الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة عمار ثليجي– الاغواط
Université Amar Thelidji- Laghouat

كلية العلوم
FACULTE : SCIENCES
DEPARTEMENT DE BIOLOGIE

MEMOIRE DE MASTER

Domaine : Sciences de la Nature et de la Vie
Filière : Sciences Biologiques
Spécialité : Pharmaco-toxicologie

Thème

Développement d'un modèle prédictif d'apprentissage automatique pour les applications de conception de médicaments

Présenté Par :

- **Melik Yamina**
- **Terbagou Amani Khouloud**

Jury de soutenance:

Promoteur : KADI Imededdine	**URPM - Université Laghouat**
Co-promoteur :	**Université Laghouat**
Président :	**Université Laghouat**
Examinateur :	**Université Laghouat**

Année universitaire 2024-2025

إهـداء

إلى خالقِ الروح والقلم،

وبارئِ الذرّ والنَسَم،

وخالقِ كلّ شيءٍ من العدم.

لكَ الحمد كما ينبغي لجلال وجهك وعظيم سلطانك.

إلى من بلّغ الرسالة، وأدّى الأمانة،

ونصح الأمة، نبيِّ الرحمة، ونور العالمين.

صلى الله عليك يا خير الورى ما ضاء نورٌ في الصَبّاح وأسفرَا ﷺ.

إلى كل من كان سبباً، ولو بكلمة...

كمن أسند القلب بشقّ تمرة، فكان العون حين قلّ الزاد.

إلى أولئك الذين لم يعرفوا أن حضورهم في حياتنا كان تِرْياقًا...

شكراً لكم.

وآخر دعوانا أنِ الحمدُ للهِ ربِّ العالمين .

أماني ، أمينة

Remerciement

Louange à Allah qui a enseigné par la plume, qui a enseigné à l'homme ce qu'il ne savait pas, et qui nous a comblés de la grâce de l'effort et de la connaissance. À Lui la louange jusqu'à ce qu'Il soit satisfait, à Lui la louange lorsqu'Il est satisfait, et à Lui la louange après la satisfaction.

À la fin de ce parcours, où les sentiments de fierté se mêlent à l'humilité, nous ne pouvons que rédiger des mots de gratitude sincères, en signe de reconnaissance envers tous ceux qui ont eu un impact sur cette modeste réalisation.

Nous exprimons notre gratitude et notre reconnaissance à notre cher professeur Dr. Imadeddine Kadi, qui n'était pas seulement un encadrant, mais aussi un mentor noble, inspirant et généreux, mérite toute notre reconnaissance pour son humilité, sa patience et ses conseils qui ont été un phare sur notre chemin scientifique.

Nous exprimons également notre profonde gratitude à tous nos professeurs de la Faculté des Sciences de la Nature et de la Vie, en particulier dans la spécialité de Pharmacie et de Toxicologie, qui ont en nous inculqué l'amour du savoir et l'esprit de recherche scientifique, et qui ont été pour nous des modèles de diligence et de dévouement.

Et n'oublions pas nos camarades de la promotion (2025/2024), vous êtes la plus belle partie du voyage. Merci pour les souvenirs, le soutien et les rires inoubliables.

En conclusion, nous dédions ce travail à tous ceux qui ont cru en nous, nous ont encouragés et nous ont poussés un pas en avant. À vous tous, nous offrons les fruits de cet effort, en demandant à Dieu de le bénir et de le rendre pur pour Sa noble cause.

Résumé

Ce travail vise à développer un modèle prédictif basé sur des techniques d'intelligence artificielle pour estimer l'activité biologique des composés chimiques vis-à-vis de l'enzyme CYP3A4, l'une des principales enzymes dans le métabolisme des médicaments dans le corps. Les données biologiques ont été extraites de la base ChEMBL, où elles ont été purifiées et converties des valeurs IC50 en pIC50, avant de classer les molécules selon leurs niveaux d'activité. Ensuite, l'outil PaDEL-Descriptor a été utilisé pour extraire les descripteurs moléculaires, suivi de la phase de construction des modèles prédictifs en utilisant plusieurs algorithmes d'apprentissage automatique, les plus importants étant Gradient Boosting, Random Forest et SVR. Après évaluation, le modèle Gradient Boosting a montré une performance exceptionnelle avec un coefficient de détermination R^2 de 0,896, surpassant les autres modèles en termes de précision et de réduction des valeurs RMSE et MAE, ce qui indique son efficacité élevée dans la prédiction de l'efficacité des composés. Ces résultats reflètent la valeur ajoutée des techniques d'apprentissage automatique dans le domaine de la conception de médicaments, en particulier en ce qui concerne l'accélération du criblage virtuel et la réduction des coûts associés aux essais biologiques traditionnels, tout en mettant en évidence la capacité de ces modèles à soutenir la prise de décision lors des phases de découverte de nouvelles composés pharmaceutiques.

ملخص

يهدف هذا العمل إلى تطوير نموذج تنبؤي يعتمد على تقنيات الذكاء الاصطناعي لتقدير النشاط البيولوجي للمركبات الكيميائية تجاه الإنزيمCYP3A4 ، أحد الإنزيمات الرئيسية في استقلاب الأدوية داخل الجسم. وقد تم استخراج البيانات الحيوية من قاعدةChEMBL ، حيث خضعت لعملية تنقية وتحويل من قيم IC50 إلىpIC50 ، قبل تصنيف الجزيئات حسب مستويات نشاطها. ثم استُخدمت أداة PaDEL-Descriptor لاستخراج الأوصاف الجزيئية، تلتها مرحلة بناء النماذج التنبؤية باستعمال خوارزميات متعددة من التعلم الآلي، أهمهاRandom Forest ، Gradient Boosting و .SVRبعد التقييم، أظهر نموذج Gradient Boosting أداءً متميزًا محققًا معامل تحديد R^2 بلغ 0.896، متفوقًا على باقي النماذج من حيث الدقة وانخفاض قيم RMSE وMAE، مما يدل على كفاءته العالية في توقع فعالية المركبات. تعكس هذه النتائج القيمة المضافة لتقنيات التعلم الآلي في مجال تصميم الأدوية، خاصة في ما يتعلق بتسريع عملية الفرز الافتراضي وتقليل التكاليف المرتبطة بالتجارب البيولوجية التقليدية، كما تبرز قدرة هذه النماذج على دعم اتخاذ القرار في مراحل اكتشاف المركبات الدوائية الجديدة.

Abstract

This work aims to develop a predictive model based on artificial intelligence techniques to estimate the biological activity of chemical compounds towards the CYP3A4 enzyme, one of the main enzymes in drug metabolism within the body. The biological data were extracted from the ChEMBL database, where they underwent a purification and conversion process from IC50 values to pIC50, before classifying the molecules according to their activity levels. Then, the PaDEL-Descriptor tool was used to extract molecular descriptors, followed by the model building phase using various machine learning algorithms, the most important of which are Gradient Boosting, Random Forest, and SVR. After evaluation, the Gradient Boosting model demonstrated outstanding performance, achieving an R^2 coefficient of 0.896, surpassing other models in terms of accuracy and lower RMSE and MAE values, indicating its high efficiency in predicting the efficacy of compounds. These results reflect the added value of machine learning techniques in the field of drug design, particularly in terms of accelerating the virtual screening process and reducing the costs associated with traditional biological experiments. They also highlight the ability of these models to support decision-making in the stages of discovering new pharmaceutical compounds.

Table des matières

Résumé 4

ملخص 5

Abstract 6

Introduction 9

Chapitre 1 :**Machine Learning**

1. Introduction 13

2. Définition d'apprentissage automatique 14

3. Les types de système de Machine Learning 14

3.1. Apprentissage supervisé 14

3.2. Apprentissage non supervisé 15

3.3. Apprentissage semi supervisé 16

3.4. Apprentissage par renforcement 16

4. Les étapes de l'apprentissage automatique 17

4.1. Collecte et préparation des données 17

4.2. Choix de l'algorithme et du modèle 18

4.3. Entraînement du modèle 18

4.4. Évaluation et validation 18

4.5. Optimisation 19

4.6. Déploiement et surveillance du modèle 19

5. Algorithmes d'apprentissage automatique 20

5.1. Régression linéaire (Linear Regression) 20

5.2. Régression logistique (Logistic Regression) 20

5.3. Arbres de décision (Decision Trees) 20

5.4. Forêts aléatoires (Random Forest) 20

5.5. Machine à vecteurs de support (SVM) 20

5.6. K plus proches voisins (K-Nearest Neighbors, KNN) 21

5.7. Naïve Bayes 21

5.8. K-means 21

5.9. Gradient Boosting 21

5.10. Réseaux de neurones (Neural Networks, MLP & Deep) 21

Chapitre 2 : **Drug Design**

1. Introduction 21

2. Définition de la conception des médicaments (Drug Design) 21
3. Le processus de découvert de médicament (Drug discovery process) 21
3.1. Identification de la cible biologique 22
3.2. Criblage des composés 22
3.3. Optimisation des pistes 22
3.4. Études précliniques 22
3.5. Essais cliniques 23
3.6. Approbation réglementaire et mise sur le marché 23
3.7. Repositionnement des médicaments 24
4. Cibles thérapeutiques en drug design 24
5. Les approches de drug design 25
5.1. Drug Design rationnel (RDD) 25
5.2. Approches basées sur la structure (Structure-Based Drug Design, SBDD) 25
5.3. Approches basées sur le ligand (Ligand-Based Drug Design, LBDD) 25
5.4. Drug Design de novo 26
6. Méthodes de représentation moléculaire pour l'apprentissage automatique 26
Chapitre 3 : **Matériels et Méthodes**
1. Récupération des données de la base CHEMBEL 28
2. Pré-traitement de données 29
3. Normalisation des données 31
4. Convertir IC_{50} en pIC_{50} 31
5. Exploration des données analysées 32
6. Téléchargement de PaDEL-Descriptors 33
7. Préparation des matrices de données X et Y 34
8. Construction des modèles de régression 35
Chapitre 4 :Résultats et Discussion
1. Récupération des données de la base CHEMBEL 38
2. Distribution des unités de calcul de IC50 38
3. Classification des molécules sélectionnées 39
4. La normalisation de données de l'activité pharmacologique 40
5. La Conversion de IC_{50} en pIC_{50} 40
6. Calcul et filtration des fingerprints moléculaires 42
7. Séparation des données et construction du model 44

Conclusion 48
Références bibliographiques.....49

Liste des figures

N°	Titre Figure	Page
Fig. I.01	Structuration des domaines de l'intelligence artificielle.	**13**
Fig. I.02	Apprentissage supervisé.	**15**
Fig. I.03	Apprentissage non supervisé.	**15**
Fig. I.04	Apprentissage semi supervisé.	**16**
Fig. I.05	Apprentissage par renforcement.	**17**
Fig. III.01	Répartition des molécules étudiées pour la cible cytochrome P450 3A.	**38**
Fig. III.02	Répartition des molécules étudiées Selon l'unité de mesure de l'IC50.	**39**
Fig. III.03	Classification des molécules selon le niveau de l'activité. pharmacologique	**40**
Fig. III.04	Variation des valeurs de pIC50 en fonction des molécules étudiées.	**41**
Fig. III.05	Distribution des valeurs de pIC50 pour l'ensemble des molécules.	**41**
Fig. III.06	Répartition des valeurs de pIC50 selon les classes de bio-activité.	**42**
Fig. III.07	Matrice de corrélation entre les descripteurs moléculaires (fingerprints)	**43**
Fig. III.08	Répartition de descripteurs après élimination des corrélations élevées.	**43**
Fig. III.09	Précision des modèles prédictifs testés.	**44**
Fig. III.10	Présentation de la performance des modèles étudiés.	**45**
Fig. III.11	Présentation Radar des performances normalisées.	**45**
Fig. III.12	Corrélation du modèle gradient Boostig.	**46**
Fig. III.13	Corrélation du modèle SVR.	**46**
Fig. III.14	Corrélation du modèle Random forest.	**47**

Liste des abréviations

Abréviation	Signification
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
QSAR	Quantitative Structure–Activity Relationship
SMILES	Simplified Molecular Input Line Entry System
LLMs	Large Language Models
XAI	Explainable Artificial Intelligence
FDA	Food and Drug Administration
EMA	European Medicines Agency
RDD	Rational Drug Design
SBDD	Structure-Based Drug Design
LBDD	Ligand-Based Drug Design
RMN	Nuclear Magnetic Resonance
IC50	Inhibitory Concentration 50%
pIC50	Negative log of IC50
CSV	Comma-Separated Values
API	Application Programming Interface
PaDEL	Pharmacological Descriptors
OLS	Ordinary Least Squares
KPI	Key Performance Indicator
GPU	Graphics Processing Unit
R^2	Coefficient of Determination
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
CHEMBL	ChEMBL Database
CYP3A4	Cytochrome P450 3A4
CYP3A	Cytochrome P450 3A

Liste des tableaux

N°	Titre Tableaux	Page
Tab. III. 01	Résumé des molécules sélectionnées et leurs unités de mesure	**39**

Introduction

Le *drug design* représente un pilier fondamental de la recherche pharmaceutique moderne, visant à concevoir des molécules thérapeutiques par une approche rationnelle intégrant biologie structurale, chimie médicinale et modélisation computationnelle. Historiquement, cette discipline a évolué des méthodes empiriques vers des stratégies ciblées, où l'identification de cibles moléculaires spécifiques et l'optimisation des propriétés pharmacocinétiques sont devenues centrales. Ce paradigme permet non seulement d'accélérer la découverte de principes actifs, mais aussi de réduire les échecs coûteux en phase clinique, notamment grâce à la prédiction des interactions médicament-cible et des profils de toxicité.(**Mak, Wong&Pichika, 2024 ; Mak &Pichika, 2019).**

L'intégration de l'IA et du machine learning (ML) révolutionne le drug design par plusieurs mécanismes clés à savoir la prédiction de cibles thérapeutiques ou les algorithmes analysent des données omiques (génomiques, transcriptomiques) pour identifier de nouvelles cibles biologiques pertinentes dans des pathologies complexes, et la conception *de novo* de molécules dans lesquelles des réseaux antagonistes génératifs et des modèles de deeplearning génèrent des structures chimiques innovantes optimisées pour l'affinité de liaison et la sécurité, explorant des espaces chimiques inaccessibles aux méthodes traditionnelles. Parallèlement, l'optimisation des essais précliniques ou le ML prédit la toxicité et l'efficacité des candidats-médicaments via l'analyse de bases de données massives, réduisant la dépendance aux criblages in vitro coûteux**(Patel & Shah, 2022 ; Jiménez-Luna, Grisoni, Weskamp& Schneider, 2021).**

Parmi les cibles thérapeutiques d'intérêt majeur dans le domaine du *drug design* figure la superfamille des cytochromes P450, et en particulier l'isoforme CYP3A4. Ces enzymes, localisées principalement dans le foie, jouent un rôle central dans le métabolisme de la majorité des médicaments actuellement sur le marché, en catalysant leur oxydation et leur élimination **(Guengerich, 2019).**Cette cible est responsable à lui seul du métabolisme de près de 50% des principes actifs, ce qui en fait une cible clé pour la prédiction des interactions médicamenteuses et la sécurité d'emploi des nouveaux composés. Les variations génétiques, l'induction ou l'inhibition de cette enzyme par d'autres substances, ainsi que la diversité des substrats qu'elle peut métaboliser, expliquent la complexité de son étude et l'importance de disposer d'outils prédictifs performants pour anticiper les effets indésirables et optimiser le profil pharmacocinétique des candidats-médicaments (**Zanger& Schwab, 2013).**

Dans cette perspective, ce projet fin d'études vise à développer un modèle prédictif de l'activité des molécules vis-à-vis du cytochrome P450, en particulier du CYP3A4, en s'appuyant sur les techniques les plus récentes d'intelligence artificielle et de machine learning. L'idée centrale est de collecter et de curer une base de données exhaustive comprenant des molécules de structures variées, dont l'activité métabolique vis-à-vis du

CYP3A4 a été caractérisée expérimentalement. À partir de cette base, des descripteurs moléculaires, à la fois 2D et 3D, seront extraits et utilisés pour entraîner différents algorithmes d'apprentissage automatique. L'objectif est de permettre au modèle d'apprendre à reconnaître les motifs structuraux et les propriétés physico-chimiques qui déterminent l'affinité ou l'inhibition de la cible. Une étape essentielle consistera à valider la robustesse du modèle par des méthodes croisées et à comparer ses prédictions à des résultats expérimentaux ou à des simulations. Ce modèle prédictif pourra ensuite être utilisé pour cribler virtuellement de nouvelles bibliothèques de composés, anticiper les risques d'interactions médicamenteuses et guider la conception de candidats-médicaments présentant un profil métabolique optimisé. À terme, ce projet s'inscrit dans la dynamique actuelle de la recherche pharmaceutique, où l'intégration de l'IA et de la biologie structurale permet d'accélérer et de fiabiliser la découverte de nouveaux traitements, tout en réduisant les coûts et les délais de développement.

