REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR

ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE SALAH BOUBNIDER CONSTANTINE 3

FACULTE DE GENIE DES PROCEDES

DEPARTEMENT DE GENIE PHARMACEUTIQUE

Mémoire de Master

Filière : Génie des procédés **Spécialité : Génie Pharmaceutique**

Développement d'un Modèle QSAR Avancé pour la Prédiction d'Activités Ames de Molécules Nitro-aromatique.

Encadrant :
Dr. GHORAB Hamida
Dr. BOUHDJAR Khalid

Présenté par :
DEROUAZ Ahmed Seif Ed
ROBAI Safouane
BOUDERMINE Amir Nidjed
LABED Younes

Année Universitaire 2023/2024
Session : Juin 2024


بسم الله الرحمن الرحيم

Remerciement

On remercie Dieu le tout puissant de nous avoir donné la santé, la volonté et la patience d'entamer et de terminer ce mémoire.

A nos encadreurs

Nous souhaitons remercier vivement notre encadrant Mme Hamida Ghorab, pour sa patience, sa disponibilité, et Mr Khalid BOUHEDJAR Chercheur au centre de Recherche en Biotechnologie (CRBt) pour son aide, ses conseils et sa patience pendant les travaux.

A l'équipe du (CRBt)

Nous tenons à montrer notre gratitude à toutes les personnes qu'on a croisées au (CRBt), ce qui a contribué à créer une si bonne atmosphère au cours de notre séjour dans le laboratoire 18.

Aussi, j'adresse évidemment mes sincères remerciements à l'ensemble de membres de jury.

Nous adressons nos sincères remerciements à tous les enseignants qui ont assuré notre formation durant les cinq ans d'étude et toutes les personnes qui ont contribué de près ou de loin durant tous notre cycle d'étude.

Dédicace

Nous dédions ce mémoire À nos parents, pour leurs amours et leurs sacrifices qui nous ont permis de poursuivre nos rêves. À nos familles et à nos amis, dont le soutien inconditionnel et les encouragements constants ont rendu ce travail possible

ملخص:

يهدف العمل المقدم في هذه المدكرة إلى تطوير نماذج QSAR موثوقة ومستقرة وقابلة للتنبؤ بخصائص logAT100 للمركبات النيتروأروماتية.

تم استخدام نوعين من نماذج QSAR ، الانحدار والتصنيف، لنمذجة سمية المواد التي تمثلها logAT100 للمركبات النيتروأروماتية.

تم ربط نموذج الانحدار، معR^2 = 72.96 ٪ وQ^2 = 71.58 ٪، بأربعة موصوفات جزيئية RCI :، SM2_Dz (i)، ATSC1m، و .MaxddsN تتوافق هذه النتائج مع التوصيات الرئيسية لـ Golbraikh و.Tropsha

حقق نموذج التصنيف باستخدام التعلم الآلي، Logistic Regression (LRegression)، حساسية 85.71٪، وخصوصية 80.43٪، ودقة 76.92٪. تشير هذه النتائج إلى أن هذا النموذج يتمتع بمعايير تنبؤية داخلية وخارجية ممتازة، بالإضافة إلى القوة والاستقرار.

كلمات مفتاحية : تجربة العالم اَمس, الطفرة, الانحدار الخطي المتعدد, التعلم الآلي, عطرو-اَروماتي.

Résumé :

Le travail présenté dans ce mémoire vise à développer des modèles QSAR fiables, stables et prédictifs pour la prédiction des propriétés logAT100 des composés nitro-aromatiques.

Deux types de modèles QSAR, régression et classification, ont été utilisés pour modéliser la toxicité des substances représentée par le logAT100 des composés nitro-aromatiques.

Le modèle de régression, avec un R^2 de 72,96 % et un Q^2 de 71,58 %, a été corrélé avec quatre descripteurs moléculaires : RCI, SM2_Dz(i), ATSC1m et MaxddsN. Ces résultats sont conformes aux principales recommandations de Golbraikh et Tropsha.

Le modèle de classification par apprentissage automatique, Logistic Regression (LRegression), a obtenu une sensibilité de 85,71 %, une spécificité de 80,43 % et une précision de 76,92 %. Ces résultats indiquent que ce modèle possède d'excellents paramètres prédictifs internes et externes, ainsi qu'un caractère robuste et stable.

Mots clés : Test d’AMES, Mutagénicité, MLR, Apprentissage Automatique, Nitro-aromatique.

Abstract:

The work presented in this thesis aims to develop reliable, stable, and predictive QSAR models for predicting logAT100 properties of nitroaromatic compounds.

Two types of QSAR models, regression and classification, were used to model the toxicity of substances represented by the logAT100 of nitroaromatic compounds.

The regression model, with an R^2 of 72.96% and a Q^2 of 71.58%, was correlated with four molecular descriptors: RCI, SM2_Dz(i), ATSC1m, and MaxddsN. These results are in accordance with the main recommendations of Golbraikh and Tropsha.

The machine learning classification model, Logistic Regression (LRegression), achieved a sensitivity of 85.71%, a specificity of 80.43%, and an accuracy of 76.92%. These results indicate that this model has excellent internal and external predictive parameters, as well as robustness and stability.

Keywords: AMES Test, Mutagenicity, MLR, Machine Learning, nitro-aromatic.

Liste des tableaux

Liste des Annexes

Liste des figures

Liste des Symboles et abréviations

Abreviation	Signification
$R`^2_0$	Coefficient of determination: observed versus predicted activities.
R^2_0	Coefficient of determination (determined the predicted versus observed activities).
2D	Structure deux dimensions.
3D	Structure tridimensionnelles.
3R	Réduction, raffinement, Remplacement
4D	Structure quatre dimensions.
ADN	Acide désoxyribonucléique.
AMM	Autorisation de mise sur le marché.
AEM	L'Agence européenne des médicaments.
AM1	Austin Mode 1.
AT100	Souche pour le test d'Ames
CIRC	Centre international de recherche sur le cancer.
DDT	Dichlorodiphényltrichloroéthane
DL_{50}	La dose qui cause la mort de 50% de la population.
F	Indice de Fisher.
GA	Algorithme génétique
HAP	Hydrocarbures aromatiques polycycliques
HPV	Human Papillomavirus
HUMO	Orbitale moléculaire occupée de plus haute énergie.
ICH	Conseil international d'harmonisation
IC_{50}	Concentration inhibitrice.
K	Slope: predicted versus observed activities regression lines through the origin.
k'	Slope: observed versus predicted activities regression lines through the origin.

LD_{50}	La dose nécessaire pour la moitié des membres d'une population testée après une durée de test spécifiée.
LOO	Validation croisée par omission d'une observation : Cross-validation by leaveone-out.
LUMO	Orbitale moléculaire non –occupée de plus basse énergie.
MLR	Régression linéaire multiple.
MM+	Mécanique Moléculaire (+).
OCDE	Organisation de Coopération et de Développement Economiques.
MCO	Moindres carrés ordinaires.
PCA	L'analyse en composantes principales
PCB	Polychlorobiphényles
PM3	Parametrization Method 3.
PRESS	Somme des carrés des erreurs de prédiction
Q	Cross-validated correlation coefficient.
Q^2	Coefficient de validation externe.
Q^2_{LMO}	Coefficient de prédiction (Leave-Many-Out).
Q^2_{LOO}	Coefficient de prédiction (Leave-One-Out).
QSAR	Relation quantitative-structure-activité.
QSPR	Relation quantitative-propriété-activité.
R	Coefficient de corrélation.
R^2	Coefficient de détermination.
R^2_{ext}	Coefficient de détermination externe.
R^2_{adj}	Coefficient de détermination ajusté.
RMN	Résonnance Magnétique Nucléaire.
SAR	Relation -structure-activité.
SCE	Somme des Carrés Expliquée
SCR	Somme des Carrés Résidus
SCT	Somme des Carrés Totale
Tr	Training.

UV	Ultraviolet
VHB	Le virus de l'hépatite B
VHC	Le virus de l'hépatite C
VIH	Le virus de l'immunodéficience humaine
$\bar{y}$ (*i*)	Valeur prédite.
$\hat{y}$ *i*	Valeur estimée.
yi	Valeur observée.

Sommaire

Table des matières

Sommaire

Introduction Générale

Introduction générale

L'évaluation du potentiel mutagène des produits de consommation constitue une préoccupation majeure pour les agences réglementaires internationales, étant donné les implications graves des mutations de l'ADN sur la santé humaine. Le développement du cancer, par exemple, est étroitement lié aux mutations résultant d'interactions chimiques. Pour répondre à ces préoccupations, l'essai d'Ames a été largement adopté comme méthode standard pour détecter la mutagénèse induite par des produits chimiques [1]. Cet essai, qui utilise des bactéries pour identifier les mutations génétiques potentielles, telles que les décalages de cadre et les substitutions de paires de bases, est mandaté par les agences réglementaires internationales pour évaluer le risque mutagène des produits de consommation.

Cependant, l'essai d'Ames ne suffit souvent pas à lui-même. Conformément aux directives du Conseil international d'harmonisation (ICH), des approches in silico viennent compléter cet essai in vitro. Ces méthodes, qui incluent notamment les modèles de relation quantitative structure-activité (QSAR), permettent de prédire les résultats du test d'Ames sur la base de la structure chimique des composés [2], [3]. En intégrant ces modèles, il est possible de réaliser une évaluation plus complète et efficace de la sécurité chimique, réduisant ainsi la nécessité d'expérimentations in vitro et in vivo, et favorisant une approche plus éthique et économique dans l'évaluation des risques.

Ce mémoire se structure en deux grandes parties principales. La première consiste en une synthèse bibliographique sur le cancer, la mutagénicité et la méthodologie QSAR. La deuxième est réservée aux études expérimentales, regroupant les matériaux utilisés dans ce travail tels que les composés organiques étudiés (Nitro Aromatiques), les approches statistiques, les logiciels utilisés, ainsi que la discussion des résultats.

En somme, cette recherche vise à approfondir notre compréhension des méthodes de détection des mutations génétiques induites par des composés chimiques, en combinant des

techniques expérimentales éprouvées et des outils de modélisation avancés pour offrir une évaluation plus robuste et prédictive des risques mutagènes.

Partie I : Synthèse bibliographique

Chapitre 1 : La mutagénicité et drug design

1. Le cancer

1.1. Généralités

Le cancer, appelé également tumeur maligne ou néoplasme, est une maladie caractérisée par la prolifération incontrôlée de cellules, liée à un échappement aux mécanismes de régulation qui assurent le développement harmonieux de notre organisme et la coexistence entre les cellules normales entre elles. En se multipliant de façon anarchique et en modifiant leur environnement, les cellules cancéreuses donnent naissance à des tumeurs de plus en plus grosses qui se développent en envahissant puis détruisant les zones qui les entourent (organes). Les cellules cancéreuses peuvent également essaimer à distance d'un organe pour former une nouvelle tumeur, ou circuler sous forme libre. [4]

1.2 Les facteurs de risque des cancers

Le cancer n'est pas lié à une cause unique. Il résulte d'un ensemble de facteurs pouvant interagir entre eux. Ces facteurs de risque peuvent être internes ou externes.

a. Les facteurs de risque internes

Ces facteurs sont liés à l'âge ou à l'histoire familiale. En effet, même si des cancers peuvent apparaître à tout âge, ils deviennent de plus en plus fréquents au fil des années, notamment après l'âge de 60 ans. Cela est dû au cumul des agressions subies par les cellules au fil de la vie et, probablement, à une moindre efficacité des mécanismes de réparation de l'ADN présente dans les cellules. [5]

Toutefois, certaines personnes présentent plus de risques de développer un cancer que d'autres parce qu'à leur naissance, elles portent certaines mutations dans un ou plusieurs de leurs gènes. Moins d'un cancer sur dix aurait une origine héréditaire.

b. Les facteurs de risque externes

Ces facteurs peuvent être liés à l'environnement**.** Les agressions répétées de l'ADN des cellules par des rayonnements (d'origine nucléaire ou solaire) ou par des produits industriels favorisent l'apparition de cellules cancéreuses. [6]

Le tabagisme est le principal facteur de risque de cancer, responsable de plus de 80 % des cancers du poumon, ainsi que d'autres cancers des voies aérodigestives supérieures et de la vessie.

L'excès d'alcool est associé à plusieurs types de cancer, notamment le cancer du sein, le cancer colorectal, la cavité buccale, le foie, l'œsophage et le larynx. [7]

L'alimentation joue également un rôle : l'excès de viande rouge et de charcuterie augmente le risque de cancer colorectal, tandis qu'une alimentation riche en fruits et légumes peut être protectrice. [6]

De plus, l'exposition au soleil (rayons UV) peut causer des cancers de la peau tels que le mélanome. Certains virus, comme le papillomavirus humain (HPV), sont associés à des cancers, comme le cancer du col de l'utérus [6]. La figure1 montre les principaux facteurs externes du cancer.

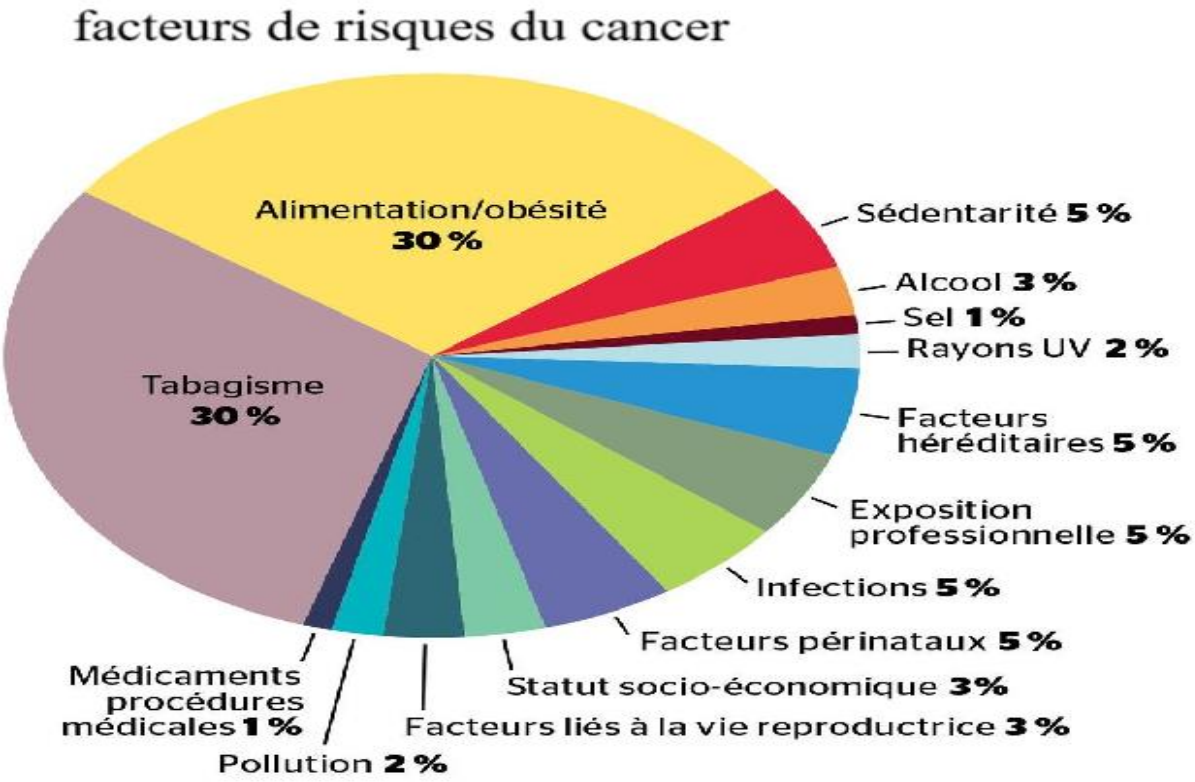


Figure 1 : les facteurs de risques du cancer [7]

1.3. La mutagénicité

1.3.1. La Mutation

Une mutation est une modification de l'information génétique contenue dans l'ADN. Elle affecte la séquence par le remplacement d'un ou plusieurs nucléotides, l'insertion ou la délétion de quelques nucléotides. Elle peut être due à l'instabilité du génome, à des erreurs de copie ou de réparation lors de la multiplication des cellules ou de la reproduction sexuée, à des coupures de l'ADN, à des conditions environnementales ou à l'action ciblée du sélectionneur [8]. On distingue différents types de mutations:

a. Les mutations germinales

Ce sont des changements héréditaires car elles atteignent les cellules reproductrices. Elles peuvent être perpétuées tant par la multiplication végétative que sexuée et correspond à l'apparition d'individus nouveaux ou mutants.

b. Les mutations géniques

La mutation génique correspond à une altération de la séquence nucléotidique de1'ADN de manière à soit arrêter complètement la synthèse d'une protéine, ou la modifier produisant ainsi une protéine inactive. La mutation génique la plus fréquente est la substitution, qui consiste à remplacer un nucléotide par un autre [9]. L'addition d'une base unique **(**insertion) ou la perte d'une base (délétion) unique font aussi partie de ce type de mutations illustrées dans la figure 02 [10]. Comme le 2-Aminofluoréne: qui provoque la formation d'adduits a l'ADN à la position C- 8 de la guanine. [11]

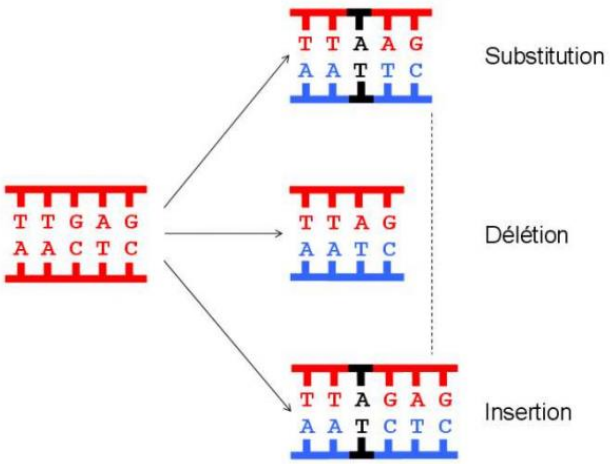


Figure 02 : différents types de mutation génique. [12]

c. Les Mutations chromosomiques

Correspondent à la perte (délétion) ou à l'addition (insertion) de fragments chromosomiques, à l'échange des fragments entre chromosomes non homologues et à la duplication ou à l'inversion d'un segment chromosomique

La mutagénicité du cancer est un domaine de recherche crucial qui explore les mécanismes par lesquels certaines substances et agents environnementaux peuvent endommager l'ADN.

De nombreuses publications ont abordé les stratégies de test pour la mutagénicité. Les évaluations de la sécurité des substances en ce qui concerne la mutagénicité sont essentielles pour garantir que les produits chimiques utilisés dans divers secteurs, tels que les produits pharmaceutiques, les pesticides, et les additifs alimentaires, ne présentent pas de risques génétiques pour les êtres humains et l'environnement. [13], [14]

1.4. Sources de mutagènes

Les sources de substances mutagènes sont variées et peuvent inclure des éléments tels que:

1.4.1. Agents chimiques

a. ***Hydrocarbures aromatiques polycycliques (HAP)* :** Les HAP sont présents dans les fumées de tabac, les gaz d'échappement des véhicules et les émissions industrielles. Ils sont associés à un risque accru de cancer, notamment de cancer du poumon.[15]

b. ***Composés organochlorés :*** Certains composés organochlorés, tels que les pesticides comme le DDT (dichlorodiphényltrichloroéthane) et les PCB (polychlorobiphényles), sont connus pour être des cancérigènes potentiels et peuvent être présents dans l'environnement. [16]

1.4.2. Agents physiques

a. ***Radiations ionisantes***

Les radiations ionisantes, telles que les rayons X, les rayons gamma et les rayons cosmiques, peuvent endommager l'ADN et augmenter le risque de cancer chez les personnes exposées. [17]

b. ***Rayonnements ultraviolets (UV)***

Les rayonnements UV du soleil sont associés à un risque accru de cancer de la peau, notamment le carcinome basocellulaire, le carcinome spinocellulaire et le mélanome. [18]

1.4.3. Agents biologiques

a- Virus oncogènes : Certains virus, tels que le virus de l'hépatite B (VHB), le virus de l'hépatite C (VHC) et le virus de l'immunodéficience humaine (VIH), sont associés à un risque accru de cancer, notamment de cancer du foie et de lymphomes. [19]

1.5. Les enjeux de la mutagénicité

Les enjeux de la mutagénicité sont nombreux et ont des implications importantes dans différents domaines, notamment la santé publique, la protection de l'environnement et le développement de produits.

1.5.1 Santé publique

a. Prévention du cancer : La compréhension des agents mutagènes et cancérigènes ainsi que de leurs sources permet de développer des stratégies de prévention du cancer, telles que la réglementation des substances dangereuses, l'éducation sur les comportements à risque et la promotion de modes de vie sains. [20]

b. Surveillance épidémiologique : La surveillance des expositions aux agents mutagènes et cancérigènes ainsi que des taux de cancer dans la population est essentielle pour évaluer les tendances, identifier les groupes à risque et orienter les politiques de santé publique. [21]

1.5.2 Protection de l'environnement

Gestion des déchets : La présence de substances mutagènes et cancérigènes dans les déchets industriels et domestiques pose des risques pour l'environnement et la santé humaine. La gestion appropriée des déchets, y compris leur traitement et leur élimination sûre, est donc cruciale pour réduire ces risques. [22]

1.5.3. Développement de produits

Évaluation de la sécurité des produits : L'évaluation de la mutagénicité et de la carcinogénicité est une étape clé dans le développement et la commercialisation de produits chimiques, pharmaceutiques et autres. Les réglementations exigent souvent des tests approfondis pour évaluer ces risques avant l'autorisation de mise sur le marché. [23]

2. La conception de médicaments ou Drug Design

2.1 Les étapes de développement du médicament

La conception d'un nouveau médicament est un processus long, coûteux et complexe, débutant par l'identification d'une cible thérapeutique, suivie de la recherche de composés actifs, jusqu'aux essais cliniques. En moyenne, cela prend de 12 à 15 ans et nécessite un investissement d'environ 1 milliard de dollars pour amener un nouveau médicament sur le marché. [24], [25]

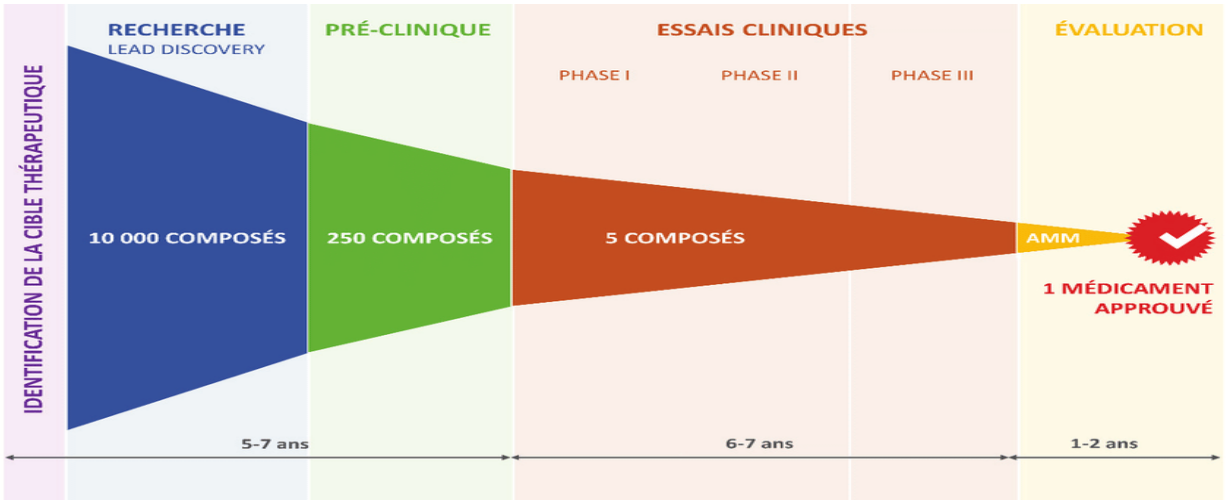


Figure 3 : Vue d'ensemble sur la recherche et le développement d'un nouveau médicament [26]

Celui-ci comprend quatre étapes majeures :

- 1ère étape : La recherche.
- 2ème étape : Les études précliniques.
- 3ème étape : Les études cliniques.
- 4ème étape : Enregistrement et autorisation de mise sur le marché.

