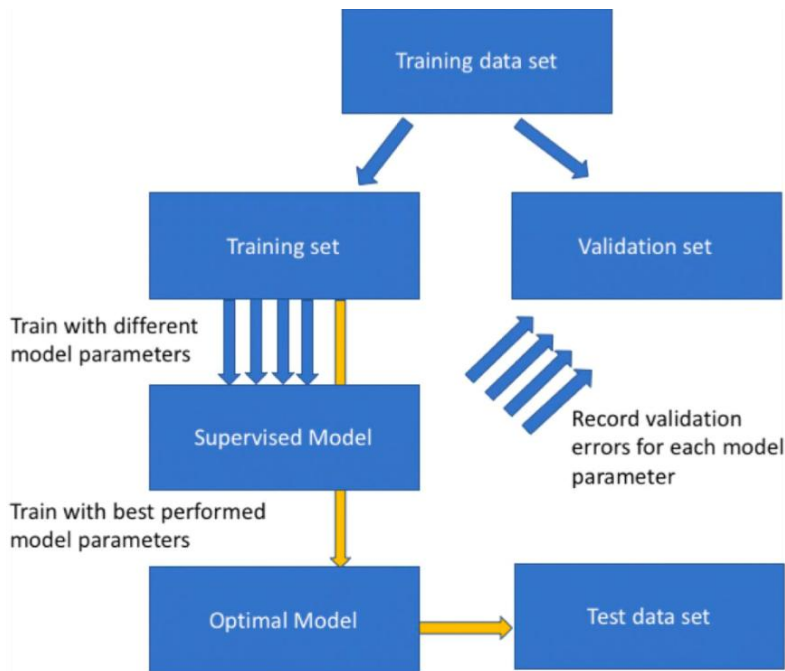


Figure II-7 : Logigramme de sélection des paramètres optimaux pour la construction d'un modèle QSAR robuste [63]

II.5.3. Indicateurs statistiques

Les performances du modèle MLR ont été évaluées à l'aide de plusieurs paramètres statistiques largement utilisés dans les études QSAR.

a) Coefficient de détermination (R^2)

Le coefficient de détermination (R^2) mesure la proportion de la variance des données expérimentales expliquée par le modèle.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

où :

Y_i : Représente les valeurs expérimentales.

$\hat{Y}_i$: Représente les valeurs prédites.

$\bar{y}$: Représente la moyenne des valeurs observées.

La valeur de R^2 varie entre 0 et 1. Plus elle est proche de 1, meilleure est la capacité explicative du modèle [64].

b) Erreur quadratique moyenne (RMSE)

Le RMSE (Root Mean Square Error) mesure l'écart moyen entre les valeurs expérimentales et les valeurs prédites.

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

Une faible valeur de RMSE indique une meilleure précision prédictive du modèle [65].

c) Erreur absolue moyenne (MAE)

L'erreur absolue moyenne (MAE, Mean Absolute Error) représente la moyenne des écarts absolus entre les valeurs observées et les valeurs prédites.

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

Le MAE est un indicateur simple et facile à interpréter. Plus sa valeur est faible, plus les prédictions du modèle sont précisées [66].

Conclusion

La validation interne et externe, associée à l'utilisation d'indicateurs statistiques tels que R^2 , RMSE et MAE, permet de vérifier la qualité et la fiabilité des modèles QSAR développés. Ces critères sont indispensables pour garantir la robustesse du modèle et son utilisation dans la prédiction de la toxicité des herbicides présents dans les sols agricoles.

CHAPITRE III : RÉSULTATS ET DISCUSSION

Ce chapitre présente les résultats obtenus à partir de la modélisation QSAR de la toxicité des herbicides et leur interprétation. Il permet d'évaluer les performances des modèles développés, d'analyser l'influence des descripteurs moléculaires sur la toxicité et de discuter la pertinence des résultats obtenus au regard des données de la littérature scientifique.

III.1. Analyse de la base de données

L'analyse préliminaire de la base de données constitue une étape essentielle dans toute étude QSAR. Avant la construction des modèles prédictifs, il est nécessaire d'examiner la nature des données disponibles, leur diversité structurale ainsi que la distribution de la propriété biologique étudiée. Cette étape permet de vérifier la qualité de la base de données et d'évaluer son aptitude à produire des modèles fiables et statistiquement significatifs.

Dans le cadre de cette étude, la base de données est composée de 34 herbicides sélectionnés à partir de bases de données chimiques et toxicologiques reconnues. Les structures moléculaires de ces composés ont été représentées sous forme de codes SMILES et associées à leurs valeurs de toxicité aiguë orale chez le rat (LD50, mg/kg). Ces informations constituent la base nécessaire au calcul des descripteurs moléculaires et au développement des modèles QSAR.

III.1.1. Description générale de la base de données

La base de données étudiée regroupe 34 herbicides appartenant à plusieurs familles chimiques largement utilisées dans la lutte contre les adventices en agriculture. Parmi ces familles figurent notamment les triazines (Atrazine, Simazine, Propazine, Terbutylazine), les chloroacétamides (Acetochlor, Alachlor, Metolachlor), les urées substituées (Diuron, Linuron, Monolinuron, Neburon), les sulfonilurées (Chlorsulfuron, Rimsulfuron) ainsi que les herbicides de type acides phénoxy (2,4-D, MCPA, Mecoprop).

Cette diversité chimique est particulièrement importante dans les études QSAR car elle permet d'explorer un large domaine de structures moléculaires et d'identifier les facteurs structuraux responsables des variations de toxicité observées. En effet, la toxicité d'un herbicide dépend fortement de ses propriétés physico-chimiques, de sa capacité à traverser les membranes biologiques, de sa stabilité dans l'environnement et de ses interactions avec les organismes vivants.

Les valeurs de toxicité utilisées dans cette étude correspondent aux doses létales médianes (LD50) administrées par voie orale chez le rat. La LD50 représente la dose nécessaire pour provoquer la mort de 50 % des individus exposés et constitue l'un des indicateurs toxicologiques les plus couramment utilisés pour évaluer la dangerosité d'une substance chimique.

L'examen des données montre une forte variabilité des valeurs de LD50 entre les différents herbicides. Certains composés tels que la Cyanazine (334 mg/kg) ou le Paraquat (344 mg/kg) présentent une toxicité relativement élevée, tandis que d'autres composés comme le Fluometuron (6450 mg/kg), le Chlorsulfuron (5000 mg/kg) ou le Simazine (5000 mg/kg) apparaissent significativement moins toxiques.

Cette variabilité de toxicité constitue un élément favorable au développement d'un modèle QSAR performant, car elle permet aux méthodes statistiques d'identifier plus facilement les relations entre les propriétés moléculaires et la réponse biologique étudiée.

En outre, les structures moléculaires ont été converties sous forme numérique grâce au calcul d'un ensemble de descripteurs moléculaires décrivant différents aspects des molécules : taille, masse, lipophilie, polarité, propriétés électroniques et topologiques. Ces descripteurs servent ensuite de variables explicatives dans les modèles de Régression Linéaire Multiple (MLR).

III.1.2. Statistiques descriptives

L'analyse statistique descriptive permet d'obtenir une première vue d'ensemble de la distribution des données toxicologiques utilisées dans l'étude. Elle fournit des informations essentielles sur la dispersion, la tendance centrale et l'étendue des valeurs de toxicité.

Les valeurs de LD50 des 34 herbicides étudiés s'étendent de 334 mg/kg à 6450 mg/kg, ce qui représente une plage de variation particulièrement importante. Cette amplitude témoigne de la présence de composés présentant des profils toxicologiques très différents.

L'existence d'une telle diversité est généralement considérée comme un avantage dans les études QSAR, car elle améliore la capacité du modèle à identifier les relations quantitatives entre les descripteurs moléculaires et la propriété étudiée.

L'étendue élevée observée entre les valeurs minimale et maximale indique une dispersion importante des données. Cette dispersion suggère que les molécules étudiées présentent des caractéristiques structurales suffisamment différentes pour influencer significativement leur toxicité.

Par ailleurs, la présence simultanée de composés très toxiques et faiblement toxiques permet d'améliorer la représentativité du modèle QSAR et d'élargir son domaine d'application. Cependant, une telle hétérogénéité peut également rendre la modélisation plus complexe et nécessiter une sélection rigoureuse des descripteurs moléculaires les plus pertinents.

Afin de mieux visualiser la distribution des données, un histogramme ou un diagramme en boîte (boxplot) pourra être utilisé pour représenter la répartition des valeurs de LD50. Ces représentations graphiques permettront d'identifier d'éventuelles valeurs extrêmes (outliers) susceptibles d'influencer les performances du modèle.