Chapitre 1

Machine Learning

1. Introduction

L'intelligence artificielle, ou IA, est la simulation de l'intelligence humaine à travers des machines et des programmes conçus pour penser et agir de manière similaire aux humains. Ces systèmes sont programmés pour apprendre et s'adapter aux nouvelles informations, ainsi que pour résoudre des problèmes et prendre des décisions en fonction des données disponibles. L'intelligence artificielle a le potentiel de révolutionner de nombreux secteurs et comprend deux domaines principaux : l'apprentissage automatique (Machine Learning) et l'apprentissage profond (Deep Learning)(**Fig. I. 01**).

L'apprentissage automatique ou ce qu'on appelle le machine learning est un sous-domaine de l'intelligence artificielle qui vise à permettre aux ordinateurs d'apprendre à partir de données, sans être explicitement programmés. L'objectif de l'apprentissage automatique est de créer des algorithmes capables de recevoir des données d'entrée et de les utiliser pour faire des prédictions ou prendre des mesures afin d'atteindre un objectif spécifique. **(EL MASSARI, 2023).**

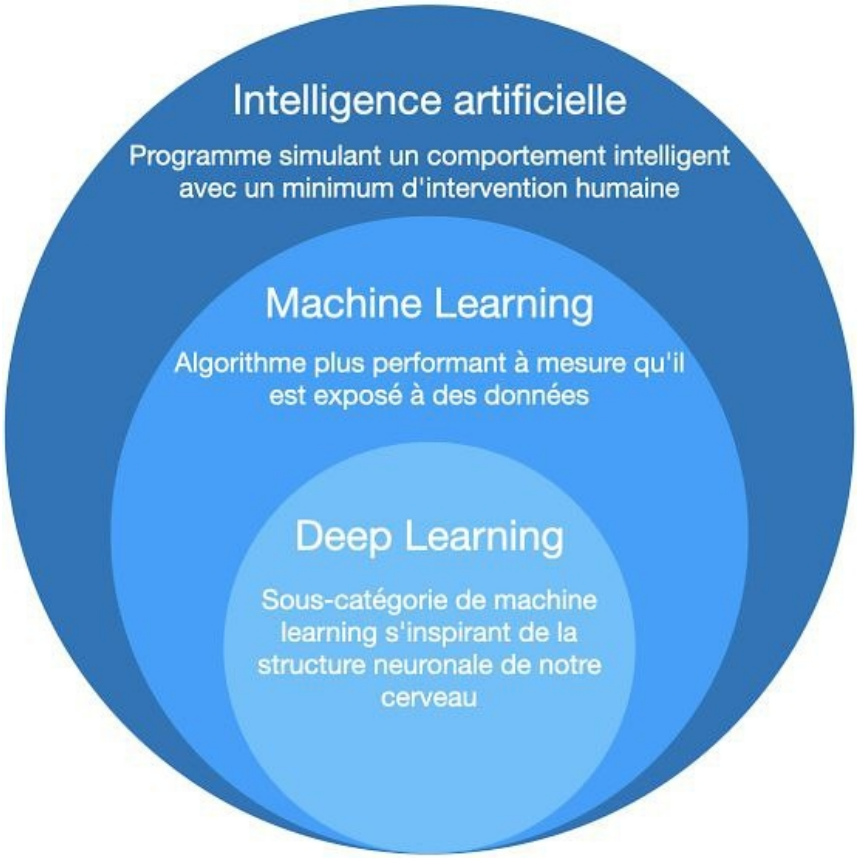

Fig. II.01 : Structuration des domaines de l'intelligence artificielle **(Pawlak, 2022)**

2. Définition d'apprentissage automatique

En 1959, Arthur Samuel, un pionnier du domaine de l'apprentissage automatique, a fait l'énoncé suivant « L'apprentissage automatique est le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés »**(Géron, 2020)**.

L'apprentissage automatique (AA), ou Machine Learning (ML), est une branche de l'intelligence artificielle qui permet aux machines d'apprendre à effectuer des tâches sans être explicitement programmées. Il s'agit d'un ensemble d'outils statistiques, géométriques et d'algorithmes informatiques qui automatisent la construction d'une fonction de prédiction à partir d'un ensemble d'observations appelé ensemble d'apprentissage. Le terme machine learning désigne les techniques utilisées pour déduire des modèles à partir de grandes quantités de données, ainsi que la capacité à faire des prédictions basées sur l'analyse des données disponibles. L'objectif principal de l'apprentissage automatique est de permettre aux systèmes informatiques d'acquérir des capacités d'apprentissage autonomes, facilitant ainsi leur indépendance vis-à-vis de la programmation explicite**(Olivier, 2023).**

3. Les types de système de Machine Learning

Il existe plusieurs types de système d'apprentissage automatique et cela varie en fonction du type de problème que l'on se pose. Il est alors utile de les classer en différentes catégories. Les systèmes de machine learning peuvent-être classés en fonction de l'importance et de la nature de la supervision qu'ils requièrent durant la phase d'entraînement. Le système d'apprentissage automatique peut être classé en quatre grandes catégories **(Olivier, 2023)** :

- Apprentissage supervisé.
- Apprentissage non supervisé.
- Apprentissage semi supervisé.
- Apprentissage par renforcement.

3.1.Apprentissage supervisé

Les algorithmes ou méthodes d'apprentissage supervisé sont les algorithmes ML les plus couramment utilisés **(Olivier, 2023)**. L'apprentissage supervisé est une branche de l'apprentissage automatique où une machine apprend à accomplir des tâches en s'appuyant sur un ensemble de données d'entraînement composé d'exemples déjà étiquetés. Chaque exemple dans cet ensemble est constitué d'une paire (entrée, sortie) **(CHEFROUR, & SOUICI-MESLATI, 2013)**. L'objectif principal de l'apprentissage supervisé est d'apprendre une association entre les échantillons de données d'entrée et les sorties correspondantes après avoir effectué plusieurs instances de donnée d'entraînement **(Olivier, 2023).** Par exemple, nous avons x comme variable d'entrée et y la variable de sortie. L'objectif d'un algorithme d'apprentissage supervisé est de trouver une fonction f de mise en correspondance de la variable d'entrée (x) avec la variable de sortie (Y), c'est-à-dire une expression du type

Y=f(x).Ce qui permettra d'obtenir de nouvelles données d'entrée (x), nous pouvons facilement prédire la variable de sortie (Y) pour ces nouvelles données d'entrée **(Juvénal, 2024).**

Ce type d'apprentissage inclut deux principales méthodes, la classification, où l'on connaît les entrées et l'objectif est de prédire les sorties, et la régression, où l'on connaît les sorties et cherche à estimer les entrées correspondantes **(CHEFROUR, & SOUICI-MESLATI, 2013)**.

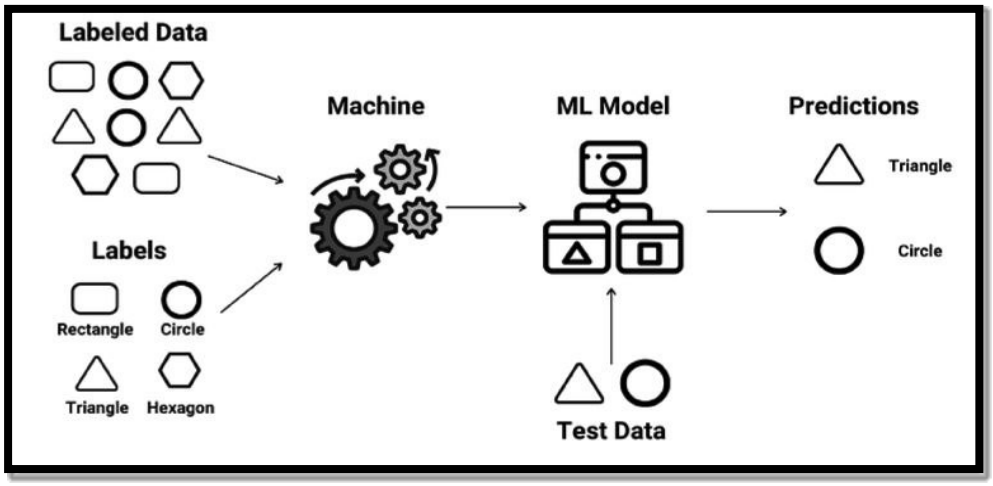

Fig. I.02 : Apprentissage supervisé **(Bright, 2024)**

3.2.Apprentissage non supervisé

L'apprentissage non supervisé est un type d'apprentissage automatique dans lequel les modèles sont entraînés à l'aide de jeux de données non étiquetés, permettant aux modèles de traiter ces données sans supervision directe. Ce type d'apprentissage vise à extraire de nouvelles informations à partir du jeu de données d'entraînement, dont on ne dispose d'aucune connaissance préalable. Dans ce contexte, la catégorie à laquelle appartiennent les exemples d'entraînement n'est pas connue. L'objectif principal de l'apprentissage non supervisé est de diviser les données d'entraînement en groupes caractérisés par une similarité des propriétés au sein de chaque groupe et une différence entre les différents groupes**(CHEFROUR, & SOUICI-MESLATI, 2013).**

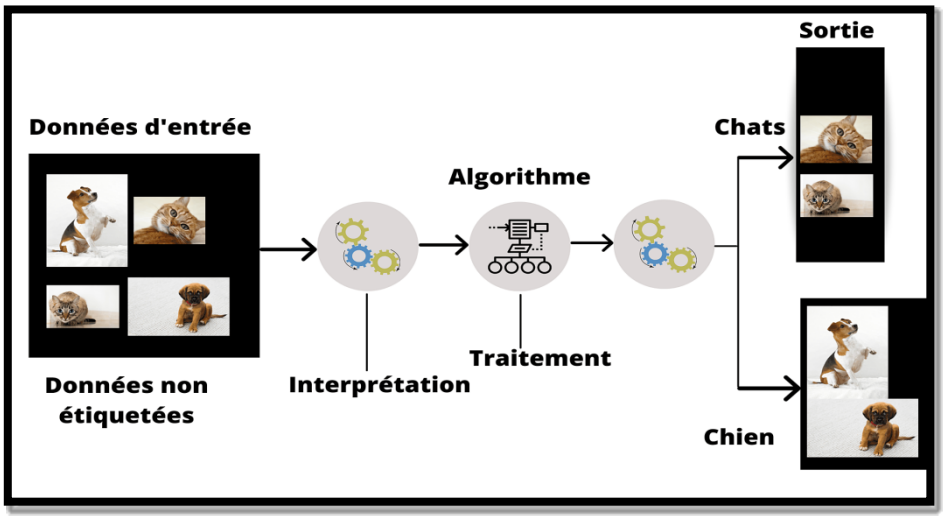

Fig. I.03 : Apprentissage non supervisé **(Juvénal, 2024).**

3.3.Apprentissage semi supervisé

L'apprentissage semi-supervisé est une combinaison des méthodes d'apprentissage supervisé et non supervisé, où l'ensemble de données utilisé comprend à la fois des échantillons étiquetés et non étiquetés. Ce type d'apprentissage vise à traiter des problèmes pour lesquels la quantité de données étiquetées est limitée, tandis que celle des données non étiquetées est importante. L'un des principaux avantages de l'apprentissage semi-supervisé est la réduction du temps et des efforts nécessaires pour annoter les données, comparé à l'apprentissage supervisé complet. Des études ont démontré que la combinaison des données non étiquetées avec les données étiquetées améliore significativement la qualité du modèle et sa capacité de généralisation. L'objectif principal de ce type d'apprentissage est de classer ou d'interpréter les données non étiquetées en s'appuyant sur les informations contenues dans les données étiquetées. L'apprentissage semi-supervisé est une option idéale dans les cas où il est difficile d'obtenir suffisamment de données étiquetées. Les graphiques suivants illustrent de manière détaillée le concept de l'apprentissage semi-supervisé **(Géron, 2020).**

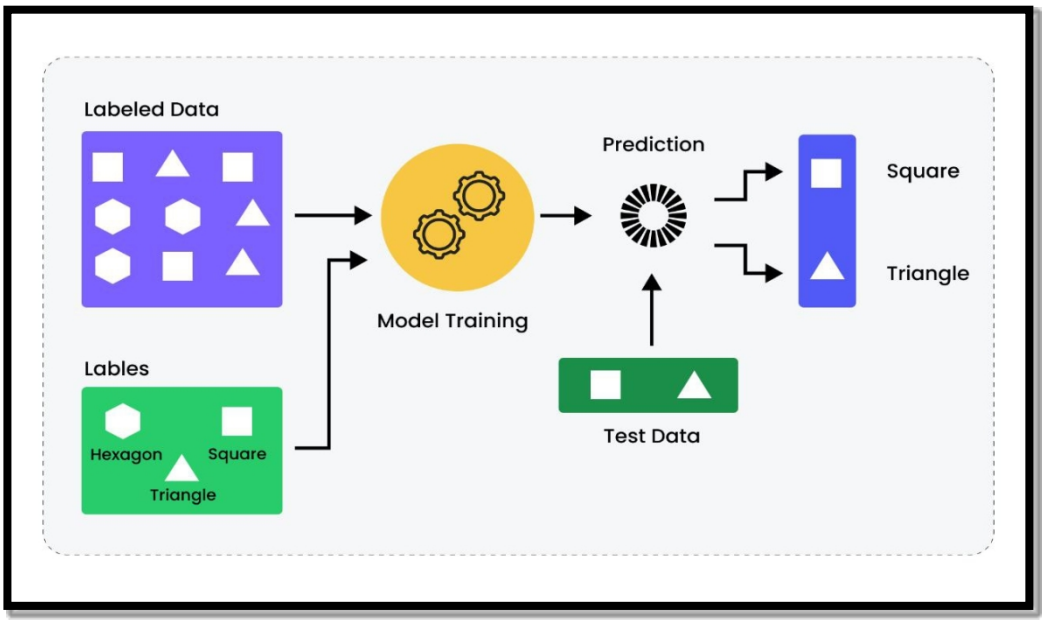

Fig. I.04 : Apprentissage semi supervisé **(Géron, 2020).**

3.4.Apprentissage par renforcement

L'apprentissage par renforcement est un type unique et totalement différent de systèmes d'apprentissage. Dans ce contexte, le système d'apprentissage est appelé « agent », qui a la capacité d'observer son environnement, de choisir et d'exécuter des actions, et de recevoir des récompenses ou des punitions (sous forme de récompenses négatives). L'agent doit apprendre par lui-même la meilleure stratégie, appelée « politique », dans le but d'obtenir le maximum de récompenses possibles sur le long terme. La politique définit les actions que l'agent doit choisir dans chaque situation qu'il rencontre.Par exemple, de nombreux robots utilisent des algorithmes d'apprentissage par renforcement pour apprendre à marcher. Le programme AlphaGo de DeepMind est également un bon exemple d'apprentissage par renforcement **(Géron, 2020).**

