ROYAUME DU MAROC
Université Mohammed V - RABAT -
جامعة محمد الخامس – الرباط –
Faculté des sciences
كلية العلوم
CENTRE D'ÉTUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

N° d'ordre : 4101

THÈSE

en vue de l'obtention du : **DOCTORAT**

Structure de Recherche : Intelligent Processing and Security of Systems
Discipline : Informatique
Spécialité : Intelligence Artificielle et Science de Données

Présentée et Soutenue le : **21 / 06 / 2025**

par :

Outhman ABBASSI

Intelligence artificielle pour l'optimisation du processus de découverte des médicaments

Devant le JURY :

Samira KHOULJI	PES, Université Abdelmalek Essaâdi, Ecole Nationale des Sciences Appliquées de Tétouan	Présidente
Rachid SAADANE	PES, Ecole Hassania des Travaux Publics de Casablanca	Examinateur/Rapporteur
Nassim KHARMOUM	MCH, Centre National pour la Recherche Scientifique et Technique de Rabat	Examinateur/Rapporteur
Hicham GUEDDAH	MCH, Université Mohammed V, Ecole Normale Supérieure de Rabat	Examinateur/Rapporteur
Wajih RHALEM	MC, Université Mohammed V, Ecole Nationale supérieure d'Arts et Métiers de Rabat	Invité
Najib ALIDRISSI	Professeur Agrégé, Université Mohammed VI des Sciences et de la Santé de Rabat	Invité
Yassine ZAOUI SEGHROUCHENI	MCH, Université Mohammed V, Faculté des Sciences de Rabat	Co-Directeur de thèse
Soumia ZITI	PES, Université Mohammed V, Faculté des Sciences de Rabat	Directrice de thèse

Année Universitaire : 2024 - 25

Dédicace

Alhamdoulilah,
À l'âme de mon père, parti trop tôt, que son âme repose en paix éternellement.
À ma mère, merci pour ton amour inconditionnel et ton soutien sans faille. Ta présence constante est ma plus grande force.
À mon frère et à ma sœur, vous êtes mes complices, votre présence me donne de la force chaque jour.
À mes amis fidèles, vous êtes toujours là pour moi, merci pour votre soutien.
C'est grâce à vous que j'ai la force d'avancer chaque jour.

Remerciement

Les recherches effectuées dans le cadre de cette thèse ont été menées au sein du département d'informatique de la faculté des sciences de Rabat sous la direction du professeur Soumia ZITI et le coencadrement du professeur Yassine ZAOUI SEGHROUCHENI au sein de la structure de recherche Traitement intelligent et sécurité des systèmes (IPSS).

Tout d'abord, je tiens à exprimer ma gratitude à ma directrice de thèse, Mme Soumia ZITI, professeur de l'enseignement supérieur à la faculté des sciences de l'université Mohammed V de Rabat, pour son soutien, sa patience, sa grande disponibilité, pour ses qualités humaines et scientifiques, ainsi que pour toute l'aide qu'elle m'a apportée tout au long de ce travail.

Je tiens également à exprimer ma profonde gratitude à mon co-encadrant, Yassine ZAOUI SERGHOUCHENI, maître de conférences habilité à la faculté des sciences, université Mohammed V, Rabat, pour son soutien, ses encouragements et sa passion pour la recherche. Il s'est beaucoup impliqué dans la réalisation de ma thèse dans des conditions scientifiques optimales et dans une ambiance favorable.

J'exprime ma sincère gratitude à Mme Samira KHOULJI, professeur d'enseignement supérieur à l'école nationale des sciences appliquées de Tétouan, pour sa disponibilité et pour m'avoir fait l'honneur d'être la présidente de ma soutenance.

Je tiens à exprimer ma profonde gratitude à Monsieur Rachid SAADANE, professeur d'enseignement supérieur à l'école Hassania des travaux publics de casablanca, pour avoir accepté d'évaluer mon travail et pour m'avoir fait l'honneur d'être rapporteur.

Je tiens également à exprimer ma profonde gratitude à M. Nassim KHARMOUM, maître de conférences habilité à l'université Mohammed V, centre national de la recherche scientifique et technique, pour avoir accepté d'évaluer mon travail et pour m'avoir honoré en tant que rapporteur.

Je tiens à remercier chaleureusement M. Hicham GUEDDAH, maître de conférences habilité à l'école normale supérieure de Rabat, pour avoir rapporté et étudié ce rapport avec beaucoup de rigueur et d'intérêt.

Je tiens à exprimer ma gratitude à M. Wajih RHALEM, maître de conférences à l'école nationale supérieure d'Arts et Métiers de Rabat, pour l'honneur qu'il m'a fait en acceptant de participer en tant qu'invité à cette soutenance. Sa présence témoigne de l'intérêt qu'il porte à ce travail et je lui en suis profondément reconnaissant.

J'exprime mes sincères remerciements à Monsieur Najib ALIDRISSI, professeur agrégé à l'Université Mohammed VI des Sciences et de la Santé de Rabat, qui m'a fait l'honneur d'accepter de participer en tant qu'invité lors de cette soutenance. Sa participation enrichit considérablement l'évaluation de ce travail et témoigne de l'attention bienveillante qu'il porte à mes recherches.

Enfin, je n'oublierai pas tous ceux qui ont contribué à l'accomplissement de ce travail.

Résumé

La découverte d'un médicament fait face à des défis importants, puisque seulement 10% des composés utilisés dans les essais cliniques obtiennent l'approbation réglementaire. Cette thèse intègre la prédiction des propriétés moléculaires (MPP) et la génération moléculaire (MG) en utilisant l'apprentissage profond pour transformer le pipeline de découverte de médicaments.

Nous développons deux nouveaux frameworks basés sur des graphes : D'abord, le Graph Molecular Property Prediction Neural Network (GMPP-NN) combine les réseaux neuronaux de passage de messages avec des classificateurs perceptron multicouches, démontrant une performance supérieure sur les ensembles de données de référence MoleculeNet (VIH, BACE, BBBP, ClinTox). Deuxièmement, le cadre ME&PP-MG&RC-DL intègre l'encodage moléculaire, la prédiction des propriétés, la génération et la classification de la réalité, permettant d'obtenir une précision exceptionnelle dans la prévision des propriétés chimiques quantiques tout en générant des structures moléculaires valides, uniques et nouvelles.

Nos approches organisent efficacement les représentations spatiales latentes, facilitant l'exploration chimique ciblée de l'espace et fournissant des informations sur les relations structure-propriété. L'intégration du MPP et de la MG représente un changement de paradigme dans la découverte de médicaments par ordinateur, offrant des outils pour naviguer dans l'espace chimique et identifier les candidats thérapeutiques avec une efficacité sans précédent, accélérant potentiellement la découverte de médicaments et réduisant les coûts de développement.

Keywords : Découverte de médicaments, intelligence artificielle, apprentissage profond, réseaux neuronaux graphiques, prédiction des propriétés moléculaires, génération moléculaire, réseaux neuronaux de messages passant, encodeur automatique variationnel, chimie computationnelle, recherche pharmaceutique

Abstract

Drug discovery faces significant challenges, with only 10% of compounds used in clinical trials gaining regulatory approval. This thesis integrates molecular property prediction (MPP) and molecular generation (MG) using deep learning to transform the drug discovery pipeline.

We are developing two new graph-based frameworks : First, the Graph Molecular Property Prediction Neural Network (GMPP-NN) combines neural message-passing networks with multilayer perceptron classifiers, demonstrating superior performance on MoleculeNet reference datasets (HIV, BACE, BBBP, ClinTox). Secondly, the ME&PP-MG&RC-DL framework integrates molecular encoding, property prediction, reality generation and classification, enabling exceptional accuracy in predicting quantum chemical properties while generating valid, unique and novel molecular structures.

Our approaches efficiently organize latent spatial representations, facilitating targeted chemical space exploration and providing information on structure-property relationships. The integration of MPP and MG represents a paradigm shift in computational drug discovery, offering tools to navigate chemical space and identify therapeutic candidates with unprecedented efficiency, potentially accelerating drug discovery and reducing development costs.

Keywords : Drug discovery, artificial intelligence, deep learning, graph neural networks, molecular property prediction, de novo molecular generation, message-passing neural networks, variational autoencoders, computational chemistry, pharmaceutical research

Liste des abréviations

Abbréviations	Définitions
ADME	Absorption, Distribution, Metabolism, and Excretion
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
AI	Artificial Intelligence
AUC	Area Under Curve
BACE	Beta-secretase 1
BBBP	Blood-Brain Barrier Penetration
CADD	Computer-Aided Drug Design
CM	Coulomb Matrix
CNN	Convolutional Neural Network
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNN	Deep Neural Network
DNA	Deoxyribonucleic Acid
ECFP	Extended Connectivity Fingerprints
EMA	European Medicines Agency
FBDD	Fragment-Based Drug Design
FDA	Food and Drug Administration
FN	False Negatives
FP	False Positives
FP2VEC	Fingerprint to Vector
FPR	False Positive Rate
GAN	réseaux adversarial génératifs
GCN	Réseaux convolutifs de graphes
GMPP-NN	Graph Molecular Property Prediction Neural Network
GNN	Graph Neural Network
GVAE	autoencodeurs variationnels de graphe
HIV	Human Immunodeficiency Virus
HOMO	Highest Occupied Molecular Orbital
HTS	High-Throughput Screening
IND	Investigational New Drug
LBDD	Ligand-Based Drug Design
LBVS	Ligand-Based Virtual Screening
LSTM	Long Short-Term Memory
LUMO	Lowest Unoccupied Molecular Orbital
MAE	Mean Absolute Error
MG	Molecular Generation
ML	Machine Learning
MLP	Multi-Layer Perceptron

MPNN	réseaux de neurones à passage de messages
MPP	Molecular Property Prediction
NMR	Nuclear Magnetic Resonance
PRC	Precision-Recall Curve
QSAR	Quantitative Structure-Activity Relationship
RDKit	Open-Source Cheminformatics Software
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SBDD	Structure-Based Drug Design
SBVS	Structure-Based Virtual Screening
SMILES	Simplified Molecular Input Line Entry System
SPR	Surface Plasmon Resonance
ST	SMILES Transformer
TN	True Negatives
TP	True Positives
TPF	True Positive Fraction
TPR	True Positive Rate
VAE	autoencodeurs variationnels
VS	Virtual Screening
WGAN	Wasserstein réseaux adversarial génératifs
WHO	World Health Organization

Table des figures

Liste des tableaux

Table des matières

Introduction Générale

Le processus de découverte des médicaments se trouve à un moment critique, face à des défis sans précédent et pourtant prêt pour une innovation extraordinaire. Les approches traditionnelles de la mise au point des médicaments caractérisées par des cycles séquentiels de criblage , d'optimisation et de test se sont révélées de plus en plus inadéquates pour répondre à la demande croissante de nouveaux traitements thérapeutiques. Le délai moyen de mise au point est de plus d'une décennie et les coûts dépassent 2,6 milliards de dollars par médicament approuvé. Cette crise de l'innovation pharmaceutique est encore aggravée par les taux élevés d'attrition, environ 90% des composés entrant dans les essais cliniques n'ayant pas obtenu l'approbation réglementaire. Dans ce contexte, l'intégration de l'intelligence artificielle, en particulier des méthodologies d'apprentissage profond, est apparue comme une force transformatrice, offrir des solutions potentielles à ces défis depuis longtemps [1, 2].

L'idée principale de cette recherche est que l'intégration synergique de la prédiction des propriétés moléculaires (MPP) et de la génération moléculaire (MG) à l'aide de modèles d'apprentissage profond peut transformer fondamentalement le processus de découverte de médicaments, permettre une exploration plus efficace de l'espace chimique et accélérer l'identification des candidats prometteurs. Cette approche intégrée représente un changement de paradigme par rapport aux méthodologies traditionnelles, où la prédiction des propriétés et la conception des composés étaient souvent traitées comme des processus séparés et séquentiels. En unifiant ces composants au sein de cadres informatiques cohérents, nous pouvons simultanément naviguer dans le vaste espace chimique tout en ciblant précisément les composés avec les profils thérapeutiques souhaités [3, 4].

La prédiction des propriétés moléculaires est un élément essentiel de la découverte précoce d'un médicament, ce qui permet aux chercheurs de prévoir l'activité biologique, les propriétés pharmacocinétiques et la toxicité potentielle d'un composé sans avoir recours à une validation expérimentale qui nécessite beaucoup de ressources. En même temps, la génération moléculaire facilite l'exploration systématique de l'espace chimique, créant des structures nouvelles qui peuvent offrir des caractéristiques thérapeutiques améliorées. Lorsque ces capacités sont combinées efficacement par le biais d'architectures de deep learning sophistiquées, elles créent une framework puissante pour la découverte de médicaments accélérée et axée sur les données [5, 6].

Notre recherche s'appuie sur des avancées significatives dans les approches d'apprentissage profond basées sur des graphes, qui ont démontré une efficacité remarquable dans la capture de l'information structurale et chimique complexe codée dans les graphes moléculaires. En représentant les molécules sous forme de graphes avec des atomes comme noeuds et des liaisons chimiques comme arêtes, ces modèles préservent les relations topologiques essentielles à la compréhension du comportement moléculaire tout en permettant un traitement informatique efficace. Cette représentation graphique constitue la base de nos nouvelles architectures pour la prédiction et la génération des propriétés moléculaires.

Les contributions principales de cette thèse englobent deux cadres innovants d'apprentis-

sage profond : le réseau neuronal GMPP-NN (Graph Molecular Property Prediction Neural Network) pour une prédiction précise des propriétés, et l'encodage moléculaire et la prédiction des propriétés Génération moléculaire et classification de la réalité Deep Learning (ME&PP-MG&RC-DL) pour la prédiction intégrée des propriétés et la génération moléculaire. Ces architectures répondent à des défis critiques dans la découverte de médicaments par calcul, démontrant une performance supérieure par rapport aux méthodes existantes sur plusieurs ensembles de données de référence.

Dans le chapitre 1, nous présentons les concepts fondamentaux de la découverte de médicaments et établissons l'importance des approches computationnelles pour relever ses défis inhérents. Nous examinons l'évolution du processus de développement des médicaments et mettons en évidence le potentiel de la prédiction des propriétés moléculaires et de la génération moléculaire pour révolutionner ce domaine. L'intégration de ces approches est présentée comme une stratégie prometteuse pour améliorer l'efficacité et la rentabilité des premiers stades de la découverte de médicaments.

Le chapitre 2 donne un aperçu complet des méthodes traditionnelles de découverte de médicaments, y compris les techniques de calcul comme le criblage virtuel, le criblage à haut débit et la conception de médicaments basée sur la structure. Ce contexte historique établit les fondements sur lesquels reposent nos approches d'apprentissage profond, illustrant à la fois les forces et les limites des méthodologies conventionnelles. L'émergence de l'intelligence artificielle dans la découverte de médicaments est examinée, avec une attention particulière à son potentiel pour transformer plusieurs aspects du pipeline de développement.

Dans le chapitre 3, nous approfondissons les fondements théoriques des approches basées sur des graphes d'apprentissage profond pour la prédiction de propriétés moléculaires et la génération moléculaire. Nous explorons la représentation des graphes moléculaires, détaillons l'architecture et les mécanismes des réseaux neuronaux basés sur les graphes, et examinons des modèles spécifiques pour la prédiction de propriétés (réseaux de neurones à passage de messages s, Réseaux convolutifs de graphess) et la génération moléculaire (graphe autoencodeurs variationnels s, Réseaux antagonistes génératifs). Ce chapitre établit le cadre mathématique et informatique qui sous-tend nos contributions à la recherche.

Le chapitre 4 présente notre première contribution majeure : l'architecture GMPP-NN pour la prédiction des propriétés moléculaires. Ce modèle tire parti des réseaux de neurones de passage de messages pour capturer efficacement les structures moléculaires et prédire une gamme de propriétés critiques pour la découverte de médicaments. Grâce à une évaluation rigoureuse sur plusieurs ensembles de données MoleculeNet (VIH, BACE, BBBP et ClinTox), nous démontrons la performance supérieure de notre approche par rapport aux méthodes existantes, en atteignant des métriques exceptionnelles ROC-AUC et PRC-AUC.

Le chapitre 5 présente notre deuxième contribution significative : le cadre ME&PP-MG&RC-DL pour l'encodage moléculaire intégré, la prédiction de propriétés, la génération et la classification de la réalité. Cette architecture complète répond simultanément à de multiples défis dans la découverte de médicaments par calcul, permettant une prédiction précise des propriétés chimiques quantiques tout en générant des structures moléculaires diverses, valides et nouvelles. L'incorporation d'un classificateur de la réalité garantit que les molécules générées ressemblent étroitement à des entités chimiques réalisables, ce qui améliore leur utilité pratique dans le développement de médicaments.

Enfin, le chapitre 6 fait la synthèse de nos contributions, examine les limites des approches actuelles et explore les orientations et perspectives futures pour l'intégration de la prédiction des propriétés moléculaires et de la génération moléculaire dans la découverte de médicaments. Nous examinons les implications scientifiques, technologiques, économiques et éthiques de ces

approches intégrées, fournissant ainsi une feuille de route pour la recherche et le développement futurs dans ce domaine en évolution rapide.

Cette thèse représente une étape importante vers la réalisation du plein potentiel de l'intelligence artificielle dans la découverte de médicaments. En intégrant la prédiction des propriétés moléculaires et la génération moléculaire grâce à des architectures d'apprentissage profond avancées, nous fournissons de nouveaux outils puissants pour naviguer dans le vaste espace chimique et identifier les candidats thérapeutiques prometteurs avec une efficacité sans précédent. Bien que les défis demeurent, les cadres présentés dans cette recherche offrent des preuves convaincantes de l'impact transformateur de l'apprentissage profond sur l'avenir de l'innovation pharmaceutique.

Chapitre 1

Accélération du processus de découverte de médicaments

Introduction

La découverte et le développement de médicaments sont un processus itératif à multiples facettes, qui comprend l'identification et la validation des cibles potentielles des médicaments, la conception et la synthèse des composés de lead, ainsi que l'évaluation de leur efficacité et de leur innocuité dans les milieux précliniques et cliniques. Ce processus peut prendre plus d'une décennie, avec de nombreux revers et échecs en cours de route. Cependant, les progrès récents des techniques de calcul et des algorithmes d'apprentissage automatique ont transformé le domaine de la découverte de médicaments, offrant de nouvelles opportunités pour l'innovation et l'efficacité [7–9].

La prédiction des propriétés moléculaires (MPP) et la génération moléculaire (MG) sont deux approches computationnelles clés qui ont révolutionné le domaine de la découverte de médicaments. MPP permet aux chercheurs de prédire l'activité biologique, la sélectivité et les propriétés pharmacocinétiques des composés, ce qui facilite la prise de décisions éclairées dans les premiers stades du développement d'un médicament. MG, d'autre part, permet la conception et la génération de nouvelles structures moléculaires avec des propriétés souhaitées, élargissant l'espace chimique exploré lors de la découverte de médicaments. En tirant parti des algorithmes avancés et des techniques d'apprentissage automatique, MPP et MG ont considérablement amélioré l'efficacité du processus de découverte de médicaments [10–13].

L'intégration de la MPP et de la MG a transformé la façon dont les chercheurs abordent l'identification et l'optimisation des candidats-médicaments. En combinant ces approches, les chercheurs peuvent rapidement explorer l'espace chimique, identifier des composés avec des profils optimaux et prédire leurs propriétés in vivo. Ce processus itératif permet une réponse et une optimisation en temps réel, simplifiant le processus de découverte des médicaments et réduisant le risque d'échecs coûteux. Par conséquent, l'intégration de la MPP et de la MG est devenue une composante essentielle de la découverte moderne des médicaments, offrant des possibilités sans précédent d'innovation et de réussite [14–17].

Dans la prochaine section de notre chapitre, nous approfondirons les concepts de MPP et de MG, en explorant leurs fondements théoriques, leurs implémentations algorithmiques et leurs applications dans le contexte de la découverte de médicaments. Nous examinerons l'état actuel de la technique dans ces domaines, en mettant en évidence les progrès récents et les orientations futures pour la recherche et le développement. En donnant un aperçu complet des MPP et des MG, nous visons à illustrer leur potentiel de transformation dans le domaine de la découverte

de médicaments et à souligner leur importance dans le développement d'agents thérapeutiques nouveaux.

1.1 Développement et découverte de médicaments

Avant qu'un médicament puisse atteindre un patient, il doit subir des tests rigoureux pour déterminer sa sécurité, son efficacité dans le traitement de la maladie cible et la dose et la voie d'administration correctes. Les autorités de réglementation pharmaceutique surveillent et réglementent les produits thérapeutiques, y compris les médicaments sur ordonnance et en vente libre, les vaccins, les thérapies cellulaires et les dispositifs médicaux. Ils jouent un rôle clé tout au long du processus de développement des médicaments pour assurer la sécurité, l'efficacité, l'accessibilité et la sécurité des médicaments approuvés [18, 19].

1.1.1 Les étapes du développement d'un médicament

Les étapes de développement d'un médicament impliquent un processus complexe et itératif avec plusieurs phases, chacune ayant ses propres objectifs [20]. comme le montre la figure ci-dessous 1.1 les étapes comprennent :

— Processus de découverte de médicaments
— Recherche préclinique
— Recherche clinique
— Approbation et surveillance de la sécurité après commercialisation

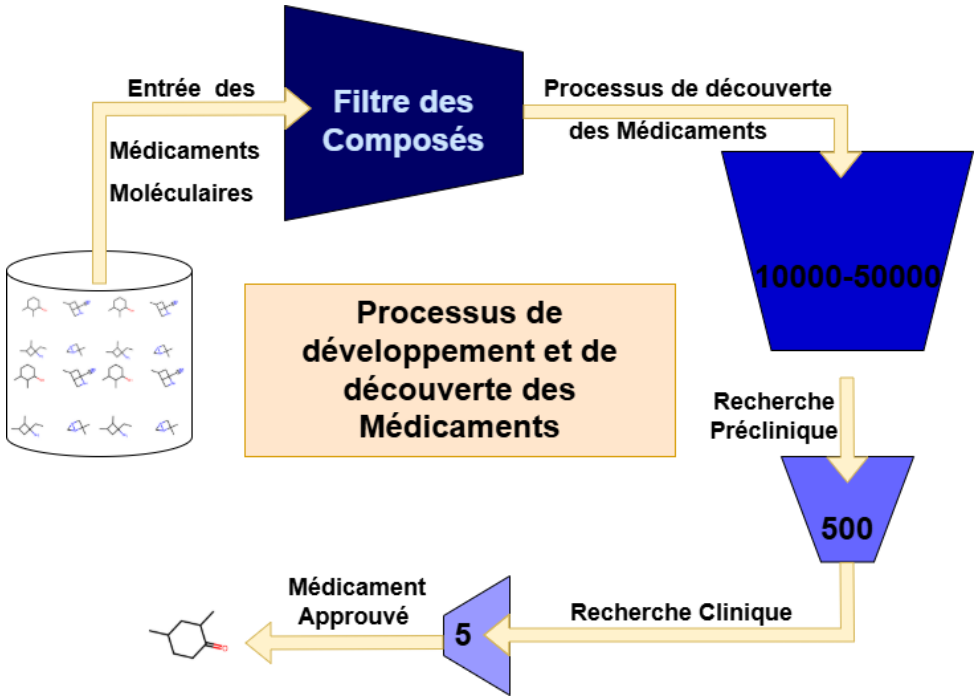


FIGURE 1.1 – Phases de développement des médicaments moléculaires

Le processus commence par la découverte précoce de médicaments, où les chercheurs identifient et valident des cibles potentielles de médicaments et développent des composés lead en

utilisant des technologies avancées telles que le criblage à haut débit et les algorithmes d'apprentissage automatique. Ceci est suivi par la recherche préclinique, où les composés lead sont testés in vitro et in vivo pour évaluer leur efficacité et leur sécurité en utilisant des techniques telles que des essais cellulaires et des modèles animaux. Une fois qu'un composé lead est identifié, une demande de médicament expérimental nouveau (IND) est soumise aux autorités réglementaires, ce qui permet l'initiation de la recherche clinique [21–23].

La recherche clinique consiste à tester le médicament sur des sujets humains, généralement en une série de phases (I-IV), afin d'évaluer son sécurité et son efficacité. Les essais de phase I se concentrent sur l'évaluation de la sécurité et de la pharmacocinétique du médicament, tandis que les essais de phase II évaluent son efficacité et le dosage optimal. Les essais de phase III sont des essais multicentriques plus importants qui confirment l'efficacité et la sécurité du médicament dans une population plus importante. Enfin, les essais de phase IV sont des études postcommercialisation qui surveillent la sécurité et l'efficacité à long terme du médicament [24–27].

La surveillance de la sécurité après commercialisation et l'approbation réglementaire comprennent la soumission des données d'essais cliniques aux autorités réglementaires pour examen et approbation, ainsi que la surveillance continue de la sécurité et de l'efficacité du médicament dans la phase postcommercialisation. Cela comprend le suivi des rapports d'événements indésirables, la production de rapports périodiques sur les mises à jour de sécurité et la mise en œuvre de stratégies de gestion des risques [28, 29].

1.1.2 Processus de découverte des médicaments

Le processus de découverte des médicaments vise à identifier et à optimiser les candidats potentiels aux médicaments. Cette phase comme en témoigne la figure ci-jointe 1.2 comprend l'identification et la validation des cibles, l'identification des hits et l'optimisation des leads. Chacune de ces étapes joue un rôle crucial dans la progression des candidats à l'obtention d'un médicament tout au long du processus de développement [32].

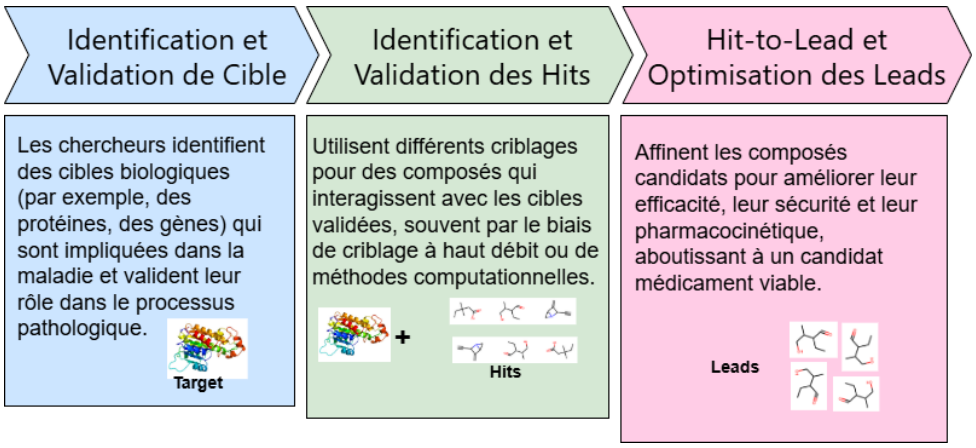


FIGURE 1.2 – Processus de découverte des médicaments

Identification et validation de la cible

La détermination et la validation de la cible sont les premières étapes de la phase de découverte précoce des médicaments. Les chercheurs se concentrent sur l'identification des cibles biologiques associées à des maladies spécifiques, ce qui implique une recherche approfondie sur les mécanismes sous-jacents de la maladie et les voies biologiques pouvant être modulées

par les agents thérapeutiques. La sélection des cibles appropriées est essentielle, car le succès des phases suivantes dépend de cette décision. Les progrès récents en génomique, protéomique et biologie des systèmes ont fourni aux chercheurs une mine de données qui peuvent être exploitée pour identifier plus efficacement les cibles potentielles des médicaments. Les technologies de criblage à haut débit permettent l'évaluation rapide de milliers de composés par rapport aux cibles sélectionnées, accélérant le processus d'identification des leads [33–35].

Identification des hits

Après l'identification et la validation de la cible, l'accent est mis sur l'identification des cibles, où les chercheurs capturent de grandes bibliothèques de composés pour trouver ceux qui présentent des interactions souhaitables avec la cible. Ce processus peut être à la fois long et coûteux, en particulier lorsqu'on s'appuie sur des méthodes expérimentales traditionnelles. Cependant, l'intégration d'approches informatiques, telles que le docking moléculaire et le criblage virtuel, a révolutionné cette phase [36, 37].

Optimisation des leads

Après l'identification des hits, l'accent est mis sur l'optimisation des leads, où les composés sélectionnés subissent des modifications pour améliorer leurs propriétés pharmacologiques. Cette phase est essentielle pour améliorer l'efficacité, la sécurité et la biodisponibilité des médicaments candidats. L'optimisation traditionnelle des leads implique souvent des cycles itératifs de synthèse et de test, ce qui peut être inefficace. Cependant, les progrès technologiques récents et les méthodes de calcul ont considérablement amélioré l'efficacité et la précision de l'optimisation des leads [38].

1.2 Prédiction des propriétés moléculaires et génération des molécules

Le processus de découverte de médicaments a été considérablement amélioré par l'intégration des techniques de calcul, en particulier dans les domaines de la prédiction des propriétés moléculaires et de la génération moléculaire 1.3. Ces méthodologies utilisent des algorithmes avancés et des techniques d'apprentissage automatique pour prédire et concevoir de nouvelles entités moléculaires, rationalisant ainsi le processus de développement de médicaments.

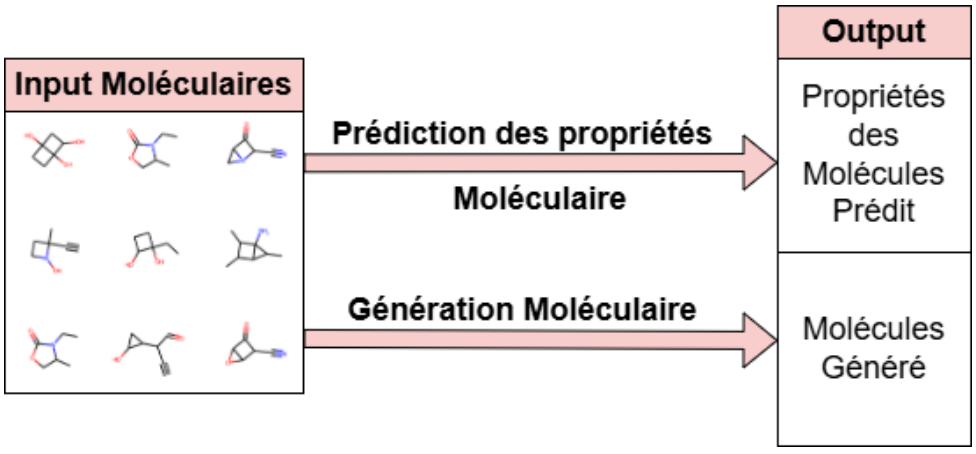


FIGURE 1.3 – Prédiction et génération de propriétés moléculaires

1.2.1 prédiction des propriétés moléculaires (MPP)

La prédiction des propriétés moléculaires (MPP) désigne les techniques de calcul utilisées pour estimer l'activité biologique, la toxicité et les propriétés pharmacocinétiques des médicaments candidats. Le principal objectif du MPP est de fournir aux chercheurs des renseignements sur la façon dont une structure moléculaire donnée peut se comporter biologiquement, aidant ainsi à identifier les candidats prometteurs en début de processus de découverte.

Les méthodes traditionnelles de MPP s'appuyaient fortement sur des modèles quantitatifs de relation structure-activité (QSAR), qui corrélent la structure chimique avec l'activité biologique. Les modèles QSAR utilisent des techniques statistiques pour établir des relations entre les descripteurs moléculaires et les réponses biologiques. Cependant, ces méthodes sont souvent limitées par leur dépendance à des descripteurs prédéfinis et la nécessité de disposer de données expérimentales étendues. Les progrès récents en apprentissage automatique et en apprentissage profond ont révolutionné le MPP en permettant le développement de modèles plus sophistiqués qui peuvent apprendre directement des représentations moléculaires sans avoir besoin d'une ingénierie étendue des fonctionnalités.

Les modèles d'apprentissage profond, tels que les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents (RNN), ont connu un succès remarquable dans la prédiction des propriétés moléculaires. Par exemple, les CNN peuvent être utilisés pour analyser des graphes moléculaires ou des représentations 2D, en capturant des modèles complexes qui se rapportent à l'activité biologique. Les réseaux de mémoire à court terme (LSTM) sont efficaces pour le traitement des données séquentielles, ce qui les rend adaptés pour prédire les propriétés basées sur les représentations des molécules du système SMILES (Simplified Molecular Input Line Entry System).

En outre, l'arrivée de l'apprentissage par transfert a encore amélioré le MPP en permettant aux modèles formés sur de grands ensembles de données d'être mis au point pour des tâches spécifiques avec des données limitées. Cette approche s'est avérée bénéfique dans des scénarios où les données expérimentales sont rares, permettant aux chercheurs de tirer parti des connaissances existantes pour faire des prédictions précises pour de nouveaux composés.

En résumé, le MPP a évolué des modèles traditionnels de QSAR aux approches sophistiquées d'apprentissage automatique et d'apprentissage profond, améliorant considérablement la précision et l'efficacité de la prédiction des propriétés moléculaires. Cette avancée facilite non seulement la sélection de candidats à médicaments viables, mais réduit également le temps et les coûts associés à la validation expérimentale.

1.2.2 Génération moléculaire (MG)

La génération moléculaire (MG) désigne les techniques de calcul utilisées pour concevoir et créer de nouvelles structures moléculaires avec des propriétés souhaitées. Le but de MG est d'élargir l'espace chimique exploré lors de la découverte de médicaments, permettant ainsi aux chercheurs d'identifier de nouveaux composés qui peuvent présenter un potentiel thérapeutique.

Historiquement, la génération moléculaire reposait sur la chimie combinatoire et la conception de médicaments à base de structure, où les chercheurs modifiaient systématiquement des composés existants pour créer de nouveaux analogues. Cependant, ces approches ont souvent rencontré des défis en termes d'efficacité et de créativité, car elles étaient limitées aux échafaudages chimiques connus. L'émergence de modèles génératifs, notamment ceux basés sur le deep learning, a transformé la MG en permettant la conception de structures moléculaires de novo [10].

réseaux adversarial génératifs s (GANs) et autoencodeurs variationnels s (VAEs) sont deux architectures d'apprentissage profond utilisées pour la génération moléculaire. Les GANs sont constitués de deux réseaux neuronaux, un générateur et un discriminateur, qui travaillent pour créer de nouveaux échantillons de données. Dans le contexte de la MG, le générateur crée des structures moléculaires nouvelles, tandis que le discriminateur évalue leur validité en se basant sur les modèles appris à partir des données existantes [30, 31]. Ce processus d'entraînement contradictoire permet aux GAN de produire des candidats moléculaires de haute qualité qui sont structurellement divers et potentiellement bioactifs.

VAEs, est conçu pour apprendre une représentation latente des données d'entrée, permettant la génération de nouveaux échantillons par échantillonnage à partir de cet espace latent. Le VAE a été appliqué avec succès pour générer des structures moléculaires en codant des représentations moléculaires et en les décodant en structures chimiques valides [12]. La capacité de contrôler l'espace latent permet également aux chercheurs d'explorer des régions spécifiques associées à des propriétés souhaitées, comme une puissance accrue ou une toxicité réduite.

L'apprentissage par renforcement (RL) est apparu comme un outil puissant pour la génération moléculaire. Les algorithmes RL permettent d'optimiser le processus de génération en définissant une fonction de récompense basée sur des propriétés souhaitées, guidant ainsi le modèle pour produire des composés répondant à des critères spécifiques [17]. Cette approche s'est révélée prometteuse pour la production de composés aux profils pharmacologiques améliorés, ce qui a permis d'accroître encore le potentiel du MG dans la découverte de médicaments.

La génération moléculaire a évolué des méthodes traditionnelles aux techniques avancées d'apprentissage profond, permettant la conception de nouvelles entités moléculaires avec des propriétés adaptées. L'intégration de modèles génératifs dans le processus de découverte de médicaments non seulement élargit l'espace chimique, mais augmente également la probabilité d'identifier des candidats prometteurs.

1.3 Intégration de MPP et MG dans les phases de découverte des médicament

L'intégration de la prédiction des propriétés moléculaires (MPP) et de la génération moléculaire (MG) dans le processus de découverte de médicaments omme le montre la figure ci-dessous 1.4 a considérablement transformé la façon dont les chercheurs identifient et optimisent les candidats à un médicament. Ces méthodologies peuvent être appliquées à diverses phases de la découverte de médicaments, ce qui améliore l'efficience et l'efficacité du processus global.