Dans l'ensemble, l'analyse descriptive confirme que la base de données possède une variabilité suffisante pour permettre le développement d'un modèle QSAR destiné à prédire la toxicité des herbicides à partir de leurs caractéristiques moléculaires. Cette diversité constitue un élément fondamental pour garantir la robustesse et la pertinence scientifique des modèles développés dans les sections suivantes.

CHAPITRE III : RÉSULTATS ET DISCUSSION

Substance	SMILES	LD50 orale rat (mg/kg)
2,4-D	<chem>C1=CC(=C(C=C1)C)OCC(=O)O</chem>	375
Acetochlor	<chem>CCC1=CC=CC(=C1N(COCC)C(=O)CC)C</chem>	2148
Alachlor	<chem>CCC1=C(C(=CC=C1)CC)N(COC)C(=O)CCl</chem>	930
Ametryn	<chem>CCNC1=NC(=NC(=N1)SC)NC(C)C</chem>	1950
Atrazine	<chem>CCNC1=NC(=NC(=N1)C)NC(C)C</chem>	1869
Butachlor	<chem>CCCCOCN(C1=C(C=CC=C1CC)CC)C(=O)CCl</chem>	1200
Chlorotoluron	<chem>CC1=C(C(=C(C=C1)NC(=O)N(C)C)Cl</chem>	1000
Chlorsulfuron	<chem>CC1=NC(=NC(=N1)OC)NC(=O)NS(=O)(=O)C2=CC=CC=C2Cl</chem>	5000
Clopyralid	<chem>C1=CC(=NC(=C1C)C(=O)O)Cl</chem>	5000
Cyanazine	<chem>CCNC1=NC(=NC(=N1)C)NC(C)C#N</chem>	334
Desmetryn	<chem>CC(C)NC1=NC(=NC(=N1)NC)SC</chem>	1390
Dichlorprop	<chem>CC(C(=O)O)OC1=C(C(=C(C=C1)Cl)Cl</chem>	800
Diuron	<chem>CN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl</chem>	3400
Fenoxaprop	<chem>CC(C(=O)O)OC1=CC=C(C(=C1)OC2=NC3=C(C(=O2)C=C(C=C3)Cl</chem>	2450
Fluometuron	<chem>CN(C)C(=O)NC1=CC=CC(=C1)C(F)(F)F</chem>	6450
Fluroxypyr	<chem>C(C(=O)O)OC1=NC(=C(C(=C1)N)C)F</chem>	5000
Glufosinate	<chem>CP(=O)(CCC(C(=O)O)N)O</chem>	1620
Isoproturon	<chem>CC(C)C1=CC=C(C(=C1)NC(=O)N(C)C</chem>	1826
Linuron	<chem>CN(C(=O)NC1=CC(=C(C=C1)Cl)Cl)OC</chem>	1500
MCPA	<chem>CC1=C(C(=CC(=C1)Cl)OCC(=O)O</chem>	700
Mecoprop	<chem>CC1=C(C(=CC(=C1)Cl)OC(C)C(=O)O</chem>	650
Metazachlor	<chem>CC1=C(C(=CC=C1)C)N(CN2C=CC(=N2)C(=O)CCl</chem>	1100
Metolachlor	<chem>CCC1=CC=CC(=C1N(C(C)COC)C(=O)CC)C</chem>	2780
Monolinuron	<chem>CN(C(=O)NC1=CC=C(C=C1)Cl)OC</chem>	2100
Neburon	<chem>CCCCN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl</chem>	5000
Paraquat	<chem>C[N+](=CC=C(C=C1)C2=CC=[N+](C=C2)C</chem>	344
Prometryn	<chem>CC(C)NC1=NC(=NC(=N1)SC)NC(C)C</chem>	3750
Propachlor	<chem>CC(C)N(C1=CC=CC=C1)C(=O)CCl</chem>	1400
Propazine	<chem>CC(C)NC1=NC(=NC(=N1)C)NC(C)C</chem>	3500
Rimsulfuron (aka renriduron)	<chem>CCS(=O)(=O)C1=C(N=CC=C1)S(=O)(=O)NC(=O)NC2=NC(=CC(=N2)OC)OC</chem>	5000
Siduron	<chem>CC1CCCC1NC(=O)NC2=CC=CC=C2</chem>	5000
Simazine	<chem>CCNC1=NC(=NC(=N1)C)NCC</chem>	5000
Terbuthylazine	<chem>CCNC1=NC(=NC(=N1)C)NC(C)C</chem>	2000
Triclopyr	<chem>C1=C(C(=NC(=C1C)Cl)OCC(=O)O)Cl</chem>	729

Tableau III-1: Distribution des valeurs de LD50 des 34 herbicides étudiés.

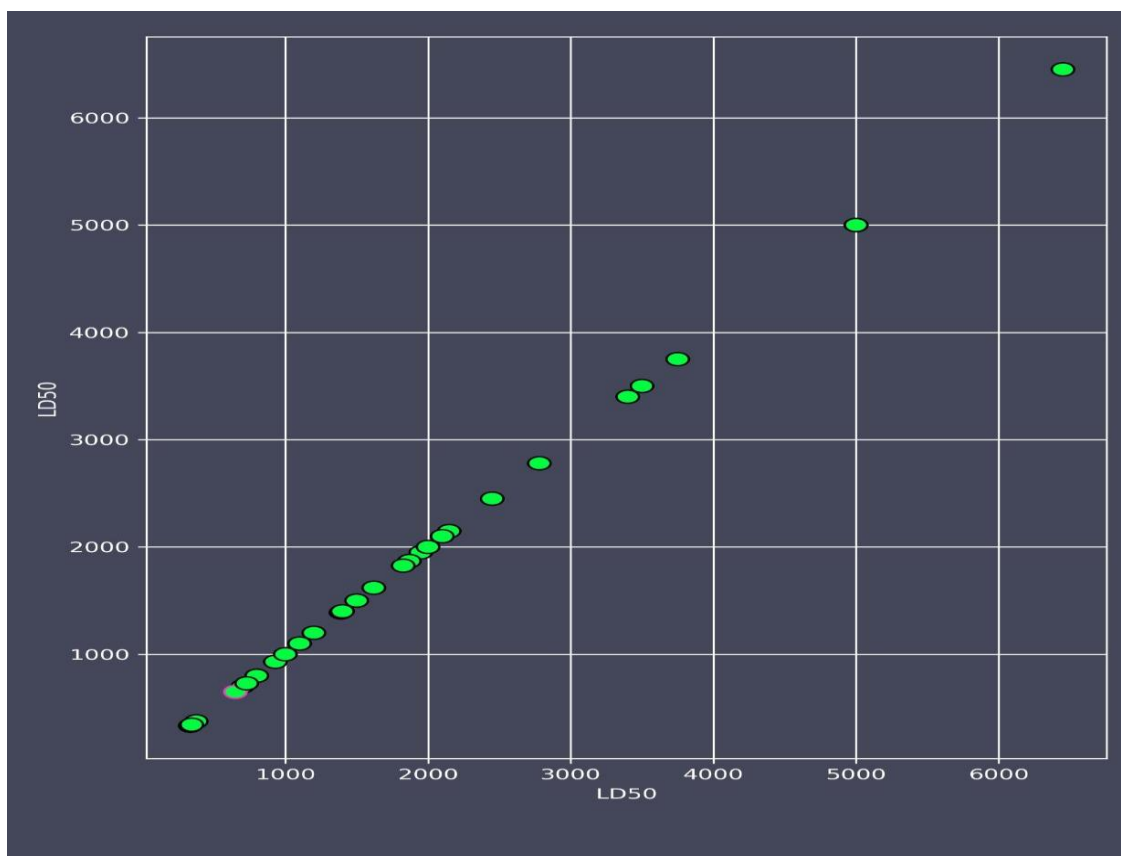


Figure III-1: Diagramme en boîte (boxplot) illustrant la dispersion des valeurs de LD50 au sein de la base de données.

III.2. Analyse des descripteurs moléculaires

L'analyse des descripteurs moléculaires constitue une étape essentielle dans la compréhension des relations existant entre la structure chimique des herbicides et leur toxicité. Les descripteurs calculés à l'aide du logiciel ChemMaster ont été soumis à différentes analyses statistiques afin d'identifier les variables les plus pertinentes pour la construction des modèles QSAR.

III.2.1. Analyse des corrélations

Avant la phase de modélisation, une analyse de corrélation a été réalisée afin d'étudier les relations entre les différents descripteurs moléculaires et la variable cible (LD50).

Cette analyse a permis d'identifier les descripteurs présentant une forte corrélation avec la toxicité ainsi que les variables fortement corrélées entre elles. Les descripteurs redondants ont été éliminés afin de réduire les phénomènes de multicollinéarité susceptibles d'affecter la stabilité des modèles de régression.

