الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université frères Mentouri Constantine 1

Faculté des sciences de la nature et de la vie

جامعة الاخوة منتوري قسنطينة 1

كلية علوم الطبيعة والحياة

قسم الكيمياء الحيوية والبيولوجيا الخلوية والجزيئية

Département de Biochimie et Biologie Cellulaire et Moléculaire

Mémoire en vue de l'obtention du Diplôme de Master

Filière : Sciences Biologiques

Spécialité : Biochimie Appliquée

THÈME

Une approche QSAR basée sur Deep Learning pour la sélection des descripteurs Explicatifs.

Présenté par :

GUENDOUZ MOHAMED

MIRI ROUMEISSA

Soutenu le :

Devant le jury :

Président de jury : Pr. BENSEGUENI Abderrahmane

Encadreur : Dr. BOUKELIA Abdelbasset

Examinateur : MCB. Mokrani El Hassen

Examinateur : Mr. DEMS MOHAMED ABDESSELAM. MRA CRBt Constantine

Année universitaire 2020-2021

REMERCIEMENTS

On tient à exprimer toute notre reconnaissance à notre encadreur de mémoire, Dr. Boukelia Abdelbasset, on le remercie de m'avoir encadré, orienté, aidé et conseillé.

Nos sincères salutations vont également vers le Pr.Bensegueni Abderrahmane et le Dr.Mokrani El Hassen et Mr DEMS Mohamed Abdesselam qui ont accepté de constituer notre jury ainsi qu'à examiner de plus près notre travail.

On adresse nos sincères remerciements à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé nos réflexions et participé à notre formation.

On remercie nos très chers parents, qui ont toujours été là pour nous.

Enfin, on remercie nos frères, sœurs et amis qui ont toujours été là pour nous. Leur soutien inconditionnel et leurs encouragements ont été précieux.

À tous ces intervenants, je présente mes remerciements, mon respect et ma gratitude.

Résumés

Contenu

Résumé

Depuis les découvertes et les progrès de techniques de séquençages haut débit (NGS) et la chromatographie en phase liquide à haute performance (HPLC), les chercheurs se focalisent sur le traitement des cellules tumorales, en utilisant de nouvelles techniques thérapeutiques basées sur les tendances de la découverte des médicaments. Cette dernière peut être décrite comme le processus d'identification des entités chimiques. Les chercheurs se focalisent sur le traitement des cellules tumorales, en utilisant nouvelles techniques thérapeutiques. La relation quantitative structure-activité (QSAR) est un domaine important dans la conception et de la découverte de médicaments, la recherche des renseignements sur la structure chimique des activités biologiques et pharmaceutiques. Cette approche exige de bons descripteurs moléculaires représentatifs des caractéristiques moléculaires responsables de l'activité moléculaire pertinente.

Dans ce travail, nous allons proposer une technique de l'apprentissage approfondi (Deep Learning) basée sur les réseaux de neurones à convolution (CNN) pour construire un modèle de régression QSAR comme une première partie de travail. Ensuite, nous allons intégrer un modèle des ilots généralisés qui est un modèle d'optimisation et de recherche coopérative, afin de trouver un pattern de descripteurs pertinents pour les molécules de NSCLC.

Les résultats expérimentaux obtenus à partir le modèle de régression basé sur CNN sont très prometteuse, avec un coefficient de détermination supérieur à 80,51%. Ainsi, nous avons obtenu plusieurs patterns de descripteurs de l'activité biologique ciblé.

Mots clés : QSAR, Descripteur moléculaire, Deep Learning, Optimisation coopérative.

Abstract

Following discoveries and advances in high-throughput sequencing (NGS) techniques and high-performance liquid chromatography (HPLC), researchers are focusing on the treatment of tumor cells, using new therapeutic techniques based on drug discovery trends. The latter can be described as the process of identifying chemical entities. The researchers focus on the treatment of tumor cells, using new therapeutic techniques. Quantitative structure-activity relationship (QSAR) is an important area in drug design and discovery, seeking information on the chemical structure of biological and pharmaceutical activities. This approach requires good molecular descriptors representative of the molecular characteristics responsible for the relevant molecular activity.

In this work, we will propose a deep learning technique (Deep Learning) based on convolution neural networks (CNN) to build a QSAR regression model as the first part of the work. Next, we will integrate a model of generalized islets that is a model of optimization and cooperative research, to find a pattern of relevant descriptors for NSCLC molecules.

The experimental results obtained from the CNN-based regression model are very promising, with a coefficient of determination greater than 80.51%. Thus, we obtained several patterns of descriptors of the targeted biological activity.

Keywords: QSAR, Molecular Descriptor, Deep Learning, Cooperative Optimization.

ملخص

بعد الاكتشافات والتطورات في تقنيات التسلسل عالي الإنتاجية (NGS) والكروماتوجرافيا السائلة عالية الأداء (HPLC) في أوائل القرن الحادي والعشرين ، يركز الباحثون على علاج الخلايا السرطانية ، باستخدام تقنيات علاجية جديدة تعتمد على اتجاهات اكتشاف الأدوية. يمكن وصف هذا الأخير بأنه عملية تحديد الكيانات الكيميائية. يركز الباحثون على علاج الخلايا السرطانية باستخدام تقنيات علاجية جديدة. تعتبر علاقة التركيب والنشاط الكمي (QSAR) مجالًا مهمًا في تصميم واكتشاف الأدوية ، والبحث عن معلومات حول التركيب الكيميائي للأنشطة البيولوجية والصيدلانية. يتطلب هذا النهج واصفات جزيئية جيدة ممثلة للخصائص الجزيئية المسؤولة عن النشاط الجزيئي ذي الصلة.

في هذا العمل ، سنقترح تقنية التعلم العميق بناءً على الشبكات العصبية الالتفافية (CNN) لبناء نموذج انحدار QSAR كجزء أول من العمل. بعد ذلك ، سنقوم بدمج نموذج للجزر المعممة يمثل نموذجًا للتحسين والبحث التعاوني ، للعثور على نمط من الواصفات ذات الصلة لجزيئات NSCLC.

النتائج التجريبية التي تم الحصول عليها من نموذج الانحدار المعتمد على CNN واعدة للغاية ، مع معامل تحديد أكبر من 80.51٪. وهكذا ، حصلنا على عدة أنماط من واصفات النشاط البيولوجي المستهدف.

الكلمات المفتاحية: QSAR ، الواصف الجزيئي ، التعلم العميق ، التحسين التعاوني.

LISTES DES FIGURES

Sommaire

Chapitre 3 : Contribution

Chapitre 4 : Résultats & Discussion

Conclusion générale

Introduction Générale

Contenu

1. Motivation & Problématique

À la fin du XIXe siècle, le développement de microscopes de meilleure qualité a non seulement permis de documenter et de définir les organismes pathogènes, mais aussi d'examiner les cellules et l'activité cellulaire. L'étude des tissus cancéreux et des tumeurs a révélé que l'apparence des cellules cancéreuses était nettement différente de celle des cellules normales des tissus environnants ou des cellules dont elles provenaient. Les chercheurs ont commencé à se concentrer sur des questions telles que l'origine de ces anomalies.

Au début du XXe siècle, de grands progrès ont été réalisés dans la compréhension des structures, des fonctions et de la chimie des organismes vivants. La recherche sur le cancer dans la culture cellulaire, les cancérogènes chimiques, les techniques de diagnostic et la chimiothérapie a fermement établi l'oncologie comme science. Les anomalies chromosomiques ont également été étudiées comme causes possibles de cancer.

Après les découvertes et les progrès de techniques de séquençages haut débit (NGS) et la chromatographie en phase liquide à haute performance (HPLC) dans le Au début du XXIe siècle, les chercheurs se focalisent sur le traitement des cellules tumorales, en utilisant de nouvelles techniques thérapeutiques basées sur les tendances de la découverte des médicaments.

La découverte de médicaments peut être décrite comme le processus d'identification des entités chimiques qui ont le potentiel de devenir des agents thérapeutiques. L'un des principaux objectifs de la découverte de médicaments ; est de reconnaître les nouvelles entités moléculaires qui peuvent être utiles dans le traitement de maladies en respectant des phases bien défini comme la modélisation des médicaments et les essais cliniques.

La modélisation de médicaments a un but de former des relations qualitatives ou semi-quantitatives entre la structure moléculaire et l'activité. Pour tester ces hypothèses, les chercheurs ont constamment utilisé des outils pharmacologiques traditionnels comme les modèles in vivo et in vitro. De plus en plus au cours de la dernière décennie, des méthodes computationnelles (in silico) ont été développées et appliquées au développement et aux tests d'hypothèses pharmacologiques. Ces méthodes in silico comprennent des bases de données, des relations quantitatives structure-activité, des pharmacophores, des modèles d'homologie et d'autres approches de modélisation moléculaire et en intégrant les techniques de l'apprentissage automatique, l'exploration de données, des outils d'analyse de réseau et des outils d'analyse de données qui utilisent un ordinateur.

La relation quantitative structure-activité (QSAR) est un domaine important dans la conception et de la découverte de médicaments, la recherche des renseignements sur la structure chimique des activités biologiques et

pharmaceutiques. Cette approche exige de bons descripteurs moléculaires représentatifs des caractéristiques moléculaires responsables de l'activité moléculaire pertinente. L'utilité de ces descripteurs dans les études QSAR a été largement démontrée, et ils ont également été utilisés comme mesure de la similitude ou de la diversité structurelle. La sélection de ces descripteurs vise à éliminer ceux qui sont redondants, bruyants ou non pertinents pour les tâches de construction envisagées, de telle sorte que la dimensionnalité de l'espace d'entrée peut être réduite sans perte d'informations importantes. Il est très difficile de sélectionner les descripteurs appropriés pour les analyses QSAR, car il n'y a pas de règles absolues qui régissent cette sélection.

2. Contribution

Dans notre projet de fin d'étude, il était considéré d'utiliser les techniques avancées de l'intelligence artificielle pour l'analyse et l'exploitation des molécules anticancer précisément le cancer de poumon non à petites cellules (Non Small-Cell Lung Cancer (NSCLC)), afin de construire un modèle QSAR performant et sélectionner et interpréter les descripteurs pertinents et importants.

Dans ce travail, nous allons proposer une technique de l'apprentissage approfondi (Deep Learning) basé sur les réseaux de neurones à convolution (CNN) pour construire un modèle de régression QSAR comme une première partie de travail. Ensuite, nous allons intégrer un modèle des ilots généralisés qui est un modèle d'optimisation et de recherche coopérative, afin de trouver un pattern de descripteurs pertinents pour les molécules de NSCLC.

Les résultats expérimentaux obtenus à partir du modèle de régression basé sur CNN sont très prometteuse, avec un coefficient de détermination supérieur à 80,51%. Ainsi, nous avons obtenu plusieurs patterns de descripteurs de l'activité biologique ciblé.

3. Organisation de mémoire

En plus de l'introduction et la conclusion générales, le mémoire est principalement structuré en deux grandes parties. Une partie qui décrit le contexte scientifique et état de l'art, elle regroupe le chapitre 1 et le chapitre 2. Une seconde partie constituée du chapitre 3 et chapitre 4 est consacrée à la description de la contribution et les résultats expérimentaux obtenus.

- **Chapitre 1 :** concerne à la modélisation des médicaments et les différents types des descripteurs et les travaux bio-informatiques pour la sélection des descripteurs.
- **Chapitre 2 :** aborde les différents outils de l'intelligence artificielle nécessaires à la compréhension de l'approche proposée, en commençant par les concepts généraux de l'apprentissage automatique, suivie par l'apprentissage profond.
- **Chapitre 3 :** consiste à notre approche QSAR proposée qui est basée sur Deep Learning pour la sélection des descripteurs explicatifs.
- **Chapitre 4 :** Finalement, dans le quatrième chapitre, nous d´écrivons les méthodes de validation proposées, les logiciels et les bibliothèques et on termine par les résultats et discussion.

Chapitre 1 : QSAR

Contenu

1. Introduction

Considérant qu'un large nombre de substances chimiques influencent notre vie (application industrielle, recherche scientifique, consommation domestique), ça serait intéressant de développer une base qui nous permet d'expliquer leur comportement, une fois cette base est établie, elle nous permettrait de modifier ces comportement en introduisant des changements structuraux raisonnables, de ce fait, l'objectif d'un chimiste sera le développement d'une substance chimique avec la manifestation comportementale désirée, c'est ce qu'on appelle : l'activité. Dans le cas d'une substance chimique avec une activité biologique déclarée, le chimiste tente d'améliorer l'efficacité de la molécule tout en réduisant ses effets toxiques.

Récemment, le développement d'une substance chimique se fait ; soit par la réalisation d'une nouvelle substance ou bien la modification d'une existante. Dans les deux cas, les chimistes ont besoin d'avoir des connaissances suffisantes concernant la nature de la substance et son potentiel d'interaction avec le system biologique. C'est évident que l'activité biologique (y compris la toxicité) d'une substance chimique (substance pharmaceutique, cancérigène et médicaments…etc.) dépend de son interaction avec le système biologique en question, de surcroit, le but essentiel d'un designer de molécules chimiques repose sur l'établissement d'une explication rationnelle du mécanisme d'action de cette molécule, ce qui mène vers la déduction d'une base théorique convenable et ainsi permet l'adaptation de sa structure pour obtenir une réponse biologique optimale.

L'activité biologique obtenue par la substance chimique est attribuée aux différentes interactions de la molécule au niveau du site de réaction du system biologique en question. Le ligand (molécule) est reconnu par un récepteur particulier suivi par la formation du complexe ligand-récepteur impliquant de différentes forces physicochimiques. Ensuite, le complexe subit quelques changements de conformation qui mènent vers une série d'évènements donnant lieu à l'activité biologique.