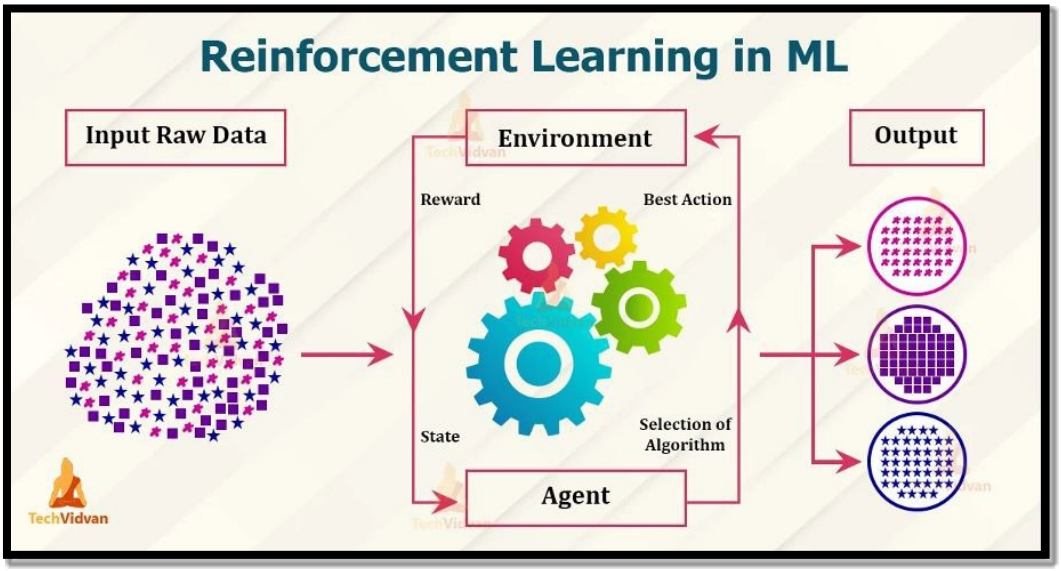

Fig. I.05 : Apprentissage par renforcement (Bright, 2024)

4. Les étapes de l'apprentissage automatique

4.1.Collecte et préparation des données

La première étape du processus d'apprentissage automatique est la collecte des données, car les données sont l'élément vital et fondamental dans ce domaine. Les données constituent la pierre angulaire de l'apprentissage automatique, car la qualité et la quantité des données influencent directement la capacité du modèle à apprendre et à performer. Sans données précises et appropriées, il devient difficile d'obtenir des résultats fiables. **(Stroppa, 2005)**

Il est donc essentiel de fournir aux machines des données de qualité pour commencer. Ces données doivent être appropriées, Comporter un minimum de valeurs manquantes ou répétées,et couvrir un large éventail des différentes sous-catégories ou classes.

La phase suivante, appelée Phase de prétraitement des données, consiste à manipuler soigneusement les informations recueillies afin d'assurer l'efficacité maximale du processus d'apprentissage. Cela peut se faire par :

- rassembler toutes les données disponibles et les disperser de manière aléatoire afin d'assurer une répartition homogène et éviter que l'ordre n'entrave le processus d'apprentissage,
- supprimer les données indésirables, les valeurs vides ou répétées, convertir les types de données si nécessaire, ainsi que réorganiser l'ensemble de données en modifiant les lignes et les colonnes si besoin,
- visualiser les données pour comprendre leur structure et les relations entre les différentes variables et classes,
- diviser les données épurées en deux ensembles : un ensemble d'entraînement à partir duquel le modèle apprend, et un ensemble de test pour vérifier la précision et la performance du modèle (**Badillo, 2020).**

4.2.Choix de l'algorithme et du modèle

Nous avons vu précédemment qu'il existe de nombreux algorithmes d'apprentissage automatique. La question est donc maintenant, lequel utiliser ? Le premier critère de choix concerne le type d'apprentissage souhaité ou disponible. En fonction de cela, on peut choisir l'apprentissage supervisé, non supervisé, par renforcement ou autre, selon ce qui est possible pour la tâche requise et les données disponibles.

La théorie du « No Free Lunch » indique qu'il n'existe pas d'algorithme qui surpasse tous les autres dans tous les problèmes. En d'autres termes, si un algorithme est performant pour une tâche donnée, il peut être moins efficace qu'un autre algorithme pour une autre tâche. Par conséquent, il n'est pas possible d'utiliser le même algorithme dans tous les cas.

Le choix du type d'apprentissage et du modèle dépend des hypothèses que nous faisons sur le problème. Ces hypothèses peuvent ne pas s'appliquer à toutes les tâches, ce qui influence la précision et l'efficacité du modèle, que ce soit positivement ou négativement. Pour cette raison, le concepteur doit avoir del'expérience pour analyser la tâche, les données et les contraintes techniques (comme la mémoire et la puissance de calcul), ou bien recourir à des techniques modernes telles queAutoML pour formuler des hypothèses logiques facilitant le choix de l'algorithme approprié **(Matteis, 2022).**

4.3.Entraînement du modèle

Lors de la phase d'entraînement, un modèle utilise un ensemble de données d'entraînement afin d'apprendre et de détecter les motifs et les relations présents dans les données. Cela se fait à l'aide d'algorithmes de machine learning adaptés à la nature de la tâche (classification ou régression), permettant ainsi au modèle d'améliorer progressivement ses prédictions ou décisions futures **(Masari Hakim, 2023).**

Il est également important de diviser les données au départ en différentes parties, notamment les données d'entraînement et les données de test. Les données d'entraînement servent à former le modèle, tandis que les données de test sont utilisées pour évaluer ses performances après la fin de l'entraînement. Il est essentiel de ne pas utiliser les données de test pendant la phase d'apprentissage, afin de garantir l'objectivité et la fiabilité de l'évaluation finale **(Aiboud & Laskri, 2020).**

4.4.Évaluation et validation

Après avoir terminé l'entraînement du modèle, il est essentiel d'évaluer son efficacité avant d'envisager son utilisation pratique.Cette étape consiste à mesurer les performances du modèle à l'aide de données qu'il n'a jamais vues auparavant, afin de vérifier sa capacité à généraliser et à faire des prédictions précises dans de nouvelles situations.

Plusieurs indicateurs sont utilisés pour cette évaluation, tels que :

- La précision (Accuracy) dans les cas de classification.
- La matrice de confusion (Confusion Matrix), qui donne une vue détaillée des prédictions correctes et incorrectes,

Ainsi que les indicateurs clés de performance (KPI), définis selon les objectifs du projet.

L'évaluation est réalisée sur un ensemble de validation ou de test, afin de s'assurer que les performances du modèle ne se limitent pas aux données d'entraînement. Le modèle est également comparé à un modèle de référence simple (Baseline) pour déterminer s'il apporte réellement une valeur ajoutée ou s'il serait préférable d'envisager une autre approche (**Badillo, 2020).** Il convient également d'analyser ce qui est efficace dans le modèle, ce qui doit être amélioré et ce qui est encore en cours de développement **(Çelik, 2018).**

4.5.Optimisation

Après avoir évalué le modèle, il peut être nécessaire d'ajuster les hyperparamètres afin d'améliorer ses performances. Ce processus est connu sous le nom de réglage des hyperparamètres ou optimisation des hyperparamètres.

Cette étape est cruciale dans le processus d'apprentissage automatique, car elle vise à trouver la meilleure combinaison de valeurs qui permet au modèle d'atteindre des performances optimales sur les données. Les hyperparamètres ne sont pas appris automatiquement pendant l'entraînement, mais sont définis à l'avance, comme :

- le taux d'apprentissage (learning rate).
- le nombre d'arbres dans une forêt aléatoire (randomforest), ou encore le nombre de couches et de neurones dans un réseau de neurones.

Parmi les techniques les plus courantes de réglage des hyperparamètres, on trouve :

- la recherche en grille (GridSearch).
- la validation croisée (Cross-Validation / Validation Croisée).

Pour résoudre les problèmes de surajustement (overfitting) et de sous-ajustement (underfitting), il est essentiel de mettre en place des techniques appropriées afin de garantir que le modèle fonctionne correctement avec de nouvelles données **(EL MASSARI, 2023)**.

Il existe également des méthodes plus avancées, comme la recherche aléatoire (Random Search),et l'optimisation bayésienne (Bayesian Optimization), qui peuvent être plus efficaces lorsqu'on traite un grand nombre d'hyperparamètres

En résumé, l'ajustement des hyperparamètres est une étape essentielle pour améliorer la précision et la fiabilité du modèle avant d'envisager son déploiement ou son intégration dans un environnement de production. Il doit être traité avec sérieux dans les différentes phases de développement de tout système d'apprentissage automatique intégré **(Matteis, 2022).**

4.6.Déploiement et surveillance du modèle

Une fois le modèle optimisé, il est déployé dans un environnement de production, que ce soit sur le cloud, sur des serveurs locaux, ou sous forme de service via une interface de programmation d'applications (API). Le modèle est alors intégré au système existant pour exécuter des tâches spécifiques, telles que la reconnaissance automatique de produits à l'aide de la vision par ordinateur.

Lors du déploiement, il est essentiel de définir l'infrastructure matérielle appropriée (comme la mémoire vive, la capacité de stockage et la puissance de calcul) afin d'assurer l'efficacité de l'inférence et la rapidité des performances. L'expérience utilisateur est ensuite évaluée à l'aide de tests A/B, et le processus de déploiement doit être fluide et soigneusement planifié afin de garantir l'acceptation des modifications.

Une fois le modèle mis en production, une phase de surveillance continue est mise en place. Celle-ci repose sur des mécanismes automatisés permettant de détecter toute anomalie ou dégradation des performances, avec l'envoi de notifications aux équipes concernées. En cas de baisse de performance, il peut être nécessaire de réentraîner le modèle sur de nouvelles données ou de modifier son architecture. Dans certains cas exceptionnels, il peut être requis de repenser l'ensemble du cycle de vie du système, en mettant à jour les données, les modèles, ainsi que l'infrastructure technique **(Suresh & Guttag, 2021).**

5. Algorithmes d'apprentissage automatique

5.1.Régression linéaire (Linear Regression)

La régression linéaire est un algorithme statistique utilisé pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Elle repose sur la minimisation de la somme des carrés des écarts entre les valeurs observées et les valeurs prédites (méthode des moindres carrés ordinaires, OLS). Ses avantages incluent la simplicité et la facilité d'interprétation, et elle est largement utilisée pour la prédiction et l'analyse de données continues **(Chikhi & Zitouni, 2024)**.

5.2.Régression logistique (Logistic Regression)

Algorithme de classification utilisé pour estimer la probabilité d'un événement binaire (par exemple oui/non) à l'aide d'une fonction logistique (sigmoïde). Contrairement à la régression linéaire, elle ne produit pas des valeurs continues, mais des probabilités comprises entre 0 et 1 **(Belghiti, 2021).**

5.3.Arbres de décision (Decision Trees)

Ils représentent un modèle de prise de décision sous forme d'un flux où les données sont progressivement divisées en nœuds contenant certaines caractéristiques jusqu'à atteindre une décision finale. Faciles à interpréter, ils sont cependant sujets au surapprentissage (overfitting) s'ils ne sont pas correctement régularisés **(Benmaamar, 2023)**.

5.4.Forêts aléatoires (Random Forest)

Basées sur l'agrégation d'un grand nombre d'arbres de décision, chaque arbre étant construit sur un échantillon différent des données. Des techniques de vote ou de moyenne sont utilisées pour la prédiction finale. Cela améliore la précision du modèle et réduit le surapprentissage **(Boukertouta, 2022)**.

5.5.Machine à vecteurs de support (SVM)

Cet algorithme vise à trouver un hyperplan optimal séparant les différentes classes dans l'espace des caractéristiques. Il est très efficace pour la classification en haute dimension et peut gérer des cas non linéaires grâce à l'utilisation de noyaux (kernels) **(Chikhi & Zitouni, 2024)**.

5.6.K plus prochesvoisins (K-Nearest Neighbors, KNN)

Cet algorithme se base sur le calcul de la distance entre les points et classe un nouvel échantillon selon la majorité des classes des k points les plus proches. Simple, il est cependant sensible au bruit et coûteux en calcul lorsque les données sont volumineuses **(Belghiti, 2021).**

5.7.Naïve Bayes

Algorithme de classification basé sur le théorème de Bayes et supposant l'indépendance des caractéristiques. Malgré sa simplicité, il est très efficace pour la classification de textes et les applications avec des données à haute dimension **(Boucher, 2023)**.

5.8.K-means

Algorithme de clustering non supervisé visant à partitionner les données en k groupes de sorte que la distance intra-groupe soit minimisée. Il repose sur la sélection initiale des centres et leur amélioration par itérations successives **(Benmaamar, 2023)**.

5.9.Gradient Boosting

Méthode d'agrégation construisant des modèles de manière séquentielle, chaque nouveau modèle cherchant à corriger les erreurs des modèles précédents. Parmi ses applications les plus célèbres figurent XGBoost et LightGBM, largement utilisés dans les compétitions de science des données **(Pawlak, 2022).**

5.10. Réseaux de neurones (Neural Networks, MLP &Deep)

Modèle inspiré du fonctionnement du cerveau humain, constitué de couches de nœuds (neurones). Il inclut à la fois les perceptrons multicouches (MLP) simples ainsi que les réseaux profonds tels que les réseaux convolutifs (CNN) et récurrents (RNN). Ces architectures se révèlent particulièrement efficaces pour le traitement d'images, l'analyse de textes et les prédictions complexes **(Boucher, 2023) et (Pawlak, 2022).**

Chapitre 2

Le Drug Design

1. Introduction

La découverte et la conception de médicaments sont des processus complexes et coûteux qui prennent de nombreuses années. En moyenne, le développement d'un nouveau médicament nécessite plus de dix ans et un coût pouvant atteindre des chiffres énormes. Ce processus se caractérise par des taux de réussite faibles, avec moins de 10 % des médicaments atteignant le marché après les essais cliniques de phase I. Ces dernières années,les techniques d'intelligence artificielle ont provoqué une transformation radicale dans le domaine de la découverte de médicaments, grâce à des applications telles que le criblage virtuel et la conception de médicaments de novo, ce qui a permis d'accélérer la recherche et d'améliorer la précision des résultats. La conception de médicaments repose principalement sur la compréhension de la cible thérapeutique, qu'il s'agisse d'une enzyme ou d'un récepteur, et sur le développement de petites molécules capables d'activer ou d'inhiber cette cible pour obtenir l'effet thérapeutique souhaité. Pour augmenter l'efficacité de la découverte, des techniques de criblage à haut débit utilisent de vastes bibliothèques chimiques et l'automatisation avancée, tandis que le criblage virtuel s'appuie sur l'informatique pour identifier les composés actifs avant les tests en laboratoire. Les médicaments sont également classés selon leur mécanisme d'action en agonistes et antagonistes,et leur efficacité est évaluée par la mesure de leur affinité et de leur effet biologique. Avec les progrès continus de l'intelligence artificielle, il est désormais possible de concevoir des médicaments plus précis et efficacesouvrant ainsi de nouvelles perspectives pour traiter les maladies de manière personnalisée et efficace.