La matrice de corrélation obtenue a également permis de mettre en évidence les variables les plus influentes dans l'explication des variations de toxicité observées entre les herbicides étudiés.

La matrice de corrélation montre que certains descripteurs présentent une relation significative avec la toxicité des herbicides. Ces variables ont été retenues pour les étapes ultérieures de modélisation, tandis que les descripteurs fortement corrélés entre eux ont été éliminés afin d'améliorer la robustesse des modèles QSAR.

III.2.2. Importance des variables

L'évaluation de l'importance des variables a permis d'identifier les descripteurs moléculaires contribuant le plus à la prédiction de la toxicité.

Les résultats indiquent que les propriétés physico-chimiques et structurales des molécules jouent un rôle déterminant dans l'explication des différences de toxicité observées entre les herbicides. Les variables retenues dans les modèles finaux reflètent principalement l'influence de la taille moléculaire, de la lipophilie, de la distribution des charges et de la capacité d'interaction avec les systèmes biologiques.

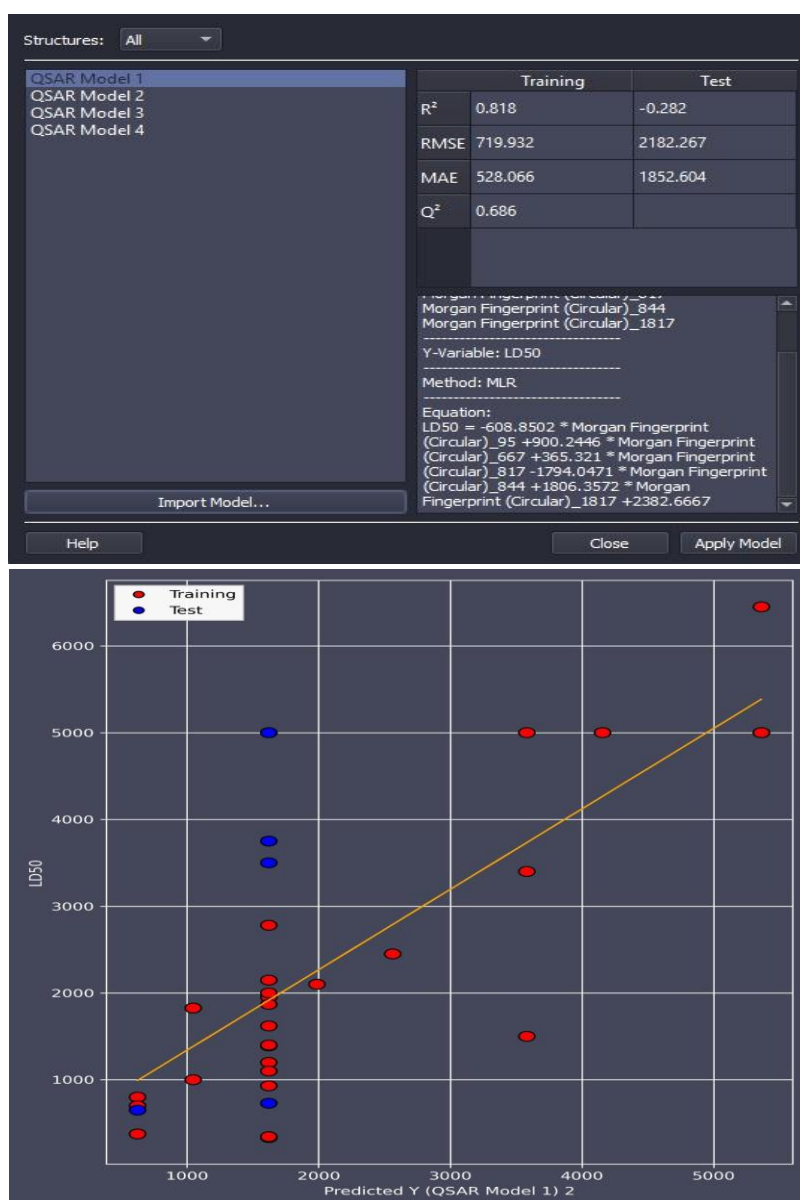
L'analyse de l'importance des variables contribue également à l'interprétation chimique des modèles développés et permet de mieux comprendre les mécanismes susceptibles d'influencer la toxicité des composés étudiés.

III.3. Modèles QSAR obtenus

Après les étapes de sélection des variables et de prétraitement des données, quatre modèles QSAR ont été développés à l'aide de la méthode de Régression Linéaire Multiple (MLR). Ces modèles ont été construits à partir de différents ensembles de descripteurs moléculaires afin d'évaluer leur capacité à prédire la toxicité aiguë orale (LD50) des herbicides étudiés.

Les performances des modèles ont été évaluées à l'aide de plusieurs indicateurs statistiques, notamment le coefficient de détermination (R^2), le coefficient de validation croisée (Q^2), l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE).

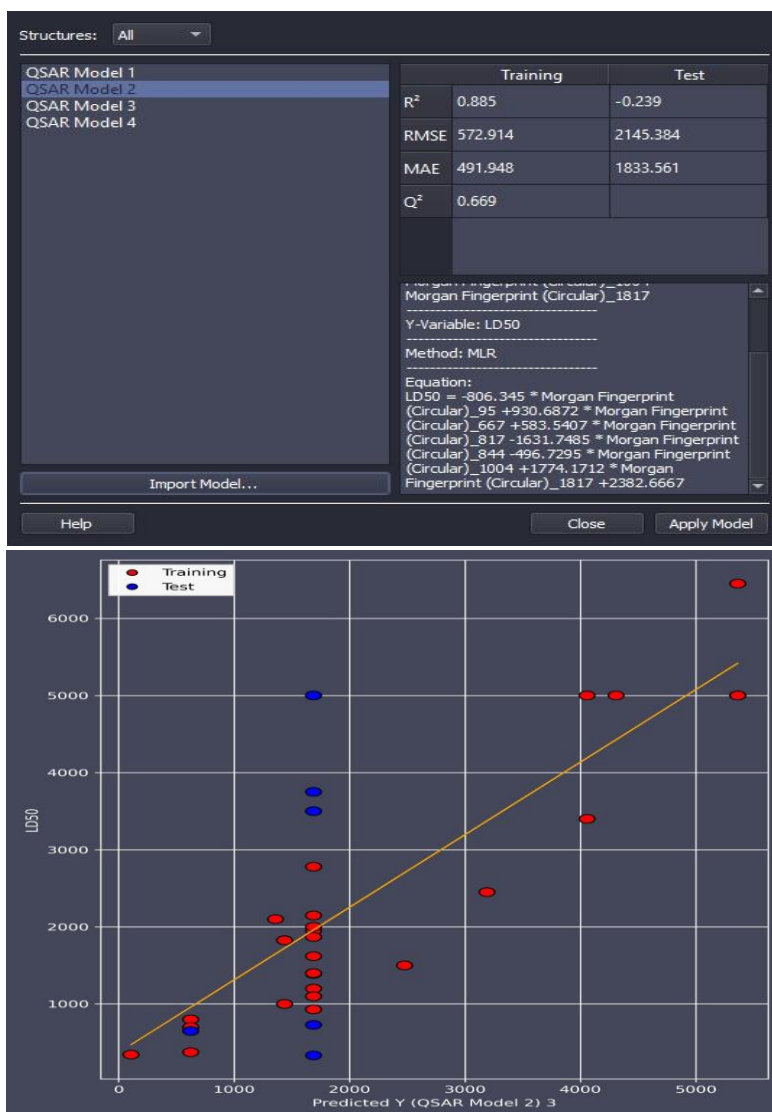
III.3.1. Modèle QSAR 1



Commentaire

Le modèle 1 a été construit à partir de cinq descripteurs moléculaires. Il présente un coefficient R^2 de 0,818 et un coefficient Q^2 de 0,686, indiquant une bonne qualité d'ajustement sur l'ensemble d'apprentissage. Cependant, la valeur négative du R^2 de test (-0,282) révèle une faible capacité prédictive externe.