L'obtention d'une réponse biologique est contrôlée par la façon dont cette molécule chimique réagit au niveau du site actif du système biologique. En d'autres termes, le système biologique joue un rôle crucial dans la détermination des caractéristiques structurales nécessaires pour l'obtention d'une réponse désirée. Avant l'obtention de cette réponse, les médicaments et les molécules bioactives connaissent un chemin complexe à l'intérieur du système biologique gouvernées par leur comportement pharmacocinétiques et pharmacodynamiques. Après l'administration, les molécules d'un médicament sont soumises aux aléas résultant de l'absorption, métabolisme, excrétion et la marche aléatoire (Random walk) en direction du site critique de réaction, ou ils vont adhérer à une molécule réceptrice appropriée. [1]

C'était Ehrlich et Himmelweit qui ont introduit le terme de récepteur ou substance réceptive, et ont envisagé qu'un composé organique (qui peut être un médicament) exerce son action à travers la liaison avec des substances réceptives spécifiques. La notion de ''spécificité du récepteur'' aide au développement d'une base fondamentale pour l'activité des substances chimiques. Les récepteurs sont des sites macromoléculaires avec une structure tridimensionnelle (3D) bien définie. En général, un récepteur peut faire référence aux différents sites de reconnaissance de l'action du médicament et ça inclut des différentes enzymes. Certaines molécules chimiques ont une affinité pour un type de récepteurs donné, d'autres non, en plus de l'affinité, un autre facteur est à prendre en considération, il s'agit du nombre de récepteurs qui joue un rôle essentiel à l'obtention d'une réponse biologique. [1]

Les 2 fonctions majeures des récepteurs sont (1) la reconnaissance spécifique du ligand et (2) la transduction du signal en réponse, par conséquence, les récepteurs sont possédés par un domaine de fixation pour le ligand et un domaine effecteur pour permettre les changements de conformation requis. Maintenant, la question qui se pose est : comment le ligand se lie-t-ils à un récepteur spécifique, la réponse réside dans la structure chimique de ce ligand. C'est évident que pour poursuivre une stratégie rationnelle dans le design d'un ligand biologiquement actif, le designer doit avoir des connaissances suffisantes en regard de l'interaction de ce ligand avec le récepteur cible (s'il est disponible) et autres caractéristiques structurales. Les études de modélisation prédictive ont pour but d'explorer tous les ces caractéristiques qui affectent l'activité du ligand. [1]

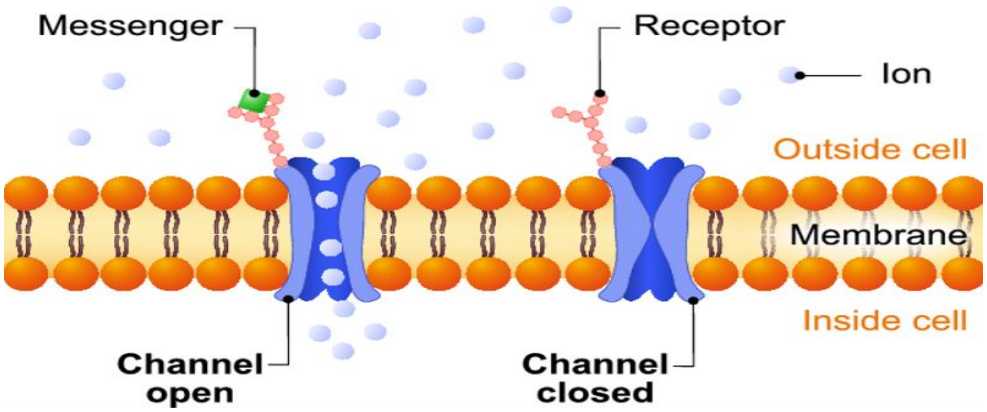


Figure 1. 1 Récepteur ionotrope [2]

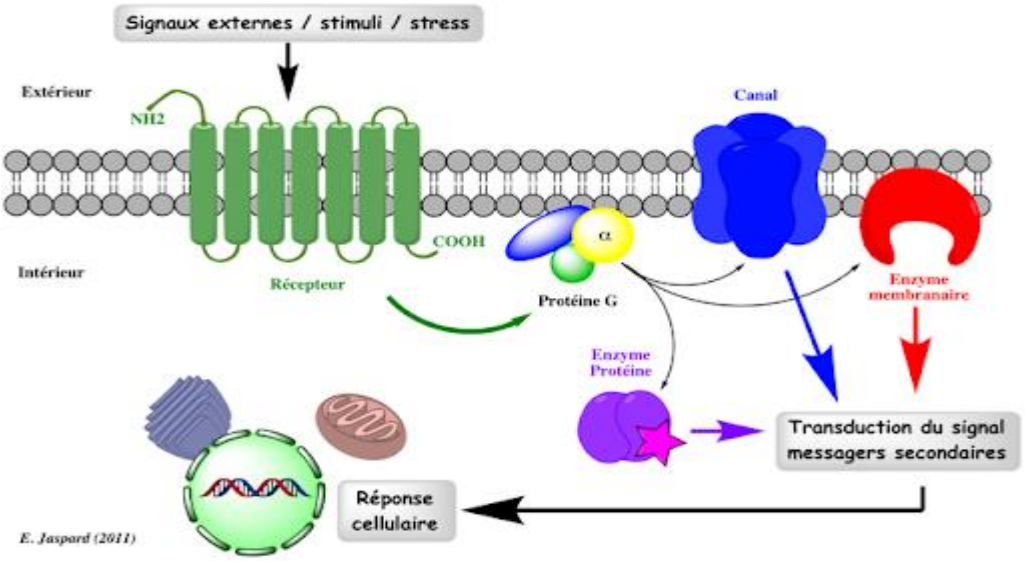


Figure 1. 2: Récepteur couplé à la protéine G [3]

2. Modélisation des médicaments

La modélisation moléculaire décrit la génération, manipulation ou la représentation des structures 3D des molécules ainsi que les propriétés physico-chimique associées. Elle inclut une série de techniques informatisées basées sur des méthodes chimiques théoriques

et des données expérimentales afin de prédire les propriétés moléculaires et biologiques. La modélisation est un outil pour la chimie, elle permet une meilleure compréhension de la chimie en fournissant de meilleurs outils pour investiguer, interpréter, expliquer et découvrir de nouveau phénomènes. La modélisation moléculaire est facile à réaliser avec les logiciels actuels mais la difficulté réside dans l'obtention du bon model et de la bonne interprétation. Actuellement, deux stratégies majeures de modélisation sont utilisées pour la conception de nouveaux médicaments. [4] Elles sont :

- **La conception directe :** Dans l'approche directe, les caractéristiques tridimensionnelles d'un site récepteur connu sont déterminées à partir de la cristallographie aux rayons X pour la conception de la molécule finale (a lead molecule). Dans la conception directe, la géométrie du site récepteur est connue, le problème est de trouver une molécule qui satisfait certaines contraintes géométriques et qui a une bonne correspondance chimique. Après avoir trouvé de bons molécules-candidats selon ces critères, une étape de Docking avec une minimisation d'énergie peut être pratiquée pour prédire la force de liaison. [4]
- **La conception indirecte** : L'approche indirecte inclut l'analyse comparative des caractéristiques structurales des molécules actives et inactives connues et qui sont complémentaires avec un site récepteur hypothétique. Si la géométrie du site est inconnue, comme c'est souvent le cas, le designer doit fonder sa conception sur d'autres molécules ligands qui se lient bien avec le site récepteur. [4]

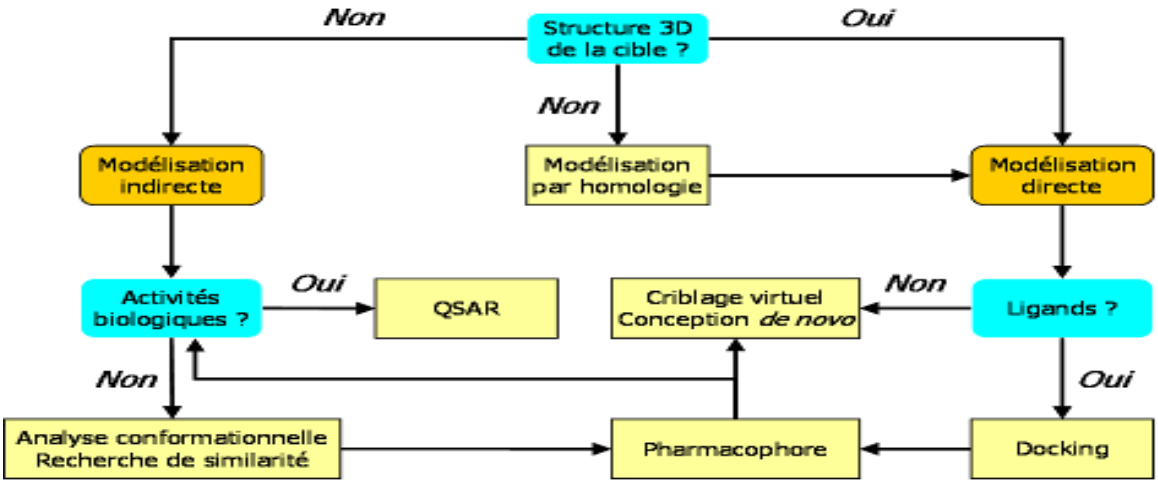


Figure 1. 3 : modélisation moléculaire

2.1. L’application de la modélisation moléculaire

Le point de départ de beaucoup d’études de modélisation moléculaire assistée par un ordinateur est en général un dessin bidimensionnel d’une molécule requise, ensuite, les structures bidimensionnelles (2D) sont transformées en représentations tridimensionnelles (3D) pour étudier les propriétés chimiques. Voici quelques applications de la modélisation moléculaire assistée par un ordinateur :

2.1.1. La génération des structures chimiques

Les structures moléculaires peuvent être générées à l’aide de plusieurs logiciels. Les structures moléculaire 3D peuvent être crées par plusieurs fonctions communes de construction, comme : make-bond, break-bond, fuserings, delete-atom, add-atom-hydrogens, invest chiral center ...etc. la modélisation informatique permet aux chimistes de construire des modèles dynamiques de composés qui leur permettent à leur tour la visualisation de la géométrie moléculaire et la démonstration des principes chimiques. [4]

2.1.2. La visualisation de la structure moléculaire

Le concept le plus important dans la modélisation moléculaire est la visualisation des intéractions et les structures moléculaires. Les molécules sont visualisées en 3D par différentes représentations comme : les bâtonnets connectés, les modèles boules-bâtonnets et les représentations compactes. [4]

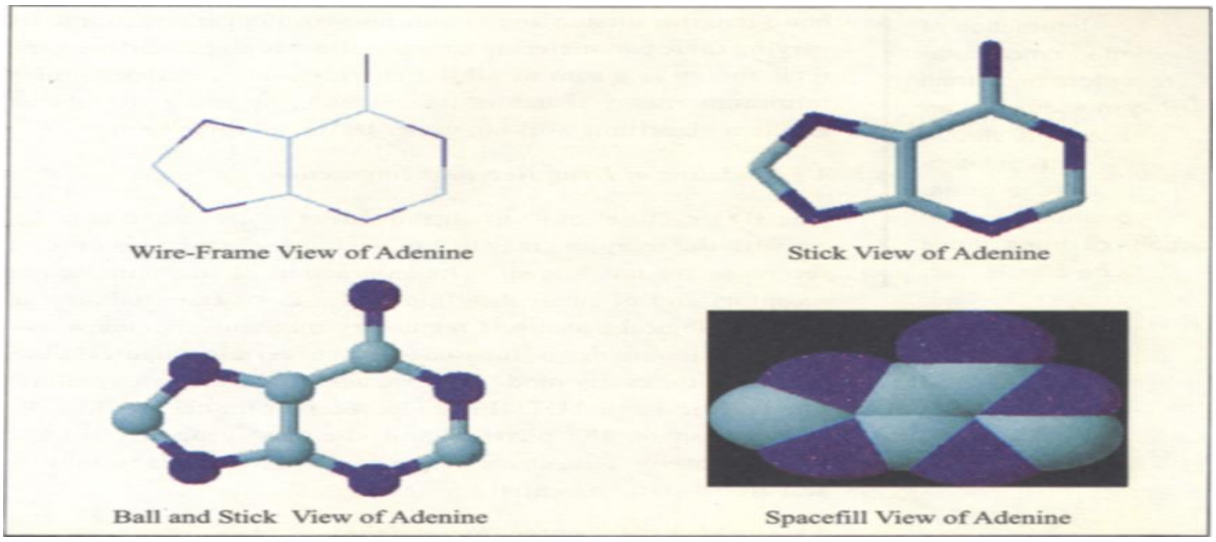


Figure 1. 4 : Vu d'adénine sous différentes représentations [4]

2.1.3. La génération des conformations

La plupart des molécules existent sous différentes conformations. La conformation préférée d'une molécule est la caractéristique structurale qui survient comme une réponse à la force d'attraction et de répulsion. La forme doit être considérée principalement dans la détermination de l'intéractions de la molécule avec le récepteur. [4]

2.1.4. La modélisation des intéractions médicament-récepteur

Les structures 3D des molécules d'un médicament qui interagissent avec les récepteurs peuvent être connues mais la structure de la plupart des récepteurs est inconnue. L'interaction des récepteurs macromoléculaires et les petites molécules du médicament est une étape essentielle dans beaucoup de processus biologiques : mécanismes régulatrices, actions pharmacologiques des médicaments, les effets toxiques de certains composés chimiques, etc. la cavité du récepteur est construite à l'aide de programmes comme : RECEPS et AUTOFIT. Le modèle du récepteur fournit une information 3D sur les propriétés physiques et chimiques de la cavité du récepteur, sa taille, sa forme, la prévisibilité des liaisons hydrogènes et le potentiel électrostatique. [4]

2.1.5. Docking

La modélisation de l'interaction du médicament avec son récepteur est un problème complexe. Plusieurs forces sont inclus à l'association intermoléculaire : dispersion hydrophobe ou van der Waals, liaison hydrogène et électrostatique. La force majeure de la liaison

semble être l'interaction hydrophobique, mais la spécificité de la liaison parait être contrôler par la liaison hydrogène et les intéractions électrostatiques. La modélisation des intéractions du complexe ligand-protéine est difficile parce qu'il y a plusieurs degrés de liberté et des connaissances limitées de l'effet du solvant sur la constante de liaison.[4]

2.1.6. La détermination des propriétés moléculaires

Les propriétés moléculaires sont des indicateurs importants d'une variété de molécules chimiques, y compris les médicaments. Les propriétés moléculaires sont normalement classifiées comme : physiques, chimiques et biologiques. Les 3 méthodes informatiques majeures pour le calcul des propriétés moléculaires sont :

Empirique (mécanique moléculaire)

Les méthodes de mécanique moléculaire sont moins compliquées, rapide et sont capable de gérer de très grandes entités biologiques, y compris les enzymes. La mécanique moléculaire est un formalisme qui a pour but la reproduction des géométries moléculaires, les énergies et autres caractéristiques en ajustant la longueur des liaisons, leurs angles et les angles de torsions aux valeurs d'équilibre qui dépendent de l'hybridation d'un atome et de son schéma de liaison. [4]

Dynamique moléculaire

Lorsqu'elle est combinée avec les données issues des études de Résonance magnétique nucléaire (RMN) , elle est utilisée pour dériver les structures 3D des peptides et petites protéines dans les cas où la cristallographie aux rayons X est impraticable. De surcroit, les données structurelles, dynamiques et thermodynamique de la dynamique moléculaire ont apporté des renseignements à la relation structure-fonction, les affinités de liaison, la stabilité et la mobilité des protéines, acides nucléiques et autres macromolécules qui ne peuvent pas être obtenus à partir des modèles statiques. [4]

Mécanique quantique

Dans sa forme la plus pure, la théorie du quantum utilise des constantes physiques connues, comme : la vitesse de la lumière, les valeurs de masses et charges des particules nucléaires et les équations différentielles afin de calculer directement les propriétés moléculaires et géométriques. Ce formalisme est appelé : ab initio en référence aux premiers principes dont la mécanique quantique en fait partie. En

général, les méthodes ab initio sont capables de reproduire les mesures de laboratoire des propriétés telle que : potentiel d'ionisation, UV et spectre visible et géométrie moléculaire.[4]

2.2. Criblage virtuel (Virtuel screening)

L'étude QSAR permet le criblage des chimiothèques qui contiennent un large nombre de composés. L'information mathématique obtenue à partir d'une analyse QSAR peut être employé comme un critère de criblage des molécules d'intérêt, le criblage virtuel peut être réalisé en employant l'information du ligand aussi bien que la structure du récepteur. La méthodologie QSAR peut être employée dans la recherche des substances chimiques en se basant sur le ligand (ligand-based searches), deux aspects de cette opération peuvent être identifiés : (1) le criblage des composés qui ont des caractéristiques structurales similaires au model développé, et (2) le criblage de la base des données chimiques des composés dont la réponse est connue (exemple : la base de données des anticancéreux) pour prédire et vérifier la valeur de la réponse en utilisant le model. Il existe plusieurs chimiotech qui contiennent une large réserve de composés, notamment : ZINC, DUD benchmark, PUBChem, ChemBank.[1][6]

2.3. QSAR/QSPR

La notion SAR (Structure-Activity Relationship) (relation entre la structure et l'activité) tente à établir une relation entre les différentes caractéristiques comportementales des composés chimiques (exemple : activité et toxicité) et les informations issues de leurs structures chimiques, en d'autres termes, l'étude SAR offre la possibilité d'établir une équation mettant en jeu l'activité/propriété/toxicité spécifique des produits chimiques en utilisant des informations sur leurs structures chimiques, du coup l'expression quantitative de l'activité d'un produit chimique définit l'étude QSAR. L'axiome central de la modélisation QSAR repose sur la présentation de la réponse chimique en termes de propriétés moléculaires, sachant que chaque propriété contenant une information chimique significative peut être employer comme descripteur. Une fois l'équation est établie, la méthode QSAR nous permet la prédiction de l'activité du produit chimique étudié, en outre, la méthode QSAR met l'accent sur la modification de la structure chimique pour obtenir les produits chimiques d'intérêt avec les valeurs de réponse désirées. L'appellation est influencée par le point final modélisé, par conséquence le processus peut être défini comme suit : QSAR/QSPR/QSTR pour activité/propriété/toxicité.