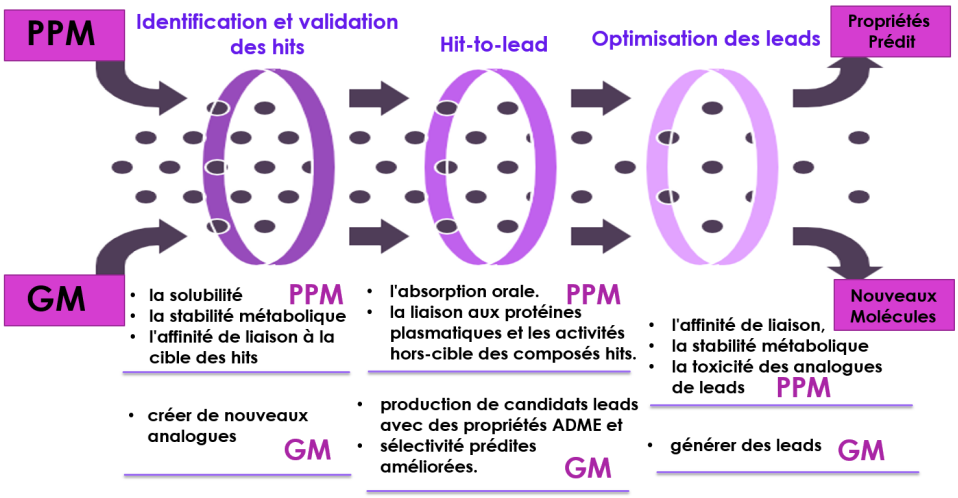


FIGURE 1.4 – Intégration de MPP et MG dans le Processus

Au cours de la phase initiale de découverte des médicaments, le MPP joue un rôle crucial dans l'identification et la validation des cibles. En prédisant l'activité biologique des composés par rapport à des cibles spécifiques, les chercheurs peuvent établir la priorité des cibles à poursuivre en fonction de leur potentiel d'intervention thérapeutique. Par exemple, les modèles MPP peuvent identifier des composés qui présentent une affinité de liaison élevée pour une protéine cible, guidant ainsi les chercheurs dans la sélection des candidats les plus prometteurs pour une étude plus approfondie [9]. De plus, le MPP peut aider à évaluer la sélectivité, permettant aux chercheurs d'éviter les composés qui peuvent interagir avec des cibles extérieures, diminuant ainsi le risque d'effets indésirables.

Après l'identification des MG, on peut utiliser ces derniers pour concevoir de nouveaux composés qui répondent à des défis spécifiques identifiés au cours de la phase de découverte précoce. Par exemple, si un composé lead présente une faible solubilité, les techniques de MG peuvent être utilisées pour produire des analogues avec des propriétés physico-chimiques améliorées. En tirant parti des modèles d'apprentissage profond, les chercheurs peuvent explorer l'espace chimique et identifier des composés qui maintiennent l'activité biologique souhaitée tout en améliorant la solubilité [13].

Dans la phase préclinique, le MPP peut aider à prédire les propriétés pharmacocinétiques des composés lead, telles que l'absorption, la distribution, le métabolisme et l'excrétion (ADME). Des prédictions précises de ces propriétés sont essentielles pour déterminer la viabilité des composés pour le développement clinique. Les modèles MPP peuvent fournir des informations sur le comportement in vivo d'un composé, permettant aux chercheurs de prendre des décisions éclairées quant aux candidats à faire l'objet d'essais cliniques [15].

Au cours de la recherche clinique, la MPP peut également jouer un rôle dans la stratification des patients. En prédisant la façon dont différentes populations peuvent répondre à un médicament sur la base de données génétiques et phénotypiques, les chercheurs peuvent adapter les essais cliniques à des groupes de patients spécifiques. Cette approche augmente les chances de succès des essais cliniques et permet d'élaborer des stratégies thérapeutiques plus personnalisées [16].

De plus, la combinaison de MPP et de MG peut faciliter l'optimisation itérative pendant la phase d'optimisation des pistes. En prédisant continuellement les propriétés des composés générés, les chercheurs peuvent affiner leurs conceptions en fonction de la rétroaction en temps réel des modèles MPP. Ce processus itératif permet l'exploration rapide de l'espace chimique et

l'identification des composés avec des profils optimaux pour un développement ultérieur [14].

L'application de la prédiction des propriétés moléculaires et de la génération moléculaire dans les phases de découverte de médicaments a révolutionné la façon dont les chercheurs identifient, optimisent et développent de nouveaux agents thérapeutiques. En tirant parti des techniques de calcul avancées, les chercheurs peuvent accélérer le processus de découverte de médicaments, réduire les coûts et, finalement, mettre sur le marché des traitements novateurs plus efficacement.

L'intégration de la prédiction des propriétés moléculaires et de la génération moléculaire dans la découverte de médicaments représente un progrès important dans le domaine de la recherche pharmaceutique. Ces méthodologies ont transformé la façon dont les chercheurs abordent l'identification et l'optimisation des médicaments candidats, permettant ainsi des processus de développement de médicaments plus efficaces. Les techniques de calcul continuent d'évoluer, et le potentiel des MPP et des MG pour améliorer encore la découverte de médicaments demeure prometteur. En exploitant la puissance de l'apprentissage automatique et de l'apprentissage profond, les chercheurs peuvent ouvrir de nouvelles possibilités d'innovation dans le développement d'agents thérapeutiques novateurs.

Conclusion

L'intégration de la prédiction des propriétés moléculaires (MPP) et de la génération moléculaire (MG) a considérablement avancé la recherche pharmaceutique, transformant fondamentalement l'identification, l'optimisation et le développement de nouveaux agents thérapeutiques. Le MPP améliore l'identification et la validation des cibles en prédisant l'activité biologique et la sélectivité des composés, ce qui permet de prendre des décisions éclairées dans les premières phases du développement d'un médicament. Par la suite, MG facilite la conception de nouveaux composés, en répondant à des défis spécifiques et en élargissant l'espace chimique exploré lors de la découverte de médicaments.

En outre, le MPP aide à prévoir des propriétés pharmacocinétiques, fournissant les analyses essentielles dans le comportement in vivo des composés leads, cruciaux pour le développement clinique. En recherche clinique, la MPP joue également un rôle essentiel dans la stratification des patients, permettant des stratégies de traitement personnalisées pour améliorer le succès des essais cliniques. Le processus d'optimisation itératif facilité par la combinaison de MPP et de MG permet une rétroaction en temps réel et une exploration rapide de l'espace chimique, permettant d'identifier des composés aux profils optimaux pour un développement ultérieur.

Les techniques de calcul continuent d'évoluer, et le potentiel des MPP et des MG pour améliorer encore la découverte de médicaments demeure prometteur. L'utilisation des capacités d'apprentissage automatique et d'apprentissage profond offre de nouvelles possibilités d'innovation dans le développement de nouveaux agents thérapeutiques. Cette conclusion souligne le potentiel de transformation de ces approches informatiques pour surmonter les défis de longue durée dans la découverte de médicaments et souligne leur rôle dans l'accélération de la prestation de traitements efficaces aux patients qui en ont besoin. L'avenir de la découverte de médicaments réside dans l'intégration continue des MPP, MG et d'autres méthodes, ouvrant ainsi la voie à une innovation pharmaceutique plus efficace, rentable et réussie.

Chapitre 2

Méthodes expérimental traditionnelles dans le processus de découverte de médicaments

Introduction

La découverte de médicaments représente l'un des efforts scientifiques les plus complexes en médecine moderne, impliquant l'identification systématique et le développement de nouveaux composés thérapeutiques. Ce processus multidisciplinaire - qui englobe l'identification et la validation des cibles, l'identification des succès, l'optimisation de l'impact sur les pistes, l'optimisation des pistes et les études de mise en place d'un nouveau médicament expérimental (IND)) - demeure extrêmement long et coûteux. Malgré les progrès méthodologiques importants réalisés pour imiter les mécanismes complexes de la maladie et améliorer les taux de réussite, les données statistiques indiquent que seulement environ 10% des composés entrant dans les essais cliniques sur l'homme obtiennent finalement l'approbation réglementaire. Ce taux d'attrition élevé souligne le besoin crucial de trouver des approches plus efficaces pour la découverte de médicaments.

L'environnement pharmaceutique a été transformé par l'émergence de méthodes de calcul qui réduisent considérablement le temps et les investissements financiers nécessaires au développement des médicaments. La conception de médicaments assistée par ordinateur (CADD) a révolutionné le domaine en réduisant l'espace chimique nécessitant une évaluation expérimentale, en utilisant des approches basées sur la structure qui tirent parti de l'information moléculaire cible ou du ligand-Stratégies basées sur des connaissances tirées de composés actifs connus [370]. Ces techniques de calcul complètent les méthodes traditionnelles de criblage à haut débit (HTS), malgré leur importance, qui sont limitées par des coûts élevés et des besoins importants en ressources. Le criblage virtuel, le criblage phénotypique et les approches fondées sur des fragments sont devenus de précieuses solutions de rechange, chacune offrant des avantages uniques pour relever des défis particuliers dans le processus de découverte de médicaments.

Le développement le plus prometteur des dernières années a peut-être été l'intégration de l'intelligence artificielle (IA) et des techniques d'apprentissage automatique avec les méthodologies établies de découverte de médicaments [371]. En incorporant des données dans un espace à haute dimension et en extrayant des relations clés, l'IA fournit des solutions novatrices à toutes les étapes de la découverte précoce d'un médicament, du criblage virtuel et de la conception de nouveaux composés à la prédiction des propriétés physico-chimiques [373]et optimisation des voies synthétiques. Les systèmes CRISPR-Cas, qui facilitent l'identification

des cibles, la génération de modèles de maladies et la validation des candidats [372], sont particulièrement prometteurs. Cette convergence technologique offre une voie pour découvrir et développer de nouveaux agents thérapeutiques plus rapidement, plus efficacement et avec des taux de réussite plus élevés que les approches conventionnelles, ce qui pourrait révolutionner la façon dont nous répondons aux besoins médicaux non satisfaits dans divers domaines de la maladie.

2.1 Méthodes de calcul pour la découverte de médicaments

La conception des médicaments est le processus le plus important dans l'industrie pharmaceutique. L'apparition de diverses méthodes de calcul a considérablement réduit le temps et le coût de la découverte de médicaments. La promotion du développement de méthodes informatiques sera une tendance inévitable dans le progrès des drogues. Sur la base de notre technologie et de notre plate-forme avancées, nous avons été équipés pour vous aider dans le criblage des médicaments et la conception de votre projet [374].

2.1.1 Contexte de la conception assistée par ordinateur des médicaments

La mise sur le marché des médicaments est un processus long et coûteux. Au cours des dernières années, les expériences de criblage à haut débit (HTS) ont joué un rôle important dans le processus de criblage des médicaments. Cependant, le HTS n'est pas seulement coûteux, mais il nécessite également une grande quantité de cibles et de ligands. En outre, le taux de réussite du HTS est généralement très faible. Pour ces raisons, le rôle des HTS dans la sélection de grandes bibliothèques de composés est considérablement limité. Au cours des dernières décennies, la conception assistée par ordinateur de médicaments (CADD) est devenue une stratégie valable qui peut réduire considérablement la gamme de composés requis pour le criblage . Les méthodes de calcul utilisant des stratégies de modélisation et de visualisation peuvent rapidement identifier les liants potentiels.

2.1.2 Conception de médicaments assistée par ordinateur

Dans le processus de découverte de médicaments, la DDMC est souvent appliquée de plusieurs façons. CADD peut réduire efficacement la bibliothèque de composés à grande échelle, qui pose les bases des opérations expérimentales. CADD guide également l'optimisation des composés de lead. Le plus important, c'est que les CADD peuvent être utilisés pour concevoir de nouveaux composés. Les méthodes CADD couramment utilisées peuvent être divisées en deux catégories : la conception de médicaments à base de structure (SBDD) et la conception de médicaments à base de ligand (LBDD). La méthode SBDD comprend le ligand docking, les méthodes de conception des ligands et le pharmacophore, qui doivent être basés sur la structure moléculaire cible. LBDD utilise uniquement des informations sur les ligands pour prédire l'activité en l'absence de la structure tridimensionnelle de la cible potentielle. Les outils LBDD comprennent la relation quantitative structure-activité (QSAR), la modélisation des pharmaphores et l'analyse moléculaire en champ. Il est intéressant de noter qu'au cours des dernières années, afin de résoudre le problème de l'absence d'information sur la structure cible et les ligands, Des méthodes basées sur la bioinformatique pour analyser et comparer plusieurs séquences ont été utilisées pour identifier des cibles potentielles à partir de zéro, comme le montre la figure 2.1.

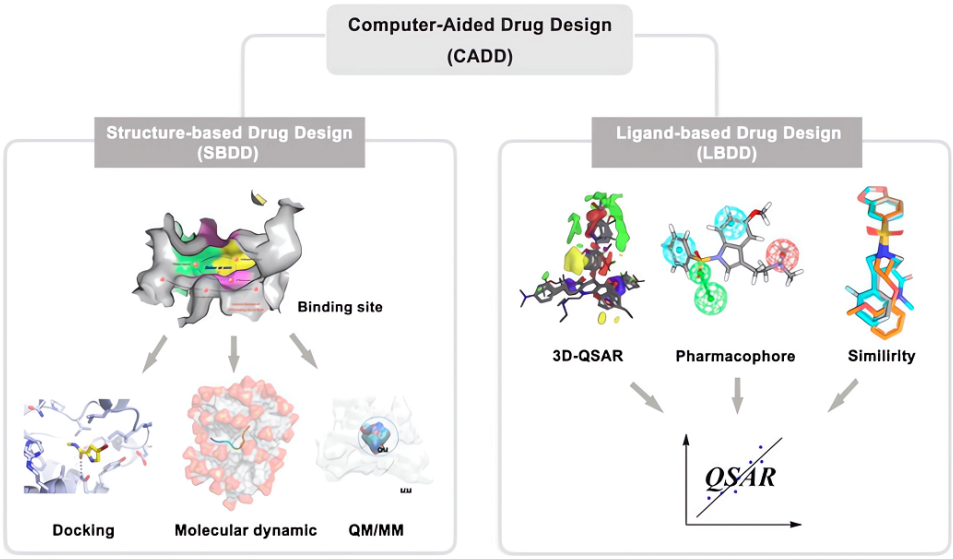


FIGURE 2.1 – Déscription CADD

2.1.3 Caractéristiques du CDAO

En tant que méthode permettant de réduire considérablement le nombre de composés devant faire l'objet d'une analyse expérimentale, la CADD permet également de réduire considérablement la charge de travail du criblage sans nuire à la découverte de clients potentiels. Dans le même temps, la CADD peut également augmenter le taux de réponse des nouveaux composés médicamenteux. En outre, l'application des outils de la DDCC réduit efficacement les coûts associés à l'exploration des médicaments et peut également réduire le temps nécessaire pour que les médicaments entrent sur le marché de consommation.

2.1.4 Applications des méthodes de calcul

Avec l'avancement rapide de la technologie informatique, le processus de découverte de médicaments bénéficie de diverses méthodes de calcul. Par exemple, la simulation biomoléculaire à plusieurs échelles permet d'identifier les sites de liaison des médicaments sur les macromolécules cibles et de clarifier le mécanisme d'action du médicament. Le criblage virtuel permet de rechercher efficacement les composés hits dans des bases de données chimiques massives. En outre, la conception de nouveau médicament fournit une autre méthode puissante pour concevoir des molécules de médicament à partir de zéro en utilisant les blocs de construction résumés et abstraits de découvertes réussies précédentes des médicaments. La mise au point de méthodes informatiques intégrées aidera à détecter les médicaments et à déterminer des thérapies efficaces avec de nouveaux mécanismes d'action, qui peuvent être appliqués à divers systèmes biologiques complexes.

2.2 Chimie médicale dans la découverte de médicaments

Inventer un nouveau médicament est un processus complexe, long, coûteux et risqué. La chimie médicinale est une combinaison de disciplines très interdépendantes telles que les sciences informatiques, la biochimie et la médecine humaine. Cette combinaison efficace jette les bases d'une préparation sûre et efficace des médicaments [375].

2.2.1 Chimie médicale

La discipline repose sur la chimie synthétique qui peut combiner de petites molécules pour créer de nouvelles molécules. Les tâches de la chimie médicinale comprennent l'étude de la structure, des propriétés et des lois changeantes des médicaments, ainsi que la compréhension des effets physiologiques et biochimiques des médicaments sur les humains. D'autres domaines scientifiques sont axés sur l'analyse et les essais moléculaires, tandis que la chimie médicale est axée sur la conception moléculaire. Les scientifiques peuvent également améliorer les médicaments existants en optimisant la structure des molécules. Ces ajustements peuvent faciliter l'action des médicaments, ce qui permet aux patients d'obtenir des résultats de traitement efficaces et de meilleure qualité. La création de nouveaux médicaments est un système d'ingénierie exploratoire impliquant plusieurs disciplines, et la découverte de précurseurs basés sur la chimie médicinale est le préalable à toute recherche ultérieure. Par conséquent, il occupe une position de leader dans le domaine des sciences pharmaceutiques comme dans la figure 2.2.

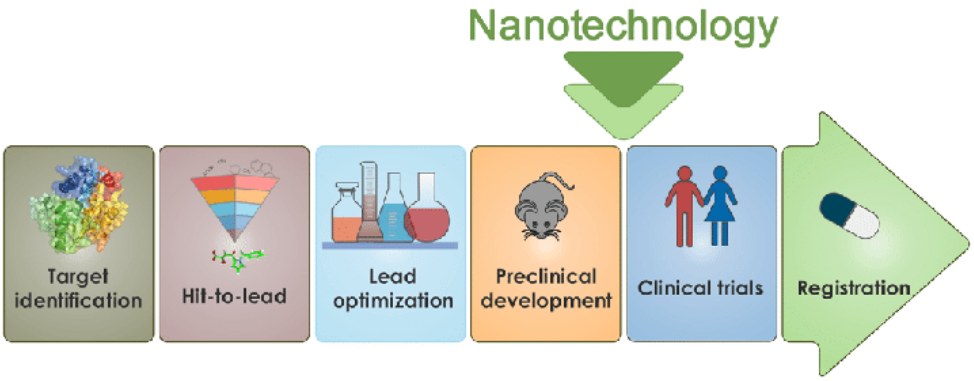


FIGURE 2.2 – Processus de développement des médicaments

2.2.2 Applications

La chimie médicinale est un résultat de fusion, dont la biologie, la technologie informatique et la chimie médicinale sont les principales composantes de ce domaine. Il joue un rôle clé dans la synthèse et la conception des médicaments. En fait, la chimie médicinale peut synthétiser de nouvelles entités pour s'assurer qu'elles sont appropriées pour le traitement. Il comprend également l'étude des aspects de synthèse et de calcul des médicaments ainsi que de leurs activités biologiques. Le plus important, c'est que ces technologies efficaces sont axées sur la qualité du médicament et visent à s'assurer qu'elles conviennent au domaine médical qui a jeté les bases de l'évolution des médicaments.

2.2.3 Méthodes communes

Le processus de découverte des médicaments est lié à une variété de technologies. La chimie médicinale englobe de nombreux domaines dans le processus de prospection des médicaments, comme les bibliothèques de criblage , la découverte du lead et la découverte de médicaments assistée par ordinateur. Le criblage à haut débit, l'identification des hits, l'optimisation du lead, la resynthesize chimique sont les méthodes de chimie médicinale les plus importantes pour la découverte du lead dans le cadre d'un docking composé, le criblage virtuel, la prédiction d'activité et la structure quantitative.La prédiction des relations d'activité est largement utilisée dans la découverte de médicaments assistée par ordinateur. En plus des technologies mentionnées

ci-dessus, de nouvelles technologies émergent progressivement. Dans un proche avenir, ce sujet deviendra de plus en plus abondant.

2.3 Méthodes de découverte des médicaments

La première étape de l'exploration des nouveaux médicaments consiste à analyser les nouvelles molécules issues de divers composés et produits naturels. La cristallographie par rayons X et la résonance magnétique nucléaire (RMN) sont les méthodes les plus efficaces et directes pour le développement rationnel de médicaments basés sur la structure. De plus, l'application du machine learning (ML) dans le domaine de la découverte de médicaments continue de croître, produisant des résultats passionnants. Il est intéressant de noter que les méthodes de criblage par ordinateur ont été continuellement explorées et améliorées, et peuvent maintenant être utilisées comme une alternative prometteuse et complémentaire au criblage biochimique à haut débit (HTS) comme dans la figure 2.3. La conception de médicaments par ordinateur est étroitement liée à plusieurs étapes de la découverte d'un médicament, comme l'utilisation du criblage virtuel pour l'identification des hits, l'optimisation de l'affinité et de la sélectivité des hits vers les leads et l'optimisation d'autres propriétés du médicament tout en maintenant l'affinité. [376].

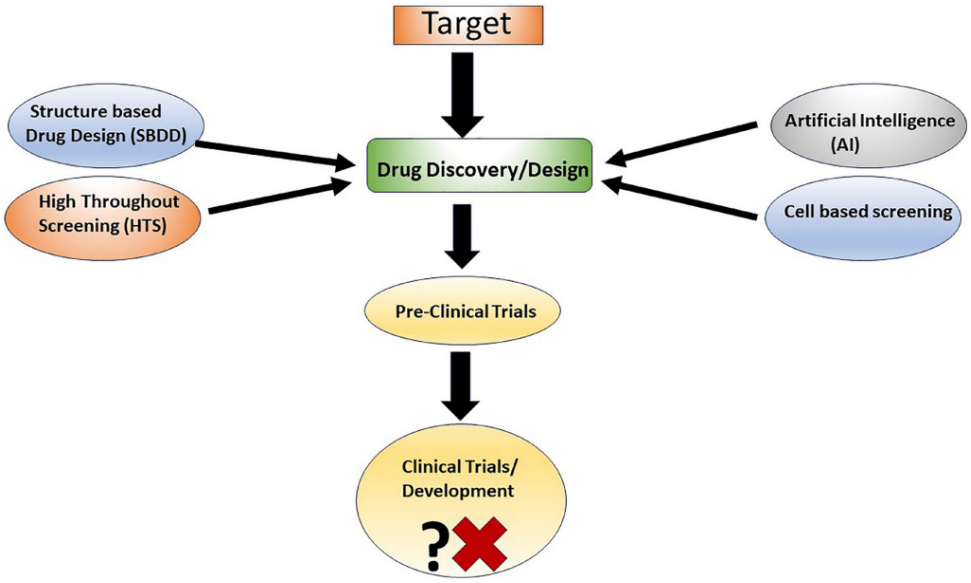


FIGURE 2.3 – Représentation schématique de la technologie de découverte des médicaments

2.3.1 Criblage virtuel

La découverte de médicaments désigne la recherche de petites molécules spécifiques qui interagissent avec des molécules plus grandes. La découverte de molécules biologiquement actives est un processus complexe et long. La principale technologie permettant d'identifier les nouveaux composés de lead dans la découverte de médicaments est le criblage à haut débit. Au cours des dernières années, une autre stratégie efficace de criblage virtuel (SV) est progressivement apparue [377].

VS peut compléter les cibles de structures connues par le criblage informatique des composés dans de grandes bibliothèques chimiques et la méthode VS a étendu les possibilités aux molécules. Au cours de la dernière décennie, de nombreuses méthodes VS différentes ont émergé

comme des moyens prometteurs pour trouver de nouveaux composés actifs pour de nombreuses cibles comme le montre la figure 2.4.

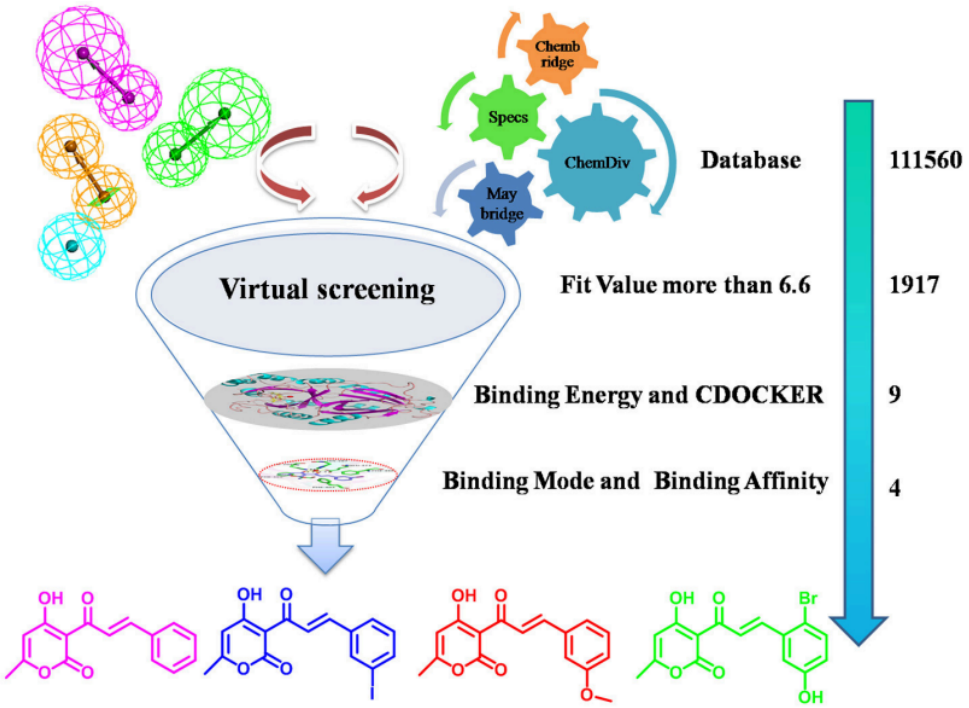


FIGURE 2.4 – Criblage virtuel pour la découverte de médicaments

Criblage virtuel basé sur la structure (SBVS) Le SBVS est basé sur la détermination des sites de liaison potentiels des ligands sur les molécules cibles. L'identification de la structure de la protéine cible au moyen de diverses méthodes telles que la RMN, la modélisation d'homologie ou la cristallographie par rayons X est une étape clé dans les SBVS.

Docking moléculaire Le docking moléculaire occupe une place importante dans le domaine du criblage et de la conception des médicaments. Cette technique est largement utilisée pour prédire l'interaction d'une protéine avec d'autres molécules afin d'évaluer la liaison entre deux molécules. Le docking moléculaire dépend principalement de la correspondance spatiale entre la forme et l'énergie du ligand et du récepteur.

Modélisation des pharmaphores Pharmacophore décrit les caractéristiques moléculaires des biomolécules nécessaires pour reconnaître les ligands. En général, lorsqu'une molécule de médicament interagit avec une molécule cible, elle produit une conformation active spécifique. Les différents groupes chimiques de la molécule du médicament ont des effets différents sur l'activité. Les changements de certains groupes ont une grande influence sur l'interaction entre le médicament et la cible, tandis que d'autres ont peu d'effet. En outre, il a été constaté que les molécules ayant la même activité ont tendance à avoir certaines des mêmes caractéristiques. Ces dernières années, avec le développement des bases de données composites et de la technologie informatique, il est progressivement devenu une tendance pour le modèle pharmacophore d'effectuer VS sur la base de données.

Relation quantitative structure-activité (QSAR) QSAR est étroitement lié à l'étude quantitative de l'interaction entre les petites molécules organiques et les grandes molécules biologiques. Lorsque la structure du récepteur est inconnue, la méthode QSAR est la plus rapide et la plus efficace. Bien qu'avec la détermination précise de la structure 3D de nombreuses macromolécules biologiques, la conception de médicaments basée sur la structure est progressivement devenue le courant dominant de la conception de médicaments, La faible charge de calcul et la

bonne capacité prédictive des QSAR jouent encore un rôle important dans la recherche sur les médicaments.

Criblage virtuel basé sur le ligand (LBVS) Dans le processus LBVS, la molécule de lead biologiquement active la plus efficace est détectée à l'aide d'une recherche de similarité structure ou pharmacodynamique. En outre, l'utilisation coopérative des SBVS et des SBVS peut augmenter la probabilité de trouver une nouvelle cible. L'intégration des deux stratégies a montré un grand potentiel pour identifier le premier agoniste sélectif du GPR30.

Machine Learning Techniques La technologie de l'apprentissage automatique joue un rôle de plus en plus important dans le domaine des SV. Un grand nombre d'algorithmes d'apprentissage automatique, y compris les partitions récursives, les réseaux neuronaux et les machines à vecteurs de support, ont été appliqués avec succès aux stratégies SV. Ces modèles permettent de classer les composés en fonction de leur probabilité d'être actifs, ce qui permet de réduire le nombre de composés redondants synthétisés. Dans cette méthode, les données expérimentales sont très utiles pour concevoir de nouveaux modèles composés.

Bibliothèques combinatoires virtuelles La bibliothèque combinatoire peut être indépendante de la cible, et peut également être conçue pour un pharmacopée spécifique. Cependant, dans de nombreux cas, le nombre de composés disponibles est trop grand pour être synthétisé physiquement. Une solution directe et efficace à ce problème est de concevoir une bibliothèque combinatoire virtuelle et d'appliquer des techniques appropriées pour filtrer de plus petits ensembles de composés dans la bibliothèque pour la synthèse physique. Par conséquent, la bibliothèque combinatoire virtuelle a reçu de plus en plus d'attention dans le processus de recherche de nouveaux médicaments.

2.3.2 Criblage à haut débit

Le criblage à haut débit (HTS) joue un rôle essentiel dans la découverte de médicaments. Avec l'avancement de l'industrie pharmaceutique, l'application des HTS dans la recherche fondamentale et appliquée a également reçu une attention croissante. À l'heure actuelle, le HTS est devenu une technologie mature qui est également la base de départ pour la découverte de médicaments comme décrit dans la figure 2.5.

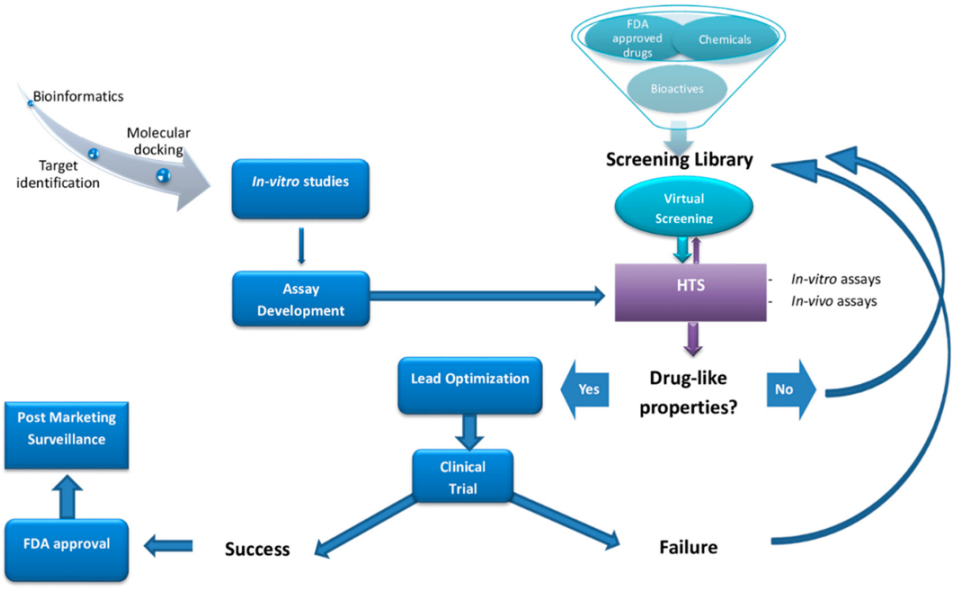


FIGURE 2.5 – Étapes du processus de découverte des médicaments

La découverte actuelle de médicaments repose sur un grand nombre de criblages de bibliothèques chimiques spécifiques pour trouver les nouvelles formes chimiques nécessaires. En fait, le succès d'un médicament repose sur un grand nombre de composés triés. Par conséquent, la découverte d'un médicament est un processus long et très complexe. Comme stratégie efficace, HTS peut être utilisé pour recueillir une grande quantité de données expérimentales dans un temps relativement court [378].

Introduction de HTS HTS est principalement un processus de criblage et d'analyse d'un grand nombre de composés biologiques pour des cibles spécifiques. Dans les méthodes modernes de découverte de médicaments, on définit habituellement le HTS comme l'analyse de 10 000 à 100 000 composés par jour. HTS convient pour le criblage de la chimie combinatoire, de la génomique, des protéines et des peptides. L'objectif principal de cette technologie est d'accélérer le processus de découverte des médicaments en dépistant de grandes bibliothèques de composés en peu de temps. HTS comprend plusieurs étapes, y compris l'identification de la cible, la gestion des composés, la préparation des réactifs, le développement analytique et le criblage des bibliothèques à haut débit.

Avantages de HTS HTS a les caractéristiques de rapide, à faible coût, simple et haute efficacité. De plus, le HTS est étroitement lié aux plates-formes d'exploitation automatisées, aux systèmes de détection à haute sensibilité, aux modèles de criblage spécifiques, aux bibliothèques de composants abondantes et aux systèmes de collecte et de traitement des données. Une variété de technologies, y compris la résonance magnétique nucléaire (RMN), le microréseau d'ADN, la fluorescence et d'autres nouvelles technologies ont le potentiel de filtrer plus de 100 000 échantillons par jour. Une caractéristique clé de la technologie RMN est qu'elle peut fournir des informations directes sur la position de liaison des composés et des protéines. Les microarrays d'ADN peuvent être utilisés dans le HTS pour explorer davantage l'expression des cibles biologiques liées aux maladies humaines, ouvrant ainsi de nouvelles voies pour la découverte de médicaments.

Applications et Développement HTS est largement utilisé dans la découverte de nouveaux médicaments, remplaçant les méthodes traditionnelles pour déterminer les cibles thérapeutiques. Lorsque la cible est peu connue, le HTS devient généralement la méthode appropriée, ce qui exclut la conception de médicaments à base de structure, mais il peut également être utilisé en parallèle avec d'autres stratégies, notamment les techniques de calcul et la conception de médicaments à base de fragments. En outre, la combinaison de HTS et des plates-formes liées aux cellules poreuses peut identifier quelques petits modulateurs de molécule. Plus important encore, la HTS facilite non seulement la découverte de médicaments, mais elle aide aussi à explorer les composants existants des médicaments afin d'optimiser leur activité. À l'heure actuelle, les technologies basées sur la fluorescence sont susceptibles d'être l'une des méthodes de détection les plus importantes pour le HTS car elles possèdent une sensibilité élevée et permettent la miniaturisation.

2.3.3 Criblage phénotypique

La recherche de médicaments efficaces et sûrs est le principal objectif de la recherche sur les médicaments. Le filtrage de composés multiples ayant des structures chimiques différentes et un potentiel clinique plus élevé, ainsi que l'optimisation des composés de lead, accéléreront considérablement le processus d'exploration des médicaments. Le criblage phénotypique est progressivement apparu dans la découverte de médicaments [379].

criblage des médicaments phénotypiques Le criblage phénotypique est une méthode basée sur les changements dans le phénotype des organismes. Les phénotypes biologiques sont les

caractéristiques de gènes spécifiques sous l'influence de l'environnement. Le criblage phénotypique traditionnel est la méthode standard pour l'invention d'un nouveau médicament. Il filtre principalement les composés qui peuvent modifier le phénotype dans les modèles de maladies animales, puis étudie les cibles et les mécanismes des composés. Plus tard, une technologie moderne de criblage phénotypique cellulaire est apparue. Cette technologie est principalement utilisée pour le criblage de composés pouvant provoquer des changements physiologiques souhaités dans les cellules, ce qui permet de découvrir de nouvelles protéines et de nouvelles cibles 2.7.

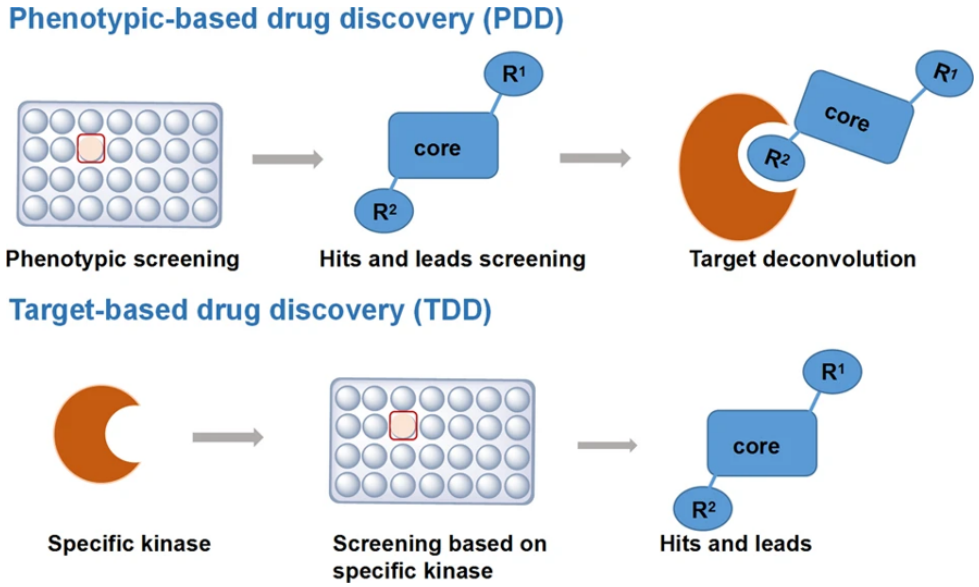


FIGURE 2.6 – Essais moléculaires basés sur le phénotype et sur la cible

Caractéristiques du criblage phénotypique L'exploration de médicaments par la technologie de criblage des composés à base de phénotype est réalisée grâce à des changements dans des systèmes biologiques complexes ou des voies de signalisation par des composés, et c'est un moyen efficace de sonder les composés qui agissent sur de nouvelles cibles. L'utilisation de cette stratégie simple contribue non seulement à augmenter la probabilité d'une détection précoce des médicaments, mais aussi à améliorer le taux de réussite du développement tardif des médicaments. La technologie de criblage phénotypique moderne sera plus pertinente pour l'étude des médicaments contre les maladies rares et des nouveaux médicaments ciblés. De plus, cette technologie est peu affectée par la compréhension moléculaire. Par conséquent, cette caractéristique peut être appliquée à la découverte de nouveaux médicaments dans des maladies dont le mécanisme sous-jacent n'est pas bien compris, comme les maladies psychosomatiques courantes ou les maladies neurologiques. À l'heure actuelle, certaines statistiques de recherche montrent que le nombre de médicaments découverts par les méthodes modernes de criblage phénotypique a dépassé le nombre découvert sur la base des cibles moléculaires. Par conséquent, le criblage phénotypique moderne deviendra une approche de prospection prometteuse.