2. Définition de la conception des médicaments (Drug Design)

La conception de médicaments est une branche de la chimie médicinale qui vise à développer de petites molécules capables d'interagir de manière sélective et efficace avec des cibles biologiques spécifiques, dans le but de modifier ou d'inhiber une fonction biologique liée à une maladie **(Benamar, 2021).**

Ce processus repose sur une compréhension approfondie de la structure de la cible moléculaire (comme les protéines ou les enzymes) et de son mécanisme d'action, et s'appuie sur les principes de la biochimie et de la biologie moléculaire afin d'améliorer les propriétés du composé en termes d'efficacité, d'absorption, de distribution, de métabolisme et de toxicité (ADMET) **(Xue et al, 2021).**

3. Le processus de découvert de médicament (Drug discovery process)

Le processus de découverte de médicament désigne une série d'étapes scientifiques interconnectées visant à développer un composé pharmaceutique efficace et sûr pour traiter une pathologie spécifique. Cette démarche implique une synergie de disciplines variées telles que la biologie, la chimie, les mégadonnées (big data) et l'intelligence artificielle. Ce qui suit est une présentation rigoureuse et détaillée des principales phases de ce processus moderne.

3.1. Identification de la cible biologique

Le processus commence par l'identification d'une cible biologique clairement définie généralement une protéine, une enzyme ou un récepteur cellulaire associée au mécanisme de la maladie. Cette cible est déterminée à travers l'étude des voies de signalisation cellulaire et l'analyse des interactions moléculaires impliquées dans le déséquilibre pathologique. L'intelligence artificielle est mobilisée pour analyser de vastes bases de données génomiques, protéomiques et transcriptomiques, afin d'identifier les corrélations entre cibles et pathologies. Les modèles explicables (XAI) aident à justifier le choix de la cible, renforçant ainsi la validité scientifique de cette étape critique **(Alizadehsani et al, 2023)**.

3.2. Criblage des composés

Cette phase consiste à tester des milliers de composés chimiques ou biologiques contre la cible identifiée, dans le but de repérer ceux présentant une activité initiale prometteuse. Grâce à l'apprentissage automatique, il est désormais possible de prédire la capacité d'un composé à interagir avec la cible avant même les essais expérimentaux, économisant ainsi temps et ressources **(Ma et al, 2024)**.

Par ailleurs, les grands modèles de langage (LLMs) permettent de générer de nouveaux composés structurés spécifiquement pour s'adapter à la cible moléculaire via des algorithmes d'IA, marquant ainsi une avancée significative dans la conception de médicaments «de novo » **(Wang et al, 2024)**.

3.3. Optimisation des pistes

Une fois les composés prometteurs identifiés, une phase d'optimisation est engagée pour améliorer leurs propriétés physico-chimiques et pharmacocinétiques. Cela inclut la modification de la structure moléculaire afin d'accroître leur efficacité, de réduire leur toxicité et d'augmenter leur stabilité biologique. L'IA joue ici un rôle crucial pour prédire les effets des modifications chimiques, réduisant le besoin de tests expérimentaux extensifs **(Liu et al, 2024)**.

Les LLMs comme BioGen offrent en outre la possibilité de modifier la structure des composés de manière interactive et rapide, selon les critères du médicament idéal, facilitant l'atteinte d'un candidat cliniquement viable **(Wang et al, 2025)**.

3.4. Études précliniques

L'objectif de cette phase est d'évaluer l'efficacité et la sécurité du composé sur des modèles non humains, comme des animaux ou des cultures cellulaires. Elle inclut l'analyse de l'absorption, la distribution, le métabolisme et l'excrétion (ADME), ainsi que l'évaluation de la toxicité aiguë et chronique. Les modèles informatiques sont de plus en plus utilisés pour simuler le comportement pharmacologique, réduisant ainsi la dépendance aux essais animaux **(Benamar, 2021)**.

L'IA permet également d'identifier précocement des signaux de toxicité, ce qui renforce l'efficacité du filtrage préclinique et diminue les échecs en phases ultérieures **(Alizadehsani et al, 2023)**.

3.5. Essais cliniques

Les essais cliniques représentent la phase la plus critique et la plus coûteuse du processus. Ils visent à tester le composé sur des sujets humains à travers trois phases successives :

- Phase I

Évalue la sécurité, la toxicité et la posologie optimale du médicament. Réalisée sur un petit groupe de volontaires sains, elle permet d'observer l'absorption, la distribution et les premiers effets secondaires **(Singh et al, 2022).**

- Phase II

Teste l'efficacité thérapeutique initiale sur des patients et continue de surveiller la tolérance. L'IA y est utilisée pour analyser les réponses individuelles et les relier à des biomarqueurs, facilitant une médecine plus personnalisée **(Liu et al, 2024).**

- Phase III

Impliquant un large échantillon de patients sur plusieurs centres, elle confirme l'efficacité et élargit l'évaluation de la sécurité. Des comparaisons sont établies avec les traitements existants ou des placebos. Le traitement des big data y est essentiel pour une analyse rapide et précise **(Liu et al, 2024).**

3.6. Approbation réglementaire et mise sur le marché

Après le succès des essais cliniques, un dossier complet est soumis aux autorités réglementaires (FDA, EMA…) pour approbation. Ce dossier contient les résultats des études précliniques et cliniques, ainsi que les plans de fabrication, d'emballage et de pharmacovigilance. L'intelligence artificielle permet d'organiser ces données complexes de manière cohérente, accélérant le processus décisionnel **(Alizadehsani et al, 2023).**

Les systèmes AI contribuent également à évaluer les bénéfices et risques, augmentant ainsi les chances d'acceptation du médicament dès la première soumission **(Xue et al, 2021).**

3.7. Repositionnement des médicaments

Le repositionnement consiste à réutiliser des médicaments déjà approuvés pour traiter de nouvelles maladies, sur la base de similarités mécanistiques. Cette approche est économique et rapide, car la sécurité du composé est déjà établie. L'IA est utilisée pour explorer les bases de données cliniques et pharmaceutiques, révélant des liens inédits entre médicaments existants et pathologies émergentes **(Xue et al, 2021).**

De plus, les modèles génératifs permettent d'identifier de nouvelles corrélations entre structures moléculaires et effets thérapeutiques, offrant ainsi une seconde vie à des médicaments abandonnés**. (Ma et al, 2024).**

4. Cibles thérapeutiques en drug design

La cible thérapeutique est considérée comme la pierre angulaire de la conception et du développement des médicaments modernes, car elle représente une molécule biologique spécifique (comme une protéine, une enzyme, un récepteur ou un acide nucléique) dont l'intervention médicamenteuse est supposée pouvoir modifier un certain parcours pathologique. La compréhension de la structure et de la fonction précise de cette molécule au sein des réseaux biologiques complexes permet aux chercheurs de concevoir des molécules médicamenteuses capables d'interagir avec elle de manière sélective, réduisant ainsi les interactions indésirables avec d'autres molécules et renforçant l'efficacité thérapeutique. Cette approche de ciblage précis a révolutionné le traitement des maladies, du cancer aux maladies auto-immunes, en réalisant des "médicaments intelligents" qui ciblent les mécanismes moléculaires fondamentaux de la maladie au lieu de se limiter à traiter les symptômes **(Alizadehsani et al, 2022).**

Dans le cadre de cette étude, nous avons choisi le Cytochrome P450 3A (CHEMBL340) comme cible thérapeutique pour appliquer la méthodologie de conception de médicaments. Le Cytochrome P450 3A (CYP3A) est l'un des enzymes les plus importants de la famille des cytochromes P450, responsable du métabolisme d'une large gamme de composés médicamenteux et chimiques dans le foie et les intestins. Cet enzyme comprend plusieurs isoformes humaines telles que CYP3A4 et CYP3A5, le CYP3A4 étant spécifiquement l'enzyme le plus abondant dans le foie humain. Sa fonction principale est de faciliter les réactions d'oxydation biochimique qui contribuent à transformer les médicaments en formes facilement éliminables.

Le CYP3A est classé parmi les cibles de découverte de médicaments en raison de son rôle crucial dans la détermination de la biodisponibilité, de la toxicité et des interactions médicamenteuses potentielles. Il est référencé dans la base de données ChEMBL sous l'identifiant CHEMBL340 en tant que cible protéique centrale utilisée pour tester l'efficacité et la sécurité des nouveaux composés médicamenteux **(Zanger & Schwab, 2013).**

5. Les approches de drug design

Les approches de drug design (conception de médicaments) sont diverses et reposent principalement sur des méthodes rationnelles et computationnelles visant à identifier ou créer des molécules thérapeutiques efficaces. Voici un résumé des principales approches :

5.1. Drug Design rationnel (RDD)

Des méthodes de conception rationnelle de médicaments ont été utilisées pour explorer et inventer de nouvelles molécules contre les maladies ou tout dysfonctionnement dans le corps humain, en se concentrant principalement sur la découverte des cibles pour les molécules actives et les molécules leaders, ainsi que sur l'optimisation des candidats ayant des propriétés pharmacologiques, en intégrant des données complètes sur les propriétés biochimiques et structurales de la cible protéique. Des techniques telles que la résonance magnétique nucléaire (RMN) et la cristallographie aux rayons X aident à étudier les propriétés structurelles des cibles protéiques, ce qui est très utile dans la découverte des cibles **(Mahapatra & Karuppasamy, 2022).**

5.2. Approches basées sur la structure (Structure-Based Drug Design, SBDD)

La conception de médicaments basée sur la structure tridimensionnelle (SBDD) repose sur l'analyse de la structure biologique tridimensionnelle de la cible à étudier, en utilisant des données dérivées de méthodes computationnelles telles que la modélisation par homologie (homology modeling) ou des expériences en laboratoire, qui peuvent inclure des protéines, des récepteurs ou des enzymes. Cette méthode étudie en détail la capacité des molécules de liaison sélectionnées à se lier au site cible à l'intérieur de la cible, et prédit également les sites clés ou les cavités auxquels ces molécules peuvent se lier. Ensuite, la force d'affinité de ces liaisons avec leur cible biologique moléculaire respective est évaluée. Les informations extraites sur la cible et les molécules de liaison sont d'une grande importance dans la conception de molécules efficaces avec des propriétés pharmacologiques et toxicologiques calculées et appropriées **(Mahapatra & Karuppasamy, 2022).**

5.3. Approches basées sur le ligand (Ligand-Based Drug Design, LBDD)

La conception de médicaments basée sur le ligand (LBDD) est principalement utilisée dans la découverte de médicaments lorsque la structure tridimensionnelle de la cible n'est pas disponible. Elle s'appuie sur des techniques avancées telles que la modélisation pharmacophore et la relation quantitative structure-activité en 3D (3D QSAR) pour développer des modèles prédictifs aidant à optimiser les molécules leaders et à comprendre les interactions avec les cibles médicamenteuses. Le LBDD est une méthode indirecte qui contribue à accélérer le développement de médicaments efficaces en étudiant les interactions potentielles entre les candidats médicamenteux et la cible. Cette approche est également liée à la conception assistée par ordinateur (CADD), qui vise à réduire le temps nécessaire à la découverte, à la caractérisation et à l'optimisation structurelle de nouvelles molécules médicamenteuses. Elle peut aussi être appliquée à l'amélioration des promédicaments (prodrugs) pour augmenter leur biodisponibilité et leur spécificité thérapeutique **(Mahapatra & Karuppasamy, 2022).**

5.4. Drug Design de novo

La conception de médicaments de novo est une méthode visant à créer de nouvelles molécules chimiques en se basant uniquement sur les informations disponibles concernant la cible biologique (comme le récepteur) ou les molécules actives connues qui s'y lient et possèdent une bonne affinité ou activité inhibitrice. Les étapes principales de cette méthode incluent la description du site de liaison actif du récepteur ou la modélisation pharmacophore des ligands associés, suivies de la construction des nouvelles molécules (échantillonnage) puis de l'évaluation des molécules générées **(Mouchlis et al, 2021).**

6. Méthodes de représentation moléculaire pour l'apprentissage automatique

La représentation des molécules et la prédiction des propriétés moléculaires constituent une étape cruciale dans le processus de découverte et de conception de médicaments. Elles peuvent être utilisées dès les premières phases de la découverte de médicaments pour identifier les molécules actives aux propriétés optimales et éliminer celles qui sont inadaptées. La qualité de la représentation influence directement la précision et l'efficacité des modèles dans la prédiction des propriétés des molécules chimiques et de leurs interactions biologiques, ce qui accélère la découverte de nouveaux médicaments et réduit les coûts associés aux expérimentations en laboratoire **(Wang, Jiang, Wang, & Xuan, 2024).**

Les principales méthodes de représentation moléculaire utilisées en apprentissage automatique et modèles prédictifs sont les suivantes :

- **Représentation par chaînes textuelles (Sequence-based) :** Les méthodes de représentation basées sur la séquence apprennent efficacement la représentation moléculaire. Elles visent à capturer des informations globales plus larges et fournissent un encodage unique pour chaque atome ou liaison, mais leur capacité de représentation reste toutefois fortement limitée. Parmi ces méthodes, on trouve notamment Smiles-Bert et Smiles **(Wang, Jiang, Wang, & Xuan, 2024).**

- **Représentation graphique (Graph-based) :** Une molécule peut être représentée sous forme de graphe où les nœuds correspondent aux atomes et les arêtes aux liaisons, reflétant ainsi la topologie moléculaire comme la connectivité des atomes, le nombre et la taille des cycles. Les méthodes basées sur les graphes, telles que les réseaux de neurones graphiques, exploitent cette topologie pour agréger les informations et améliorer la représentation moléculaire, ce qui a prouvé son efficacité dans des tâches comme la génération de molécules et la prédiction des propriétés **(Wang, Jiang, Wang, & Xuan, 2024).**

- **Représentation géométrique (Geometry-based) :** Ces méthodes se concentrent davantage sur les informations moléculaires au niveau géométrique. Autrement dit, cette approche prend en compte la structure tridimensionnelle de la molécule, y

compris les distances et les angles entre les atomes, ce qui est essentiel pour comprendre avec précision les propriétés physiques et chimiques **(Wang, Jiang, Wang, & Xuan, 2024).**

Chapitre 3

Matériels et Méthodes

1. Récupération des données de la base CHEMBEL

Les données ont été récupéré de la base de sonnées CHEMBEL, une ressource majeure et publique dédiée à la bioactivité des molécules d'intérêt pharmaceutique,en utilisant des scripts python sur googlecolab. ChEMBL regroupe des données sur les propriétés chimiques, la bioactivité, ainsi que les cibles moléculaires associées à de nombreux composés, facilitant ainsi la recherche et le développement de nouveaux médicaments**(1)**

Le modulechembl_webresource_clientest installé pour permettre l'accès aux services web de ChEMBL. Ensuite, les bibliothèquespandaspour la manipulation des données tabulaires et le client ChEMBL sont importés.

```
! pipinstallchembl_webresource_client
import pandas aspd
fromchembl_webresource_client.new_clientimportnew_client
```

Accès à l'API et requête sur les cibles en utilisant l'objet target qui est instancié via le client, permettant d'interroger la base sur les cibles moléculaires. La méthode search('CHEMBL340')effectue une recherche ciblée sur l'identifiant d'inetret ChEMBL340, qui correspond à une cible moléculaire précise (Cytochrome P450 3A).

```
target = new_client.target
target_query = target.search('CHEMBL340')
targets = pd.DataFrame.from_dict(target_query)
targets
```

```
selected_target = targets.target_chembl_id [0]
selected_target
```

Les résultats de la requête sont convertis en un DataFrame, ce qui permet une manipulation aisée des données, leur exploration et leur analyse statistique.