L'équation mathématique : Réponse = f (propriétés structurale/chimiques) [1]

L'analyse QSAR a été créé afin de remplir les objectifs suivants : (a) la prédiction de nouveaux analogues avec une meilleure activité, (b) améliorer la compréhension et l'investigation du mode d'action de produits chimiques et pharmaceutiques, (c) l'optimisation de la molécule type en congénères moins toxiques, (d) la rationalisation des expérimentations humides (QSAR offre une alternative économique et rapide aux essais in vitro à débit moyen ainsi qu'aux essais in vivo à faible débit) [1]

2.4. ADMET

ADMET représente : Absorption, Distribution, Métabolisme, Excrétion et Toxicité. La prédiction des propriétés ADMET joue un rôle important dans le processus de la conception du médicament parce que ces propriétés comptent pour l'échec d'environ 60/100 des médicaments pendant les phases cliniques. Pendant que traditionnellement les outils ADME ont été appliqués à la fin du pipeline du développement des médicaments, actuellement ADME est appliqué dans une phase précoce du processus du développement des médicaments, dans le but d'éliminer les molécules dont les propriétés ADME sont mauvaises et permettre d'importantes économies en termes de couts de recherches et de développement.

Les propriétés physicochimiques d'un médicament ont un impact très important sur son avenir pharmacocinétique et métabolique à l'intérieur de l'organisme. Donc une bonne compréhension de ces propriétés, leur calcul et leur prédiction sont cruciaux dans le succès d'un programme de développement de médicaments. [7]

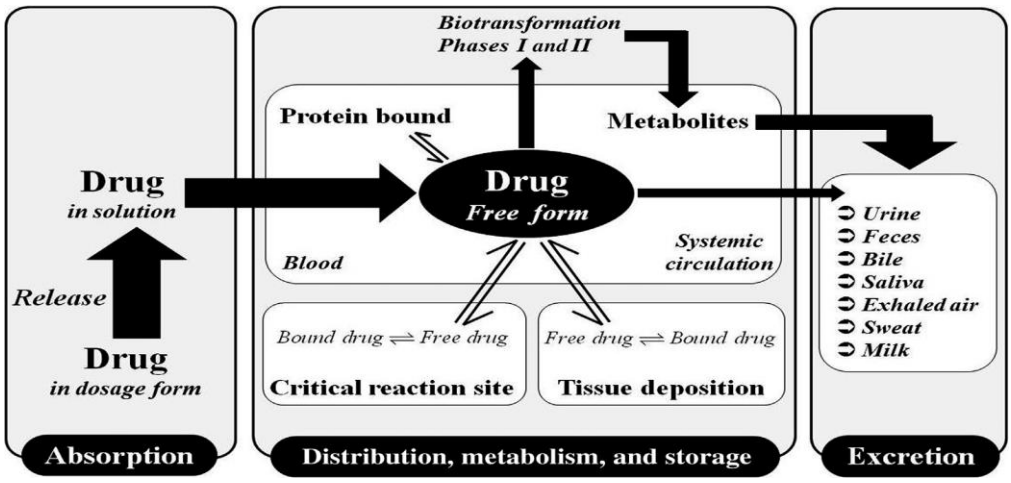


Figure 1. 5 : Les différentes étapes pharmacocinétiques des ligans bioactifs au sein d'un vivant [1]

3. Sélection des descripteurs

L'utilisation des descripteurs pour définir la structure moléculaire des composés biologiquement actifs est la méthode principale pour découvrir de nouvelles molécules.

Les descripteurs moléculaires représentent les caractéristiques les plus signifiants à la modélisation QSAR/QSPR. L'information codée par les descripteurs dépend généralement du type de représentation moléculaire et l'algorithme définie pour leur calcul.

Parmi ces descripteurs, on a les indices topologiques, les descripteurs géométriques, constitutionnels et physicochimiques.

3.1. Descripteurs

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

Les descripteurs moléculaires sont les traits communs les plus considérables de structure moléculaire qui peut être utilisée pour développer la « Relation Structure – Activité » avec le but de prédire l'activité biologique et propriétés physico-chimique des molécules.[1] [8]

3.2. Types de descripteurs

Il y a beaucoup de descripteurs moléculaires qui ont été répertoriés, qu'ils soient dérivés de la théorie ou tirés des approches différentes.

3.2.1. Descripteurs constitutionnels

Les descripteurs constitutionnels sont directement liés à la formule brute de la molécule, à l'aide de la composition moléculaire, c'est-à-dire les atomes qui le constituent, Il s'agit de : La masse molaire.

Les nombres absolus et relatifs d'atomes (C, H, O, S, N, F, Cl, Br, I, P.…).

Les nombres absolus et relatifs de groupes fonctionnels (NH2, COOH, OH. . .).

Les nombres absolus et relatifs de liaisons (simples, doubles, aromatiques. . .).

Les nombres absolus et relatifs de cycles (aromatiques ou non).

Ces descripteurs sont très utilisés du fait de leur extrême simplicité non seulement d'un point de vue conceptuel mais surtout calculatoire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution. C.-à-d., si on développe des modèles avec ce type de descripteurs seulement, ils peuvent poser problème pour l'interprétation des mécanismes d'interaction mis en jeu pour la propriété étudiée.[9]

3.2.2. Descripteurs topologiques

Les descripteurs topologiques "ou indices topologiques", décrivent la connectivité atomique dans la molécule, ils sont obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Ces descripteurs s'inspirent de la théorie des graphes appliquée à la table de connectivité qui n'est autre qu'une représentation compacte de la connectivité interatomique au sein de la molécule. Les indices topologiques les plus fréquemment utilisés sont l'indice de Wiener, l'indice de Randic, l'indice de connectivité de valence de Kier-Hall et l'indice de Balaban. [9]

3.2.3. Descripteurs géométriques

Les descripteurs géométriques d'une molécule sont évalués à partir des positions relatives de ses atomes dans l'espace, et décrivent des caractéristiques plus complexes ; leurs calculs nécessitent de connaître, la géométrie 3D de la molécule, par modélisation moléculaire empirique ou ab initio, Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés qui dépendent de la structure 3D. On distingue plusieurs descripteurs importants, le volume moléculaire, la surface accessible au solvant, le moment d'inertie.

Le volume moléculaire est le volume occupé par la molécule en appliquant une grille

3D de cubes dans la boîte parallélépipédique dont les dimensions X_{max}, Y_{max} et Z_{max}. La surface accessible au solvant SAS, ou la zone de surface accessible est la surface d'une molécule qui est accessible à un solvant, généralement mesuré en unités d'angströms carrés. Le moment d'inertie est une grandeur physique qui caractérise la distribution de masse dans la molécule. [9]

3.2.4. Descripteurs électrostatiques

Ces descripteurs reflètent les caractéristiques de la distribution de charge de la molécule. Les charges partielles empiriques dans la molécule sont calculées en utilisant l'approche proposée par Zefirov. Cette méthode est basée sur l'échelle d'électronégativité. Sur la base de ces charges partielles les descripteurs électrostatiques suivantes sont calculés comme suivants :

- Les charges partielles minimales et maximales dans la molécule (qmin, qmax).
- Les charges partielles minimales et maximales pour l'atome (C, N, O.…).
- Les indices électroniques topologiques.

Ces descripteurs sont responsables sur des interactions entre les molécules polaires. [9]

3.2.5. Descripteurs thermodynamiques

Les descripteurs thermodynamiques sont calculés sur la base de la fonction de partition totale Q de la molécule. La fonction de partition

commode la façon avec laquelle l'énergie d'un système de molécules est répartie parmi les individus moléculaires. Sa valeur dépend du poids moléculaire, de la température, du volume moléculaire, des distances inter nucléaires, des mouvements moléculaires et des forces intermoléculaires. La fonction de partition est le point le plus commode entre les propriétés microscopiques des molécules individuelles (niveaux d'énergie, moments d'inertie) avec les propriétés macroscopiques (chaleur spécifique, entropie). La molécule peut accroître son énergie de translation, de vibration, de rotation de façon pratiquement indépendante. [9]

Figure 1. 6: Représentation des descripteurs moléculaires utilisés à la modélisation QSAR [8]

4. Travaux bioinformatique pour la sélection des descripteurs

Dans cette section nous allons présenter les principales approches pour la sélection de descripteurs.

4.1. Les méthodes de filtres

Les méthodes filtres sont les méthodes les plus simples pour la sélection des caractéristiques, ces approches utilisent les données d'entrainement (Training Data) afin de sélectionner les caractéristiques

sans appliquer des algorithmes ou les techniques de l'apprentissage automatique.[10]

4.2. Les méthodes de cohérence

Ce sont des méthodes basées sur la robustesse des données d'entrainement et évaluent essentiellement la cohérence de l'ensemble des caractéristiques sélectionnées. Même si cette procédure est simple permettant de réaliser de petit sous-ensemble, elle a des inconvénients, en fait, elle s'applique uniquement avec des caractéristiques discrètes (discontinues), et si le sous-ensemble consiste de caractéristiques continues, elles doivent être discrétisées en 1er lieu. [10]

4.3. Méthodes d'information

Ceux-ci sont des méthodes qui comparent principalement l'information obtenue par la nouvelle caractéristique par rapport à la précédente. [10]

4.4. Les méthodes de dépendance

Les méthodes de dépendance évaluent comment la valeur d'une variable peut être prédite utilisant une valeur d'une autre variable, dans ce cas, la méthode sélectionne la caractéristique qui corrèle le plus avec la classe cible sélectionnée. [10]

4.5. Les méthodes de la distance

Elles constituent une grande classe des Méthodes FS, d'un point de vue général, elles utilisent des distances conventionnelles pour mesurer la similarité entre deux échantillons. [10]

4.6. La sélection en avant (Forward sélection)

C'est une méthode très largement utilisée pour FS. C'est un type spécifique de la régression pas à pas qui commence par un sous ensemble de variables vides et ajoute des caractéristiques une par une à chaque étape. La caractéristique sélectionnée est celle qui permet une meilleure amélioration du model. Cette procédure continue jusqu'à ce qu'il n'y ait aucune caractéristique capable d'améliorer le model. L'inconvénient principal de cette méthode c'est qu'elle tend vers un sur-apprentissage grâce auquel il est important d'avoir un strict critère d'arrêt. [10]

4.7. Élimination en arrière

Cette méthode fonctionne d'une manière opposée à la méthode de la sélection en avant (Forward selection), en fait, elle commence en incluant toutes les caractéristiques, puis elle élimine une par une à chaque étape tout en évaluant la contribution de la caractéristique à l'amélioration du model. Cette procédure n'est pas très utilisée en raison de la production de modèles surchargés.[10]

4.8. La sélection progressive

Celle-ci est probablement la plus utilisée pour FS dans QSAR. C'est une méthode hybride basée sur les deux méthodes précédentes (sélection en avant et élimination en arrière). L'avantage majeur de cette méthode c'est que la variable qui rentre dans le model peut être supprimer après si elle s'est avérée impertinente. En vérité, le processus commence par l'ajout de la variable la plus en corrélation avec le point terminal. à chaque étape, la variable avec la plus de corrélation est ajouté jusqu'à ce qu'il n'y aura plus de variables significatives parmi l'ensemble des caractéristiques.[10]

5. Conclusion

Dans ce chapitre nous avons présenté les principaux concepts de base de modélisation de médicaments. Nous avons focalisé sur les principaux concepts comme QSAR, les descripteurs. Nous avons cité aussi quelques travaux dans le domaine de sélection de descripteurs.

Dans le chapitre prochain, nous allons présenter les principaux concepts de l'intelligence artificielle.

Chapitre 2 : Intelligence Artificielle

Contenu

1 Introduction

L'intelligence artificielle (IA) est la discipline scientifique qui s'intéresse à la création des programmes informatiques qui ont la simple fonction d'effectuer des opérations comparables à celles de l'esprit humain (l'apprentissage ou le raisonnement logique), c'est un domaine scientifique qui cherche à résoudre des problèmes logique, algorithmique est plus généralement constituer des diapositifs imitant ou remplacent l'être humain.

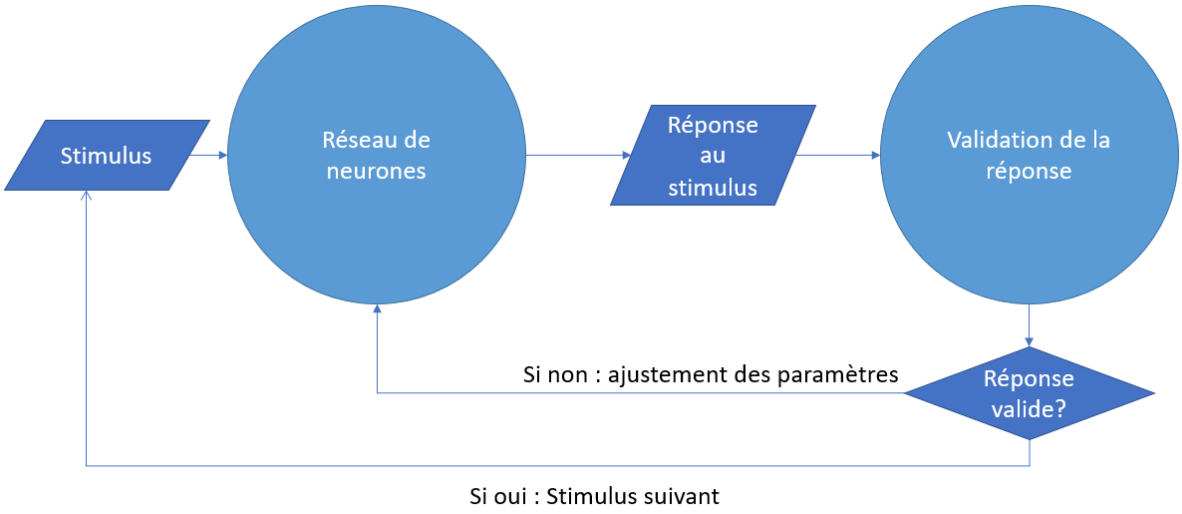


Figure 2. 1: mécanisme d'apprentissage d'un réseau de neurone. [11]

Il y'a deux approches d'IA appliquées au domaine de la santé :

La première est plutôt "agnostique" et tend à appliquer des algorithmes afin de faire apparaître des inférences entre les données, en se concentrant sur l'efficacité des résultats obtenus et la valeur médicale de biomarqueurs de prévention.