2.1.1. La recherche

L'orientation de la recherche dans l'industrie pharmaceutique est influencée par divers facteurs, notamment les avancées de la recherche fondamentale, les exigences médicales actuelles et les orientations stratégiques internes des entreprises. Cette phase initiale du processus de développement de médicaments englobe plusieurs étapes essentielles. Tout d'abord, il y a l'identification et la validation de la cible thérapeutique, où les scientifiques ciblent des biomarqueurs ou des processus biologiques spécifiques associés à une maladie donnée. Ensuite, intervient l'identification de principes actifs, souvent appelée découverte du "hit", où des milliers de composés chimiques sont testés pour leur activité contre la cible thérapeutique. Enfin, l'optimisation moléculaire ou l'identification du "lead" consiste à améliorer les propriétés pharmacologiques et la sélectivité du composé actif initial, en vue de développer un médicament potentiel. Ces étapes sont cruciales pour définir la voie de recherche et pour sélectionner les candidats médicaments les plus prometteurs pour la phase suivante du développement. [27], [28]

2.1.2. Les études précliniques

Les études précliniques jouent un rôle essentiel dans le développement de médicaments en évaluant l'efficacité, la sécurité et la pharmacocinétique des candidats médicaments potentiels avant leur évaluation chez les humains lors des essais cliniques. Cette phase comprend une série d'expériences in vitro et in vivo visant à caractériser le profil pharmacologique des composés candidats. Les études in vitro permettent d'explorer les interactions entre les composés et leurs cibles biologiques, ainsi que leur toxicité potentielle sur les cellules. Parallèlement, les études in vivo utilisent des modèles animaux pour évaluer l'efficacité du composé candidat dans des conditions physiologiques plus complexes, ainsi que sa distribution, son métabolisme et son élimination dans l'organisme. Ces études précliniques sont cruciales pour identifier les candidats médicaments les plus prometteurs et pour orienter la conception des essais cliniques ultérieurs. [29],[30]

2.1.3. Les études cliniques

Les études cliniques sont soumises à une réglementation stricte et ne peuvent être entreprises qu'après avoir obtenu l'autorisation des autorités sanitaires compétentes, telles que l'Agence européenne des médicaments (AEM). Une fois les tests précliniques réussis, une série d'évaluations est menée pour démontrer l'innocuité et l'efficacité des futurs médicaments chez l'homme. Ces évaluations sont cruciales pour garantir que les médicaments proposés répondent aux normes de sécurité et d'efficacité requises avant leur utilisation généralisée. [31]

Phase I : Cette phase évalue principalement la sécurité et la tolérabilité du médicament chez un petit groupe de volontaires sains. L'objectif est de déterminer la dose maximale tolérée et de recueillir des informations préliminaires sur la pharmacocinétique. [32]

Phase II : Dans cette phase, l'efficacité du médicament est évaluée chez un groupe plus large de patients atteints de la maladie cible. Des données supplémentaires sur la sécurité et la posologie optimale sont également recueillies pour guider les essais ultérieurs. [33]

Phase III : Cette phase vise à confirmer l'efficacité et la sécurité du médicament chez un grand nombre de patients dans des conditions de pratique clinique. Les essais de cette phase sont souvent

randomisés et contrôlés par placebo pour évaluer de manière plus précise les avantages du médicament par rapport aux traitements existants ou à l'absence de traitement. [34]

Phase IV : Après l'approbation réglementaire et la mise sur le marché, cette phase consiste en une surveillance post-commercialisation pour évaluer l'efficacité à long terme et détecter d'éventuels effets secondaires rares ou tardifs. [35]

2.1.4. Enregistrement et Autorisation de Mise sur le Marché (AMM)

L'enregistrement et l'autorisation de mise sur le marché (AMM) d'un médicament représentent une étape cruciale dans le processus de développement pharmaceutique. Après les essais cliniques réussis en phases I, II et III, une demande d'AMM est soumise aux autorités réglementaires compétentes telles que la Food and Drug Administration (FDA) aux États-Unis, l'Agence Européenne des Médicaments (EMA) en Europe, ou d'autres agences nationales de régulation. Cette demande inclut des données complètes provenant des essais précliniques et cliniques, des informations sur la fabrication, les contrôles de qualité, et la stabilité du produit. Les autorités examinent rigoureusement ces informations pour s'assurer que le médicament est sûr, efficace et de qualité constante pour les patients. Le processus d'évaluation peut prendre plusieurs mois, voire des années, en fonction de la complexité du dossier et de la nécessité d'informations supplémentaires ou de nouvelles études. Une fois l'AMM accordée, le médicament peut être commercialisé et mis à la disposition des patients, sous réserve de la surveillance continue post-commercialisation pour détecter tout effet indésirable non observé lors des essais cliniques. [32]

2.2. Les approches de la conception de médicaments

Il existe plusieurs approches dans la conception de médicaments, chacune avec ses propres méthodologies et stratégies. Voici quelques-unes des approches couramment utilisées :

2.2.1. Approche basée sur la cible

Cette approche consiste à identifier une cible biologique spécifique, telle qu'une protéine ou un enzyme, impliquée dans une maladie, puis à concevoir des médicaments qui modulent cette cible de manière sélective pour traiter la maladie. [36]

2.2.2. Approche basée sur la structure

Dans cette approche, les médicaments sont conçus en se basant sur la structure tridimensionnelle de la cible biologique, souvent obtenue par cristallographie aux rayons X ou

résonance magnétique nucléaire (RMN). Cela permet de concevoir des médicaments qui se lient de manière spécifique à la cible et modulent son activité de manière efficace. [37]

2.2.3. Approche basée sur les ligands

Cette approche consiste à identifier des composés chimiques (ligands) qui se lient de manière sélective à la cible biologique, puis à optimiser ces ligands pour améliorer leur affinité et leur sélectivité, afin de développer des médicaments efficaces. [38]

2.2.4. Approche basée sur le criblage à haut débit

Cette approche utilise des techniques de criblage à haut débit pour tester de grandes bibliothèques de composés chimiques afin d'identifier ceux qui modulent l'activité de la cible biologique. Les composés les plus prometteurs sont ensuite optimisés pour développer des médicaments. [39]

2.2.5. Approche basée sur le pharmacophore

Cette approche repose sur l'identification des caractéristiques structurelles clés nécessaires pour l'interaction entre un ligand et sa cible biologique. Ces caractéristiques sont utilisées pour concevoir de nouveaux composés chimiques ayant un profil pharmacologique optimal. [40]

Chapitre 2 : La méthodologie QSAR

1. Chimio-bio-informatique

La bio-informatique consiste à conceptualiser la biologie en termes de molécules (au sens de la chimie physique) et à appliquer des "techniques informatiques" (dérivées de disciplines telles que les mathématiques appliquées, l'informatique et la statistique) pour comprendre et organiser les informations associées à ces molécules, à grande échelle. En bref, la bio-informatique est un système d'information de gestion pour la biologie moléculaire et possède de nombreuses applications pratiques. [41]

La bio-informatique se dirige vers un degré plus élevé d'automatisation, de parallélisme et de fiabilité dans la collecte, le stockage et l'élaboration d'informations biologiques ou d'origine biologique [42], S'intéressent de plus en plus à l'utilisation du calcul pour comprendre les phénomènes biologiques et pour acquérir et exploiter des données biologiques. Données à grande échelle. [43]

La chimio-bio-informatique fournit les outils et les ressources nécessaires pour développer et appliquer des modèles QSAR, en utilisant des bases de données moléculaires, des algorithmes d'apprentissage automatique et d'autres techniques de modélisation. Ainsi, la chimio-bio-informatique et le QSAR travaillent en tandem pour prédire les activités biologiques des composés chimiques, ce qui est crucial dans la découverte de médicaments, la conception de nouveaux produits chimiques et d'autres applications en chimie et en biologie.

2. La représentation moléculaire

2.1. **Fingerprints**

Fingerprints moléculaires sont des méthodes couramment employées pour chercher des similitudes. Elles sont constituées de différents descripteurs codés en chaînes de bits. Comme illustré dans la figure 4, on calcule et compare quantitativement les chaînes de bits des composés de la requête et de la base de données en utilisant des mesures de similarité. Les empreintes digitales qui se chevauchent entre les composés testés sont perçues comme une évaluation de la similarité moléculaire. Par conséquent, lorsque le coefficient sélectionné atteint une valeur seuil préétablie, les molécules comparées sont considérées comme étant communes. [44]

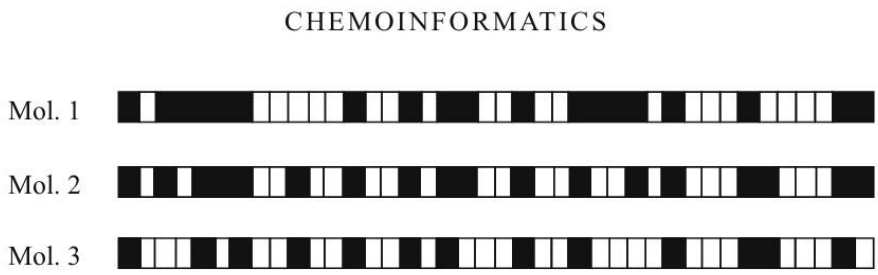


Figure 4 : Fingerprints de modèles et comparaisons Tc. [44]

Fingerprints se présente sous la forme d'une chaîne de bits, constituée d'une séquence d'uns et de zéros où un ou un zéro dans une position précise indique la présence ou l'absence de structure (figure 5). Ce format présente l'avantage d'augmenter la vitesse de calcul et de diminuer l'espace de stockage. [45]

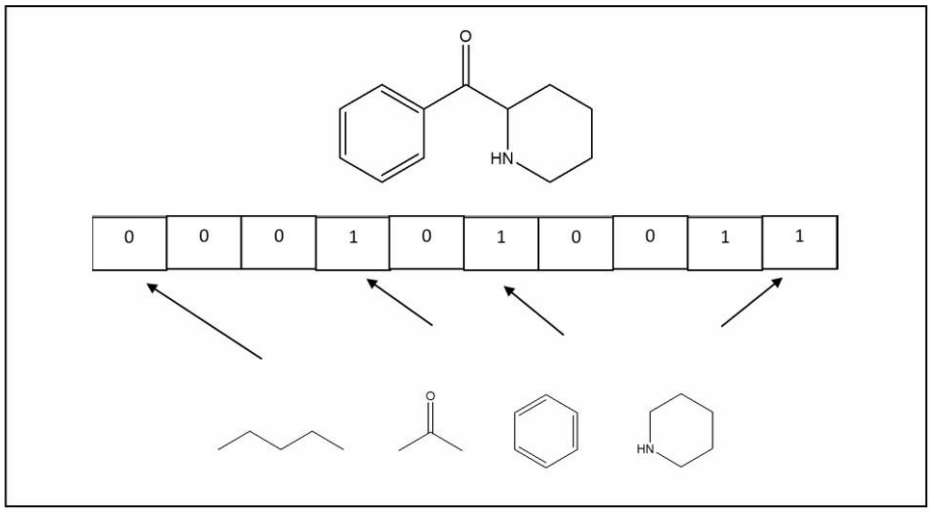


Figure 5 : Codage des structures moléculaires dans une chaîne de bits. [45]

Chaque position de bit dans de nombreux modèles d'empreintes digitales correspond à une caractéristique moléculaire particulière, ce qui active le bit si cette caractéristique est présente. En outre, il est possible de codifier d'autres descripteurs moléculaires de manière incrémentielle en utilisant des chaînes de bits. Les empreintes numériques peuvent être très grandes, pouvant atteindre des milliers de positions de bits. Ces représentations opèrent dans des environnements chimiques de référence et sont comparables à d'autres approches de classification basées sur la

similitude. Un vecteur dans un espace de descripteurs à n dimensions est créé par une empreinte digitale représentant n descripteurs à l'aide de n bits, où chaque dimension est soit nulle, soit égale à l'unité. [44]

Fingerprints en chimie bio-informatique sont souvent classées en 2D ou en 3D selon la dimensionnalité des descripteurs qu'elles codent. Les empreintes digitales 2D captent les fragments structurels, tandis que les empreintes digitales pharmacophores sont considérées comme 3D, représentant chaque position de bit comme un pharmacophore multipoint spécifique. Bien que les méthodes 3D puissent sembler plus réalistes, les difficultés à prédire avec précision les conformations actives des composés et les propriétés 3D dépendantes de la conformation compromettent souvent leur exactitude. Par conséquent, les méthodes 2D plus simples mais plus robustes obtiennent souvent de meilleurs résultats. La performance relative des approches 2D et 3D dépend non seulement de la conception des algorithmes, mais aussi des spécificités des applications. [44]

3. La biologiques des produits chimiques et toxicologie

La biologie des produits chimiques se concentre sur la compréhension des interactions entre les composés chimiques et les systèmes biologiques, y compris les réponses cellulaires, les mécanismes de signalisation et les effets sur la santé humaine et environnementale. [46]

La toxicologie étudie les effets néfastes des substances chimiques sur les organismes vivants, en examinant notamment les doses toxiques, les voies d'exposition et les mécanismes de toxicité. Cela inclut souvent des études sur les animaux pour évaluer les effets potentiels sur la santé humaine. [47]

Les études sur les animaux, telles que les souris, les rats et d'autres espèces, sont souvent utilisées pour évaluer les effets toxiques des produits chimiques sur des systèmes biologiques complexes.[48] Les modèles animaux et notamment les rongeurs sont encore largement utilisés en toxicologie expérimentale, même si l’extrapolabilité des mécanismes de toxicité chez l’animal n’est pas toujours pertinente pour l’homme.[49]

Il est désormais clair que les connaissances fournies par les études sur les animaux ne sont souvent pas transposables aux humains, ce qui explique le taux d'échec très élevé observé lorsque

de nouveaux médicaments sont évalués dans le cadre d'essais cliniques. Aucune conclusion claire ne peut en être tirée.[50]

Les expériences sur les animaux sont cruelles, coûteuses et les résultats ne sont parfois pas applicables aux humains et des problèmes éthiques se posent également. Les scientifiques du monde entier ont donc développé et utilisé des méthodes pour étudier les maladies et tester des produits qui suivent le principe des 3R (réduction, raffinement, Remplacement) et sont réellement pertinents pour la santé humaine. Ces alternatives à l'expérimentation animale incluent des méthodes in vitro sophistiquées comme les cellules et tissus humains, en modes silico comme les techniques avancées de modélisation informatique. [51]

Parmi les techniques avancées de modélisation informatique Le QSAR permet de prédire l'activité biologique des composés chimiques in silico, c'est-à-dire par des méthodes informatiques, sans avoir besoin de tests expérimentaux sur des animaux. Cela peut contribuer à réduire le nombre d'animaux utilisés dans la recherche préclinique.

4. Essais biologiques de toxicité

La toxicité biologique se réfère à la capacité d'une substance à causer des dommages ou des effets nocifs sur des organismes vivants, et son étude est essentielle pour évaluer les risques pour la santé humaine et l'environnement.

Le processus d'utilisation des animaux pour évaluer la toxicité des produits chimiques, également connu sous le nom d'expérimentation animale en toxicologie, est un domaine complexe et réglementé de la recherche scientifique. La relation entre la dose et ses effets sur l'organisme exposé est d'une grande importance en matière de toxicologie. Le processus d'utilisation des animaux pour évaluer la toxicité des produits chimiques a été défini comme suit:

- **Planification de l'étude :** Les chercheurs conçoivent un protocole expérimental détaillé, déterminant les objectifs de l'étude, les doses à tester, les voies d'administration, les espèces animales à utiliser, la durée de l'exposition, les critères d'évaluation des effets toxiques. [52]
- **Sélection des animaux :** Les animaux utilisés dans les études de toxicité peuvent être des rats, des souris, des lapins, des cobayes, des chiens, des singes, etc. Le choix dépend

souvent de la disponibilité, de la pertinence pour le modèle de toxicité étudié et des réglementations locales. [53]

- **Administration de la substance chimique** : Les substances chimiques à tester sont administrées aux animaux selon le protocole expérimental, souvent par voie orale, cutanée, inhalée ou intraveineuse. [54]
- **Observation des effets :** Les animaux sont observés attentivement pour détecter tout signe d'intoxication, de malaise, de changement de comportement, de symptômes physiologiques ou de mortalité. [53]
- **Analyse des résultats :** Les données recueillies sont analysées statistiquement pour évaluer la dose-réponse, déterminer la concentration sans effet observable (NOAEL) et la dose létale médiane (LD50), et identifier les effets toxiques spécifiques associés à différentes doses et voies d'administration. [51]
- **Rapports et interprétation des résultats :** Les résultats de l'étude sont documentés dans un rapport scientifique détaillé, qui est ensuite interprété pour évaluer les risques potentiels pour la santé humaine et l'environnement associés à l'exposition à la substance chimique testée. [55]

La relation dose-réponse décrit comment l'effet sur un organisme varie en fonction des différents niveaux de doses d'un produit chimique après un certain temps d'exposition. En toxicologie, cette relation est souvent représentée graphiquement pour illustrer comment l'effet (par exemple, la réponse biologique ou la toxicité) change avec la dose administrée. [46], [56]

- Sur l'axe des x, nous avons la dose mesurée, tandis que sur l'axe des y, nous avons la réponse. Dans ce scénario, la courbe est généralement en forme de sigmoïde, avec la pente la plus prononcée au milieu. L'exemple présenté dans la figure 6 illustre une courbe dose-réponse pour la DL50.
- La réponse peut prendre la forme d'une réaction physiologique ou biochimique. La DL50 est employée en toxicologie humaine, tandis que la concentration inhibitrice IC50 et sa contrepartie, la concentration efficace EC50, sont utilisées en pharmacologie.

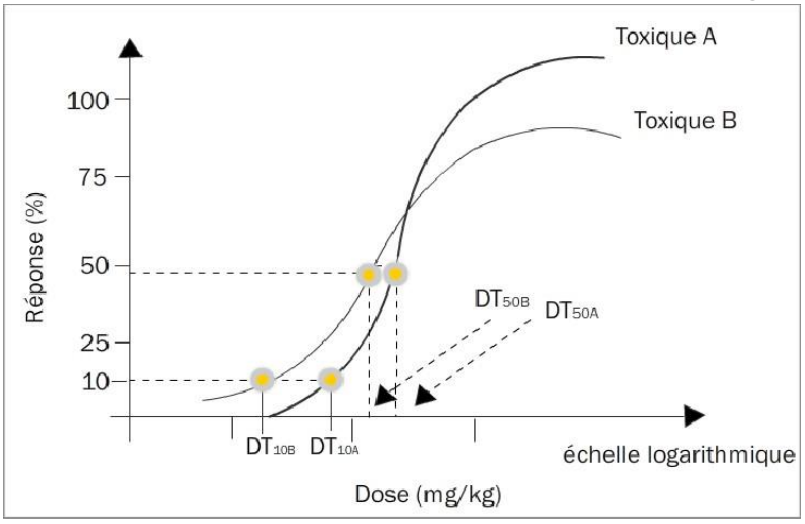


Figure 6 : Exemple de courbe dose-réponse pour la DT50 [57]

La connaissance d'un risque toxique (probabilité et magnitude) pour des individus d'une espèce donnée est obtenue par la mise en œuvre de méthodes, le plus souvent expérimentales prospectives qui utilisent généralement des individus d'une autre espèce. La variabilité interspécifique étant grande, il est donc indispensable d'avoir une estimation de la concordance. Trois approches sont envisageables, les méthodes allométriques, les méthodes utilisant les facteurs de sécurité et le recours à des méthodes sur cellules humaines (essais in vitro).

Les méthodes in vivo et in vitro utilisent toutes deux la substance chimique ; Dans le cas des étapes initiales de la conception de nouveaux médicaments ou de nouveaux matériaux, la substance chimique est seulement conçue, mais pas encore produite. D'autres méthodes qui ne nécessitent que les structures chimiques sont nécessaires, et elles sont collectivement nommé in silico.[58]

5. Méthodes in silico méthodes alternatives

Les méthodes in silico, également appelées méthodes informatiques, sont des approches utilisant des outils informatiques et des simulations numériques pour prédire, modéliser ou analyser des phénomènes biologiques, chimiques ou physiques. Ces méthodes tirent parti de la puissance de calcul des ordinateurs pour simuler des processus qui seraient difficiles ou coûteux à étudier expérimentalement. [59]

Les méthodes in silico les plus connues sont les méthodes de QSAR basées sur le principe que la structure moléculaire est responsable de toutes les activités et visent à trouver la dose ou la sous-structure responsable de l'activité

Utilisation de modèles informatiques pour établir des relations quantitatives entre la structure chimique des composés et leur activité biologique ou toxicologique. Ces modèles sont utilisés pour prédire les propriétés biologiques ou les effets toxiques de nouveaux composés sur la base de leur structure moléculaire. [60]

5.1. QSAR

Les modèles de relation structure-activité quantitative (QSAR) établissent une corrélation entre une variable de réponse spécifique et des descripteurs moléculaires, qu'ils soient calculés ou mesurés à partir des molécules elles-mêmes. Ces méthodes, développées par Corwin Hansch dans les années 1940, reposent sur l'équation QSAR suivante [61], [62] :

$$\textbf{Log } 1/C = a\,p + b\,s + c\,Es + \text{const} \qquad \textbf{(eq 1)}$$

- [C] : concentration de l'effet.
- [p] : coefficient de partage octanol-eau.
- [s] : constante de substitution de Hammett (électronique).
- [Es] : constante de substituant de Taft.0

Le coefficient de partage octanol-eau (Log 1/C) mesure la différence de solubilité d'un composé dans ces deux solvants. Une valeur élevée de (Log 1/C) indique que la substance est hydrophobe et se répartit préférentiellement dans des compartiments hydrophobes tels que les membranes cellulaires, tandis que l'hydrophile se trouve dans des compartiments hydrophiles comme le sérum sanguin. De nos jours, les valeurs de (Log 1/C) sont souvent prédites plutôt que mesurées. [63], [64]

Bien que les définitions de la structure chimique et de la fonction restent un défi, la relation entre la structure et la propriété est largement utilisée dans la découverte de médicaments et l'évaluation des risques. Les méthodes QSAR, parfois appelées QSPR (relations quantitatives structure-propriété), sont des modèles statistiques qui permettent de prédire les réponses pour des points de données invisibles sans nécessiter l'utilisation réelle du produit chimique, même avant sa synthèse.[65],[66]

5.1.1. Descripteurs Moléculaire

Le descripteur moléculaire est le résultat final d'une procédure logique et mathématique qui transforme l'information chimique chiffrée dans une représentation symbolique d'une molécule à un nombre utile ou le résultat de quelques expériences standard.

Les descripteurs moléculaires sont les traits communs les plus considérables de structure moléculaire qui peut être utilisée pour développer la « Relation Structure – Activité » avec le but de prédire l'activité biologique et propriétés physico-chimique des molécules. [67], [68]

5.1.2. Les types de descripteurs moléculaires

Actuellement, plus de 10 000 de ces descripteurs existent, chacun quantifiant des caractéristiques physico-chimiques ou structurelles propres aux molécules. Leur origine peut être empirique ou non empirique, mais privilégions les descripteurs calculés plutôt que mesurés. Pourquoi ? Parce qu'ils nous permettent de prédire sans avoir à synthétiser les molécules réelles, un objectif essentiel en modélisation.

Cependant, quelques descripteurs sont bel et bien mesurés. Ils correspondent généralement à des données expérimentales plus accessibles que la propriété ou l'activité que nous cherchons à prédire. Prenons l'exemple du coefficient de partage eau-octanol [69], de la polarisabilité ou encore du potentiel d'ionisation. Ces valeurs mesurées nous éclairent sur les propriétés des molécules.

Enfin, notons que les descripteurs moléculaires sont souvent classés en fonction de la dimensionnalité de la représentation moléculaire sur laquelle ils sont calculés. Ainsi, nous distinguons les descripteurs 1D, 2D et 3D ou 4D [70].

a- Les descripteurs 1-D

En effet, les descripteurs moléculaires 1D, également appelés descripteurs constitutionnels, sont des caractéristiques qui peuvent être calculées rapidement et facilement à partir de la formule brute d'une molécule. Ils fournissent des informations globales sur le composé, telles que sa composition atomique et sa masse molaire. Cependant, ils ne permettent pas de distinguer les isomères, qui ont la même formule brute mais des structures moléculaires différentes. Pour une analyse plus approfondie, il est souvent nécessaire d'utiliser des descripteurs 2D ou 3D qui tient compte de la connectivité atomique et de la géométrie moléculaire. [71]

b- Les descripteurs 2-D

Sont dérivés de la structure plane de la molécule. Ils contiennent des informations essentielles sur la connectivité atomique et certains fragments moléculaires, ainsi que des estimations des propriétés physico-chimiques. [72] Voici une exploration approfondie de ces descripteurs et de leurs principales catégories :

o Les indices topologiques .

o Les indices constitutionnels.

c- Les descripteurs 3-D

Les descripteurs moléculaires 3D sont des outils essentiels pour explorer les propriétés complexes des molécules. Leur calcul repose sur la géométrie tridimensionnelle de la molécule, obtenue par modélisation moléculaire empirique ou ab-initio.[73] Voici une exploration approfondie de ces descripteurs et de leurs principales catégories :

o Les descripteurs géométriques .

o Les descripteurs électrostatiques.

o Les descripteurs physico-chimiques.

En somme, les descripteurs moléculaires 3D, bien que coûteux en temps de calcul, apportent des informations cruciales pour la modélisation et la prédiction des propriétés chimiques.[74]

5.1.3. Construction de modèles de QSAR

- La construction de modèle (QSAR) commence par la préparation d'un ensemble de données contenant des informations expérimentales sur le point final d'intérêt. Cependant, cette étape est souvent complexe, car les ensembles de données disponibles dans la littérature sont rarement complets et validés. Dans de nombreux cas, les modèles QSAR sont élaborés à partir d'un nombre limité de données de haute qualité. Cependant, dans des applications réelles, il est essentiel de collecter des données à partir de sources primaires, telles que des études de laboratoire, et de les vérifier pour garantir leur cohérence et leur comparabilité.[75]
- En somme, la construction de modèles QSAR nécessite une approche rigoureuse pour garantir la fiabilité des résultats et leur applicabilité dans des contextes pratiques.[76]

- La prochaine étape implique le calcul et le choix des descripteurs. De nos jours, il est possible de calculer facilement des centaines ou des milliers de descripteurs 2D. Il est souvent préférable d'en calculer un grand nombre et ensuite de choisir les quelques descripteurs les plus pertinents en utilisant des méthodes simples comme les algorithmes génétiques, l'accumulation (incorporation d'un descripteur à la fois), la suppression d'une variable à la fois, ou des méthodes d'optimisation pour la sélection des variables.
- Peu importe l'algorithme utilisé pour créer des modèles prédictifs, il est crucial de prendre en considération les statistiques concernant la qualité des modèles et de garantir l'utilisation d'une méthodologie de modélisation adéquate afin d'éviter les débordements. Il existe de nombreuses statistiques sur les modèles qui peuvent donner une indication sur la possibilité d'appliquer de nouveaux points de données, à partir desquels les réponses doivent être prédictées, au modèle. [77]

Il existe deux types de modèles d'apprentissage supervisé qui sont intéressants :

• Les méthodes de classification qui attribuent la molécule cible à deux ou plusieurs classes, généralement biologiquement actives ou inactives ;

• Les méthodes de régression qui cherchent à utiliser des données continues, telles qu'une variable de réponse biologique mesurée, pour corréler les molécules avec ces données et prédire une valeur numérique continue pour de nouvelles molécules et inconnues en utilisant la génération modélisée.