Pour sélectionner les lignes rapportant des valeurs d'IC50, il faut filtrer le DataFrame obtenu à partir de la requête sur deux critères, la colonnestandard_typedoit être égale à "IC_{50}" et la colonnestandard_unitsdoit être égale à "µM". Ce filtrage permet d'isoler uniquement les résultats d'activité mesurés en micromolaires pour l'inhibition de la cible sélectionnée.

```
activity = new_client.activity
res = activity. filter (target_chembl_id=selected_target).filter(standard_type="IC50")
df1 = pd.DataFrame.from_dict(res)
df1
```

Ce code retourne uniquement les lignes où l'activité est mesurée en IC_{50} et exprimée en µM, ce qui est essentiel pour garantir l'homogénéité des données lors de l'analyse et de la comparaison des valeurs d'inhibition.

2. Pré-traitement de données

La filtration du Data Frame df1 obtenuest crucial afin de ne conserver que les lignes contenant des valeurs non manquantes dans la colonne standard value (qui représente la valeur IC50). Ce dernier permet d'éviter les erreurs de calcul qui ne traitent pas les valeurs manquantes (NaN) et garantit la qualité des données pour une utilisation ultérieure sans problèmes secondaires.

```
df2 = df1[df1.standard_value.notna()]
df2
```

Nous sélectionnons uniquement certaines colonnes du DataFrame df2 pour créer un nouveau DataFrame df3 contenant uniquement ces colonnes. Cela est nécessaire pour simplifier les données (ce qui rend leur manipulation et leur analyse plus faciles) et réduire la taille du DataFrame (ce qui accélère les opérations et économise de la mémoire), car le DataFrame d'origine contient de nombreuses colonnes non nécessaires.

```
selection = ['molecule_chembl_id','canonical_smiles','standard_value', 'units']
df3 = df2[selection]
df3
```

Ensuite, Le nombre d'occurrences de chaque unité de mesure dans la colonne « units » du DataFrame df3 est calculé, ce qui permet de comprendre la distribution des unités, de détecter les incohérences et d'aider à la préparation des données.

```
unit_counts = df3['units'].value_counts()
print("Nombre de lignes pour chaque unité:")
print(unit_counts)
```

Le DataFrame df3 a été filtré pour ne conserver que les lignes où la valeur de la colonne « units » est égale à 'µM'(micromolaire), en supprimant toutes les lignes contenant des unités différentes de 'µM'. Ce qui aide à se concentrer sur une unité particulière et à nettoyer les données pour faciliter leur manipulation.

```
df3 = df3[df3['units'] == 'nM']
df3 = df3.reset_index(drop=True)
```

Ensuite, on fait exporter le DataFrame df3 dans un fichier au format CSV. Ce dernier permet de sauvegarder les données traitées précédemment pour l'utilisation dans l'étape suivante, ce qui est également très important pour le stockage et l'archivage des résultats du traitement et de l'analyse.

```
df3.to_csv('df3_DATA')
```

Ce fichier CSV est de nouveau lu et chargé dans un DataFrame nommé df4 à l'aide de la bibliothèque pandas, dans le but de récupérer les données précédemment enregistrées pour les analyser et les traiter.

```
df4 = pd.read_csv('/content/df3_DATA')
df4
```

Nous sélectionnons uniquement certaines colonnes du DataFrame df4 pour créer un nouveau DataFrame df5 qui ne contient que les colonnes dont nous avons besoin. Cela est nécessaire pour simplifier le DataFrame (lecture et traitement plus faciles et amélioration des performances des tâches ultérieures) et se concentrer sur les plus importantes. Il aide également à trier et nettoyer les données (prétraitement des données).

```
selection = ['molecule_chembl_id','canonical_smiles','standard_value', 'units']
df5 = df4[selection]
df5
```

L'étape suivante est portée sur la classification des molécules en trois catégories, actif, inactif et intermédiaire. Basé sur les valeurs de son activité biologique en utilisant une mesure numérique, ce regroupement facilite ainsi l'analyse et la présentation. Aussi, la préparation des données pour la modélisation, en plus d'accélérer la compréhension des résultats biologiques selon les limites de son activité.

```
bioactivity_class = []
for i in df5.standard_value:
 if float(i) >= 10000:
  bioactivity_class.append("Inactive")
 elif float(i) <= 1000:
  bioactivity_class.append("Active")
 else:
  bioactivity_class.append("Intermediate")
df5
```

Le Ce code ci-dessous est utilisé pour calculer le nombre total de valeurs manquantes (NaN) dans l'ensemble du DataFrame df5 rapidement. Cette étape est très importante pour évaluer la qualité des données et décider s'il est nécessaire de nettoyer les données ou de remplacer ces valeurs manquantes.

```
df5.isna().sum().sum()
```

Après ça, on exporte le DataFrame df5 vers un fichier au format CSV

```
df5.to_csv('df5_DATA')
```

Le script en dessous est utilisé pour ajouter une nouvelle colonne appelée « bioactivity_class » à notreDataFrame df6. C'est utile lorsque nous avons des résultats ou des classifications calculés séparément et que nous voulons les intégrer dans le même tableau,

ce qui facilite ensuite l'analyse, la présentation ou la construction de modèles en utilisant toutes les données combinées.

```
df6 = pd.read_csv('df5_DATA')
df6
```

```
bioactivity_class = pd.Series(bioactivity_class, name='bioactivity_class')
df7 = pd.concat([df6, bioactivity_class], axis=1)
df7
```

Nous passons ensuite à l'enregistrement du data dans un DataFrame df7, un fichier au format CSV

```
df7.to_csv('df7_DATA', index=False)
```

3. Normalisation des données

Cette fonction norm_valueest utilisée pour normaliser les valeurs présentes dans la colonne 'standard_value' du notre DataFrame df7. Plus précisément, elle fixe un plafond ou une limite maximale pour les valeurs afin qu'elles ne dépassent pas un seuil maximal. Puis, les valeurs modifiées sont stockées dans une nouvelle colonne en supprimant la colonne originale afin de traiter les valeurs extrêmes et d'uniformiser les données pour réduire les valeurs aberrantes, et donc obtenir des données plus propres et plus adaptées à l'analyse et à la réalisation du modèle prédictif.

```
defnorm_value(df7):
  norm = []
  for i in df7['standard_value']:
    if i >100000000:
      i = 100000000
    norm.append(i)
  df7['IC50_nM_norm'] = norm
  x = df7.drop('standard_value', axis=1)
  return x
```

```
df_norm = norm_value(df7)
df_norm
```

```
df_norm.IC50_nM_norm.describe()
```

Le but de ce code est de normaliser les données présentes dans df7 et de préparer les données pour un traitement ultérieur en garantissant que toutes les variables soient sur une échelle uniforme et comparable.

4. Convertir IC_{50} en pIC_{50}

Pour convertir les valeurs IC_{50} en pIC_{50}, une échelle logarithmique plus adaptée à l'analyse, en utilise le script python en dessous. On stocke ensuite les résultats dans le cadre

de données df8, qui inclut maintenant la nouvelle colonne pIC50. Cette conversion permet une analyse facile des données biologiques et chimiques en raison de son importance dans les modèles prédictifs et la mesure de l'efficacité des molécules.

```
import pandas aspd
importnumpyasnp
```

```
def calculate_pIC50(IC50_nM_norm):
  if IC50_nM_norm > 0:
    return -np.log10(IC50_nM_norm * 1e-9)
  else:
    return np.nan

df_norm['pIC50'] = df_norm['IC50_nM_norm'].apply(calculate_pIC50)

df8 = df_norm
df8
```

```
df8.columns
```

Cette opération est suivie par exportation des données présentes dans df8 vers un fichier au format CSV.

```
df9.to_csv('df10_DATA.csv', index=False)
```

5. Exploration des donnéesanalysées

Afin d'explorer nos notre data, on commence par une préparation des bibliothèques nécessaires, à savoir Matplotlib (qui est la bibliothèque principale pour le tracé de graphiques) et Seaborn (qui est une bibliothèque basée sur Matplotlib et qui applique un style visuel spécifique pour améliorer automatiquement l'esthétique des graphiques). Ce script permet d'analyser les données visuellement et de présenter les résultats de manière professionnelle dans les rapports ou les modèles.

```
importseabornassns
sns.set(style='ticks')
import matplotlib.pyplot as plt
```

On crée donc un histogramme affichant la distribution et le nombre de molécules (la fréquence) dans chaque catégorie d'activité biologique en utilisant les deux bibliothèques mentionnées précédemment, puis enregistrer ce graphique en tant que fichier image png.

```
plt.figure(figsize=(5.5, 5.5))
sns.countplot(x='bioactivity_class', data=df8, edgecolor='black')
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
```

```
plt.ylabel('Frequency', fontsize=14, fontweight='bold')
plt.savefig('plot_bioactivity_class.pdf')
```

Également, le script ci-dessous est également utilisé pour créer un nouveau graphique de type boxplot (boîte à moustaches) à partir des mêmes bibliothèques. Ce type de graphique permet de comparer la distribution des valeurs de pIC50 entre les différentes catégories d'activité biologique, et d'évaluer la qualité des données ainsi que l'efficacité des composés étudiés.

```
plt.figure(figsize=(5.5, 5.5))
sns.boxplot(x = 'bioactivity_class', y = 'pIC50', data = df8)
plt.xlabel('Bioactivity class', fontsize=14, fontweight='bold')
plt.ylabel('pIC50 value', fontsize=14, fontweight='bold')
plt.savefig('plot_ic50.pdf')
```

6. Téléchargement de PaDEL-Descriptors

Dans cette étape on passe au calcul des descripteurs de chaque molécule étudiée, en installant préalablement à partir du dépôt GitHub vers l'environnement de travail Google Colab :

- Le programme PaDEl-Descriptorqui est utilisé pour extraire les descripteurs chimiques etles empreintes des composés à partir de fichiers de molécules tels que SMILES ou SDF.
- Script Shell qui est un programme simple utilisé pour exécuter le programme PaDEL automatiquement.

Ces fichiers sont nécessaires pour analyser les composés biochimiques et extraire les données (descripteurs) que nous utilisons ensuite pour construire des modèles d'apprentissage automatique.

```
! wget https://github.com/dataprofessor/bioinformatics/raw/master/padel.zip
! wget https://github.com/dataprofessor/bioinformatics/raw/master/padel.sh
```

```
! unzip padel.zip
```

Pour bien organiser, un nouveau script qui permet d'extraire des colonnes spécifiques (la structure moléculaire, l'identifiant de la molécule chEMBL ID et la valeur d'activité pIC50) de df8 a été appliqué, puis de les stocker dans un nouveau DataFrame df10. Ensuite, il est exporté dans un nouveau fichier nommé 'molecule.smi' pour l'étape suivante pour la construction de modèles QSAR.

```
selection = ['canonical_smiles','molecule_chembl_id','pIC50']
df10 = df9[selection]
df10.to_csv('molecule.smi', sep='\t', index=False, header=False)
```

```
! cat molecule.smi | head-5
```

Le contenu du data est récupéré dans un nouveau DataFrame df11 en utilisant la bibliothèque pandas. Ce fichier « df11 » est affiché partiellement pour vérifier que le fichier a été enregistré correctement et est prêt pour la poursuite du traitement ou de l'analyse.

```
df11 = pd.read_csv('molecule.smi')
df11
```

Ensuite, on commence à extraire les propriétés moléculaires en utilisant le script suivant :

```
! cat padel.sh
```

```
! bash padel.sh
```

```
! ls -l
```

7. Préparation des matrices de données X et Y

Dans cette partie, nous importons les descripteurs moléculaires calculées et enregistrés, et les convertir en un DataFrame df12, puis les afficher pour examiner et vérifier leurs qualités.

```
df12 = pd.read_csv('/content/descriptors_output.csv')
df12
```

Ensuite on supprime la colonne nommée 'Name' de df12 en raison de son insignifiance dans l'analyse numérique ou la modélisation, et crée un nouveau DataFrame df13 sans cette colonne, puis l'affiche. Cette étape est utile pour nettoyer et réduire la taille du DataFrame et le préparer pour les algorithmes d'apprentissage automatique qui traitent généralement uniquement des données numériques.

```
df13 = df12.drop(columns=['Name'])
df13
```

Parallèlement, on se concentre sur l'extraction et l'isolement de la colonne pIC50 uniquement du DataFrame df8 et son stockage dans une nouvelle variable appelée df14. Le but de ce code est d'analyser la colonne pIC50 de manière indépendante et de l'utiliser comme variable cible dans le modèle d'apprentissage automatique.

```
df14= df10_1 ['pIC50']
df14
```

Afin de finaliser notre jeu de données, on procède a la fusion deux tableaux [df13, df14] côte à côte (c'est-à-dire fusionne les colonnes) de manière horizontale, afin d'obtenir un nouveau DataFrame nommé df15 qui combine les caractéristiques moléculaires numériques et les valeurs de l'activité biologiques de chaque molécule. L'objectif de ce code est de préparer les données pour les utiliser dans l'entraînement de notre modèle d'apprentissage automatique.

```
df15 = pd.concat([df13,df14], axis=1)
df15
df15.to_csv('df15_DATA.csv', index=False)
```

8. Construction des modèles de régression

On commence cette partie par l'importation des bibliothequesnecessaires, pour traiter les données (pandas), visualiser et analyser les données (Seaborn), et diviser les données en un ensemble d'entraînement et un ensemble de test en utilisant la fonction train_test_split importée de la bibliothèque scikit-learn, tout en important le Random Forest Regressor, qui est un algorithme d'apprentissage automatique puissant et efficace pour prédire les valeurs numériques (régression). Tout cela vise à préparer un environnement logiciel pour entraîner un modèle d'apprentissage automatique solide.

```
import pandas aspd
importseabornassns
fromsklearn.model_selectionimport train_test_split
fromsklearn.linear_modelimportLinearRegression, Ridge, Lasso
fromsklearn.ensembleimportRandomForestRegressor, GradientBoostingRegressor
fromsklearn.svmimport SVR
fromsklearn.preprocessingimportStandardScaler
fromsklearn.metricsimport r2_score, mean_squared_error, mean_absolute_error
```

Ensuite, on commence par supprimer les colonnes non numériques (telles que pIC50 et canonical_smiles) pour ne conserver que les descripteurs moléculaires utilisables par les algorithmes d'apprentissage.

```
features = df.drop(columns=['pIC50', 'canonical_smiles'], errors='ignore')
```