L'autre approche est davantage guidée par un effort de modélisation à l'échelle cellulaire, tissulaire, voire organique, afin de guider, de "superviser".

Dans ce chapitre on va découvrir les principes de base de l'intelligence computationnelle (IC), les mécanismes d'IC pour les problèmes d'optimisation et le Deep Learning.

2 Principaux concepts

Avant d'entamer les méthodes de l'intelligence computationnelle (IC) on va aborder d'abord les notions de bases utilisées dans le cadre de notre travail.

2.1 Définition de l'IC

L'intelligence computationnelle est un ensemble des méthodes et d'approches de calcul inspirées de la nature pour traiter des problèmes complexes du monde réel auxquels la modélisation mathématique ou traditionnelle peut s'avérer inutile pour plusieurs raisons ; les processus peuvent être trop complexes pour le raisonnement mathématique.

L'expression intelligence computationnelle (IC) désigne généralement la capacité d'un ordinateur à apprendre une tâche spécifique à partir de données ou d'observations expérimentales. [12]

2.2 Algorithme d'optimisation et de recherche

Un algorithme d'optimisation est une procédure mathématique qui permet d'obtenir les minimums (ou maximums)[1] d'une fonction réel f (que l'on appelle fonction objective)

Min $x \in \Re^n$ f(x)

Les algorithmes d'optimisations ont besoin en général des dérivées de premier et deuxième degré de la fonction Pour le calcul du gradient d'une fonction, on peut utiliser la dérivation directe, approximation par différences finies…Par exemple, la méthode de

descente de gradient a besoin juste des 1eres dérivées; la méthode de Newton nécessite les 2èmes dérivées de la fonction objective ; sans dérivée, on peut trouver les méthodes d'algorithme du 'simplexe' , 'simulated annealing' , 'neural networks' et 'algorithmes génétiques'.[13]

2.3 Méta-heuristique

On parle de *méta*, du grec « au-delà » (comprendre ici « à un plus haut niveau »), *heuristique*, du grec ευρισκειν / *heuriskein*, qui signifie « trouver ». En effet, ces algorithmes se veulent des méthodes génériques pouvant optimiser une large gamme de problèmes différents, sans nécessiter de changements profonds dans l'algorithme employé.

Une métaheuristique est un algorithme d'optimisation visant à résoudre des problèmes d'optimisation difficile (souvent issus des domaines de la recherche opérationnelle, de l'ingénierie ou de l'intelligence artificielle) pour lesquels on ne connaît pas de méthode classique plus efficace. Ils progressent vers un optimum global, c'est-à-dire l'extremum global d'une fonction, par échantillonnage d'une fonction objective. Elles se comportent comme des algorithmes de recherche, tentant d'apprendre les caractéristiques d'un problème afin d'en trouver une approximation de la meilleure solution (d'une manière proche des algorithmes d'approximation). [14]

Blum et Roli ont décrit neuf propriétés de métaheuristiques [15] [16] :

- Les métaheuristiques sont des stratégies qui guident le processus de recherche.
- L'objectif est d'explorer efficacement l'espace de recherche afin de trouver des solutions quasi optimales.
- Les métaheuristiques font des procédures de recherche locales simples aux processus d'apprentissage complexes.
- Les métaheuristiques sont approximatives et généralement non déterministe.
- Les concepts de bases de la métaheuristiques permettent une description abstraite.
- Les métaheuristiques peuvent incorporer des mécanismes pour éviter de se retrouver piégées dans des zones de l'espace de recherche.
- Les métaheuristiques ne sont pas spécifiques à un seul problème.

- Les métaheuristiques peuvent utiliser des connaissances spécifiques à un domaine sous forme d'heuristique contrôlée par une stratégie de niveau supérieur.
- Aujourd'hui les métaheuristiques les plus avancées utilisent l'expérience de recherche (incorporée dans une forme de mémoire) pour guider la recherche.

3 Apprentissage automatique

L'apprentissage automatique en anglais machine Learning, ou l'apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fond sur des approches statistiques afin de donner aux ordinateurs la capacité d'apprendre à partir d'un ensemble de données. C'est-à-dire d'améliorer leurs performances pour résoudre des taches sans être explicitement programmés par un expert. [17]

On a 4 méthodes de l'apprentissage automatique : le supervisé, non supervisé, semi supervisé, par renforcement.

3.1 Apprentissage automatique supervisé

L'apprentissage supervisé utilise certaines variables pour prédire des valeurs inconnues dans le futur d'autres variables par rapport à la variable cible qui se divise en deux types, la régression et la classification. [17]

- **la classification :** cela fait référence à la possibilité de classer quelque chose dans un ensemble distinct de classe ou de catégories.[15]
- **La régression :** cela fait référence à la capacité de prédire les valeurs d'une variable continue. Par exemple, un modèle qui permet de compter les cellules dans une image microscopique.[15]

3.2 L'apprentissage non supervisé

L'apprentissage non supervisé sert à trouver des formes interprétantes par un algorithme permettant de regrouper les données sans utiliser de variables cible a priori. Le clustering est un type de l'apprentissage non supervisé qui se base sur une mesure de similarité, qui est généralement mesurée en termes de distances. [17]

3.3 L’apprentissage automatique semi supervisé

L'apprentissage semi supervisé effectue de manière probabiliste ou non, il vise à faire apparaitre la distribution sous-jacente dans leur espace de description. Il est mis en œuvre quand des données manquent, Le modèle doit utiliser des exemples non étiquetés pouvant néanmoins renseigner. [17]

3.4 L’apprentissage automatique par renforcement

L’apprentissage par renforcement équivaut à apprendre à jouer à un jeu. Les règles et les objectifs sont clairement définis. Cependant, le résultat de chaque jeu dépend du jugement du joueur qui doit ajuster son approche en fonction de l’environnement, des compétences et des actions du candidat sortant.

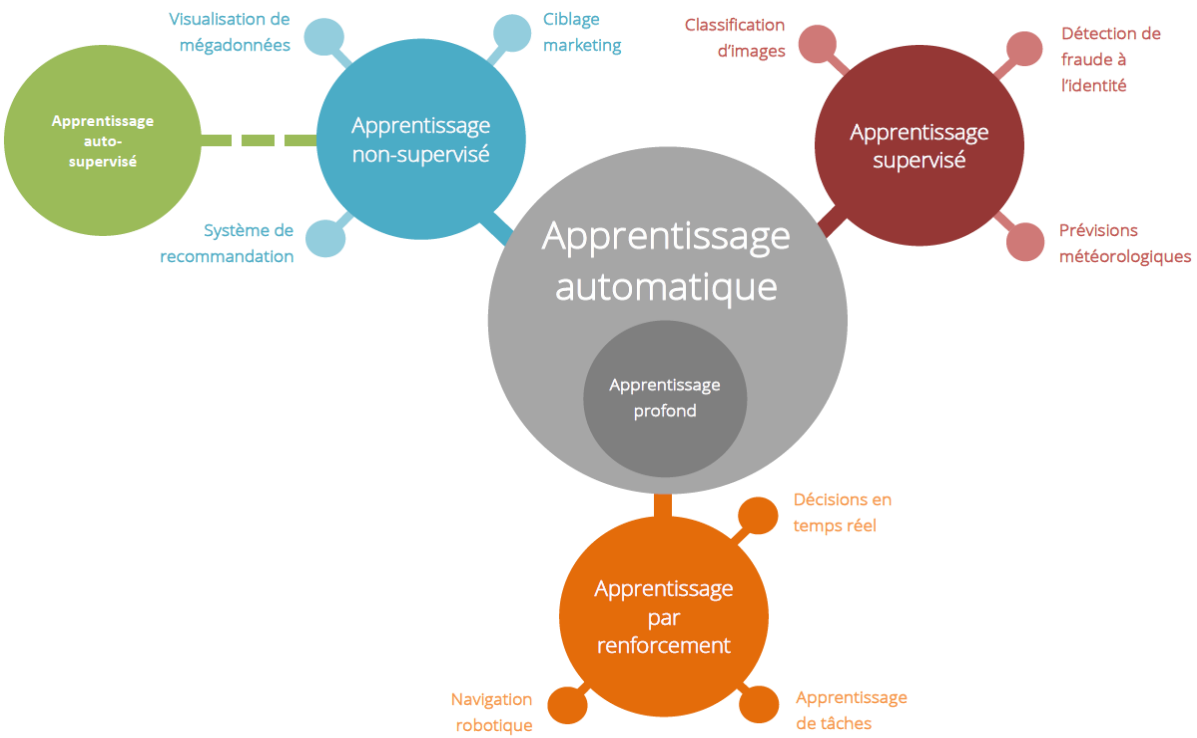


Figure 2. 2 : les différents types de l'apprentissage automatique

4 Méthodes IC pour les problèmes d'optimisation et de recherche

Il s'agit d'un outil indispensable pour de nombreux praticiens dans divers domaines. De nouvelles techniques d'intelligence computationnelle, telle que les Algorithmes stochastiques, les algorithmes évolutionnaire, l'intelligence par essaims et le modèle de ilots se sont relevées efficaces pour résoudre les problèmes d'optimisation globaux.

4.1 Algorithmes d'optimisation stochastique

Les méthodes d'optimisation stochastique s'appuient sur des mécanismes de transition probabiliste et aléatoire. Cette caractéristique indique que plusieurs exécutions successives de ces méthodes peuvent conduire à des résultats différents pour une même configuration initiale d'un problème d'optimisation. [19]

4.2 Algorithmes évolutionnaires

Ces algorithmes portent sur les systèmes inspirés de la sélection naturelle et sont appliqués dans la recherche et l'optimisation. Parmi les algorithmes évolutionnaires que nous utilisons le plus, on peut citer : Les algorithmes génétiques (AG) et Évolution différentielle (ED). [15]

Algorithme génétique (AG)

C'est un métaheuristique bioinspiré et évolutionnaire dont le but est de trouver des solutions aux problèmes difficiles, elle fait partie des méthodes d'optimisation globale.

L'idée des AG est inspirée des chaines d'ADN, qui composent tout organisme vivant. Les étapes menées dans ce type d'algorithme sont :

- La génération d'une population initiale aléatoire contenant une solution individuelle appelé : Chromosome.
- Évoluer cette population en utilisant des répétition, nommées : générations, et durant cette étape qu'on évalue chaque chromosome par la mesure de la condition objective, appelé : l'adéquation.

- La création d'une autre génération : de nouveaux chromosomes sont formés en fusionnant les chromosomes de la génération actuelle en les croisant (imitation du phénomène du Crossover) ou en modifiant le chromosome.
- La sélection de la nouvelle génération (pour maintenir la taille de la population constante).
- Les chromosomes ajustés ont une grande probabilité d'être sélectionnés et après plusieurs générations générées, l'algorithme converge vers le meilleur chromosome qui constitue la solution optimale ou sous-optimale du problème.[15]

L'évolution différentielle

Ce type d'algorithme est basé sur le maintien d'une population de solutions candidates soumises à des itérations de recombinaison, d'évaluation et de sélection. La recombinaison : implique la création de nouveaux composants de solutions candidates.[15]

4.3 Intelligence par essaims

C'est un paradigme qui considère l'intelligence collective comme un comportement qui émerge de l'interaction et de la coopération d'un grand nombre d'agents moins intelligents.

Il renferme 2 catégories :

- Optimisation de la colonie de fourmis (algorithmes probabilistes inspirés du comportement de recherche de fourmis).
- Optimisation de l'essaim de particules (algorithmes probabilistes inspirés par le comportement des oiseaux et de poissons).

Les techniques basées sur l'intelligence en essaims sont des stratégies adaptatives et sont utilisées dans la recherche et l'optimisation. [15]

5 L'apprentissage en profondeur (Deep Learning)

C'est une technique d'apprentissage automatique moderne qui utilise des réseaux de neurones extrêmement avancés et qui permettent de générer des modèles plus complexes et plus nettes. Il faut noter aussi que les modèles d'apprentissage en profondeur ont une grande capacité d'absorption des données, le Deep Learning est appliqué dans plusieurs

domaines, on peut citer : le traitement des images, la traduction du langage…etc. [15]

5.1 Notions fondamentales

Il existe des concepts fondamentaux utilisés dans les algorithmes de Deep Learning, nous allons citer :

Réseau de neurones et perceptron multicouche

Un modèle motivé par le fonctionnement du cerveau et qui est utilisé dans une grande variété de domaines tel que la reconnaissance de l'écriture manuscrite et la détection des visages. [15]

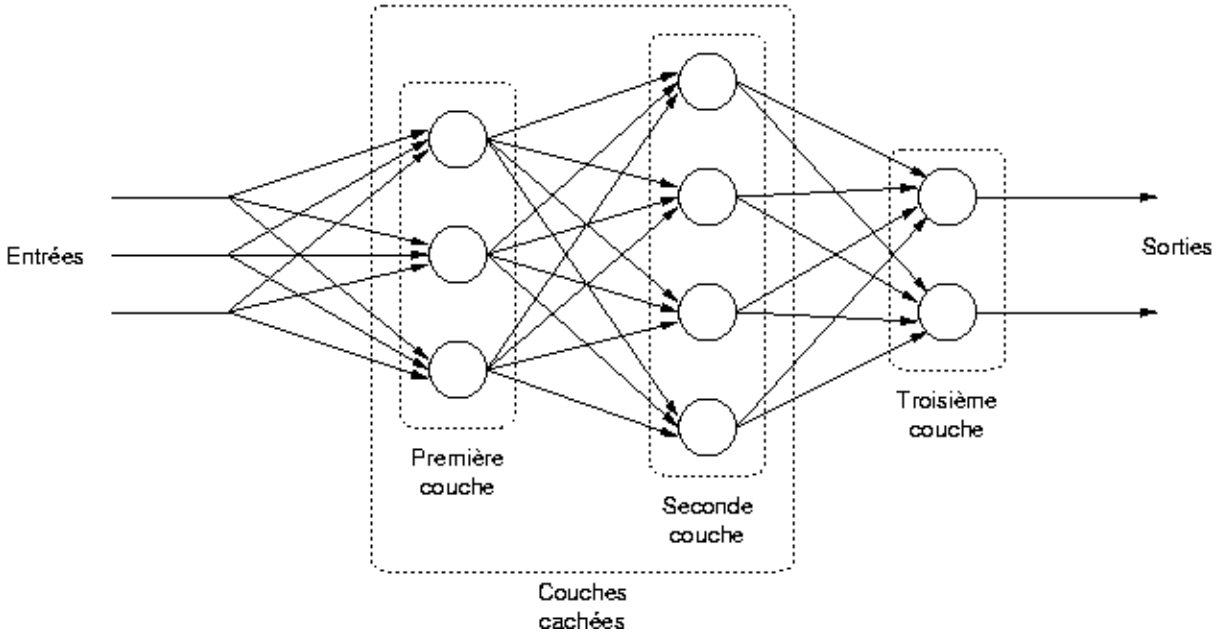


Figure 2. 3Réseau de neurones multicouche[20]

Convolution

C'est un concept très important dans le monde de l'apprentissage automatique, c'est une technique qui automatise l'extraction et la synthèse des caractéristiques importantes et nécessaires à l'indentification des classes cible. [15]

Le prétraitement (Data Preprocessing)

Les algorithmes de l'apprentissage automatiques apprennent à partir des données qui leur sont fournies, par conséquent, si ces données sont de mauvaise qualité, erronées, incomplètes ou redondantes, l'algorithme qui en résulte sera lui-même assez mauvais, puisqu'il est censé reflété ce qu'il voit dans les données. Dans ce cas il est impératif de bien préparer nos données avant leur introduction dans la machine, en les nettoyant, les filtrant et les normalisant, c'est cette étape qu'on appelle : Le prétraitement.