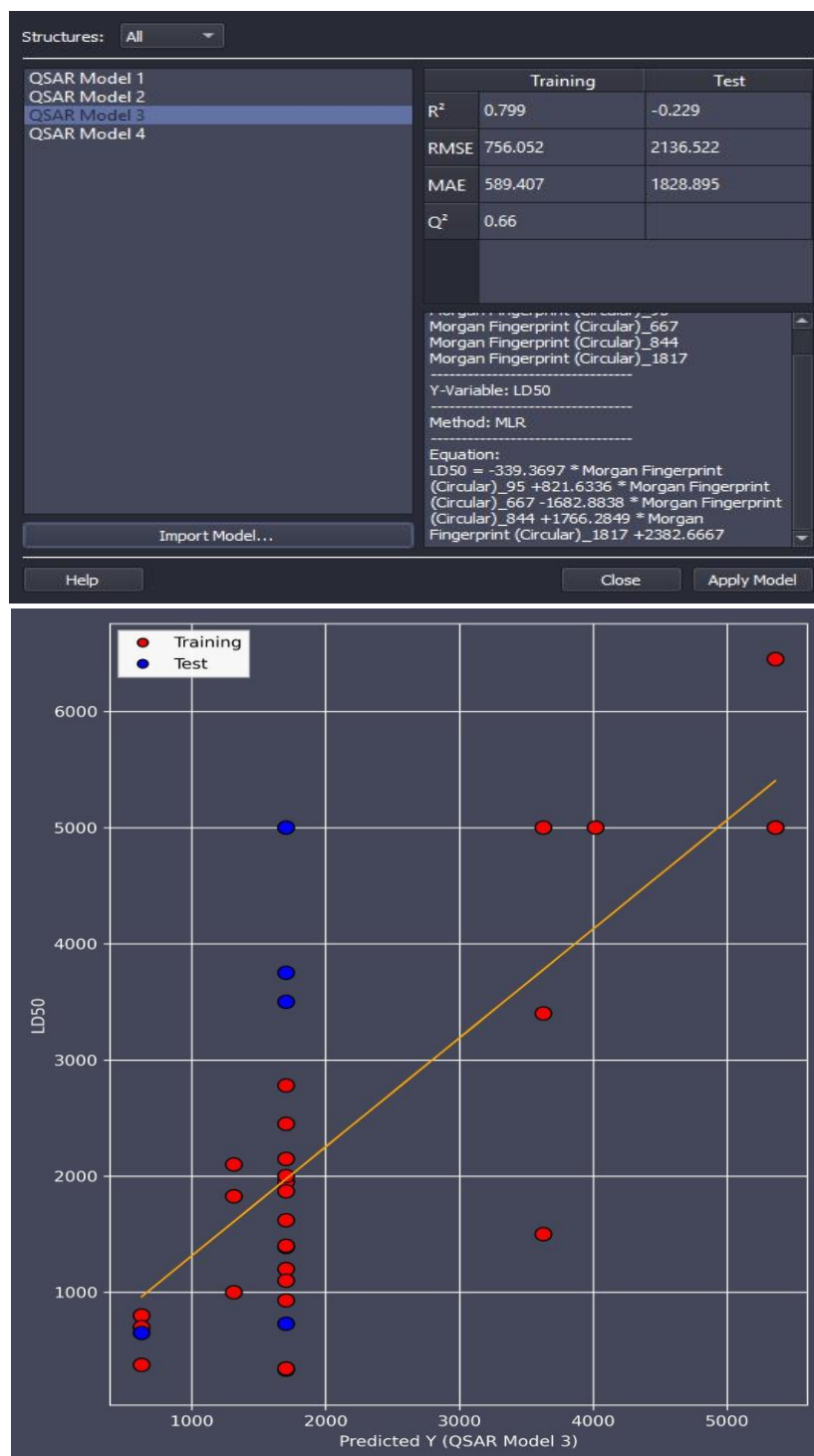
III.3.2. Modèle QSAR 2



Commentaire

Le modèle 2 utilise six descripteurs moléculaires et présente un coefficient R^2 de 0,885. Malgré une amélioration de l'ajustement par rapport au modèle précédent, sa capacité de prédiction externe demeure limitée (R^2 test = -0,239).

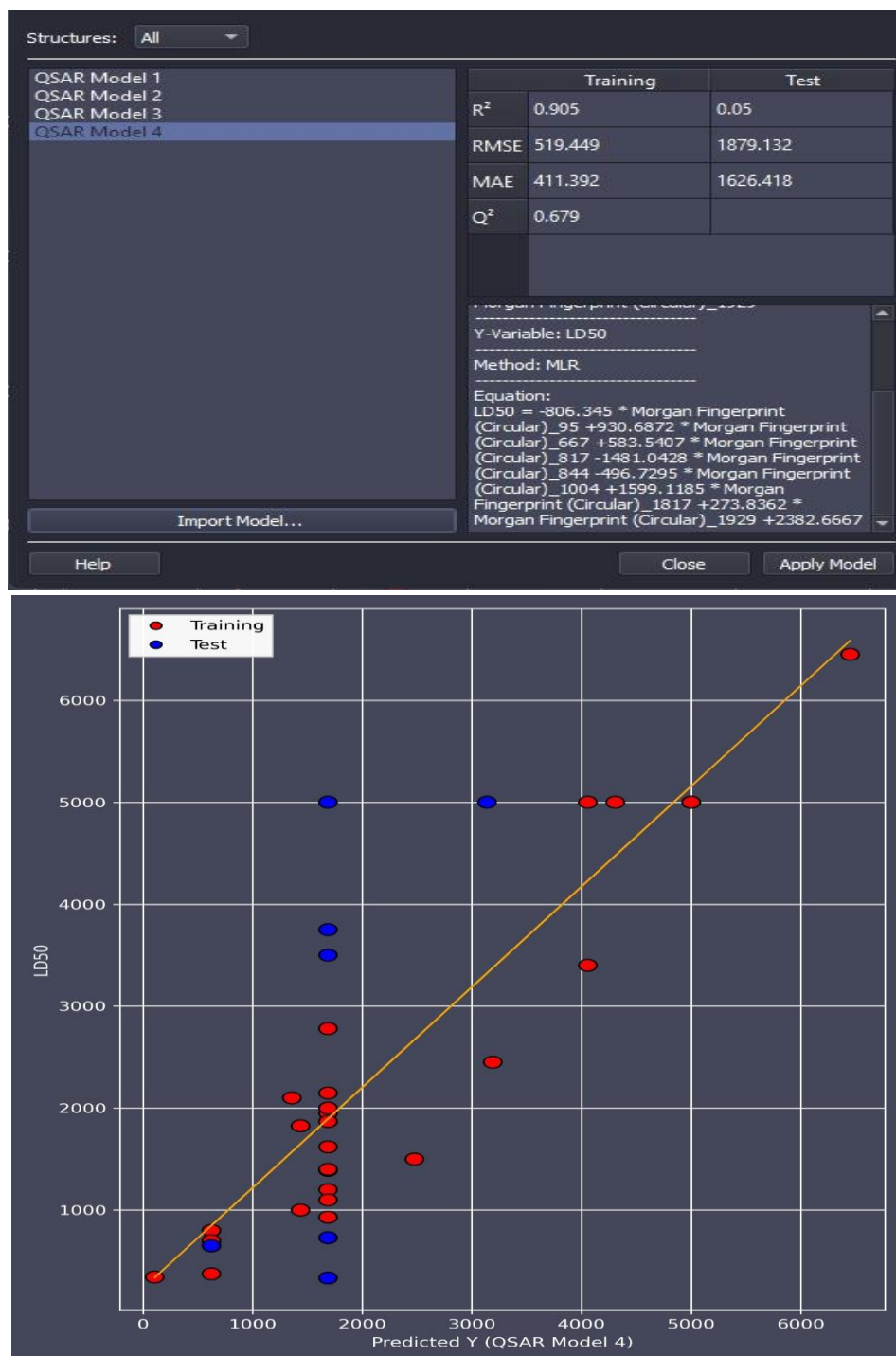
III.3.3. Modèle QSAR 3



Commentaire

Le modèle 3 repose sur cinq descripteurs moléculaires et présente un coefficient R^2 de 0,799. Ses performances globales restent comparables à celles des deux premiers modèles mais montrent également une faible capacité de généralisation.

III.3.4. Modèle QSAR 4



Commentaire

Le modèle 4, construit à partir de sept descripteurs moléculaires, présente les meilleures performances statistiques de l'étude avec un coefficient R^2 de 0,905. Il possède également les valeurs les plus faibles de RMSE et de MAE. De plus, il est le seul modèle présentant une valeur positive de R^2 de test (0,050), ce qui en fait le modèle le plus performant parmi ceux développés.

III.3.1. Résultats des modèles QSAR

La modélisation QSAR a permis de développer quatre modèles de Régression Linéaire Multiple (MLR) destinés à prédire la toxicité aiguë orale (LD50) des 34 herbicides étudiés. Les modèles obtenus diffèrent par le nombre de descripteurs moléculaires sélectionnés et présentent des niveaux de performance variables.