On distingue également :

• les méthodes de similarité, qui cherchent à regrouper des molécules similaires ;

• les méthodes de caractéristiques, qui utilisent un ensemble de caractéristiques moléculaires (descripteurs) pour élaborer la fonction de classification.

Il existe de nombreux algorithmes et méthodes pour la modélisation, tels que la régression statistique (analyse discriminante linéaire, moindres carrés partiels et régression linéaire multiple) et les méthodes d'apprentissage (classificateur bayésien naïf, arbres de décision, partitionnement récursif, forêt aléatoire, réseaux neuronaux artificiels et machines à vecteurs de support). [78]-[79]

Dans le domaine du QSAR, de nombreuses méthodes dérivées de l'intelligence artificielle ont évolué au cours des deux dernières décennies, car elles ne sont pas paramétriques et permettent de trouver des fonctions de modélisation sans prendre de décisions préalables [80].

5.1.4. Interprétation du Modèle

L'analyse des modèles est un élément essentiel à considérer lors de la validation des modèles pour la prédiction dans un contexte réglementaire.

Les chercheurs tentent de saisir les modèles en se basant sur les principes fondamentaux connus. En général, on considère que le nombre limité de descripteurs utilisés et leur rôle dans des équations linéaires simples sont indispensables pour accepter les résultats QSAR.

Toutefois, si le but principal de QSAR est de prédire, il est essentiel de mettre l'accent sur la qualité du modèle plutôt que sur son interprétation. Le QSAR prédictif, qui se focalise sur la précision prédictive, est souvent distingué du QSAR descriptif (SAR), qui se focalise sur l'interprétation. [81], [82]

Selon Polishchuk, cette conception du « QSAR » et du « SAR » est en quelque sorte dépassée, car le concept d'interprétation des modèles QSAR a connu une évolution.

Les modèles statistiques prédictifs SAR et QSAR rencontrent des difficultés dans la théorie de l'apprentissage statistique. Les modèles les plus populaires possèdent de nombreuses fonctionnalités qui peuvent réaliser la tâche d'induire des modèles, dont la performance et la simplicité. [83], [84]

5.1.5. Stratégie globale d'une étude QSAR

La stratégie de développement d'un modèle Quantitative Structure-Activity Relationship (QSAR), et le modèle Quantitative Structure-Property Relationship (QSPR), s'articule autour de six étapes essentielles :

a. **Constitution de la Base de Données Structure-Propriété** :

- Collecter des mesures quantitatives, fiables et normalisées de la propriété cible pour chaque composé.
- Sélectionner des descripteurs chimiques pertinents en relation avec la propriété cible.

b. **Division de l'Ensemble de Données** :

- Séparer l'ensemble de données en un jeu d'apprentissage et un ensemble de test.

c. **Construction des Modèles**:

- Utiliser des outils statistiques pour construire des modèles à partir de l'ensemble d'apprentissage.
- Caractériser ces modèles en évaluant leurs indices de validation internes et vérifier leur robustesse par un test d'hasardisation.

d. **Validation des Modèles**:

- Tester les modèles avec l'ensemble de test.
- Calculer leur indice de corrélation externe pour évaluer leur performance.

e. **Répétition de la Division** :

- Répéter la division pour obtenir d'autres ensembles d'apprentissage et de test.
- Chercher la division optimale qui permet d'obtenir un petit ensemble d'apprentissage capable de prédire efficacement sur un grand ensemble de test.

f. **Exploration et Exploitation des Modèles Validés** :

- Analyser les modèles validés pour comprendre les mécanismes sous-jacents.
- Utiliser les modèles pour faire des prévisions.

En somme, cette approche méthodique permet de développer des modèles QSAR robustes et informatifs pour la prédiction des propriétés chimiques.

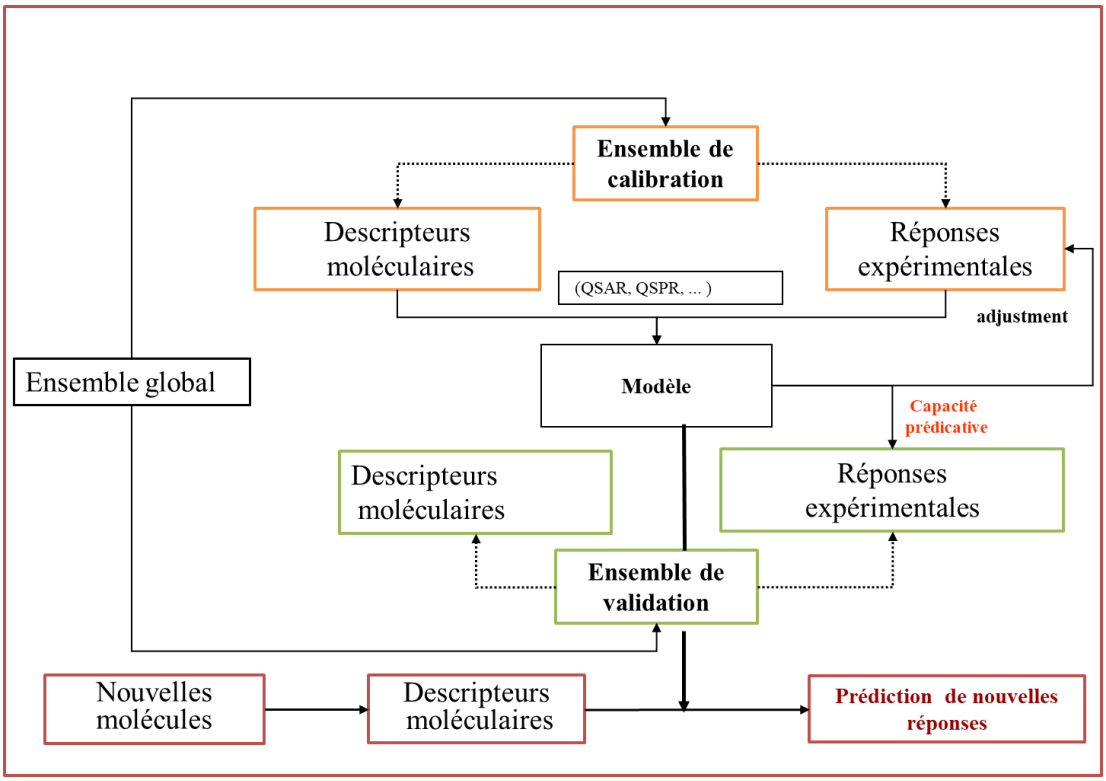


Figure 7 : Stratégie globale d'une étude QSAR [85]

6. Conclusion

En conclusion, les méthodes QSAR se révèlent être des outils précieux pour la recherche scientifique, en particulier dans les domaines de la découverte de médicaments, de la toxicologie et de la chimie.

Leur capacité à réaliser des simulations complexes, à faciliter la conception moléculaire, à faire des prédictions précises et à expliquer les mécanismes d'action des substances en fait des outils indispensables pour les scientifiques.

De plus, les modèles QSAR constituent une méthode inestimable pour filtrer et analyser de grandes quantités de données, ce qui les rend particulièrement utiles dans le cadre de recherches à grande échelle.

Il est important de noter que les méthodes QSAR ne remplacent pas entièrement les techniques expérimentales classiques, mais plutôt qu'elles les complètent et les enrichissent.

En combinant les approches expérimentales et computationnelles, les chercheurs peuvent obtenir une compréhension plus complète et plus précise des relations entre la structure chimique et l'activité biologique des composés.

Alors que les méthodes QSAR continuent d'évoluer et de se perfectionner, elles sont appelées à jouer un rôle encore plus important dans la recherche scientifique et dans le développement de nouveaux produits.

Partie II : Etude Expérimentale

Matériels et méthodes

1. Matériels et méthodes

1.1 Les composés nitro-aromatiques

Les composés nitro-aromatiques sont des composés organiques comportant un groupe nitro (-NO2) attaché à un noyau aromatique. Le groupe nitro confère des propriétés spécifiques à ces composés, les rendant précieux dans divers domaines comme la pharmacologie, la chimie des explosifs et la chimie des colorants. La présence du groupe nitro leur confère une certaine réactivité, permettant une variété de transformations chimiques. [86], [87], [88]

1.1.1 Stratégies de synthèse des composés nitro-aromatiques

La synthèse des composés nitro-aromatiques implique généralement des réactions de nitration, où un groupe nitro (-NO2) est introduit sur un noyau aromatique. Une des méthodes les plus couramment utilisées est la nitration électrophilique, qui consiste à traiter un composé aromatique avec un mélange d'acide nitrique concentré et d'acide sulfurique. Cette réaction conduit à la formation du composé nitro-aromatique souhaité en tant que produit principal. Cependant, cette méthode peut être limitée par la formation d'un mélange de produits et la génération de sous-produits indésirables. Pour surmonter ces limitations, des techniques de contrôle de sélectivité et des catalyseurs spécifiques ont été développés pour favoriser la formation du produit souhaité. D'autres méthodes de synthèse des composés nitro-aromatiques incluent la réduction de composés nitroso, la réaction de Sandmeyer et d'autres réactions de substitution aromatique. Chaque méthode présente ses propres avantages et limitations, et le choix de la stratégie de synthèse dépend souvent des exigences spécifiques du composé cible et des conditions réactionnelles. La littérature scientifique regorge de nombreuses études décrivant ces différentes stratégies de synthèse des composés nitro-aromatiques, offrant ainsi un vaste ensemble de méthodes disponibles pour les chercheurs et les chimistes synthétiques. [89], [90], [91]

1.1.2. Le profile pharmacologique des composés nitro-aromatiques

Les composés nitro-aromatiques ont suscité un intérêt croissant en raison de leur profil pharmacologique diversifié et de leur potentiel dans diverses applications thérapeutiques. Leur activité pharmacologique a été étudiée dans le cadre de différentes maladies, notamment le cancer, les infections bactériennes et les troubles inflammatoires. Par exemple, des études ont exploré l'activité anticancéreuse d'un composé nitro-aromatique sur des lignées cellulaires de cancer du poumon, mettant en évidence son efficacité dans l'inhibition de la croissance tumorale in vitro. De

même, d'autres travaux ont examiné l'activité antibactérienne d'un dérivé nitro-aromatique contre des souches bactériennes résistantes aux antibiotiques, révélant son potentiel en tant qu'agent antibactérien prometteur. Ces études démontrent le large spectre d'activités pharmacologiques des composés nitro-aromatiques et leur importance potentielle dans le développement de nouvelles thérapies médicamenteuses. [92], [93]

1.1.3 Les mécanismes d'action des mutagènes chimiques

Les mécanismes d'action des mutagènes chimiques se réfèrent aux processus par lesquels ces substances induisent des mutations génétiques, altérant ainsi le matériel génétique des cellules. Ces mécanismes peuvent inclure l'interaction directe avec l'ADN, la formation de liaisons covalentes avec des bases nucléiques, la génération de radicaux libres ou d'autres espèces réactives, et la perturbation des processus cellulaires normaux de réparation de l'ADN. Ces altérations génétiques peuvent conduire à des mutations ponctuelles, des délétions, des insertions ou d'autres modifications de l'ADN, pouvant entraîner des effets mutagènes et potentiellement carcinogènes. [94]

Les composés nitroaromatiques subissent un processus métabolique complexe pour former des métabolites toxiques qui peuvent induire des effets mutagènes et potentiellement cancérogènes. Le mécanisme implique généralement une réduction initiale du groupe nitro pour former un radical nitro-anion stabilisé. Ce radical peut réagir avec l'oxygène moléculaire pour produire des espèces réactives de l'oxygène, telles que l'anion superoxyde. En milieu anaérobie, le radical nitro-anion peut être réduit davantage pour former des espèces telles que les nitroso, hydroxylamine et amine. Ces métabolites réactifs peuvent interagir avec l'ADN et d'autres biomolécules, entraînant des dommages génétiques et des altérations cellulaires. [95]

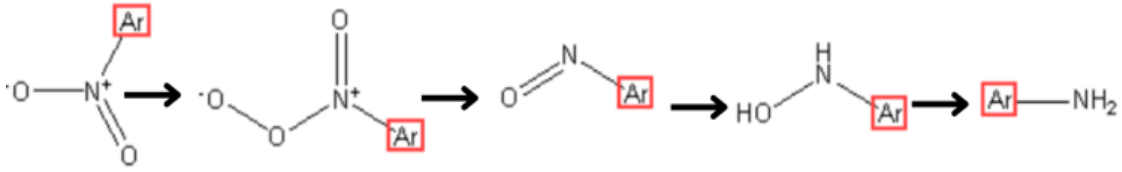


Figure 8 : Voies d'activation métabolique des nitroarènes. Ar = aryle [95]

1.2 Test d'Ames

1.2.1 Définition du test d'Ames

Le test Microbien Ames est un test bactérien simple, rapide et robuste composé de différentes souches et applications de *Salmonella typhimurium/E. coli*, utilisé pour vérifier le potentiel mutagène. En 1975, Ames et ses partisans ont standardisé le protocole de test traditionnel d'Ames et l'ont réévalué dans les années 1980 (Maron et Ames, 1983). L'induction de nouvelles mutations remplaçant les mutations existantes permet de restaurer la fonction des gènes. Les cellules mutantes nouvellement formées peuvent se développer en l'absence d'histidine et former des colonies, c'est pourquoi ce test est également appelé « test de réversion » (Ames, 1971). Alors que le test d'Ames traditionnel est assez laborieux et prend du temps pour la surveillance initiale des composés mutagènes, la miniaturisation de la suspension liquide a eu un impact significatif sur la facilité d'utilisation en la rendant plus pratique. Les doses standard (2 µl, 5 µl, 10 µl, 50 µl et 100 µl) ont été définies pour évaluer la mutagénicité d'une concentration inférieure à une concentration supérieure (Hayes, 1982). [96]

1.2.2 Protocole d'activité du teste d'Ames et origine des données

Le test d'incorporation sur plaque standard de Salmonella de Maron et Ames (1983) a été utilisé pour déterminer la mutagénicité des produits chimiques testés. Deux souches de Salmonella typhimuriurn, TA98 et TA100, obtenues auprès de Bruce N. Ames (Berkeley, CA) ont été utilisées pour mesurer l'induction de mutations rétroactives. Les enzymes microsomales et cytosoliques du foie (fraction S9) de rats induits par l'Aroclor. Les contrôles positifs comprenaient les mutagènes méthanesulfonate de méthyle et 2,4,7- trinitro-9-fluorénone pour les souches d'essai TA100. Les promutagènes 7,12-diméthylbenz[a]anthracène et 2-aminofluorène ont également été utilisés pour déterminer l'activité de la préparation S9. Tous les composés testés ont été dissous dans du sulfoxyde de diméthyle et testés avec 3 plaques par niveau de dose à 4 doses ou plus dans au moins 3 déterminations indépendantes. Les doses allaient de 1 à 1000 µg/plaque dissoutes dans 100 µl de DMSO et ajoutées à la gélose supérieure contenant 0,1 ml d'une culture d'une nuit (8 h) de la souche testée et 0,2 ml de la solution de 0,5 mM d'histidine-0,5 mM de biotine (Maron et Ames, 1983). 500 µl du mélange standard S9 contenant 20 µl de la fraction S9 (32 mg de protéines/ml) ont été ajoutés juste avant de verser le mélange de gélose supérieur sur la plaque de gélose à milieu minimal. Le nombre de colonies révertantes a été noté après 48 heures d'incubation. [96], [97]

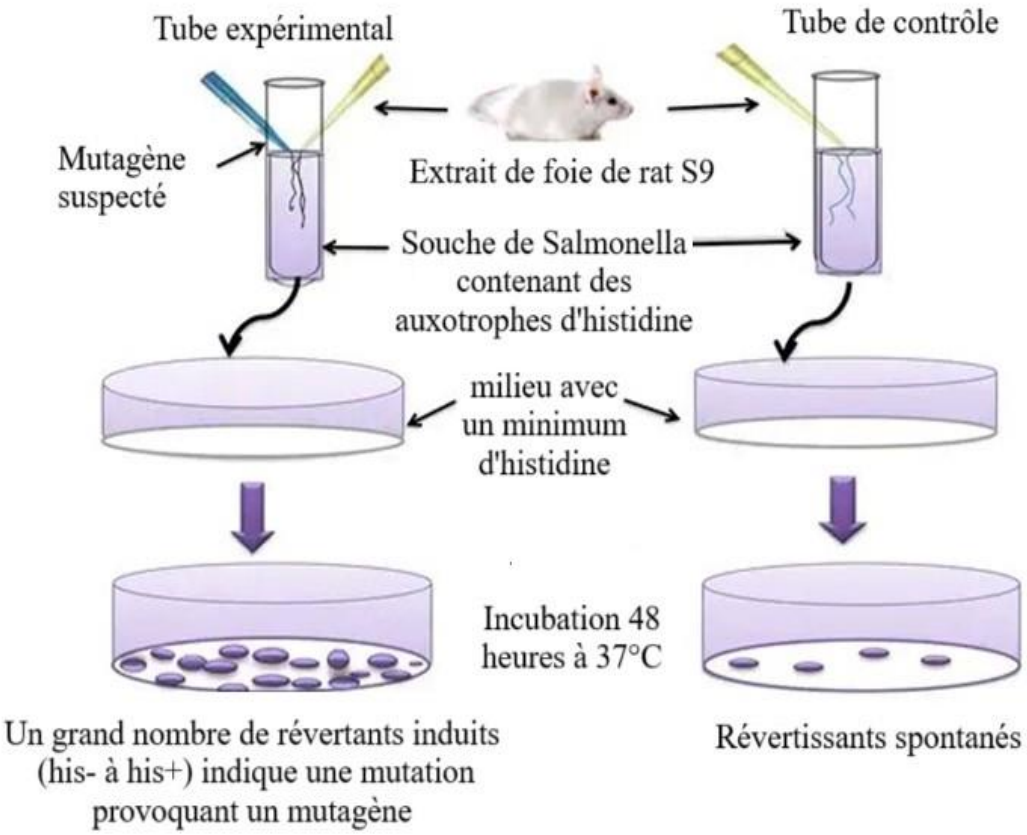


Figure 9: Protocole d'activité du teste d'AMES. [98]

1.3. Traitements des données

Le développement du modèle QSAR nécessite le passage de notre base de composés chimiques par trois étapes :

- Etape 1 Préparation de l'ensemble de données.
- Etape 2 : Calcule des descripteurs.
- Etape 3 : La modélisation.

En utilisant plusieurs logiciels de traitement :

- **ChemDraw** : C'est une solution complète pour les chimistes et les biologistes, intégrant une variété d'outils intelligents. Il s'est imposé comme une référence pour le dessin des structures des composés chimiques. Il est simple à utiliser, puissant et permet de dessiner de manière intuitive et efficace en deux et trois dimensions. [99], [100]

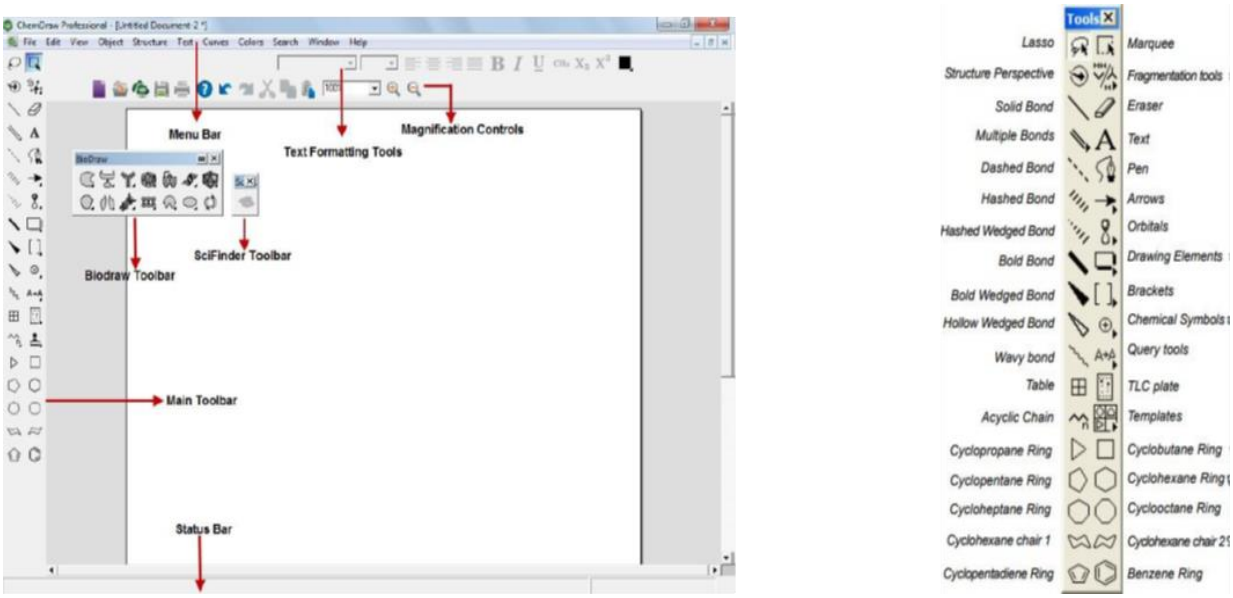


FIGURE 10 : le programme ChemDraw [99], [100]

- **HyperChem** : C'est un environnement sophistiqué de modélisation moléculaire, reconnu pour sa qualité, sa flexibilité et sa facilité d'utilisation. Il offre différentes méthodes d'optimisation (PM3, MM+, AM1, etc.). En optimisant la molécule, les descripteurs moléculaires tels que le volume moléculaire, les énergies HUMO et LUMO, etc. peuvent être calculés. Les descripteurs moléculaires théoriques ont été calculés en suivant ce processus : les structures moléculaires ont été pré-optimisées par le champ de force de la mécanique moléculaire MM+ issu de ce logiciel de modélisation moléculaire. [101], [102]

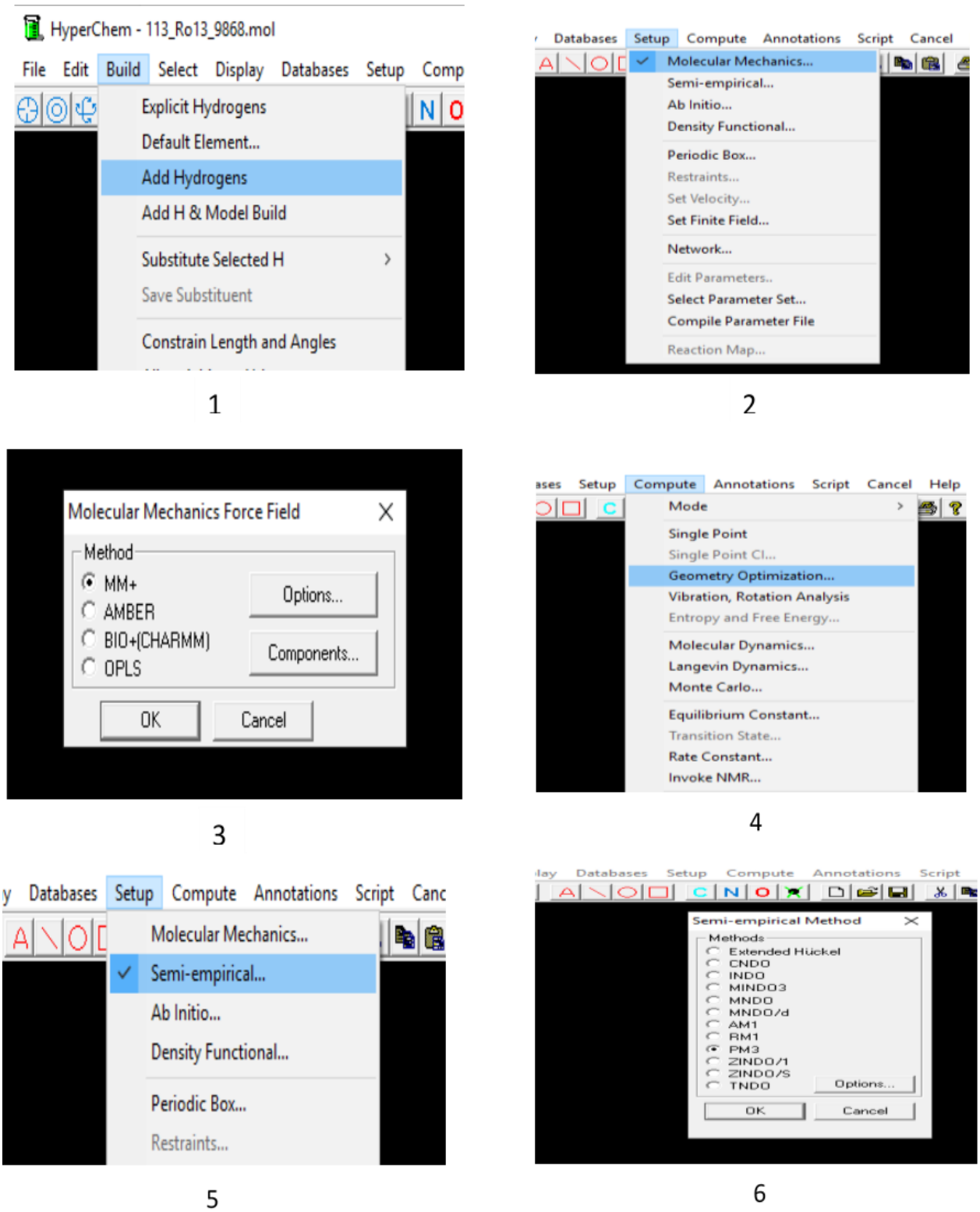


FIGURE 11 : le programme HyperChem [101], [102]