5.2 CNN

Un réseau de neurones à convolution est un modèle d'apprentissage approfondi supervisé utilisé pour effectuer une classification ou une régression, ce modèle s'inspire de la disposition des neurones dans le cortex visuel des animaux. [15]

5.3 Auto-encodeur (AE)

Les auto-encodeurs (AEs) sont des modèles de réseaux neuronaux non supervisés destinés à l'apprentissage destinés à l'apprentissage de représentation significatives de données. Il est composé de 2 parties : le codeur et le décodeur.

Le modèle vise à réduire l'erreur entre l'échantillon d'entrée et celui reconstitué afin d'apprendre les caractéristiques représentatives des données d'entrée. [15][21]

5.4 RNN

Les réseaux de neurones récurrents (RNNs) sont les modèles de réseaux de neurones les plus compatibles avec le traitement des donnés séquentielles. Ces données séquentielles peuvent être trouvés dans les données d'ADN qui contiennent des milliers de protéines de bases.[15]

6 Conclusion

L'intelligence artificielle est un domaine large qui via ses longs racines (apprentissage en profondeur, apprentissage automatique, etc.) nous offre un large choix permettant la résolution d'une grande variété

de problèmes et ce en se basant sur des algorithmes. Dans ce chapitre nous avons présenté simplement les différentes approches de description et d'analyse des données et leur application. La bio-informatique est sans doute le plus grand domaine d'application de l'intelligence computationnelle.

Dans le prochain chapitre nous allons expliquer notre contribution dans l'application des approches de l'intelligence computationnelle dans la sélection des descripteurs pertinents des molécules existantes dans notre base de données afin de choisir les meilleures molécules candidates.

Chapitre 3 : Contribution

Contenu

6. Introduction

Au cours des dernières années, l'apprentissage en profondeur a conduit à de très bonnes performances sur une variété de problèmes, tels que Drug Discovery pharma et médecine. Parmi les différents types de réseaux neuronaux profonds, les réseaux de neurones convolutionnels ont été le plus étudiés. Tirer parti de la croissance rapide de la quantité de données annotées et des grandes améliorations des forces des unités de traitement.

Dans ce chapitre, nous allons présenter notre approche proposée, qui consiste à la modélisation QSAR en intégrant le réseau de neurones à convolution CNN pour la construction d'un modèle de régression pour prédire l'IGC50 d'u jeux de données QSAR., Aussi nous allons proposer une méthode de recherche pour extraire les descripteurs moléculaires descriptifs à partir d'un modèle CNN entrainé.

7. Approche proposée

L'approche proposée est basée sur les réseaux de neurones à convolution (CNN). Les architectures des réseaux de neurones se sont améliorées et la puissance de calcul disponible et la disponibilité des données. En effet, le grand enjeu pour le Deep Learning reste la capacité à être correctement entraîné et à avoir à disposition un nombre virtuellement infini d'exemples pour parfaire le modèle à construire. Notre approche proposée consiste de trois phases, extraction de descripteurs, l'apprentissage d'une fonction régression à l'aide d'un modèle CNN proposé et la sélection de descripteurs explicatifs à l'aide d'une technique d'optimisation coopérative.

La figure suivante illustre le Framework complet de notre approche.

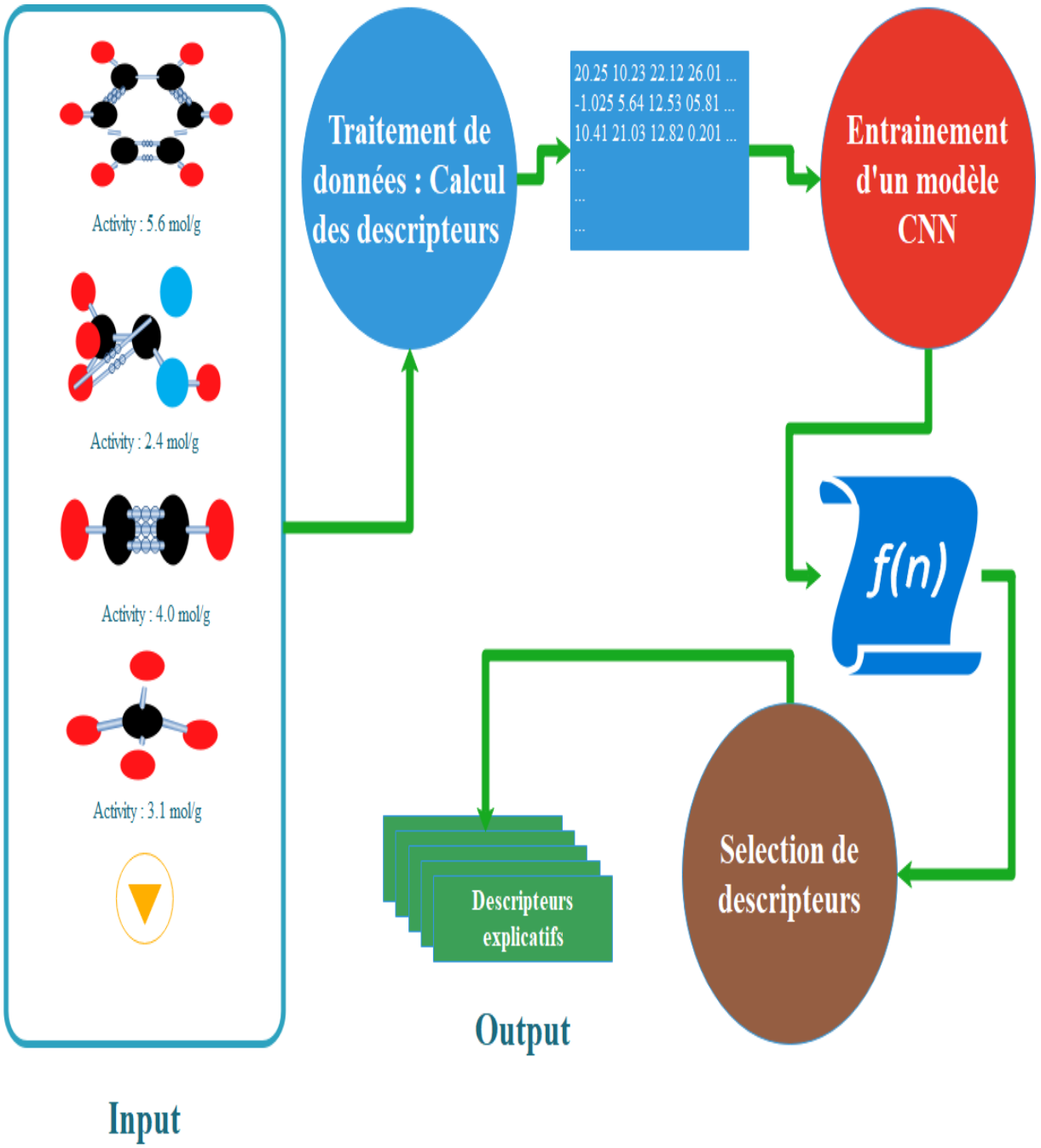


Figure 3. 1 Vue global de l'approche proposée

8. Calcul de descripteurs

Un modèle QSAR est dépendant des données expérimentales de référence, le choix de la base de données est un point critique dans le développement de ces modèles. Dans la plupart des cas, les données expérimentales sont issues de la littérature, et pour être de qualité, une base de données doit être composée de données expérimentales aussi fiables que possible, puisque les barres d'erreurs sur celles-ci se propageront dans le modèle final, étant donné que les paramètres de ce dernier sont ajustés par rapport à ces données.

Il est donc important de choisir des données présentant des incertitudes faibles, afin de limiter les barres d'erreur expérimentales. De plus, les données doivent être obtenues suivant un protocole expérimental unique. En effet, les conditions expérimentales ont une forte influence sur les valeurs obtenues.

La définition de la propriété en termes de conditions expérimentales est d'ailleurs un point important de la démarche. Un ensemble de données d'entrée typique pour l'analyse QSAR est, dans sa forme de base, une matrice contenant les produits chimiques étudiés dans les rangées et la des valeurs de descripteurs correspondantes.

Dans cette partie nous allons aborder le processus de l'extraction des descripteurs à partir un jeu de données contenant des molécules en utilisant une bibliothèque python appelée « MORDRED ». C'est une plateforme dédie pour le calcul des descripteurs moléculaires, pouvant générer plus de 1800 descripteurs 2D et 3D avec une grande performance même pour les grandes molécules, ce qui n'est pas l'apanage des autres plateformes connues tels que : ChemDes, PaDEL, ChempPy, etc.

8.1. Motivation de choix

Le prétraitement des molécules affecte les résultats d'extraction de descripteurs dans la plupart des logiciels. Cependant, pour chaque descripteur, MORDRED prétraite automatiquement les molécules (en ajoutant ou en supprimant des atomes d'hydrogènes, en performant la kekulisation, et en détectant l'aromaticité moléculaire). Ainsi que, tous les descripteurs sont automatiquement testés pour vérifier si le calcul des résultats est précis, en utilisant les valeurs de références des descripteurs moléculaires, les valeurs références sont collectées à partir des études

publiées, des constatations consensuelles de multiples programmes de calcul des descripteurs, et les résultats des contrôles manuels selon les algorithmes de calcul publiés.

8.2. Conception de MODRED

Mordred consiste en deux classes majeures ; la classe de descripteurs et la classe de calculateur.

- **La classe de descripteurs** : les algorithmes des descripteurs moléculaires sont mis en place dans une sous-classe à l'intérieur de la classe de descripteurs.

  Parce qu'il y a des interdépendances entre les calculs de descripteur moléculaire, une sous-classe peut dépendre d'une autre et ceci a pour but d'augmenter l'efficacité.

Plusieurs algorithmes de calculs de descripteurs donnent de multiples résultats ce qui complique le traitement des valeurs de descripteurs, cependant, chaque copie de descripteur donnée aux utilisateurs renvoie une seule valeur à MORDRED, ce qui le rend, le meilleur outil de calcul de descripteurs qui soit.

- **La classe de calculateur** gère la dépendance entre les descripteurs, cache les résultats, traite les erreurs et permet le calcul parallèle. Pour calculer les descripteurs d'une seule molécule, une copie de calculateur peut être utilisée en tant que fonction, tandis que pour calculer les descripteurs de multiples molécules, c'est la méthode 'map' qui est utilisée.

La figure suivante illustre comment MORDRED calcule les descripteurs.

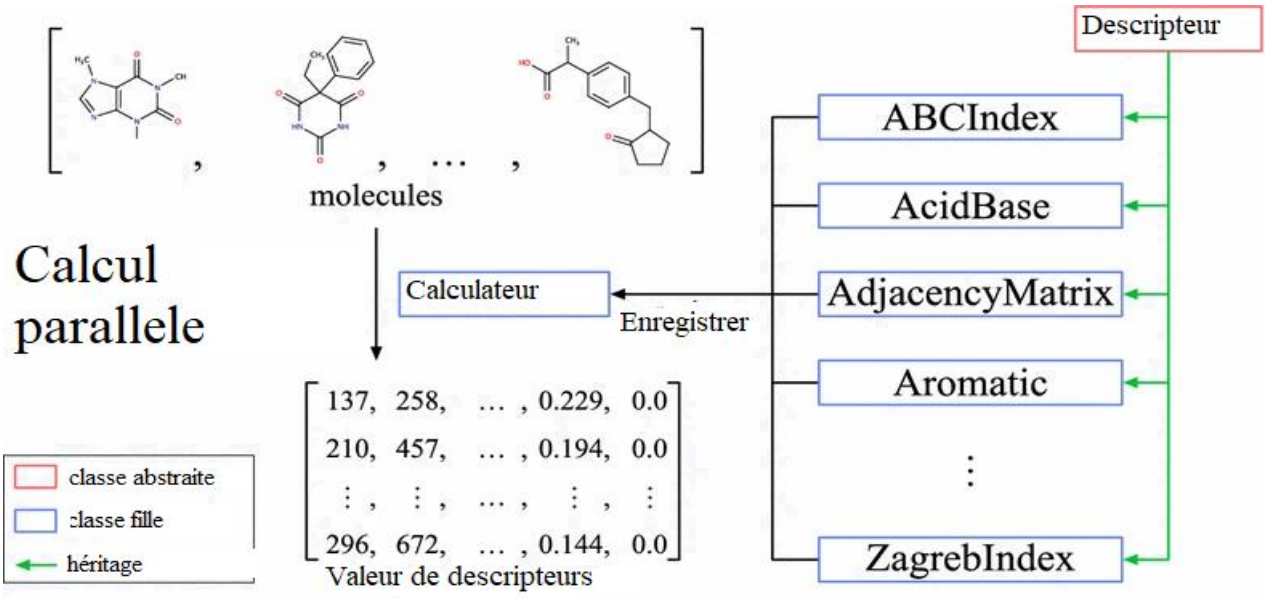


Figure 3. 2 Fonctionnement de la bibliothèque Mordred

A la fin de cette étape, nous obtenons une matrice contenant les descripteurs de chaque molécule avec une taille de **1613**.

Ci-dessus le code source utilisé pour calculer les descripteurs des molécules en utilisant la bibliothèque MORDRED.

```
from rdkit import Chem
from mordred import Calculator, descriptors
calc = Calculator(descriptors, ignore_3D=True)
file = open ("LNS-NCI-H522.txt")
i=0
smiles = []
for line in file :
  if(i!=0):
    mol = Chem.MolFromSmiles(line.split("\t")[1])
    smiles.append(mol)
  i = i + 1
df = calc.pandas(smiles)
df.to_csv("descriptors.csv")
```


Code Source 1: code source python pour utiliser la bibliothèque MORDRED pour calculer les descripteurs

9. Modèle d'apprentissage

La grande émergence des réseaux de neurones à convolutions (CNN) en bioinformatique et dans le domaine de conception de médicaments, conduit à les intégrer dans plusieurs modèles de QSAR et de criblage virtuel pour ses performances dans l'analyse et l'exploitation de données complexe et de Big Data. Par conséquent, dans notre approche, nous avons construit un modèle CNN pour faire une fonction de régression QSAR à la base des descripteurs calculés dans la partie 1 de notre approche.