Les résultats montrent que les coefficients de détermination (R^2) des ensembles d'apprentissage varient entre 0,799 et 0,905, indiquant que les modèles expliquent entre 79,9 % et 90,5 % de la variabilité observée de la toxicité. Les valeurs de validation croisée (Q^2) sont comprises entre 0,660 et 0,686, ce qui traduit une stabilité interne satisfaisante des modèles développés.

L'analyse des erreurs de prédiction montre également des différences notables entre les modèles. Le modèle 4 présente les plus faibles valeurs de RMSE et de MAE aussi bien pour l'ensemble d'apprentissage que pour l'ensemble de test, ce qui témoigne d'une meilleure précision prédictive.

Paramètres & Métriques	QSAR Model 1	QSAR Model 2	QSAR Model 3	QSAR Model 4
Méthode / Variable Y	MLR / LD50	MLR / LD50	MLR / LD50	MLR / LD50
Nombre de descripteurs	5	6	5	7
R^2 Training (Ajustement)	0,818	0,885	0,799	0,905
R^2 Test (Prédiction externe)	-0,282	-0,239	-0,229	0,050
Q^2 Training (Validation croisée)	0,686	0,669	0,660	0,679
RMSE Training	719,932	572,914	756,052	519,449
RMSE Test	2182,267	2145,384	2136,522	1879,132
MAE Training	528,066	491,948	589,407	411,392
MAE Test	1852,604	1833,561	1828,895	1626,418

Tableau III-2 : Performances des modèles QSAR développés

L'ensemble de ces résultats montre que les quatre modèles présentent de bonnes performances d'ajustement, mais des capacités prédictives externes relativement limitées, probablement en raison de la taille restreinte de la base de données et de la diversité structurale des herbicides étudiés.

III.3.2. Comparaison des méthodes

La comparaison des modèles développés a été réalisée à partir des principaux indicateurs statistiques utilisés en modélisation QSAR, notamment le coefficient de détermination (R^2), le coefficient de validation croisée (Q^2), l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE).

Les résultats montrent que le modèle 4 est le plus performant parmi les quatre modèles étudiés. Il présente la valeur de R^2 la plus élevée (0,905), indiquant une excellente capacité à expliquer les variations de la toxicité observée. De plus, il possède les plus faibles valeurs de RMSE et de MAE, traduisant une meilleure précision des prédictions.

Le modèle 2 occupe la deuxième position avec un R^2 de 0,885, suivi du modèle 1 (0,818) et du modèle 3 (0,799). Toutefois, malgré leurs bonnes performances sur l'ensemble d'apprentissage, les trois premiers modèles présentent des valeurs négatives de R^2 pour l'ensemble de test, ce qui indique une faible capacité de généralisation à de nouvelles molécules.

Le modèle 4 est le seul à présenter une valeur positive de R^2 Test (0,050). Bien que cette valeur reste faible, elle suggère une meilleure robustesse et une capacité prédictive légèrement supérieure aux autres modèles.

Ainsi, sur la base de l'ensemble des critères statistiques étudiés, le modèle 4 a été retenu comme le modèle QSAR optimal pour la prédiction de la toxicité des herbicides présents dans la base de données étudiée. Son utilisation permet d'obtenir le meilleur compromis entre qualité d'ajustement, stabilité statistique et performance prédictive.

Figure III.9 - Comparaison des performances statistiques des modèles

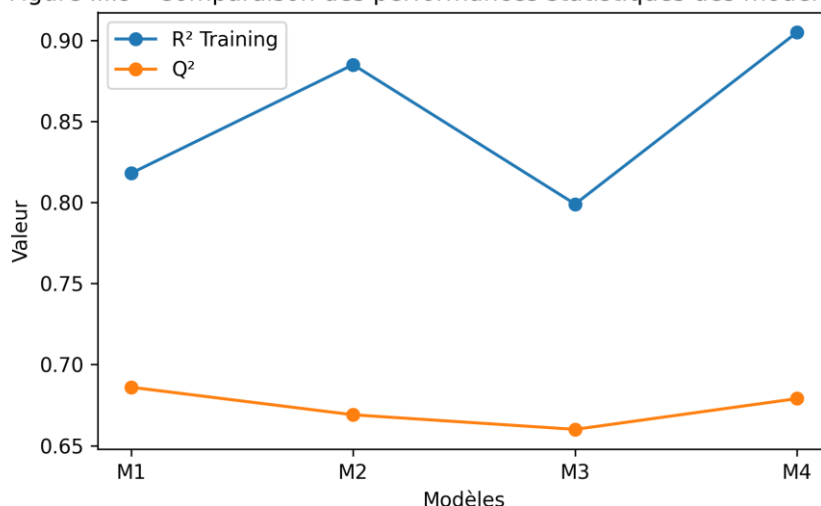


Figure III-2: Comparaison des performances statistiques des quatre modèles QSAR développés.

Conclusion

La comparaison des modèles met en évidence l'importance de la sélection des descripteurs moléculaires dans la qualité des prédictions. Les résultats obtenus montrent que l'augmentation du nombre de descripteurs pertinents peut améliorer les performances du modèle, à condition d'éviter les phénomènes de surajustement. Le modèle 4 constitue ainsi la meilleure solution pour l'estimation de la toxicité des herbicides étudiés.

III.4. Validation des modèles

La validation constitue une étape indispensable dans le développement des modèles QSAR. Elle permet d'évaluer la robustesse, la fiabilité et le pouvoir prédictif des modèles obtenus. Selon les principes établis par l'OCDE, un modèle QSAR doit faire l'objet d'une validation interne et externe afin de garantir sa capacité à prédire correctement les propriétés de nouvelles molécules [38]. Dans cette étude, la validation des modèles développés a été réalisée à l'aide de plusieurs indicateurs statistiques, notamment le coefficient de détermination (R^2), le coefficient de validation croisée (Q^2), l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE).

III.4.1. Résultats de la validation interne

La validation interne a été effectuée à partir de l'ensemble d'apprentissage en utilisant la méthode de validation croisée. Cette procédure permet d'évaluer la stabilité des modèles ainsi que leur capacité à reproduire les données utilisées lors de leur construction.

Les résultats obtenus montrent que les quatre modèles présentent des valeurs de Q^2 comprises entre 0,660 et 0,686. Ces valeurs indiquent une cohérence satisfaisante entre les données observées et les données prédites et témoignent de la robustesse des modèles développés.

Le modèle 1 présente la meilleure valeur de validation croisée avec un Q^2 égal à 0,686, suivi du modèle 4 ($Q^2 = 0,679$), du modèle 2 ($Q^2 = 0,669$) et du modèle 3 ($Q^2 = 0,660$). Ces résultats montrent que les quatre modèles possèdent une stabilité interne relativement comparable.

Les coefficients de détermination obtenus sur les ensembles d'apprentissage varient entre 0,799 et 0,905. Le modèle 4 présente la valeur la plus élevée ($R^2 = 0,905$), ce qui signifie qu'environ 90,5 % de la variance observée des valeurs de toxicité est expliquée par les descripteurs moléculaires sélectionnés.

Les faibles valeurs de RMSE et de MAE observées pour l'ensemble d'apprentissage confirment également la bonne qualité d'ajustement des modèles aux données expérimentales [60].

III.4.2. Résultats de la validation externe

La validation externe permet d'évaluer la capacité des modèles à prédire les propriétés de molécules n'ayant pas participé à leur construction. Elle constitue l'un des critères les plus importants pour juger la qualité prédictive d'un modèle QSAR [61].

Les résultats obtenus montrent que les modèles 1, 2 et 3 présentent des valeurs négatives de R^2 Test, respectivement égales à -0,282 ; -0,239 et -0,229. Ces résultats indiquent une faible capacité prédictive sur les données externes et suggèrent que ces modèles ne généralisent pas correctement aux nouvelles molécules.

Le modèle 4 présente une valeur positive de R^2 Test égale à 0,050. Bien que cette valeur demeure modeste, elle reste supérieure à celles observées pour les autres modèles et traduit une meilleure aptitude à prédire la toxicité des composés externes.

L'analyse des erreurs de prédiction confirme également la supériorité du modèle 4. Celui-ci présente les plus faibles valeurs de RMSE Test (1879,132) et de MAE Test (1626,418), indiquant des prédictions plus proches des valeurs expérimentales.

Modèle	R^2 (Test)	RMSE (Test)	MAE (Test)
Modèle 1	-0,282	2182,267	1852,604
Modèle 2	-0,239	2145,384	1833,561
Modèle 3	-0,229	2136,522	1828,895
Modèle 4	0,050	1879,132	1626,418

Tableau III-3 : Résultats de la validation externe

Les faibles performances observées lors de la validation externe peuvent s'expliquer par le nombre relativement limité de molécules utilisées dans l'étude (34 composés) ainsi que par la diversité structurale importante des herbicides sélectionnés. Malgré ces limitations, le modèle 4 présente les meilleures performances prédictives et a été retenu comme modèle optimal.