- **AlvaDesk :** est un logiciel puissant conçu pour le calcul et l'analyse des descripteurs moléculaires et des empreintes digitales. Il offre une gamme complète de fonctionnalités pour explorer les représentations mathématiques des produits chimiques, permettant aux chercheurs et aux scientifiques de construire des modèles prédictifs pour les propriétés chimiques. AlvaDesc se distingue par sa capacité à gérer des structures moléculaires complexes, telles que les sels, les mélanges, les liquides ioniques et les complexes métalliques, offrant ainsi une flexibilité d'analyse inégalée. En plus du calcul des descripteurs et des empreintes digitales, AlvaDesc propose des outils avancés pour explorer les ensembles de données chimiques, tels que la vérification de la structure moléculaire, la visualisation des structures, l'analyse en composantes principales (PCA) et l'analyse de corrélation. [103]

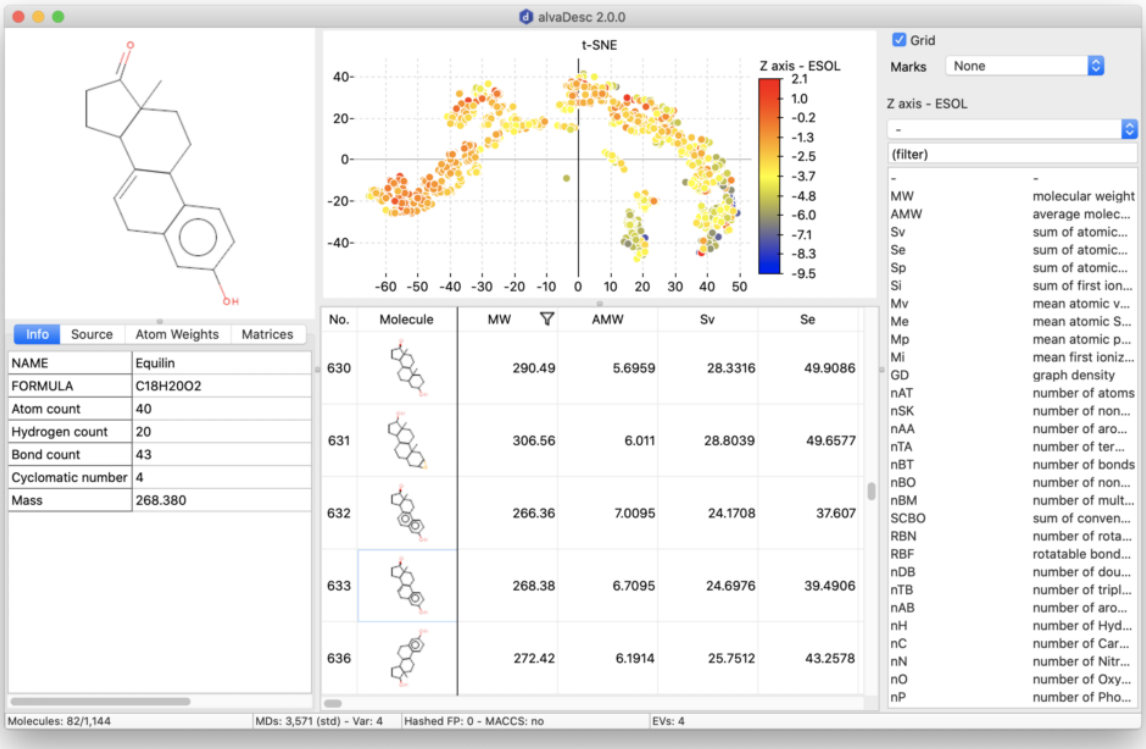


FIGURE 12 : le programme AlvaDesc [103]

- **QSARINS (QSAR-INSUBRIA)** : est un logiciel récent qui permet de créer et de valider des modèles de régression linéaire multiple (MLR) en utilisant les moindres carrés ordinaires (MCO) et un algorithme génétique pour choisir les variables. L'objectif principal de ce programme est de

valider les modèles QSAR. On met en place la réduction des descripteurs moléculaires d'entrée, la division de l'ensemble de données en ensembles d'apprentissage et de prédiction, la détection de valeurs aberrantes et de prédictions interpolées ou extrapolées, la validation interne et externe par divers paramètres, la modélisation consensuelle et différents graphiques pour les visualisations. QSARINS est une plateforme facile à utiliser pour la modélisation QSAR conformément aux Principes de l'OCDE et pour évaluer la fiabilité des données prédictives recueillies. [104], [105]

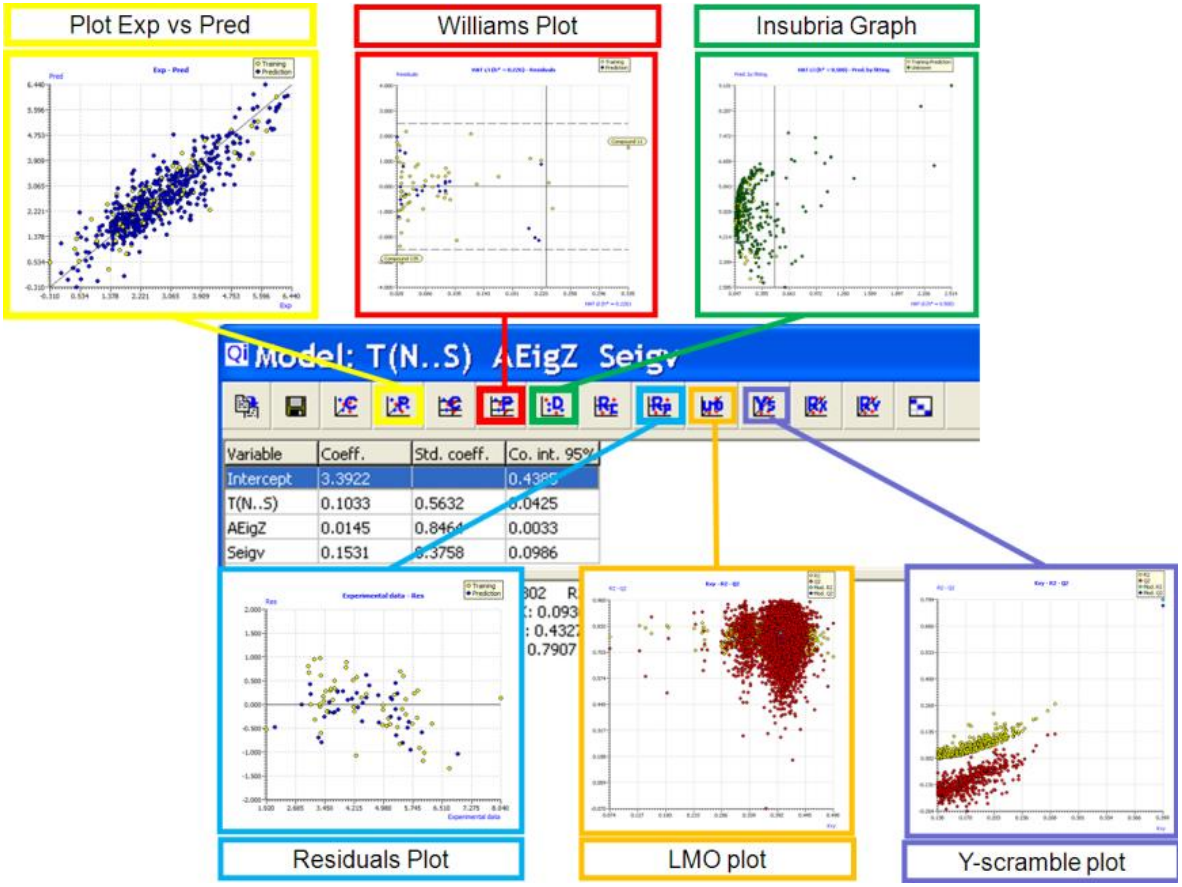


FIGURE 13 : le programme QSARINS [104], [105]

1.4. Plate-forme KNIME pour la modélisation QSAR

KNIME [106] (Konstanz Information Miner) est une technologie de flux de travail open-source avec une interface utilisateur graphique basée sur des collections de noeuds connus sous le nom d'« extensions » qui permettent le traitement des données et leur transport via des connexions entre ces nœuds [107]. KNIME fournit un atelier d'assemblage visuel facile qui permet aux scientifiques de créer et de visualiser facilement des flux de travail complexes [108]. En outre, il

ne se limite pas à la capacité d'analyser les résultats, il peut donc comprendre plusieurs étapes de traitement et d'analyse, y compris l'analyse statistique, la visualisation des données et l'exploration des données sur le plan expérimental [109]. Il prend également en charge un large éventail de fonctionnalités et dispose d'une communauté active dans le domaine de la chimie et de la bioinformatique. Puisque l'analyse manuelle de données volumineuses demande un coût et un temps considérables, il est devenu indispensable d'améliorer le flux de travail, tels que l'utilisation de KNIME pour les procédures de curation des structures chimiques, ou la création de modèles de flux de travail basés sur plusieurs nœuds de la communauté chimique, afin de calculer et prédire les propriétés physicochimiques et les activités biologiques.

1.4.1. Atelier KNIME (The KNIME Workbench)

Le Workbench apparaît après le lancement de KNIME. Les principales sections de l'atelier KNIME sont illustrées à la figure 10 ; une brève description de chacune d'entre elles est fournie ci-dessous. Pour plus de détails, voir le Guide de l'atelier KNIME [106].

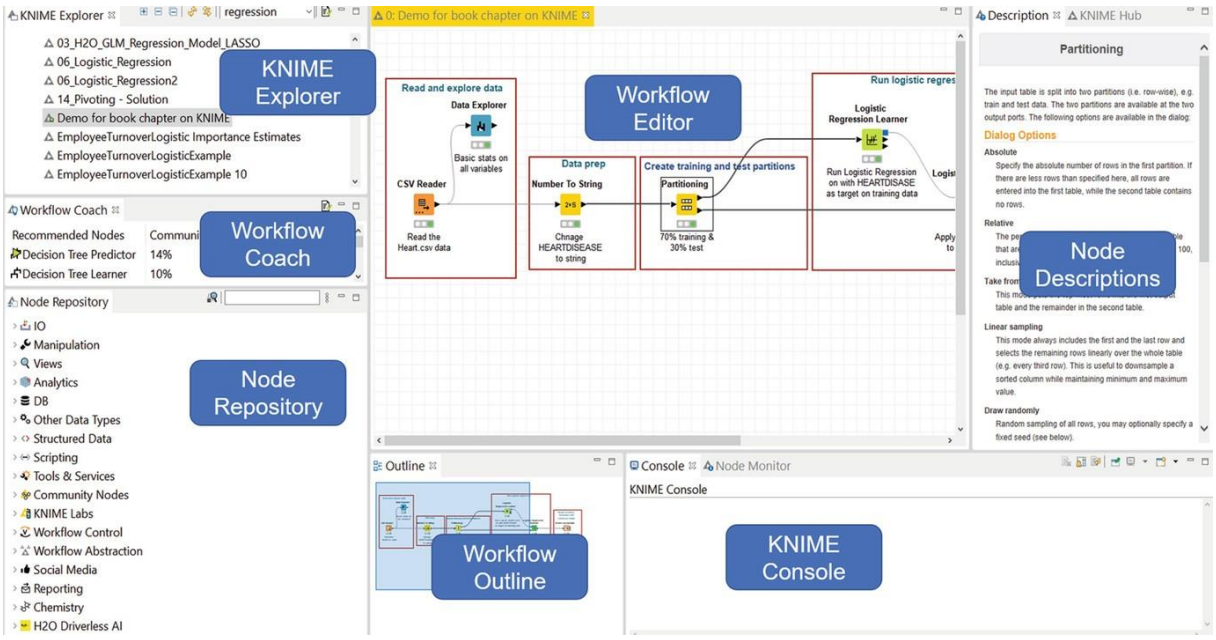


Figure 14: L'atelier KNIME(The KNIME Workbench).[110]

1.4.2 Eléments de l'atelier KNIME

KNIME Explorer : Ce lien permet d'accéder aux workflows disponibles sur votre machine et à ceux disponibles sur les serveurs KNIME, y compris les exemples de workflows de KNIME et de la communauté KNIME.

Workflow Coach : Il s'agit d'un outil pratique pour vous aider à construire un flux de travail analytique. Le Coach suggère de connecter des nœuds et des processus à n'importe quel nœud sélectionné dans votre flux de travail. Les suggestions du Coach sont basées sur des workflows réels construits par des utilisateurs de la Communauté KNIME.

Node Repository : Les nœuds d'analyse installés sur votre machine sont répertoriés dans cette zone. Les nœuds sont disponibles pour lire et écrire des fichiers, explorer et transformer des données, exécuter des analyses de base et avancées, et créer des visualisations. Un ensemble de nœuds de base est inclus lors de l'installation de KNIME, mais des milliers de nœuds supplémentaires sont disponibles et faciles à installer. Les nœuds sont organisés par catégories, mais vous pouvez également utiliser le champ de recherche en haut du référentiel de nœuds pour trouver des nœuds.

Workflow Editor : Il s'agit du canevas permettant de modifier le flux de travail actuellement actif. Le flux de travail est créé en faisant glisser des nœuds depuis le référentiel de nœuds et en les reliant de manière interactive.

Workflow Outline : Cette zone affiche une petite vue d'ensemble du flux de travail actuel.

KNIME Console : Elle montre le traitement qui a lieu lors de l'exécution d'un flux de travail. Elle fournit également des avertissements et des messages d'erreur.

Node Descriptions : Pour chaque nœud sélectionné dans un flux de travail, une description détaillée est fournie, y compris la fonction générale du nœud, les paramètres disponibles et les ports d'entrée et de sortie.

1.5. Machine Learning pour la classification (Méthodes d'apprentissage automatique)

La modélisation computationnelle, incluant les méthodes d'apprentissage automatique (Machine Learning), occupe une place de plus en plus importante dans le domaine de la chimioinformatique, notamment dans la modélisation QSAR. L'analyse QSAR contemporaine

bénéficie de plusieurs algorithmes d'apprentissage automatique sophistiqués. En général, le rôle de l'apprentissage automatique consiste à extraire les caractéristiques les plus importantes en analysant les combinaisons de descripteurs. Diverses méthodes d'apprentissage automatique ont déjà été employées dans différents domaines de la modélisation QSAR, telles que les forêts aléatoires (Random Forest), le boosting par gradient (Gradient Boosting) et la régression logistique (Logistic Regression) [111].

1.5.1. Arbre aléatoire (*forêts aléatoires* RF)

Est une méthode d'apprentissage automatique puissante et polyvalente largement utilisée dans le domaine de la modélisation QSAR. Il s'agit d'une technique d'ensemble qui combine les prédictions de plusieurs arbres de décision générés de manière aléatoire pour produire une prédiction finale plus précise et robuste. RF a démontré sa capacité à exceller dans diverses tâches de modélisation QSAR, notamment la classification de la toxicité [112].

- **Principes *fondamentaux de RF***

RF fonctionne en créant une forêt d'arbres de décision, où chaque arbre est formé sur un sous-ensemble aléatoire des données d'entraînement. Cette approche d'échantillonnage aléatoire permet de réduire le sur-apprentissage et d'améliorer la généralisation du modèle. De plus, RF utilise la technique du "bagging" pour combiner les prédictions des arbres individuels. Lors du bagging, chaque point de données est classé par tous les arbres de la forêt, et la classe majoritaire est attribuée comme prédiction finale [112].

1.5.*2* Gradient Boosted Trees (GBoost)

Gradient Boosted Trees (GBoost) est une méthode d'apprentissage automatique puissante et largement utilisée pour la modélisation QSAR. Il s'agit d'une technique d'apprentissage par ensemble qui construit un modèle prédictif en séquence, où chaque nouvel arbre de décision est formé pour corriger les erreurs de prédiction des arbres précédents. Cette approche itérative permet d'améliorer progressivement les performances du modèle et d'obtenir des prédictions plus précises et robustes [113].

- **Principes fondamentaux de GBoost**

GBoost fonctionne en créant une séquence d'arbres de décision, où chaque arbre est formé sur un sous-ensemble résiduel des données d'entraînement. Le sous-ensemble résiduel est constitué des points de données que les arbres précédents ont mal prédits. L'objectif de chaque nouvel arbre

est de minimiser l'erreur de prédiction globale en se concentrant sur les points de données les plus difficiles à prédire [113].

1.5.3. Naïve Bayes (NBayes)

Est un algorithme de classification probabiliste simple et largement utilisé dans le domaine de la modélisation QSAR. Il est basé sur le théorème de Bayes et suppose que les caractéristiques des données d'entraînement sont indépendantes les unes des autres. Bien que cette hypothèse soit souvent simplifiée dans la pratique [114].

- **Principes fondamentaux de NBayes**

NBayes fonctionne en calculant la probabilité qu'un point de données donné appartienne à chaque classe possible. Cette probabilité est calculée en utilisant le théorème de Bayes et en supposant que les caractéristiques sont indépendantes. Pour chaque classe, NBayes calcule la probabilité conjointe de toutes les caractéristiques étant donné la classe, puis normalise les probabilités pour obtenir la distribution finale des classes [114].

1.5.4. Régression Logistique (LRegression)

Contrairement à la régression linéaire, qui prédit des valeurs continues, LRegression modélise la probabilité qu'une observation donnée appartienne à une classe particulière, généralement binaire (par exemple, actif/inactif, toxique/non toxique). Cette approche probabiliste en fait un outil précieux pour diverses tâches de classification QSAR, notamment la classification de la toxicité, la prédiction de l'activité biologique et l'étude de la relation structure-activité (SAR) [115].

- **Principes fondamentaux de LRegression**

LRegression utilise une fonction logistique, également connue sous le nom de fonction sigmoïde, pour mapper les valeurs d'entrée (descripteurs moléculaires) à une probabilité de sortie (appartenance à une classe). La fonction logistique transforme les valeurs d'entrée entre 0 et 1, représentant la probabilité que l'observation appartienne à la classe positive. LRegression ajuste ensuite les paramètres de la fonction logistique en utilisant une méthode d'optimisation, telle que la descente du gradient, pour minimiser l'erreur de prédiction sur les données d'entraînement [115].

1.5.5 Les arbres de décision (DTree)

Sont une méthode d'apprentissage automatique populaire et intuitive largement utilisée dans le domaine de la modélisation QSAR. Ils fonctionnent en construisant une arborescence hiérarchique où chaque nœud représente une décision basée sur une caractéristique des données

d'entraînement. Les branches issues de chaque nœud représentent les différentes valeurs possibles pour cette caractéristique, et chaque branche mène à un nœud enfant suivant. Le processus se poursuit jusqu'à ce qu'une feuille soit atteinte, qui représente la classe prédite pour l'observation donnée [116].

- **Principes fondamentaux des arbres de décision**

La construction d'un arbre de décision implique un processus itératif de sélection de la caractéristique la plus discriminante à chaque nœud. Cette sélection se fait généralement à l'aide de mesures d'information, telles que le gain d'information ou le gain de Gini, qui quantifient la réduction de l'incertitude apportée par la prise en compte de cette caractéristique. L'arbre se développe ensuite en divisant les données d'entraînement en fonction des valeurs de la caractéristique sélectionnée, et le processus se répète jusqu'à ce que les critères d'arrêt soient atteints, tels qu'une profondeur maximale d'arbre ou un nombre minimum d'observations par nœud [116].

1.5.6 paramètres d'évaluation pour de classification

En règle générale, les modèles qualitatifs sont évalués à l'aide de la statistique de Cooper. Dans le cas simple d'une classification il y a deux classes, comme toxique (positif) ou non (négatif). Les résultats d'un peuvent donc être regroupés en quatre catégories : les composés toxiques prédits comme toxiques (True Positive ou TP) ou comme non toxiques (False Negative ou FN) ainsi que les composés non toxiques prédits comme non toxiques (vrai négatif ou TN) ou toxiques (faux positif ou FP). Ces quatre classes sont généralement représentées dans la matrice dite de confusion [117].

- **La sensibilité:** utilisée pour évaluer la performance d'un test de classification binaire. Elle indique la proportion des instances positives qui sont correctement identifiées par le modèle. La sensibilité est particulièrement utile pour comprendre la capacité du modèle à détecter les instances positives.[118]

$$\text{Sensibilité} = \frac{VP}{VP+FN} \qquad \textbf{(eq 1)}$$

- **La spécificité :** pour déterminer sa capacité à identifier correctement les instances négatives. Elle est définie comme le rapport entre le nombre de vrais négatifs (VN) et le nombre total de véritables négatifs (VN + faux positifs (FP)). En d'autres termes, la spécificité mesure la proportion des exemples négatifs qui sont correctement identifiés par le modèle.[119]

$$\text{Spécificité} = \frac{VN}{VN+FP} \qquad \textbf{(eq 2)}$$

- **Le rappel (Recall) :** Le rappel est défini comme le rapport entre le nombre de vrais positifs (VP) et le nombre total de véritables positifs (VP + faux négatifs (FN)). En d'autres termes, il mesure la proportion des exemples positifs qui sont correctement identifiés par le modèle.[120]

$$\text{Recall} = \frac{VP}{VP+FN} \qquad \textbf{(eq 3)}$$

- **La précision :** La précision, également appelée valeur prédictive positive, Elle indique la proportion des instances identifiées comme positives par le modèle qui sont réellement positives. La précision est particulièrement utile pour comprendre la qualité des prédictions positives faites par le modèle.[121]

$$\text{Précision} = \frac{VP}{VP+FP} \qquad \textbf{(eq 4)}$$

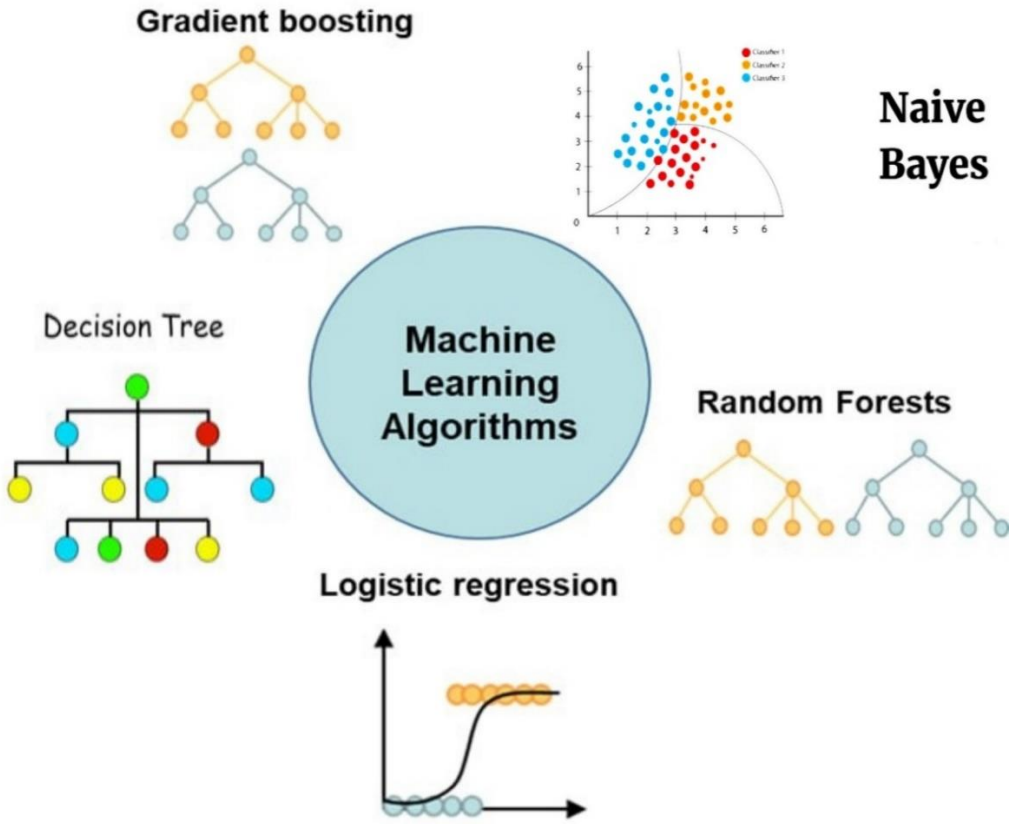


Figure 15 : Algorithmes d'apprentissage automatique : Un aperçu complet des différentes techniques [122]

1.6 Méthodes statistiques des modèles

L'analyse statistique vise précisément à « diviser » ces descripteurs et à déterminer ceux qui sont liés à la variable cible, qui génèrent du signal, et ceux qui ne le sont pas, qui génèrent du bruit. De plus, l'analyse statistique permet de repérer les descripteurs qui sont liés les uns aux autres afin de ne conserver que les principaux et de diminuer la redondance d'informations [123].

Les corrélations entre les descripteurs et la variable cible sont déterminées et mesurées par l'analyse statistique. Elle mentionne aussi l'apport relatif de chaque descripteur dans l'explication globale de l'exploitation. La valeur de la variable cible est calculée en fonction de la somme des valeurs pondérées des descripteurs dans le modèle statistique [123].

❖La régression linéaire multivariée (Multivariate Linear Regression - MLR) est l'un des principaux outils statistiques utilisés pour obtenir un modèle.