Notre modèle est construit de deux parties. La première partie consiste à l'apprentissage de caractéristique. La deuxième partie consiste à l'apprentissage d'une fonction de régression. La figure ci-dessous décrit, l'architecture du CNN proposé :

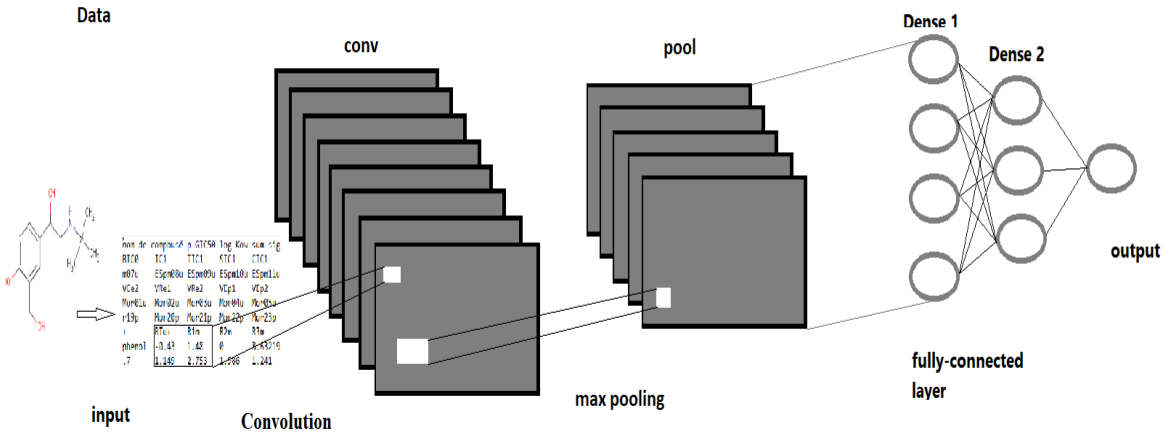


Figure 3. 3 l'architecture de modèle CNN proposée

9.1. Apprentissage de caractéristiques

Cette partie de modèle consiste à l'apprentissage de caractéristique afin d'extraire un pattern de caractéristique à partir les vecteurs d'entrés représentant les descripteurs. Afin de résoudre le problème de la taille des jeux de données d'entré, nous avons appliqué les étapes suivantes :

- **Convolution :** Nous avons appliqué la couche Conv2D après avoir remodelé les vecteurs des descripteurs 1-dimension en 2-dimension. Ensuite, les filtres

dans cette couche, utilisent l'algorithme de fenêtre coulissante, afin de glisser sur les données d'entrée 2D, en effectuant une multiplication dans le sens des éléments. Par conséquent, les filtres seront additionnés dans une seule matrice de sortie. Chaque fois que le noyau glisse sur un emplacement, il effectuera la même opération, transformant les vecteurs d'entrés en une matrice de caractéristiques différentes qui sont nommées la carte des caractéristiques ou le pattern des caractéristiques.

- **Rectification :** Nous avons appliqué une fonction d'unités linéaires rectifiées (ReLu) aux vecteurs d'entrée en raison de leur insaturation. ReLu maintient le gradient toujours au niveau élevé dans les couches activées. En outre, par rapport à la fonction sigmoïde ou à des fonctions d'activation similaires comme *tanh*, ReLu est extensible dans l'entrainement de modèle sur l'ensemble de descripteurs de grande taille. L'équation suivante décrit la formule de ReLu :

$$\boldsymbol{ReLu(x) = \max(0, x)}$$

- **Sous-échantillonnage :** En raison de la taille énorme des données, une stratégie de sous-échantillonnage est utilisée basée sur la couche MaxPool2D. Comme Conv2D, cette couche applique une technique de fenêtre coulissante en utilisant un noyau sur la carte de caractéristiques extraite de l'étape précédente, pour réduire la taille en choisissant la valeur maximale de chaque noyau.

En conséquence de cette étape d'entrainement de caractéristiques, une carte de caractéristiques réduite contenant une représentation liée aux descripteurs importantes est obtenu et elle sera utilisé pour la tâche de régression.

En sortie de cette étape, nous allons obtenir une carte de caractéristique représente les descripteurs pertinents pour prédire IGC50.

9.2. Apprentissage de régression

À cette étape, nous formerions un modèle de régression fondé sur un perceptron multicouche. Les étapes suivantes sont effectuées pour vérifier cette tâche :

- **Aplatie (flatting) :** Afin de traiter les matrices de sortie à 2 dimensions, nous avons appliqué une couche aplatie pour remodeler ces matrices en vecteurs à 1 dimension.
- **MLP :** Nous avons utilisé deux couches de réseaux de neurones classique pour couvrir et améliorer la tâche d'apprentissage de la régression. Les nœuds de cette couche appliquent la fonction d'activation sigmoïde, qui est définie comme suit :

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

- **Dropout :** Les sorties des couches entièrement connectées sont passées par une couche d'abandon, qui effectue une mise à zéro aléatoire des entrées des demi-nœuds vers la couche de sortie pendant la phase d'entraînement. Cela régularise le réseau et évite over-fitting.
- **Couche de sortie :** nous avons intégré une couche de sorite afin de prédire la valeur de IGC50, elle utilise sigmoïde comme fonction d'activation.

```
import numpy as np
from keras import backend as K
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import pandas as pd
from keras.layers import Conv2D
from keras.layers import Input
from keras.layers import MaxPooling2D, UpSampling2D
from keras.models import Model
from keras.wrappers.scikit_learn import KerasRegressor
from keras.utils import np_utils
from keras.metrics import *
from keras.layers import *
from keras.models import Sequential
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import cross_val_score
from sklearn.pipeline import Pipeline
```

Code Source 3. 1 : L'appel de bibliothèques utilisés pour implémenter le modèle CNN

Le code source de cette phase est décrit ci-dessus :

```
def defCNN():
  model = Sequential()
  model.add(Conv2D(8, kernel_size=1,input_shape=(1613,1,1),activation="relu"))
  model.add(MaxPool2D(pool_size=1))
  model.add(Flatten())
  model.add(Dense(128))
  model.add(Dense(64))
  model.add(Dropout(0.5))
  model.add(Dense(1,activation='sigmoid'))
  model.compile(optimizer='adam', loss='mse')
  model.summary()
  return model
```

Code Source 3. 2 : implémentation de modèle CNN proposé

La configuration de CNN est illustrée dans la table suivante :