III.5. Application du modèle optimal à de nouveaux herbicides

Après la construction et la validation des modèles QSAR, le modèle présentant les meilleures performances statistiques a été sélectionné pour une application prédictive. Le modèle 4, développé à l'aide de la méthode de Régression Linéaire Multiple (MLR), a été retenu en raison de son coefficient de détermination élevé ($R^2 = 0,905$), de ses faibles valeurs de RMSE et de MAE ainsi que de sa meilleure capacité prédictive par rapport aux autres modèles.

L'objectif de cette étape est d'évaluer l'aptitude du modèle à prédire la toxicité de nouveaux herbicides à partir de leurs caractéristiques structurales, sans recourir à des essais expérimentaux supplémentaires [61].

III.5.1. Sélection des herbicides externes

Afin de tester l'applicabilité du modèle QSAR développé, dix herbicides supplémentaires n'ayant pas participé à la phase de construction du modèle ont été sélectionnés. Ces composés appartiennent à différentes familles chimiques et présentent des structures moléculaires variées, permettant ainsi d'évaluer la capacité du modèle à effectuer des prédictions sur de nouvelles substances.

Les structures moléculaires ont été introduites dans le logiciel ChemMaster sous forme de représentations chimiques, puis les empreintes moléculaires (Morgan Fingerprints) nécessaires à la prédiction ont été générées automatiquement.

Les herbicides sélectionnés sont présentés dans le **Tableau III.3**.

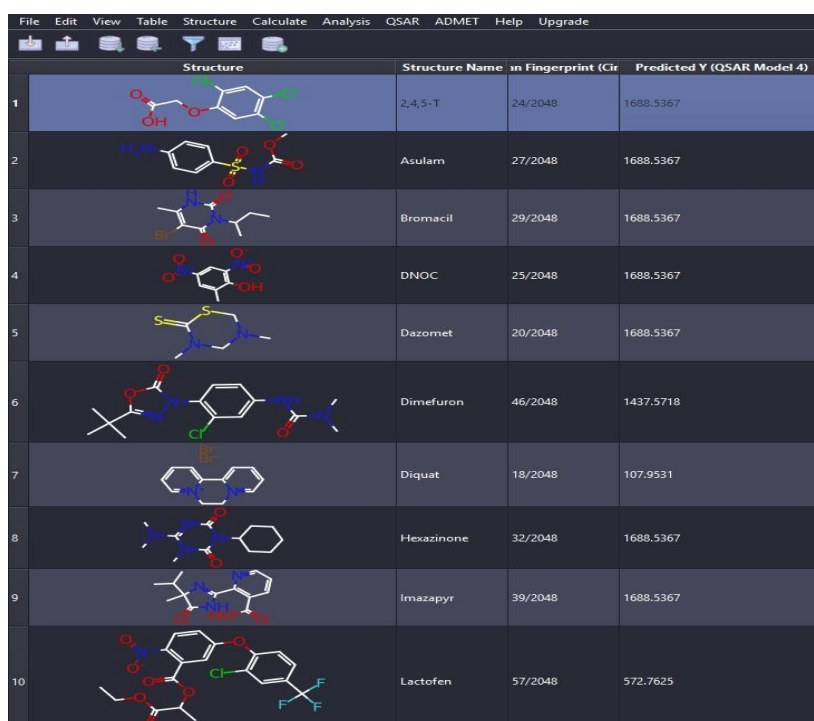
N°	Herbicide
1	2,4,5-T
2	Asulam
3	Bromacil
4	DNOC
5	Dazomet
6	Dimefuron
7	Diquat

CHAPITRE III : RÉSULTATS ET DISCUSSION

8	Hexazinone
9	Imazapyr
10	Lactofen

Tableau III-4: Herbicides utilisés pour l'application du modèle QSAR

Interface du logiciel ChemMaster montrant les herbicides utilisés pour l'application du modèle QSAR.



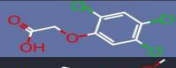
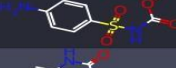
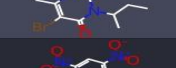
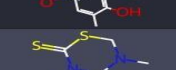
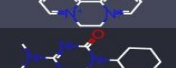
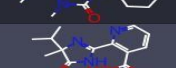
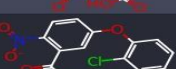
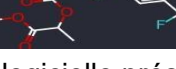
	Structure	Structure Name	in Fingerprint (Clr	Predicted Y (QSAR Model 4)
1		2,4,5-T	24/2048	1688.5367
2		Asulam	27/2048	1688.5367
3		Bromacil	29/2048	1688.5367
4		DNOC	25/2048	1688.5367
5		Dazomet	20/2048	1688.5367
6		Dimetufuron	46/2048	1437.5718
7		Diquat	18/2048	107.9531
8		Hexazinone	32/2048	1688.5367
9		Imazapyr	39/2048	1688.5367
10		Lactofen	57/2048	572.7625

Figure III-3 : Interface logicielle présentant les structures 2D des herbicides et leurs valeurs de toxicité prédites par le modèle QSAR 4.

III.5.2. Prédiction des valeurs de LD50

Le modèle QSAR optimal a été appliqué aux dix herbicides sélectionnés afin de prédire leurs valeurs de toxicité aiguë orale (LD50). Les prédictions ont été réalisées automatiquement par le logiciel ChemMaster à partir des empreintes moléculaires générées pour chaque composé [56].

Les résultats obtenus sont présentés dans le **Tableau III.5**

Herbicides	LD50 prédite (mg/kg)
2,4,5-T	1688,54
Asulam	1688,54
Bromacil	1688,54
DNOC	1688,54
Dazomet	1688,54
Dimefuron	1437,57
Diquat	107,95
Hexazinone	1688,54
Imazapyr	1688,54
Lactofen	572,76

Tableau III-5 : Valeurs de LD50 prédites par le modèle QSAR 4

Les résultats montrent une variabilité importante des valeurs prédites, traduisant des différences significatives de toxicité entre les molécules étudiées

III.5.3. Discussion des résultats obtenus

L'application du modèle QSAR développé a permis d'obtenir une estimation rapide de la toxicité des dix herbicides sélectionnés. Les valeurs prédites mettent en évidence des profils toxicologiques différents selon la nature chimique des composés.

Le composé Diquat présente la valeur de LD50 prédite la plus faible (107,95 mg/kg), ce qui suggère une toxicité plus élevée par rapport aux autres molécules étudiées. Ce résultat est cohérent avec la littérature scientifique qui classe le Diquat parmi les herbicides présentant une toxicité relativement importante pour les organismes vivants.

À l'inverse, plusieurs composés tels que 2,4,5-T, Asulam, Bromacil, DNOC, Dazomet, Hexazinone et Imazapyr présentent des valeurs de LD50 similaires, proches de 1688 mg/kg. Ces résultats indiquent une toxicité modérée selon les prédictions du modèle.

Le Lactofen présente une valeur intermédiaire (572,76 mg/kg), traduisant une toxicité plus importante que celle observée pour la majorité des composés étudiés, mais inférieure à celle prédite pour le Diquat.

Les résultats obtenus démontrent l'intérêt des modèles QSAR dans l'évaluation préliminaire des risques toxicologiques. En effet, ces approches permettent de prédire rapidement les propriétés toxicologiques de nouvelles molécules à partir de leur structure chimique, tout en réduisant le recours aux expérimentations animales [38].

Cependant, les prédictions obtenues doivent être interprétées avec prudence, notamment lorsque les molécules étudiées présentent des caractéristiques structurales éloignées de celles utilisées pour construire le modèle. L'amélioration future de la base de données ainsi que l'intégration d'un plus grand nombre de composés pourraient permettre d'augmenter la précision des prédictions.

III.6. Discussion générale

L'objectif principal de cette étude était de développer des modèles QSAR capables de prédire la toxicité aiguë des herbicides à partir de leurs caractéristiques moléculaires. Pour cela, une base de données composée de 34 herbicides a été utilisée afin de construire plusieurs modèles de Régression Linéaire Multiple (MLR). Les résultats obtenus ont permis d'évaluer les performances des modèles développés et d'examiner leur potentiel pour la prédiction de la toxicité de nouveaux composés.

III.6.1. Forces du modèle

Les modèles QSAR développés dans cette étude présentent plusieurs avantages. Tout d'abord, les valeurs élevées du coefficient de détermination obtenues pour les ensembles d'apprentissage (R^2 compris entre 0,799 et 0,905) montrent que les descripteurs moléculaires sélectionnés expliquent une part importante de la variabilité observée de la toxicité des herbicides.