1.6.1. La Régression Linéaire multiple MLR

La méthode statistique de modélisation la plus simple et la plus utilisée dans les études de la relation structure-activité est la régression linéaire multiple (MLR). Hansch a popularisé cette méthode en associant l'activité biologique aux caractéristiques expérimentales de la lipophile, de l'électronicité et de la stérilité pour des séries de composés. [124]

La relation linéaire entre une variable dépendante Y (Activité) et des variables indépendantes X (descripteurs moléculaires) est modélisée par la MLR. La méthode des moindres carrés (ordinary least squares) est utilisée pour la MLR : le modèle est modifié de manière à réduire la somme des carrés entre les différentes valeurs réelles et prédites. Les coefficients de régression (R2) sont estimés par MLR en utilisant l'ajustement des moindres carrés. [126]

L'expression de régression prend la forme :

$$Y = a_0 + \sum_{i=1}^{n} a_i X_i \qquad \textbf{(eq 5)}$$

Avec :

Y: la variable dépendante (Activité);

xi: les variables indépendantes (descripteur);

n : le nombre de variables ;

a0: la constante de l'équation du modèle;

ai: les coefficients de descripteurs dans l'équation du modèle

1.6.2. Paramètres d'évaluation des modèles

Dans les études statistiques, la validation des modèles QSAR demeure une étape extrêmement cruciale pour évaluer leur importance. En tant que résultat d'une analyse statistique, il est essentiel d'interpréter et d'appliquer un modèle dans le cadre très spécifique du domaine étudié. [126]

On utilise plusieurs paramètres statistiques pour évaluer la qualité d'un modèle, comme :

• Le coefficient de détermination multiple qui représente le degré de corrélation entre les données de la propriété observées et prévues de l'ensemble de test.

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \textbf{(eq 6)}$$

- Où Yi et $Y\ i$ sont les données de la propriété observée et prédite pour les composés de test,
- Est la valeur moyenne des valeurs observées pour l'ensemble de calibrage.

Les modèles qui ont des valeurs de R2 supérieures à 0,5 sont considérés comme étant très prédictifs.

En général, on examine la stabilité du modèle QSAR publié en utilisant : • "La validation croisée par omission d'une observation" (LOO : cross-validation by leave-one-out) : qui implique de recalculer le modèle sur (n - 1) composés de calibrage. Le modèle obtenu servira à estimer la valeur de la propriété du composé éliminé noté ŷ(i). Le même procédé est répété pour chaque n composé de l'ensemble de calibrage. [126], [127]

On l'utilise pour définir le coefficient de prédiction:

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \qquad \textbf{(eq 7)}$$

À la différence de R^2 qui augmente en fonction du nombre de paramètres du modèle, le facteur Q^2_{LOO} présente une courbe obtenue pour un certain nombre de descripteurs, puis diminue de manière perpétuelle. Le coefficient Q^2_{LOO} est donc d'une grande importance.

Une valeur Q^2_{LOO} > 0,5 est jugée satisfaisante, tandis qu'une valeur supérieure à 0,9 est considérée comme excellente. [128]

Si nous disposons de données adéquates qui n'ont pas été utilisées dans la création du modèle ou après avoir recueilli de nouvelles données, il est possible ou nécessaire de procéder à une validation externe.

Notée que Q^2_{ext}, est calculée comme suit :

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_{(i)})^2 \Big/ n_{ext}}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2 \Big/ n_{tr}} \qquad \textbf{(eq 8)}$$

Si la valeur de Q^2_{LOO}, est élevée, une valeur élevée de Q^2_{ext} suggère une capacité prédictive élevée du modèle.

Depuis dix ans, une grande controverse a été soulevée concernant l'utilisation de la validation interne ou externe pour garantir la prédiction des modèles [129]. La validation interne vise principalement à vérifier la stabilité des modèles, mais ce type de validation ne peut pas être utilisé pour une validation externe réelle. [130]

- **Les conditions de Golbraikh et Tropsha**

Selon les résultats, il a été démontré que le pouvoir prédictif des modèles QSAR ne peut être attribué que si le modèle a été efficacement appliqué pour prédire les composés de l'ensemble de test externe (les composés qui n'ont pas été utilisés lors de la formation du modèle). Diverses mesures sont mises en place pour évaluer la qualité des prévisions, en se basant sur les critères de Golbraikh et de Trospsha [131], [132] ont été proposées [126], [131], [133], comme Q^2_{F1}, Q^2_{F2}, Q^2_{F3} et r^2_m qui sont détaillées par les équations suivantes :

$$\boldsymbol{Q^2_{F1}} = 1 - \frac{\sum(\boldsymbol{Y_{obs(test)} - Y_{pred(test)}})\,\boldsymbol{2}}{\sum(\boldsymbol{Y_{obs(test)} - \bar{Y}_{training}})\,\boldsymbol{2}} \qquad \textbf{(eq 9)}$$

Q^2_{F2} Proposé par Schuurmann *et al*. [170] , donnée par:

$$\boldsymbol{Q^2_{F2}} = 1 - \frac{\sum(\boldsymbol{Y_{obs(test)} - Y_{pred(test)}})\,\boldsymbol{2}}{\sum(\boldsymbol{Y_{obs(test)} - \bar{Y}_{test}})\,\boldsymbol{2}} \qquad \textbf{(eq 10)}$$

$$\boldsymbol{Q^2_{F3}} = 1 - \frac{\sum(\boldsymbol{Y_{obs(test)} - Y_{pred(test)}})\,\boldsymbol{2/n_{test}}}{\sum(\boldsymbol{Y_{obs(train)} - \bar{Y}_{train}})\,\boldsymbol{2/n_{test}}} \qquad \textbf{(eq 11)}$$

$$\boldsymbol{r^2_m} = r^2.\,(1-\sqrt{(\boldsymbol{r^2 + r^2_0})} \qquad \textbf{(eq 12)}$$

Selon ces auteurs, les modèles sont considérés comme satisfaisants, si l'ensemble des conditions suivantes sont simultanément réalisées : $R^2 > 0.6$, $Q^2 > 0.5$.

Résultats et discutions

2. Résultat et discutions

2.1. Source des données de la méthode de régression

Avant le développement du modèle, il était extrêmement important de nettoyer les données pour les rendre adaptées à l'analyse QSAR. Dans cette étude, les données ont été nettoyées en supprimant les composés inorganiques, les sels et les composés ayant des poids moléculaires supérieurs à 700 (en se concentrant uniquement sur les petites molécules).

Ces 277 composés ont été divisés en un ensemble d'apprentissage (222composés) et un ensemble de test (55composés) en utilisant la randomisation ainsi que la diversité chimique et biologique.

2.2. Construction model AlvaDesc descripteur

Le modèle QSAR a été élaboré en utilisant le logarithme de TA100 comme variable dépendante. Les variables explicatives ont été sélectionnées parmi trois modèles :

- Model deux descripteur (MPC06 , ATSC1m)
- Model trois descripteur (RCI , ATSC1m , SpMax8_Bh(v))
- Model quatre descripteur (RCI , SM2_Dz(i) , ATSC1m , MaxddsN).

Tableau 1 : Les descripteurs intervenant dans les modèles

Descripteurs	Classe	Signification
MPC06	Comptage des marches et des chemins	Nombre de voies moléculaires d'ordre 6.
ATSC1m	Autocorrélations 2D	Autocorrélation de Broto-Moreau centrée sur le lag 1 pondéré par la masse.
RCI	Descripteurs d'anneaux	Indice de complexité de l'anneau
SpMax8_Bh(v)	Valeurs propres du fardeau	Plus grande valeur propre n. 8 de la matrice de Burden pondérée par le volume de van der Waals.
SM2_Dz(i)	Descripteurs basés sur une matrice 2D	Le moment spectral d'ordre 2 de la matrice de Barysz pondéré par le potentiel d'ionisation.
MaxddsN	Indices d'état E de type atome	Maximum ddsN.

2.2.1 Analyse de la régression

Après la préfiltration des valeurs constantes et des descripteurs fortement intercorrélés, un total de 1830 descripteurs AlvaDesc restants ont été utilisés pour établir les modèles QSAR. La mutagénicité représentée par logTA100 de 277 dérivés de nitro-aromatique a été utilisée dans la première partie pour une régression linéaire. Ensuite, la GA combinée à la procédure MLR a été utilisée pour sélectionner les meilleures variables de modélisation, et 100 modèles ont été générés.

Parmi les modèles sélectionnés par l'algorithme génétique nous avons retenu le modèles à quatre descripteurs :

2.2.1.1 Model pour quatre descripteurs RCI SM2_Dz(i) ATSC1m MaxddsN

Équation du modèle :

Log AT100 = -11.9713(±1.6004) + 4.9621 (±0.7724) **RCI** + 1.2984(±0.1986) **SM2_Dz(i)** -0.611 (±0.105) **ATSC1m +** 2.4012**(**±1.2135) **MaxddsN**

Le meilleur model pour 4 dimensions a été obtenu grâce à utilisation de l'algorithme génétique avec RCI qui représente l'indice de complexité de l'anneau et SM2_Dz(i) qui représente le moment spectral d'ordre 2 de la matrice de Barysz pondéré par le potentiel d'ionisation et ATSC1m qui représente l'Autocorrélation de Broto-Moreau centrée sur le lag 1 pondéré par la masse et MaxddsN qui représente Maximum ddsN. Les paramètres statistiques sont regroupés dans le tableau 2.

Tableau 2 *:* Paramètres statistiques du modèle 1

MODEL	R^2	Q^2_{LOO}	Q^2_{LMO} 10%	Q^2_{LMO} 30%	Q^2_{LMO} 50%	R^2_{ext}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	KΔ	F	S
(RCI , SM2_Dz(i) ,ATSC1m , MaxddsN)	72.96	71.58	69.23	71.16	69.49	70.03	70.37	68.31	63.25	0.21	146.36	0.99

D'après les résultats de ce tableau (R^2 =*72.96* %) (Q^2_{LOO} =*71.58* %) (R^2_{ext}=70.03 %)

2.2.2 Qualité de l'ajustement

Les valeurs des paramètres statistiques (Tableau1) prennent en considération la corrélation entre les quatre descripteurs (RCI, SM2_Dz(i) , ATSC1m , MaxddsN) des 277 dérivés de nitro-aromatique.

La valeur élevée de F = 146.35 plus les valeurs de (R^2 =*72.96*%) et de (Q^2 =*71.58%*) montrent la bonne qualité de l'ajustement, et la valeur du coefficient de détermination R^2 signifie que *72.96*% de la variabilité de *log AT100* est expliquée par ces descripteurs.

Le tableau 2 et la figure 12, représentent la bonne qualité d'ajustement, la robustesse et le pouvoir prédictif de notre modèle (R^2, Q^2_{LOO}, R^2_{ext}), les valeurs des erreurs relativement faibles, et celles de la dispersion des points autour de la droite, montrent que les valeurs prédites sont en adéquation avec les valeurs expérimentales, caractéristique d'un bon ajustement.

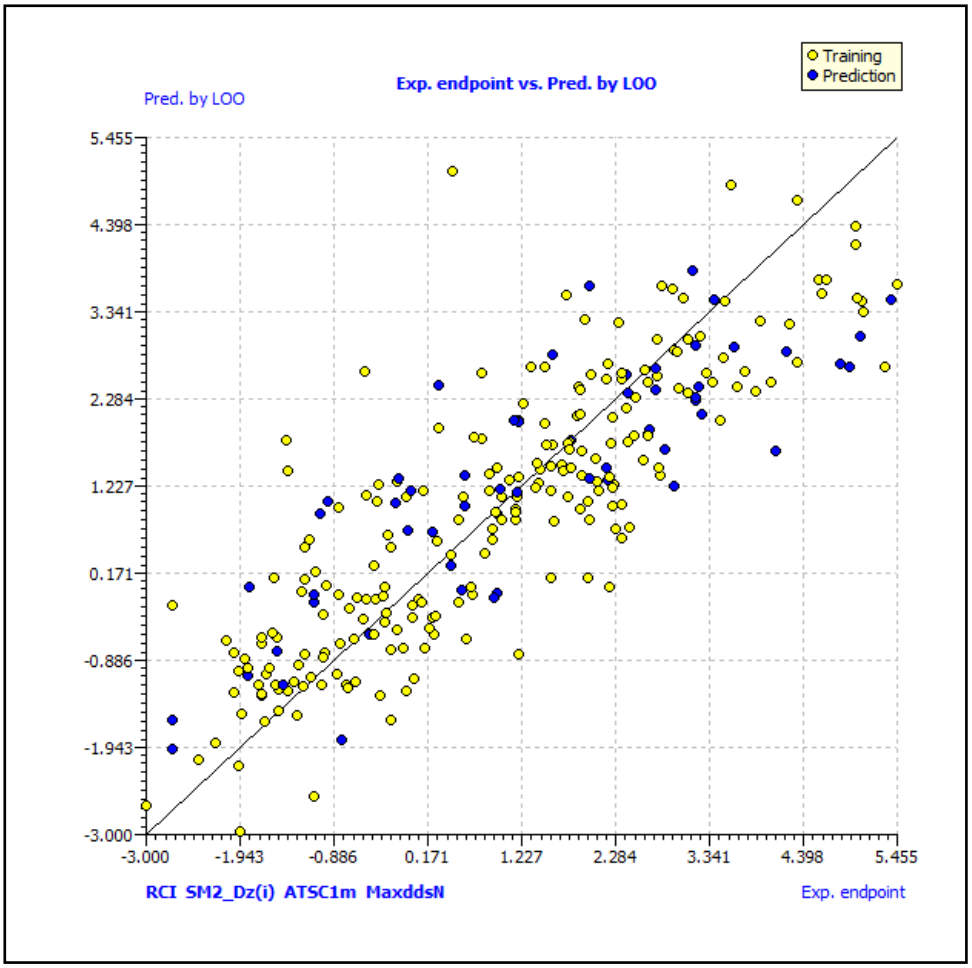


Figure 16 : Droites d'ajustement pour le modèle (03) des valeurs expérimentales et prédits de la notation *log AT100* pour le modèle de QSAR.

2.2.3 Validation

Dans cette étude l'ensemble des données a été divisé en deux sous-ensembles d'entrainement (80 % = 222 molécules) pour la construction du model et la validation interne, et un sous-ensemble de Test (20 % = 55 molécules) pour la validation externe.

2.2.3.1 Validation interne

Le coefficient de détermination de la validation croisée du modèle Q^2_{LOO} = 71.58 %, témoigne d'une bonne corrélation entre l'activité prédite et l'activité réelle, et reflète une précision du modèle.

2.2.3.2. Validation externe

La validation interne des composés nouveaux ne peut pas être suffisante pour évaluer la capacité prédictive du modèle élaboré. Il est nécessaire de réaliser une validation externe afin de prévoir correctement les composés qui n'ont pas été utilisés lors de la création du modèle.
Les paramètres de la validation externe R^2_{ext} = 70.03 % confirme la bonne capacité prédictive du modèle pour les composés non impliqués dans les calculs.

2.2.3.3. Test de randomisation (Y scrambling)

A fin de prévenir toute corrélation causée par le hasard et confirmer le modèle MLR calculé, on a utilisé le test de randomisation. Le mélange des valeurs de la réponse Y entre elles , est effectué sans modifier la valeur des descripteurs Xi (figure 13). [134], [135]

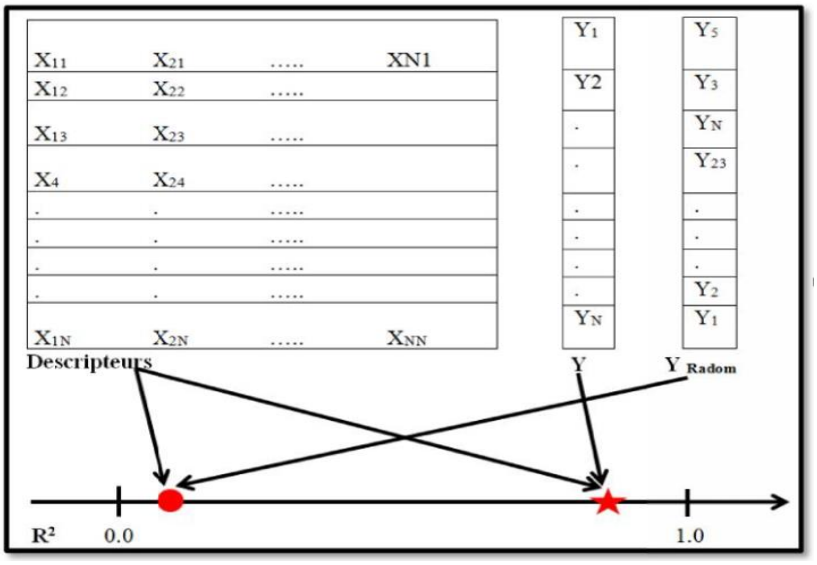


Figure 17 : Principe du test de randomisation. [136]

Les 100 modèles ont été créés, et la figure 14 illustre les faibles valeurs de Q^2 et R^2 obtenues après chaque mélange, illustrant ainsi que les résultats du modèle initial ne sont pas aléatoires.

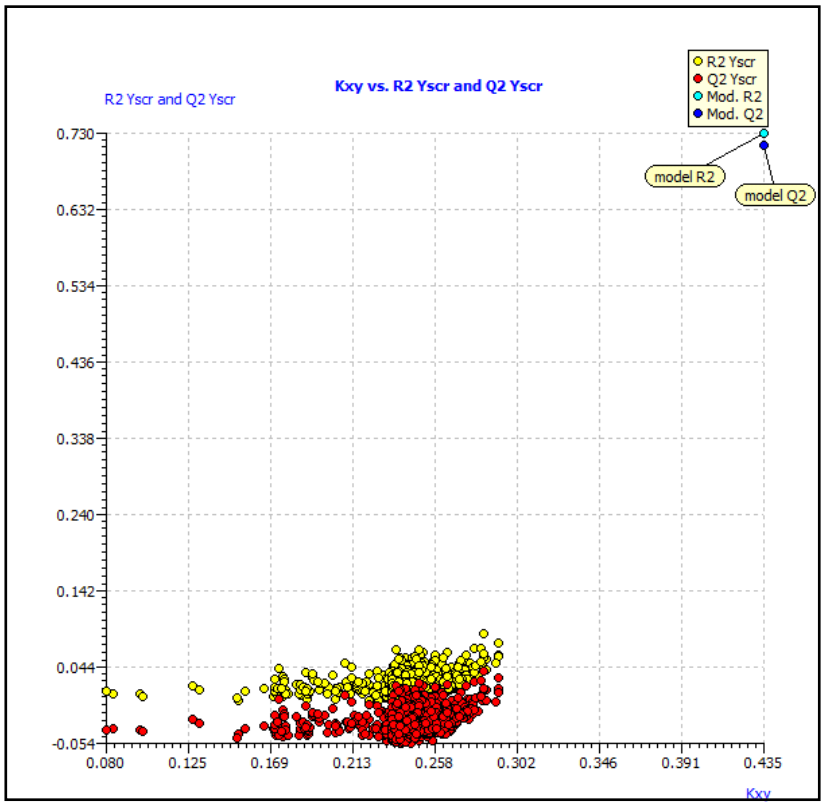


Figure 18 : Tests de randomisation

2.2.4. Le domaine d'application

Le diagramme de Williams représente les composés situés dans le domaine d'application de notre modèle (Figure 15). La majorité des composés de l'ensemble de données se situent dans l'intervalle ±3 du domaine d'application, sauf 3 molécules qui représentes des points aberrant (33/141/175), les 12 composés qui dépassent le seuil h*= 0,067 (points de levier). (16/17/95/113/161/171/212/216/273/274/275/276). Il serait donc possible d'utiliser le modèle proposé afin d'analyser les bases de données déjà existantes.

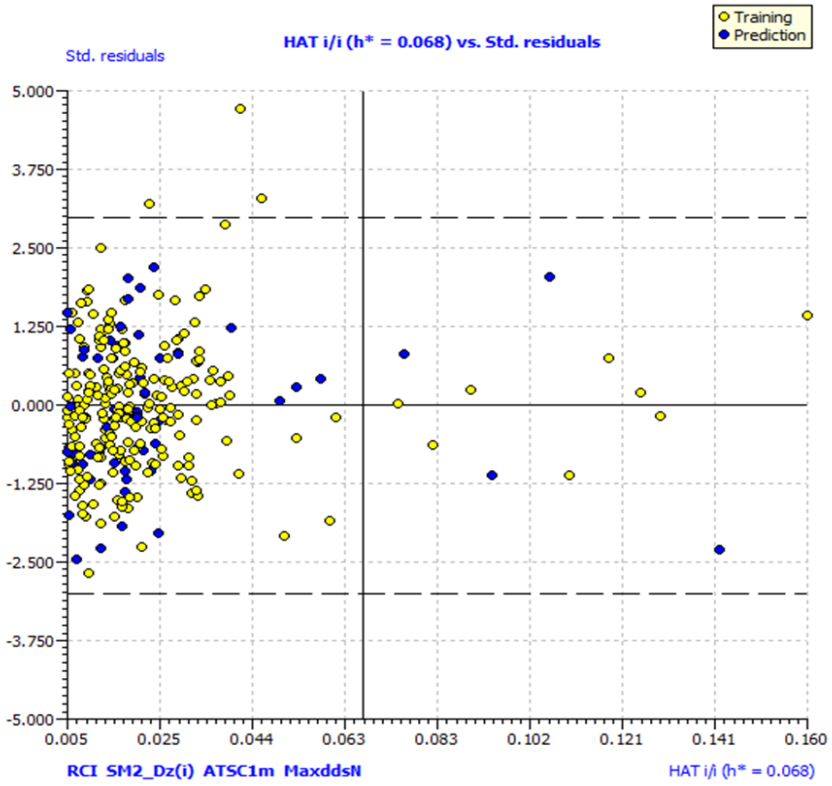


Figure 19 : Diagramme de Williams

2.2.5 Conclusion

Une équation de QSAR pratique qui relie des descripteurs théoriques au logAT100 expérimental de 277 molécules a été développée.

Plus de 1830 descripteurs sont calculés pour chaque composé à l'aide du logiciel AlvaDesc. Toutes les molécules ont été séparées en deux parties, à savoir « Training » et « Prédiction », avec un choix aléatoire. Ensuite, les meilleurs descripteurs ont été choisis à l'aide de l'algorithme génétique de QSARINS.

Lorsque le modèle est obtenu avec une qualité statistique élevée et fiable, ainsi qu'une faible erreur de prédiction, il est possible de conclure que les techniques de modélisation combinées ont pour effet d'améliorer les modèles linéaires pour cet ensemble de données.

2.3. Classification de l'activité Ames

Dans l'étude de classification, le travail se concentre sur la création de modèles prédictifs pour la mutagénicité des composés nitroaromatiques en utilisant des méthodes de machine learning et des empreintes moléculaires.

Un ensemble de données de 404 composés nitroaromatiques a été utilisé pour la classification, basée sur leur mutagénicité mesurée par la souche Salmonella typhimurium TA100. Les composés nitroaromatiques sont classés en deux catégories : "H" pour les composés hautement mutagènes et "L" pour les composés faiblement mutagènes, en fonction de leurs valeurs de logTA100.

Les fingerprints sont des représentations numériques des structures moléculaires qui capturent des informations sur la connectivité atomique, les motifs structurels et les propriétés chimiques des molécules. Les fingerprints sont utilisés comme descripteurs d'entrée pour les modèles de machine learning afin de prédire la mutagénicité des composés nitroaromatiques. Ils permettent de représenter de manière concise et informatique les caractéristiques structurales des molécules.

Workflow KNIME

Les méthodes de Machine Learning utilisées dans la partie classification sont construites sur la plateforme KNIME. La figure 16 montre la Flow des m méthode Classifieur bayésien naïf, Gradient Boosting, Forêt Aléatoire, Arbre de Décision, et Régression Logistique.

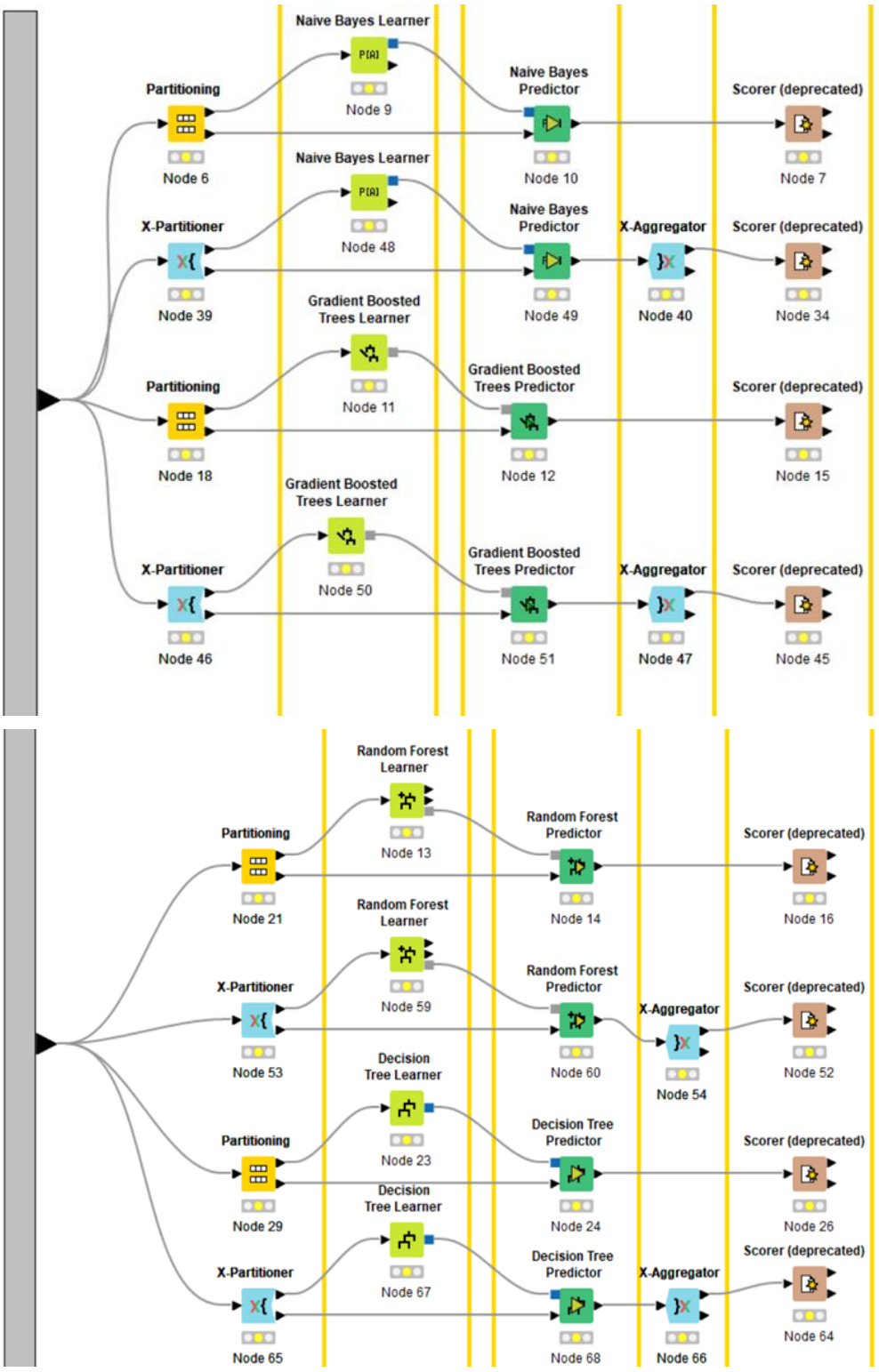


Figure 20 : Les méthodes de Machine Learning utilisées.