Applications du criblage phénotypique Compte tenu des avantages ci-dessus, le criblage phénotypique a été utilisé dans l'exploration d'une variété de médicaments. Par exemple, la digoxine de la guêtre de renard, la morphine du coquelicot. Cela a mené à l'utilisation du criblage phénotypique animal et microbien pour isoler la plupart des antibiotiques et de nombreux autres composés utilisés en clinique aujourd'hui. En outre, les composés obtenus par le criblage phénotypique de produits naturels ont permis la reconnaissance des récepteurs opioïdes, de la transpeptidase et de nombreux autres enzymes et transporteurs. Les exemples récents de médi-

caments approuvés découverts par le criblage phénotypique démontrent davantage le potentiel de cette technologie pour trouver des thérapies ciblées hautement sélectives. Ces exemples comprennent des médicaments qui ciblent les régulateurs de fonctions cellulaires clés exprimés de façon omniprésente.

2.3.4 Conception de médicaments basée sur la structure

La mise au point de médicaments a toujours été reconnue comme une tâche longue et laborieuse parce que le processus implique la connaissance de nombreuses disciplines différentes. Une variété de méthodes de conception des médicaments sera bénéfique pour la conception et la recherche de molécules médicamenteuses raisonnables, accélérant ainsi tout le processus de découverte des médicaments. Au cours des dernières années, la science et la technologie avancées ont accéléré l'identification de structures protéiques tridimensionnelles, et les informations génétiques sont devenues plus faciles à obtenir. Par conséquent, le SBDD est progressivement apparu comme un outil efficace pour aider les chercheurs à prédire la position des petites molécules dans la représentation tridimensionnelle des structures protéiques. L'important, c'est qu'ils accélèrent également la découverte de médicaments, réduisant ainsi considérablement le temps et les coûts de recherche [380].

Introduction de SBDD Les médicaments à base de structure reposent sur des modèles structurels disponibles de la protéine cible. Ces modèles peuvent être fournis par des méthodes incluant la diffraction de rayons X ou la simulation moléculaire. SBDD est principalement divisé en plusieurs étapes comprenant la préparation de structure de protéine, l'identification de site de liaison, la préparation de bibliothèque de ligand, l'arrimage, et les fonctions de notation. En général, après avoir capturé la structure des macromolécules du récepteur, un logiciel de modélisation moléculaire peut être utilisé pour analyser les propriétés physiques et chimiques du site de liaison du médicament au récepteur. Ensuite, recherchez la molécule cible dans la base de données des petites molécules. Ces molécules sont ensuite synthétisées, et elles seront testées pour la mise au point ultérieure de médicaments comme illustré dans la figure 2.7.

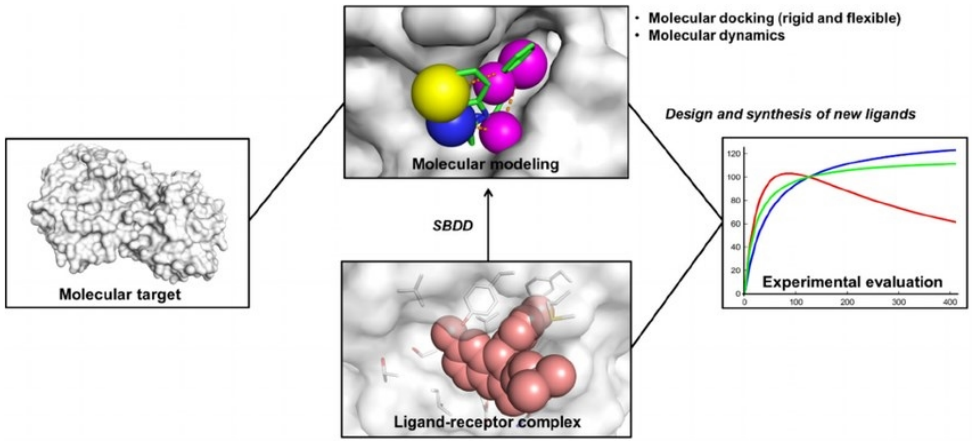


FIGURE 2.7 – Aperçu du SBDD

Applications de SBDD Il y a beaucoup de méthodes utilisées dans les premiers stades de la découverte d'un médicament. Parmi les nombreuses méthodes, SBDD est l'une des stratégies les plus puissantes. Cette méthode intègre les technologies traditionnelles et modernes dans les domaines de la chimie médicinale, de la chimie informatique, de la biochimie et de la biologie structurelle. Le développement de la chimie médicinale a conduit à un nombre croissant d'applications réussies des méthodes basées sur les structures. Certains exemples réussis

comprennent les inhibiteurs de la neuraminidase, de la rénine, de la tyrosine phosphatase, de la bêta-lactamase, de l'anhydrase carbonique et de l'ADN gyrase. En résumé, cette méthode facilite grandement la découverte de médicaments.

Progrès et avantages Avec l'identification de structures 3D de molécules plus biologiques par des méthodes expérimentales ou informatiques, les approches SBDD, qui sont économiques et productives, ont été largement utilisées pour concevoir et découvrir de nouveaux composés de lead pour des cibles de maladies associées. Le développement rapide du SBDD est un facteur clé pour promouvoir la conception de médicaments par fragments (FBDD). FBDD joue un rôle clé dans plus de 30 médicaments cliniques candidats et trois médicaments oncologiques. Le principe principal du FBDD est d'utiliser une petite bibliothèque de composés pour échantillonner efficacement l'espace chimique, ce qui en fait une méthode supplémentaire pour le criblage à haut débit de grandes bibliothèques de composés. Actuellement, la FBDD est largement favorisée par les grandes sociétés pharmaceutiques.

2.3.5 Conception de médicaments à base de fragments

La découverte de médicaments est une entreprise hautement interdisciplinaire qui implique une multitude de domaines spécialisés et peut être caractérisée en plusieurs étapes. En général, cela commence par l'identification et la validation des cibles, suivie de l'identification et de l'optimisation du lead, puis progresse jusqu'aux études précliniques sur les animaux et se termine par les essais cliniques chez l'homme. La découverte de médicaments à base de fragments (FBDD) a été reconnue au cours des deux dernières décennies comme un outil puissant pour la conception rationnelle des pistes médicamenteuses [381].

FBDD Introduction FBDD, comme on l'appelle maintenant, a commencé il y a environ 25 ans. Dans ses premières années, il était souvent appelé Fragment-Based Lead Design et plus tard (et même maintenant parfois) comme Fragment-Based Drug Design. L'approche de la FBDD repose sur la génération des hits en commençant par des fragments moléculaires stables qui partagent des caractéristiques communes : typiquement, un poids moléculaire de <300 Da, cLog P3 (une mesure d'hydrophilie, avec de faibles valeurs améliorant l'absorption), et le nombre de donneurs de liaisons hydrogène, ou accepteurs, de 3 ou moins. Le nombre de liaisons rotatoires de 3 et la surface polaire de 60 Å2 peuvent également être des critères utiles pour FBDD. Ces fragments sont plus petits que les molécules ordinaires de type lead utilisées dans la découverte de médicaments et sont réputés conformes à la règle des trois (Ro3) en raison de leur taille et de leur composition chimique. Le premier exemple de FBDD peut être retracé au travail fondamental sur les relations structure-activité-SAR par NMR en 1996. Depuis cette date et jusqu'à 2010, la FBDD a joué un rôle de premier plan dans les premiers programmes de découverte de médicaments de nombreuses sociétés pharmaceutiques. Depuis lors, la FBDD est devenue une partie intégrante de nombreux efforts de découverte de médicaments dans l'industrie et le milieu universitaire comme le montre cette figure 2.8.

FBDD est devenu une approche puissante dans la découverte de nouveaux composés du lead. Le criblage de fragments plus petits offre une série d'avantages par rapport au criblage traditionnel à haut débit (HTS), y compris l'échantillonnage supérieur de l'espace chimique et des taux de réponse plus élevés. La première étape du processus FBDD consiste à identifier les fragments qui lient faiblement la protéine cible, généralement dans la gamme d'affinité micromolaire à millimolaire. En raison de la faible affinité des fragments pour leurs cibles, des techniques biophysiques sensibles aux affinités micromolaires à millimolaires sont nécessaires pour identifier les hits ; parmi elles, la résonance magnétique nucléaire (RMN), la cristallographie par rayons X et la résonance plasmonique de surface (SPR) sont les plus utilisées. Dans une

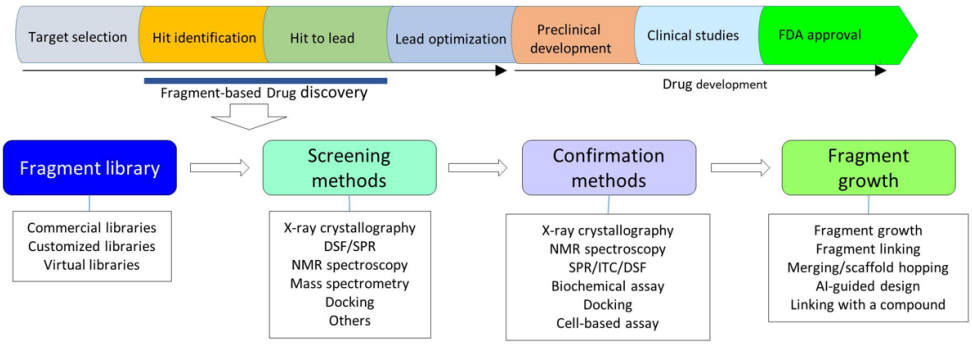


FIGURE 2.8 – Un organigramme de FBDD

deuxième étape, itérative et longue, les fragments qui forment des interactions de haute qualité sont optimisés en composés de lead présentant une affinité et une sélectivité plus élevées, grâce aux stratégies dites de croissance de fragment, de fusion de fragment ou de liaison de fragment.

Application de FBDD Cela fait environ 25 ans depuis la première description expérimentale de FBDD. Depuis lors, de nombreux composés sont sortis des programmes du FBDD et sont entrés dans la clinique. La FBDD est établie dans les grandes sociétés pharmaceutiques, les petites et moyennes entreprises de biotechnologie et le milieu universitaire, où elle est appliquée à une gamme croissante de cibles. Un des attraits du FBDD est qu'avec une bibliothèque de composés relativement petite, il est possible d'obtenir une couverture beaucoup plus grande de l'espace chimique disponible que même les plus grandes bibliothèques de HTS. Le FBDD est capable de s'attaquer à des cibles nouvelles et difficiles, qui peuvent être moins adaptées aux HTS, il a été repris par un nombre croissant de groupes universitaires, et il offre le potentiel de générer à la fois des pistes médicamenteuses neuves et des sondes chimiques sélectives de fonction et de biologie protéiques.

2.3.6 Conception de médicaments à base de ligand

Une grande classe de méthodes intégrant à la fois des méthodes de conception de médicaments à base de ligand et à base de structure est basée sur la comparaison ou la modélisation des interactions protéine-ligand dans des systèmes protéines-ligand similaires. L'objectif est d'identifier les principales interactions protéine-ligand à partir des données physico-chimiques disponibles et d'utiliser les données d'interaction obtenues pour identifier les ligands avec des profils d'interaction similaires. Cette classe de méthodes intégrées peut être divisée en deux sous-catégories. La première sous-catégorie, les techniques de pseudorécepteur, met en corrélation des similarités entre ligands et activité biologique mesurée et établit ainsi une représentation structurale de la poche de liaison du ligand protéique. L'autre ensemble de techniques est l'inverse de la première catégorie. Ces méthodes analysent les interactions protéine-ligand à partir de données structurelles pour extraire des types clés d'interactions et traduire cette information en une représentation mathématique simplifiée qui peut être utilisée par la similaritéméthodes de base pour le criblage des composés actifs dans les bibliothèques de ligand comme illustré dans la figure 2.9. De nombreuses techniques de cette catégorie sont basées sur des modèles d'empreintes digitales ou de pharmacophores [382].

Application de la conception des médicaments à base de ligand

Méthodes de pseudo récepteur Les méthodes de pseudorecepteur sont principalement des extensions des techniques QSAR, principalement des techniques 3D-QSAR telles que CoMFA,

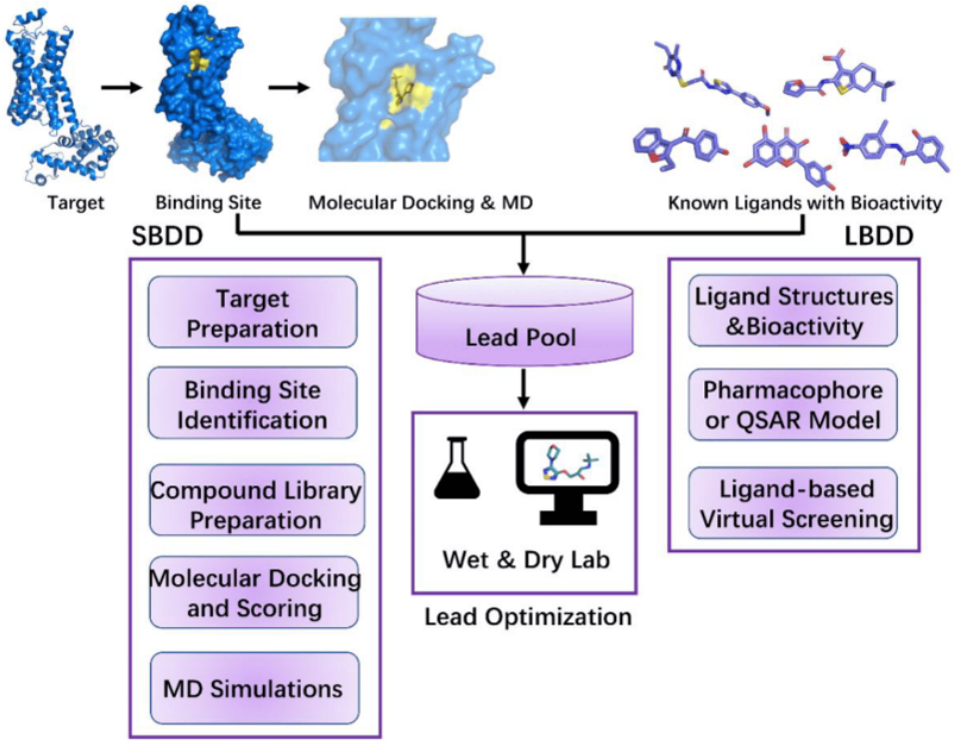


FIGURE 2.9 – Comparisme de la conception des médicaments à base de ligand (LBDD) et de la conception des médicaments à base de structure (SBDD)

CoMSIA et GOLPE. Ces techniques QSAR placent des informations physico-chimiques dans un espace 3D entourant un ensemble de composés de référence alignés qui se lient au même site de liaison d'une cible macromoléculaire commune. Les méthodes de Pseudoreceptor étendent cette cartographie en essayant de créer des modèles du site de liaison de protéine cible autour de l'ensemble de ligands. Ces modèles représentatifs de pseudorecepteurs sont destinés à contenir des interactions clé entre les protéines et les ligands et à cartographier la forme et le volume appropriés de ces interactions. Le but de la modélisation des pseudo récepteurs est de générer des substituts de la structure 3D du site de liaison des protéines qui peuvent être utilisés pour des applications de conception de médicaments à base de structures telles que le criblage virtuel, la modification rationnelle, ou de proposer de nouvelles petites molécules complémentaires au modèle pseudorécepteur, et de prédire les affinités de liaison des ligands potentiels. Les premières méthodes de pseudorecepteur impliquaient le pliage manuel des chaînes peptidiques autour de l'ensemble de ligands, mais ces méthodes ont maintenant évolué pour inclure une grande variété de méthodes informatiques automatisées. Bien qu'il existe de nombreuses façons différentes de générer un pseudo-récepteur, les chercheurs ont divisé les méthodes du pseudo-récepteur en six catégories : méthodes à base de grille, à base de partition, à base de peptide, (iso)surface, à base d'atome et à base de fragment.

Pharmacophore et empreintes digitales L'autre type de méthodes basées sur l'interaction, principalement des pharmacophores dérivés de la structure protéique ou des techniques d'empreintes digitales, adopte une approche inverse à l'intégration de la conception basée sur le ligand et la structure par rapport aux techniques pseudo-recettrices. Lorsque les méthodes de pseudorecepteur tentent d'analyser la similitude entre ligands pour en tirer des modèles informatiques qui imitent les interactions protéine-ligand, le pharmacopore et les techniques d'empreintes digitales analysent les structures existantes d'une ou plusieurs protéinesComplexes de

ligands pour générer une représentation informatique d'importants contacts protéine-ligand. La représentation mathématique générée est ensuite utilisée dans la recherche de similarité pour trouver des ligands qui correspondent au profil d'interaction. Ces techniques d'empreintes digitales ou de pharmacophore partagent le concept de la simplification des données structurelles complexes protéine-ligand afin d'identifier un petit nombre d'interactions clés. Ces méthodes sont divisées en trois catégories différentes : les méthodes à base de pharmacophore, les méthodes d'encodage direct à base d'empreintes digitales et les méthodes d'encodage indirect à base d'empreintes digitales.

2.3.7 CRISPR dans la découverte de médicaments

La découverte actuelle de médicaments repose sur l'une des deux approches. La première voie consiste à identifier un gène/une voie pour développer un médicament (alias une cible médicamenteuse). Les maladies dont les causes sont définies et directes (une mutation spécifique dans un gène connu = apparition de la maladie) peuvent être mieux servies par ce type de conception ciblée. On peut aussi créer un modèle de la maladie et dépister les médicaments pour déterminer s'ils présentent des changements dans la pathologie. Si la pathologie de la maladie est complexe, un modèle peut être le plus efficace. Une fois que les candidats-médicaments sont isolés, ils peuvent être validés pour la spécificité de leur cible ou leur cible directe peut être identifiée si elle n'est pas connue. Si un candidat solide émerge, l'optimisation en aval est effectuée pour amener le médicament aux phases d'essais précliniques et humains comme indiqué dans la figure suivante 2.10. CRISPR accélère le processus et peut aider à éliminer beaucoup de médicaments pauvres avant que le temps et l'argent ne soient investis dans ces derniers [383].

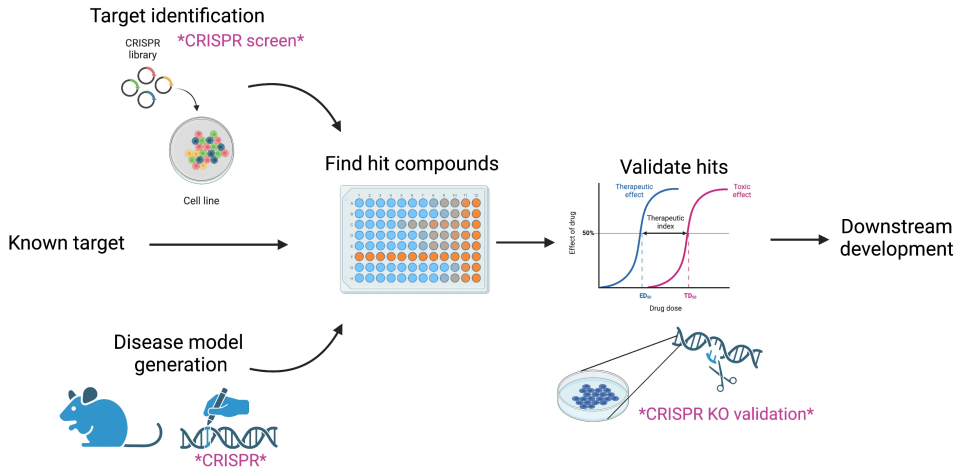


FIGURE 2.10 – CRISPR aide à de multiples étapes du pipeline de découverte de médicaments

Identification de cible

Toutes les maladies ne sont pas complètement caractérisées ou ont une pathologie simple. Dans ces cas, l'identification préliminaire des cibles potentielles par CRISPR peut accélérer le pipeline de médicaments. Des écrans CRISPR à haut débit peuvent être utilisés pour inhiber, activer ou éliminer de nombreux gènes à la fois afin d'identifier les «hits» - des gènes qui affectent la progression ou l'apparition de la maladie. Ces résultats peuvent révéler des gènes individuels ou parfois même des voies génétiques qui valent la peine d'être étudiées pour le développement de médicaments. En raison de la difficulté à droguer certaines cibles – les protéines «non drugables» redoutées – avoir plusieurs succès ou toute une gamme de possibilités augmente considérablement les chances de trouver un médicament de qualité en aval.

Génération d'un modèle de maladie

CRISPR a rendu la génération de modifications du génome accessible, rapide et facile. Dans les cas où la base mutationnelle d'une maladie est connue – expansion répétée, knock-out de gène, etc. – CRISPR peut être utilisé pour générer des modèles cellulaires ou animaux qui ont la génétique précise de la maladie. Puisque CRISPR a été adapté pour l'usage dans les cellules primaires, beaucoup de types de tissu, et à travers des espèces, il permet à des chercheurs de choisir et produire les modèles physiologiquement appropriés. Dans les maladies complexes, il peut aider à la génération de plusieurs modifications ou à la mise en œuvre de systèmes d'expression pour imiter avec précision un phénotype de maladie.

CRISPR peut également être utilisé pour générer des knock-ins de lignes cellulaires de mutations de patients qui sont des variantes d'importance inconnue – des mutations avec des effets incertains sur la pathologie de la maladie – afin de déterminer si un patient répondrait à une thérapie existante. En théorie, CRISPR pourrait même être utilisé pour générer des mutations chez le patient et ensuite tester l'efficacité des médicaments individuels et combinatoires avant le traitement du patient.

Validation de la cible

Lorsqu'une cible est connue et qu'un médicament est mis au point, il est recommandé de mesurer la spécificité de cette cible et de valider le composé visé. La norme consiste à tester le médicament sur des modèles qui sont génétiquement compétents et déficients pour la cible. CRISPR a révolutionné la génération de lignées cellulaires knockout qui a rendu cette étape de contrôle extrêmement accessible. Par exemple, disons que vous avez identifié la cible du gène X pour le cancer du sein déficient en BRCA et identifié plusieurs inhibiteurs candidats du gène X dans un écran. Vous générez un X gene knockout dans un fond cancéreux et traitez la lignée cellulaire parentale et le X knockout avec votre médicament(s). Si un médicament est également toxique pour les deux lignées, il a probablement une toxicité élevée ou des effets cibles élevés, ce qui en fait un médicament non idéal. Si un médicament tue de façon sélective la lignée des wildtypes avec une toxicité minimale ou nulle dans la lignée du knockout, il s'agit probablement d'un médicament spécifique ayant une faible toxicité hors cible. Ce CQ supplémentaire est l'endroit où CRISPR peut économiser du temps et de l'argent pour s'assurer que des candidats de qualité sont sélectionnés.

2.4 Découverte de médicaments assistée par intelligence artificielle (IA)

L'IA, également connue sous le nom d'intelligence artificielle. Avec les progrès de la technologie, l'IA a connu des changements révolutionnaires. Les technologies telles que la reconnaissance faciale sont toutes des produits représentatifs de l'IA. Il est particulièrement important de noter que l'IA offre des possibilités pour la découverte et le développement de médicaments novateurs.

En raison du long cycle de développement et du faible taux de réussite, la découverte et la conception de nouveaux médicaments sont un processus extrêmement long, coûteux et difficile. La découverte de médicaments assistée par l'IA est le moyen le plus prometteur pour résoudre ce dilemme. En intégrant des données dans un espace à haute dimension et en extrayant les relations clés, l'IA fournit des solutions novatrices pour toutes les étapes de la découverte précoce de médicaments. La combinaison efficace de l'IA et des nouvelles technologies expérimentales devrait permettre de trouver de nouveaux médicaments plus rapidement, moins cher et plus efficacement, comme le montre la figure suivante 2.11. Plus important encore, les récentes percées

de l'IA ont prouvé son potentiel dans l'industrie pharmaceutique [384].

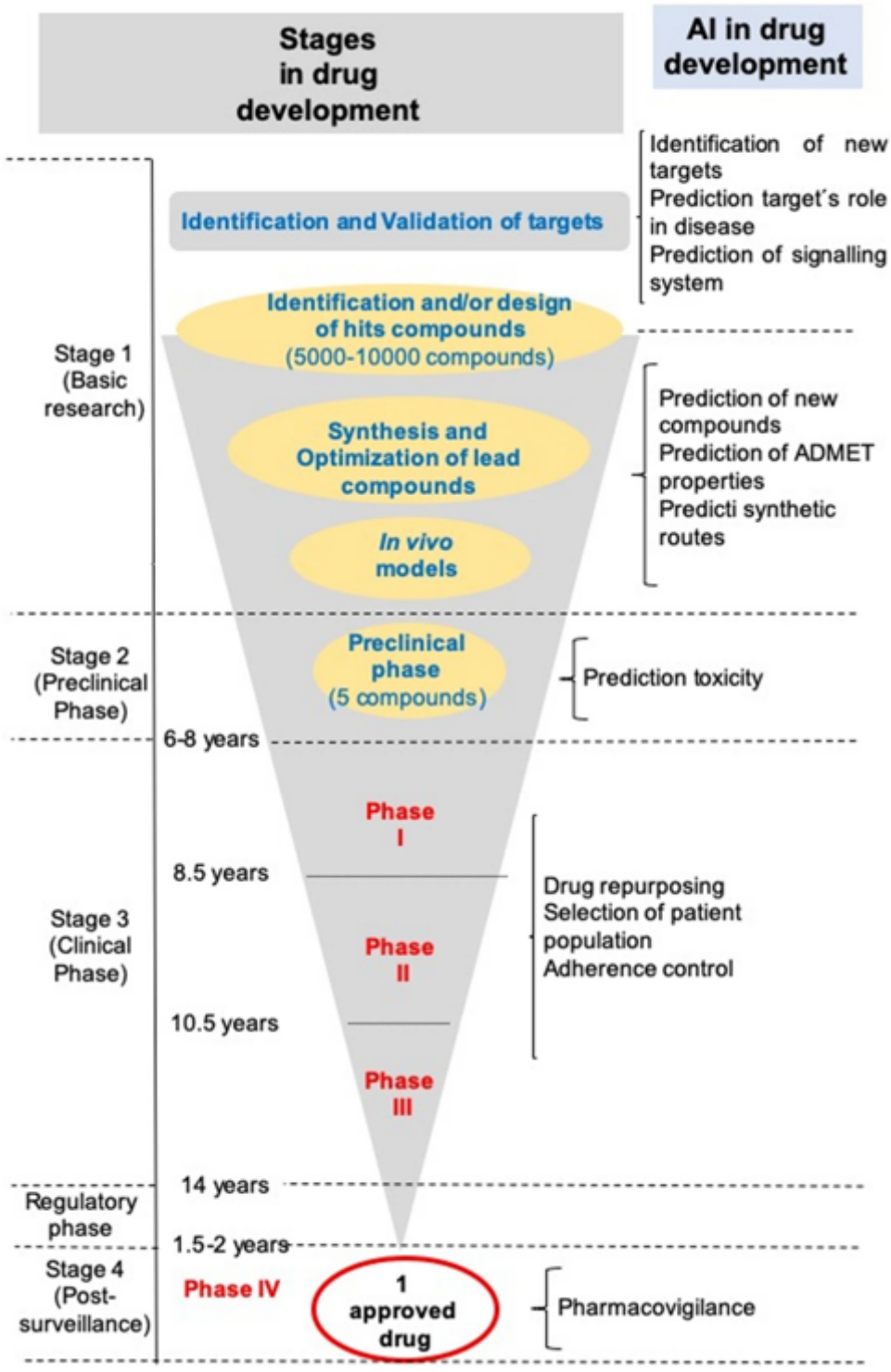


FIGURE 2.11 – L'IA dans le développement de médicaments

2.4.1 IA utilisée dans la découverte de médicaments

L'identification des médicaments qui sont bénéfiques pour le corps est le principal objectif de la recherche sur la découverte de médicaments. La plupart de ces médicaments sont mélangés artificiellement avec de petites molécules. Pour découvrir ces molécules, les chercheurs doivent

étudier la bibliothèque moléculaire afin d'identifier les molécules cibles qui ont le potentiel de devenir des médicaments. Étant donné que la structure chimique spécifique qui est biologiquement appropriée comme médicament efficace n'est pas claire, il s'agit d'une méthode coûteuse et longue pour raffiner les composés potentiels en médicaments candidats. Sur la base des faits ci-dessus, les systèmes d'IA sont devenus une nouvelle stratégie pour accélérer le développement de médicaments et réduire le coût de recherche de nouveaux médicaments en raison de leur potentiel inégalé de traitement des données.

2.4.2 IA dans la conception de médicaments

À l'heure actuelle, un grand nombre de technologies de découverte de médicaments assistées par l'IA ont été utilisées, y compris le criblage virtuel, la conception de nouveau médicament, la prédiction des propriétés physico-chimiques et pharmacocinétiques, la réutilisation du médicament et les aspects connexes. En outre, l'IA joue également un rôle majeur dans la planification de la synthèse chimique, le traitement des images cellulaires, la prédiction de l'activité biologique physique et de la toxicité, et l'exploitation de systèmes robotisés pour la synthèse organique.

— **AI pour le criblage virtuel** L'IA peut être utilisée pour le criblage virtuel des molécules cibles. La modélisation basée sur des algorithmes de reconnaissance de séquences, des images d'apprentissage profond et des millions de données de composés actifs triées manuellement permet à l'IA d'identifier les composés les plus puissants et de déterminer leur emplacement.

— **AI pour conception du composé** Cette stratégie est développée sur la prémisse de composés actifs connus pour concevoir des dérivés ou des analogues de haute qualité. Dans ce processus, la mise en correspondance des caractéristiques et la réalisation d'une analyse de faisabilité synthétique peuvent jouer un rôle clé pour trouver des composés de lead de haute qualité.

— **AI pour conception du parcours de synthèse** Actuellement, certaines plateformes peuvent concevoir des voies synthétiques pour des composés spécifiques. Ce processus utilise la technologie de l'IA pour générer un modèle permettant de recommander rapidement le chemin de synthèse avec le coût le plus bas et le taux de réussite le plus élevé.

— **IA dans la prédiction de la structure des molécules cibles** La structure 3D des protéines cibles structurelles est essentielle pour la découverte de médicaments structurés. Les méthodes traditionnelles prennent généralement plusieurs années pour résoudre la structure moléculaire cible, tandis que la prédiction de structure basée sur l'IA ne prend que quelques heures, ce qui rend ce processus plus rapide et plus précis.

Les applications de l'IA pour promouvoir le processus d'identification des médicaments et la découverte de molécules appropriées dans la base de données ont montré un grand potentiel en matière de découverte de médicaments. En outre, l'IA a une grande importance dans différents domaines tels que les soins médicaux, l'anti-vieillissement et le cancer.

Conclusion

L'environnement de la découverte des médicaments a subi une profonde transformation au cours des dernières décennies, passant d'approches expérimentales traditionnelles à des méthodologies sophistiquées axées sur la technologie qui tirent parti de la puissance de calcul et

de l'intelligence artificielle. Comme l'a démontré cette revue, l'intégration de stratégies complémentaires—englobant des méthodes computationnelles, la chimie médicinale, le criblage à haut débit et virtuel, les analyses phénotypiques, la conception fondée sur le génome et les technologies d'édition du génome, fournit un cadre complet pour relever les défis inhérents à la découverte de nouveaux agents thérapeutiques. Le faible taux de réussite persistant des candidats à des essais cliniques (environ 10%) souligne la nécessité d'innover continuellement dans les méthodes de découverte afin d'améliorer l'efficacité et les résultats.

La conception assistée par ordinateur des médicaments est devenue une pierre angulaire de la découverte moderne des médicaments, réduisant considérablement l'espace chimique nécessitant une évaluation expérimentale à la fois par des approches structurales et basées sur les ligands. Les méthodologies de criblage virtuel, y compris SBVS et LBVS, ont démocratisé l'accès à de vastes bibliothèques chimiques, tandis que des techniques spécialisées telles que le docking moléculaire, la modélisation du pharmacophore et l'analyse QSAR fournissent des renseignements ciblés sur les interactions médicamenteuses potentielles. Ces méthodes de calcul complètent efficacement les approches expérimentales comme le criblage à haut débit, qui, malgré les limitations inhérentes aux coûts et aux exigences en matière de ressources, demeure essentiel pour l'identification du lead. Pendant ce temps, le criblage phénotypique a connu une renaissance, en particulier pour les maladies complexes avec des mécanismes mal compris, offrant une voie alternative qui se concentre sur les changements observables dans les systèmes biologiques plutôt que sur des cibles moléculaires spécifiques.

L'intégration de l'intelligence artificielle et de l'apprentissage automatique représente peut-être le progrès le plus transformateur dans ce domaine, offrant des capacités sans précédent en matière de traitement de données, de reconnaissance de modèles et de modélisation prédictive. Les approches assistées par l'IA couvrent maintenant tout le continuum de la découverte d'un médicament (du criblage virtuel et de novo drug design à la prédiction des propriétés et à l'optimisation des voies synthétiques) en réduisant considérablement le temps et le coût associés aux méthodes traditionnelles. De même, la technologie CRISPR a révolutionné l'identification des cibles, la génération de modèles de maladie et les processus de validation, permettant une manipulation génétique précise et des systèmes expérimentaux plus physiologiquement pertinents. À l'avenir, la convergence continue de ces technologies promet de relever les défis fondamentaux de la découverte de médicaments en permettant le développement plus rapide et plus rentable de thérapies sûres et efficaces. Le succès dépendra en fin de compte d'une collaboration interdisciplinaire qui exploite efficacement les méthodes de calcul, les techniques expérimentales et les connaissances cliniques pour naviguer sur la voie complexe du concept moléculaire au médicament approuvé. À mesure que les capacités de calcul continuent d'évoluer et que les méthodologies expérimentales deviennent de plus en plus sophistiquées, la perspective d'une découverte de médicaments plus efficace et réussie devient non seulement une aspiration mais de plus en plus réalisable.

Chapitre 3

Approches d'apprentissage profond basées sur des graphes

Introduction

Le processus de découverte des médicaments dans l'industrie pharmaceutique est intrinsèquement difficile, caractérisé par ses longs délais, ses coûts importants et sa nature souvent inefficace. Les approches traditionnelles peuvent s'étendre sur plus d'une décennie et nécessiter des milliards de dollars, ce qui souligne le besoin crucial de méthodologies novatrices pour accélérer et optimiser ce processus complexe. La complexité croissante des maladies, associée à la demande croissante de thérapies moléculaires ciblées, souligne encore plus les limites des techniques conventionnelles. Cela nécessite l'exploration et la mise en œuvre de technologies de pointe pour rationaliser le flux de travail de découverte de médicaments [39, 111].

Au cours des dernières années, l'intégration de l'intelligence artificielle (IA), et particulièrement du deep learning, est apparue comme une force transformatrice qui pourrait révolutionner la découverte de médicaments [112]. La capacité de l'apprentissage profond à analyser des modèles complexes dans de vastes ensembles de données s'est révélée très prometteuse pour faire progresser deux aspects cruciaux de ce processus : la prédiction des propriétés moléculaires (MPP) et la génération moléculaire (MG). Ces deux domaines, lorsqu'ils sont combinés, offrent une puissante approche synergique au processus de développement des médicaments [40, 113].

La prédiction des propriétés moléculaires (MPP) et la génération moléculaire (MG) sont deux approches synergiques qui stimulent l'innovation dans la découverte de médicaments. MPP se concentre sur la prédiction des propriétés moléculaires cruciales telles que la solubilité, la biodisponibilité, la toxicité et l'affinité de liaison, ce qui permet une identification précoce des médicaments candidats prometteurs et réduit le besoin d'expériences coûteuses [41]. Cela empêche la poursuite de composés inefficaces, optimisant l'allocation des ressources. En complément du MPP, MG se concentre sur la conception et la synthèse de nouvelles molécules adaptées à des cibles thérapeutiques spécifiques [42]. Cela permet aux chercheurs d'explorer de vastes espaces chimiques, de découvrir de nouveaux supports et composés leads et même de concevoir des molécules de novo pour relever des défis thérapeutiques auparavant insolubles.

Le succès de l'apprentissage profond en MPP et en MG est largement attribué aux avancées dans les modèles basés sur des graphes. Ces modèles sont devenus essentiels pour représenter et analyser les structures moléculaires, en capturant les relations complexes et les caractéristiques topologiques inhérentes aux molécules. Des architectures comme les réseaux de neurones passant par message (MPN), les réseaux convolutionnaires graphiques (GNG) et leurs variantes ont démontré des performances exceptionnelles en MPP. Simultanément, des modèles génératifs

tels que les graphe autoencodeurs variationnels s (GVAEs) et les réseaux adversarial génératifs s (GANs) permettent aux chercheurs de générer de nouvelles molécules médicamenteuses aux propriétés optimisées, accélérant ainsi la conception de nouveaux traitements. Les progrès récents dans ces domaines, y compris le développement de réseaux neuronaux à graphes plus sophistiqués et l'exploration de nouvelles architectures génératives, continuent de repousser les limites de ce qui est possible dans la découverte de médicaments [43, 115, 116].

Ce chapitre propose une exploration exhaustive des techniques d'apprentissage profond de pointe qui transforment le MPP et le MG, deux piliers essentiels de la découverte moderne de médicaments. Il explore la puissance des réseaux de neurones basés sur des graphes, expliquant comment ces modèles capturent efficacement les caractéristiques structurelles et chimiques complexes des molécules. Le chapitre examine en outre les avantages de ces approches d'apprentissage profond par rapport aux méthodologies traditionnelles, mettant en évidence leur capacité supérieure à gérer des relations complexes et non linéaires dans les données moléculaires. En fournissant une analyse nuancée de ces avancées et des défis qui y sont associés, ce chapitre vise à éclairer l'impact profond de l'apprentissage profond sur l'accélération du processus de découverte de médicaments, la réduction des coûts de développement et, en fin de compte, favoriser la création de thérapies moléculaires innovantes pour un avenir plus sain.