Le modèle 4 s'est particulièrement distingué par ses performances statistiques supérieures, avec un coefficient R^2 de 0,905 ainsi que les plus faibles valeurs de RMSE et de MAE. Ces résultats témoignent d'une bonne capacité d'ajustement et d'une représentation satisfaisante des relations entre la structure moléculaire et la toxicité.

Par ailleurs, l'application du modèle à dix nouveaux herbicides a démontré son utilité pratique pour l'estimation rapide de la toxicité de composés n'ayant pas été testés expérimentalement. Cette approche constitue un outil intéressant pour le criblage préliminaire des substances chimiques et l'évaluation des risques environnementaux.

Enfin, l'utilisation du logiciel ChemMaster a permis d'automatiser le calcul des descripteurs moléculaires et la construction des modèles, facilitant ainsi l'analyse de données complexes et la génération de prédictions fiables.

III.6.2. Limites de l'étude

Malgré les résultats encourageants obtenus, certaines limitations doivent être prises en considération.

La principale limite concerne la taille relativement réduite de la base de données utilisée. En effet, l'étude repose sur seulement 34 herbicides, ce qui peut limiter la capacité du modèle à

capturer l'ensemble de la diversité chimique des substances herbicides présentes sur le marché.

Une autre limitation réside dans les performances modestes observées lors de la validation externe. Les valeurs de R^2 Test obtenues restent faibles, voire négatives pour certains modèles. Ces résultats indiquent que la capacité de généralisation des modèles demeure limitée et que les prédictions réalisées sur de nouvelles molécules doivent être interprétées avec prudence.

La diversité structurale importante des herbicides étudiés constitue également un facteur susceptible d'influencer les performances des modèles. Des composés appartenant à des familles chimiques très différentes peuvent présenter des mécanismes toxicologiques distincts qui ne sont pas toujours correctement capturés par un modèle linéaire.

Enfin, la méthode MLR utilisée dans cette étude repose sur l'hypothèse d'une relation linéaire entre les descripteurs moléculaires et la toxicité. Or, les phénomènes biologiques impliqués dans les mécanismes toxiques sont souvent complexes et non linéaires.

III.6.3. Comparaison avec la littérature

Les résultats obtenus sont globalement cohérents avec les observations rapportées dans la littérature scientifique consacrée à la modélisation QSAR de la toxicité des pesticides.

Plusieurs études ont montré que des descripteurs liés à la masse moléculaire, à la lipophilie (LogP), aux propriétés électroniques et à la polarité jouent un rôle majeur dans la prédiction des propriétés toxicologiques des composés organiques (Roy et al., 2015 ; Gramatica, 2007).

Les coefficients de détermination obtenus dans cette étude sont comparables à ceux rapportés dans certaines études QSAR utilisant des méthodes de régression linéaire multiple pour la prédiction de la toxicité environnementale. Toutefois, des performances supérieures sont parfois observées lorsque des méthodes d'apprentissage automatique plus avancées sont utilisées, telles que les forêts aléatoires (Random Forest), les machines à vecteurs de support (Support Vector Machine) ou les réseaux de neurones artificiels.

Malgré ces différences, les résultats obtenus confirment l'intérêt des approches QSAR comme outil complémentaire à l'expérimentation classique pour l'évaluation des risques liés aux substances chimiques.

III.6.4. Perspectives d'amélioration

Plusieurs pistes peuvent être envisagées afin d'améliorer les performances des modèles développés dans cette étude

L'une des premières améliorations consisterait à augmenter la taille de la base de données en intégrant un plus grand nombre d'herbicides et de composés apparentés. Une base de

données plus importante permettrait d'élargir le domaine chimique couvert et d'améliorer la robustesse statistique des modèles.

Il serait également intéressant d'explorer d'autres catégories de descripteurs moléculaires, notamment les descripteurs tridimensionnels, quantiques ou topologiques susceptibles d'apporter des informations complémentaires sur les propriétés des molécules.

L'utilisation de techniques de modélisation plus avancées pourrait également permettre d'améliorer les performances prédictives. Des méthodes telles que les forêts aléatoires, les réseaux de neurones artificiels ou les algorithmes de gradient boosting sont aujourd'hui largement utilisées dans le domaine du QSAR et offrent souvent de meilleures capacités de généralisation.

Enfin, une validation sur des bases de données externes plus importantes permettrait de confirmer la robustesse des modèles et d'évaluer leur applicabilité dans des contextes réels d'évaluation du risque environnement.

CONCLUSION GÉNÉRALE

CONCLUSION GÉNÉRALE

L'objectif principal de ce mémoire était de développer des modèles QSAR capables de prédire la toxicité des herbicides présents dans les sols agricoles à partir de leurs caractéristiques moléculaires. Cette étude s'inscrit dans une démarche visant à améliorer l'évaluation des risques environnementaux tout en réduisant le recours aux expérimentations biologiques traditionnelles.

Une base de données composée de 34 herbicides a été constituée à partir de données toxicologiques disponibles dans des bases de données scientifiques reconnues. Les descripteurs moléculaires ont été calculés à l'aide du logiciel ChemMaster puis sélectionnés après différentes étapes de prétraitement et d'analyse statistique. Quatre modèles QSAR ont ensuite été développés en utilisant la méthode de Régression Linéaire Multiple (MLR).

L'évaluation des performances des modèles a montré que le modèle 4 présente les meilleurs résultats statistiques avec un coefficient de détermination (R^2) de 0,905 ainsi que les plus faibles valeurs de RMSE et de MAE. Les procédures de validation interne et externe ont permis de vérifier la robustesse des modèles et de mettre en évidence leurs capacités prédictives.

Afin de démontrer l'applicabilité du modèle optimal, celui-ci a été utilisé pour prédire les valeurs de toxicité (LD50) de dix herbicides supplémentaires n'ayant pas participé à sa construction. Les résultats obtenus ont confirmé l'intérêt des approches QSAR comme outils d'aide à la prédiction de la toxicité de nouvelles substances chimiques.

Les résultats de cette étude permettent ainsi de répondre à la problématique initiale en montrant qu'il est possible d'établir des relations quantitatives entre la structure moléculaire des herbicides et leur toxicité. Bien que certaines limitations subsistent, notamment la taille relativement réduite de la base de données et les performances limitées de la validation externe, les modèles développés constituent une base intéressante pour l'évaluation préliminaire des risques toxicologiques.

Ce travail met également en évidence l'intérêt des méthodes QSAR dans le domaine de la chimie environnementale. Ces approches offrent une alternative rapide, économique et respectueuse des principes de réduction de l'expérimentation animale pour l'évaluation des substances chimiques.

Dans une future perspective, il serait intéressant d'élargir la base de données à un plus grand nombre d'herbicides, d'intégrer de nouveaux types de descripteurs moléculaires et d'explorer des techniques de modélisation plus avancées telles que les forêts aléatoires (Random Forest), les machines à vecteurs de support (SVM) ou les réseaux de neurones artificiels. Ces améliorations pourraient contribuer à augmenter la précision et la capacité prédictive des modèles QSAR destinés à l'évaluation des risques environnementaux liés aux pesticides.

REFERENCES

BIBLIOGRAPHIQUES

RÉFÉRENCES BIBLIOGRAPHIQUES

- [1] KOSTIĆ, D., 2018. Tractor spraying pesticides at soybean field
- [2] Matsunaka, S. (1976). Evolution of herbicides in modern agriculture. *Journal of Pesticide Science*, 1(1), 3–10.
- [3] Peterson, M. A., McMaster, S. A., Riechers, D. E., Skelton, J., & Stahlman, P. W. (2016). 2,4-D past, present, and future: A review. *Weed Technology*, 30(2), 303–345.
- [4] Singh, B. K., Walker, A., & Wright, D. J. (2006). Bioremedial potential of fenamiphos and chlorpyrifos degrading isolates. *Chemosphere*, 65, 618–628.
- [5] Duke, S. O., & Powles, S. B. (2008). Glyphosate: a once-in-a-century herbicide. *Pest Management Science*, 64(4), 319–325.
- [6] Aktar, W., Sengupta, D., & Chowdhury, A. (2009). Impact of pesticides use in agriculture: their benefits and hazards. *Interdisciplinary Toxicology*, 2(1), 1–12.
- [7] VectorStock. (2021). Sources of groundwater pollution diagram [Infographie vectorielle].
- [8] Hassall, K. A. (1990). *The Biochemistry and Uses of Pesticides*. Macmillan Press.
- [9] Ware, G. W., & Whitacre, D. M. (2004). *The Pesticide Book*. MeisterPro Information Resources.
- [10] Duke, S. O. (2012). Why have no new herbicide modes of action appeared in recent years? *Pest Management Science*, 68(4), 505–512.
- [11] Kraehmer, H., Laber, B., Rosinger, C., & Schulz, A. (2014). Herbicides as weed control agents: state of the art. *Plant Physiology*, 166(3), 1119–1131.
- [12] Shimabukuro, R. H., & Swanson, H. R. (1969). Atrazine metabolism, selectivity, and mode of action. *Journal of Agricultural and Food Chemistry*, 17(2), 199–205.
- [13] Devine, M. D., Duke, S. O., & Fedtke, C. (1993). *Physiology of Herbicide Action*. Prentice Hall.
- [14] Cobb, A. H., & Reade, J. P. H. (2010). *Herbicides and Plant Physiology*. Wiley-Blackwell.
- [15] Giddings, J. M., Anderson, T. A., Hall, L. W., Hosmer, A. J., Kendall, R. J., Richards, R. P., Solomon, K. R., & Williams, W. M. (2005). Atrazine in North American surface waters: a probabilistic aquatic ecological risk assessment. *Environmental Toxicology and Chemistry*, 24(4), 815–827.
- [16] Islam, F., Wang, J., Farooq, M. A., Khan, M. S. S., Xu, L., Zhu, J., Zhao, M., Muñoz, S., Li, Q. X., & Zhou, W. (2018). Potential impact of the herbicide 2,4-dichlorophenoxyacetic acid on human and ecosystems. *Environment International*, 111, 332–351.