Egalement, on élimine automatiquement les descripteurs présentant une corrélation supérieure à 80 % entre eux, afin de réduire la redondance et d'éviter les problèmes de multicolinéarité pouvant nuire à la performance et à l'interprétabilité du modèle. Cette opération permet ainsi de sélectionner un ensemble de variables plus pertinent et moins redondant pour l'entraînement du modèle QSAR

```
defremove_correlated_features(data, threshold=0.8):
  corr_matrix = data.corr().abs()
  upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(bool))
  to_drop = [column for column in upper.columns if any(upper[column] > threshold)]
  return data.drop(columns=to_drop)
```

```
features_filtered = remove_correlated_features(features, threshold=0.8)
print(f"Nombre initial de descripteurs : {features.shape[1]}")
print(f"Nombre de descripteurs après filtrage : {features_filtered.shape[1]}")
```

Par la suite, nous appliquons une série de scripts dédiés à la visualisation graphique, afin de faciliter l'analyse et l'interprétation des descripteurs sélectionnés. Ces visualisations incluent la matrice de corrélation entre les descripteurs, qui permet d'identifier visuellement les relations linéaires et les éventuelles redondances persistantes. Nous présentons également la distribution des coefficients de corrélation, offrant un aperçu global du niveau de dépendance entre les variables retenues. Enfin, un graphique illustrant le nombre de

descripteurs sélectionnés après filtrage est généré, ce qui permet de suivre l'impact du prétraitement sur la réduction de la dimensionnalité.

```
features_filtered = remove_correlated_features(features, threshold=0.8)
print(f"Nombre initial de descripteurs : {features.shape[1]}")
print(f"Nombre de descripteurs après filtrage : {features_filtered.shape[1]}")
```

On passe maintenant à la Séparation des variables explicatives (X) et de la cible (y) ou les descripteurs moléculaires sont isolés dans X tandis que la variable cible pIC50 est stockée dans y. Les données sont séparées en ensembles d'entraînement et de test avec une répartition 90/10, en fixant une graine aléatoire pour la reproductibilité.

```
X = df_filtered.drop(columns=['pIC50'])
y = df_filtered['pIC50']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.10, random_state=42)
```

On définit trois modèles de régression et donc trois algorithmes sont configurés :

- RandomForestRegressor avec 100 arbres
- GradientBoostingRegressor avec 100 itérations
- SVR (Support VectorRegression) avec paramètres par défaut

```
models = {
  "RandomForest": RandomForestRegressor(n_estimators=100, random_state=42),
  "GradientBoosting": GradientBoostingRegressor(n_estimators=100, random_state=42),
  "SVR": SVR()
}
```

On finalise cette étape par l'évaluation des performances pour chaque prédiction sur le test set, trois métriques sont calculées :

- **R^2** (coefficient de détermination)
- **RMSE** (racine de l'erreur quadratique moyenne)
- **MAE** (erreur absolue moyenne)

Suivie d'une visualisation graphique des résultats ce qui permet une comparaison immédiate de l'efficacité des modèles.

Chapitre 4

Résultats et Discussion

1. Récupération des données de la base CHEMBEL

CHEMBL est une base de données utilisée comme ressource de référence pour extraire des données sur les molécules bioactives. Dans notre étude, la cible sélectionnée, le cytochrome P450 3A (CHEMBL340), est d'une grande importance pharmacologique car elle est impliquée dans le métabolisme de nombreux médicaments.

Parmi les **34 417** molécules récupérées initialement pour cette cible, seules **11 166** ont des valeurs de IC50 déterminées expérimentalement, soit environ **32,4 %** du total. Cela signifie qu'environ deux tiers des molécules **(67,6 %)** manquent de cette information critique et ont été supprimées de l'analyse ultérieure comme le montre la figure III.01.

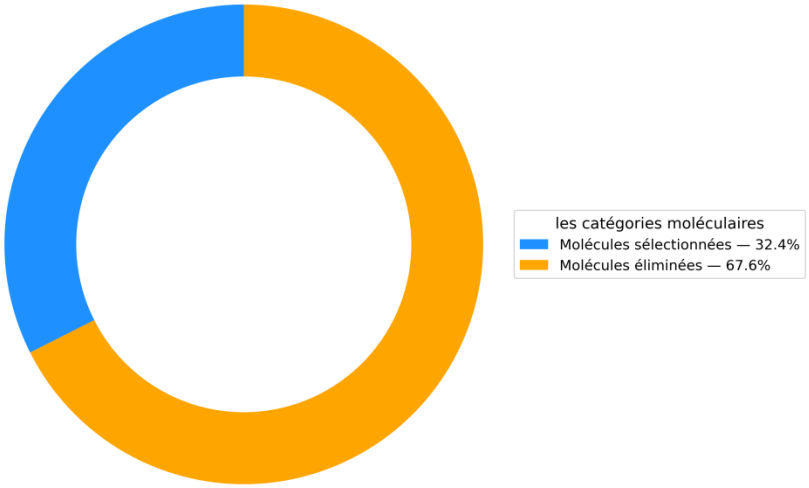

Fig. III.01 : Répartition des molécules étudiées pour la cible cytochrome P450 3A.

Le fait que seulement un tiers des molécules ait une valeur IC50 exploitable met en lumière une limitation courante dans les bases de données publiques ou beaucoup de composés manquent de données expérimentales complètes. Cette sélection rigoureuse est pourtant cruciale pour assurer la qualité des modèles prédictifs QSAR qui ont besoin de valeurs quantitatives fiables pour leur construction et leur validation

2. Distribution des unités de calcul de IC50

Parmi les **11166** molécules sélectionnées, la majorité écrasante (**83,9 %**) présente une IC_{50} exprimée en micromolaire (µM), tandis qu'une proportion nettement moindre (**15,8 %**) l'exprime en nanomolaire (nM), et une fraction résiduelle (**0,3 %**) dans d'autres unités. Pour garantir l'homogénéité et la comparabilité des données, seules les valeurs en micromolaire ont été retenues pour l'analyse, les autres unités ayant été éliminées. Ce choix méthodologique assure la cohérence des résultats, bien qu'il réduise la diversité des données exploitées, mais il reste essentiel pour la robustesse des analyses quantitatives ultérieures. L'ensemble des détails sont organisés dans le tableau Tab.III.01.

Tab. III. 01 : Résumé des molécules sélectionnées et leurs unités de mesure

Cible sélectionnée	Cytochrome P450 3A		
CHEMBL ID	CHEMBL340		
Nombre total de molécules étudiées	34 417		
Molécules présentant une valeur IC_{50}	11 166		
Unités de mesure (IC_{50})	Nanomollaires (nM)	1739	Eliminées
	Micromollaire (µM)	9213	Sélectionnées
	Autres	214	Eliminées

La figure (Fig.III.01) présente une illustration graphique des unités de mesure de l'activité pharmacologique, mettant en avant **9213** molécules qui ont été sélectionnées pour poursuivre le parcours.

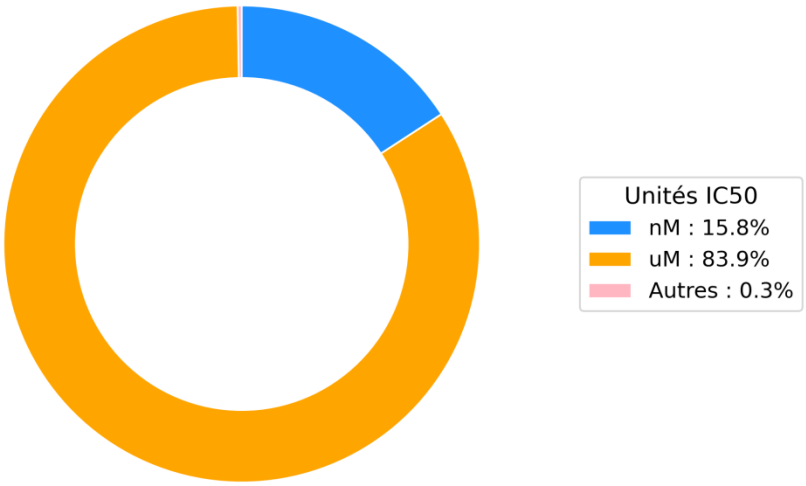

Fig. III.02 : Répartition des molécules étudiées Selon l'unité de mesure de l'IC_{50}.

3. Classification des molécules sélectionnées

L'analyse des **9213** molécules présentant une activité pour le Cytochrome P450 3A, selon le seuil prédéterminé, révèle une nette domination des composés inactifs, qui représentent presque deux tiers de l'ensemble (**5795**, soit **62,90%**). Environ un quart (**2223**, soit **24,24%**) des molécules donnent une activité intermédiaire, tandis qu'une petite fraction (**1185**, soit **12.86%**)est active. Cette répartition déséquilibrée, mise en lumière par l'histogramme (Fig. III. 03), souligne la difficulté d'identifier des inhibiteurs puissants pour cette enzyme polyvalente. Cela met également en avant l'importance d'adapter nos méthodes de modélisation pour mieux prendre en compte la sous-représentation des molécules actives et ainsi améliorer l'efficacité des modèles prédictifs.

L'utilisation de l'approche multi-classes de bio-activité (active, intermédiaire et inactive) permet d'améliorer la qualité du modèle prédictif en offrant une vision plus nuancée et réaliste de la diversité des réponses biologiques, et permet ainsi au modèle d'être plus robuste et pertinent.

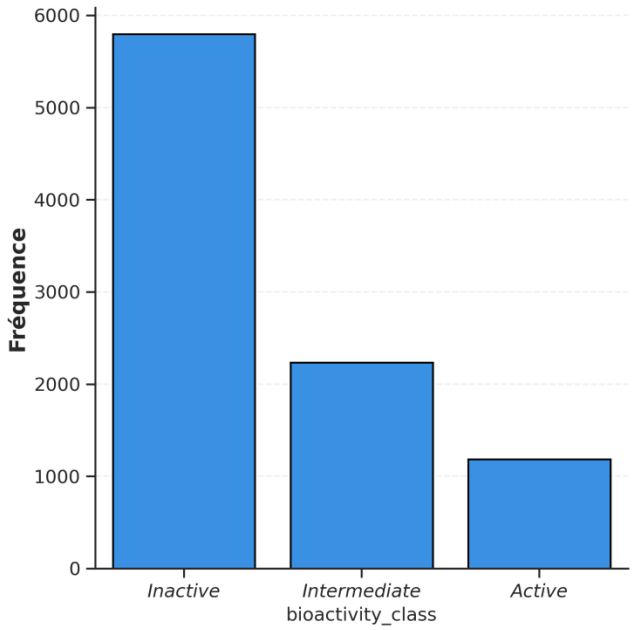

Fig. III.03 : Classification des molécules selon le niveau de l'activité pharmacologique

4. La normalisation de données de l'activité pharmacologique

La normalisation des valeurs d'IC_{50} est une étape clé dans la création de modèles QSAR en machine Learning. Elle permet d'uniformiser l'échelle des données, de réduire l'impact des valeurs extrêmes et d'assurer une meilleure comparabilité entre les différents composés analysés. Ce prétraitement renforce la stabilité, la robustesse et la performance prédictive des modèles, tout en aidant les algorithmes d'apprentissage automatique à converger plus facilement. C'est essentiel pour obtenir des prédictions fiables et généralisables concernant l'activité biologique de nouvelles molécules.

5. La Conversion de IC_{50} en pIC_{50}

La conversion des valeurs d'IC_{50} en pIC_{50} constitue une étape fondamentale dans la préparation des données, car elle permet de linéariser et de normaliser la distribution des activités biologiques, facilitant ainsi l'analyse statistique et la comparaison des composés. En effet, le pIC50, défini comme le logarithme négatif décimal de l'IC_{50} exprimé en molaire (M), inverse l'échelle de puissance : plus le pIC50 est élevé, plus la molécule est puissante, ce qui rend la lecture des résultats plus intuitive et réduit l'impact des valeurs extrêmes.

Les figures **Fig. III.04** et **Fig. III.05** montre que la distribution des pIC50 est asymétrique, avec une majorité de composés faiblement actifs et une minorité de molécules très puissantes, ce qui met en évidence l'intérêt de cette transformation pour obtenir une meilleure répartition des données et optimiser la performance des algorithmes

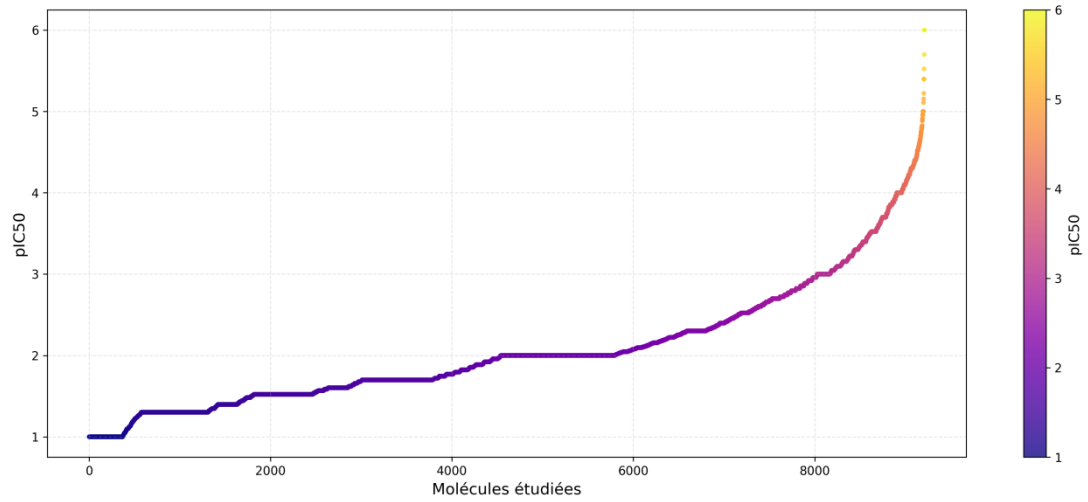

Fig. III.04 : Variation des valeurs de pIC50 en fonction des molécules étudiées

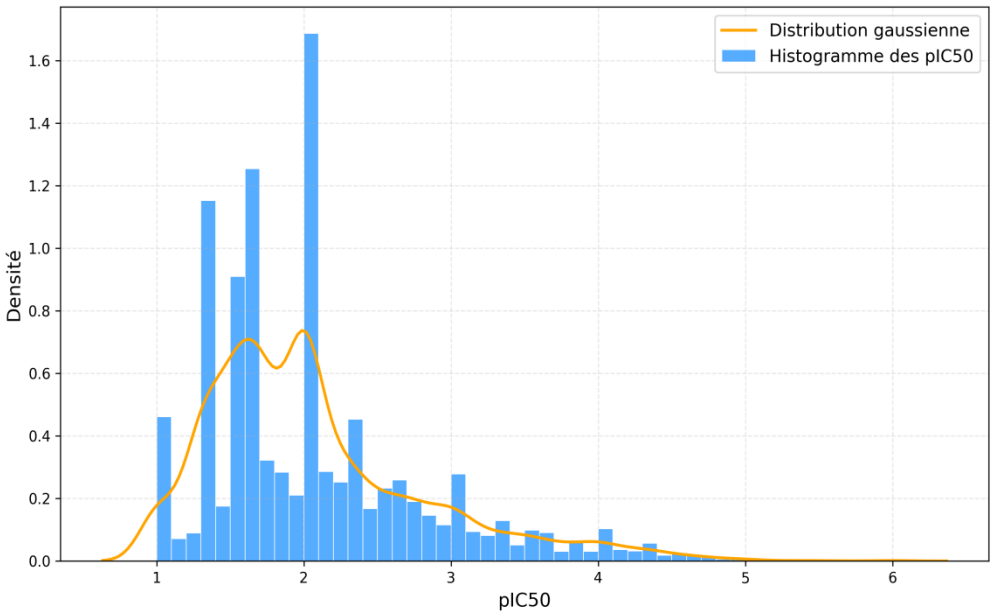

Fig. III.05 : Distribution des valeurs de pIC50 pour l'ensemble des molécules

En parallèle, la figure **Fig.III.06** présente la distribution des valeurs de pIC50 pour les trois classes de bio-activité. On observe que la médiane et l'étendue des valeurs de pIC50 augmentent progressivement de la classe inactive vers la classe active, ce qui reflète une augmentation de la puissance biologique des composés. Les boîtes montrent la dispersion des valeurs, avec quelques points aberrants particulièrement visibles dans la classe active, traduisant la présence de molécules exceptionnellement puissantes.