2.3.1 Validation du modèle

a-Validation interne (validation croisée)

La validation croisée est une technique essentielle pour évaluer la robustesse et la fiabilité des modèles prédictifs en machine Learning, car elle permet d'estimer la performance du modèle sur des données non testées et de détecter tout problème de surajustement.

b-Validation externe

Évaluation de la performance du modèle à l'aide de mesures telles que Sensibilité, la Précision, et le Spécificité sur l'ensemble de test pour vérifier sa capacité à généraliser sur de nouvelles données.

Les résultats de la validation externe et interne sont présentés dans les deux tableaux ci-dessous :

Tableau 3 : Comparaison des performances de 05 modèles sur l'ensemble de test

Nom externe	Sensibilité (%)	Spécificité (%)	Précision (%)
Apprenant Naïf Bayésien	100	2,17	43,75
Gradient Boosté	80	78,26	73,68
Forêt Aléatoire	77,14	80,43	75
Arbre de Décision	80	73,91	70
Régression Logistique	85,71	80,43	76,92

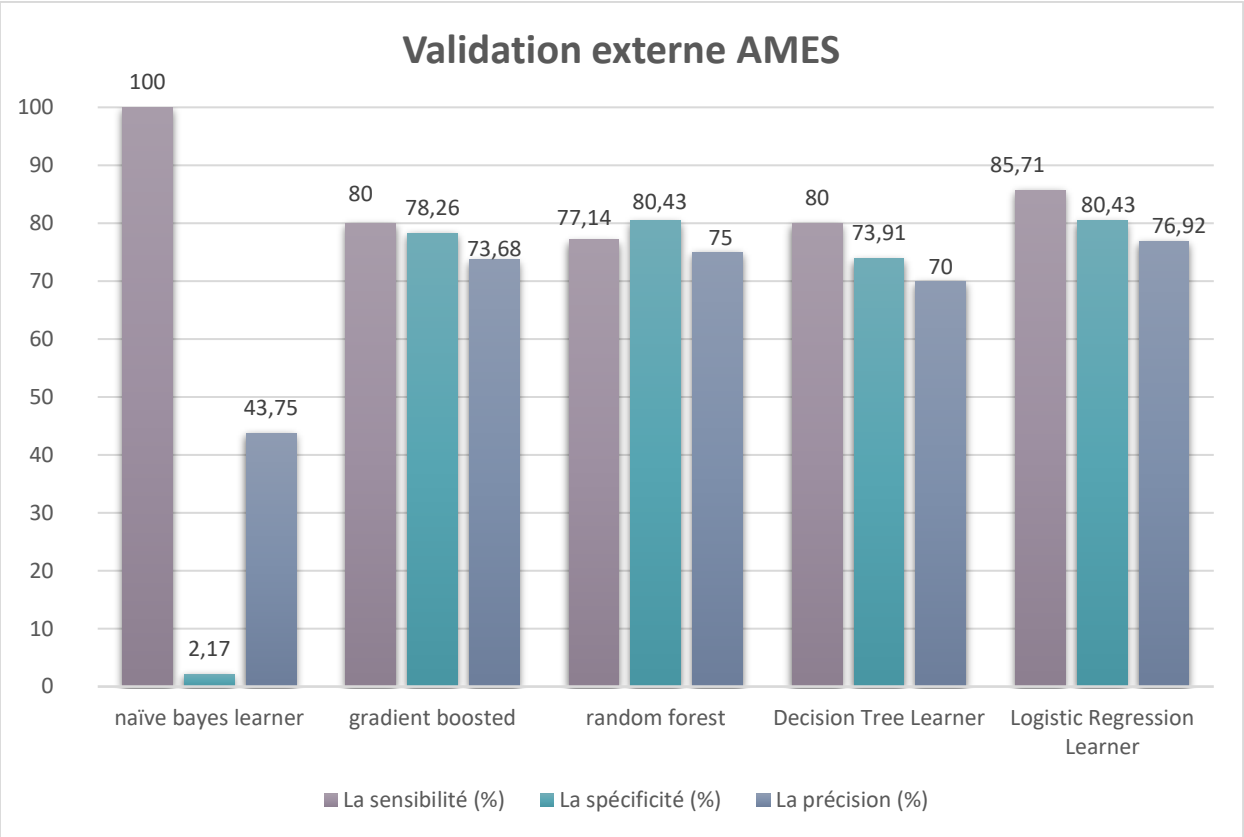


Figure 21 : Spécificité, sensibilité et précision de 5 modèles pour l'ensemble de test.

Tableau 4 : Comparaison des performances de six modèles pour la validation croisée

Nom interne	La sensibilité (%)	La spécificité (%)	La précision (%)
Apprenant Naïf Bayésien	100	0,97	49,13
Gradient Boosté	73,1	79,61	77,42
Forêt Aléatoire	74,62	83,98	81,67
Arbre de Décision	72,08	68,45	68,6
Régression Logistique	76,14	77,18	76,14

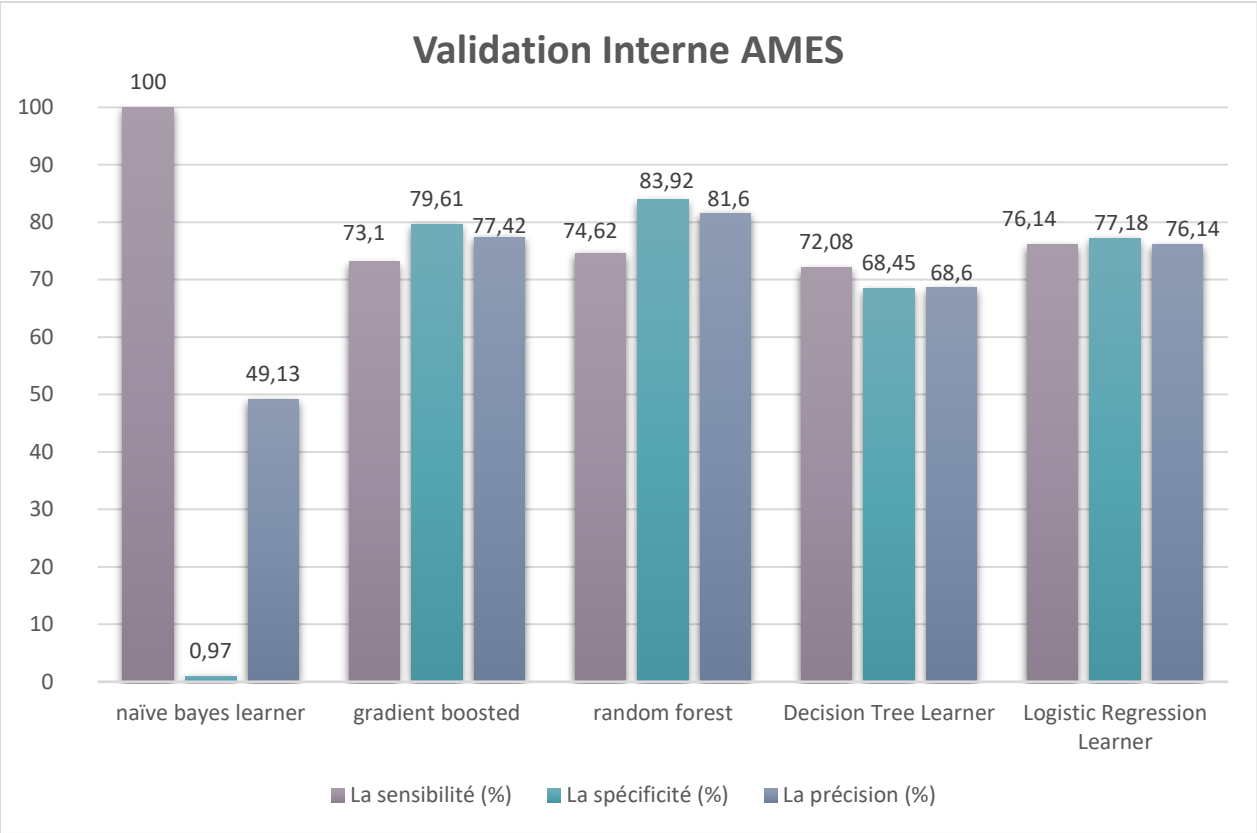


Figure 22 : Spécificité, sensibilité et précision de 5 modèles pour la validation croisée.

2.3.2. Interprétation des résultats

Le modèle Naïf Bayésien présente une sensibilité parfaite (100%) mais une spécificité extrêmement faible (2,17%) et une précision modérée (43,75%). En revanche, le modèle de Régression Logistique se distingue par la sensibilité la plus élevée (85,71%), une spécificité élevée (80,43%) et la précision la plus élevée (76,92%), en faisant le modèle le plus fiable pour la validation externe. Le modèle Gradient Boosté affiche une sensibilité de 80%, une spécificité de 78,26% et une précision de 73,68%, indiquant de bonnes performances équilibrées. Le modèle Forêt Aléatoire présente également de bonnes performances avec une sensibilité de 77,14%, une spécificité de 80,43% et une précision de 75%, le rendant également fiable. Le modèle Arbre de Décision, avec une sensibilité de 80%, une spécificité de 73,91% et une précision de 70%, montre une fiabilité modérée par rapport aux autres modèles. En conclusion, pour la validation externe, le modèle de Régression Logistique est le plus performant, suivi de près par les modèles Forêt Aléatoire et Gradient Boosté, tandis que le modèle Naïf Bayésien montre des performances inégales malgré une sensibilité parfaite.

2.3.3. Conclusion

L'étude a démontré l'efficacité des modèles QSAR et de classification basée sur des descripteurs et des fingerprints moléculaires pour prédire la mutagénicité des nitroaromatiques, offrant ainsi des perspectives prometteuses pour la prédiction des effets toxiques des composés chimiques.

En conclusion, chaque modèle à ses forces et ses faiblesses, et le choix du modèle dépendra de l'importance relative de la sensibilité, de la spécificité et de la précision pour notre application spécifique. Par exemple, si la détection de tous les vrais positifs est plus importante que l'évitement des faux positifs, un modèle avec une sensibilité plus élevée pourrait être préférable. De même, si l'évitement des faux positifs est plus important, un modèle avec une spécificité plus élevée pourrait être préférable. Enfin, si l'exactitude globale de la prédiction est la plus importante, un modèle avec une précision plus élevée serait le meilleur choix.

Conclusion générale

Conclusion générale

La modélisation moléculaire représente l'un des principaux outils permettant de prédire l'activitéé des substances en se fondant sur la relation entre leur activité et leur structure moléculaire.

Dans ce contexte, l'objectif de ce mémoire était de développer des modèles QSAR fiables pour prédire les logAT100 à partir de séries de composés nitro-aromatiques par la régression et par la classification.

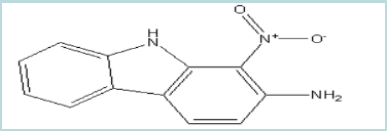
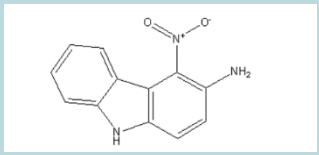
Dans la première partie, portant sur une série de 277 composés, l'utilisation des descripteurs du logiciel QSARINS a conduit à la création de modèles QSAR en recourant à une approche algorithmique génétique. Il est crucial que les résultats obtenus ainsi que les paramètres statistiques des modèles soient acceptables. Pour ce faire, l'algorithme génique a été utilisé conjointement avec une analyse de régression linéaire multiple (MLR) afin d'établir une corrélation entre le logAT100 et quatre descripteurs sélectionnés : (RCI, SM2_Dz(i), ATSC1m, MaxddsN). L'objectif principal était de dériver l'équation finale de QSAR. Les principales techniques de validation (interne, externe, la randomisation des Y...) ont été utilisées. Le modèle QSAR obtenu, conforme aux cinq principes de l'OCDE, a été évalué à l'aide de ces données ; les critères d'acceptabilité du modèle de Golbarikh et Tropsha ont également été vérifiés. Le modèle QSAR a été déterminé avec la variable dépendante étant le logarithme de TA100 et les variables explicatives étant les descripteurs. Selon les critères de Golbreich et Tropsha, le modèle trouvé présentait un R^2 de 72,96 % et un Q^2 de 71,58 %.

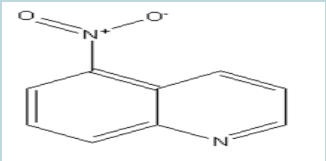
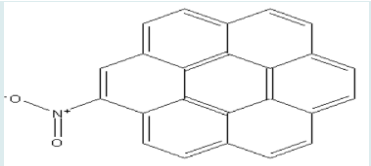
Dans la deuxième partie, portant sur une série de 404 composés, l'utilisation des descripteurs de fingerprint et des méthodes d'apprentissage automatique (Random Forest (RForest), Gradient Boosted Trees (GBoost), Naïve Bayes (NBayes), Logistic Regression (LRegression), Decision Tree (DTree)) sur plateforme KNIME a été réalisée. La validation interne et externe de AMES nous a permis de choisir le modèle dépendant de l'importance relative de la sensibilité, le modèle équilibré étant le meilleur pour ces modèles. La Logistic Regression (LRegression) a présenté une sensibilité de 85,71 %, une spécificité de 80,43 % et une précision de 76,92 %.

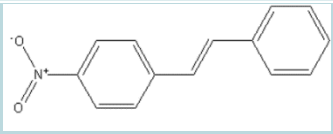
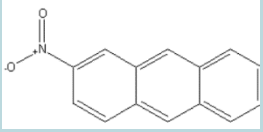
Cet objectif principal nous a ouvert des perspectives prometteuses dans cette discipline, où nous envisageons de continuer à développer des modèles en explorant d'autres méthodes telles que le Machine Learning, le Deep Learning et les graphes machines pour poursuivre nos recherches.

Annexes

Annexes 01 : Les valeurs expérimentales prédit et calculé par l'approche AlvaDesc-QSARINS pour les 277 dérivés de Nitroaromatique.

	SMILE	Ensemble	logAT100 Exp	logAT100 Pred
1		Training	2,36	0,6195
2	[N+](=O)([O-])C1=CC=2NC3=CC=CC=C3C2C=C1NC(C)=O	Training	-0,07	1,0948
3	[N+](=O)([O-])C1=CC=2NC3=CC=CC=C3C2C=C1N	Training	0,91	0,7234
4	C(C)(=O)NC=1C=CC=2NC3=CC=CC=C3C2C1[N+](=O)[O-]	Training	-0,52	1,1007
5		Training	-0,27	0,6395
6	CN1C2=CC=CC=C2C=2C=CC(=CC12)[N+](=O)[O-]	Training	2,2989	0,7335
7	CN1C2=CC=CC=C2C=2C=C(C=CC12)[N+](=O)[O-]	Training	2,4456	0,7481
8	CC1=CC(=C(C=2C3=CC=CC=C3N(C12)C)C)[N+](=O)[O-]	Training	2,2304	0,0622
9	CC1=CC(=C(C=2C3=CC=CC=C3NC12)C)[N+](=O)[O-]	Training	0,8129	0,432
10	CC1=CC(=C(C=2C3=CC(=CC=C3NC12)C)C)[N+](=O)[O-]	Training	1,5682	0,1603
11	CC1=CC(=C(C=2C3=CC(=CC=C3NC12)O)C)[N+](=O)[O-]	Training	0,2788	0,5694
12	[N+](=O)([O-])C1=CC=2C=CC3=CC=CC=C3C2C=C1	Training	2,11	1,1818
13	[N+](=O)([O-])C=1C=CC=C2C=CC=NC12	Training	-1,24	-0,0623
14	[N+](=O)([O-])C1=C2C=CC=NC2=CC=C1	Training	-0,96	0,0109

15	[N+](=O)([O-])C1=CC=C2C=CC=C3C4=CC=CC=C4C1=C23	Prediction	2,74	2,4022
16	*FC1=C(C=C(C=C1)F)[N+](=O)[O-]*	Prediction	-0,79	-1,841
17		Training	4,99	4,435
18	CC=1C=C(C=CC1[N+](=O)[O-])C1=CC=CC=C1	Training	-0,1	-0,7189
19	C(C)(=O)OC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Training	1,86	2,0827
20	[N+](=O)([O-])C1=CC2=C(N=CN2)C=C1	Prediction	-1,83	0,0093
21	CC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Training	2,36	1,033
22	[N+](=O)([O-])C=1C2=CC=CC=C2C=2C=CC=CC2C1	Training	2,25	1,0031
23	NC1=CC=C2C=CC3=CC=C(C4=CC=C1C2=C34)[N+](=O)[O-]	Prediction	2,43	2,3714
24	CN1N=CC2=CC=C(C=C12)[N+](=O)[O-]	Prediction	-1,1	-0,1669
25	[N+](=O)([O-])C1=CC=C2C=CC3=CC=CC4=CC=C1C2=C34	Training	2,758	2,5705
26	[N+](=O)([O-])C=1C=CC=2C3=CC=CC=C3C3=CC=C(C1C23)[N+](=O)[O-]	Prediction	3,62	2,925
27	[N+](=O)([O-])C1=C(C=CC=C1N)N	Training	-3	-2,6435
28	[N+](=O)([O-])C=1C=CC=2C(C3=CC=CC=C3C2C1)=O	Training	2,6557	1,8772
29	[N+](=O)([O-])C=1C=C2C=CC=NC2=CC1	Training	-1,08	0,1838
30	C(#N)C1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Training	2,51	2,3097
31	[N+](=O)([O-])C1=CC=C(C=C1)C1=CC(=CC=C1)[N+](=O)[O-]	Prediction	0,23	0,68
32		Training	0,45	4,8574
33	OC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Training	1,68	1,488

34	[N+](=O)([O-])C1=CC=C2C=CC=3C=C4C(=C5C=CC1=C2C53)C=CC=C4	Training	2,7557	3,0175
35	FC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Prediction	2,68	1,9228
36	[N+](=O)([O-])C1=CC=2C(C3=CC(=CC=C3C2C(=C1)[N+](=O)[O-])[N+](=O)[O-])=O	Prediction	3,41	3,4942
37	FC(C(=O)NC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-])(F)F	Training	2,81	3,6311
38	[N+](=O)([O-])C1=CC2=CC=CC=C2C=C1	Training	-0,3	0,006
39	NC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Training	1,56	1,1802
40	CC=1C(=CC2=CC=CC=C2C1)[N+](=O)[O-]	Training	0	-0,2084
41		Training	0,69	-0,0661
42	[N+](=O)([O-])C1=CC=C(C=C1)C1=CC=CC=C1	Training	-0,3	-0,4176
43	COC=1C=C2C=CC=NC2=C(C1)[N+](=O)[O-]	Training	-1,21	0,0957
44	[N+](=O)([O-])C1=CC=CC2=CC=CC=C12	Training	-0,61	-0,1294
45	BrC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Training	3,06	3,5054
46	CN1N=C2C=CC(=CC2=C1)[N+](=O)[O-]	Prediction	-1,1	-0,0752
47	[N+](=O)([O-])C1=CC=2C3=CC=CC=C3C3=CC=CC(=C1)C23	Training	3,01	2,4256
48	[N+](=O)([O-])C1=CC=2CCC3=CC=CC=C3C2C=C1	Training	1,99	0,844
49	[N+](=O)([O-])C1=CC=C(C=C1)[N+](=O)[O-]	Training	0,15	-0,7166
50	[N+](=O)([O-])C1=CC=2CCC=3C=C(C=C4CCC(=C1)C2C43)[N+](=O)[O-]	Training	3,5	2,8079
51		Prediction	2,95	1,2276

52	[N+](=O)([O-])C1=CC2=NC3=CC(=CC=C3N=C2C=C1)[N+](=O)[O-]	Prediction	2,75	2,6681
53	[N+](=O)([O-])C1=CC=CC2=C(C=CC=C12)[N+](=O)[O-]	Training	0,52	0,8316
54	[N+](=O)([O-])C1=CC=C2C=CC=3C(=CC=C4C5=C(C1=C2C43)CCCC5)[N+](=O)[O-]	Prediction	2,41	2,595
55	[N+](=O)([O-])C1=CC(=C2C=CC=3C=CC=C4C5=C(C1=C2C43)CCCC5)[N+](=O)[O-]	Training	2,41	2,1953
56	[N+](=O)([O-])C1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Prediction	3,22	2,4501
57	[N+](=O)([O-])C1=CC2=C(OC3=C(O2)C=C(C(=C3)Cl)Cl)C=C1	Training	1,73	3,4848
58	CC1=C(C=CC2=CC=CC=C12)[N+](=O)[O-]	Training	-0,7	-0,2565
59	[N+](=O)([O-])C1=CC2=NC3=CC=CC=C3N=C2C=C1	Training	2,06	1,5745
60	ClC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Training	3,11	2,3714
61	NC1=C(C=C(C=C1)[N+](=O)[O-])O	Training	-2,4	-2,098
62	[N+](=O)([O-])C1=CC=2C(C3=CC(=CC=C3C2C=C1)[N+](=O)[O-])=O	Prediction	3,19	2,9436
63	[N+](=O)([O-])C1=CC(=CC2=CC=CC=C12)[N+](=O)[O-]	Prediction	-0,05	0,6909
64	NC1=CC=C(C=C1)C1=CC(=CC=C1)[N+](=O)[O-]	Training	-1,52	-0,6046
65	OC1=C(C=2C3=CC=CC=C3C3=CC=CC(=C1)C23)[N+](=O)[O-]	Training	2,26	2,068
66	[N+](=O)([O-])C1=C2C=CC=3C=C4C(=C5C=CC(C=C1)=C2C53)C=CC=C4	Training	3,1126	3,0175
67	[N+](=O)([O-])C1=CC=2CC3=CC=CC(=C3C2C=C1)[N+](=O)[O-]	Prediction	3,2	2,2805
68	[N+](=O)([O-])C=1C=C2C(C(NC2=CC1)=O)=O	Prediction	-0,94	1,0563

69	[N+](=O)([O-])C1=CC=C2C=CC=3C=CC(=C4C5=C(C1=C2C43)CCCC5)[N+](=O)[O-]	Training	2,19	2,5325
70	[N+](=O)([O-])C1=CC=2CC3=CC=CC=C3C2C=C1	Training	1,43	1,2714
71	NC=1C=C(C=CC1)C1=C(C=CC=C1)[N+](=O)[O-]	Training	-2	-0,8012
72	[N+](=O)([O-])C1=C2C=CC=C3C=4C=C5C(=CC4C(C=C1)=C32)C=CC=C5	Training	2,76	3,012
73	[N+](=O)([O-])C1=CC=C(C=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	1,17	0,8281
74	ClC1=C(C=CC=C1)[N+](=O)[O-]	Training	-1,72	-1,1919
75	NC1=C(C=CC=C1)C1=CC(=CC=C1)[N+](=O)[O-]	Prediction	-1,52	-0,7693
76	CC1=C(C=CC=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-0,23	-0,7421
77	[N+](=O)([O-])C1=CC2=C(OC3=C(O2)C=CC=C3)C=C1	Prediction	1,79	1,7912
78	[N+](=O)([O-])C1=C2C3=CC=CC4=CC=CC(C2=CC=C1)=C43	Training	1,87	2,4433
79	[N+](=O)([O-])C1=CC(=CC=C1)[N+](=O)[O-]	Training	0,03	-1,0615
80	[N+](=O)([O-])C1=CC=2NC3=CC=CC=C3C2C=C1	Training	1,01	1,0975
81	[N+](=O)([O-])C1=CC=C(C=C)C=C1	Training	-1,3	-1,5461
82	[N+](=O)([O-])C1=CC2=C(OC3=C2C=CC=C3)C=C1	Training	1,4473	1,4334
83	CN1N=CC2=CC=CC(=C12)[N+](=O)[O-]	Training	-1	-0,34
84	[N+](=O)([O-])C1=CC=C2C=CC=3C=CC=C4C5=C(C1=C2C43)C=CC=C5	Prediction	1,59	2,8352
85	[N+](=O)([O-])C1=CC=2CCC=3C=CC=C4CCC(=C1)C2C43	Training	1,58	1,7363
86	[N+](=O)([O-])C=1C=CC=2C3=C(C4=CC=CC=5C=CC1C2C45)C=CC=C3	Training	2,95	2,8837
87	[N+](=O)([O-])C=1C2=CC=CC3=CC=C4C=CC=C(C1)C4=C32	Training	3,3918	2,5209