RÉFÉRENCES BIBLIOGRAPHIQUES

- [17] Tomlin, C. D. S. (2009). *The Pesticide Manual*. British Crop Production Council.
- [18] Weber, J. B., & Miller, C. T. (1989). Organic chemical movement over and through soil. *Advances in Soil Science*, 11, 305–334.
- [19] Sparks, D. L. (2003). *Environmental Soil Chemistry*. Academic Press.
- [20] Alexander, M. (1999). *Biodegradation and bioremediation*. Academic Press.
- [21] Fenner, K., Canonica, S., Wackett, L. P., & Elsner, M. (2013). Evaluating pesticide degradation in the environment. *Science*, 341(6147), 752–758.
- [22] Lapworth, D. J., Baran, N., Stuart, M. E., & Ward, R. S. (2012). Emerging organic contaminants in groundwater. *Environmental Pollution*, 163, 287–303.
- [23] VectorStock. (2021). Sources of groundwater pollution diagram [Infographie vectorielle]. VectorStock.
- [24] Newman, M. C., & Unger, M. A. (2003). *Fundamentals of Ecotoxicology*. CRC Press.
- [25] Biologie-Schule. (2020). Mechanismen der Phytoremediation: Phytoextraktion, Phytostabilisierung und Phytodegradation [Schéma conceptuel]. Biologie-Schule Deutschland.
- [26] Khan, M. A., & Iqbal, M. (2006). Factors affecting pesticide behavior in soil. *Environmental Chemistry Letters*, 4, 99–104.
- [27] Pimentel, D. (2005). Environmental and economic costs of the application of pesticides primarily in the United States. *Environment, Development and Sustainability*, 7, 229–252.
- [28] Newman, M. C. (2014). *Fundamentals of Ecotoxicology*. CRC Press.
- [29] Rand, G. M. (1995). *Fundamentals of Aquatic Toxicology*. Taylor & Francis.
- [30] Klaassen, C. D. (2013). *Casarett and Doull's Toxicology: The Basic Science of Poisons*. McGraw-Hill.
- [31] Designua. (2020). Lethal Dose 50 percent (LD50) toxicity diagram [Infographie vectorielle]. Shutterstock.
- [32] Walker, C. H., Hopkin, S. P., Sibly, R. M., & Peakall, D. B. (2012). *Principles of Ecotoxicology*. CRC Press.
- [33] OECD (2006). *Current Approaches in the Statistical Analysis of Ecotoxicity Data*. OECD Publishing.

RÉFÉRENCES BIBLIOGRAPHIQUES

- [34] Edwards, C. A., & Bohlen, P. J. (1996). *Biology and Ecology of Earthworms*. Chapman & Hall.
- [35] Domínguez, J., & Pérez-Losada, M. (2010). *Eisenia fetida* and *Eisenia andrei* as model organisms in earthworm ecotoxicology [Photographie]. ResearchGate.
- [36] Johnsen, K., Jacobsen, C. S., Torsvik, V., & Sørensen, J. (2001). Pesticide effects on bacterial diversity in agricultural soils. *Biology and Fertility of Soils*, 33, 443–453.
- [37] Solomon, K. R., Carr, J. A., Du Preez, L. H., Giesy, J. P., Kendall, R. J., Smith, E. E., & Van Der Kraak, G. J. (2008). Effects of atrazine on fish, amphibians, and aquatic reptiles. *Critical Reviews in Toxicology*, 38(9), 721–772.
- [38] Roy, K., Kar, S., & Das, R. N. (2015). *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. Academic Press.
- [39] Cronin, M. T. D., & Schultz, T. W. (2003). Pitfalls in QSAR. *Journal of Molecular Structure: THEOCHEM*, 622(1–2), 39–51.
- [40] Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. Wiley-VCH.
- [41] El-Gharbi, M. (2021). *Principes de la chémoinformatique et développement de modèles QSAR pour la prédiction de la toxicité environnementale* [Thèse de Master, Faculté des Sciences]. DSpace Repository.
- [42] Hugo Kubinyi. *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications*, Ludwigshafen, Germany
- [43] Boukhari, S. (2022). *Modélisation moléculaire et approches chimioinformatiques : Étude historique et applications des modèles QSAR/ QSPR* [Mémoire de Master, Département de Chimie]. Archive Ouverte Scientifique.
- [44] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R. D., & Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry*, 57(12), 4977–5010.
- [45] Leo, A., Hansch, C., & Elkins, D. (1971). Partition coefficients and their uses. *Chemical Reviews*, 71(6), 525-616.
- [46] Todeschini, R., & Consonni, V. (2000). *Handbook of Molecular Descriptors*. Wiley-VCH.
- [47] Gramatica, P. (2007). Principles of QSAR models validation. *QSAR & Combinatorial Science*, 26(5), 694–701.
- [48] Roy Kunal, Kar Supratik, Rudra Narayan Das. *A Primer on QSAR/QSPR Modeling Fundamental Concepts*.

RÉFÉRENCES BIBLIOGRAPHIQUES

- [49] Amari, L. (2023). Apports de la modélisation chimioinformatique dans l'évaluation écotoxicologique des contaminants de l'environnement [Mémoire de Master, Département de Chimie Analytique]. Plateforme des Thèses et Mémoires.
- [50] Mark T. D. Cronin. Tomasz Puzyn, Jerzy Leszczynski, Mark T. Cronin. Recent Advances in QSAR Studies: Methods and Applications. Page: 4.
- [51] Kar, S., Sanderson, H., Roy, K., Leszczynski, J., & Cronin, M. T. (2018). Workflow and best practices for the development of robust QSAR models [Schéma méthodologique]. *Environmental Science and Pollution Research*, 25(12), 11234-11248.
- [52] National Center for Biotechnology Information (NCBI). (2026). PubChem Substance and Compound Databases [Capture d'écran]. National Library of Medicine (US)
- [53] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2023). PubChem 2023 update. *Nucleic Acids Research*, 51(D1), D1373–D1380.
- [54] Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., et al. (2017). The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics*, 9(1), 61.
- [55] Qin, J., & Zhang, L. (2024). Development of intelligent consensus QSAR models for predicting acute lethal toxicity of organic chemicals to mayfly [Schéma méthodologique]. *Journal of Hazardous Materials*, 465, 133450.
- [56] Gramatica, P. (2007). Principles of QSAR model validation. *QSAR & Combinatorial Science*, 26(5), 694–701.
- [57] Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., & Wold, S. (2013). Multi- and Megavariate Data Analysis. Umetrics Academy.
- [58] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.
- [59] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* [Schéma conceptuel]. Springer
- [60] Gramatica, P. (2020). Principles of QSAR modeling. *International Journal of Quantitative Structure–Property Relationships*, 5(3), 61–97.
- [61] OECD. (2014). *Guidance Document on the Validation of (Q)SAR Models*. OECD Publishing.
- [62] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.)* [Schéma conceptuel]. O'Reilly Media.

RÉFÉRENCES BIBLIOGRAPHIQUES

- [63] Bishop, C. M. (2006). Pattern Recognition and Machine Learning [Schéma conceptuel]. Springer.
- [64] Draper, N. R., & Smith, H. (1998). Applied Regression Analysis. Wiley.
- [65] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
- [66] Willmott, C. J., & Matsuura, K. (2005). Advantages of the MAE over the RMSE in assessing average model performance. *Climate Research*, 30, 79–82.