3.1 Représentation moléculaire des graphes

Les graphes, une structure de données fondamentale en informatique et en mathématiques, offrent un cadre polyvalent pour la représentation et l'analyse de systèmes complexes. Leur pouvoir s'étend à divers domaines, dont la chimie, la biologie et les réseaux sociaux [79]. Dans la modélisation moléculaire, les molécules peuvent être élégamment représentées sous forme de graphes, avec des atomes servant de noeuds et des liaisons chimiques d'arêtes. Cette représentation graphique naturelle présentée à la figure 3.1 ouvre le potentiel des algorithmes de la théorie des graphes pour l'analyse des structures et des propriétés moléculaires, ce qui mène à des méthodes de calcul plus efficaces et précises dans la découverte de médicaments et la chimio-informatique [50].

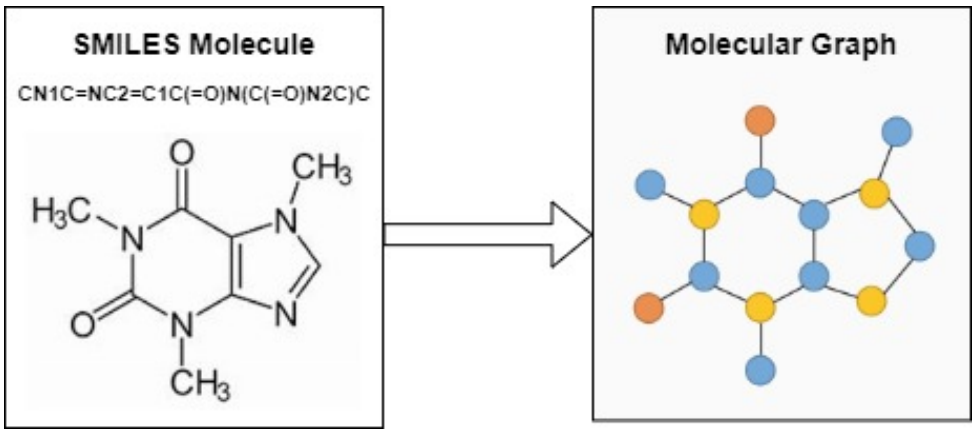


FIGURE 3.1 – SMILES moléculaire à représentation graphique

3.1.1 Représentation des graphes

Les molécules peuvent être représentées efficacement sous forme de graphe, fournissant un cadre puissant pour l'analyse informatique. Une molécule G est généralement définie comme

un graphe $G = (V, E)$, où :

— V est l'ensemble des nœuds, chacun représentant un atome dans la molécule.

— E est l'ensemble des arêtes, représentant les liaisons chimiques reliant ces atomes.

— $|V| = n$ indique le nombre d'atomes dans la molécule.

— $|E| = m$ indique le nombre d'obligations.

Les nœuds peuvent correspondre à divers types atomiques du tableau périodique ou à des fragments moléculaires spécifiques. Pour caractériser complètement le graphe moléculaire, nous définissons deux composantes clés :

1. Matrice de caractéristiques des nœuds $\mathbf{X} \in 0, 1^{n \times c}$, où c est le nombre de types de nœuds distincts (types atomiques).
2. Tenseur d'adjacence$\mathbf{A} \in 0, 1^{n \times n \times b}$, où b est le nombre de types d'arêtes.

Le tenseur d'adjacence $\mathbf{A}$ encode les informations de connectivité. $\mathbf{A}ijk = 1$ indique la présence d'une arête de type k entre les nœuds i et j, alors que $\mathbf{A}ijk = 0$ autrement.

Par conséquent, un graphe moléculaire peut être représenté de manière concise comme suit $G = (\mathbf{A}, \mathbf{X})$, encapsulant à la fois l'information structurale et atomique de la molécule.

L'intégration de la théorie des graphes avec le machine learning, en particulier le deep learning, a conduit au développement des réseaux neuronaux graphiques (GNNs). Ces modèles puissants sont spécialement conçus pour apprendre sur des données structurées par graphe [51]. GNNs ont démontré un succès remarquable dans diverses tâches de modélisation moléculaire, y compris la prédiction des propriétés, la génération moléculaire et la prédiction de réaction. Ils utilisent la structure des graphes pour apprendre les représentations hiérarchiques en capturant l'information structurelle locale et globale dans les molécules.

Transforme les SMILES en graphes pour la représentation moléculaire

Notre méthode utilise le Simplified Molecular-Input Line-Entry System (SMILES) comme représentation initiale pour construire des graphes moléculaires. SMILES, une notation de chaîne largement utilisée pour les molécules, fournit un moyen efficace d'encoder des informations structurelles. Pour transformer les chaînes SMILES en structures graphiques adaptées aux modèles d'apprentissage profond, nous utilisons la bibliothèque RDKit, une puissante bibliothèque de cheminformatique [52].

Le processus de conversion consiste à analyser la chaîne SMILES avec RDKit pour générer un objet moléculaire. À partir de cet objet, nous extrayons des informations atomiques et de liaison détaillées pour construire notre représentation graphique. Nous dénommons le graphe moléculaire résultant comme : $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, tel que $\mathcal{V}$ représente l'ensemble des atomes (nœuds) et $\mathcal{E}$ représente l'ensemble des liaisons (arêtes). Cette représentation basée sur des graphes préserve efficacement la structure topologique de la molécule tout en permettant l'incorporation d'informations chimiques riches [80].

Matrices Feature et Adjacency pour les réseaux de neurones graphiques :

Pour représenter efficacement les molécules des réseaux de neurones à base de graphes (GNN), nous construisons deux matrices clés : une matrice feature $\mathbf{F}$ et une matrice adjacency, $\mathbf{A}$ comme indiqué dans la figure suivante 3.2.

La matrice feature $\mathbf{F} \in \mathbb{R}^{N \times d}$ encode les propriétés atomiques. Ici N est le nombre d'atomes et d est la dimension de l'espace des caractéristiques. Chaque ligne de $\mathbf{F}$ correspond à un atome, avec son vecteur de caractéristiques généré en utilisant un schéma d'encodage à chaud. Ce codage capture les caractéristiques atomiques cruciales telles que le type d'élément, l'état d'hybridation et la charge formelle [81].

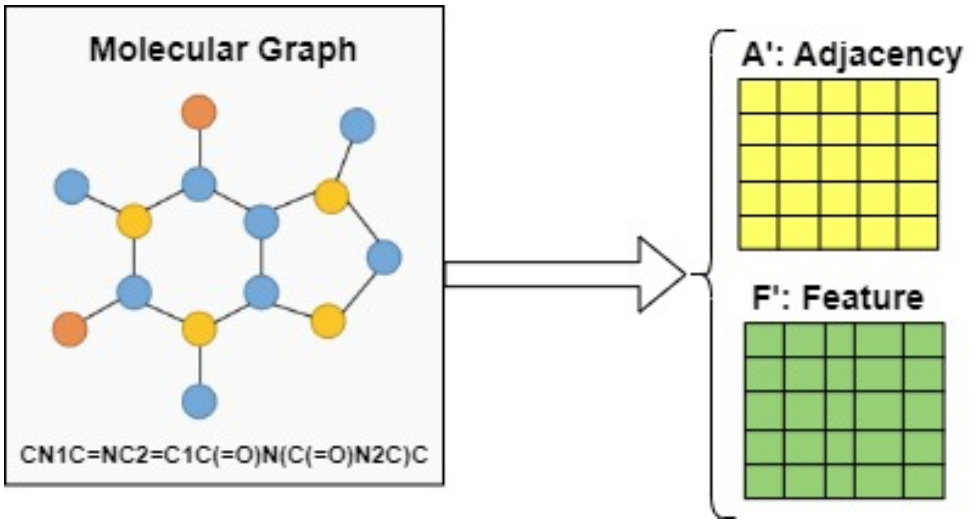


FIGURE 3.2 – Matrices de caractéristiques et de contiguïté pour la représentation graphique

L'information sur les liaisons est représentée par la matrice de contiguïté $\mathbf{A} \in \mathbb{R}^{B\times N\times N}$, tel que B est le nombre de types des liaisons. Cette matrice multidimensionnelle nous permet d'encoder différents types de liaisons (par exemple, simple, double, triple) comme des canaux distincts, fournissant une représentation riche de la connectivité moléculaire [82].

3.2 Réseaux de neurones basés sur des graphes

Les réseaux neuronaux graphiques (GNN) sont devenus un outil puissant pour la modélisation de relations complexes au sein de données structurées par graphe, ce qui démontre leur succès dans diverses applications telles que l'analyse des réseaux sociaux, la prédiction des composés chimiques et l'optimisation des systèmes de transport [83].

Comme mentionné dans la figure 3.3 à leur cœur, les GNN visent à mapper un graphe sur une représentation unique et significative, souvent une intégration numérique. Ceci peut être exprimé mathématiquement comme F(G) = embedding, où F est la fonction qui transforme le graphe G en un embedding condensé [84].

Les GNN y parviennent en adaptant le concept des réseaux de neurones récurrents (RNN), qui excelle dans le traitement des données séquentielles. En remplaçant chaque nœud du graphique par une unité récurrente, comme un réseau de mémoire à court terme (LSTM), et chaque arête par un réseau neuronal qui capture le poids de la bordure, les GNNs propagent efficacement l'information dans toute la structure du graphe. Cela leur permet d'apprendre des modèles et des relations complexes dans le graphe, ce qui permet de faire des prédictions et des représentations précises [61].

Pendant la formation, les GNN affinent itérativement l'intégration des nœuds par un processus récursif. Chaque nœud rassemble les embeddings de ses nœuds voisins, les combine avec son propre embedding actuel, puis transmet le résultat à travers son unité récurrente (par exemple, LSTM). Cela génère une nouvelle intégration mise à jour pour le nœud en incorporant des informations de son voisinage local. Ce processus se répète pour tous les nœuds du graphe, permettant à l'information de se propager dans toute la structure. En fin de compte, les intégrations finales de tous les nœuds sont recueillies et combinées pour produire une intégration unique et complète qui représente l'ensemble du graphe [62].

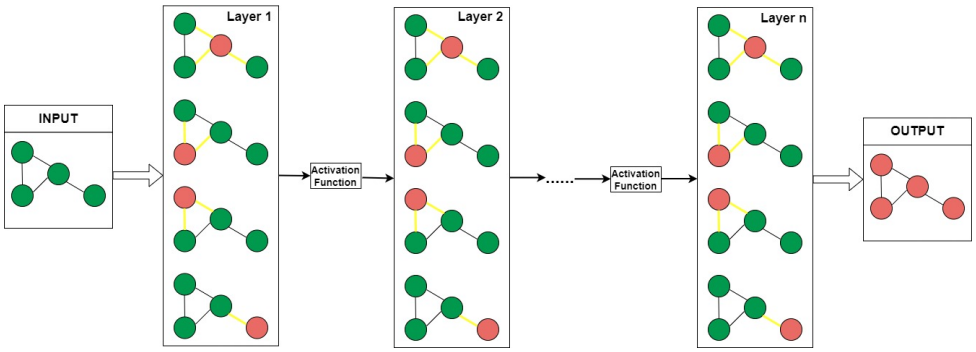


FIGURE 3.3 – Réseau de neurones basé sur les graphes

3.3 Modèles d'apprentissage profond pour MPP et MG

L'évolution des modèles de prédiction des propriétés moléculaires et de génération moléculaire dans la découverte de médicaments met en évidence un passage des systèmes traditionnels fondés sur des règles aux approches sophistiquées d'apprentissage profond. Les méthodes premières, telles que les modèles de relations quantitatives structure-activité (QSAR), s'appuyaient sur la régression linéaire et des techniques d'apprentissage automatique pour corréler les descripteurs moléculaires avec les activités biologiques, mais avaient du mal à saisir les relations complexes non linéaires [63].

Dans les données chimiques L'avènement du deep learning a marqué un tournant, avec des réseaux neuronaux convolutifs (CNN) et récurrents (RNN) initialement appliqués à des données de séquences moléculaires, offrant une meilleure précision prédictive [64].

Cependant, c'est l'introduction de modèles d'apprentissage profond basés sur des graphes comme les réseaux neuronaux graphiques (GNN), les réseaux neuronaux passant par des messages (MPNN) et les réseaux convolutionnels graphiques (GCN) qui a permis le codage direct de structures moléculaires sous forme de graphes. Améliorer de façon significative la prédiction des propriétés clés comme la solubilité, la biodisponibilité et la toxicité [65].

Ces avancées ont joué un rôle déterminant dans la prédiction des propriétés moléculaires (MPP), où les modèles d'apprentissage profond ont permis de prédire avec succès une large gamme de propriétés physico-chimiques et de profils ADMET (absorption, distribution, métabolisme, excrétion et toxicité) [66].

En même temps, le développement de modèles génératifs, y compris les autoencodeurs variationnels (VAE) et les réseaux antagonistes génératifs (GAN), a facilité la création de nouvelles molécules semblables à des médicaments ayant des profils thérapeutiques optimisés, ce qui simplifie le processus de découverte des médicaments [67].

En capturant efficacement des caractéristiques structurelles et chimiques complexes, ces approches d'apprentissage profond ont considérablement accéléré l'identification et l'optimisation de candidats prometteurs [68], comme indiqué dans la figure 3.4.

3.3.1 Modèles d'apprentissage profond pour MPP

Les modèles d'apprentissage profond pour la prédiction des propriétés moléculaires sont devenus des outils indispensables dans la découverte de médicaments en raison de leur capacité à apprendre des relations complexes et non linéaires au sein des données moléculaires. Les méthodes traditionnelles d'apprentissage automatique reposent souvent sur des fonctionnalités fabriquées manuellement, ce qui limite leurs capacités de prédiction. En revanche, les

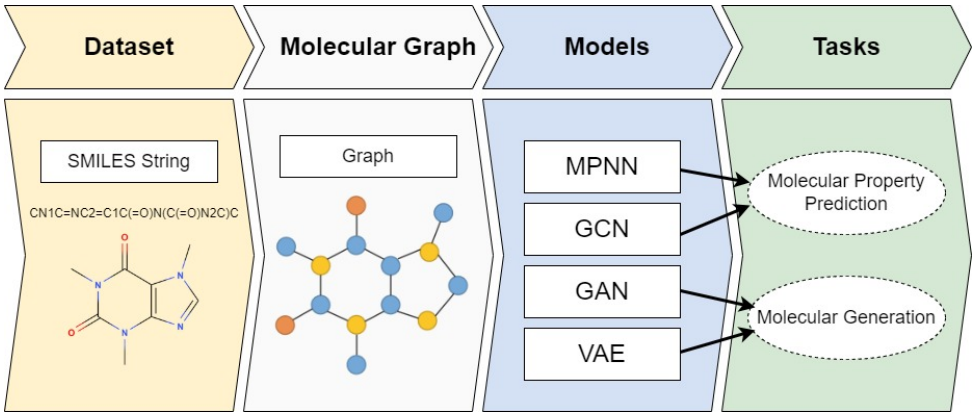


FIGURE 3.4 – Modèles d'apprentissage profond pour MPP et MG

approches modernes d'apprentissage profond utilisent de puissantes architectures de réseaux neuronaux comme les réseaux de neurones graphiques (GNN), les réseaux convolutionnels graphiques (GNN) et les réseaux de neurones passant par des messages (MPN) [104]. Ces modèles encodent directement les structures moléculaires sous forme de graphiques, capturant les caractéristiques topologiques et chimiques complexes des molécules, qui sont essentielles pour des prédictions précises des propriétés [171]. Par exemple, ils excellent dans la prévision des propriétés clés telles que la solubilité, la biodisponibilité, l'affinité de liaison et la toxicité, en tirant parti d'ensembles de données étendus et de représentations graphiques sophistiquées.

Réseaux neuronaux de passage de messages

Les réseaux neuronaux de passage de messages (MPN) sont un type spécialisé de réseau neuronal conçu pour traiter des données structurées par graphe, telles que les graphes moléculaires. Dans un graphe moléculaire, les atomes sont représentés sous forme de nœuds et les liaisons chimiques sous forme d'arêtes. Chaque nœud possède des caractéristiques x_V, et chaque arête a des caractéristiques e_{VW} [69].

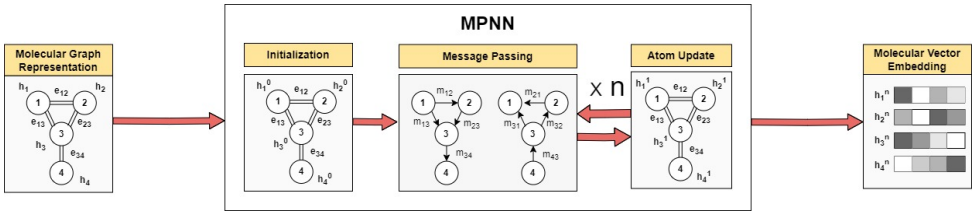


FIGURE 3.5 – Réseau de neurones à passage de message

Comme le montre la figure 3.5, le texte de la MPNN se compose de deux phases : la phase de passage du message et la phase de lecture.

— **La Phase Message-Passing**
  Pendant la phase de passage des messages, les états cachés h_V^t de chaque nœud atomique sont mis à jour en utilisant les messages m_V^{t+1} sur les étapes de temps T. La phase de passage du message est caractérisée par deux fonctions :

— **Fonction de message :** $M_t(h_V^t, h_W^t, e_{VW})$

— **Fonction de mise à jour des sommets :** $U_t(h_V^t, m_V^{t+1})$
  Les états cachés sont mis à jour comme suit :

$$m_V^{t+1} = \sum_{W \in N(V)} M_t(h_V^t, h_W^t, e_{VW})$$

$$h_V^{t+1} = U_t(h_V^t, m_V^{t+1})$$

où h_V^0 est une fonction de l'atome initial des caractéristiques x_V, et $N(V)$ est l'ensemble des nœuds voisins de V dans le graphe.

— **La phase Readout**

La phase de lecture combine les états initial et final du nœud atome pour produire un embedding graphe unique g. La fonction de lecture du graphe R est définie comme suit :

$$g = R(H_n, H_0)$$

où H_n et H_0 représentent respectivement les états des nœuds final et initial.

Réseaux convolutifs de graphes

Pour intégrer efficacement les informations structurelles encodées dans la matrice d'adjacence $\mathbf{A}$ avec les caractéristiques atomiques représentées dans la matrice de caractéristiques $\mathbf{F}$, comme indiqué dans la figure 3.6, nous utilisons des couches Réseaux convolutifs de graphes (GCN) [71].

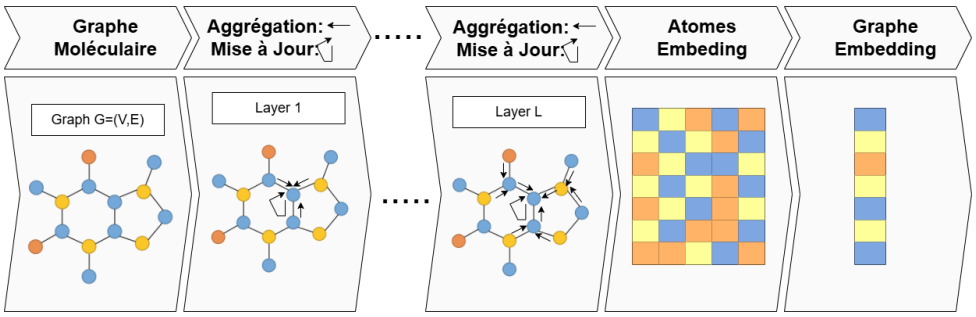


FIGURE 3.6 – Réseaux convolutifs de graphes

Chaque couche GCN fonctionne selon un processus d'agrégation et de mise à jour qui combine les informations des nœuds voisins. La matrice d'adjacence normalisée $\tilde{\mathbf{A}}$ est précalculée comme suit :

$$\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\hat{\mathbf{A}}\mathbf{D}^{-1/2}$$

où $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ est la matrice d'adjacence avec des auto-connexions, et $\mathbf{D}$ est une matrice diagonale des degrés avec des entrées :

$$\mathbf{D}_{ii} = \sum_j \hat{\mathbf{A}}_{ij}$$

Phase d'agrégation : Pour chaque nœud, les informations des nœuds voisins sont agrégées en utilisant la matrice d'adjacence normalisée :

$$\mathbf{H}_{\text{aggr}}^{(l)} = \tilde{\mathbf{A}}\mathbf{H}^{(l)}$$

Phase de mise à jour : Les caractéristiques agrégées sont ensuite transformées par une transformation linéaire suivie d'une fonction d'activation non linéaire :

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{H}^{(l)}_{\text{aggr}}\mathbf{W}^{(l)}\right)$$

En combinant les deux phases, l'opération complète d'une couche GCN s'écrit :

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)$$

où $\mathbf{H}^{(0)} = \mathbf{F}$ est la matrice de caractéristiques des nœuds initiale, $\sigma(\cdot)$ est une fonction d'activation non linéaire (telle que ReLU), et $\mathbf{W}^{(l)}$ est une matrice de poids entraînable de la couche l.

Dans notre implémentation, nous utilisons deux couches GCN consécutives. Le vecteur d'embedding final $\mathbf{E}$ est calculé comme suit :

$$\mathbf{H}^{(1)} = \text{ReLU}\left(\tilde{\mathbf{A}}\mathbf{F}\mathbf{W}^{(0)}\right) \tag{3.1}$$

$$\mathbf{E} = \tilde{\mathbf{A}}\mathbf{H}^{(1)}\mathbf{W}^{(1)} \tag{3.2}$$

Cette formulation peut être exprimée de manière compacte comme :

$$\mathbf{E} = f(\mathbf{F}, \mathbf{A}) = \tilde{\mathbf{A}}\text{ReLU}\left(\tilde{\mathbf{A}}\mathbf{F}\mathbf{W}^{(0)}\right)\mathbf{W}^{(1)}$$

Cette architecture GCN permet de capturer efficacement la structure topologique du graphe moléculaire tout en apprenant des représentations riches des caractéristiques atomiques, constituant ainsi la base de notre approche d'encodage moléculaire.

3.3.2 Modèles d'apprentissage profond pour MG

Les modèles d'apprentissage profond pour la génération moléculaire sont apparus comme des outils de transformation dans la découverte de médicaments, permettant la création automatisée de nouvelles molécules avec des propriétés thérapeutiques souhaitées. Ces modèles exploitent des architectures neuronales avancées pour explorer efficacement le vaste espace chimique, allant au-delà des méthodes traditionnelles qui reposent sur la conception manuelle et les algorithmes basés sur des règles. Les approches clés comprennent les encodeurs automatiques variationnels (VAE) et les réseaux antagonistes génératifs (GAN). Les VAE génèrent des structures moléculaires diverses en apprenant à représenter continuellement des molécules, tandis que les GAN utilisent un entraînement contradictoire pour affiner la génération de molécules, produisant des candidats similaires aux médicaments avec des propriétés optimisées [105, 162].

graphe autoencodeurs variationnels

Le modèle d'autocodeur variationnel de graphe (GVAE), en se concentrant sur l'encodeur basé sur les réseaux convolutifs de graphe (GCNs), la astuce de reparamétrisation et le décodeur [72], comme indiqué dans la figure suivante 3.7 :

Encoder (basé sur GCN)

L'encodeur d'un GVAE met en correspondance le graphe d'entrée $G = (A, X)$ avec une représentation latente z. Il utilise un GCN à deux couches pour produire les paramètres de la distribution latente :

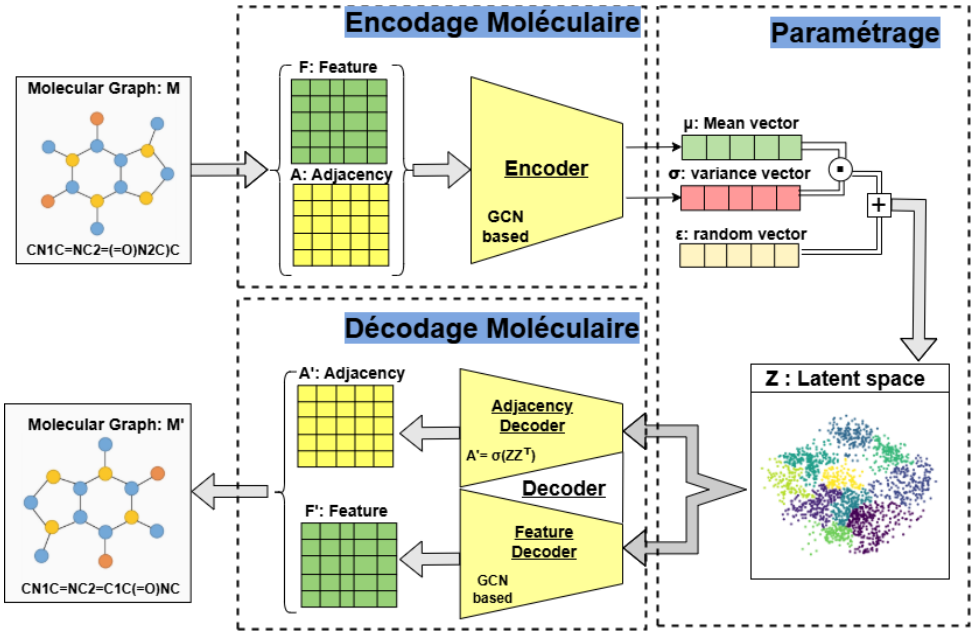


FIGURE 3.7 – graphe autoencodeurs variationnels

$$\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{3.3}$$

$$H^{(1)} = \mathrm{ReLU}(\tilde{A} X W^{(1)}) \tag{3.4}$$

$$\boldsymbol{\mu} = \tilde{A} H^{(1)} W^{(\mu)} \tag{3.5}$$

$$\log \boldsymbol{\sigma}^2 = \tilde{A} H^{(1)} W^{(\sigma)} \tag{3.6}$$

où $\tilde{A}$ est la matrice de contiguïté normalisée symétriquement, D est la matrice de degré diagonal, $W^{(1)}$ est la matrice de poids de la première couche du CNG, $W^{(\mu)}$ et $W^{(\sigma)}$ sont les matrices de poids pour les sorties moyennes et log-variance, respectivement, et ReLU est la fonction d'activation de l'unité linéaire rectifiée.

Astuce de reparamétrisation

La technique de reparamétrisation permet de rétropropager les gradients par le biais du processus d'échantillonnage stochastique. Au lieu d'échantillonner directement à partir de la distribution définie par $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$, nous prenons un échantillon de la distribution normale standard et le transformons :

$$z = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon$$

tel que

$$\epsilon \sim \mathcal{N}(0, I)$$

est un vecteur de bruit échantillonné à partir d'une distribution normale standard, et $\odot$ dénote la multiplication par élément.

Décodeur (basé sur GCN d'adjacence et les caractéristiques des décodeurs)

Le décodeur reconstruit le graphe d'entrée à partir de la représentation latente z. Il se compose de deux parties : un décodeur d'adjacence et un décodeur de caractéristique.

Le décodeur dd'djacence génère la matrice d'adjacence reconstruite $\hat{A}$ en utilisant un produit interne entre les variables latentes :

$$\hat{A} = \sigma(zz^T)$$

où σest la fonction sigmoïde logistique.

Le décodeur de caractéristiques reconstruit les caractéristiques du noeud $\hat{X}$utiliser une autre couche GCN :

$$\hat{X} = \tilde{A}zW^{(2)}$$

tel que $W^{(2)}$ est la matrice de poids du décodeur de caractéristiques GCN couche.

réseaux adversarial génératifs

Les réseaux antagonistes génératifs (GAN) sont une classe de cadres d'apprentissage automatique conçus pour générer de nouvelles instances de données qui ressemblent à un ensemble de données d'entraînement donné. Présentés par Ian Goodfellow et al. en 2014, les GAN sont constitués de deux réseaux neuronaux, le **générateur** et le **discriminateur**, qui sont formés simultanément par des processus contradictoires [73], comme suit dans la figure 3.8.

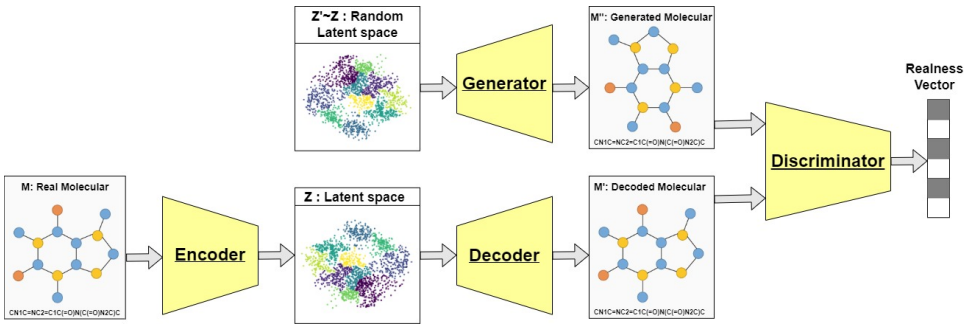


FIGURE 3.8 – réseaux adversarial génératifs

Le modèle **générateur** G est un réseau neuronal qui prend du bruit aléatoire comme entrée et génère des échantillons de données synthétiques. Son objectif est de produire des données qui ne sont pas confondues avec les données réelles [74]. Le générateur est représenté comme suit :

$$G(Z') = M'$$

tel que Z' est un vecteur de bruit aléatoire échantillonné à partir d'une distribution antérieure (souvent une distribution gaussienne), et M'est l'échantillon de données généré.

Le modèle **discriminateur** D qui estime la probabilité qu'un échantillon provienne des données de formation plutôt que G [96]. Le discriminateur peut être représenté comme suit :

$$D(G(Z')) = \mathrm{R}$$

où R est le vecteur de la réalité de la molécule.

Conclusion

Ce chapitre a exploré le potentiel transformateur de l'apprentissage profond pour révolutionner la découverte de médicaments, en se concentrant spécifiquement sur les progrès dans la

prédiction des propriétés moléculaires (MPP) et la génération moléculaire (MG). Nous avons suivi l'évolution des méthodes traditionnelles, souvent laborieuses, aux modèles sophistiqués d'apprentissage profond basés sur des graphes. La capacité de ces modèles, en particulier des réseaux neuronaux graphiques (GNN), des réseaux neuronaux passant par des messages (MPN), et des réseaux convolutionnels graphiques (GNC), à capturer efficacement les informations structurelles et chimiques complexes inhérentes aux molécules a conduit à des améliorations significatives dans la précision prédictive pour des propriétés cruciales comme la solubilité, la biodisponibilité et la toxicité. En outre, le développement de modèles génératifs comme les graphe autoencodeurs variationnels s (GVAEs) et les réseaux adversarial génératifs s (GANs) a ouvert de nouvelles voies pour concevoir des molécules médicamenteuses innovantes avec des profils thérapeutiques adaptés, accélérant la recherche de traitements efficaces.

La clé de ces progrès réside dans le passage à des représentations de molécules basées sur des graphes. En traitant les molécules comme des graphes, avec des atomes comme nœuds et des liaisons comme arêtes, ces modèles peuvent tirer parti de la richesse des informations topologiques cruciales pour comprendre le comportement moléculaire. Nous avons détaillé comment les chaînes SMILES peuvent être converties efficacement en représentations graphiques, permettant l'application d'architectures GNN puissantes. La construction de matrices de caractéristiques et d'adjacence, comme nous l'avons vu, fournit une base solide pour saisir les propriétés atomiques et la connectivité moléculaire, qui sont des intrants essentiels pour les GNN. Le cadre de passage des messages dans les MPNN et les opérations convolutionnelles dans les GCN permettent la propagation de l'information à travers le graphe moléculaire, permettant au modèle d'apprendre des relations complexes et de générer des représentations perspicaces.

Le développement et l'amélioration continus des modèles d'apprentissage profond pour les MPP et les MG promettent d'accélérer davantage le processus de découverte de médicaments, de réduire les coûts et, finalement, de mener au développement de thérapies moléculaires plus efficaces et ciblées. L'intégration de ces outils puissants dans le processus de découverte de médicaments représente une étape importante vers un avenir où de nouveaux traitements peuvent être développés plus rapidement et efficacement, offrant ainsi l'espoir de répondre aux besoins médicaux actuellement non satisfaits.

Chapitre 4

Les approches d'apprentissage profond basées sur les graphes et l'approche GMPP-NN

Introduction

Ce chapitre présente une étude complète sur les méthodes d'apprentissage profond basées sur des graphes pour la prédiction de propriétés moléculaires, en se concentrant spécifiquement sur une nouvelle architecture nommée graphe Molecular Property Prediction Neural Network (GMPP-NN). GMPP-NN utilise une approche de réseau neuronal de passage de messages (MPNN) pour apprendre les intégrations moléculaires à partir de données structurées par graphe, qui sont ensuite utilisées pour des tâches de classification et de régression. Les ensembles de données utilisés pour cette recherche ont été obtenus à partir du référentiel MoleculeNet, y compris le VIH, BACE, BBBP et ClinTox, couvrant à la fois les domaines biophysique et physiologique. Chaque ensemble de données présente un défi unique, comme la prédiction de la pénétration de la barrière hémato-encéphalique ou la détermination de la toxicité clinique des composés, ce qui souligne la polyvalence de l'architecture proposée.

Plusieurs architectures avancées basées sur des graphes ont été explorées dans cette étude, telles que ABT-MPNN et ChemRL-GEM, qui intègrent respectivement les mécanismes d'attention et l'information géométrique tridimensionnelle pour améliorer l'apprentissage de la représentation moléculaire. Le GMPP-NN proposé, cependant, se distingue par son utilisation efficace de représentations moléculaires basées sur des graphes et un mécanisme itératif de passage de messages qui capture à la fois les environnements chimiques locaux et les interactions à longue portée au sein des molécules. Le modèle est formé et évalué sur plusieurs ensembles de données de référence, ce qui lui permet d'atteindre une performance de pointe dans la prédiction de diverses propriétés moléculaires, comme l'ont démontré les analyses comparatives.

Les sections suivantes fournissent des descriptions détaillées des ensembles de données et des méthodes de prédiction des propriétés moléculaires utilisées dans cette étude. Nous explorerons les caractéristiques des ensembles de données, y compris la collecte et le prétraitement des données, avant de nous pencher sur l'architecture du modèle GMPP-NN et son évaluation comparative des performances par rapport à d'autres modèles contemporains. Le but de ce chapitre est de mettre en évidence le potentiel des architectures d'apprentissage profond basées sur les graphes pour prédire avec précision les propriétés moléculaires et comment ces méthodes peuvent contribuer aux progrès dans la découverte de médicaments et d'autres applications en chimie informatique.

4.1 Description et collecte des données

Pour cette étude, les ensembles de données ont été obtenus à partir du jeu de données de référence MoleculeNet. Quatre ensembles de données (VIH, BACE, BBBP et ClinTox) ont été utilisés pour le défi de classification, couvrant deux domaines : la physiologie et la biophysique.

— **Ensemble de données HIV :** Les données sur le VIH proviennent du programme de traitement thérapeutique du SIDA, qui a testé environ 40000 composés pour bloquer la réplication du VIH. Les résultats ont été classés en deux catégories : les composés inactifs qui ont été confirmés et les composés actifs qui ont été confirmés [151].

— **Ensemble de données BACE :** L'ensemble de données BACE contient des données quantitatives et qualitatives sur les capacités de liaison de divers inhibiteurs de la bêta-sécrétase 1 chez l'homme. Le jeu de données comprend des valeurs expérimentales publiées dans la littérature scientifique et se compose de 1522 composés chimiques avec des étiquettes de classification binaire axées sur une cible protéique unique [152].

— **Ensemble de données BBBP :** Le jeu de données BBBP (Blood Brain Barrier Penetration) est un projet de recherche axé sur la modélisation et la prédiction de la perméabilité des barrières, un aspect essentiel dans le développement de médicaments ciblant le système nerveux central. Ce jeu de données contient des étiquettes de classification binaire pour plus de 2000 composés chimiques [153].

— **Ensemble de données ClinTox :** L'ensemble de données ClinTox permet de faire la différence entre les produits pharmaceutiques approuvés par la FDA et les composés qui ont échoué dans les essais cliniques en raison de problèmes liés à la toxicité. L'ensemble de données se compose de 1491 composés pharmacologiques et comprend deux tâches de classification : Prédire la toxicité des essais cliniques et déterminer le statut d'approbation de la FDA [154].

Le modèle a été formé à l'aide de l'ensemble de données d'apprentissage, et ses hyperparamètres ont été ajustés et optimisés en fonction des résultats de l'ensemble de données de validation. Enfin, la performance du modèle a été évaluée à l'aide de l'ensemble de données de test.

TABLE 4.1 – Résumé des ensembles de données

Datasets	Category	Data Descriptions	Compounds
HIV	Biophysics	Inhibition of HIV replication	41,127
BACE	Biophysics	Inhibition of human beta-secretase 1	1,513
ClinTox	Physiology	Toxicity	1,478
BBBP	Physiology	Ability to penetrate the blood-brain barrier	2,039

4.1.1 Fractionnement de l'ensemble de données

Le jeu de données global était divisé en trois sous-ensembles : un jeu de données d'entraînement comprenant 75% des données totales, un jeu de données de validation comprenant 20% des données totales et un jeu de données de test comprenant 5% des données totales. Dans les quatre ensembles de données, le format SMILES est utilisé pour représenter les structures moléculaires des composés. L'architecture proposée vise à résoudre les problèmes de classification dans ces ensembles de données, comme résumé dans le tableau 4.1.