Table 3. 1 : Configuration du modèle proposé

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 1613, 1, 8)        16
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 1613, 1, 8)        0
_________________________________________________________________
flatten (Flatten)            (None, 12904)             0
_________________________________________________________________
dense (Dense)                (None, 128)               1651840
_________________________________________________________________
dense_1 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 65
=================================================================
Total params: 1,660,177
Trainable params: 1,660,177
           Non-trainable params: 0
```

10. Sélection des descripteurs explicatifs

Dans cette section, nous allons décrire comment interpréter et mesurer l'importance des descripteurs utilisées par le modèle de CNN proposé. CNN est un extracteur de caractéristiques efficace, Cependant, il n'y a pas une interprétation de pattern de caractéristiques complète et sur ce qui se passe à l'intérieur des couches cachées de CNN. C'est pour cette raison, le CNN est un black box. Pour cela, nous voulons proposer une méthode d'optimisation et de recherche pour interpréter les descripteurs sélectionnés de la part de modèle CNN entrainé.

Dans cette étape, nous avons utilisé une technique d'optimisation coopérative, ou nous avons intégré plusieurs métaheuristiques d'optimisation et de recherche pour améliorer la complexité de calcul et les résultats. En respectant les étapes suivantes, nous avons pu à sélectionner plusieurs patterns de descripteurs :

1. Initialiser aléatoirement un vecteur d'entrer *I* ;
2. Créer des ilots en contenant deux types de métaheuristiques ; (Algorithme génétique, Algorithme de différentiation évolutionnaire)
3. Mettre le coefficient de détermination comme fonction d'objectif ;
4. Appliquer les lois de croisement et mutation et sélection ;
5. Evaluer la fonction objective pour chaque ilot ;
6. Effectuer une migration : après chaque i^{eme} itération, la meilleure solution de chaque îlot est choisie pour subir la migration et elle est transmise à tous les îlots voisins de même modèle. Une fois que toutes les solutions sélectionnées sont reçues, une stratégie de recombinaison est utilisée pour former la nouvelle population. Ce processus est répété jusqu'à la condition d'arrêt soit atteinte (solution optimale ou itérations).

Ce travail a été réalisé en utilisant la bibliothèque PYGMO de python.

11. Conclusion

Dans ce chapitre, nous avons présenté notre contribution pour la sélection des descripteurs. Nous avons utilisé les descripteurs générés et calculés à partir une bibliothèque Python nommés : MORDRED. Après,

pour construire un modèle de régression QSAR, nous avons proposé et entrainé un modèle Deep Learning basé sur les réseaux de neurones à convolution. Dans l'étape final de notre contribution, nous avons utilisé une technique d'optimisation coopérative afin de mesurer les descripteurs explicatifs extraits à partir le modèle CNN entrainé.

Dans le chapitre suivant, nous allons présenter les résultats expérimentaux, ainsi que les techniques et les métriques de validation utilisées.

Chapitre 4 : Résultats & Discussion

Contenu

1 *Introduction*

Dans ce chapitre, nous allons présenter les plateformes utilisées pour réaliser notre approche proposée. Après, nous allons présenter les jeux de données utilisés ainsi que les métriques et la technique de validation utilisés. A la fin, nous allons présenter les résultats expérimentaux obtenus.

2 Plateforme d'exécution

Dans cette section, nous allons détailler les plateformes hardware et software utilisés pour implémenter notre approche.

2.1 Hardware

- Nom de l'appareil : DESKTOP-4S8L28F
- Processeur : Intel(R) Core (TM) i3-5005U CPU @ 2.00GHz 2.00 GHz
- Mémoire RAM installée : 4.00 Go
- ID de périphérique : 4C0389AF-BC65-4AB4-B824-391DEABDC96B
- ID de produit : 00331-10000-00001-AA482
- Type du système : Système d'exploitation 64 bits, processeur x64
- Stylet et fonction tactile : La fonctionnalité d'entrée tactile ou avec un stylet n'est pas disponible sur cet écran

2.2 Spécification de Windows

- Édition : Windows 10 Professionnel
- Version : 20H2
- Build du système d'exploitation : 19042.1083
- Expérience: Windows Feature Experience Pack 120.2212.3530.0

2.3 Software

Pour réaliser notre approche, nous avons utilisés les plateformes et les bibliothèques softwares suivants.

2.3.1 Python

Python est un langage de programmation open source interprété coté serveur et non compilé. Crée par Guido van Rossum, il est utilisé pour le développement web, le développement de jeux vidéo et autres logiciels, ainsi que pour les interfaces utilisateur graphique. Il a notamment été utilisé dans la création d'Instagram, de YouTube et de Spotify, et il est très important dans l'apprentissage automatique est l'un des langages de programmation officiels de Google.

Figure 4. 1 : Le logo de python.

2.3.2 Tensorflow

TensorFlow est très populaire pour ses nombreux avantages qu'on cite ci-dessous :

TensorFlow est aujourd'hui particulièrement utilisé pour l'apprentissage en profondeur et donc les réseaux de neurones. Son nom est notamment inspiré du fait que les opérations courantes sur des réseaux de neurones sont principalement faites via des tables de données multidimensionnelles, appelées Tenseurs (Tensor). Un Tensor `a deux dimensions est l´équivalent d'une matrice. Aujourd'hui, les principaux produits de Google sont basés sur TensorFlow : Gmail, Google Photos, Reconnaissance de voix…, etc.

Figure 4. 2: logo de tensorflow

2.3.3 KERAS

Est une API de réseaux de neurones de haut niveau, écrite en Python et capable de s'exécuter sur TensorFlow, CNTK ou Theano. Il a été d´développé pour permettre une expérimentation rapide. Pouvoir faire de la recherche de qualité est essentiel pour pouvoir passer de l'idée au résultat le plus rapidement possible.

Figure 4. 3: logo de keras

2.3.4 Jupyter

Jupyter Notebook est une application Web open source qui

Figure 4. 4 : logo de jupyter

permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Les utilisations comprennent : le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, le ML et bien plus encore.

2.3.5 NumPY

NumPy est un projet open source visant à permettre le calcul numérique avec Python. Il a été créé en 2005, en s'appuyant sur les premiers travaux des bibliothèques Numerical et Numarray. Il sera toujours 100% open source, gratuit pour tous. Il est développé à l'air libre sur GitHub, grâce au consensus de NumPy et de la communauté scientifique Python au sens large.

Figure 4. 5 Logo de numpy

2.3.6 PANDAS

Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

Figure 4. 6 : logo de pandas

2.3.7 PYGMO

PyGMO est une bibliothèque scientifique Python dérivée de PaGMO (Parallel Global Multiobjective Optimization), un logiciel open source développé à l'agence spatiale Européenne. Le cadre flexible et complet de PaGMO (et de son équivalent PyGMO) peut être appliqué aux problèmes d'optimisation à objectif unique, à objectifs multiples, continus, entiers, à contraintes de boite, à contraintes non linéaire, stochastiques et déterministes.

Figure 4. 7 : logo de pygmo

2.3.8 Colab

Colaboratory, souvent raccourci en (Colab), est un produit de Google Research. Colab permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté au machine Learning, à l'analyse de données et à l'éducation. En termes plus techniques, Colab est un service hébergé de notebooks Jupyter qui ne nécessite aucune

configuration et permet d'accéder gratuitement à des ressources informatiques.

Figure 4. 8: logo de colab

2.3.9 Mordred

Mordred a été conçu pour être un logiciel facile à installer et à utiliser, prenant en charge de nombreux descripteurs moléculaires, doté d'une vitesse de calcul élevée et comprenant des tests automatisés. Les programmes de calcul de descripteur moléculaire ont généralement de nombreux logiciels dépendants qui doivent être installés manuellement.

3 La base de données

Les approches computationnelles modernes et les techniques de l'apprentissage automatique sont en train d'accélérer l'invention de nouveaux médicaments pour plusieurs maladies comme le cancer.

Dans cette étude, la maladie ciblée est le cancer du poumon, ce dernier constitue la cause la plus commune de décès par le cancer dans le monde entier. Il est aussi connu sous le nom de carcinome pulmonaire, c'est une tumeur maligne caractérisée par une croissance cellulaire incontrôlée au niveau du tissu pulmonaire. Cette croissance peut se propager vers les tissus voisins ou vers d'autres parties du corps par un processus que l'on dénomme : métastase.

Histologiquement parlant, il y a deux sous types du cancer de poumon :

- Cancer du poumon à petites cellules (CPPC) (constitue 15% des cas)
- Cancer du poumon non à petites cellules (CPNPC) (constitue 85% des cas)

Ici nous sommes concernés par le cancer du poumon non à petites cellules (non small cell lung cancer) qui est la forme la plus commune du cancer du poumon. Cette maladie commence lorsque les cellules saines du poumon changent et croissent d'une manière incontrôlable, formant une masse appelée ; Tumeur, lésion ou nodule. Une tumeur peut commencer n'importe ou au niveau du poumon et peut être cancéreuse ou bénigne. Une tumeur cancéreuse est maligne, ce qui veut dire qu'elle peut se propager vers d'autres parties du corps. Une tumeur bénigne veut dire une tumeur qui peut croitre mais ne se propage pas vers d'autres tissus.

Le CPNPC commence au niveau des cellules épithéliales et ses différents types sont :

- Adénocarcinome
- Carcinome épidermoïde (ou malpighien)
- Carcinome pulmonaire à grande cellules
- CPNPC indifférencié (ou non classé par ailleurs)

Pour cibler ces cellules, nous avons utilisé des molécules importées à partir d'une base de données appelée : LNS-NCI-H522 contenant plus de 700 molécules anticancer NSCC du poumon.

La lignée cellulaire NCI-522 est responsable de l'adenocarcinome pulmonaire [22], de ce fait, nous allons nous intéresser à un récepteur dont l'une de ses fonctions principales est la prolifération cellulaire, et que la dérégulation de cette fonction a une implication importante dans la carcinogenèse, notamment, le cancer du poumon non à petites cellules (non small cell lung cancer) dont l'adénocarcinome pulmonaire (sous-type) fait partie , il s'agit de : HER2.

3.1 Aperçu général sur le HER2

Récepteur 2 pour le facteur de croissance épidermique humain (HER2), connu aussi sous le nom : ERBB2, est une protéine avec un domaine tyrosine kinase intracellulaire et un domaine de liaison du ligand extracellulaire. La famille HER inclut 4 membres, HER1(ERBB1, connu aussi sous le nom EGFR), HER2(ERBB2), HER3(ERBB3) et HER4(ERBB4). Le HER2 est le seul qui n'a pas de ligand spécifique identifié, c'est le partenaire préféré pour former un hétérodimère avec un autre membre du HER. HER2 qui est impliqué dans l'hétérodimérisation constitue la voie de signalisation la plus puissante parmi tous les dimères

formés par la famille HER. HER2 joue un role important dans la croissance cellulaire, la survie et la différentiation. La voie de signalisation majeure induite par HER2 implique la voie de mitogen-activated protein kinase (MAPK) et la voie de la phosphatiylinositol 3-kinase (PI3K).

3.2 Structure du HER2

Similaire à tous les récepteurs de la famille HER, HER2 est une glycoprotéine transmembranaire de type 1 composée de 3 région distinctes : le domaine extracellulaire N-terminal (ECD), un domaine transmembranaire α-hélice unique (TM), et un domaine tyrosine kinase intracellulaire. Le domaine extracellulaire N-terminal est la plus grande partie et il contient approximativement 600 résidus (90~110 kD), c'est un domaine formé par 4 subdomaines (I–IV). Le subdomaine I et III peuvent former un site de liaison pour des ligands potentiels. Tandis que les subdomaines riches en cysteine II et IV sont impliqués à l'homodimérisation et l'hétérodimérisation. Le subdomaine II qui contient un bras de dimérisation, semble être le contributeur majeur de la dimérisation.[23]

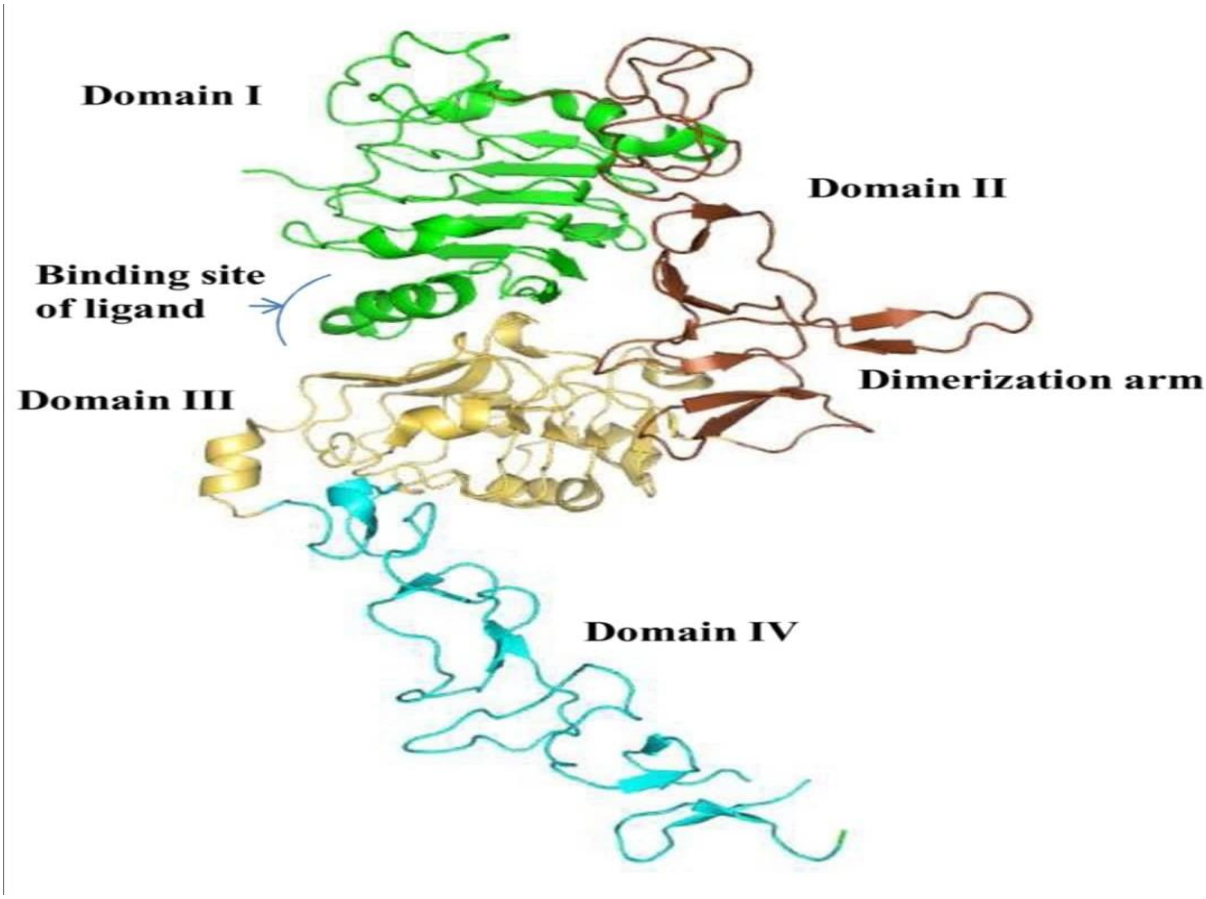


Figure 4. 9: La structure du domaine extracellulaire du HER2

Le domaine TM (transmembranaire) du récepteur HER2 est un α-hélice unique constitue en 23 acides aminés. Le séquençage d'alignement de la famille HER montre qu'il y a deux motifs avec une séquence conservée de 5 résidus dans le domaine TM. Un point de mutation (VAL-664→GLU) au niveau du motif Sternberg-Gullick du gène oncogène neu du rat est connu pour induire une transformation oncogène. Le motif G***G se retrouve dans le domaine TM du HER1, HER2 et HER4, et non dans HER3. Ces deux motifs au niveau du domaine TM pourraient être la force motrice principale pour la dimérisation du récepteur. [23]

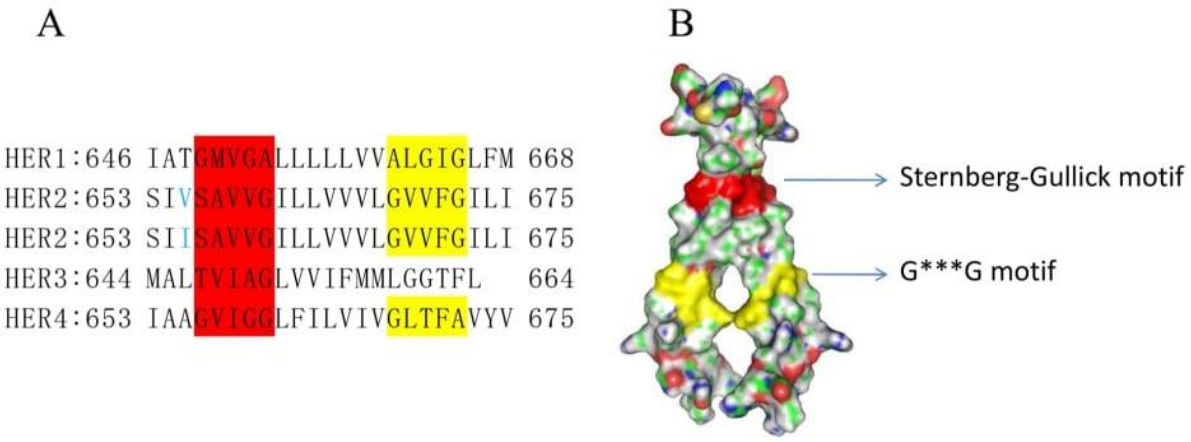


Figure 4. 10 : La séquence et la structure du domaine transmembranaire du récepteur HER[23]

Le domaine intracellulaire est approximativement constitue de 500 résidus et consiste en un coupleur cytoplasmique juxtamembranaire (cytoplasmic juxtamembrane linker), un domaine tyrosine kinase (TyK) et une queue terminale carboxylique. Le JM linker est un coupleur felxible et court qui connecte le domaine TM et le domaine Tyrosine Kinase. Au niveau de la partie la plus compliquée du récepteur HER2, le domaine TyK contient plusieurs boucles importantes qui forment le site actif de l'enzyme. La queue terminale carboxylique a six résidus tyrosine qui sont disponible pour la transphosphorylation, et joue le rôle d'un site d'amarrage pour les molécules de signalisation qui contiennent Src homology 2 (SH2) ou le domaine de liaison phosphotyrosine (PTB). [23]

3.3 La voie de signalisation du récepteur HER2

Après la dimérisation, le récepteur HER2 peut signaler à travers 3 voies différentes ; PI3K, MAPK, et le phospholipase C-γ (PLCγ). Le type de dimérisation influence significativement les voies de signalisation en aval, contrairement à la voie PI3K, tous les HER2 impliqués dans la dimérisation (HER1/HER2, HER2/HER3, et HER2/HER4) peuvent activer la voie MAPK.

Les voies de signalisation PI3K et MAPK sont des voies de signalisation clés pour promouvoir la prolifération cellulaire et prévenir l'apoptose. [23][24]

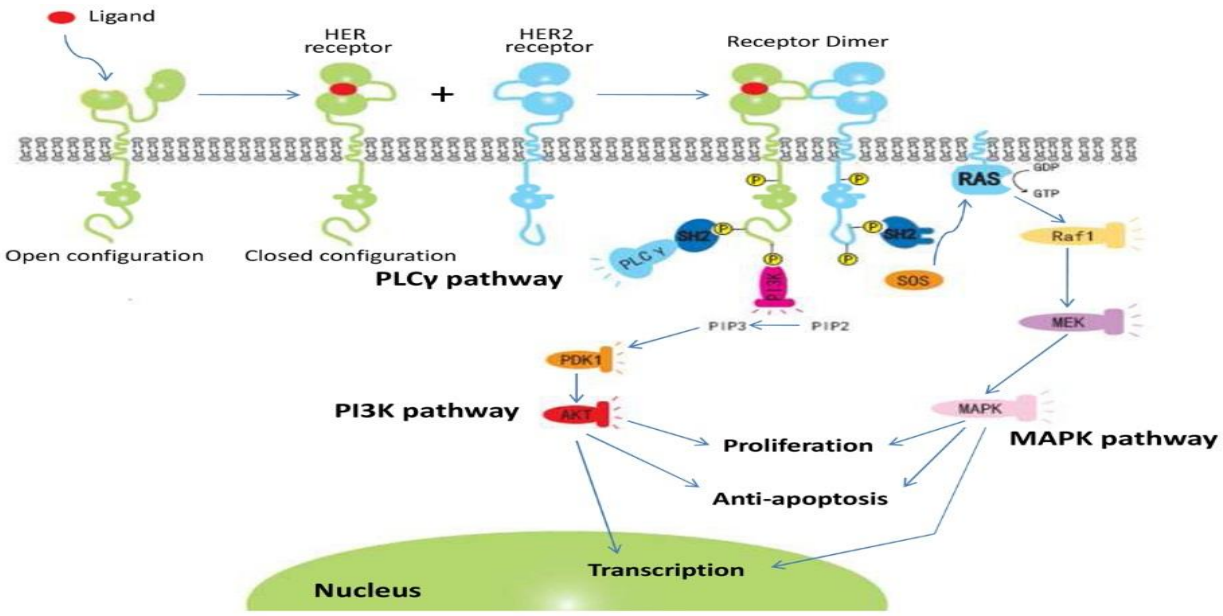


Figure 4. 11: La voie de signalisation du récepteur HER2 (après l'activation dépendante du ligand du récepteur HER, HER2 se dimérise avec le récepteur HER activé ce qui mène à la phosphorylation des résidus tyrosine et la transduction du signal.[23]

Remarque : PLCγ, PI3K et MAPK sont les cascades les plus communes, tandis que PI3K et MAPK sont les voies majeures impliquées dans la croissance du tumeur et l'antiapoptose. [23]

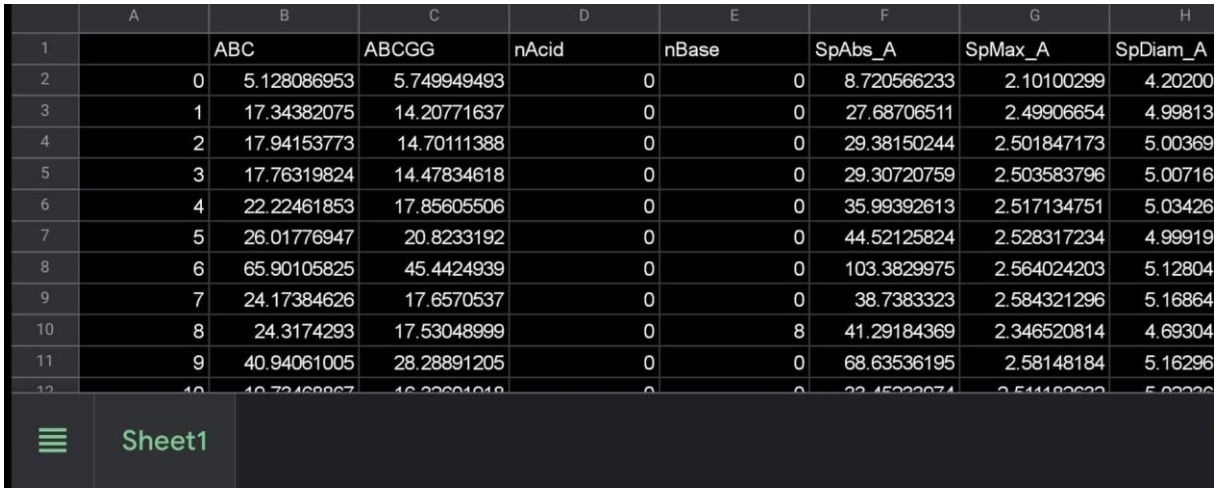
3.4 Rôle du HER2 dans la carcinogenèse

Dans les cellules normales, HER2 joue un role important tout au long des stages de développement cellulaire, toutefois, la mutation ou la surexpression du HER2 pourrait directement mener à la cancérogenèse ainsi qu'à la métastase. Même si la mutation du gène neu est requise à la cancérogenèse chez les rongeurs, le HER2 humain semble contenir un potentiel tumorigène à travers la surexpression du gène HER2 wild-type. La surexpression du HER2 augmente et prolonge les signaux qui déclenche la transformation des cellules.[23]

Les mutations du HER2 dans le cancer non à petites cellules du poumon constitue une cible moléculaire importante.

3.5 Taille de descripteurs