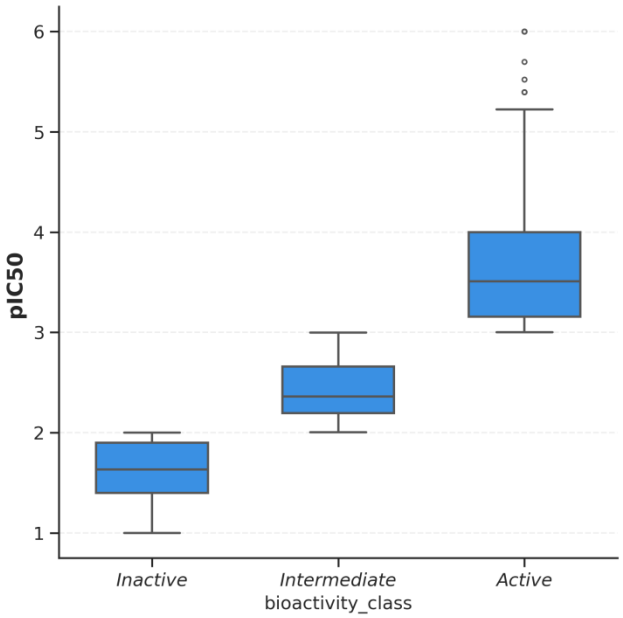

Fig. III.06 : Répartition des valeurs de pIC50 selon les classes de bio-activité

6. Calcul et filtration des fingerprints moléculaires

Le calcul des descripteurs moléculaires *fingerprints* à l'aide de l'outil PaDEL est essentiel dans la construction de modèles prédictifs QSAR, permettant de traduire des structures chimiques en données quantitatives exploitables par des algorithmes de machine learning. PaDEL calcule **881** descripteurs (incluant des propriétés 1D/2D comme la polarisabilité, les indices topologiques ou les comptes de sous-structures, ainsi que des descripteurs 3D tels que les moments d'inertie). Son architecture multithread optimise les calculs pour des jeux de données volumineux, tandis que sa compatibilité avec plus de 90 formats de fichiers moléculaires (SDF, SMILES, etc.) et son intégration via des wrappers Python facilitent son utilisation dans les pipelines de modélisation. Ces descripteurs, combinés à des méthodes d'apprentissage automatique, permettent de prédire l'activité biologique des molécules en identifiant des motifs structuraux corrélés à leur efficacité, tout en assurant reproductibilité et standardisation des données.

La figure **Fig. III.07** représente une matrice de corrélation des fingerprints moléculaires calculés pour l'ensemble des composés étudiés. Chaque case de la matrice indique le degré de corrélation entre deux descripteurs. On observe une prédominance de bleu, traduisant une faible redondance globale entre la majorité des descripteurs, mais aussi la présence de quelques zones orange, qui signalent des groupes de descripteurs fortement corrélés entre eux. Cette visualisation met en évidence la diversité des informations apportées par les fingerprints, tout en soulignant la nécessité d'identifier et éventuellement d'éliminer les descripteurs redondants afin d'optimiser la qualité et la performance du modèle prédictif.

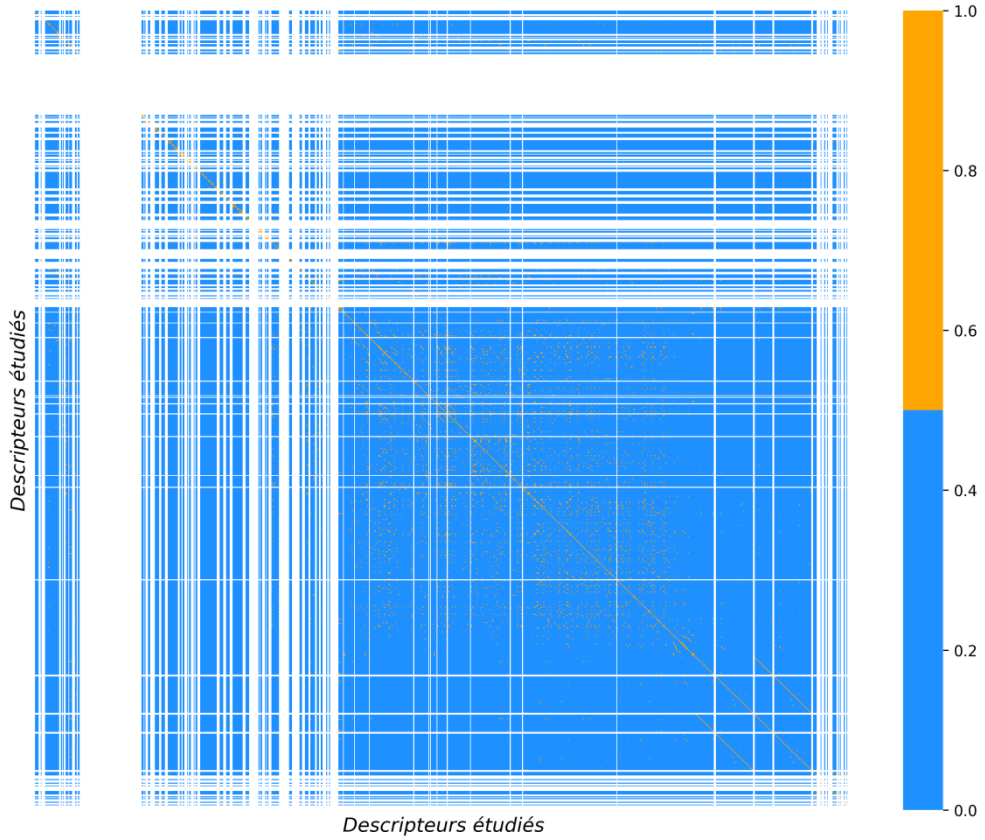

Fig. III.07 : Matrice de corrélation entre les descripteurs moléculaires (fingerprints)

Le nombre de descripteurs moléculaires utilisés pour la modélisation prédictive après l'élimination des descripteurs fortement corrélés (corrélation ≥ **80 %**) est présenté dans la figure **Fig. III.08.** Avant filtration, **881** descripteurs étaient pris en compte, alors qu'après cette étape de sélection, seuls **555** descripteurs indépendants ont été conservés, soit **63 %** du total initial. Cette réduction significative permet de limiter la redondance des informations, d'améliorer la performance et la robustesse du modèle prédictif, tout en facilitant son interprétation.

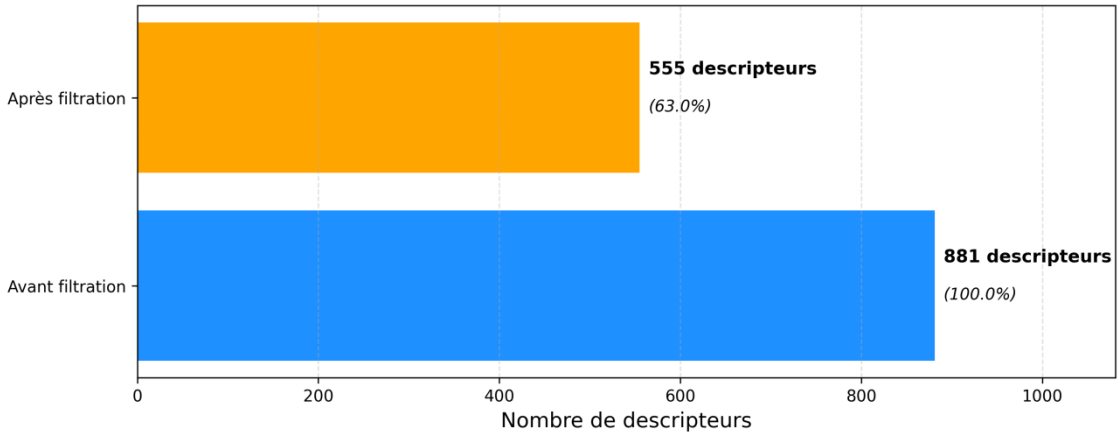

Fig. III.08 : Répartition de descripteurs après élimination des corrélations élevées

7. Séparation des données et construction du model

La construction d'un modèle prédictif en machine learning repose sur une séparation rigoureuse des données en ensembles d'entraînement et de test. Notre méthode a adopté le *train-test split* aléatoire (90-10%) pour séparer les données. La validation externe sur un ensemble test indépendant est cruciale pour vérifier la robustesse du modèle, avec des métriques convenables à déterminer.

Les résultats obtenus montrent que le modèle Gradient Boosting a été le meilleur par rapport aux autres algorithmes testés, avec un coefficient de détermination R^2 de **0,896**, indiquant une excellente capacité à expliquer la variance des données. Ce modèle a donné également les plus faibles valeurs de RMSE (**0,217**) et de MAE (**0,176**), traduisant une meilleure précision et une moindre erreur moyenne absolue dans les prédictions. Le Random Forest a été positionné en second, avec un R^2 de **0,87** et des erreurs légèrement supérieures (RMSE de **0,243** et MAE de **0,193**), ce qui confirme sa robustesse mais suggère qu'il capture moins bien la complexité des relations non linéaires par rapport au Gradient Boosting. Enfin, ش‌le modèle SVR (Support VectorRegression) présente les performances les plus modestes parmi les trois, avec un R^2 de **0,839** et les erreurs les plus élevées (RMSE de **0,270** et MAE de **0,202**), indiquant qu'il est moins adapté à ce jeu de données ou nécessite un ajustement plus poussé des hyperparamètres. L'ensemble des résultats obtenus sont organisés dans les figures **Fig. III.09**, **Fig. III.10**et **Fig. III.11**.

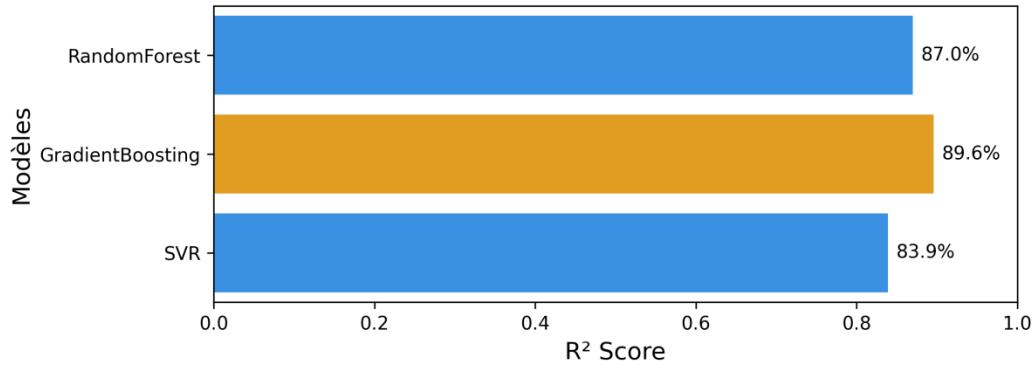

Fig. III.09 : Précision des modèles prédictifs testés

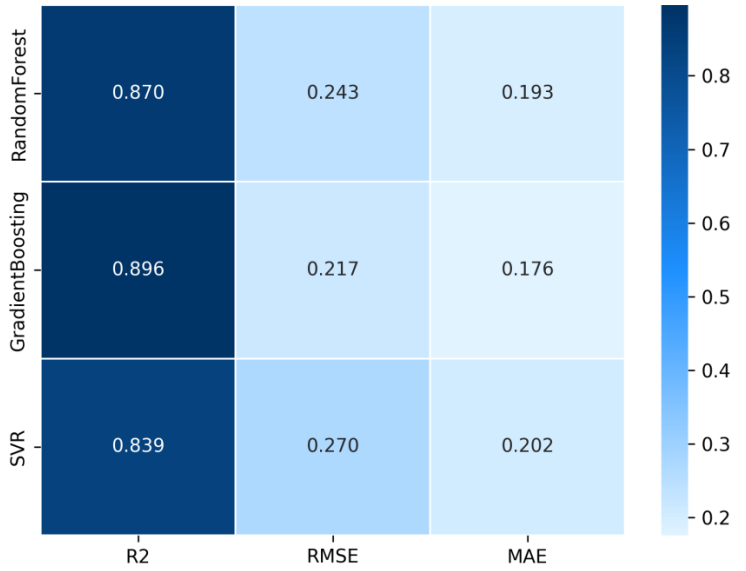

Fig. III.10 : présentation de la performance des modèles étudiés

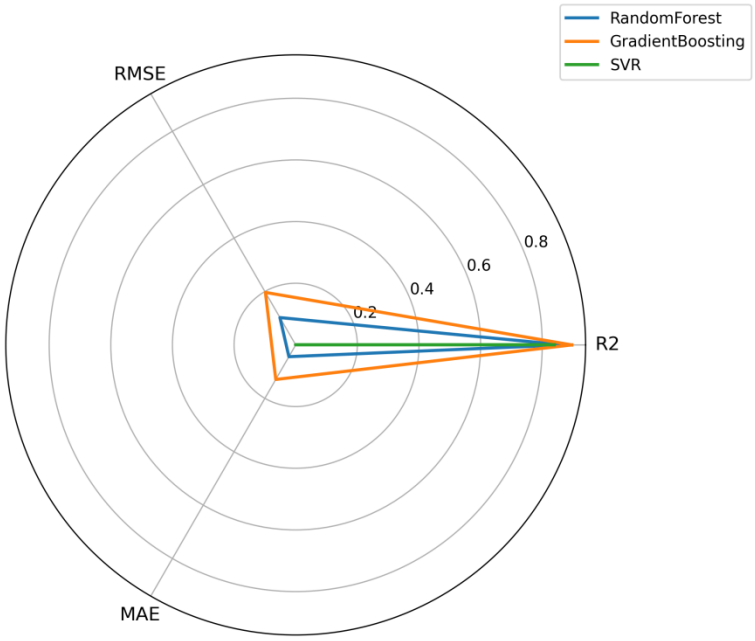

Fig. III.11 : présentation Radar des performances normalisées

Les performances des modèles Gradient Boosting, Random Forest et SVR ont été évaluées à l'aide de métriques telles que le coefficient de détermination (R^2), l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE). Ces indicateurs permettent d'estimer la justesse des prédictions par rapport aux valeurs réelles. Une forte corrélation entre les valeurs prédites et expérimentales traduit la capacité du modèle à capturer la relation entre les descripteurs moléculaires et l'activité biologique. Ainsi, l'analyse graphique de la corrélation pour chaque modèle fournira une visualisation claire de leur précision respective et mettra en évidence les éventuels écarts ou biais dans les prédictions.