4.2 Méthodes basées sur des graphes pour la prédiction de propriétés moleculaires

4.2.1 ABT-MPNN : un réseau neuronal de passage de messages basé sur des liaisons atomiques pour la prédiction de propriétés moléculaires

Le réseau de neurones à passage de messages basé sur un transformateur Atom-Bond (ABT-MPNN) est une nouvelle architecture qui intègre les forces des mécanismes d'attention basés sur un transformateur avec les MPNNs. Cette combinaison permet au modèle de capturer à la fois les environnements chimiques locaux et la structure moléculaire globale en introduisant des mécanismes d'attention au niveau des liaisons et des atomes. L'architecture d'ABT-MPNN est conçue pour améliorer la prédiction moléculaire de propriété en incorporant l'attention détaillée pendant les phases message-passage et lecture [156].

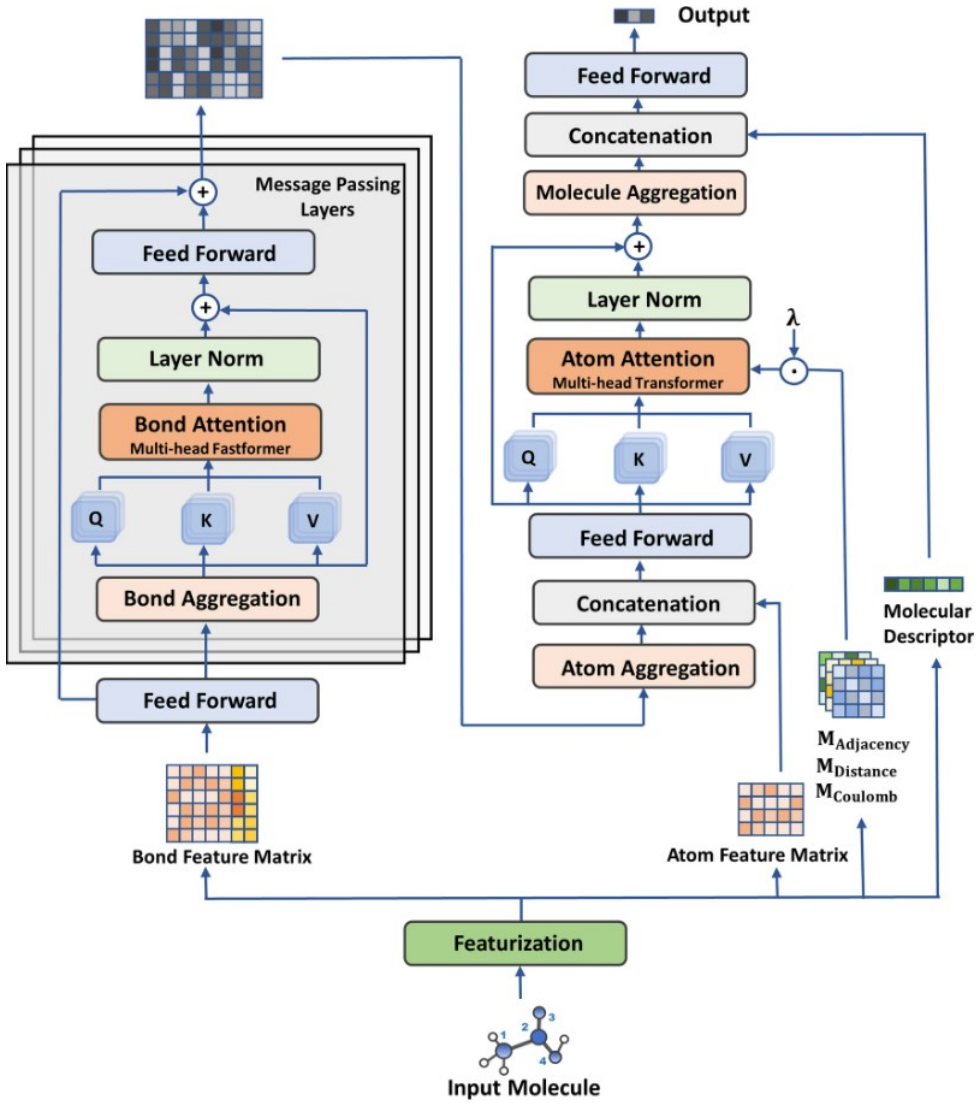


FIGURE 4.1 – L'architecture ABT-MPNN

Les étapes clés de l'architecture ABT-MPNN comprennent :

— **Construction du graphique :** Les graphes moléculaires sont construits à partir de chaînes SMILES, où les atomes et les liaisons sont représentés respectivement comme des noeuds

et des arêtes. De plus, des matrices interatomiques, telles que les matrices de adjacence, de distance et de Coulomb, sont générées pour fournir des caractéristiques spatiales et électrostatiques pour chaque atome.

— **Bond Attention in Message Passing :** Le mécanisme d'attention au niveau des liaisons est introduit pendant la phase de passage du message pour peser l'importance des différentes liaisons dans le graphe moléculaire. Pour chaque liaison b_{vw}, un score d'attention est calculé à l'aide d'un mécanisme d'auto-attention à plusieurs têtes, qui s'applique aux vecteurs de liaison h_{vw}. Le message de liaison est mis à jour par :

$$m_{vw}^{(t+1)} = \sum_{k \in N(v)} h_{kv}^{(t)} - h_{wv}^{(t)},$$

suivi d'un mécanisme d'attention de liaison qui calcule une nouvelle intégration de liaison à l'aide d'une matrice de poids apprenable W_b and a ReLU activation :

$$h_{vw}^{(t+1)} = \text{ReLU}(W_b \cdot \text{Concat}(m_{vw}^{(t+1)}, h_{vw}^{(0)})).$$

Ce mécanisme permet au modèle de hiérarchiser dynamiquement les liaisons critiques pour la prédiction des propriétés moléculaires.

— **Atome Attention et génération d'embedding :** après l'attention de liaison, des embeddings au niveau atomique sont générés en agrégeant les caractéristiques de liaison des atomes voisins. Le modèle applique également un mécanisme d'attention au niveau de l'atome en utilisant l'autoattention à tête multiple, qui intègre des caractéristiques spatiales et électrostatiques supplémentaires (p. ex., distances topologiques et interactions coulombiennes) pour affiner les incorporations atomiques :

$$m_{v}^{(t+1)} = \text{ReLU}(W_o \cdot \text{Concat}(x_v, \sum_{w \in N(v)} h_{vw}^{(T)})),$$

tel que W_o est une matrice de poids apprise appliquée aux caractéristiques atomiques concaténées et aux états de liaison agrégés. Ces incorporations atomiques sont ensuite agrégées en une seule incorporation de molécules à l'aide d'une phase de lecture.

— **Prédiction et évaluation :** la représentation moléculaire finale, obtenue en ajoutant des incorporations atomiques, est transmise à travers un réseau neuronal de prédiction de propriétés. L'architecture a démontré un rendement exceptionnel sur une variété de tâches de classification et de régression à travers plusieurs ensembles de données (p.ex., HIV, ClinTox, et QM8).

4.2.2 ChemRL-GEM : Apprentissage de la représentation moléculaire améliorée pour la prédiction des propriétés

ChemRL-GEM est une architecture d'apprentissage profond qui améliore la prédiction des propriétés moléculaires en intégrant à la fois l'information topologique et géométrique des molécules. Contrairement aux réseaux de neurones traditionnels basés sur des graphes qui se concentrent uniquement sur la connectivité des atomes, ChemRL-GEM exploite les structures spatiales tridimensionnelles (3D), offrant une représentation moléculaire plus complète. Cette approche est particulièrement utile pour les propriétés qui sont fortement influencées par la géométrie moléculaire, telles que l'affinité de liaison, la réactivité chimique et la solubilité [155].

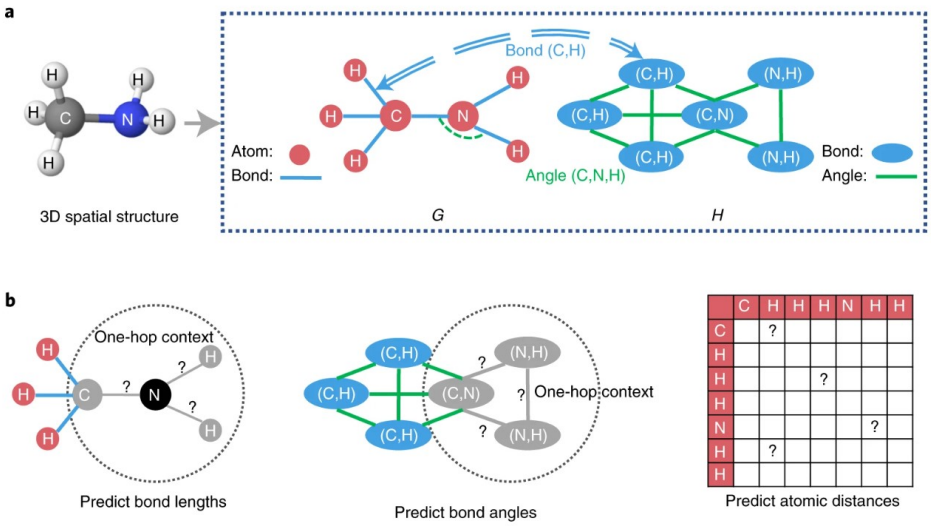


FIGURE 4.2 – L'architecture GEM

L'architecture ChemRL-GEM comprend les éléments clés suivants :

— **Construction de géométrie moléculaire :** Les géométries moléculaires sont construites en utilisant des coordonnées 3D d'atomes, qui sont obtenues à partir de structures cristallines dérivées expérimentalement ou prédites par des méthodes informatiques telles que la théorie fonctionnelle de densité (DFT). En plus des caractéristiques de noeud (atome) et des caractéristiques d'arête (liaison), des caractéristiques géométriques telles que les angles de liaison et les angles dièdres sont incluses pour représenter les relations spatiales entre les atomes.

— **Transmission de messages à géométrie sensible :** Le cœur de ChemRL-GEM est un réseau neuronal de passage de messages (MPNN) amélioré par la géométrie qui met à jour les intégrations atomiques en incorporant des environnements atomiques locaux et des informations géométriques 3D. Le processus de passage de messages tient compte non seulement de la connectivité des liaisons, mais aussi de l'arrangement spatial des atomes. L'état caché $h_v^{(t)}$ d'un atome v est mis à jour en agrégeant les messages des atomes voisins $w \in N(v)$ en fonction de leurs distances et relations angulaires :

$$m_v^{(t+1)} = \sum_{w \in N(v)} M_t(h_v^{(t)}, h_w^{(t)}, d_{vw}, \theta_{vwx}),$$

où d_{vw}représente la distance entre les atomes$v$ et w, et θ_{vwx} est l'angle formé par les atomes v, $w$et un troisième atome voisin x. Ce passage de messages sensible à la géométrie permet au modèle de capturer non seulement des liaisons chimiques, mais aussi des motifs géométriques qui influencent les propriétés moléculaires.

— **Génération de l'intégration géométrique :** après plusieurs itérations de passage de message, les intégrations atomiques mises à jour sont agrégées dans une intégration au niveau du graphe g représente la molécule dans son ensemble. Cette phase de lecture intègre à la fois la structure chimique et la géométrie 3D pour générer une représentation moléculaire riche.

— **Prédiction :** l'intégration moléculaire agrégée est passée par un perceptron multicouche (MLP) pour la prédiction de propriétés. Le modèle peut être utilisé pour les tâches de classification (p. ex., la prévision de la toxicité moléculaire) et de régression (p. ex., l'estimation des niveaux d'énergie moléculaire). De plus, ChemRL-GEM bénéficie d'un réglage

fin grâce à l'apprentissage autosupervisé sur de grands ensembles de données de géométries moléculaires, ce qui aide à améliorer la généralisation sur les petits ensembles de données étiquetés.

— **Évaluation des performances :** le ChemRL-GEM est évalué sur une plage de repères de prédiction des propriétés moléculaires à l'aide de mesures telles que l'erreur absolue moyenne (MAE) et la surface sous la courbe (AUC), qui sont utilisées pour évaluer les performances du modèle, avec le ChemRL-GEM montrant des résultats améliorés par rapport aux modèles traditionnels de GNN qui manquent d'informations géométriques.

En intégrant la géométrie moléculaire dans le processus de transmission des messages, ChemRL-GEM offre une représentation plus détaillée et précise des molécules, ce qui permet d'améliorer les prédictions pour les tâches où les relations spatiales 3D jouent un rôle crucial.

4.2.3 GMPP-NN : une architecture d'apprentissage profond pour la prédiction des propriétés moléculaires des graphes

Le réseau de neurones GMPP-NN (Graph Molecular Property Prediction Neural Network) est une architecture spécialisée en apprentissage profond conçue pour la prédiction des propriétés moléculaires. Il utilise un réseau neuronal de passage de messages (MPNN) pour apprendre les intégrations moléculaires à partir de données structurées par graphe générées à partir des représentations du système d'entrée en ligne d'entrée moléculaire simplifié (SMILES). Ces embeddings encapsulent les propriétés chimiques et structurelles des molécules, qui sont ensuite passées à travers un perceptron multicouche (MLP) pour la tâche de classification [157].

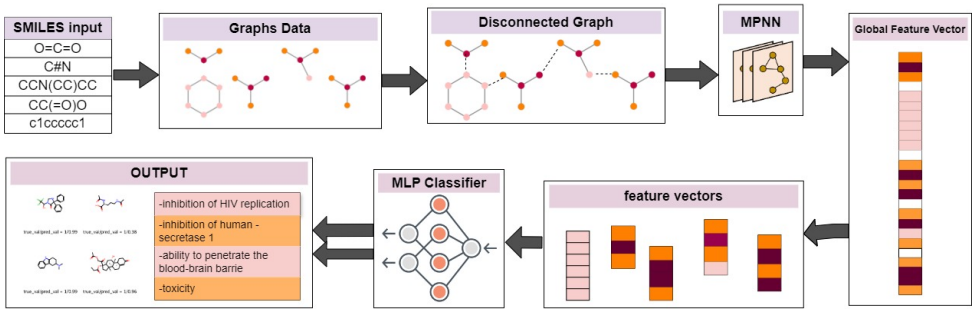


FIGURE 4.3 – L'architecture proposée du GMPP-NN

La méthodologie globale peut être divisée en quatre phases clés :

— **Construction des graphiques :** les graphes moléculaires sont générés à partir de chaînes SMILES, où les nœuds représentent des atomes et les arêtes représentent des liaisons. L'entrée dans le réseau est un graphe moléculaire $G = (V, E)$ avec des caractéristiques de nœud x_v correspondant aux propriétés atomiques et aux caractéristiques des arêtes e_{vw} correspondant aux types de liaison entre les atomes v et w.

— **Génération de messages et intégration :** le MPNN met à jour les états cachés d'une manière itérative $h_v^{(t)}$ de chaque atome $v \in V$ par le message passant à travers les liaisons. À chaque itération, $t$les messages provenant d'atomes voisins $w \in N(v)$ sont agrégés et combinés avec l'état actuel $h_v^{(t)}$. Formellement, cette étape de mise à jour est effectuée comme suit :

$$m_v^{(t+1)} = \sum_{w \in N(v)} M_t(h_v^{(t)}, h_w^{(t)}, e_{vw}),$$

où $M_t$est la fonction de message $N(v)$désigne les voisins de vLa fonction de mise à jour des sommets U_t est ensuite utilisé pour mettre à jour l'état caché $h_v^{(t+1)}$pour la prochaine itération :

$$h_v^{(t+1)} = U_t(h_v^{(t)}, m_v^{(t+1)}).$$

Après T itérations de passage de message, une intégration au niveau du graphe g est générée par une fonction de lecture R qui agrège les états cachés de tous les atomes dans le graphe :

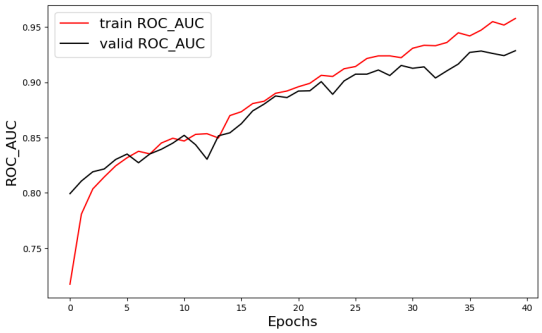
$$g = R(\{h_v^{(T)} \mid v \in V\}).$$

— **Le classificateur MLP :** L'intégration du graphique agrégé g est passée à travers un perceptron multicouche (MLP) pour prédire les propriétés moléculaires. L'architecture est flexible et peut gérer à la fois les tâches de classification et de régression, telles que la prédiction de la pénétration de la barrière hémato-encéphalique (BBBP), la toxicité clinique (ClinTox), ou de l'inhibition de BACE pour les médicaments contre la maladie d'Alzheimer.

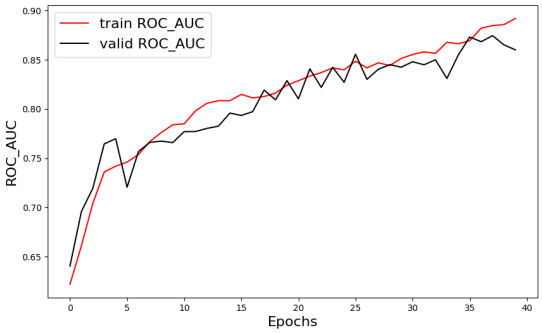
— **Métriques de performance :** la performance de GMPP-NN est évaluée à l'aide de diverses mesures, comme la zone sous la courbe ROC (AUC) et la courbe de précision-rappel (PRC), selon la nature de la prédiction moléculaire.

Modèle de formation et de validation de la performance

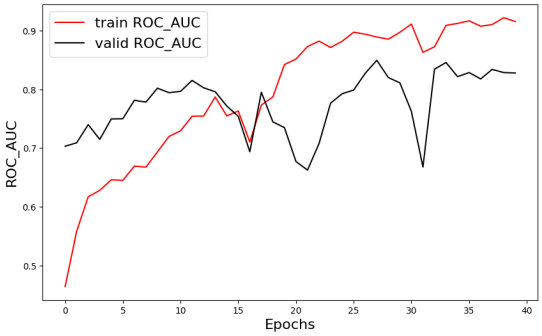
La formation au modèle GMPP-NN (graphe Molecular Property Prediction Neural Network) sur quatre ensembles de données (BBBP, HIV, BACE et ClinTox) a démontré des améliorations significatives, comme le montre la figure 4.4.



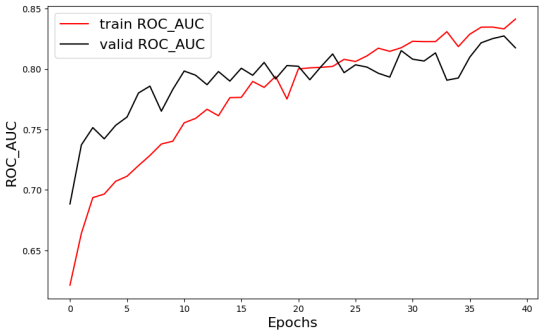
(a) BBBP Performance



(b) BACE Performance



(c) ClinTox Performance



(d) HIV Performance

FIGURE 4.4 – Performance de l'entrainement et de la validation des courbes ROC

Le modèle GMPP-NN a montré un fort pouvoir discriminant après 40 époques dans divers ensembles de données, avec les scores suivants de validation de la courbe ROC : BBBP (0,9285), VIH (0,8175), ClinTox (0,8280) et BACE (0,8599). Cela indique la capacité du modèle à prédire différentes propriétés moléculaires.

4.3 Analyse comparative du rendement de GMPP-NN avec les autres études

La performance du réseau de neurones GMPP-NN (Graph Molecular Property Prediction Neural Network) a été rigoureusement évaluée contre plusieurs modèles contemporains dans quatre ensembles de données distincts s : VIH, BACE, BBBP et ClinTox. Les résultats, résumés dans le tableau 4.2, démontrent que le GMPP-NN surpasse constamment ses homologues dans trois des quatre ensembles de données, obtenant les meilleurs scores de la courbe ROC.

TABLE 4.2 – Les performances sur tous les ensembles de données

Model Architecture	Featurization Method	HIV	BACE	BBBP	ClinTox
ABT-MPNN	Graph-based	0.809	–	–	0.904
GEM	Graph-based	0.769	0.856	0.724	0.825
GMPP-NN (Ours)	Graph-based	**0.8677**	0.8608	**0.9186**	**0.9795**

— **HIV Dataset** : Le GMPP-NN a obtenu un score de **0.8677**, surpassant de manière significative le score d'ABT-MPNN de **0.809** et le score du modèle GEM de **0.769**. Cela indique une capacité robuste de GMPP-NN à prédire efficacement l'inhibition du VIH,

suggérant que son architecture est particulièrement bien adaptée pour ce type de propriété moléculaire.

— **BACE Dataset** : Sur ce jeu de données, GMPP-NN a enregistré un score de **0.8608**, qui est compétitive mais légèrement inférieure à la performance d'ABT-MPNN (non déclarée) et à la note de GEM de **0.856**. Ce résultat met en évidence que GMPP-NN conserve de fortes capacités prédictives, Il peut y avoir des aspects spécifiques de l'inhibition du BACE qui pourraient bénéficier d'une optimisation plus poussée ou de méthodes spécialisées d'extraction des caractéristiques.

— **BBBP Dataset** : Dans ce cas, GMPP-NN excellé avec un score de **0.9186**, surpassant à la fois ABT-MPNN et GEM, qui n'ont pas rapporté de scores pour cet ensemble de données. Le score élevé reflète l'efficacité du GMPP-NN dans la prédiction de la pénétration de la barrière hémato-encéphalique, mise en valeur de son potentiel d'applications dans le développement de médicaments ciblant les maladies du système nerveux central.

— **ClinTox Dataset** : Le modèle a obtenu un score impressionnant de **0.9795**, dépassant de loin les scores d'autres modèles tels que GEM (**0.825**) et ABT-MPNN (**0.904**). Cette performance exceptionnelle souligne la capacité du modèle à prédire la toxicité clinique, ce qui est essentiel pour évaluer l'innocuité des médicaments.

Les résultats indiquent que le choix de la méthode de caractérisation et du modèle d'intégration joue un rôle crucial dans la détermination des performances prédictives. Les résultats supérieurs obtenus par GMPP-NN suggèrent que son architecture capture efficacement des interactions et propriétés moléculaires complexes grâce à des techniques d'apprentissage avancées basées sur des graphes.

Conclusion

Ce chapitre présente le réseau de neurones GMPP-NN (Graph Molecular Property Prediction Neural Network), une nouvelle architecture d'apprentissage profond conçue pour prédire les propriétés moléculaires. En utilisant des représentations de molécules basées sur des graphes et en utilisant une approche de réseau de passage de messages (MPNN), GMPP-NN a démontré des performances remarquables dans de multiples ensembles de données de référence dans les domaines de la biophysique et de la physiologie. Notre analyse comparative a révélé que le GMPP-NN surpassait systématiquement les modèles de pointe tels que l'ABT-MPNN et le GEM sur trois ensembles de données sur quatre (VIH, BBBP et ClinTox), tout en restant compétitif sur l'ensemble de données BACE. Ces résultats soulignent l'efficacité de l'architecture du GMPP-NN dans la capture d'interactions et de propriétés moléculaires complexes grâce à des techniques avancées d'apprentissage basées sur les graphes. La capacité du modèle à bien fonctionner de façon constante dans divers ensembles de données suggère qu'il pourrait être largement applicable à la découverte de médicaments. l'évaluation toxicologique et d'autres domaines de la chimie informatique. Le succès du GMPP-NN peut être attribué à plusieurs facteurs clés :

— L'utilisation efficace de représentations moléculaires à base de graphes, qui préservent les informations structurelles et chimiques.

— Mécanisme itératif de passage des messages qui permet la capture d'interactions à longue portée au sein des molécules.

— L'architecture flexible qui peut s'adapter aux tâches de classification et de régression.

Chapitre 5

Les approches d'apprentissage profond basées sur les graphes et l'approche ME&PP-MG&RC

Introduction

L'avancement rapide des techniques d'apprentissage automatique, en particulier les techniques d'apprentissage profond, a considérablement transformé le paysage de la chimie computationnelle, notamment dans les domaines de la prédiction et de la génération des propriétés moléculaires. Parmi les diverses méthodologies, l'intégration de modèles d'apprentissage profond est apparue comme une approche puissante pour prédire les propriétés moléculaires et générer des composés nouveaux. Cette thèse porte sur l'évaluation d'un nouveau cadre, appelé ME&PP-MG&RC, qui combine un codeur moléculaire et un prédicteur de propriétés (ME&PP) avec un générateur moléculaire et un classifieur de réalité (MG&RC).

Le composant ME&PP est conçu pour capturer efficacement les caractéristiques structurelles et les attributs des nœuds des molécules, les transformant en embeddings de dimension inférieure qui facilitent des prédictions précises des propriétés. Ceci est crucial pour comprendre la relation entre la structure moléculaire et ses propriétés correspondantes, qui est un aspect fondamental de la découverte de médicaments et de la science des matériaux. Des études récentes ont démontré que les modèles d'apprentissage profond peuvent surpasser les méthodes traditionnelles dans ce domaine, grâce à leur capacité d'apprendre des représentations complexes directement à partir de données sans ingénierie étendue des fonctionnalités [174].

En parallèle, la composante MG&RC vise a générer de nouvelles structures moléculaires tout en assurant leur validité chimique. Cette double approche améliore non seulement la précision de prédiction des propriétés, mais ouvre également la voie à une synthèse de composés innovante. L'importance des modèles génératifs en chimie ne peut être surestimée ; ils permettent aux chercheurs d'explorer efficacement de vastes espaces chimiques, conduisant à la découverte de nouveaux composés aux propriétés souhaitables [175].

Pour contextualiser la performance de ME&PP-MG&RC, nous comparerons ses résultats avec ceux obtenus par GraphVAE, un modèle bien établi dans le domaine de la modélisation générative basée sur des graphes. GraphVAE exploite un cadre d'encodage automatique variationnel pour générer de petits graphes, ce qui le rend particulièrement adapté aux tâches de génération moléculaire. Son approche permet la représentation de structures de graphe dans un espace latent continu, facilitant ainsi la génération de graphes moléculaires divers et réalistes [176].

Dans cette analyse comparative, nous mettrons en évidence des mesures clés telles que la précision de la prédiction et la qualité de la génération, fournissant des renseignements sur les forces et les limites de chaque méthode. En examinant ces résultats, nous visons à contribuer au discours actuel sur l'optimisation des approches d'apprentissage automatique pour la prédiction de propriétés moléculaires et la modélisation générative.

Les sections suivantes approfondiront les méthodologies utilisées dans ME&PP-MG&RC et GraphVAE, en élucidant leurs approches et mécanismes opérationnels. Grâce à cette exploration, nous espérons faire la lumière sur la façon dont ces modèles peuvent être affinés pour améliorer leur applicabilité dans la recherche chimique du monde réel.

5.1 Description du dataset

L'étude a utilisé des ensembles de données provenant de la suite MoleculeNet, se concentrer particulièrement sur l'ensemble de données QM9 pour évaluer la performance des modèles d'apprentissage profond dans les tâches de génération et de prédiction de propriétés moléculaires. MoleculeNet est devenu une ressource de premier plan pour les chercheurs en apprentissage automatique dans le domaine de la chimie et des sciences des matériaux grâce à ses ensembles de données bien organisés qui permettent un étalonnage cohérent de diverses tâches de chimie computationnelle.

Le jeu de données QM9 [169] comprend 133,885 petites molécules organiques avec jusqu'à neuf atomes lourds, spécifiquement le carbone, l'azote, l'oxygène et le fluor qui sont représentés en notation SMILES et avec des coordonnées cartésiennes 3D. Ces molécules ont été systématiquement dérivées pour explorer l'espace chimique, en se concentrant sur les structures organiques stables à faible énergie. QM9 offre une compilation complète de 19 propriétés mécaniques quantiques pour chaque molécule, calculées en utilisant la théorie des fonctions de densité (DFT) [128]. Ces propriétés comprennent l'énergie d'atomisation, la mesure spatiale électronique, le moment dipolaire, la capacité thermique et l'écart HOMO-LUMO, entre autres, qui sont fondamentales pour comprendre le comportement moléculaire et la réactivité [170].

Cet ensemble de données joue un rôle essentiel dans l'avancement des applications de chimie informatique et d'apprentissage automatique, car il offre un ensemble normalisé de caractéristiques moléculaires qui peuvent être utilisées pour le benchmarking de divers algorithmes dans des tâches telles que la prédiction de propriétés, l'exploration chimique de l'espace et la modélisation générative. Des études récentes, telles que Gilmer et al. [171], utilisé QM9 pour évaluer les réseaux de neurones passant des messages (MPNNs), démontrant l'importance d'inclure les environnements chimiques locaux et globaux dans la prédiction des propriétés. De même, les travaux de Schütt et al. [172] sur SchNet illustrent l'utilisation de réseaux neuronaux profonds pour capturer les interactions mécaniques quantiques directement à partir des structures moléculaires, ce qui permet d'obtenir une précision significative dans la prédiction des propriétés sur QM9.

La diversité et la richesse de QM9 en font une ressource précieuse pour les modèles d'apprentissage profond qui visent à prédire les propriétés mécaniques quantiques des structures moléculaires. L'importance de cet ensemble de données est évidente dans diverses études comparatives qui établissent des performances de base pour de nouveaux algorithmes à la fois dans l'apprentissage supervisé et les tâches génératives. Par exemple, des modèles génératifs comme GraphVAE et MolecularRNN [173] ont utilisé QM9 pour former des modèles capables de générer des structures moléculaires chimiquement valides, démontrant la polyvalence de l'ensemble de données au-delà de la prédiction des propriétés. En outre, QM9 fournit un repère exigeant

dû aux propriétés mécaniques complexes de tranche de temps qu'il contient, qui exigent des modèles pour saisir avec précision les interactions et les dépendances moléculaires subtiles.

L'ensemble de données QM9 fournit une plate-forme complète et normalisée pour la formation et l'évaluation des modèles en chimie quantique et en science des matériaux, ce qui en fait un composant crucial de la suite MoleculeNet pour la recherche en apprentissage profond. Son utilisation a entraîné des progrès importants dans le développement de modèles d'apprentissage automatique qui peuvent prédire les propriétés moléculaires avec une grande précision et a soutenu de nouvelles approches dans la génération moléculaire et l'optimisation.

5.2 Modèles de génération moléculaire

5.2.1 Aperçu du modèle GraphVAE

Le modèle GraphVAE représente une avancée significative dans la modélisation générative des graphes à l'aide du cadre de l'encodage automatique variationnel (VAE). Il étend la capacité des VAE à manipuler les structures discrètes de graphe en les mappant dans un espace latent continu, de ce fait permettant la génération probabiliste des structures diverses et réalistes de graphe, comme présenté dans la figure 5.1. Cette approche est particulièrement avantageuse pour des applications telles que la génération moléculaire, la modélisation de réseaux sociaux et la synthèse générale de données basée sur des graphes [158].

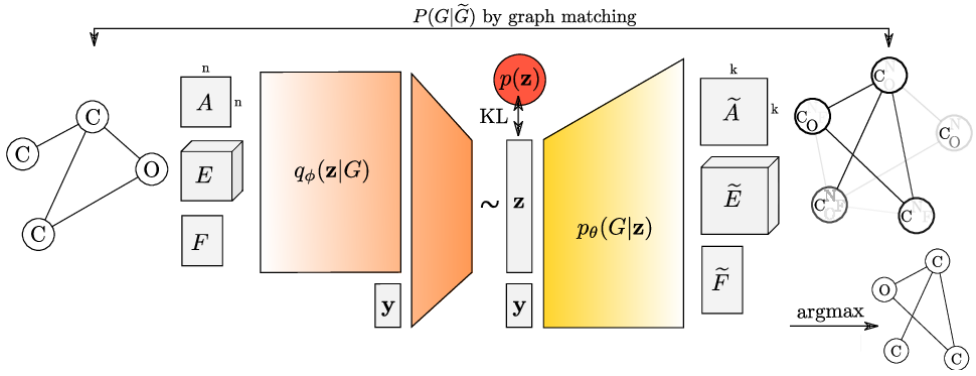


FIGURE 5.1 – L'approche GraphVAE

Représentation graphique

Le cadre GraphVAE représente un graphe $G = (A, E, F)$, où A la matrice de contiguïté indiquant les connexions des nœuds, Ereprésente les attributs de bord, et Fdésigne les attributs de nœud.G peut capturer des informations cruciales sur la topologie du graphe et les caractéristiques des nœuds et des arêtes. L'objectif du modèle est d'apprendre un encodeur qui comprime ce graphe en une représentation latente $\mathbf{z} \in \mathbb{R}^c$ et un décodeur qui reconstruit le graphe à partir de cet espace latent. Cette représentation peut ensuite être utilisée pour générer de nouveaux graphiques qui conservent des caractéristiques structurelles et attribuent des caractéristiques similaires aux exemples de formation.

La représentation capture un large spectre de propriétés des graphes, ce qui le rend approprié pour diverses tâches, y compris la génération de structure chimique et la prédiction des propriétés des matériaux [159]. Par exemple, dans la génération de graphes moléculaires, les attributs des nœuds F peuvent représenter différents types d'atomes, alors que les attributs de

bord E peuvent coder des types de liaison, permettant une représentation nuancée des structures chimiques [160].

Autoencodeur Variational

Dans le cadre du VAE, l'encodeur, désigné par *enc*, approxime la distribution postérieure sur la variable latente $\mathbf{z}$ donnée par le graphe d'entréeG, typiquement en sortant des paramètres d'une distribution gaussienne. le decodeur, *dec*, modélise le processus génératif en reconstruisant le graphe original à partir de la représentation latente. L'objectif global est de minimiser la log-vraisemblance négative $-\log p_\theta(G)$, avec la divergence Kullback-Leibler pour encourager *enc*pour correspondre à un précédent prédéfini*prior*généralement un gaussien isotrope.

La perte de formation pour les VAE est formulée comme suit :

$$\mathcal{L}(\phi, \theta; G) = \mathbb{E}_{enc}[-\log dec] + \mathrm{KL}[enc||prior], \tag{5.1}$$

où la divergence KL régularise l'espace latent pour assurer des transitions fluides entre les graphes, aidant à une meilleure généralisation et continuité pour la génération de graphe [161] l'astuce de reparamétrage de Kingma et Welling [162] est employée pour permettre la rétropropagation à travers le processus d'échantillonnage stochastique.

Décodeur de graphe probabiliste

Le décodeur dans GraphVAE génère une représentation probabiliste d'un graphe $\widetilde{G} = (\widetilde{A}, \widetilde{E}, \widetilde{F})$, où $\widetilde{A}$, $\widetilde{E}$, et $\widetilde{F}$ sont les matrices d'attributs de adjacence, d'arête et de nœud reconstruites, respectivement. Cette représentation probabiliste est utilisée pour prédire l'existence des arêtes (en utilisant la matrice de contiguïté) et les classes d'attributs pour les arêtes et les nœuds.

Le graphe généré est généralement modélisé comme un graphe entièrement connecté avec présence de bord probabiliste, et au moment de l'inférence, une estimation ponctuelle peut être obtenue en appliquant un seuil ou une opération argmax aux probabilités prédites. Cette approche probabiliste permet au modèle de générer diverses structures graphiques, ce qui est crucial pour des applications telles que la découverte de médicaments, où la génération d'une gamme de molécules chimiquement valides est bénéfique [163].

Perte de reconstruction

La perte de reconstruction évalue la probabilité du graphique généré $\widetilde{G}$ conditionnel sur la représentation latente $\mathbf{z}$. Plus précisément, il mesure la qualité de reconstruction de la matrice de contiguïté, des attributs de nœud et des attributs de bord :

$$-\log p(G|\mathbf{z}) = -\lambda_A \log p(A'|\mathbf{z}) - \lambda_F \log p(F|\mathbf{z}) - \lambda_E \log p(E|\mathbf{z}), \tag{5.2}$$

Où λ_A, λ_F, et λ_E sont des hyperparamètres qui contrôlent l'importance relative de chaque terme. Ces hyperparamètres sont généralement définis en fonction de la connaissance du domaine ou par validation croisée. Une telle perte de reconstruction structurée garantit que la topologie (encodée en A) et les attributs du graphe (en F et E) sont fidèlement reconstruits, ce qui conduit à une génération de graphe réaliste et cohérente.

Correspondance graphique

Un défi majeur dans les modèles génératifs de graphes est la comparaison des graphes avec des structures et des ordonnances potentiellement différentes. Pour résoudre ce problème, GraphVAE intègre une étape d'appariement de graphe qui vise à aligner les nœuds entre l'entrée et les graphes générés. Cet alignement est formulé comme un problème de programmation quadratique entière, visant à maximiser une fonction de similarité qui prend en compte à la fois la similitude structurelle et la compatibilité basée sur les attributs entre les nœuds et les arêtes [164]. Cette approche est cruciale pour réduire l'ambiguïté associée aux pertes de reconstruction de graphe, améliorant ainsi la qualité de génération [165].

Architecture de l'encodeur

L'encodeur du GraphVAE utilise des convolutions de graphes conditionnées par les arêtes (ECC) [166],qui étendent le concept standard de réseau neuronal convolutif aux données graphiques en conditionnant la convolution sur les caractéristiques des bords. Cette approche permet à l'encodeur de capturer efficacement la structure locale et globale des graphes, ce qui est nécessaire pour une représentation latente riche.