Ils constituent la partie déterminante de l'activité des molécules qu'on explore, notre model renferme 1613 descripteurs, il y a les descripteurs constitutionnels, thermodynamiques, topologiques, géométriques et électrostatiques.



	A	B	C	D	E	F	G	H
1		ABC	ABCGG	nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A
2	0	5.128086953	5.749949493	0	0	8.720566233	2.10100299	4.20200
3	1	17.34382075	14.20771637	0	0	27.68706511	2.49906654	4.99813
4	2	17.94153773	14.70111388	0	0	29.38150244	2.501847173	5.00369
5	3	17.76319824	14.47834618	0	0	29.30720759	2.503583796	5.00716
6	4	22.22461853	17.85605506	0	0	35.99392613	2.517134751	5.03426
7	5	26.01776947	20.8233192	0	0	44.52125824	2.528317234	4.99919
8	6	65.90105825	45.4424939	0	0	103.3829975	2.564024203	5.12804
9	7	24.17384626	17.6570537	0	0	38.7383323	2.584321296	5.16864
10	8	24.3174293	17.53048999	0	8	41.29184369	2.346520814	4.69304
11	9	40.94061005	28.28891205	0	0	68.63536195	2.58148184	5.16296

Sheet1

Figure 4. 12 : une capture d'écran de notre fichier "molécule/descripteurs"

4 Validation

La validation a un objectif de vérifier que toutes les étapes de fabrication d'un model aboutiront à un model conforme et efficace.

4.1 Métriques de validation

Ces métriques permettent de bien évaluer notre model, et on en a utilisé deux :

4.1.1 Coefficient de détermination

Le coefficient de détermination est un indicateur qui permet de juger la qualité d'une régression linéaire simple. Ce coefficient varie entre 0 et 1, soit entre un pouvoir de prédiction faible et un pouvoir de prédiction fort. Il mesure l'adéquation entre le modèle et les données observées ou encore à quel point l'équation de régression est adaptée pour décrire la distribution des points. Si le R^2 est nul, cela signifie que

l’équation de la droite de régression détermine 0 % de la distribution des points. Cela signifie que le modèle mathématique utilisé n’explique absolument pas la distribution des points. Si le R^2 vaut 1, cela signifie que l’équation de la droite de régression est capable de déterminer 100 % de la distribution des points. Cela signifie alors que le modèle mathématique utilisé, ainsi que les paramètres a et b calculés sont ceux qui déterminent la distribution des points. En bref, plus le coefficient de détermination se rapproche de 0, plus le nuage de points se disperse autour de la droite de régression. Au contraire, plus le R^2 tend vers 1, plus le nuage de points se resserre autour de la droite de régression. Quand les points sont exactement alignés sur la droite de régression, alors $R^2 = 1$. [25]

$$r^2 = 1 - \frac{\Sigma (y - y')^2}{\Sigma (y - \overline{y}')^2}$$

4.1.2 Erreur quadratique moyenne (Mean squared error)

MSE est généralement une métrique d'évaluation de régression plus populaire que MAE. L'accent supplémentaire mis par MSE sur les erreurs importantes est souhaitable car nous voulons produire un modèle qui se généralise bien et produit des prédictions avec de faibles erreurs sur l'ensemble de données.

Comme MAE, une valeur MSE plus proche de zéro indique de meilleures performances du modèle. MSE est utilisé moins fréquemment pour l'interprétation humaine car il n'est pas représenté dans des unités facilement interprétables, mais il est très populaire pour une utilisation dans l'optimisation de l'apprentissage automatique. [27]

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

$$y_i = actual\ value$$
$$\hat{y}_i = predicted\ value$$
$$n = \#\ of\ observations$$

4.2 Technique de validation

Après avoir entraîné **un modèle de Machine Learning** sur des données étiquetées, celui-ci est supposé fonctionner sur de nouvelles données. Toutefois, il est important de s'assurer de l'exactitude des prédictions du modèle en production. Pour ce faire, il est nécessaire de valider le modèle. Le processus de validation consiste à décider si les résultats numériques quantifiant les relations hypothétiques entre les variables sont acceptables en tant que descriptions des données. On le teste sur de nouvelles données et en fonction des performances des modèles sur des données inconnues, on peut déterminer s'il est sous-ajusté, sur-ajusté, ou " bien généralisé ".

L'une des techniques utilisées pour tester l'efficacité d'un modèle de Machine Learning est la " cross-validation " ou validation croisée. Cette méthode est aussi une procédure de " re-sampling " (rééchantillonnage) permettant d'évaluer un modèle même avec des données limitées. Pour effectuer une " CV " (cross-validation), il est nécessaire d'écarter en amont une partie des données du dataset d'entraînement. Ces données ne seront pas utilisées pour entraîner le modèle, mais plus tard pour tester et valider le modèle. On l'utilise en Machine Learning pour comparer différents modèles et sélectionner le plus approprié pour un problème spécifique. Elle est à la fois simple à

comprendre, simple à implémenter et moins biaisée que les autres méthodes. Découvrons à présent les principales techniques de validation croisée. [28]

5 Résultats Expérimentaux

Après l'entrainement du modèle nous avons obtenu des résultats pour évaluer le modèle proposé. Les résultats expérimentaux vont être discuté dans cette partie.

5.1 Résultats d'évaluation du modèle

Ci-dessus, la table suivante représente les résultats expérimentaux de validation du modèle en comparant avec des algorithmes de machine learning traditionnelle comme : perceptron multicouches (MLP), Support Vector Machine (SVM), Random Forest (RF) et la régression linéaire (LR).

Table 4. 1 : Comparaison entre différents algorithmes de machine learning en utilisant deux métriques

Modèle	Coefficient de détermination %	Erreur quadratique moyenne %
CNN	**80,51**	**25**
MLP	**62.01**	**62**
SVM	**67.87**	**51.31**
RF	**25.12**	**78.12**
LR	**-21.05**	**255.21**

La table 4.1 montre que la méthode de régression basée sur le CNN est très performante avec un R2 score égale à 80.51% et un MSE égale à 0.25. Parce que CNN peut éliminer les descripteurs non pertinents qui font de flou pour l'apprentissage et extraire un pattern de caractéristiques qui représente les descripteurs explicatifs.

5.2 Résultats de sélection de descripteurs

Après l'évaluation du modèle, nous avons lancé une optimisation coopérative décrite dans le chapitre 3. Nous avons obtenu 3 patterns de descripteurs explicatifs. La table suivante représente les patterns extraits et le score du coefficient de détermination pour chaque pattern.

Table 4. 2 : Les patterns de descripteurs sélectionnés

Pattern de descripteurs	R2-score %
VE3_Dzm, AETA_eta_RL, SpAbs_Dzv, MAXssssAs, Xc-3d, LogEE_Dzm	**71,09**
VE3_A, AETA_eta_RL, SpAbs_Dzv, MAXssssAs, Xp-0d, MATS8i, NsSeH	**71.02**
AATS2s, n9aHRing, n10aHRing, n11aHRing, nRot, RotRatio, SLogP, SMR , SpMax_Dzm, AATS1d, GATS7dv.	**72,58**

La table 4.2 décrit des ensembles de descripteurs sélectionnés avec un coefficient de détermination. Ce qui montre que le meilleur au terme de r2-score est le troisième ensemble, mais il contient plus de descripteurs.

6 Conclusion

Dans ce chapitre nous avons présenté les principaux outils et plateformes utilisés pour réaliser notre projet. Nous avons aussi discuté les métriques et les techniques de validations. A la fin, nous avons présenté les résultats expérimentaux obtenus.

Conclusion Générale

Contenu

1. Conclusion

Le QSAR est une procédure prédictive très importante, elle exige de bons descripteurs afin de nous livrer une bonne prédiction des nouvelles molécules thérapeutiques, une bonne compréhension sur leur mode d'action, sur leur structure et par conséquent, une amélioration de leur activité.

Après avoir calculer les descripteurs des molécules disponibles dans notre base de données et sélectionner les meilleurs, nous avons proposé une technique de l'apprentissage approfondi (Deep Learning) basée sur les réseaux de neurones à convolution (CNN) pour construire un modèle de régression QSAR, ensuite nous avons intégré un modèle des ilots généralisé qui est un modèle d'optimisation et de recherche coopérative afin de trouver un pattern de descripteurs pertinents pour les molécules de NSCLC.

Les résultats expérimentaux obtenus à partir du modèle de régression basé sur CNN sont très prometteuse, avec un coefficient de détermination supérieur à 80,51%. Ainsi, nous avons obtenu plusieurs patterns de descripteurs de l'activité biologique ciblé.

2. Perspectives

Les résultats de ce modeste travail constituent les bases d'un travail à poursuivre et à améliorer pour une étude beaucoup plus approfondie qui pourra faire l'objet d'un projet de recherche, surtout quand on s'aperçoit que les études portées sur l'implication du récepteur HER2 dans le NSCLC sont encore à leur début. Ainsi, les perspectives futures seraient une étude sur des bases de données et avec des autres types de données.

Bibliographie

Références

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

[2] "Ligand-Gated Ion Channels Library - 100% quality control. Custom Design." https://www.chemdiv.com/ligand-gated-ion-channels-library/ (accessed Mar. 22, 2021).

[3] "Voie signalisation recepteur second messager transduction signal Relation structure fonction proteine protein structure function relationship Enseignement recherche Biochimie Universite Angers Emmanuel Jaspard biochimej." http://biochimej.univ-angers.fr/Page2/COURS/7RelStructFonction/2Biochimie/5Signalisation/4RCPGetProteinesG/1RCPGetProtG.htm (accessed Mar. 22, 2021).

[4] R. R. Nadendla, "Molecular modeling: A powerful tool for drug design and molecular docking," *Resonance*, vol. 9, no. 5, pp. 51–60, 2004, doi: 10.1007/bf02834015.

[5] "Plateforme Modélisation Moléculaire." http://ea4481.univ-lille2.fr/fr/plateformes/plateforme-modelisation-moleculaire.html (accessed Mar. 22, 2021).

[6] S. Cosconati, S. Forli, A. L. Perryman, R. Harris, D. S. Goodsell, and A. J. Olson, "Virtual screening with AutoDock : theory and practice," pp. 597–607, 2010.

[7] H. van de Waterbeemd and E. Gifford, "ADMET in silico modelling: Towards prediction paradise?," *Nat. Rev. Drug Discov.*, vol. 2, no. 3, pp. 192–204, 2003, doi: 10.1038/nrd1032.

[8] Danishuddin and A. U. Khan, "Descriptors and their selection methods in QSAR analysis: paradigm for drug design," *Drug Discov. Today*, vol. 21, no. 8, pp. 1291–1302, 2016, doi: 10.1016/j.drudis.2016.06.013.

[9] E. T. D. E. La and R. Scientifique, "Élaboration des modèles QSPR prédictifs des propriétés physico- chimiques à l'aide des descripteurs moléculaires.," 2015.

[10] K. Roy, *Ecotoxicological QSARs*. 2020.

[11] "Comment intégrer l'intelligence artificielle dans vos projets - Nat System." https://www.natsystem.fr/comment-integrer-lia-dans-vos-projets (accessed Aug. 10, 2021).

[12] "Intelligence computationnelle." https://www.24pm.com/component/tags/tag/intelligence-computationnelle (accessed Aug. 10, 2021).

[13] O. Toolbox, “X . Algorithmes d ’ optimisation,” no. x, pp. 1–14, 2010.

[14] “🔎 Métaheuristique - Définition et Explications.” https://www.techno-science.net/glossaire-definition/Metaheuristique.html (accessed Aug. 10, 2021).

[15] B. Pr, A. Boukelia, M. C. Batouche, and M. Belguidoum, “Méthodes d’intelligence computationnelle pour l’analyse et l’exploration des données épigénétiques massives,” 2020.

[16] C. Blum and A. Roli, “Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison,” *ACM Comput. Surv.*, vol. 35, no. 3, pp. 268–308, 2003, doi: 10.1145/937503.937505.

[17] K. Arfara, “Table Des Mati Ères,” *Théâtralités Contemp.*, pp. 1–80, 2016, doi: 10.3726/978-3-0352-0111-6/9.

[18] “Glossaire.” https://www.coe.int/fr/web/artificial-intelligence/glossary (accessed Aug. 10, 2021).

[19] O. Hajji, “Contribution au développement de méthodes d’optimisation stochastique, Application à la conception des dispositifs électrotechniques,” pp. 1–171, 2003.

[20] “3. Le perceptron multicouche.” https://www.becoz.org/these/memoirehtml/ch06s04.html (accessed Aug. 10, 2021).

[21] L. Deng, D. Yu, and B. -Delft, “Deep Learning: Methods and Applications,” doi: 10.1561/2000000039.

[22] “NCI-H522 (Human Non-small Cell Lung Adenocarcinoma) Whole Cell Lysate – LY500004 | OriGene.” https://www.origene.com/catalog/proteins/cell-line-lysates/ly500004/nci-h522-human-non-small-cell-lung-adenocarcinoma-whole-cell-lysate (accessed Aug. 26, 2021).

[23] W. Tai, R. Mahato, and K. Cheng, “The role of HER2 in cancer therapy and targeted drug delivery,” *J. Control. Release*, vol. 146, no. 3, pp. 264–275, 2010, doi: 10.1016/j.jconrel.2010.04.009.

[24] Y. Yarden and M. X. Sliwkowski, “Untangling the ErbB signalling network,” *Nat. Rev. Mol. Cell Biol.*, vol. 2, no. 2, pp. 127–137, 2001, doi: 10.1038/35052073.

[25] “coefficient de détermination | Lexique de mathématique.” https://lexique.netmath.ca/coefficient-de-determination/ (accessed Aug. 26, 2021).

[26] “coefficient of determination | Barrons Dictionary | AllBusiness.com.” https://www.allbusiness.com/barrons_dictionary/dictionary-coefficient-

of-determination-4950051-1.html (accessed Aug. 31, 2021).

[27] “Métriques d’évaluation communes pour l’analyse de régression.” https://ichi.pro/fr/metriques-d-evaluation-communes-pour-l-analyse-de-regression-82886198762157 (accessed Aug. 28, 2021).

[28] “Cross-Validation : définition et importance en Machine Learning.” https://datascientest.com/cross-validation (accessed Aug. 28, 2021).

Année universitaire : 2020 / 2021

Présenté par :
GUENDOUZ MOHAMED
MRI ROUMEISSA

Une approche QSAR basée sur Deep Learning pour la sélection des descripteurs Explicatifs.

Mémoire de fin de cycle pour l'obtention du diplôme de master en biochimie appliquée

Résumé :

Depuis les découvertes et les progrès de techniques de séquençages haut débit (NGS) et la chromatographie en phase liquide à haute performance (HPLC), les chercheurs se focalisent sur le traitement des cellules tumorales, en utilisant de nouvelles techniques thérapeutiques basées sur les tendances de la découverte des médicaments. Cette dernière peut être décrite comme le processus d'identification des entités chimiques. Les chercheurs se focalisent sur le traitement des cellules tumorales, en utilisant nouvelles techniques thérapeutiques. La relation quantitative structure-activité (QSAR) est un domaine important dans la conception et de la découverte de médicaments, la recherche des renseignements sur la structure chimique des activités biologiques et pharmaceutiques. Cette approche exige de bons descripteurs moléculaires représentatifs des caractéristiques moléculaires responsables de l'activité moléculaire pertinente.

Dans ce travail, nous allons proposer une technique de l'apprentissage approfondi (Deep Learning) basée sur les réseaux de neurones à convolution (CNN) pour construire un modèle de régression QSAR comme une première partie de travail. Ensuite, nous allons intégrer un modèle des ilots généralisés qui est un modèle d'optimisation et de recherche coopérative, afin de trouver un pattern de descripteurs pertinents pour les molécules de NSCLC.

Les résultats expérimentaux obtenus à partir le modèle de régression basé sur CNN sont très prometteuse, avec un coefficient de détermination supérieur à 80,51%. Ainsi, nous avons obtenu plusieurs patterns de descripteurs de l'activité biologique ciblée.

Mots clés : QSAR, Descripteur moléculaire, Deep Learning, Optimisation coopérative.

Laboratoire de recherche : laboratire de l'informatique N° 18 au niveau du centre de recherche de la biotechnologie (CR.B.T) Constantine

Devant le jury :

Président de jury : Mr. BENSEGUENI ABDERRAHMANE Pr. UFM. Constantine 1
Encadreur : Dr. BOUKELIA ABDELBASSET MRA CRBt Constantine
Examinateur : Mr. MOKRANI EL HASSEN MAA UFM. Constantine 1
Examinateur : Mr. DEMS MOHAMED ABDESSELAM. MRA CRBt Constantine