Les différents graphiques de corrélation de chaque modèle sont représentés dans les figures **Fig. III.12**, **Fig. III.13** et **Fig. III.14**.

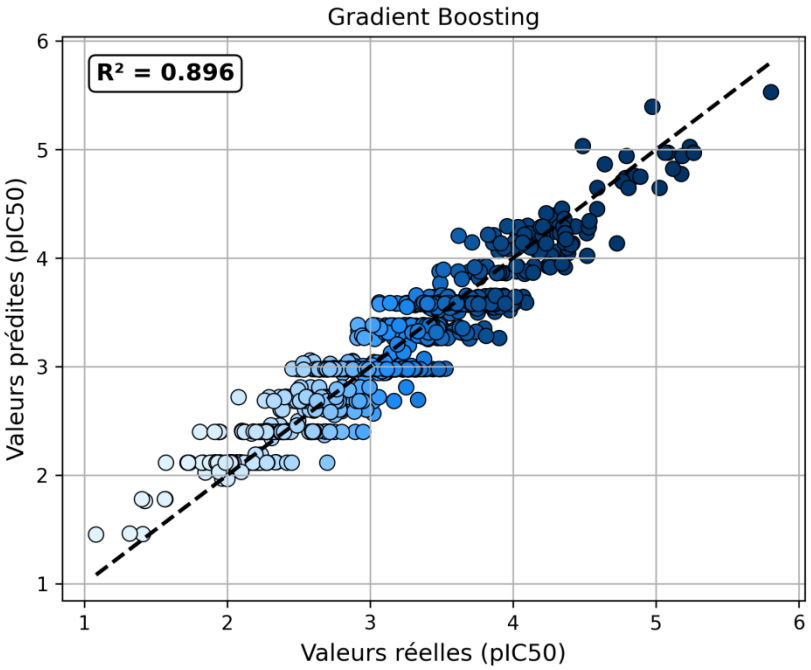

Fig. III.12 : corrélation du modèle gradient Boostig

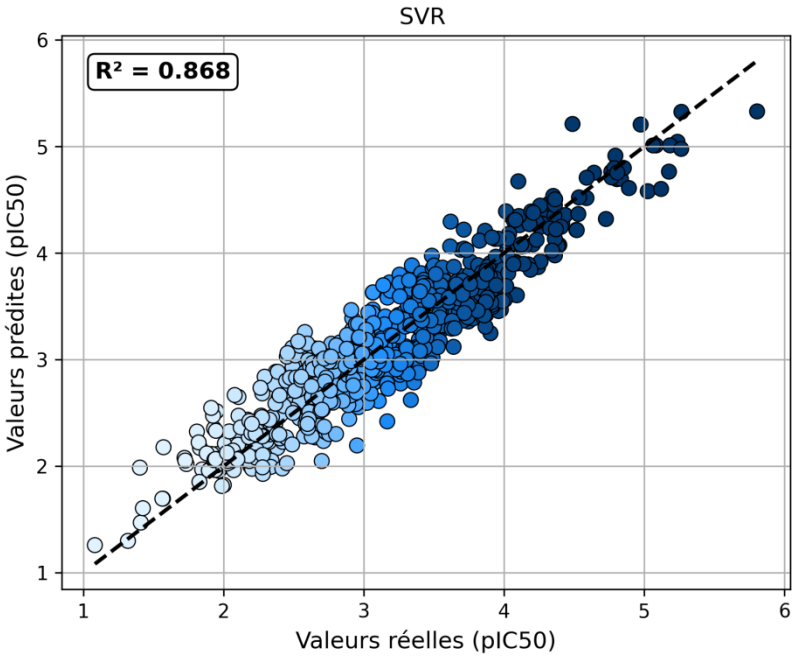

Fig. III.13 : corrélation du modèle SVR

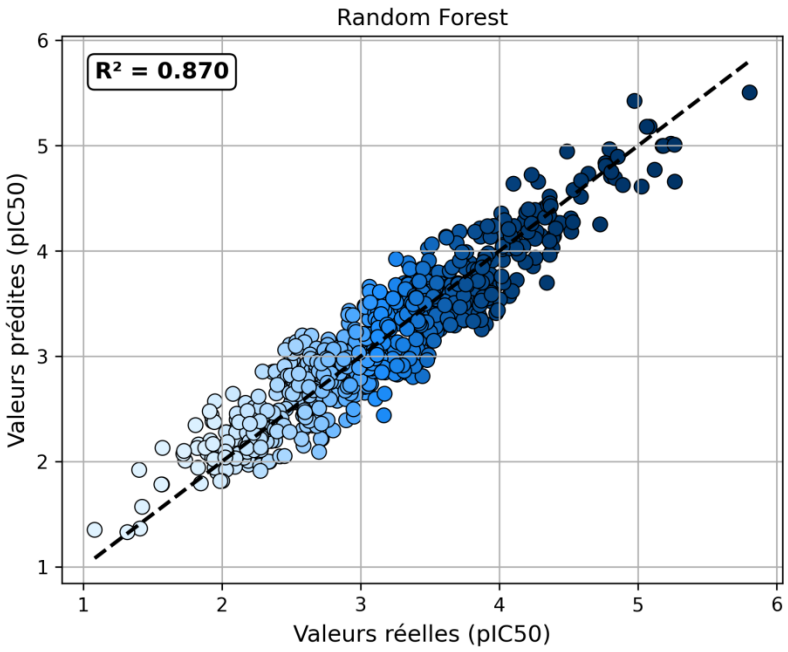

Fig. III.14 : corrélation du modèle Random forest

Conclusion

En conclusion, ce travail s'inscrit dans le cadre du développement de modèles QSAR prédictifs reposant sur des techniques d'apprentissage automatique, dans le but de prédire l'activité biologique de nouvelles molécules à partir de leurs descriptions chimiques. Nous avons développé, à travers cette étude, un modèle prédictif de l'activité des composés chimiques vis-à-vis de l'enzyme CYP3A4, l'une des enzymes clés dans le métabolisme des médicaments, en nous basant sur des données extraites de la base ChEMBL et en les traitant à l'aide d'outils spécialisés pour extraire les propriétés structurelles et moléculaires.

En comparant les performances de plusieurs algorithmes d'apprentissage automatique, nous avons pu identifier les méthodes les plus efficaces, en tête desquelles l'algorithme Gradient Boosting, qui a montré une grande capacité à modéliser la relation complexe entre la structure du composé et son activité biologique, obtenant ainsi les meilleurs résultats en termes de précision et d'indicateurs de performance. Cette approche contribue à accélérer le processus de découverte des composés biologiquement actifs, tout en réduisant la dépendance aux expériences de laboratoire coûteuses et chronophages. Elle souligne également l'importance de la vérification rigoureuse et de l'analyse comparative des modèles pour garantir la fiabilité des prédictions.

Ces résultats confirment le rôle croissant des technologies de l'intelligence artificielle dans l'accélération des phases de découverte des médicaments, en réduisant les coûts et en améliorant la qualité des décisions aux premières étapes du développement pharmaceutique. Elle renforce également l'importance de l'exploitation des données publiques ouvertes et de la modélisation mathématique dans la fourniture de solutions innovantes en pharmacologie.

À la lumière de ces données, nous proposons dans les travaux futurs d'élargir la base de données pour inclure d'autres cibles thérapeutiques, et d'expérimenter des algorithmes plus complexes tels que les réseaux neuronaux profonds, avec la possibilité d'intégrer des techniques de vérification biologique aux informations cliniques pour obtenir des modèles plus réalistes et complets. Ces approches pourraient contribuer à la construction d'outils prédictifs efficaces soutenant la recherche pharmaceutique contemporaine et s'alignant sur la transformation mondiale vers la médecine de précision et ciblée.

Références bibliographiques

Aiboud, L., &Laskri, S. (2020). Appréciation de la qualité des leads dans le marketing numérique à l'aide de l'apprentissage profond [Mémoire de master, Université Mohamed Khider – Biskra]. DSpace UMMTO.

Alizadehsani, R., Oyelere, S. S., Hussain, S., Calixto, R. R., de Albuquerque, V. H. C., Roshanzamir, M., Rahouti, M., &Jagatheesaperumal, S. K. (2023, November 2).*Explainable artificial intelligence for drug discovery and development – A comprehensive survey* (arXiv:2309.12177v2). arXiv.

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., ... & Zhang, J. D. (2020). An introduction to machine learning. *Clinical pharmacology & therapeutics*, *107*(4), 871-885. S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., ... & Zhang, J. D. (2020). An introduction to machine learning. *Clinical pharmacology & therapeutics*, *107*(4), 871-885.

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., ... & Zhang, J. D. (2020). An introduction to machine learning. *Clinical pharmacology & therapeutics*, *107*(4), 871-885.

Belghiti, A. A. (2021). Prédiction de situations anormales par apprentissage automatique pour la maintenance prédictive: approches en transport optimal pour la détection d'anomalies (Doctoral dissertation, Université Paris-Saclay).

Benamar, N. (2021). *Utilisation de l'apprentissage automatique pour la prédiction des propriétés pharmacologiques des composés chimiques* [Mémoire de Master, Université de Mila]. DSpace Université de Mila

Benamar, N. (2021). *Utilisation de l'apprentissageautomatique pour la prédiction des propriétés pharmacologiques des composéschimiques [Mémoire de Master, Université de Mila]. DSpace Université de Mila.

Benmaamar, O. (2023). Classification des images mammographiques selon leurs densités (Mémoire de master, Université 8 Mai 1945 de Guelma, Département de l'Informatique).

Boucher, P. (2023). Introduction aux réseaux conceptuels appliqués à l'apprentissage automatique des machines (Thèse de doctorat électronique, École de technologie supérieure, Montréal).

Boukertouta, M. A. (2022). Détection des intrusions basée sur l'apprentissage automatique dans les systèmes IdO (Internet des Objets) (Mémoire de master, Université 8 Mai 1945 de Guelma, Département de l'Informatique).

Çelik, Ö. (2018). A research on machine learning methods and its applications. Journal of Educational Technology and Online Learning, 1(3), 25-40.

CHEFROUR, A., & SOUICI-MESLATI, L. (2013). Un panorama de méthodes d'apprentissage incrémental. Atelier CIDN Classification Incrémentale et Détection de Nouveauté.

Chikhi, R., &Zitouni, I. (2024). Développement d'un modèle de machine learning (apprentissage automatique) pour la classification automatique des organismes microscopiques (Mémoire de master, Université de Mouloud Mammeri, TiziOuzou). Faculté des Sciences de la Nature et de la Vie. Retrieved from

EL MASSARI, H. A. K. I. M. (2023). Proposition d'un modèle de prédiction basé sur Machine Learning et le web sémantique.

Géron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems: O'Reilly Media, Inc.; 2017.

Guengerich, F. P. (2019). Cytochrome P450 research and the journal of biological chemistry. Journal of Biological Chemistry, 294(5), 1671-1680.

Jiménez-Luna, J., Grisoni, F., Weskamp, N., & Schneider, G. (2021). Artificial intelligence in drug discovery: recent advances and future perspectives. Expert opinion on drug discovery, 16(9), 949-959.

Juvénal JVC. Introduction au Machine Learning avec Python. Consulté le 22 juillet 2024.URL.

Labiad, A. (2017). Sélection des mots clés basée sur la classification et l'extraction des règles d'association (Mémoire de maîtrise, Université du Québec à Trois-Rivières, Maîtrise en mathématiques et informatique appliquées).

Liu, S., Lu, Y., Chen, S., Hu, X., Zhao, J., Lu, Y., & Zhao, Y. (2024, November 23). *DrugAgent: Automating AI-aided drug discovery programming through LLM multi-agent collaboration* (arXiv:2411.15692v1). arXiv.

Ma, Z., Xie, Z., Liang, Y., Wang, L., Yuan, S., & Zhang, X. (2024, February 15). *Drug discovery with generative models: Progress, challenges, and opportunities* (arXiv:2402.08703v1). arXiv.

Mahapatra, M. K., &Karuppasamy, M. (2022). Fundamental considerations in drug design. In Computer aided drug design (CADD): from ligand-based methods to structure-based approaches (pp. 17-55). Elsevier.

Mak, K. K., &Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. Drug discovery today, 24(3), 773-780.

Mak, K. K., Wong, Y. H., &Pichika, M. R. (2024). Artificial intelligence in drug discovery and development. Drug discovery and evaluation: safety and pharmacokinetic assays, 1461-1498.

MASARI HAKIM EL MASSARI, H. A. K. I. M. (2023). Proposition d'un modèle de prédiction basé sur Machine Learning et le web sémantique.

Masari Hakim El Massari. (2023). Proposition d'un modèle de prédiction basé sur Machine Learning et le web sémantique.

Matteis, L.D.; JANNY, S.; NATHAN, S.; QUARTIER, W.S. (2022) 'Introduction à l'apprentissage automatique'. paris. p. 18

Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A. G., Aidinis, V., ... &Melagraki, G. (2021). Advances in de novo drug design: from conventional to machine learning methods. International journal of molecular sciences, 22(4), 1676.

Olivier, R. N. en vue de l'obtention du DIPLOME de MASTER.

Patel, V., & Shah, M. (2022). Artificial intelligence and machine learning in drug discovery and development. Intelligent Medicine, 2(3), 134-140.

Pawlak, G. (2022). Intelligence artificielle et machine learning dans les stratégies de drugdiscovery (Doctoral dissertation).

Pawlak, G. M. (2022). Intelligence artificielle et machine learning dans les stratégies de drugdiscovery (Thèse de doctorat, Université de Lille, Faculté de Pharmacie de Lille). Soutenue le 8 décembre 2022.

Singh, R., Rathi, P., & Roy, S. S. (2022, February 17). *A review on applications of artificial intelligence in drug discovery* (arXiv:2202.08320v1). arXiv.

Stroppa, N. (2005). Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles (Doctoral dissertation, Télécom ParisTech).

Suresh, H., & Guttag, J. (2021, October). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).

Unknown author. (2020–2021). Machine Learning [R17A0534] lecture notes [Course notes, B.Tech IV Year – I Sem]. Department of Computer Science and Engineering, Malla Reddy College of Engineering & Technology.

Wang, X., Zhang, H., Li, Y., & Lin, H. (2024, November 11). *Prompting LLMs for molecular generation: A comprehensive benchmark* (arXiv:2411.06009v1). arXiv.

Wang, Y., Zhang, Z., Zhao, C., Hu, H., Liu, B., & Song, X. (2025, February 27). *BioGen: Large language models for biomolecular design* (arXiv:2502.13959v1). arXiv.

Wang, Z., Jiang, T., Wang, J., &Xuan, Q. (2024). Multi-modal representation learning for molecular property prediction: sequence, graph, geometry. arXiv preprint arXiv:2401.03369 (Wang, Jiang, Wang, &Xuan, 2024).

Xue, D., Xie, L., & Chu, X. (2021, June 9).*Review of drug repositioning approaches and resources* (arXiv:2106.05386v1). arXiv.

Xue, D., Xie, L., & Chu, X. (2021, June 9).*Review ofdrug repositioning approaches and resources* (arXiv:2106.05386v1). arXiv.

Zanger, U. M., & Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacology&therapeutics, 138(1), 103-141.

Zanger, U. M., & Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacology & Therapeutics, 138(1), 103–141.