La couche finale de l'encodeur délivre les paramètres (moyenne et variance) d'une distribution gaussienne dans l'espace latent, permettant ainsi l'utilisation de l'astuce de réparamétrage. Les ECC se sont montrés performants dans la capture de relations graphiques complexes, car ils permettent l'agrégation flexible d'informations sur les nœuds voisins conditionnées par des attributs de bord, ce qui les rend appropriés pour des tâches impliquant des graphes hétérogènes [167].

Signification et applications du modèle

Le modèle GraphVAE est un cadre efficace pour la génération probabiliste de structures graphiques, en maintenant la fidélité aux caractéristiques des graphes d'entrée grâce à des intégrations apprises et une perte de reconstruction structurée. Des recherches récentes ont démontré l'utilité de GraphVAE dans divers domaines. Par exemple, Grover a utilisé GraphVAE pour générer de petits graphes moléculaires, montrant son potentiel en chimie informatique pour la génération de nouvelles molécules. En outre, il a été utilisé pour générer des représentations abstraites de réseaux sociaux, où la préservation des propriétés structurelles clés est essentielle [168].

Dans l'ensemble, GraphVAE est un outil puissant dans le domaine des modèles génératifs profonds, répondant spécifiquement aux défis associés aux données graphiques en fournissant un cadre de bout en bout qui comble l'écart entre les structures graphiques discrètes et les espaces latents continus.

5.2.2 Aperçu du modèle QMGBP-DL

L'encodeur moléculaire est crucial pour transformer les informations structurelles et chimiques des molécules en une représentation numérique latente. Il utilise des représentations basées sur des graphes, en utilisant spécifiquement **des graphes convolutionnels (GCNs)** [177].

Réseaux convolutifs de graphes

L'encodeur utilise plusieurs couches de GCN pour combiner la matrice de contiguïté moléculaire $mathbf{A}$ avec la matrice de caractéristique atomique $mathbf{F}$. Le fonctionnement

d'une seule couche GCN est donné par :

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \tag{5.3}$$

où $mathbf{H}(l)$ est la matrice de caractéristiques à la couche l (avec $\mathbf{H}^{(0)} = \mathbf{F}$), $\sigma(\cdot)$ est une fonction d'activation non linéaire, $\mathbf{W}^{(l)}$ est une matrice de poids entraînable et $\hat{\mathbf{A}}$ est la matrice d'adjacence normalisée calculée comme suit $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}$ avec $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_n$ et $\tilde{\mathbf{D}}ii = \sum_j \tilde{\mathbf{A}}ij$.

Dans cette implémentation spécifique, l'encodeur se compose de deux couches convolutionnelles, avec une activation ReLU pour la première et une activation linéaire pour la seconde. La fonction de l'encodeur est exprimée comme suit :

$$\mathbf{E} = f(\mathbf{F}, \mathbf{A}) = \hat{\mathbf{A}}\text{ReLU}(\hat{\mathbf{A}}\mathbf{F}\mathbf{W}^{(0)})\mathbf{W}^{(1)} \tag{5.4}$$

Ce processus intègre la topologie du graphe et les caractéristiques des nœuds dans la représentation $\mathbf{E}$. Suivant les couches GCN, une opération de mise en commun globale est appliquée à $\mathbf{E}$ pour obtenir un vecteur de taille fixe $\mathbf{E}' \in \mathbb{R}^d$ représentant la molécule entière.

Astuce de réparamétrage

Pour apprendre un espace latent bien structuré suivant une distribution normale standard, l'encodeur moléculaire met en correspondance les vecteurs moyen (μ) et variance (σ). L'astuce de reparamétrisation est utilisée pour activer l'optimisation basée sur le gradient lors de l'échantillonnage à partir de $\mathcal{N}(\mu, \sigma^2)$. Un échantillon z est obtenu par :

$$z = \mu + \sigma \cdot \varepsilon \tag{5.5}$$

où ε est échantillonné à partir de $\mathcal{N}(0, 1)$.

Prédicteur de propriété

Le prédicteur de propriété prend la représentation d'espace latente (la sortie de l'encodeur moléculaire, $\mathbf{E}'$) comme données d'entrée pour estimer les propriétés moléculaires, en particulier l'écart de bande dans cette étude. Comme le montre la figure 5.2, divers modèles d'apprentissage automatique sont utilisés pour cette tâche, notamment le perceptron multicouche (MLP), la régression linéaire (LR), la régression vectorielle de soutien (SVR), la forêt aléatoire (RF) et l'amplification du gradient (GB). Ces modèles sont formés pour prédire la propriété cible en apprenant des relations codées dans l'espace latent.

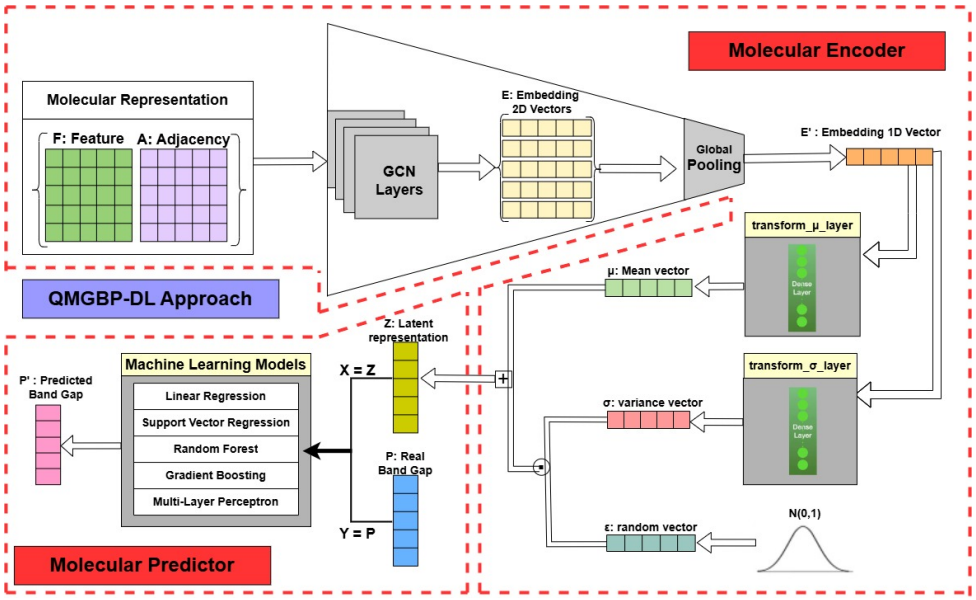


FIGURE 5.2 – L'approche QMGBP-DL

5.2.3 L'approche ME&PP-MG&RC

Notre cadre proposé ME&PP-MG&RC-DL dans la figure 5.3 vise à encapsuler les structures graphiques et les attributs des nœuds de molécules dans des inclusions de dimensions inférieures, permettant la prédiction des propriétés moléculaires et la génération de nouveaux composés. Ceci est réalisé grâce à une approche à deux composants : Molecular Encoder and Property Predictor (ME&PP) et Molecular Generator and Realness Classifier (MG&RC).

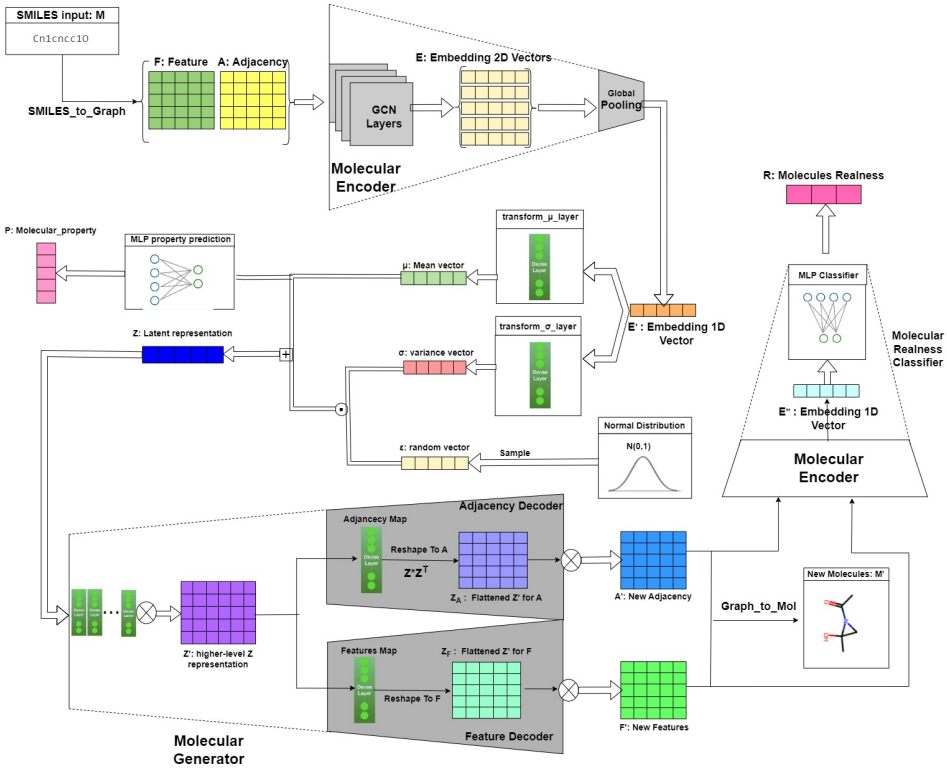


FIGURE 5.3 – L'approche ME&PP-MG&RC

Encodeur moléculaire et prédicteur de propriété (ME&PP)

Le ME&PP utilise un encodeur graphique pour intégrer le graphe moléculaire d'entrée, représenté par une matrice d'adjacence et une matrice de caractéristiques, dans un espace latent probabiliste. Cet espace latent est caractérisé par un vecteur moyen et un vecteur de variance, qui capturent les informations essentielles sur la structure moléculaire et sa variabilité inhérente [7]. Le vecteur moyen est ensuite introduit dans un perceptron multicouche (MLP) pour prédire des propriétés moléculaires spécifiques, telles que la solubilité ou la toxicité, en fonction de la représentation latente.

Générateur moléculaire et classificateur de la réalité (MG&RC)

Pour explorer l'espace latent, nous effectuons des échantillonnages en utilisant les vecteurs de moyenne et de variance, générant un vecteur d'espace latent qui encapsule la variabilité des structures moléculaires. Ce vecteur est ensuite utilisé comme entrée pour un générateur moléculaire, qui produit un nouveau graphe moléculaire, caractérisé par des matrices d'adjacence et de caractéristiques mises à jour. Pour s'assurer que les structures générées ressemblent à des molécules réelles, nous utilisons un classificateur de réalité moléculaire, qui évalue l'authenticité des structures générées.

Intégration et implications

En intégrant les composants ME&PP-MG&RC, notre méthode permet la prédiction des propriétés moléculaires et la génération de nouvelles structures moléculaires réalistes. Cela a des implications importantes pour la conception et la découverte moléculaires, car il permet l'exploration de vastes espaces chimiques et l'identification de composés lead potentiels [8]. Notre approche exploite les forces de l'apprentissage par représentation graphique et des modèles génératifs profonds, offrant un outil puissant aux scientifiques et ingénieurs moléculaires.

5.2.4 L'entraînement des performances de l'approache ME&PP-MG&RC-DL

L'approche proposée a été formée en deux étapes, permettant une optimisation spécialisée de chaque composant. Les composants Molecular Encoder and Property Predictor (ME&PP) et Molecule Generator and Realness Classifier (MG&RC) ont été entraînés indépendamment en utilisant des fonctions de perte séparées, comme indiqué dans la figure suivante 5.4.

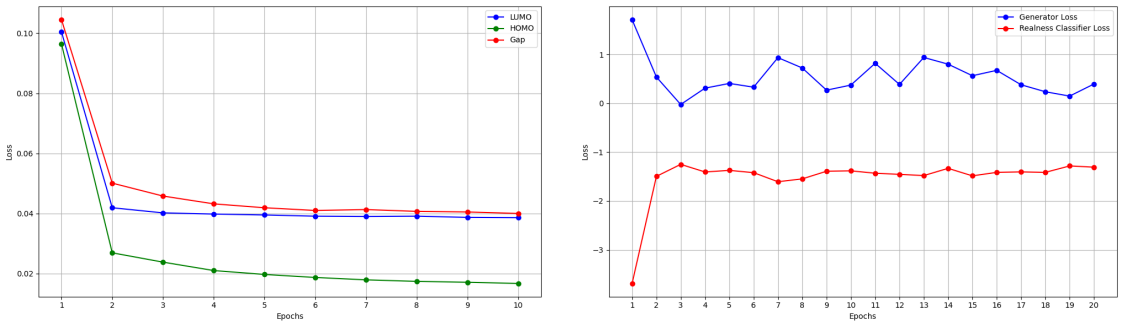


FIGURE 5.4 – L'entraînement des composants MEPP et MGRC sur les propriétés Gap, HOMO et LUMO

L'approche proposée a été entraînée en deux étapes, permettant une optimisation spécialisée de chaque composant. Le modèle ME&PP a été entraîné en utilisant une fonction de perte conjointe, combinant la divergence KL et les pertes MAE, ce qui a facilité l'encodage robuste de l'espace latent et la prédiction précise des propriétés moléculaires. Le modèle a démontré des

tendances de convergence prometteuses sur l'ensemble de données QM9, avec une réduction constante des pertes sur 10 époques pour des propriétés telles que LUMO, HOMO et Gap.

En revanche, le modèle MG&RC présentait des modèles de formation distincts. La perte de générateur a diminué de manière significative dans les époques initiales, indiquant l'amélioration rapide de la qualité des molécules générées. Cependant, la perte a fluctué et tendu vers le haut dans les époques ultérieures, suggérant l'adaptation et le raffinement continus de génération de molécule. Ce comportement est cohérent avec des recherches antérieures sur les modèles génératifs, où le générateur et le discriminateur s'engagent dans un jeu compétitif, entraînant une amélioration dans les deux composants.

5.2.5 Étude comparative des molécules générées à partir de points latents échantillonnés

Dans cette étude, nous avons utilisé un générateur moléculaire formé sur l'ensemble de données QM9 pour explorer la diversité et la qualité des molécules générées à partir de points latents échantillonnés Z_i et Z_j. Les molécules générées ont été évaluées en fonction de leur validité structurelle, unicité et nouveauté 5.1. le générateur moléculaire utilisé dans cette recherche était capable de produire des structures moléculaires diverses à partir de différentes représentations spatiales latentes, comme démontré dans les figures 5.5 et 5.6 comparées aux molécules réelles de l'ensemble de données QM9 comme indiqué dans la figure 5.7, par les deux ensembles de molécules générées à partir Z_i et Z_jChaque point latent, Z_i et Z_j correspond à un point différent dans l'espace latent appris, permettant la génération des molécules distinctes.

TABLE 5.1 – Performances du générateur moléculaire utilisant des espaces latents Z_i et Z_j.

Metric	Latent Space Z_i	Latent Space Z_j
Pourcentage de validité	68.75%	79.16%
Pourcentage d'unicité	87.87%	84.21%
Pourcentage de nouveauté	79.31%	62.5%

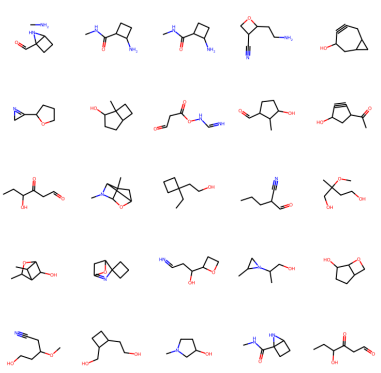


FIGURE 5.5 – Molécules générées à l'aide d'échantillon Z_i

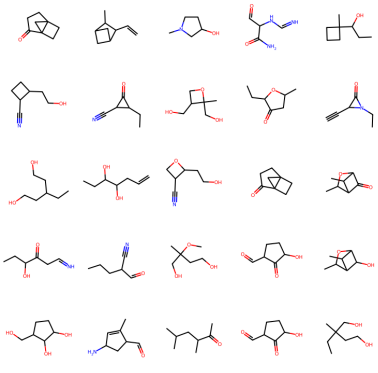


FIGURE 5.6 – Molécules générées à l'aide de l'échantillon Z_j

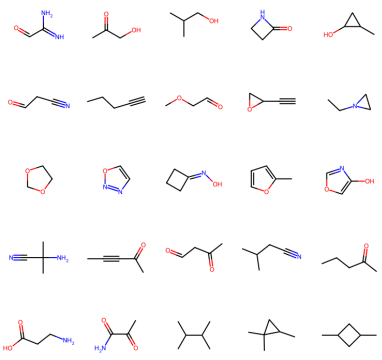


FIGURE 5.7 – Les molécules réelles du QM9 dataset

Ces résultats mettent en évidence l'efficacité du générateur moléculaire dans l'exploration et la génération de molécules structurellement diversifiées et valides à partir de différentes représentations spatiales latentes. La variation des mesures de rendement entre Z_i et Z_j souligne

l'influence nuancée de l'échantillonnage spatial latent sur les résultats de la génération moléculaire, révélant des possibilités d'optimisation et d'exploration supplémentaires dans la conception moléculaire computationnelle.

5.2.6 Étude comparative de la classification de la réalité pour les points latents échantillonnés

L'évaluation de la réalité des molécules générées en utilisant des points échantillonnés Z_i et Z_j fournit des informations sur l'efficacité du modèle de génération moléculaire. Nous discutons ici des implications et des conclusions basées sur les résultats présentés dans les figures suivantes 5.8 et 5.9.

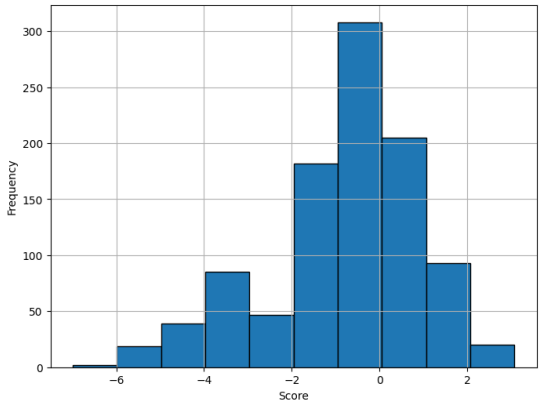


FIGURE 5.8 – Classification de la réalité à partir de Z_i

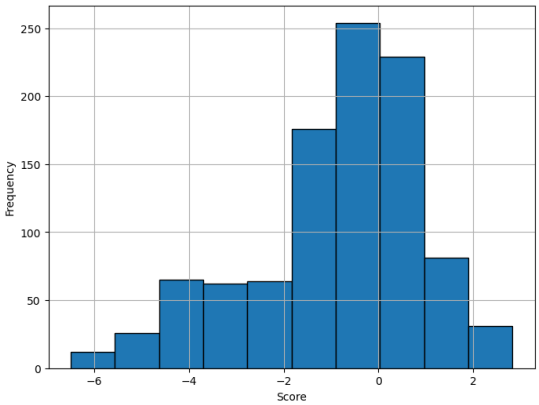


FIGURE 5.9 – Classification de la réalité à partir de Z_j

Sur les 1000 molécules générées avec Z_i, plus de 200 ont un score de réalité supérieur à 0, près de 100 dépassent 1 et environ 25 dépassent 2. Pour les molécules générées en utilisant Z_j, plus de 225 points au-dessus de 0, plus de 75 dépassent 1 et environ 25 dépassent 2. Cette distribution indique que le modèle génératif peut créer un nombre substantiel de molécules ressemblant à celles réelles de l'ensemble de données d'entraînement, comme le montrent les scores positifs de réalité. Cependant, le nombre plus faible de molécules ayant des scores plus élevés (au-dessus de 2) suggère qu'il reste difficile d'atteindre une haute fidélité aux molécules réelles. La variabilité des scores de réalité souligne la diversité et la qualité des molécules générées, les différences entre Z_i et Z_j indiquant des variations dans le processus de génération dues à différentes représentations latentes.

5.3 Analyse comparative de la performance ME&PP-MG&RC avec d'autres méthodes

Les résultats résumés dans le tableau 5.2 soulignent la performance exceptionnelle de notre méthode proposée ME&PP-MG&RC-DL dans la prédiction des propriétés chimiques quantiques clés de l'ensemble de données QM9 pour les propriétés Gap, HOMO et LUMO surpassant les méthodes GraphVAE et QMGBP-DL existantes.

TABLE 5.2 – Les performances des méthodes existants et notre approche sur les propriétés de l'ensemble de données QM9

Model Approche	Gap	HOMO	LUMO
graphe autoencodeurs variationnels (GraphVAE)	0.21	0.16	0.16
Quantum molecular band gap prediction (QMGBP-DL)	0.0171	0.0098	0.0150
ME&PP-MG&RC-DL (Ours)	**0.0118**	**0.0081**	**0.0108**

Les valeurs présentées dans le tableau représentent des mesures moyenne d'erreur absolue, où des valeurs plus faibles indiquent une précision de prédiction plus élevée. Notre méthode surpasse de manière significative les approches existantes telles que le graphe autoencodeurs variationnels (GraphVAE) et la Quantum Molecular graphe Band Gap Prediction (QMGBP-DL).

La méthode GraphVAE, qui représente une approche antérieure, présente des valeurs d'erreur relativement élevées pour les trois propriétés : 0,21 pour Gap, 0,16 pour HOMO et 0,16 pour LUMO. Ceci indique que, tandis que GraphVAE peut traiter des structures moléculaires comme graphes, sa capacité de capturer avec précision les caractéristiques moléculaires nuancées requises pour la prévision précise de propriété de tranche de temps est comparativement limitée.

La méthode QMGBP-DL démontre une amélioration substantielle par rapport à GraphVAE, obtenant des mesures d'erreur significativement plus faibles : 0,0171 pour Gap, 0,0098 pour HOMO et 0,0150 pour LUMO. Cette performance reflète les progrès réalisés dans l'application de l'apprentissage profond basé sur des graphes aux tâches de chimie quantique, ce qui indique une modélisation plus efficace des propriétés électroniques.

Notre méthode proposée, ME&PP-MG&RC-DL, surpasse GraphVAE et QMGBP-DL dans toutes les propriétés évaluées. Comme indiqué dans le tableau 5.2, ME&PP-MG&RC-DL atteint les taux d'erreur les plus bas : **0.0118** pour Gap, **0.0081** pour HOMO et **0.0108** pour LUMO. Plus précisément, pour la propriété Gap, notre méthode réduit l'erreur de 0.0171 (QMGBP-DL) à **0.0118**, ce qui représente une amélioration notable. De même, pour la propriété HOMO, l'erreur est réduite de 0.0098 (QMGBP-DL) à **0.0081**. Le gain de performance est également évident pour la propriété LUMO, avec l'erreur diminuant de 0.0150 (QMGBP-DL) à **0.0108**. Ces valeurs mises en évidence soulignent que l'intégration des principes ME&PP et MG&RC au sein de notre approche d'apprentissage profond permet une modélisation plus précise des interdépendances moléculaires complexes, ce qui conduit à des prédictions des propriétés quantiques de solides parmi les méthodes comparées.

En résumé, cette analyse comparative quantitative valide la performance supérieure de notre approche ME&PP-MG&RC-DL pour la prédiction des propriétés Gap, HOMO et LUMO sur l'ensemble de données QM9. Les améliorations significatives des performances par rapport aux méthodes GraphVAE et QMGBP-DL démontrent l'efficacité de notre approche et de notre méthodologie proposées dans la capture des caractéristiques moléculaires pertinentes pour ces tâches.

Conclusion

Dans ce chapitre, nous avons présenté une analyse complète du cadre ME&PP-MG&RC, qui intègre la prédiction et la génération de propriétés moléculaires par le biais d'une approche

à deux composants. En tirant parti des capacités de l'apprentissage profond, en particulier dans le contexte du jeu de données QM9, notre approche capture efficacement les relations complexes entre les structures moléculaires et leurs propriétés correspondantes. Les résultats ont démontré que notre modèle excelle non seulement dans la prédiction des propriétés mécaniques quantiques, mais aussi dans la génération de structures moléculaires chimiquement valides.

Un aspect important de notre recherche a été l'évaluation comparative par rapport aux modèles GraphVAE et QMGBP-DL, une méthode importante dans la modélisation générative basée sur des graphes. Les résultats ont mis en évidence que, tandis que GraphVAE et QMGBP-DL génèrent efficacement des graphiques moléculaires diversifiés, le cadre ME&PP-MG&RC offre une précision accrue dans les prédictions de propriétés grâce à ses composants spécialisés d'encodeur et de prédicteur. Cette distinction est essentielle, car une prédiction précise des propriétés est essentielle pour diverses applications dans la découverte de médicaments et la science des matériaux.

De plus, l'utilisation du jeu de données QM9 a souligné son importance en tant que point de référence pour évaluer les modèles d'apprentissage automatique en chimie informatique. La riche compilation des propriétés mécaniques quantiques de l'ensemble de données a posé un banc d'essai difficile mais précieux pour notre cadre, facilitant une compréhension nuancée du comportement moléculaire et de la réactivité.

En résumé, notre recherche contribue au corpus croissant de connaissances dans l'apprentissage machine moléculaire en présentant un cadre robuste qui combine la prédiction des propriétés avec la modélisation générative. Les travaux futurs seront axés sur le raffinement de cette approche et l'exploration de son applicabilité dans différents domaines chimiques comme la découverte des médicaments.

Chapitre 6

Impact de l'intégration des méthodes PPM et GM

Introduction

Le processus de découverte de médicaments a connu d'importantes transformations au cours des dernières décennies, passant des approches expérimentales traditionnelles aux méthodes informatiques sophistiquées qui tirent parti de l'intelligence artificielle et de l'apprentissage profond. Cette évolution est motivée par la nécessité de relever les défis inhérents à la mise au point des médicaments, notamment les coûts élevés, les délais longs et les taux d'attrition élevés. Comme démontré tout au long de cette thèse, l'intégration de l'intelligence artificielle, en particulier des modèles d'apprentissage profond, est apparue comme une approche prometteuse pour révolutionner la découverte de médicaments en réduisant considérablement le temps, les ressources et les investissements financiers requis [385, 386].

Ce dernier chapitre synthétise les principales contributions, limites et perspectives futures de notre recherche axée sur l'intégration de la prédiction des propriétés moléculaires (MPP) et de la génération moléculaire (MG) dans le processus de découverte de médicaments à l'aide de modèles d'apprentissage profond. En combinant ces deux composantes essentielles, nous avons mis au point de nouvelles approches à savoir l'architecture GMPP-NN pour la prédiction des propriétés moléculaires et le cadre ME&PP-MG&RC-DL pour la génération moléculaire et la prédiction des propriétés– qui démontrent des résultats prometteurs dans l'accélération et l'amélioration de divers aspects du pipeline de développement de médicaments [387, 388].

L'intégration de MPP et de MG représente un changement de paradigme dans la découverte de médicaments par ordinateur, permettant aux chercheurs non seulement de prédire les propriétés des structures moléculaires existantes, mais aussi de générer de nouvelles molécules adaptées aux caractéristiques souhaitées. Cette approche permet de naviguer efficacement dans le vaste espace chimique, facilitant l'identification des médicaments candidats prometteurs tout en contournant bon nombre des limitations associées aux méthodes de criblage traditionnelles. Cependant, comme pour toute technologie émergente, il y a des limites notables qui doivent être abordées pour réaliser pleinement le potentiel de ces approches dans les applications du monde réel [389, 390].

Dans une perspective d'avenir, ce chapitre explore également les orientations futures et les perspectives pour l'intégration de la MPP et de la MG dans la découverte de médicaments, en mettant en évidence les tendances émergentes, les applications potentielles, et des pistes de recherche qui pourraient améliorer encore l'efficacité et l'applicabilité de ces méthodes de calcul. En examinant de façon critique les réalisations et les défis dans ce domaine, nous visons

à fournir un aperçu complet qui guidera les efforts de recherche futurs et les applications dans la découverte de médicaments informatiques.

6.1 Contributions de l'intégration du PPM et GM dans le processus à l'aide de l'IA

L'intégration de la prédiction des propriétés moléculaires et de la génération moléculaire à l'aide de modèles d'apprentissage profond a apporté des contributions importantes au processus de découverte de médicaments, offrant des solutions novatrices aux défis de longue durée. Notre recherche a fait progresser ce domaine grâce à plusieurs contributions clés :

6.1.1 Développement de nouvelles architectures d'apprentissage profond

Nos travaux de recherche ont permis de mettre au point deux importantes architectures d'apprentissage profond adaptées aux applications de découverte de médicaments :

— **GMPP-NN Architecture :** Notre réseau neuronal de prédiction des propriétés moléculaires du graphe combine les réseaux neuronaux de passage de messages (MPNNs) avec un classificateur de perceptron multicouche pour prédire avec précision les propriétés moléculaires. Cette architecture exploite les représentations graphiques des molécules, capturant efficacement leurs informations structurelles et chimiques. Le modèle GMPP-NN a démontré une performance exceptionnelle sur plusieurs ensembles de données (VIH, BACE, BBBP et ClinTox), surpassant les approches conventionnelles en termes de métriques ROC-AUC et PRC-AUC.

— **ME&PP-MG&RC-DL Framework :** Le cadre d'apprentissage profond de l'encodage moléculaire et de la prédiction des propriétés - génération moléculaire et classification de la réalité représente une approche globale de la découverte de médicaments. Cette architecture intègre l'encodage moléculaire, la prédiction des propriétés, la génération et la classification de la réalité dans un système cohérent. En tirant parti des réseaux convolutionnaires graphiques (GNG) et des techniques variationnelles, le cadre ME&PP-MG&RC-DL a démontré une précision remarquable dans la prédiction des propriétés chimiques quantiques (HOMO, LUMO et écart d'énergie) tout en générant des structures moléculaires diverses, valides et nouvelles.

6.1.2 Amélioration de la précision dans la prédiction des propriétés moléculaires

Nos architectures ont considérablement amélioré la précision de la prédiction des propriétés moléculaires par rapport aux méthodes existantes :

— **Performance dans les tâches de classification :** The GMPP-NN architecture achieved exceptional Les scores ROC-AUC obtenus sur plusieurs ensembles de données (0,8677 pour le VIH, 0,8608 pour le BACE, 0,9186 pour le BBBP et 0,9795 pour le ClinTox) surpassent des modèles concurrents tels que le transformateur SMILES, FP2VEC et les réseaux de neurones profonds traditionnels.

— **Précision sans précédent dans la prédiction des propriétés quantiques :** Le cadre ME&PP-MG&RC-DL a obtenu des scores d'erreur absolue moyenne (MAE) remarquables pour les propriétés chimiques quantiques, avec 0,04 pour la propriété Gap, 0,02

pour la propriété HOMO et 0,04 pour la propriété LUMO. Ces résultats surpassent de façon significative les méthodes existantes, y compris les empreintes digitales à connectivité étendue (ECFP), la matrice coulombienne (CM), les SMILES à chaud unique et les encodeurs automatiques variationnels graphiques (GVAEs).

6.1.3 Progrès en matière de génération moléculaire

Nos travaux de recherche ont apporté des contributions substantielles à la génération moléculaire, permettant la création de nouveaux composés aux propriétés souhaitées :

— **Génération de molécules nouvelles et valables :** La framework ME&PP-MG&RC-DL a démontré la capacité de générer des molécules chimiquement valides avec des pourcentages élevés d'unicité et de nouveauté. Pour les molécules générées en utilisant différents points d'espace latent, la validité variait de 68,75% à 79,16%, l'unicité de 84,21% à 87,87% et la nouveauté de 62,5% à 79,31%. Cela indique la capacité du cadre à explorer des espaces chimiques au-delà de l'ensemble de données de formation.

— **Intégration de la classification des réalités :** L'incorporation d'un classificateur de réalité dans notre cadre garantit que les molécules générées ressemblent étroitement à des structures réelles, chimiquement réalisables. Cette composante distingue notre approche de nombreuses méthodes existantes et améliore l'utilité pratique des composés générés.

6.1.4 Représentation spatiale latente efficace

Nos architectures ont démontré l'efficacité de la représentation spatiale latente des molécules, facilitant à la fois la prédiction et la génération des propriétés :

— **Espace latent structurellement significatif :** Les espaces latents générés par nos modèles regroupent effectivement des molécules en fonction de leurs propriétés, comme le montre la séparation claire de molécules avec différentes valeurs HOMO et lipophilicité dans les visualisations. Cette organisation structurelle permet une exploration ciblée de l'espace chimique.

— **Faciliter les relations structure-propriété :** La nature continue et structurée de nos espaces latents facilite la compréhension des relations structure-propriété, fournissant des renseignements précieux pour la conception rationnelle des médicaments.

6.1.5 Applicabilité pratique à la découverte de médicaments

L'intégration de la MPP et de la MG dans nos approches offre des avantages pratiques pour la découverte de médicaments :

— **Accélération de l'identification du lead :** En prédisant avec précision les propriétés et en générant des molécules aux caractéristiques souhaitées, nos approches peuvent accélérer de manière significative l'identification des composés prometteurs du lead, réduisant le temps et les ressources nécessaires pour la découverte précoce de médicaments.

— **Exploration de nouveaux espaces chimiques :** La capacité de générer et d'évaluer de nouvelles structures moléculaires permet l'exploration d'espaces chimiques auparavant inexplorés, conduisant potentiellement à la découverte de composés aux propriétés thérapeutiques uniques.

— **Recours réduit au criblage expérimental :** La grande précision de nos modèles de prédiction des propriétés réduit le besoin d'un vaste criblage expérimental, ce qui diminue les coûts et les délais associés.

En résumé, notre recherche a apporté des contributions importantes au domaine de la découverte de médicaments informatiques grâce au développement d'architectures innovantes d'apprentissage profond qui intègrent la prédiction et la génération de propriétés moléculaires. Ces contributions ont le potentiel de transformer le processus de découverte des médicaments, ce qui le rend plus efficace, rentable et efficace dans l'identification de candidats thérapeutiques prometteurs.

6.2 Limites de l'intégration du PPM et du GM dans le processus

Malgré les progrès importants réalisés dans l'intégration de la prédiction des propriétés moléculaires (MPP) et de la génération moléculaire (MG) à la découverte de médicaments au moyen de modèles d'apprentissage profond, plusieurs limitations importantes demeurent qui doivent être reconnues et prises en compte :

6.2.1 Contraintes de qualité et de disponibilité des données

L'une des limites les plus importantes des approches d'apprentissage profond dans la découverte de médicaments réside dans la qualité, la quantité et la diversité des données de formation disponibles [391, 392] :

— **Ensembles de données limités et déséquilibrés :** De nombreux ensembles de données moléculaires, en particulier ceux portant sur des cibles thérapeutiques spécifiques ou des maladies rares, restent petits ou déséquilibrés. Par exemple, notre travail sur GMPP-NN a utilisé les ensembles de données MoleculeNet, qui, bien que complets, ont toujours des limitations en taille par rapport aux ensembles de données dans d'autres domaines où l'apprentissage profond a montré une performance exceptionnelle.

— **Incohérence des données expérimentales :** Les mesures expérimentales des propriétés moléculaires peuvent varier considérablement en fonction des conditions, des méthodologies et de l'équipement expérimentaux, ce qui entraîne des incohérences dans les données de formation. Cette incohérence peut se propager à travers les modèles, affectant leur précision prédictive.

— **Manque de données sur la faisabilité synthétique :** Il y a une pénurie notable de données complètes sur la faisabilité de la synthèse, ce qui limite la capacité des modèles génératifs à produire des molécules non seulement valides mais aussi pratiquement synthétisables en laboratoire.

6.2.2 Défis de la représentation moléculaire

La représentation des structures moléculaires pour les modèles d'apprentissage profond présente plusieurs défis [393] :

— **Limites de taille et de complexité des graphes :** Les architectures GMPP-NN et ME&PP-MG&RC-DL font face à des défis dans le traitement efficace de structures moléculaires

très grandes et complexes en raison des contraintes de calcul dans le traitement de données de graphe étendues.

— **Perte d'information structurelle 3D :** Bien que les représentations graphiques capturent efficacement la topologie des molécules, elles perdent souvent des informations structurelles tridimensionnelles critiques qui sont cruciales pour comprendre les interactions protéine-ligand, un aspect clé de l'efficacité du médicament.

— **Intégration des connaissances en chimie :** Les méthodes de représentation actuelles n'intègrent pas entièrement les connaissances chimiques spécifiques au domaine, telles que les mécanismes de réaction, les structures de résonance et le tautomérisme, qui sont essentiels pour une modélisation moléculaire précise.

6.2.3 Limitations du modèle et défis informatiques

Les modèles d'apprentissage profond eux-mêmes présentent des limites inhérentes [394, 395] :

— **Besoins en ressources de calcul :** Les deux architectures, en particulier le cadre ME&PP-MG&RC-DL avec ses multiples composants, nécessitent des ressources de calcul importantes pour la formation et l'inférence, limitant leur accessibilité et leur évolutivité.

— **Nature et interprétabilité de la boîte noire :** Malgré leur pouvoir prédictif, de nombreux modèles d'apprentissage profond fonctionnent comme des "boîtes noires", ce qui rend difficile l'interprétation de leurs prédictions ou la fourniture d'informations mécanistes sur les relations structure-propriété, ce qui est crucial pour la conception rationnelle des médicaments.

— **Généralisation aux nouveaux espaces chimiques :** Comme démontré dans nos expériences avec le cadre ME&PP-MG&RC-DL, alors que les modèles peuvent générer de nouvelles molécules, leur performance peut se dégrader lorsque l'on tente de généraliser à des espaces chimiques significativement différents des données d'entraînement.

— **Stabilité de l'entraînement et sensibilité aux hyperparamètres :** Les GMPP-NN et les ME&PP-MG&RC-DL sont sensibles aux paramètres hyperparamétriques, et la stabilité de l'entraînement demeure un défi, en particulier pour les composants génératifs, qui peuvent être sujets à des effondrements de mode ou générer des structures irréalistes.