88	[N+](=O)([O-])C1=CC2=CC=CC(=C2C=C1[N+](=O)[O-])[N+](=O)[O-]	Training	1,51	1,7363
89	[N+](=O)([O-])C1=CC(=C2C=CC3=C(C=CC4=CC=C1C2=C34)[N+](=O)[O-])[N+](=O)[O-]	Training	4,99	4,1992
90	COC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Training	2,79	1,3721
91	[N+](=O)([O-])C1=C2C=CN=CC2=CC=C1	Training	-1,55	0,0857
92	[N+](=O)([O-])C1=CC2=CC=CC3=CC=CC1=C23	Training	1,77	1,7683
93	NC=1C=C(C=CC1)C1=CC(=CC=C1)[N+](=O)[O-]	Training	-1,7	-0,6875
94	[N+](=O)([O-])C=1C=CC=2C3=C(C4=CC=CC=5C=CC1C2C45)CCCC3	Training	0,78	1,7827
95	[N+](=O)([O-])C1=CC=2C(C3=CC(=CC(=C3C2C(=C1)[N+](=O)[O-])[N+](=O)[O-])[N+](=O)[O-])=O	Training	2,93	3,553
96	[N+](=O)([O-])C1=CC2=CC=C3C=CC=C4C=CC(=C1)C2=C43	Training	3,3106	2,6209
97	[N+](=O)([O-])C1=CC=C2C=CC=3C=CC=C4C5=C(C1=C2C43)CCCC5	Training	0,7	1,7987
98	[N+](=O)([O-])C1=C2C=CC=NC2=C2N=CC=CC2=C1	Prediction	0,59	1,3664
99	OC=1C=C2C=CC=C3C=CC4=CC=C(C1C4=C32)[N+](=O)[O-]	Training	1,89	2,3957
100	[N+](=O)([O-])C=1C=CC=2C3=CC=CC=C3C3=CC=CC1C23	Training	3,67	2,4633
101	[N+](=O)([O-])C=1C2=CC=C3C=4C5=C(C=CC6=C(C=CC(C(=CC1)C42)=C65)[N+](=O)[O-])C=C3	Training	3,6	4,8491
102	[N+](=O)([O-])C1=CC=C(C=O)C=C1	Training	-1,64	-1,0551
103	[N+](=O)([O-])C1=C2CCC=3C=CC=C(C=C1)C32	Prediction	1	1,1999
104	NC1=C(C=CC=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-1,7	-0,6175
105	[N+](=O)([O-])C=1C=C2C=CCC2=CC1	Training	0,08	-0,1296
106	FC1=CC=C(C=C1)[N+](=O)[O-]	Training	-0,23	-1,5506

107	[N+](=O)([O-])C1=C2C=CC=3C=C4C(=C5C=CC(C=C1)=C2C53)CCCC4	Training	0,3005	1,8967
108	[N+](=O)([O-])C1=C(C=CC(=C1)[N+](=O)[O-])C1=C(C=CC=C1)[N+](=O)[O-]	Prediction	-0,19	1,0246
109	CN1N=CC2=CC(=CC=C12)[N+](=O)[O-]	Training	-0,82	-0,0897
110	[N+](=O)([O-])C1=C(C=C2C=C(C=C3C4=CC=CC=C4C1=C23)[N+](=O)[O-])[N+](=O)[O-]	Prediction	3,16	3,844
111	[N+](=O)([O-])C=1C=C2C=3C=CC=CC3C=CC2=C2C=CC=CC12	Training	1,75	1,7508
112	CC=1C=C(C=CC1[N+](=O)[O-])C1=C(C=CC=C1)C	Training	-0,84	-1,0401
113	[N+](=O)([O-])C1=C(C=C(C=C1)F)F	Training	-1,66	-1,6231
114	[N+](=O)([O-])C1=CC2=CC=C3C=CC=C4CCC(=C1)C2=C43	Prediction	3,27	2,1058
115	[N+](=O)([O-])C1=CC=CC2=NC3=CC(=CC=C3N=C12)[N+](=O)[O-]	Training	2,02	2,5723
116	[N+](=O)([O-])C=1C=C(C=CC1[N+](=O)[O-])C1=CC(=CC=C1)[N+](=O)[O-]	Training	1,92	1,3924
117	[N+](=O)([O-])C1=CC=C2C=CC=3C=CC=C1C32	Training	1,91	1,6698
118	[N+](=O)([O-])C1=C(C=CC(=C1)[N+](=O)[O-])NN	Training	-0,07	-1,2085
119	[N+](=O)([O-])C1=C(C=CC=C1)OC	Prediction	-2,7	-1,9581
120	[N+](=O)([O-])C1=CC=C(C=C1)\C=C\C(=O)C1=CC=CC=C1	Training	-1,15	0,5443
121	[N+](=O)([O-])C1=C(C=O)C=CC=C1	Training	-1,92	-1,5312
122	[N+](=O)([O-])C1=C(C2=C(OC3=C(O2)C=C(C(=C3)Cl)Cl)C=C1Cl)Cl	Training	-1,4	1,314
123	[N+](=O)([O-])C1=CC2=CC=C3C=C(C=C4CCC(=C1)C2=C43)[N+](=O)[O-]	Training	4,25	3,2245
124	ClC=1C=C(C=CC1F)[N+](=O)[O-]	Training	-1,21	-0,8088

125	OC1=CC(=C2C=CC3=CC=CC4=CC=C1C2=C34)[N+](=O)[O-]	Training	3,87	2,4165
126	[N+](=O)([O-])C1=CC=2C3=CC=CC=C3C3=CC=CC(=C1[N+](=O)[O-])C23	Training	2,62	2,6475
127	CN1N=C2C=C(C=CC2=C1)[N+](=O)[O-]	Training	-0,41	-0,1464
128	ClC1=CC(=C(N)C=C1)[N+](=O)[O-]	Training	-2	-1,2803
129	[N+](=O)([O-])C1=CC=CC2=CC=CC(=C12)[N+](=O)[O-]	Training	0,9	0,5844
130	ClC1=C(C=CC=C1Cl)[N+](=O)[O-]	Training	-1,51	-1,2321
131	[N+](=O)([O-])C=1C(=CC2=C(OC3=C(O2)C=C(C(=C3)Cl)Cl)C1)Cl	Training	-0,53	2,5613
132	[N+](=O)([O-])C1=C(C=CC=C1)OCC	Training	-2,22	-1,8794
133	ClC1=C(C=C(C=C1)Cl)[N+](=O)[O-]	Training	-1,54	-1,1819
134	NC=1C=C(C=CC1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	0,25	-0,5454
135	NC1=CC=C(C=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	0,19	-0,4709
136	CC=1NC2=C(N1)C=CC(=C2)[N+](=O)[O-]	Training	-0,51	-0,1328
137	[N+](=O)([O-])C1=C2C=CC=3C=CC=C(C=C1)C32	Training	1,77	1,691
138	[N+](=O)([O-])C1=CC2=NC3=CC=C(C=C3N=C2C=C1)[N+](=O)[O-]	Training	4,34	2,7618
139	[N+](=O)([O-])C=1C=CC=C2C=NNC12	Training	0,11	-0,1618
140	[N+](=O)([O-])C1=CC=2C3=CC=CC=C3C3=CC(=CC(=C1)C23)[N+](=O)[O-]	Training	2,32	3,213
141	[N+](=O)([O-])C1=CC=C2CCC=3C=CC=C1C32	Training	0,58	1,0992
142	[N+](=O)([O-])C=1NC2=C(N1)C=CC=C2	Training	0	-0,3578
143	[N+](=O)([O-])C1=C(C=CC(=C1)[N+](=O)[O-])OC	Training	-1,89	-0,888
144	[N+](=O)([O-])C1=CC=C2CCNC2=C1	Prediction	-0,48	-0,5677
145	OC1=C2C=CC3=CC=C(C4=CC=C(C=C1)C2=C43)[N+](=O)[O-]	Training	1,34	2,6661
146	C(C)(=O)NC1=CC=2CC3=CC(=CC=C3C2C=C1)[N+](=O)[O-]	Prediction	2,85	1,684

147	OC1=CC=C2C=CC3=CC=C(C4=CC=C1C2=C43)[N+](=O)[O-]	Training	1,49	2,6661
148	BrC1=C(N)C(=CC(=C1)[N+](=O)[O-])[N+](=O)[O-]	Training	-1,32	-1,1417
149	[N+](=O)([O-])C1=CC=C2C=CC=3C(=CC=C4C5=C(C1=C2C43)C=CC=C5)[N+](=O)[O-]	Prediction	1,99	3,669
150	[N+](=O)([O-])C1=CC2=CC=C3C=C(C=C4C=CC(=C1)C2=C43)[N+](=O)[O-]	Training	4,58	3,7497
151	[N+](=O)([O-])C1=CC(=C2C=CC3=CC=CC4=CC=C1C2=C34)[N+](=O)[O-]	Prediction	5,04	3,0524
152	[N+](=O)([O-])C=1C=C2C3=CC=CC4=CC=CC(C2=CC1)=C43	Training	4,05	2,507
153	[N+](=O)([O-])C1=CC(=C2C=CC3=CC=CC4=CC=C1C2=C34)OC(C)=O	Prediction	4,22	2,8668
154	[N+](=O)([O-])C1=CC=C2C=CC3=C(C=CC4=CC=C1C2=C34)[N+](=O)[O-]	Training	5,06	3,5017
155	[N+](=O)([O-])C1=CC=C(C=C1)OC	Prediction	-2,7	-1,5965
156	[N+](=O)([O-])C1=C(C=CC(=C1)[N+](=O)[O-])F	Training	1,2	-0,7078
157	[N+](=O)([O-])C1=CC=CC2=NC3=CC=CC=C3N=C12	Training	0,87	1,3785
158	[N+](=O)([O-])C1=CC=C2C=NNC2=C1	Training	0,66	0,0269
159	[N+](=O)([O-])C=1C=C(C=CC1[N+](=O)[O-])C1=CC=C(C=C1)[N+](=O)[O-]	Training	2,6	1,5864
160	[N+](=O)([O-])C=1C=CC=2C3=CC(=CC=C3C3=CC=CC1C23)[N+](=O)[O-]	Training	5,02	3,5432
161	[N+](=O)([O-])C=1C(=C(C2=C(OC3=C(O2)C=C(C(=C3)Cl)Cl)C1Cl)Cl)Cl	Training	-0,33	-0,1215
162	[N+](=O)([O-])C1=CC=C2C=CCC2=C1	Prediction	0,96	-0,0668

163	[N+](=O)([O-])C1=CC=CC2=NC3=CC=CC(=C3N=C12)[N+](=O)[O-]	Training	1,26	2,2195
164	CN1N=C2C(=CC=CC2=C1)[N+](=O)[O-]	Training	0,23	-0,3429
165	[N+](=O)([O-])C1=CC=C(C=C1)\C=C\C(=O)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-1,42	1,6362
166	[N+](=O)([O-])C=1C2=CC=C3C=4C5=C(C=CC6=CC=C(C(C(=CC1)C42)=C65)[N+](=O)[O-])C=C3	Training	4,33	4,6994
167	[N+](=O)([O-])C1=C(C=CC(=C1)N)N	Training	-1,11	-2,4745
168	[N+](=O)([O-])C=1C=CC=C2C=CC(=NC12)C	Training	-2,7	-0,248
169	[N+](=O)([O-])C1=CC=C2C=CC3=CC=C(C4=CC=C1C2=C34)[N+](=O)[O-]	Prediction	5,39	3,5017
170	[N+](=O)([O-])C=1C=CC=2C3=CC=CC(=C3C3=CC=CC1C23)[N+](=O)[O-]	Training	5,09	3,3794
171	[N+](=O)([O-])C=1C=C(C=CC1[N+](=O)[O-])F	Prediction	-1,84	-1,057
172	[N+](=O)([O-])C=1C=C2CCNC2=CC1	Training	-0,17	-0,4975
173	[N+](=O)([O-])C1=CC=2C3=CC=CC=C3C3=CC=CC=C3C2C=C1	Prediction	4,09	1,6663
174	[N+](=O)([O-])C=1C=CC=2NC3=CC=CC=C3C2C1	Training	1,1821	1,0997
175	[N+](=O)([O-])C=1C=C(C=CC1)C1=CC=CC=C1	Training	-1,5758	-0,5541
176	[N+](=O)([O-])C1=C(C=CC=C1)C1=CC=CC=C1	Training	-2,0986	-0,6567
177	CC1=C(C2=CC=CC=C2C=C1)[N+](=O)[O-]	Training	-0,2932	-0,3011
178	NC1=CC=C(C=C1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Training	0,5274	-0,158
179	ClC=1C=C(C=CC1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Training	0,9543	0,9013
180	ClC1=CC=C(C=C1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Training	1,1669	0,9563
181	C(#N)C=1C=C(C=CC1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Training	1,6016	0,8274

182	C(#N)C1=CC=C(C=C1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Training	1,1683	0,9214
183	COC=1C=C(C=CC1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Prediction	0,9168	-0,1187
184	COC1=CC=C(C=C1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Prediction	0,5568	-0,0312
185	[N+](=O)([O-])C=1C=C(C=CC1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Training	1,8952	0,9766
186	[N+](=O)([O-])C1=CC=C(C=C1)C=CC1=CC=C(C=C1)[N+](=O)[O-]	Training	1,989	1,0851
187	C(#N)C=1C=C(C=CC1)\C=C/C1=CC=C(C=C1)[N+](=O)[O-]	Training	1,0119	0,8274
188	C(#N)C1=CC=C(C=C1)\C=C/C1=CC=C(C=C1)[N+](=O)[O-]	Training	0,9432	0,9214
189	CSC1=CC=C(C=C1)[N+](=O)[O-]	Training	-1,0132	-1,1766
190	C(C)SC1=CC=C(C=C1)[N+](=O)[O-]	Training	-0,7447	-1,1733
191	[N+](=O)([O-])C1=CC=C(C=C1)SCCC	Training	-0,7212	-1,2119
192	C(CCC)SC1=CC=C(C=C1)[N+](=O)[O-]	Training	-0,3565	-1,2867
193	C(C=C)SC1=CC=C(C=C1)[N+](=O)[O-]	Training	0,6107	-0,611
194	C(C1=CC=CC=C1)SC1=CC=C(C=C1)[N+](=O)[O-]	Prediction	0,4472	0,2665
195	[N+](=O)([O-])C1=CC=C(C=C1)SC1=CC=CC=C1	Training	0,4472	0,4032
196	[N+](=O)([O-])C1=CC=2C=CC=C3C4=C(C=5C=CC=C1C5C32)CCCC4	Training	2,233	1,7741
197	OC1CCCC=2C1=CC=1C=CC3=CC=C(C=4C=CC2C1C34)[N+](=O)[O-]	Training	1,2041	2,0135
198	OC1CCCC=2C1=CC=1C=CC3=C(C=CC=4C=CC2C1C34)[N+](=O)[O-]	Prediction	1,2041	2,0135
199	[N+](=O)([O-])C1=CC=2C=CC=C3C4=C(C=5C=CC=C1C5C32)C=CC=C4	Training	2,9934	2,8699

200	[N+](=O)([O-])C1=CC(=C2C=CC=3C=CC=C4C5=C(C1=C2C43)C=CC=C5)[N+](=O)[O-]	Training	3,9274	3,2621
201	C(C)C=1C=C(C=CC1[N+](=O)[O-])C1=CC=CC=C1	Training	-1,27	-0,9429
202	C(C)(C)C=1C=C(C=CC1[N+](=O)[O-])C1=CC=CC=C1	Training	-1,39	-1,2517
203	C(C)C=1C=C(C=C(C1[N+](=O)[O-])CC)C1=CC=CC=C1	Training	-1,51	-1,4971
204	C(C)(C)C=1C=C(C=C(C1[N+](=O)[O-])C(C)C)C1=CC=CC=C1	Training	-1,96	-2,1356
205	C(C)C1=C(C=CC=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-1,6	-0,9844
206	CC1=CC=C(C=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-0,66	-0,6192
207	C(C)C1=CC=C(C=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-0,98	-0,7831
208	C(C)(C)C1=CC=C(C=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-1,96	-1,0372
209	C(CCC)C1=CC=C(C=C1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-1,14	-1,0859
210	CC=1C=C(C=CC1)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-0,81	-0,684
211	CC=1C=C(C=C(C1)C)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-1,84	-0,99
212	C(C)(C)(C)C=1C=C(C=C(C1)C(C)(C)C)C1=CC=C(C=C1)[N+](=O)[O-]	Training	-1,93	-2,8459
213	C1(=CC=CC=C1)C1=NC=C(C=C1)[N+](=O)[O-]	Training	-0,54	-0,3755
214	C(C)(C)(C)C1=CC=C(C=C1)C1=NC=C(C=C1)[N+](=O)[O-]	Training	-1,7	-1,3121
215	C(C)(C)(C)C1=CC=C2C=3C=CC(=CC3CC2=C1)[N+](=O)[O-]	Training	-0,43	0,2467
216	C12(CC3CC(CC(C1)C3)C2)C2=CC=C3C=1C=CC(=CC1CC3=C2)[N+](=O)[O-]	Prediction	-1,03	0,9039
217	[N+](=O)([O-])C1=CC=CC=2C3=CC=CC=C3NC12	Training	-0,832	0,9549
218	CC1=CC=C(/C=C/C2=CC=C(C=C2)[N+](=O)[O-])C=C1	Training	0,26	-0,3249

219	C(C)C1=CC=C(/C=C/C2=CC=C(C=C2)[N+](=O)[O-])C=C1	Training	-0,42	-0,5543
220	C(C)(C)C1=CC=C(/C=C/C2=CC=C(C=C2)[N+](=O)[O-])C=C1	Training	-1	-0,8462
221	C(C)(C)(C)C1=CC=C(/C=C/C2=CC=C(C=C2)[N+](=O)[O-])C=C1	Prediction	-1,4559	-1,1739
222	C(C)(CC)C1=CC=C(/C=C/C2=CC=C(C=C2)[N+](=O)[O-])C=C1	Training	-0,64	-1,1061
223	[N+](=O)([O-])C1=CC=[N+](C2=CC=CC=C12)[O-]	Training	1,9777	0,1478
224	[N+](=O)([O-])C=1C2=CC=CC=C2C=C2C=CC=CC12	Training	-0,3979	1,0311
225	[N+](=O)([O-])C1=C2C(=C3C=CC=4C=CC=C5CCC1=C3C54)C=CC=C2	Prediction	0,301	2,4665
226	[N+](=O)([O-])C1=CC=C2C=CC=C3C(=O)C4=CC=CC=C4C1=C23	Training	3,7451	2,6417
227	[N+](=O)([O-])C1=CC=2C3=CC=CC=C3C(C3=CC=CC(=C1)C23)=O	Training	2,2041	2,716
228	[N+](=O)([O-])C=1C=CC=2C3=CC=CC=C3C(C3=CC=CC1C23)=O	Training	5,3189	2,7135
229	[N+](=O)([O-])C=1C=C2C(C3=CC=CC4=CC=CC(C2=CC1)=C43)=O	Prediction	4,9288	2,6801
230	[N+](=O)([O-])C=1C=CC=C2C(C3=CC=CC4=CC=CC(C12)=C43)=O	Training	0,7782	2,5961
231	[N+](=O)([O-])C1=CC=C2C=CC=C3C(=O)C4=CC(=CC=C4C1=C23)[N+](=O)[O-]	Training	4,618	3,5897
232	[N+](=O)([O-])C=1C=CC=2C3=CC=C(C=C3C(C3=CC=CC1C23)=O)[N+](=O)[O-]	Training	4,668	3,7682

233	[N+](=O)([O-])C=1C=CC=2C3=C(C=CC=C3C(C3=CC=CC1C23)=O)[N+](=O)[O-]	Training	3,5224	3,489
234	[N+](=O)([O-])C=1C=CC=C2C=3C=CC=NC3C=CC12	Training	2,0815	1,3004
235	[N+](=O)([O-])C=1C=C2C=3N=CC=CC3C=CC2=CC1	Training	-0,1647	1,2867
236	[N+](=O)([O-])C=1C=CC=C2C=3N=CC=CC3C=CC12	Training	2,2687	1,2607
237	[N+](=O)([O-])C1=CC=CC=2C=NC3=CC=CC=C3C12	Training	1,3999	1,2168
238	[N+](=O)([O-])C1=C2C=3C=CC=CC3C=NC2=CC=C1	Training	0,1288	1,1776
239	[N+](=O)([O-])C=1C=C2C=3C=CC=CC3C=NC2=CC1	Prediction	2,2128	1,3061
240	[N+](=O)([O-])C1=CC=C2C=3C=CC=CC3C=NC2=C1	Prediction	2,0058	1,3336
241	[N+](=O)([O-])C1=C2C=3C=CC=[N+](C3C=CC2=CC=C1)[O-]	Prediction	-0,1423	1,3263
242	[N+](=O)([O-])C=1C=C2C=3C=CC=[N+](C3C=CC2=CC1)[O-]	Training	1,5567	1,4761
243	[N+](=O)([O-])C=1C=CC=C2C=3C=CC=[N+](C3C=CC12)[O-]	Training	0,9604	1,4652
244	[N+](=O)([O-])C1=C2C=3[N+](=CC=CC3C=CC2=CC=C1)[O-]	Prediction	1,18	1,1699
245	[N+](=O)([O-])C=1C=C2C=3[N+](=CC=CC3C=CC2=CC1)[O-]	Training	2,2257	1,3546
246	[N+](=O)([O-])C=1C=CC=C2C=3[N+](=CC=CC3C=CC12)[O-]	Training	1,2067	1,3401
247	[N+](=O)([O-])C1=CC=CC=2C3=CC=CC=C3[N+](=CC12)[O-]	Training	2,2637	1,2283
248	[N+](=O)([O-])C1=CC=2C=[N+](C3=CC=CC=C3C2C=C1)[O-]	Training	1,709	1,4258
249	[N+](=O)([O-])C=1C=CC=2C=[N+](C3=CC=CC=C3C2C1)[O-]	Prediction	2,1827	1,4605
250	[N+](=O)([O-])C1=C2C=3C=CC=CC3C=[N+](C2=CC=C1)[O-]	Prediction	-0,0173	1,1874

251	[N+](=O)([O-])C1=CC=[N+](C=2C3=C(C=CC=C3C=CC12)[N+](=O)[O-])[O-]	Training	3,4747	2,0643
252	[N+](=O)([O-])C1=CC=[N+](C=2C3=CC=CC(=C3C=CC12)[N+](=O)[O-])[O-]	Prediction	3,2018	2,3097
253	[N+](=O)([O-])C1=C(C)C=CC(=C1)[N+](=O)[O-]	Training	-1,2218	-1,2004
254	[N+](=O)([O-])C1=C(C)C(=CC=C1)[N+](=O)[O-]	Training	-1,699	-1,2947
255	NC1=CC=CC=2C3=CC=C(C=C3CC12)[N+](=O)[O-]	Training	1,7482	1,1024
256	C(C)(=O)NC1=CC=CC=2C3=CC=C(C=C3CC12)[N+](=O)[O-]	Training	1,415	1,5099
257	[N+](=O)([O-])C=1C=C2C=CC=3C=C4C(=C5C=CC(C1)=C2C53)C=CC=C4	Training	3,2422	3,0576
258	[N+](=O)([O-])C1=CC2=C(OC3=C2C=C(C=C3)[N+](=O)[O-])C=C1	Training	2,3667	2,5386
259	[N+](=O)([O-])C1=CC=CC=2OC3=C(C21)C=CC=C3	Training	-0,3702	1,2375
260	[N+](=O)([O-])C=1C=CC2=C(OC3=C2C=CC=C3)C1	Training	1,7829	1,4565
261	[N+](=O)([O-])C1=CC=CC2=C1OC1=C2C=CC=C1	Training	1,0996	1,3062
262	[N+](=O)([O-])C1=CC2=C(OC3=C2C=CC(=C3)[N+](=O)[O-])C=C1	Training	2,3608	2,6081
263	[N+](=O)([O-])C=1C=CC=2CC3=CC=CC=C3C2C1	Training	0,8688	1,1752
264	[N+](=O)([O-])C1=CC=2C(C3=CC=CC=C3C2C=C1)=O	Training	2,4943	1,8495
265	[N+](=O)([O-])C1=CC=C2C(=C3C(=CO2)C=CC=C3)C1=O	Prediction	1,1458	2,0336
266	[N+](=O)([O-])C1=CC2=C(SC3=C2C=CC=C3)C=C1	Training	2,4362	1,7849
267	[N+](=O)([O-])C=1C=CC2=C(SC3=C2C=CC=C3)C1	Training	2,6607	1,8521

268	[N+](=O)([O-])C=1C=C2C(=C3C(=CO2)C=CC=C3)C(C1)=O	Training	1,493	1,9846
269	[N+](=O)([O-])C1=C(C=CC(=C1)N)C1=CC=C(N)C=C1	Training	-1,2007	0,4422
270	[N+](=O)([O-])C=1C=C(C=CC1N)C1=CC=C(N)C=C1	Prediction	0,5987	1,0017
271	[N+](=O)([O-])C=1C=C(C=CC1N)C1=CC(=C(N)C=C1)[N+](=O)[O-]	Training	-0,2351	0,4635
272	[N+](=O)([O-])C1=CC2=C(C3=CC=4CCCCC4C=C3CC2)C=C1	Training	2,7853	1,5072
273	ClC=1C=C(C=CC1[N+](=O)[O-])C1=CC(=C(C=C1)[N+](=O)[O-])Cl	Training	1,9542	3,0535
274	ClC=1C=C(C=CC1N)C1=CC(=C(C=C1)[N+](=O)[O-])Cl	Training	1,9031	2,0811
275	ClC=1C=C(C=CC1NC(C)=O)C1=CC(=C(C=C1)[N+](=O)[O-])Cl	Training	2,6628	2,5263
276	ClC=1C=C(C=C(C1N)[N+](=O)[O-])C1=CC(=C(C=C1)[N+](=O)[O-])Cl	Prediction	4,8195	2,7164
277	[N+](=O)([O-])C=1C=CC=2C3=C(C4=CC=C(C=5C=CC1C2C45)[N+](=O)[O-])C=CC=C3	Training	5,4548	3,7873

Liste des références bibliographiques

Références bibliographiques

[1] *Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology*. Récupéré de https://doi.org/10.1021/cr100222q

[2] *Evaluation of QSAR Models for the Prediction of Ames Genotoxicity: A Retrospective Exercise on the Chemical Substances Registered Under the EU REACH Regulation*. Récupéré de https://doi.org/10.1080/10590501.2014.938955

[3] *Improvement of quantitative structure–activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project*. Récupéré de https://doi.org/10.1093/mutage/gey031

[4] Ligue Contre Le Cancer. (2023, Août 2). *Le cancer, définition*. Récupéré de ligue-cancer.net/articles

[5] *Facteurs de risque et prévention des cancers*. Récupéré de https://www.ameli.fr/assure/sante/themes/cancers/facteurs-risques-prevention

[6] République Française Institut Nationale Du Cancer. (2021, Janvier 6). *Facteurs de risque*. Récupéré de https://www.e-cancer.fr/Comprendre-prevenir-depister/Qu-est-ce-qu-un-cancer/Facteurs-de-risque

[7] *Les facteurs de risque du cancer - Richard Béliveau*. Récupéré de paroconseil.ca

[8] *Biotechnologies - Mutagenèse*. Récupéré de https://www.semae-pedagogie.org/sujet/biotechnologies-mutagenese/

[9] Krahn, M., Labelle, V., Borges, A., Bartoli, M., & Lévy, N. (2010). *Exclusion of Mutations in the Dysferlin Alternative Exons 1 of DYSF-v1, 5a, and 40a in a Cohort of 26 Patients*. Genetic Testing and Molecular Biomarkers, p. 153-154.