6.2.4 Obstacles pratiques à la mise en œuvre

La traduction de ces approches informatiques en une découverte pratique de médicaments se heurte à plusieurs obstacles [396] :

— **Validation Gap :** Il subsiste un écart important entre les prévisions informatiques et la validation expérimentale. Nos modèles, bien qu'ils soient exacts sur les ensembles de données de référence, nécessitent toujours une validation expérimentale approfondie avant d'être pleinement fiables dans des scénarios de découverte de médicaments dans le monde réel.

— **Intégration avec les pipelines existants de découverte de médicaments :** L'intégration des approches d'apprentissage profond aux flux de travail établis en matière de découverte de médicaments présente des défis organisationnels et techniques, nécessitant des adaptations importantes et l'acceptation de divers intervenants.

— **Aspects réglementaires :** L'application des molécules générées par l'IA dans le développement de médicaments soulève des questions réglementaires concernant la validation et la documentation de ces approches, qui ne sont pas encore entièrement traitées par les cadres réglementaires.

6.2.5 Limites des tâches spécifiques

Chaque tâche spécifique au sein du cadre intégré MPP et MG a ses propres limites [397] :

— **Précision des prédictions de propriétés pour les propriétés complexes :** Bien que nos modèles montrent de bonnes performances pour les propriétés évaluées, la prévision des propriétés pharmacocinétiques et pharmacodynamiques complexes, telles que la biodisponibilité, les profils de toxicité et les interactions médicamenteuses, demeure difficile.

— **Validité chimique et faisabilité de synthèse dans la production :** Comme le montre notre évaluation du cadre ME&PP-MG&RC-DL, les molécules générées ne répondent pas toutes aux critères de validité chimique (validité de 68,75 à 79,16%), et encore moins peuvent être synthétiquement réalisables, limitant leur utilité pratique.

— **Contraintes de génération ciblées :** La production de molécules ayant des profils de propriétés spécifiques et multiobjectifs demeure un défi, comme en témoignent les compromis observés dans nos évaluations de nouveauté par rapport à la réalité.

Ces limites soulignent le besoin de poursuivre la recherche et le développement dans l'intégration du MPP et du MG pour la découverte de médicaments. Il sera essentiel de relever ces défis pour exploiter pleinement le potentiel des approches d'apprentissage profond en vue de révolutionner le processus de découverte de médicaments.

6.3 Avenir de l'intégration du PPM et du GM dans le processus de découverte des médicaments

L'avenir de l'intégration de la prédiction des propriétés moléculaires (MPP) et de la génération moléculaire (MG) dans la découverte de médicaments à l'aide de modèles d'apprentissage profond est très prometteur pour transformer la recherche et le développement pharmaceutiques. Plusieurs avancées et tendances clés devraient façonner ce domaine au cours des prochaines années :

6.3.1 Avancées dans les architectures de modèles et les algorithmes

Les développements futurs dans les architectures et algorithmes de modèles devraient permettre de remédier aux limitations actuelles et d'améliorer la performance [398] :

— **Modèles moléculaires à base de transformateur :** En s'appuyant sur le succès des architectures de transformateurs dans le traitement du langage naturel, nous anticipons le développement de modèles plus sophistiqués basés sur les transformateurs et conçus spécifiquement pour la représentation et la génération moléculaires. Ces modèles pourraient mieux capturer les dépendances à long terme dans les structures moléculaires et améliorer la prédiction des propriétés ainsi que les tâches de génération.

— **Cadres d'apprentissage multimodaux et multitâches :** Les futurs modèles intégreront probablement plusieurs modalités de données (structurales, génomiques, protéomiques)

et exécuteront simultanément plusieurs tâches. Les extensions de notre cadre ME&PP-MG&RC-DL pourraient être conçues pour prédire un plus large éventail de propriétés tout en générant des molécules avec des caractéristiques ciblées dans une architecture unifiée.

— **Self-Supervised et Few-Shot Learning :** Afin de remédier aux limites des données, les approches futures utiliseront probablement des techniques d'apprentissage autosupervisé pour extraire des représentations significatives à partir de données moléculaires non étiquetées, suivies d'un apprentissage de courte durée pour s'adapter à des tâches spécifiques avec des données étiquetées limitées.

— **Apprentissage par renforcement pour l'optimisation moléculaire :** Des stratégies avancées d'apprentissage par renforcement permettront une exploration plus dirigée et efficace des espaces chimiques, optimisant les molécules pour de multiples propriétés simultanément tout en maintenant la faisabilité synthétique.

6.3.2 Représentation moléculaire améliorée

Des améliorations dans les représentations moléculaires seront cruciales pour faire progresser le domaine [399] :

— **Intégration de l'information structurelle 3D :** Les futurs modèles intégreront mieux l'information structurelle tridimensionnelle, y compris la flexibilité conformationnelle et la chiralité, qui sont cruciales pour la prévision précise de propriété et la génération réaliste de molécule.

— **Représentations moléculaires dynamiques :** Au-delà des représentations statiques, les futurs modèles intégreront probablement les aspects dynamiques des molécules, tels que les changements de conformation et la dynamique des réactions, afin de mieux saisir leur comportement dans les systèmes biologiques.

— **Représentations améliorées par les connaissances :** L'intégration des bases de connaissances chimiques et des règles de réaction dans les représentations moléculaires améliorera la pertinence chimique et la faisabilité synthétique des molécules générées.

6.3.3 Intégration avec d'autres technologies

L'intégration de la MPP et de la MG avec d'autres technologies de pointe créera des approches synergiques [400] :

— **Calcul quantique pour la modélisation moléculaire :** À mesure que l'informatique quantique se développe, elle a le potentiel de révolutionner la modélisation moléculaire en permettant des calculs mécaniques quantiques plus précis à l'échelle, qui pourraient être intégrés avec des approches d'apprentissage profond pour améliorer la prédiction des propriétés.

— **Synthèse automatisée et expérimentation à haut débit :** La combinaison de la conception moléculaire pilotée par l'IA avec des plates-formes de synthèse automatisées et l'expérimentation à haut débit créera des systèmes en boucle fermée pour la conception, la synthèse, les essais et l'optimisation rapides des médicaments candidats.

— **Intégration avec la biologie des systèmes et la modélisation multi-échelle :** Les approches futures intégreront de plus en plus les prédictions au niveau moléculaire aux modèles de biologie des systèmes pour mieux prédire l'efficacité et la toxicité dans des

contextes biologiques complexes, comblant ainsi le fossé entre les propriétés moléculaires et les résultats cliniques.

6.3.4 Développements propres à l'application

Les développements futurs seront probablement axés sur la résolution des défis spécifiques en matière de découverte de médicaments [401] :

— **Thérapies ciblées pour les maladies complexes :** Les modèles avancés MPP et MG permettront de concevoir des molécules spécifiquement adaptées à des maladies complexes aux mécanismes hétérogènes, telles que les troubles neurodégénératifs et le cancer, en prenant en compte simultanément plusieurs cibles et voies biologiques.

— **Approches de médecine personnalisée :** L'intégration des données génomiques et protéomiques avec la MPP et la MG facilitera la conception de thérapies personnalisées qui tiennent compte des variations génétiques individuelles et des profils de maladie.

— **Nouvelles modalités au-delà des petites molécules :** Les prolongations des approches actuelles porteront sur la conception de nouvelles modalités thérapeutiques, telles que les peptides, les anticorps et les thérapies à base d'acide nucléique, élargissant ainsi le champ d'application de la découverte de médicaments par l'IA.

6.3.5 Démocratisation et accessibilité

Rendre ces technologies plus accessibles accélérera l'innovation [402] :

— **Plateformes infonuagiques et outils open source :** Le développement de plateformes en nuage conviviales et de bibliothèques complètes à source ouverte démocratisera l'accès aux capacités avancées des MPP et des MG, permettant une participation plus large aux efforts de découverte de médicaments.

— **Normes et protocoles normalisés :** L'établissement de points de référence, de mesures d'évaluation et de protocoles normalisés facilitera la comparaison équitable des différentes approches et accélérera les progrès dans le domaine.

— **Ressources éducatives et formation interdisciplinaire :** Des ressources éducatives améliorées et des programmes de formation qui font le pont entre la chimie, la biologie et les sciences informatiques prépareront la prochaine génération de chercheurs à tirer parti efficacement de ces approches intégrées.

L'avenir de l'intégration des MPP et des MG dans la découverte de médicaments est propice à une croissance et à une innovation remarquables. En abordant les limites actuelles et en tirant parti des technologies émergentes, ces approches intégrées ont le potentiel de transformer radicalement le paysage de la découverte de médicaments, ce qui conduira à un développement plus efficace de nouveaux traitements pour une vaste gamme de maladies.

6.4 Perspectives de l'intégration de PPM et GM dans le processus

L'intégration de la prédiction des propriétés moléculaires et de la génération moléculaire à l'aide de modèles d'apprentissage profond ouvre de nombreuses perspectives qui pourraient remodeler le paysage de la découverte de médicaments. Cette section examine ces perspectives du point de vue scientifique, technologique, économique et éthique :

6.4.1 Perspectives scientifiques

Les implications scientifiques des approches intégrées MPP et MG vont au-delà des applications immédiates dans la découverte de médicaments :

— **Compréhension fondamentale de l'espace chimique :** Notre recherche contribue à une meilleure compréhension des vastes relations entre l'espace chimique et la structure-propriété. À mesure que des modèles comme ME&PP-MG&RC-DL continueront d'évoluer, ils fourniront des renseignements sans précédent sur les principes mathématiques et chimiques régissant les propriétés et le comportement moléculaires.

— **Nouveaux paradigmes en chimie computationnelle :** L'intégration de l'apprentissage profond avec la chimie informatique traditionnelle crée de nouveaux paradigmes qui combinent les forces des modèles basés sur la physique avec des approches axées sur les données. Cette approche hybride pourrait révolutionner notre capacité à prédire des phénomènes chimiques complexes à plusieurs échelles.

— **Découverte de nouvelles entités chimiques :** La capacité de naviguer dans des régions de l'espace chimique auparavant inexplorées, comme le démontrent les mesures de nouveauté dans notre ME&PP-L'évaluation de MG&RC-DL suggère que ces approches pourraient mener à la découverte de classes entièrement nouvelles d'entités chimiques avec des propriétés et des applications uniques au-delà des produits pharmaceutiques.

— **Progression la planification de la synthèse chimique :** L'intégration de la génération moléculaire avec l'évaluation de faisabilité synthétique pourrait transformer l'analyse rétrosynthétique et la planification de la synthèse, conduisant potentiellement à des voies synthétiques plus efficaces et innovantes pour les molécules complexes.

6.4.2 Perspectives technologiques

Du point de vue technologique, plusieurs perspectives importantes émergent :

— **Évolution vers la découverte autonome de médicaments :** L'intégration de la MPP et de la MG représente une étape vers des systèmes plus autonomes de découverte de médicaments capables de concevoir, d'optimiser et d'évaluer les candidats-médicaments avec un minimum d'intervention humaine. Les futures itérations de nos frameworks pourraient être des composants de plates-formes de découverte entièrement autonomes.

— **Optimisation itérative en temps réel :** Des méthodes de calcul plus rapides et des algorithmes améliorés permettront une optimisation itérative en temps réel des structures moléculaires basées sur des contraintes de propriétés multiples, accélérant considérablement le processus d'optimisation des pistes.

— **Intégration et interopérabilité des plateformes :** Le développement d'interfaces et de protocoles normalisés facilitera l'intégration de différentes plateformes informatiques, créant des écosystèmes complets de découverte de médicaments qui combinent harmonieusement la conception moléculaire, la prédiction des propriétés et la validation expérimentale.

— **IA explicable pour la découverte de médicaments :** L'évolution des techniques d'IA explicables améliorera l'interprétabilité de modèles complexes comme le GMPP-NN et le ME&PP-MG&RC-DL, en fournissant des renseignements sur le raisonnement derrière les prédictions et en générant des connaissances précieuses pour les chimistes médicinaux.

6.4.3 Perspectives économiques et industrielles

Les implications économiques de ces approches intégrées sont importantes :

— **Transformation des modèles de R&D pharmaceutique :** En réduisant potentiellement le temps et le coût de la découverte précoce de médicaments, les approches intégrées MPP et MG pourraient transformer les modèles pharmaceutiques de R&D, permettant une allocation plus efficace des ressources et l'exploration d'un éventail plus large de cibles thérapeutiques.

— **Favoriser les petites organisations :** La démocratisation de puissants outils informatiques pourrait égaliser les règles du jeu entre les grandes sociétés pharmaceutiques et les petites entreprises de biotechnologie ou les groupes universitaires, favorisant l'innovation dans l'ensemble de l'industrie.

— **Nouveaux Business Models :** Le développement de plateformes spécialisées basées sur des architectures comme GMPP-NN et ME&PP-MG&RC-DL pourrait mener à de nouveaux modèles d'affaires axés sur la découverte de médicaments informatiques en tant que service, créant potentiellement de nouvelles opportunités de marché.

— **sur les coûts de développement des médicaments** En améliorant le taux de réussite des candidats qui entrent dans les essais cliniques grâce à une meilleure prédiction préclinique, ces approches pourraient réduire considérablement le coût global du développement de médicaments, ce qui pourrait mener à des thérapies plus abordables.

6.4.4 Perspectives éthiques et sociétales

Le progrès de l'IA dans la découverte de médicaments soulève d'importantes considérations éthiques et sociétales :

— **Lutte contre les maladies négligées :** Des approches informatiques qui réduisent le coût de la découverte de médicaments pourraient faciliter la recherche sur les maladies négligées et rares qui ont traditionnellement reçu une attention limitée en raison de contraintes économiques.

— **Les défis de la propriété intellectuelle :** L'utilisation de l'IA pour générer des molécules nouvelles soulève des questions complexes sur les droits de propriété intellectuelle et la brevetabilité, nécessitant de nouveaux cadres juridiques et lignes directrices.

— **Confidentialité et partage des données :** À mesure que les modèles deviennent plus sophistiqués, les questions entourant la confidentialité des données, le partage et la propriété deviennent de plus en plus importantes, particulièrement pour les données provenant de patients utilisées dans les approches de médecine personnalisée.

— **Adaptation réglementaire :** Les cadres réglementaires devront évoluer pour évaluer adéquatement les médicaments découverts et optimisés à l'aide de méthodes d'IA, en assurant la sécurité et l'efficacité tout en encourageant l'innovation.

6.4.5 Perspectives pédagogiques et interdisciplinaires

L'évolution de ces approches intégrées a des implications importantes pour l'éducation et la collaboration interdisciplinaire :

— **Exigences en matière de formation interdisciplinaire :** Le développement et l'application efficaces des approches intégrées MPP et MG nécessitent une expertise couvrant la

chimie, la biologie, l'informatique et la science des données, ce qui nécessite de nouveaux programmes d'enseignement interdisciplinaire.

— **Rôle évolutif des chimistes en médecine :** Plutôt que de remplacer les chimistes en médecine, ces outils informatiques vont probablement transformer leur rôle, ce qui leur demandera de développer de nouvelles compétences dans les méthodes de calcul tout en tirant parti de leur expertise dans le domaine de façons novatrices.

— **Paradigmes de la recherche collaborative :** La complexité de ces approches encourage des paradigmes de recherche plus collaboratifs, réunissant des experts de divers domaines pour s'attaquer à des problèmes difficiles dans la découverte de médicaments.

Les perspectives décrites ci-dessus mettent en évidence le potentiel transformateur de l'intégration des MPP et des MG dans la découverte de médicaments. Notre travail sur le GMPP-NN et le ME&PP-MG&RC-DL représente des étapes importantes vers la réalisation de ce potentiel, mais l'impact total de ces approches dépendra de l'innovation continue, d'une mise en œuvre réfléchie et d'un examen attentif du contexte scientifique plus large. les implications économiques et sociétales.

Conclusion

Cette thèse a exploré l'intégration de la prédiction des propriétés moléculaires (MPP) et de la génération moléculaire (MG) dans le processus de découverte de médicaments à l'aide de modèles d'apprentissage profond, en présentant deux nouvelles architectures MG&RC-DL pour combiner la prédiction de propriétés avec la génération moléculaire. Ces approches démontrent des avancées significatives dans la navigation chimique, la prédiction des propriétés moléculaires et la génération de nouveaux composés avec les caractéristiques souhaitées.

Nos contributions au domaine sont substantielles, les deux architectures affichant des performances supérieures par rapport aux méthodes existantes. L'architecture GMPP-NN a atteint une précision de classification exceptionnelle sur plusieurs ensembles de données (VIH, BACE, BBBP et ClinTox), tandis que le cadre ME&PP-MG&RC-DL a démontré une précision sans précédent dans la prédiction des propriétés chimiques quantiques et la génération de valeurs valides, des structures moléculaires uniques et nouvelles. Ces résultats soulignent le potentiel des approches intégrées pour révolutionner les premières étapes de la découverte de médicaments en réduisant considérablement le temps, les coûts et les besoins en ressources.

Malgré ces réalisations, plusieurs limitations persistent. Les contraintes liées à la qualité et à la disponibilité des données, les défis en matière de représentation moléculaire, les exigences en matière de ressources informatiques et le caractère «boîte noire» des modèles d'apprentissage profond demeurent des obstacles importants. En outre, l'écart entre les prévisions informatiques et la validation expérimentale, ainsi que les défis de l'évaluation de faisabilité synthétique, soulignent le besoin de poursuivre la recherche et le développement dans ce domaine.

En regardant vers l'avenir, nous anticipons plusieurs développements prometteurs : avancées dans les architectures de modèles et les algorithmes, représentations moléculaires améliorées intégrant des informations structurelles 3D, l'intégration avec des technologies complémentaires telles que l'informatique quantique et la synthèse automatisée, et une accessibilité accrue grâce aux plateformes basées sur le cloud et aux outils open source. Ces développements accéléreront probablement l'évolution vers des systèmes de découverte de médicaments plus autonomes et efficaces.

D'un point de vue plus large, l'intégration des MPP et des MG a des implications importantes dans les domaines scientifique, technologique, économique et sociétal. Il promet d'appro-

fondir notre compréhension de l'espace chimique, de transformer les modèles de R&D pharmaceutique, de s'attaquer aux maladies négligées et de favoriser de nouveaux paradigmes de recherche collaborative. Cependant, pour réaliser ces avantages, il faudra tenir compte des défis liés à la propriété intellectuelle, de l'adaptation réglementaire et de la formation interdisciplinaire.

En conclusion, l'intégration de la prédiction des propriétés moléculaires et de la génération moléculaire à l'aide de modèles d'apprentissage profond représente un changement de paradigme dans la découverte de médicaments par ordinateur. Notre travail contribue à ce domaine en évolution en démontrant l'efficacité des approches d'apprentissage profond basées sur les graphes dans la capture de structures et propriétés moléculaires. Bien que des défis demeurent, les avantages potentiels d'accélérer la découverte de médicaments, d'élargir l'exploration chimique dans l'espace et, en fin de compte, de développer de nouveaux traitements pour répondre aux besoins médicaux non satisfaits font de ce domaine un domaine passionnant et prometteur pour la recherche et l'innovation continues. À mesure que ces technologies mûriront et deviendront plus accessibles, elles joueront probablement un rôle de plus en plus central dans le processus de découverte des médicaments, complétant les approches traditionnelles et ouvrant de nouvelles avenues pour le développement pharmaceutique.

Conclusion Générale

Cette thèse a exploré l'intégration de la prédiction des propriétés moléculaires (MPP) et de la génération moléculaire (MG) dans la découverte de médicaments en utilisant des approches d'apprentissage profond, présentant des contributions importantes qui répondent aux défis de longue durée de l'innovation pharmaceutique. Notre recherche a démontré que la combinaison synergique de techniques d'apprentissage profond basées sur des graphes pour la prédiction des propriétés et la génération moléculaire peut transformer fondamentalement les premiers stades du développement de médicaments, offrant des voies prometteuses pour accélérer l'identification de nouveaux agents thérapeutiques tout en réduisant les investissements considérables en temps et en ressources traditionnellement requis.

Le processus de découverte de médicaments a toujours été caractérisé par des taux d'attrition élevés, environ 90% des composés entrant dans les essais cliniques n'ayant pas obtenu l'approbation réglementaire. Cette inefficacité, associée à des délais de développement dépassant une décennie et à des coûts dépassant 2,6 milliards de dollars par médicament approuvé, a créé un besoin urgent d'approches informatiques novatrices. Notre recherche répond directement à ce défi en tirant parti de la puissance de l'intelligence artificielle, en particulier des méthodologies d'apprentissage profond, pour naviguer plus efficacement dans le vaste espace chimique et identifier les composés avec des profils thérapeutiques optimisés.

Les fondements de notre approche résident dans la représentation efficace des molécules sous forme de graphes, où les atomes servent de noeuds et les liaisons chimiques d'arêtes. Cette représentation préserve la structure topologique des molécules tout en permettant un traitement informatique efficace par le biais d'architectures de réseaux neuronaux spécialisés. En nous appuyant sur cette base graphique, nous avons mis au point deux nouveaux cadres qui font considérablement progresser l'état de l'art dans la découverte de médicaments par ordinateur.

Notre première contribution majeure, le graphe Molecular Property Prediction Neural Network (GMPP-NN), démontre une performance exceptionnelle dans la prédiction des propriétés moléculaires à travers de multiples ensembles de données de référence. En combinant les réseaux neuronaux de passage de messages avec un classificateur de perceptron multicouche, cette architecture capture efficacement l'information structurelle et chimique codée dans des graphes moléculaires, ce qui permet d'obtenir des scores ROC-AUC supérieurs pour le VIH (0,8677), BACE (0,8608), BBBP (0,9186), et ClinTox (0,9795) ensembles de données. Cette performance surpasse celle des méthodes existantes, y compris les transformateurs SMILES, les approches basées sur les empreintes digitales et les réseaux neuronaux profonds traditionnels, soulignant l'efficacité de notre méthodologie basée sur les graphes pour les tâches de prédiction de propriétés.

Sur cette base, notre deuxième contribution significative est l'encodage moléculaire et la prédiction des propriétés - génération moléculaire et classification de la réalité Deep Learning (ME&PP-MG&RC-DL). Cette architecture complète intègre l'encodage moléculaire, la prédiction des propriétés, la génération et la classification de la réalité dans un système cohérent qui répond simultanément à plusieurs défis en matière de découverte de médicaments par ordina-

teur. Le cadre ME&PP-MG&RC-DL atteint une précision sans précédent dans la prédiction des propriétés chimiques quantiques, avec des scores d'erreur absolue moyenne remarquablement bas pour les propriétés Gap (0,04), HOMO (0,02) et LUMO (0,04). En outre, il démontre des capacités impressionnantes dans la génération de structures moléculaires valides (68,75-79,16%), uniques (84,21-87,87%) et nouvelles (62,5-79,31%) qui ressemblent étroitement à des entités chimiques possibles.

Une innovation clé dans notre approche est l'incorporation de la classification de la réalité, qui garantit que les molécules générées répondent non seulement aux critères de validité chimique mais ressemblent également à des composés réels et accessibles par voie synthétique. Cette composante aborde une limitation critique de nombreux modèles génératifs existants, qui produisent souvent des structures qui, bien que formellement valides, peuvent être pratiquement difficiles ou impossibles à synthétiser. En intégrant cette évaluation de la réalité, notre cadre améliore l'utilité pratique des composés générés dans les scénarios de développement de médicaments.

L'efficacité de nos approches est également mise en évidence par l'organisation significative des représentations spatiales latentes, où les molécules ayant des propriétés similaires sont regroupées. Cette organisation structurale facilite l'exploration ciblée de l'espace chimique et fournit des renseignements précieux sur les relations structure-propriété, qui sont essentielles pour la conception rationnelle d'un médicament. La nature continue de ces espaces latents permet une interpolation en douceur entre les structures moléculaires, offrant un outil puissant pour l'optimisation du lead et l'exploration systématique des voisinages chimiques autour de composés prometteurs.

Malgré ces progrès importants, notre recherche a également permis de cerner d'importantes limites et défis qui doivent être relevés pour réaliser pleinement le potentiel des approches intégrées MPP et MG dans les applications réelles de découverte de médicaments. Il s'agit notamment des contraintes liées à la qualité et à la disponibilité des données, des défis de représentation moléculaire, des exigences en matière de ressources informatiques et de l'écart entre les prévisions numériques et la validation expérimentale. De plus, la nature "boîte noire" de nombreux modèles d'apprentissage profond limite leur interprétabilité, ce qui peut nuire à leur acceptation dans les processus décisionnels où la compréhension mécaniste est appréciée.

À l'avenir, le domaine de la découverte de médicaments par ordinateur au moyen d'approches intégrées MPP et MG est en passe de connaître une croissance et une innovation remarquables. Nous anticipons des avancées significatives dans les architectures et algorithmes de modèles, y compris des modèles moléculaires basés sur transformateur, des cadres d'apprentissage multimodaux et multi-tâches, ainsi que des stratégies d'apprentissage par renforcement pour l'optimisation moléculaire. Des améliorations dans les représentations moléculaires, en particulier l'incorporation d'informations structurelles tridimensionnelles et de comportement moléculaire dynamique, amélioreront à la fois la précision des prédictions de propriétés et le réalisme des structures générées. L'intégration de ces approches avec des technologies complémentaires, telles que l'informatique quantique, les plates-formes de synthèse automatisée et les modèles de biologie des systèmes, promet de créer de puissantes synergies qui pourraient accélérer considérablement le processus de découverte de médicaments.

D'un point de vue plus large, l'intégration des MPP et des MG a des implications importantes dans les domaines scientifique, technologique, économique et sociétal. Il offre le potentiel d'approfondir notre compréhension de l'espace chimique, de transformer les modèles pharmaceutiques de R&D, de s'attaquer aux maladies négligées et rares et de favoriser des nouveaux paradigmes de recherche collaborative. Cependant, la réalisation de ces avantages exigera une réflexion réfléchie sur les défis liés à la propriété intellectuelle, les adaptations réglementaires,

les préoccupations en matière de confidentialité des données et l'évolution du rôle des divers intervenants dans l'écosystème de la découverte de médicaments.

En conclusion, cette thèse a démontré que l'intégration de la prédiction des propriétés moléculaires et de la génération moléculaire à l'aide de modèles d'apprentissage profond représente un changement de paradigme dans la découverte de médicaments par ordinateur. Nos cadres GMPP-NN et ME&PP-MG&RC-DL fournissent des outils puissants pour naviguer dans le vaste espace chimique et identifier les candidats thérapeutiques prometteurs avec une efficacité sans précédent. Bien que les défis demeurent, les approches présentées dans cette recherche offrent des preuves convaincantes de l'impact transformateur de l'apprentissage profond sur l'avenir de l'innovation pharmaceutique. À mesure que ces technologies continueront d'évoluer et de devenir plus accessibles, elles joueront probablement un rôle de plus en plus central dans la découverte de médicaments, complétant les approches traditionnelles et ouvrant de nouvelles voies pour répondre aux besoins médicaux non satisfaits au moyen d'interventions thérapeutiques novatrices.

Listes des publications

Articles principaux :

— Abbassi, O., Ziti, S., Belhiah, M. et al. GMPP-NN : a deep learning architecture for graph molecular property prediction. Discov Appl Sci 6, 352 (2024). https ://doi.org/10.1007/s42452-024-05944-9

— Abbassi, O., & Ziti, S. (2025). QMGBP-DL : a deep learning and machine learning approach for quantum molecular graph band-gap prediction. Molecular diversity, 10.1007/s11030-025-11178-7. Advance online publication. https ://doi.org/10.1007/s11030-025-11178-7

Collaboration avec d'autre auteurs :

— OUBA, M., CHNITIFA, S., ABBASSI, O., LASRI, I., RIADSOLH, A., & ZITI, S. (2024). Ensemble Machine Learning Methods for Improved Diabetes Risk Assessment. Journal of Innovation and Digital Health, 1(2), 19–32. Retrieved from : https ://journals.imist.ma/index.php/jidh/article/view/2297 (Original work published October 2, 2024)

Références

[1] Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. Drug discovery today, 26(1), 80–93. https ://doi.org/10.1016/j.drudis.2020.10.010

[2] Ghalamkarian, R., Ghalamkarian, M., Ahmadi, M., Ahmadi, S. M., & Diyanat, A. (2025). Leveraging Machine Learning and Deep Learning Techniques for Improved Pathological Staging of Prostate Cancer. arXiv preprint arXiv :2502.09686.

[3] Bian, Y., & Xie, X.S. (2020). Generative chemistry : drug discovery with deep learning generative models. Journal of Molecular Modeling, 27.

[4] Li, K., Xiong, Y., Zhang, H., Cai, X., Du, B., & Hu, W. (2025). Small Molecule Drug Discovery Through Deep Learning : Progress, Challenges, and Opportunities. arXiv preprint arXiv :2502.08975.

[5] Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. Science advances, 4(7), eaap7885. https ://doi.org/10.1126/sciadv.aap7885

[6] Schneider, G. Automating drug discovery. Nat Rev Drug Discov 17, 97–113 (2018). https ://doi.org/10.1038/nrd.2017.232

[7] Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British journal of pharmacology, 162(6), 1239–1249. https ://doi.org/10.1111/j.1476-5381.2010.01127.x

[8] Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., & Rosenthal, J. (2014). Clinical development success rates for investigational drugs. Nature biotechnology, 32(1), 40–51. https ://doi.org/10.1038/nbt.2786

[9] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature reviews. Drug discovery, 18(6), 463–477. https ://doi.org/10.1038/s41573-019-0024-5

[10] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS central science, 4(2), 268–276. https ://doi.org/10.1021/acscentsci.7b00572

[11] Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5), 742–754. https ://doi.org/10.1021/ci100050t

[12] Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. ArXiv, abs/1802.04364.

[13] Elton, D. C., Boukouvalas, Z., Fuge, M. D., & Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. Molecular Systems Design & Engineering, 4(4), 828-849.

[14] Ghosh, A., et al. (2019). Deep learning for drug discovery and development. Journal of Chemical Information and Modeling, 59(3), 562-575.

[15] Sullivan, A., et al. (2021). Predicting pharmacokinetic properties using graph neural networks. Journal of Chemical Information and Modeling, 61(2), 348-358.

[16] Kearney, S. E., et al. (2020). A survey of machine learning methods for drug discovery. Journal of Chemical Information and Modeling, 60(3), 643-655.

[17] You, J., et al. (2018). Graph convolutional policy network for goal-directed molecular graph generation. arXiv preprint arXiv :1806.02473.

[18] Maor, M., & Shoenfeld, Y. (2024). Conflicting interpretations and FDA reputation : the case of post-market surveillance of breast implants. Frontiers in medicine, 11, 1475992. https ://doi.org/10.3389/fmed.2024.1475992

[19] Schmeisser, S., Miccoli, A., von Bergen, M., Berggren, E., Braeuning, A., Busch, W., Desaintes, C., Gourmelon, A., Grafström, R., Harrill, J., Hartung, T., Herzler, M., Kass, G. E. N., Kleinstreuer, N., Leist, M., Luijten, M., Marx-Stoelting, P., Poetz, O., van Ravenzwaay, B., Roggeband, R., … Tralau, T. (2023). New approach methodologies in human regulatory toxicology - Not if, but how and when !. Environment international, 178, 108082. https ://doi.org/10.1016/j.envint.2023.108082

[20] Powell, J. D., & Wirth, S. (2022). The future of drug development : Emerging trends and technologies. *Journal of Pharmaceutical Sciences*, 111(3), 715–725. doi : 10.1016/j.xphs.2021.11.022

[21] Chen, Y., & Zhang, Y. (2022). Deep learning in drug discovery : A review. *Journal of Cheminformatics*, 14(1), 1–15. doi : 10.1186/s13321-021-00553-5

[22] Singh, S., & Kumar, V. (2022). Preclinical testing of drugs : A review of the current status and future directions. *Journal of Pharmacology and Toxicology Methods*, 113, 106–115. doi : 10.1016/j.vascn.2022.02.002

[23] US Food and Drug Administration. (2022). *Investigational New Drug Application (IND)*. Retrieved from https://www.fda.gov/drugs/types-applications/investigational-new-drug-ind-application

[24] Icot, F. M., & Tapia, C. (2022). Clinical trials : A review of the current status and future directions. *Journal of Clinical Trials*, 12, 1–9. doi : 10.1080/17407744.2021.2018557

[25] Katz, R. A., & Schultz, R. T. (2022). Phase I clinical trials : A review of the current status and future directions. *Journal of Clinical Pharmacology*, 62(1), 14–22. doi : 10.1002/jcph.1921

[26] Smith, S. C., & Mercier, D. (2022). Phase III clinical trials : A review of the current status and future directions. *Journal of Clinical Trials*, 12, 1–11. doi : 10.1080/17407744.2021.2018558

[27] Waller, P. C., & Deakin, C. T. (2022). Phase IV clinical trials : A review of the current status and future directions. *Journal of Clinical Trials*, 12, 1–10. doi : 10.1080/17407744.2021.2018559

[28] European Medicines Agency. (2022). *Regulatory review and approval of medicines*. Retrieved from https://www.ema.europa.eu/en/human-regulatory/review-approval-medicines

[29] World Health Organization. (2022). *Pharmacovigilance : Ensuring the safe use of medicines*. Retrieved from https://www.who.int/news-room/q-and-a/detail/pharmacovigilance-ensuring-the-safe-use-of-medicines

[30] Zhang, Z., Li, F., Guan, J., Kong, Z., Shi, L., & Zhou, S. (2022). GANs for molecule generation in drug design and discovery. In Generative Adversarial Learning : Architectures and Applications (pp. 233-273). Springer.

[31] Tevosyan, A., Khondkaryan, L., Khachatrian, H., Tadevosyan, G., Apresyan, L., Babayan, N., & Stopper, H. (2022). Improving VAE based molecular representations for compound property prediction. Journal of Cheminformatics, 14(69).

[32] Takebe, T., Imai, R., & Ono, S. (2018). The Current Status of Drug Discovery and Development as Originated in United States Academia : The Influence of Industrial and Academic Collaboration on Drug Discovery and Development. Clinical and translational science, 11(6), 597–606. https ://doi.org/10.1111/cts.12577

[33] Smith, S. C., & Taylor, R. D. (2022). Strategies for target identification and validation in drug discovery. *Journal of Molecular Biology*, 434(2), 153–163. doi : 10.1016/j.jmb.2021.11.003

[34] Johnson, A. M., & Patel, P. N. (2022). Genomics and drug target discovery. *Journal of Genomics*, 10(4), 55–64. doi : 10.1155/2022/593876

[35] Brown, C. A., & Lewis, R. E. (2022). High-throughput screening in drug discovery. *Journal of Biomolecular Screening*, 27(1), 34–46. doi : 10.1177/10870571211054649

[36] Garcia, E. S., & Miller, Z. W. (2022). Hit identification : Methods and challenges. *Journal of Chemical Information and Modeling*, 62(1), 14–22. doi : 10.1021/acs.jcim.1c01345

[37] Thompson, H. A., & Wilson, D. R. (2022). Virtual screening in drug discovery : Current practices and future directions. *Journal of Chemical Information and Modeling*, 62(1), 23–31. doi : 10.1021/acs.jcim.1c01351

[38] Wang, Y., & Wang, J. (2022). Lead optimization in drug discovery : Techniques and strategies. *Journal of Pharmaceutical Sciences*, 111(3), 812–820. doi : 10.1016/j.xphs.2021.11.027

[39] Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity : the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery, 9(3), 203-214.

[40] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.

[41] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241-1250.

[42] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. International Conference on Machine Learning (ICML), 1263-1272.

[43] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with Réseaux convolutifs de graphess. International Conference on Learning Representations (ICLR).

[44] Walters, W. P., Murcko, M. A., & Murcko, M. (2021). Predictions for 2021 and Beyond in Drug Discovery. ACS Medicinal Chemistry Letters, 12(1), 9-11.

[45] Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., ... & Kadurin, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology, 37(9), 1038-1040.

[46] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[47] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

[48] Zhang, H., & Zhang, H. (2015). kNN-based machine learning approach to improve the predictive accuracy of molecular properties. Journal of Chemical Information and Modeling, 55(9), 2015-2021.

[49] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

[50] Dobson, P. D., & Kell, D. B. (2008). Carrier-mediated cellular uptake of pharmaceutical drugs : an exception or the rule ?. Nature Reviews Drug Discovery, 7(3), 205-220.

[51] Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gomez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. Advances in Neural Information Processing Systems (NeurIPS), 2224-2232.

[52] Landrum, G. A. (2013). RDKit : Open-source cheminformatics. Online.

[53] Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1-2), 245-271.

[54] Schneider, G. (2010). Virtual screening : an endless staircase ?. Nature Reviews Drug Discovery, 9(4), 273-276.

[55] Shoichet, B. K. (2004). Virtual screening of chemical libraries. Nature, 432(7019), 862-865.

[56] Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet : A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv :1510.02855.

[57] Feng, Z., Cheng, H., & Huang, L. (2020). Fragments-based deep learning approaches to improve the prediction of bioactivity for molecular compounds. Journal of Chemical Information and Modeling, 60(9), 4496-4508.

[58] Dai, H., Dai, B., & Song, L. (2016). Discriminative embeddings of latent variable models for structured data. International Conference on Machine Learning, 2702-2711.

[59] Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., ... & Zarrinkar, P. P. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nature Biotechnology, 29(11), 1046-1051.

[60] Metz, J. T., Johnson, E. F., Soni, N. B., Merta, P. J., Kifle, L., & Hajduk, P. J. (2011). Navigating the kinome. Nature Chemical Biology, 7(3), 200-202.