[10] Ehrenberg, L. (1960). *Chemical mutagenesis: Biochemical and chemical point of view on mechanisms of action*. Chemische Mutagènes, 124-136.

[11] Peter, D., Moore, T., Samuel, D., Rabkin, L., Osborn, I., Charles, M., King, S., & Bernard, S. (1982). *Effect of acetylated and deacetylated 2-aminofluorene adducts on in vitro DNA synthesis*. National Academic Science, 79:7166-7170.

[12] Hanna, K. (2005). *Chromosome breakage at high dose rates*. Mutation Research, 182:270-271.

[13] Eastmond, D. A., Hartwig, A., Anderson, D., Anwar, W. A., Cimino, M. C., Dobrev, I., Douglas, G. R., Nohmi, T., Phillips, D. H., & Vickers, C. (2009). *Mutagenicity testing for chemical risk assessment: update of the WHO/IPCS Harmonized Scheme*. Mutagenesis, 24(4), 341–349.

[14] Park, C. G., & Lim, H. B. (2024). *Evaluation of Antimutagenic and Antioxidant Properties in Fomes fomentarius L.: Potential Development as Functional Food*. Applied Sciences.

[15] IARC. (2010). *some non-heterocyclic polycyclic aromatic hydrocarbons and some related exposures*. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 92.

[16] IARC. (2017). *some organochlorine pesticides*. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 113.

[17] UNSCEAR. (2008). *Sources and effects of ionizing radiation*. United Nations Scientific Committee on the Effects of Atomic Radiation.

[18] IARC. (2012). *Radiation, including ultraviolet radiation, and skin cancer*. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 100D.

[19] IARC. (2012). *Biological agents. Volume 100 B: A review of human carcinogens*. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans.

[20] Wild, C. P., et al. (2020). *Cancer prevention: Insights gained from the field and future perspectives*. International Journal of Cancer, 146(6), 1537-1546.

[21] Schüz, J., et al. (2021). *Challenges in the assessment of occupational exposure to carcinogens in Europe*. International Journal of Environmental Research and Public Health, 18(3), 1276.

[22] Boesch, R., et al. (2017). *Human and environmental toxicity of chemicals in household products: Questions from the Clinician's Point of View*. International Journal of Environmental Research and Public Health, 14(2), 134.

[23] Galloway, S. M., et al. (2020). *Approaches to genotoxicity testing: A regulatory perspective*. Environmental and Molecular Mutagenesis, 61(2), 116-131.

[24] Grosdidier, A. (2007). *Conception d'un logiciel de docking et applications dans la recherche de nouvelles molécules actives*. Thèse de doctorat en Pharmacie : Université Joseph Fourier, France.

[25] Sethi, A., Joshi, K., Sasikala, K., & Alvala, M. (2020). *Molecular Docking in Modern Drug Discovery: Principles and Recent Applications*. Drug Discovery and Development - New Advances. doi: 10.5772/intechopen.85991.

[26] *Processus de développement d'un nouveau médicament de l'identification de la cible*. Récupéré de researchgate.net

[27] DiMasi, J. A., & Grabowski, H. G. (2007). *The cost of biopharmaceutical R&D: is biotech different?* Managerial and Decision Economics, 28(4-5), 469-479.

[28] Arrowsmith, J. (2011). *Trial watch: phase III and submission failures: 2007-2010*. Nature Reviews Drug Discovery, 10(2), 87-87.

[29] Smith, D. A., & van de Waterbeemd, H. (1999). *Pharmacokinetics and metabolism in drug design* (Vol. 51). John Wiley & Sons.

[30] Pritchard, J. F., & Jurima-Romet, M. (1997). *Pharmacokinetic optimization in drug research: biological, physicochemical, and computational strategies* (Vol. 46) CRC press.

[31] European Medicines Agency (EMA). (s.d.). *Clinical trials*. Récupéré de https://www.ema.europa.eu/en/human-regulatory/research-development/clinical-trials; Chow, S. C., & Liu, J. P. (2008). *Design and analysis of clinical trials: concepts and methodologies* (Vol. 2). John Wiley & Sons.

[32] Chow, S. C., & Liu, J. P. (2008). *Design and analysis of clinical trials: concepts and methodologies* (Vol. 2). John Wiley & Sons.

[33] Pocock, S. J. (1983). *Clinical trials: a practical approach*. John Wiley & Sons.

[34] Chow, S. C., & Liu, J. P. (2008). *Design and analysis of clinical trials: concepts and methodologies* (Vol. 2). John Wiley & Sons.

[35] FDA. (2006). *Guidance for industry: postmarketing studies and clinical trials—implementation of section 505(o)(3) of the federal food, drug, and cosmetic act*. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER).

[36] Hopkins, A. L., & Groom, C. R. (2002). *The druggable genome*. Nature Reviews Drug Discovery, 1(9), 727-730.

[37] Wang, R., Lu, Y., & Wang, S. (2003). *Comparative evaluation of 11 scoring functions for molecular docking*. Journal of medicinal chemistry, 46(12), 2287-2303.

[38] Keserű, G. M., & Makara, G. M. (2009). *The influence of lead discovery strategies on the properties of drug candidates*. Nature Reviews Drug Discovery, 8(3), 203-212.

[39] Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., … & Zhou, J. (2011). *Impact of high-throughput screening in biomedical research*. Nature Reviews Drug Discovery, 10(3), 188-195.

[40] Egan, W. J., Merz Jr, K. M., & Baldwin, J. J. (2000). *Prediction of drug absorption using multivariate statistics*. Journal of medicinal chemistry, 43(21), 3867-3877.

[41] Tapprich, W. E., Reichart, L., Simon, D. M., Duncan, G., McClung, W., Grandgenett, N., & Pauley, M. A. (2020). *Biochemistry and Molecular Biology Education* (Vol. 49, Issue 1). doi: 10.1002/bmb.21361.

[42] Guiducci, C. (2013). *ACM Journal on Emerging Technologies in Computing Systems* (Vol. 9, Issue 4, Article No.: 26). doi: 10.1145/2536744.

[43] Gusfield, D. (2004). *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (Vol. 1, Issue 1). doi: 10.1109/TCBB.2004.9.

[44] *CHEMOINFORMATICS: THEORY, PRACTICE, & PRODUCTS*.

[45] Fernández-de Gortari, E., García-Jacas, C. R., Martinez-Mayorga, K., & Medina Franco, J. L. (2017). *Database fingerprint (DFP): an approach to represent molecular databases*. J. Cheminform, 9. doi:10.1186/s13321-017-0195-1.

[46] Hodgson, E. (2010). *A Textbook of Modern Toxicology* (4th ed.). Wiley-Blackwell.

[47] Klaassen, C. D., Watkins III, J. B., & Casarett, L. J. (2019). *Casarett & Doull's Essentials of Toxicology* (3rd ed.). McGraw-Hill Education.

[48] Hayes, A. W. (Ed.). (2014). *Principles and Methods of Toxicology* (6th ed.). CRC Press.

[49] TETE, A., ZGHEIB, E., AL AWABDH, S., BENOIT, L., BERNAL, K., DUARTE HOSPITAL, C., LARIGOT, L., LOPEZ SUAREZ, L., ANDREAU, K., CHAUVET, C., KIM, M. J., KOUAL, M., TOMKIEWICZ-RAULET, C., COUMOUL, X., BLANC, É., AUDOUZE, K., & BORTOLI, S. (2022). *Alternatives in vitro et in silico aux modèles animaux en toxicologie*. doi: RE295 v1.

[50] Pound, P. (2020). *Safer Medicines Trust*, P.O. Box 122, Kingsbridge TQ7 9AX, UK.

[51] Kumari, M., Singla, M., & Sobti, R. C. (2022). *Advances in Animal Experimentation and Modeling Understanding Life Phenomena* (Page 87). doi:10.1016/B978-0-323-90583-1.00014-3.

[52] Balls, M., & Combes, R. D. (Eds.). (1991). *Animal Alternatives, Welfare and Ethics*. Elsevier.

[53] Treuting, P. M., & Dintzis, S. M. (2012). *Comparative Anatomy and Histology: A Mouse, Rat, and Human Atlas*. Academic Press.

[54] Greim, H., & Snyder, R. (Eds.). (2017). *Toxicology and Risk Assessment: Principles, Methods, and Applications*. Wiley.

[55] Casarett, L. J., Klaassen, C. D., & Doull, J. (Eds.). (2013). *Casarett & Doull's Toxicology: The Basic Science of Poisons* (8th ed.). McGraw-Hill Education.

[56] Haynes, W. M. (Ed.). (2014). *CRC Handbook of Chemistry and Physics* (95th ed.). CRC Press.

[57] *Courbes dose-réponse de deux toxiques A & B avec les logarithmes des doses toxiques 10 et 90*. Récupéré de researchgate.net

[58] Hayashi, M., Kamata, E., Hirose, A., Takahashi, M., Morita, T., & Ema, M. (2005). *In silico assessment of chemical mutagenesis in comparison with results of Salmonella microsome assay on 909 chemicals*. Mutat. Res. Toxicol. Environ. Mutagen., 588(2), 129–135. doi: 10.1016/j.mrgentox.2005.09.009.

[59] Schneider, G. (2018). *In Silico Methods in Drug Discovery*. Royal Society of Chemistry.

[60] Tropsha, A. (2010). *Best Practices for QSAR Model Development, Validation, and Exploitation*. Molecular Informatics, 29(6-7), 476–488.

[61] Nikolova-Jeliazkova, N., & Jaworska, J. (2005). *An Approach to Determining Applicability Domains for QSAR Group Contribution Models: An Analysis of SRC KOWWIN*. Altern. to Lab. Anim., 33(5), 461–470. doi: 10.1177/026119290503300510.

[62] HANSCH, C., MALONEY, P. P., FUJITA, T., & MUIR, R. M. (1962). *Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients*. Nature, 194(4824), 178–180. doi: 10.1038/194178b0.

[63] Ghose, A. K., & Crippen, G. M. (1987). *Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions*. J. Chem. Inf. Comput. Sci., 27(1), 21–35. doi: 10.1021/ci00053a005.

[64] Kubinyi, H. (2002). *From Narcosis to Hyperspace: The History of QSAR*. Quant. Struct. Relationships, 21(4), 348–356. doi: 10.1002/1521-3838(200210)21:4<348::AID-QSAR348>3.0.CO;2-D.

[65] Selassie, C., & Verma, R. P. (2010). *History of Quantitative Structure–Activity Relationships*. In Burger's Medicinal Chemistry and Drug Discovery. doi: 10.1002/0471266949.bmc001.pub2.

[66] Yousefinejad, S., & Hemmateenejad, B. (2015). *Chemometrics tools in QSAR/QSPR studies: A historical perspective*. Chemom. Intell. Lab. Syst., 149, 177–204. doi: 10.1016/j.chemolab.2015.06.016.

[67] Roy, K., Kar, S., & Das, R. N. (2015). *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*.

[68] Danishuddin & Khan, A. U. (2016). *Descriptors and their selection methods in QSAR analysis: paradigm for drug design*. Drug Discov. Today, 21(8), 1291–1302. doi: 10.1016/j.drudis.2016.06.013.

[69] Hansch, C., Leo, A., & Hoekmann, D. (1995). *Exploring QSAR: hydrophobic, electronic and steric constants*. Washington, DC: American Chemical Society.

[70] Dudek, A. Z., Arodz, T., & Gàlvez, J. (2006). *Computational methods in developing quantitative structure-activity relationships (qsar): A review*. Combinatorial Chemistry & High Throughput Screening, 9, 213–228.

[71] Regina Todeschini, A., & Hakomori, S. (2008). *Functional role of glycosphingolipids and gangliosides in control of cell adhesion, motility, and growth, through glycosynaptic microdomains*. Biochim. Biophys. Acta - Gen. Subj., 1780(3), 421–433. doi: 10.1016/j.bbagen.2007.10.008.

[72] Wiener, H. (1947). *Structural Determination of Paraffin Boiling Points*. J. Am. Chem. Soc., 69(1), 17–20. doi: 10.1021/ja01193a005.

[73] Labute, P. (2000). *A widely applicable set of descriptors*. J. Mol. Graph. Model., 18(4–5), 464–477. doi: 10.1016/S1093-3263(00)00068-1.

[74] Viswanadhan, V. N., Ghose, A. K., Revankar, G. R., & Robins, R. K. (1989). *Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain*. J. Chem. Inf. Model., 29(3), 163–172. doi: 10.1021/ci00063a006.

[75] Brown, N. (2009). *Chemoinformatics—an introduction for computer scientists*. ACM Comput. Surv., 41(2), 1–38. doi: 10.1145/1459352.1459353.

[76] Gadaleta, D., Lombardo, A., Toma, C., & Benfenati, E. (2018). *A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications*. J. Cheminform., 10(1), 60. doi: 10.1186/s13321-018-0315-6.

[77] Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. New York, NY: Springer New York. doi: 10.1007/978-0-387-21606-5.

[78] Tomasz Puzyn, M. T. C., & Leszczynski, J. *Recent Advances in QSAR Studies: Methods and Applications* (Volume 8 of Challenges and Advances in Computational Chemistry and Physics).

[79] Xu, L., & Yu, K. (2014). *A Note on Dynamic Roy's Identity*. Theor. Econ. Lett., 04(07), 513–516. doi: 10.4236/tel.2014.47064.

[80] Gini, G. C. (2019). *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*.

[81] Zefirov, N. S., & Palyulin, V. A. (2001). *QSAR for Boiling Points of 'Small' Sulfides. Are the 'High-Quality Structure-Property-Activity Regressions' the Real High Quality QSAR Models?* J. Chem. Inf. Comput. Sci., 41(4), 1022–1027. doi: 10.1021/ci0001637.

[82] Polishchuk, P. (2017). *Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future*. J. Chem. Inf. Model., 57(11), 2618–2639. doi: 10.1021/acs.jcim.7b00274.

[83] Wolpert, D. H., & Macready, W. G. (1997). *No free lunch theorems for optimization*. IEEE Trans. Evol. Comput., 1(1), 67–82. doi: 10.1109/4235.585893.

[84] Topliss, J. G., & Edwards, R. P. (1979). *Chance factors in studies of quantitative structure-activity relationships*. J. Med. Chem., 22(10), 1238–1244. doi: 10.1021/jm00196a017.

[85] *ETUDE QSAR DE LA TOXICITE D'UNE SERIE*. Récupéré de dspace.univ-tlemcen.dz

[86] March, J. (Ed.). (2007). *Advanced organic chemistry: Reactions, mechanisms, and structure* (6th ed.). John Wiley & Sons.

[87] Smith, M. B., & March, J. (Eds.). (2007). *March's advanced organic chemistry: Reactions, mechanisms, and structure* (7th ed.). John Wiley & Sons.

[88] Gomberg, M. (1890). *An instance of trivalent nitrogen: Triphenylmethyl*. Journal of the American Chemical Society, 12(8), 501-508.

[89] March, J. (2007). *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure* (6th ed.). John Wiley & Sons.

[90] Clayden, J., Greeves, N., Warren, S., & Wothers, P. (2012). *Organic Chemistry*. Oxford University Press.

[91] Vogel, A. I., Furniss, B. S., Hannaford, A. J., Smith, P. W. G., & Tatchell, A. R. (2001). *Vogel's Textbook of Practical Organic Chemistry* (5th ed.). Pearson Education.

[92] Smith, J., et al. (2018). *Anticancer activity of a novel nitro-aromatic compound against lung cancer cells*. Cancer Chemotherapy and Pharmacology, 82(5), 923-932.

[93] Zhang, X., et al. (2020). *Antibacterial activity of a nitro-aromatic derivative against antibiotic-resistant bacterial strains*. Journal of Antibiotics, 73(2), 113-121.

[94] Pullman, B., & Pullman, A. (1980). *Carcinogenesis: Fundamental Mechanisms and Environmental Effects*. D. Reidel: Dordrecht, the Netherlands.

[95] Nelson, S. D. (1982). *J. Med. Chem.*, 25, 753.

[96] Vijay, U., Gupta, S., Mathur, P., Suravajhala, P., & Bhatnagar, P. (2018). *Microbial Mutagenicity Assay: Ames Test*. Vol 8, Iss 06, Mar 20, 2018. doi:10.21769/BioProtoc.2763.

[97] Mullin, C. A., Rashid, K. A., & Mumma, R. O. (1987). *Mutagenic potency of some conjugated nitroaromatic compounds and its relationship to structure*. Elsevier, Pesticide Research Laboratory and Graduate Study Center, Department of Entomology, Pennsylvania State University, University Park, PA 16802 (U.S.A.) 2 March 1987. doi: 0165121887900036.

[98] *Ames test*. Récupéré de image.slidesharecdn.com

[99] YARNELL, A. (2015). *Happy Birthday, ChemDraw*. Chem. Eng. News Arch., vol. 93, no. 27, p. 3, Jul. 2015. doi: 10.1021/cen-09327-editorial.

[100] *Chemdraw Ultra*. « Molecular modelling, structure drawing, Semi-empirical calculations, structure display, MOPAC, solvation energy, MM2 ». Copyright Cambridge Soft Corporation, 2002.

[101] Hinchliffe, A. (2001). *HyperChem Release 4.5 for Windows*. Electron. J. Theor. Chem., vol. 1, no. 1, pp. 62–64, Mar. 2001. doi: 10.1002/ejtc.9.

[102] Hypercube, Inc., Gainesville, Florida, *HyperChem*. 1999. [Online]. Available: http://www.hyper.com

[103] Mauri, A., Bertola, M., & Alvascience. (2022). *A New Software Suite for the QSAR Workflow Applied to the Blood–Brain Barrier Permeability*. Int. J. Mol. Sci. 2022, 23, 12882. doi: 10.3390/ ijms232112882.

[104] Gramatica, P., Chirico, N., Papa, E., Cassani, S., & Kovarich, S. (2013). *QSARINS: A new software for the development, analysis, and validation of QSAR MLR models*. J. Comput. Chem., vol. 34, no. 24, pp. 2121–2132, Sep. 2013. doi: 10.1002/jcc.23361.

[105] Gramatica, P., Cassani, S., & Chirico, N. (2014). *QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS*. J. Comput. Chem., vol. 35, no. 13, pp. 1036–1044, May 2014. doi: 10.1002/jcc.23576.

[106] *KNIME Analytics Platform: Professional open-source software*. (n.d.). https://www.knime.org.

[107] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, P., Ohl, P., Thiel, K., & Wiswedel, B. (2009). *KNIME - The Konstanz Information Miner*. SIGKDD Explor. 11. doi:10.1145/1656274.1656280.

[108] Warr, W. A. (2012). *Scientific workflow systems: Pipeline Pilot and KNIME*. J. Comput. Aided. Mol. Des. 26. doi:10.1007/s10822-012-9577-7.

[109] Chichester, C., Digles, D., Siebes, R., Loizou, A., Groth, P., & Harland, L. (2015). *Drug discovery FAQs: Workflows for answering multidomain drug discovery questions*. Drug Discov. Today. 20. doi:10.1016/j.drudis.2014.11.006.

[110] *KNIME Analytics Platform Workbench Guide*. Récupéré de docs.knime.com

[111] Gini, G., & Zanoli, F. (2020). *Machine Learning and Deep Learning Methods in Ecotoxicological QSAR Modeling*. doi:10.1007/978-1-0716-0150-1_6.

[112] Breiman, L. (2001). *Random forests*. Machine learning 45, no. 1.

[113] Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. Annals of statistics.

[114] Zhang, H. (2001). *The optimality of Naive Bayes*. In Proceedings of the seventeenth international joint conference on Artificial intelligence. Morgan Kaufmann Publishers Inc.

[115] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

[116] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

[117] Benfenati, E. (2023). Istituto di Ricerche Farmacologiche “Mario Negri” Via Mario Negri 2, 20156, Milano (Italy). First edition.

[118] Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*. Journal of Machine Learning Technologies, 2(1).

[119] Zhou, X. H., McClish, D. K., & Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine*. Wiley-Interscience.

[120] Davis, J., & Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves*. In Proceedings of the 23rd international conference on Machine learning. ACM.

[121] Sokolova, M., & Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks*. Information Processing & Management, 45(4).

[122] *Mudrush*. Récupéré de icarn.ubc.ca

[123] Burello, E. (2017). *Review of (Q)SAR models for regulatory assessment of nanomaterials risks*. NanoImpact, vol. 8. doi: 10.1016/j.impact.2017.07.002.

[124] Besalu, E., De Julian-Ortiz, J. V., & Pogliani, L. (2007). *Trends and plot methods in MLR studies*. J. Chem. Inf. Model, 47. doi: 10.1021/ci6004959.

[125] Doreswamy, H., & Vastrad, C. M. *PREDICTIVE COMPARATIVE QSAR ANALYSIS OF AS 5-NITROFURAN-2-YL DERIVATIVES MYCO BACTERIUM TUBERCULOSIS*.

[126] Tropsha, A., Gramatica, P., & Gombar, V. (2003). *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models*. QSAR Comb. Sci., 22(1), 69–77. doi: 10.1002/qsar.200390007.

[127] Shen, M., Béguin, C., Golbraikh, A., Stables, J. P., Kohn, H., & Tropsha, A. (2004). *Application of Predictive QSAR Models to Database Mining: Identification and Experimental Validation of Novel Anticonvulsant Compounds*. J. Med. Chem., 47(9), 2356–2364. doi: 10.1021/jm030584q.

[128] Shen, M., Béguin, C., Golbraikh, A., Stables, J. P., Kohn, H., & Tropsha, A. (2004). *Application of Predictive QSAR Models to Database Mining: Identification and Experimental Validation of Novel Anticonvulsant Compounds*. J. Med. Chem., 47(9), 2356–2364. doi: 10.1021/jm030584q.

[129] Gramatica, P. (2007). *Principles of QSAR models validation: internal and external*. QSAR Comb. Sci., 26(5), 694–701. doi: 10.1002/qsar.200610151.

[130] Tropsha, A., & Golbraikh, A. (2007). *Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening*. Curr. Pharm. Des., 13(34), 3494–3504. doi: 10.2174/138161207782794257.

[131] Golbraikh, A., & Tropsha, A. (2002). *Beware of q2!*. J. Mol. Graph. Model., 20(4), 269–276. doi: 10.1016/S1093-3263(01)00123-1.

[132] Tropsha, A. (2010). *Best Practices for QSAR Model Development, Validation, and Exploitation*. Mol. Inform., 29(6–7), 476–488. doi: 10.1002/minf.201000061.

[133] Schüürmann, G., Ebert, R.-U., Chen, J., Wang, B., & Kühne, R. (2008). *External Validation and Prediction Employing the Predictive Squared Correlation Coefficient — Test Set Activity Mean vs Training Set Activity Mean*. J. Chem. Inf. Model., 48(11), 2140–2145. doi: 10.1021/ci800253u.

[134] Linusson, A., Elofsson, M., Andersson, I. E., & Dahlgren, M. K. (2010). *Statistical Molecular Design of Balanced Compound Libraries for QSAR Modeling*. Curr. Med. Chem., 17(19), 2001–2016. doi: 10.2174/092986710791233661.

[135] Roy, K., Mitra, I., Kar, S., Ojha, P. K., Das, R. N., & Kabir, H. (2012). *Comparative Studies on Some Metrics for External Validation of QSPR Models*. J. Chem. Inf. Model., 52(2), 396–408. doi: 10.1021/ci200520g.

[136] ZAKARIA, R., & AMINE, S. M. (2021). *Analyse de la relation quantitative Structure / Activité des inhibiteurs de la tyrosine kinase*. Université Frères Mentouri Constantine 1.