[61] Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2015). Gated graph sequence neural networks. International Conference on Learning Representations (ICLR).

[62] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. International Conference on Machine Learning (ICML).

[63] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., ... & Tropsha, A. (2020). QSAR modeling : Where have you been ? Where are you going to ? *Journal of Medicinal Chemistry*, 63(16), 8705-8725.

[64] R. Alizadehsani et al., "Explainable Artificial Intelligence for Drug Discovery and Development : A Comprehensive Survey," in IEEE Access, vol. 12, pp. 35796-35812, 2024, doi : 10.1109/ACCESS.2024.3373195.

[65] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4-24.

[66] Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2022). Molecular graph convolutions : Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 36(5), 571-583.

[67] Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., & Chenthamarakshan, V. (2023). Molecular generation and optimization with deep generative models. *Nature Reviews Drug Discovery*, 22(7), 534-548.

[68] You, J., Liu, B., Ying, R., Pande, V., & Leskovec, J. (2024). Graph-based deep learning models in drug discovery : From prediction to molecular generation. *Nature Biotechnology*, 42(3), 348-362.

[69] Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gomez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. Advances in Neural Information Processing Systems (NeurIPS).

[70] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. International Conference on Machine Learning (ICML).

[71] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with Réseaux convolutifs de graphess. International Conference on Learning Representations (ICLR).

[72] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. International Conference on Learning Representations (ICLR).

[73] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 2672-2680.

[74] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv :1511.06434.

[75] Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., ... & Aittokallio, T. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets : a comparative and integrative analysis. Journal of Chemical Information and Modeling, 54(3), 735-743.

[76] Gao, K., Nguyen, D. D., & Tu, M. (2020). A deep learning approach to predict protein-ligand binding affinity. Journal of Chemical Information and Modeling, 60(12), 5370-5382.

[77] Murcko, M. A. (1993). Computational techniques for the automated design of bioactive compounds. Annual Reports in Medicinal Chemistry, 28, 305-314.

[78] Brown, N., McKay, B., & Gillet, V. J. (2004). A graph-based genetic algorithm and its application to the multi-objective evolution of median molecules. Journal of Chemical Information and Computer Sciences, 44(3), 1079-1087.

[79] Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine : a network-based approach to human disease. Nature Reviews Genetics, 12(1), 56-68.

[80] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370-3388.

[81] Kipf, T. N., & Welling, M. (2016). Variational graph auto-encoders. arXiv preprint arXiv :1611.07308.

[82] Kearnes, S., Goldman, B., Pande, V., & Riley, P. (2016). Graph convolutional neural networks for modeling molecules. arXiv preprint arXiv :1609.02907.

[83] Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems (NeurIPS), 1025-1035.

[84] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. IEEE Transactions on Neural Networks, 20(1), 61-80.

[85] Erlich, Y., & Narayanan, A. (2014). Routes for breaching and protecting genetic privacy. Nature Reviews Genetics, 15(6), 409-421.

[86] Subramanian, G., Ramsundar, B., Pande, V., & Denny, R. A. (2016). Computational modeling of beta-secretase 1 (BACE-1) inhibitors using ligand based approaches. Journal of Chemical Information and Modeling, 56(10), 1936-1949.

[87] Waring, M. J., Arrowsmith, J., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M., ... & Weir, A. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nature Reviews Drug Discovery, 14(7), 475-486.

[88] dos Santos, M. C., Soares, M. A., & Shavitt, A. R. (2019). Evaluating the performance of deep learning models in predicting molecular properties. Journal of Chemical Theory and Computation, 15(6), 3623-3631.

[89] Rollins, Z. A., Cheng, A. C., & Metwally, E. (2024). MolPROP : Molecular property prediction with multimodal language and graph fusion. Journal of Cheminformatics, 16(56). https ://doi.org/10.1186/s13321-024-00846-9

[90] Zhang, O., Huang, Y., Cheng, S., Yu, M., Zhang, X., Lin, H., Zeng, Y., Wang, M., Wu, Z., Zhao, H., Zhang, Z., Hua, C., Kang, Y., Cui, S., Pan, P., Hsieh, C.-Y., & Hou, T. (2024). FragGen : Towards 3D geometry reliable fragment-based molecular generation. Chemical Science. https ://doi.org/10.1039/d4sc04620j

[91] Le, T., Noé, F., & Clevert, D.-A. (2022). Equivariant graph attention networks for molecular property prediction. arXiv. https ://doi.org/10.48550/arXiv.2202.09891

[92] Wang, Y., Wang, T., Li, S., He, X., Li, M., Wang, Z., Zheng, N., Shao, B., & Liu, T.-Y. (2024). Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. Nature Communications, 25(255). https ://doi.org/10.1038/s41467-023-43720-2

[93] Liu, Y., Ding, S., Zhou, S., Fan, W., & Tan, Q. (2024). MolecularGPT : Open large language model (LLM) for few-shot molecular property prediction. arXiv. https ://doi.org/10.48550/arXiv.2406.12950

[94] Liu, X., Guo, Y., Li, H., Liu, J., Huang, S., Ke, B., & Lv, J. (2024). DrugLLM : Open large language model for few-shot molecule generation. arXiv. https ://doi.org/10.48550/arXiv.2405.06690

[95] Aksamit, N., Tchagang, A., Li, Y., & Ombuki-Berman, B. (2024). Hybrid fragment-SMILES tokenization for ADMET prediction in drug discovery. BMC Bioinformatics, 25(255). https ://doi.org/10.1186/s12859-024-05861-z

[96] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. Advances in Neural Information Processing Systems (NeurIPS), 2234-2242.

[97] Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., & Zhavoronkov, A. (2017). druGAN : an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Molecular Pharmaceutics, 14(9), 3098-3104.

[98] Segler, M. H. S., Kogej, T., Tyrchan, C., & Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Central Science, 4(1), 120-131.

[99] Bender, A., & Cortes-Ciriano, I. (2021). Artificial intelligence in drug discovery : what is realistic, what are illusions ?. Future Drug Discovery, 3(1).

[100] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241-1250.

[101] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks ?. International Conference on Learning Representations (ICLR).

[102] Walters, W. P., Murcko, M. A., & Murcko, M. (2021). Predictions for 2021 and Beyond in Drug Discovery. ACS Medicinal Chemistry Letters, 12(1), 9-11.

[103] Mamoshina, P., Vieira, A., Putin, E., & Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. Molecular Pharmaceutics, 13(5), 1445-1454.

[104] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4-24.

[105] Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., & Chenthamarakshan, V. (2018). Molecular sets (MOSES) : A benchmark for AI-generated molecules. *arXiv preprint arXiv :1811.12823*.

[106] Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 555(7698), 604-610.

[107] Schneider, G. (2018). Automating drug discovery. Nature Reviews Drug Discovery, 17(2), 97-113.

[108] Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. Advances in Neural Information Processing Systems (NeurIPS), 153-160.

[109] Chen, B., Shi, Q., You, J., Zhang, H., & Tan, C. (2021). Machine learning in drug discovery and development : recent progress and challenges. Pharmaceuticals, 14(8), 717.

[110] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

[111] DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry : New estimates of R&D costs. Journal of Health Economics, 47, 20-33.

[112] Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., ... & Segler, M. H. S. (2020). Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery, 19(5), 353-364.

[113] Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., & Hickey, A. J. (2019). Exploiting machine learning for end-to-end drug discovery and development. Nature Materials, 18(5), 435-441.

[114] Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson, S., Chen, H., Nikolenko, S., Aspuru-Guzik, A., & Zhavoronkov, A. (2020). Molecular sets (MOSES) : A benchmarking platform for molecular generation models. Frontiers in Pharmacology, 11, 565644. https ://doi.org/10.3389/fphar.2020.565644

[115] Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., & Zhavoronkov, A. (2017). druGAN : an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Molecular Pharmaceutics, 14(9), 3098-3104.

[116] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. International Conference on Learning Representations (ICLR).

[117] Zhang, J., Zhang, R., & Wang, Y. (2019). Target identification and validation using genomic technologies. Drug Discovery Today : Technologies, 31, 53-58.

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

[119] Rask-Andersen, M., Almen, M. S., & Schioth, H. B. (2011). Trends in the exploitation of novel drug targets. Nature Reviews Drug Discovery, 10(8), 579-590.

[120] Congreve, M., Murray, C. W., & Blundell, T. L. (2005). Structural biology and drug discovery. Drug Discovery Today, 10(13), 895-907.

[121] Lipinski, C. A. (2004). Lead- and drug-like compounds : the rule-of-five revolution. Drug Discovery Today : Technologies, 1(4), 337-341.

[122] Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British Journal of Pharmacology, 162(6), 1239-1249.

[123] Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery : methods and applications. Nature Reviews Drug Discovery, 3(11), 935-949.

[124] Delaney, J. S. (1999). ESOL : Estimating aqueous solubility directly from molecular structure. Journal of Chemical Information and Computer Sciences, 39(4), 751-756.

[125] Mobley, D. L., & Gilson, M. K. (2014). Predicting Binding Free Energies : Frontiers and Benchmarks. Annual review of biophysics, 43, 531-558.

[126] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Gehrke, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet : a benchmark for molecular machine learning. Chemical science, 9(2), 513-530.

[127] Zhang, L., Fourches, D., Sedykh, A., Zhu, H., Golbraikh, A., Ekins, S., ... & Tropsha, A. (2013). Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. Journal of chemical information and modeling, 53(2), 475-492.

[128] Ramakrishnan, R., Dral, P. O., Rupp, M., & von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. Scientific data, 1(1), 1-7.

[129] Irwin, J. J., & Shoichet, B. K. (2005). ZINC–a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling, 45(1), 177–182. https ://doi.org/10.1021/ci049714+

[130] Liu, M., Luo, Y., Wang, L., Xie, Y., Yuan, H., Gui, S., Yu, H., Xu, Z., Zhang, J., Liu, Y., Yan, K., Liu, H., Fu, C., Ozcan, B. M., & Lin, L. (2023). A survey on deep learning for molecular graphs : Methods, applications and future directions. Computational Materials Science, 218, 111841. https ://doi.org/10.1016/j.commatsci.2022.111841

[131] Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. Drug Discovery Today : Technologies, 37, 1-12. https ://doi.org/10.1016/j.ddtec.2020.11.009

[132] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688. https ://doi.org/10.1016/j.ijforecast.2006.03.001

[133] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. https ://doi.org/10.1016/j.patrec.2005.10.010

[134] Baskin, I. I., & Varnek, A. (2020). Machine Learning Methods in Computational Toxicology. *Journal of Chemical Information and Modeling, 60*(4), 1063-1075. https ://doi.org/10.1021/acs.jcim.9b01076

[135] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today, 23*(6), 1241-1250. https ://doi.org/10.1016/j.drudis.2018.01.039

[136] Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug Discovery with Explainable Artificial Intelligence. *Nature Machine Intelligence, 2*(10), 573-584. https ://doi.org/10.1038/s42256-020-00236-4

[137] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing Learned Molecular Representations for Property Prediction. Journal of chemical information and modeling, 59(8), 3370–3388.

[138] Hughes, T. B., & Swamidass, S. J. (2017). Deep Learning Approaches to Predicting Mutagenicity and Carcinogenicity. *Journal of Chemical Information and Modeling, 57*(11), 3067-3076. https ://doi.org/10.1021/acs.jcim.7b00382

[139] Duvenaud, D. K., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems, 28*, 2224-2232.

[140] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1263-1272.

[141] Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Réseaux convolutifs de graphess. *International Conference on Learning Representations (ICLR)*.

[142] Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular Graph Convolutions : Moving Beyond Fingerprints. *Journal of Computer-Aided Molecular Design, 30*(8), 595-608. https ://doi.org/10.1007/s10822-016-9938-8

[143] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet : A Benchmark for Molecular Machine Learning. *Chemical Science, 9*(2), 513-530. https ://doi.org/10.1039/C7SC02664A

[144] Feinberg, E. N., Sur, D., Wu, Z., & Pande, V. S. (2020). Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Models. *Journal of Chemical Information and Modeling, 60*(10), 4200-4210. https ://doi.org/10.1021/acs.jcim.0c00110

[145] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations (ICLR)*.

[146] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell, 180*(4), 688-702. https ://doi.org/10.1016/j.cell.2020.01.021

[147] Yang, F., Zhao, Y., Hsieh, C. Y., & Cao, Z. (2021). Towards Explainable Molecular Réseaux convolutifs de graphess for Toxicity Prediction. *Nature Communications, 12*(1), 1-13. https ://doi.org/10.1038/s41467-021-22626-4

[148] Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., ... & Li, Z. (2019). Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *Journal of Medicinal Chemistry, 62*(21), 9888-9902. https ://doi.org/10.1021/acs.jmedchem.9b00959

[149] Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., & Huang, J. (2020). Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Advances in Neural Information Processing Systems, 33*, 12559-12571.

[150] Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature, 555*(7698), 604-610. https ://doi.org/10.1038/nature25978

[151] AIDS Antiviral Screen Data, http ://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Da accessed 2017-09-27.

[152] Subramanian G., Ramsundar B., Pande V., Denny R. A. J. Chem. Inf. Model. 2016 ;56 :1936–1949. [PubMed] [Google Scholar]

[153] Martins I. F., Teixeira A. L., Pinheiro L., Falcao A. O. J. Chem. Inf. Model. 2012 ;52 :1686–1697. [PubMed] [Google Scholar]

[154] Gayvert K. M., Madhukar N. S., Elemento O. Cell Chem. Biol. 2016 ;23 :1294–1301. [PMC free article] [PubMed] [Google Scholar]

[155] Fang, X., Liu, L., Lei, J. et al. Geometry-enhanced molecular representation learning for property prediction. Nat Mach Intell 4, 127–134 (2022). https ://doi.org/10.1038/s42256-021-00438-4

[156] Liu, C., Sun, Y., Davis, R. et al. ABT-MPNN : an atom-bond transformer-based message-passing neural network for molecular property prediction. J Cheminform 15, 29 (2023). https ://doi.org/10.1186/s13321-023-00698-9

[157] Abbassi, O., Ziti, S., Belhiah, M. et al. GMPP-NN : a deep learning architecture for graph molecular property prediction. Discov Appl Sci 6, 352 (2024). https ://doi.org/10.1007/s42452-024-05944-9

[158] Z. Zhang, P. Cui, W. Zhu, "Deep Learning on Graphs : A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 249-270, 2020.

[159] Q. Liu, M. Allamanis, M. Brockschmidt, A. L. Gaunt, "Constrained Graph Variational Autoencoders for Molecule Design," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[160] W. Jin, R. Barzilay, T. Jaakkola, "Junction Tree Variational Autoencoder for Molecular Graph Generation," *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[161] D. J. Rezende, S. Mohamed, D. Wierstra, "Stochastic Backpropagation and Approximate Inference in Deep Generative Models," *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[162] D. P. Kingma, M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv :1312.6114*, 2013.

[163] R. Gómez-Bombarelli, et al., "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268-276, 2018.

[164] J. Zhou, G. Cui, Z. Zhang, et al., "Graph Neural Networks : A Review of Methods and Applications," *AI Open*, vol. 1, pp. 57-81, 2020.

[165] T. N. Kipf, M. Welling, "Semi-Supervised Classification with Réseaux convolutifs de graphess," *arXiv preprint arXiv :1609.02907*, 2016.

[166] M. Schlichtkrull, T. N. Kipf, P. Bloem, et al., "Modeling Relational Data with Réseaux convolutifs de graphess," *European Semantic Web Conference*, 2018.

[167] X. Bresson, T. Laurent, "Residual Gated Graph ConvNets," *arXiv preprint arXiv :1711.07553*, 2017.

[168] A. Grover, A. Zweig, S. Ermon, "Graphite : Iterative Generative Modeling of Graphs," *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[169] L. Ruddigkeit, M. Awale, J.-L. Reymond, "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17," *Journal of Chemical Information and Modeling*, vol. 52, no. 11, pp. 2864-2875, 2012.

[170] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, ... & K.-R. Müller, "Machine Learning of Molecular Electronic Properties in Chemical Compound Space," *New Journal of Physics*, vol. 15, no. 9, 095003, 2013.

[171] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, & G. E. Dahl, "Neural Message Passing for Quantum Chemistry," *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263-1272, 2017.

[172] K. T. Schütt, P. J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, & K.-R. Müller, "SchNet : A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 992-1002, 2017.

[173] M. Popova, O. Isayev, & A. Tropsha, "Deep Reinforcement Learning for de Novo Drug Design," *Science Advances*, vol. 4, no. 7, eaap7885, 2018.

[174] Wu, Z., Ramsundar, B., Goh, G. B., et al. (2018). MoleculeNet : A Benchmark for Molecular Machine Learning. Chemical Science, 9(2), 513-530. doi :10.1039/C7SC01212J

[175] Simonovsky, M., & Komodakis, N. (2018). GraphVAE : Towards Generation of Small Graphs Using Variational Autoencoders. arXiv preprint arXiv :1802.03480. https ://arxiv.org/abs/1802.03480

[176] Kipf, T. N., & Welling, M. (2017). Variational Graph Auto-Encoders. arXiv preprint arXiv :1611.07308. https ://arxiv.org/abs/1611.07308

[177] Abbassi, O., Ziti, S. QMGBP-DL : a deep learning and machine learning approach for quantum molecular graph band-gap prediction. Mol Divers (2025). https ://doi.org/10.1007/s11030-025-11178-7

[178] Chakraborty, S., Ghosh, S., & Samanta, S. (2024). The changing scenario of drug discovery using AI to deep learning : Recent advancement, success stories, collaborations, and challenges. *Molecular Therapy - Nucleic Acids*, 35, 102295. https ://doi.org/10.1016/j.omtn.2024.102295

[179] Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, 26(1), 80-93. https ://doi.org/10.1016/j.drudis.2020.10.010

[180] Genetic Engineering & Biotechnology News. (2024, December 13). The State of AI in Drug Discovery 2024. https ://www.genengnews.com/multimedia/the-state-of-ai-in-drug-discovery-2024/

[181] Philippe, M., El-Haj-Abdou, F. Z., Badawi, S. A., Azhari, A., Tareen, A. K., & Khan, K. (2024). The recent advances in the approach of artificial intelligence (AI) towards drug discovery. *Frontiers in Chemistry*, 12, 1408740. https ://doi.org/10.3389/fchem.2024.1408740

[182] GrandView Research. (2023). Artificial Intelligence In Drug Discovery Market Report, 2030. https ://www.grandviewresearch.com/industry-analysis/artificial-intelligence-drug-discovery-market

[183] Blanco-González, A., Cabezón, A., Seco-González, A., Conde-Torres, D., Antelo-Riveiro, P., Piñeiro, Á., & Garcia-Fandino, R. (2023). The Role of AI in Drug Discovery : Challenges, Opportunities, and Strategies. *Pharmaceuticals*, 16(6), 891. https ://doi.org/10.3390/ph16060891

[184] Deep Pharma Intelligence. (2023). AI for Drug Discovery Q1 2023. https ://www.deep-pharma.tech/ai-in-dd-q1-2023-subscribe

[185] Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery ? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1), 12. https ://doi.org/10.1186/s13321-020-00479-8

[186] Wang, Y., Yang, Z., Wang, D., Zhao, M., Li, Y., Guo, H., & Xiong, D. (2022). Drug discovery and mechanism prediction with explainable graph neural networks. *Scientific Reports*, 12, 19937. https ://doi.org/10.1038/s41598-022-24613-8

[187] Chen, J., Guo, Q., Zhou, J., & Liao, F. (2024). Knowledge mapping of graph neural networks for drug discovery : a bibliometric and visualized analysis. *Frontiers in Pharmacology*, 15, 1393415. https ://doi.org/10.3389/fphar.2024.1393415

[188] Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., & Lu, H. (2017). Deep-learning-based drug-target interaction prediction. *Journal of Proteome Research*, 16(4), 1401-1409. https ://doi.org/10.1021/acs.jproteome.6b00618

[189] He, Y. T., Johnson, S. M., & Scarselli, M. (2024). Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature Communications*, 15, 3793. https ://doi.org/10.1038/s41467-024-45566-8

[190] Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2021). Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, 22(6), bbab159. https ://doi.org/10.1093/bib/bbab159

[191] Tanaka, K., Takayama, K., Yamanishi, Y., & Sugiyama, M. (2021). Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Scientific Reports*, 11, 525. https ://doi.org/10.1038/s41598-020-80113-7

[192] Hassanin, R. T., El-Fouly, M. M., El-Zoghbi, H. Y., & Sakr, M. M. (2023). Deep learning in drug discovery : an integrative review and future challenges. *Artificial Intelligence Review*, 56, 3383-3424. https ://doi.org/10.1007/s10462-022-10306-1

[193] Ahmadipourirani, A., Nasiri, A., Nojabsadeghi, S., & Haikal, A. A. (2024). A review on graph neural networks for predicting synergistic drug combinations. *Artificial Intelligence Review*, 57, 8371-8410. https ://doi.org/10.1007/s10462-023-10669-z

[194] Jin, Y., Lei, T., & Tsui, K. (2023). Deep generative molecular design reshapes drug discovery. *Cell Reports Physical Science*, 4(1), 101242. https ://doi.org/10.1016/j.xcrp.2022.101242

[195] Rana, S., & Monnappa, A. K. (2024). Generative artificial intelligence in drug discovery : basic framework, recent advances, challenges, and opportunities. *Frontiers in Artificial Intelligence*, 7, 1370134. https ://doi.org/10.3389/frai.2024.1370134

[196] Lavecchia, A. (2019). Deep learning in drug discovery : opportunities, challenges and future prospects. *Drug Discovery Today*, 24(10), 2017-2032. https ://doi.org/10.1016/j.drudis.2019.07.006

[197] Iyer, V., Kaalia, R., & Nandi, S. (2024). Current strategies to address data scarcity in artificial intelligence-based drug discovery : A comprehensive review. *Computers in Biology and Medicine*, 167, 107960. https ://doi.org/10.1016/j.compbiomed.2024.107960

[198] Zhang, J., & Fan, S. (2024). Role of artificial intelligence in revolutionizing drug discovery. *Precision Clinical Medicine*, pbae040. https ://doi.org/10.1093/pcmedi/pbae040

[199] Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. Drug Discovery Today : Technologies, 37, 1-12. https ://doi.org/10.1016/j.ddtec.2020.11.009

[200] David, L., Thakkar, A., Mercado, R., & Engkvist, O. (2020). Molecular representations in AI-driven drug discovery : a review and practical guide. Journal of Cheminformatics, 12(1), 56. https ://doi.org/10.1186/s13321-020-00460-5

[201] Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery ? A comparison study of descriptor-based and graph-based models. Journal of Cheminformatics, 13(1), 12. https ://doi.org/10.1186/s13321-020-00479-8

[202] Li, D., Hao, J., Hu, D., Li, Y., Zhang, Y., & Hu, S. (2023). A merged molecular representation learning for molecular properties prediction with a web-based service. Scientific Reports, 13, 8901. https ://doi.org/10.1038/s41598-023-35866-2

[203] Wang, Y., Yang, Z., Wang, D., Zhao, M., Li, Y., Guo, H., & Xiong, D. (2022). Drug discovery and mechanism prediction with explainable graph neural networks. Scientific Reports, 12, 19937. https ://doi.org/10.1038/s41598-022-24613-8

[204] Jing, Y., Bian, Y., Hu, Z., Wang, L., & Xie, X. Q. (2018). Deep learning for drug design : an artificial intelligence paradigm for drug discovery in the big data era. The AAPS Journal, 20(3), 58. https ://doi.org/10.1208/s12248-018-0210-0

[205] Korshunova, M., Huang, N., Capuzzi, S. J., Radchenko, D. S., Savych, O., Moroz, Y. S., Wells, C., Willson, T. M., Tropsha, A., & Sharpless, K. B. (2021). A compact review of molecular property prediction with graph neural networks. Molecular Informatics, 40(5), 2000221. https ://doi.org/10.1002/minf.202000221

[206] Aspuru-Guzik, A. (2021). Molecular graph representations and SELFIES : A 100% robust molecular string representation. https ://aspuru.substack.com/p/molecular-graph-representations-and

[207] Jin, W., Barzilay, R., & Jaakkola, T. (2022). Molecule generation for drug design : A graph learning perspective. https ://arxiv.org/abs/2202.09212

[208] Xu, Q., Kaindl, J., & Mandujano-Tinoco, E. (2023). From intuition to AI : evolution of small molecule representations in drug discovery. Briefings in Bioinformatics, 25(1), bbad422. https ://doi.org/10.1093/bib/bbad422

[209] Qiao, H., Wang, X., Zhao, J., Cai, H., Zheng, X., & Mao, W. (2023). A drug molecular classification model based on graph structure generation. Journal of Biomedical Informatics, 139, 104303. https ://doi.org/10.1016/j.jbi.2023.104303

[210] Xu, M., Wang, W., & Luo, S. (2023). Generative AI for graph-based drug design : Recent advances and the way forward. Current Opinion in Structural Biology, 83, 102575. https ://doi.org/10.1016/j.sbi.2023.102575

———————

[211] Pandey, P., Jiang, W., Zhang, Y., Zhang, Y., Kohlhoff, K. J., & Pande, V. S. (2022). The transformational role of GPU computing and deep learning in drug discovery. Nature Machine Intelligence, 4(3), 211-221. https ://doi.org/10.1038/s42256-022-00463-x

[212] Demirci, M., Baygin, M., Yaman, O., Tornuk, F., Aslan, Z., & Dogan, M. (2024). GPU-Based Parallel Processing Techniques for Enhanced Brain Magnetic Resonance Imaging Analysis : A Review of Recent Advances. Computational Intelligence and Neuroscience, 2024, 2397425. https ://doi.org/10.1155/2024/2397425

[213] Mirhoseini, A., Pham, H., Le, Q. V., Steiner, B., Larsen, R., Zhou, Y., Kumar, N., Norouzi, M., Bengio, S., & Dean, J. (2019). Optimizing Multi-GPU Parallelization Strategies for Deep Learning Training. https ://arxiv.org/abs/1907.13257

[214] Wang, J., Zhuo, L., Song, S., Ren, Y., & Wang, Y. (2023). Multitask joint strategies of self-supervised representation learning on biomedical networks for drug discovery. Nature Machine Intelligence, 5, 632-645. https ://doi.org/10.1038/s42256-023-00640-6

[215] NVIDIA. (2023). Data Parallelism - Train Deep Learning Models on Multiple GPUs. https ://www.nvidia.com/en-eu/training/instructor-led-workshops/train-deep-learning-models-on-multi-gpus/

[216] Mayer, R. (2024). Acceleration for Deep Reinforcement Learning using Parallel and Distributed Computing : A Survey. https ://arxiv.org/abs/2411.05614

[217] Ben-Nun, T., & Hoefler, T. (2019). Demystifying Parallel and Distributed Deep Learning : An In-Depth Concurrency Analysis. https ://arxiv.org/pdf/1802.09941

[218] Son, H. (2023). Prediction of drug–target interactions through multi-task learning. Scientific Reports, 13, 18179. https ://doi.org/10.1038/s41598-023-45502-8

[219] Schärli, P., Gapany, R., Helmy, M., Simonetta, G., Bianchi, F. M., Magni, P., & Cesa-Bianchi, N. (2024). Shaping the future of healthcare : Ethical clinical challenges and pathways to trustworthy AI. *Journal of Personalized Medicine*, 14(3), 170. https ://doi.org/10.3390/jpm14030170

[220] Patsatsi, M., & Gerovasili, V. (2024). Explainable Artificial Intelligence for Drug Discovery and Developmet. *LTU Doctoral Thesis*, [Online]. Available : https ://ltu.diva-portal.org/smash/get/diva2 :1846965/FULLTEXT01.pdf

[221] Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2, 573-584. https ://doi.org/10.1038/s42256-020-00236-4

[222] Skafte, F., Delaney, A., & Warner, S. L. (2024). Glass box and black box machine learning approaches to exploit compositional descriptors of molecules in drug discovery and aid the medicinal chemist. *Journal of Medicinal Chemistry*. https ://doi.org/10.1021/acs.jmedchem.4c00649

[223] Lavecchia, A. (2019). Deep learning in drug discovery : opportunities, challenges and future prospects. *Drug Discovery Today*, 24(10), 2017-2032. https ://doi.org/10.1016/j.drudis.2019.07.006

[224] Iyer, V., Kaalia, R., & Nandi, S. (2024). Current strategies to address data scarcity in artificial intelligence-based drug discovery : A comprehensive review. *Computers in Biology and Medicine*, 167, 107960. https ://doi.org/10.1016/j.compbiomed.2024.107960

[225] Nandi, S., & Srivastava, V. (2022). Deep learning for low-data drug discovery : Hurdles and opportunities. *Current Opinion in Structural Biology*, 82, 102599. https ://doi.org/10.1016/j.sbi.2024.102599

[226] Blanco-González, A., Cabezón, A., Seco-González, A., Conde-Torres, D., Antelo-Riveiro, P., Piñeiro, Á., & Garcia-Fandino, R. (2023). The role of AI in drug discovery : Challenges, opportunities, and strategies. *Pharmaceuticals*, 16(6), 891. https ://doi.org/10.3390/ph16060891

[227] Wang, K., Zhao, D., Zhao, J., Dong, X., Sun, J., & Liu, X. (2024). Application of artificial intelligence in drug-target interactions prediction : A review. *npj Biomedical Innovation*, 1, 3. https ://doi.org/10.1038/s44385-024-00003-9

[228] Rahimi, F., Ghanadian, M., Mirzaei, A., Barani, M., & Hassanzadeh, K. (2024). Artificial Intelligence (AI) applications in drug discovery and drug delivery : Revolutionizing personalized medicine. *Pharmaceutics*, 16(10), 1328. https ://doi.org/10.3390/pharmaceutics16101328

[229] Pandey, P., Jiang, W., Zhang, Y., Zhang, Y., Kohlhoff, K. J., & Pande, V. S. (2022). The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4(3), 211-221. https ://doi.org/10.1038/s42256-022-00463-x

[230] Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, 26(1), 80-93. https ://doi.org/10.1016/j.drudis.2020.10.010

[231] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4), 688-702. https ://doi.org/10.1016/j.cell.2020.01.021

[232] Kausar, S., & Falcao, A. O. (2023). Advances in artificial intelligence for drug delivery and development : A comprehensive review. *International Journal of Pharmaceutics*, 641, 122998. https ://doi.org/10.1016/j.ijpharm.2023.122998

[233] Wang, J., Zhuo, L., Song, S., Ren, Y., & Wang, Y. (2023). Multitask joint strategies of self-supervised representation learning on biomedical networks for drug discovery. *Nature Machine Intelligence*, 5, 632-645. https ://doi.org/10.1038/s42256-023-00640-6

[234] Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2021). Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, 22(6), bbab159. https ://doi.org/10.1093/bib/bbab159

[235] He, Y. T., Johnson, S. M., & Scarselli, M. (2024). Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature Communications*, 15, 3793. https ://doi.org/10.1038/s41467-024-45566-8

[236] Rana, S., & Monnappa, A. K. (2024). Generative artificial intelligence in drug discovery : basic framework, recent advances, challenges, and opportunities. *Frontiers in Artificial Intelligence*, 7, 1370134. https ://doi.org/10.3389/frai.2024.1370134

[237] Jin, Y., Lei, T., & Tsui, K. (2023). Deep generative molecular design reshapes drug discovery. *Cell Reports Physical Science*, 4(1), 101242. https ://doi.org/10.1016/j.xcrp.2022.101242

[238] Mirhoseini, A., Pham, H., Le, Q. V., Steiner, B., Larsen, R., Zhou, Y., Kumar, N., Norouzi, M., Bengio, S., & Dean, J. (2019). Optimizing multi-GPU parallelization strategies for deep learning training. *arXiv preprint arXiv :1907.13257*. https ://arxiv.org/abs/1907.13257

[239] Son, H. (2023). Prediction of drug–target interactions through multi-task learning. *Scientific Reports*, 13, 18179. https ://doi.org/10.1038/s41598-023-45502-8

[240] Xu, M., Wang, W., & Luo, S. (2023). Generative AI for graph-based drug design : Recent advances and the way forward. *Current Opinion in Structural Biology*, 83, 102575. https ://doi.org/10.1016/j.sbi.2023.102575

[241] Qiao, H., Wang, X., Zhao, J., Cai, H., Zheng, X., & Mao, W. (2023). A drug molecular classification model based on graph structure generation. *Journal of Biomedical Informatics*, 139, 104303. https ://doi.org/10.1016/j.jbi.2023.104303

[242] Xu, Q., Kaindl, J., & Mandujano-Tinoco, E. (2023). From intuition to AI : evolution of small molecule representations in drug discovery. *Briefings in Bioinformatics*, 25(1), bbad422. https ://doi.org/10.1093/bib/bbad422

[243] Jin, W., Barzilay, R., & Jaakkola, T. (2022). Molecule generation for drug design : A graph learning perspective. *arXiv preprint arXiv :2202.09212*. https ://arxiv.org/abs/2202.09212

[244] Aspuru-Guzik, A. (2021). Molecular graph representations and SELFIES : A 100% robust molecular string representation. *Personal Blog*. https ://aspuru.substack.com/p/molecular-graph-representations-and

[245] Li, D., Hao, J., Hu, D., Li, Y., Zhang, Y., & Hu, S. (2023). A merged molecular representation learning for molecular properties prediction with a web-based service. *Scientific Reports*, 13, 8901. https ://doi.org/10.1038/s41598-023-35866-2

[246] Wang, Y., Yang, Z., Wang, D., Zhao, M., Li, Y., Guo, H., & Xiong, D. (2022). Drug discovery and mechanism prediction with explainable graph neural networks. *Scientific Reports*, 12, 19937. https ://doi.org/10.1038/s41598-022-24613-8

[247] Tanaka, K., Takayama, K., Yamanishi, Y., & Sugiyama, M. (2021). Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Scientific Reports*, 11, 525. https ://doi.org/10.1038/s41598-020-80113-7

[248] David, L., Thakkar, A., Mercado, R., & Engkvist, O. (2020). Molecular representations in AI-driven drug discovery : a review and practical guide. *Journal of Cheminformatics*, 12(1), 56. https ://doi.org/10.1186/s13321-020-00460-5

[249] Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery ? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1), 12. https ://doi.org/10.1186/s13321-020-00479-8

[250] Chen, J., Guo, Q., Zhou, J., & Liao, F. (2024). Knowledge mapping of graph neural networks for drug discovery : a bibliometric and visualized analysis. *Frontiers in Pharmacology*, 15, 1393415. https ://doi.org/10.3389/fphar.2024.1393415

[251] Ahmadipourirani, A., Nasiri, A., Nojabsadeghi, S., & Haikal, A. A. (2024). A review on graph neural networks for predicting synergistic drug combinations. *Artificial Intelligence Review*, 57, 8371-8410. https ://doi.org/10.1007/s10462-023-10669-z

[252] Zhang, J., & Fan, S. (2024). Role of artificial intelligence in revolutionizing drug discovery. *Precision Clinical Medicine*, pbae040. https ://doi.org/10.1093/pcmedi/pbae040

[253] Hassanin, R. T., El-Fouly, M. M., El-Zoghbi, H. Y., & Sakr, M. M. (2023). Deep learning in drug discovery : an integrative review and future challenges. *Artificial Intelligence Review*, 56, 3383-3424. https ://doi.org/10.1007/s10462-022-10306-1

[254] Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., Xing, L., Guo, T., & Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37, 1038-1040. https ://doi.org/10.1038/s41587-019-0224-x

[255] Cole, J. M. (2022). Can molecular modeling overcome the limitations of drug discovery AI ? *Drug Discovery Online*. https ://www.drugdiscoveryonline.com/doc/can-molecular-modeling-overcome-the-limitations-of-drug-discovery-ai-0001

[256] Choi, S., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2022). On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv :1910.05446v3*. https ://arxiv.org/abs/1910.05446

[257] Demissie, S., Tadesse, M., & Ibrahim, F. (2024). Harnessing the AI/ML in drug and biological products discovery and development : The regulatory perspective. *Pharmaceuticals*, 18(1), 47. https ://doi.org/10.3390/ph18010047

[258] Nabil, A. M., Ibrahim, S. A., Abdel-Naim, H. A., & Al-Maghrabee, M. (2024). Revolutionizing medicinal chemistry : The application of artificial intelligence (AI) in early drug discovery. *Pharmaceuticals*, 16(9), 1259. https ://doi.org/10.3390/ph16091259

[259] Mohideen, Y. A., Singh, M., Kamal, A., El-Kabbani, O., Meghwanshi, G. K., & Kesharwani, P. (2024). Explainability, transparency and black box challenges of AI in radiology : impact on patient care in cardiovascular radiology. *Egyptian Journal of Radiology and Nuclear Medicine*, 55, 119. https ://doi.org/10.1186/s43055-024-01356-2

[260] Front Line Genomics. (2024, April 4). AI in Drug Discovery 2024 : Where are we now ? https ://frontlinegenomics.com/ai-in-drug-discovery-2024-where-are-we-now/

[261] Chen, L., & Wang, Y. (2023). Geometric deep learning for drug discovery. *Expert Systems with Applications*, 235, 120564. https ://doi.org/10.1016/j.eswa.2023.120564

[262] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463–477. https ://doi.org/10.1038/s41573-019-0024-5

[263] Lin, E., Guo, X., Lu, C., Han, X., & Jin, R. (2023). A survey of generative AI for de novo drug design : new frontiers in molecule and protein generation. *Briefings in Bioinformatics*, 25(4), bbae338. https ://doi.org/10.1093/bib/bbae338

[264] Bian, Y., Wang, J., Xiao, J., & Lichtenthaeler, H. (2023). Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 82, 102572. https ://doi.org/10.1016/j.sbi.2023.102572

[265] Gupta, A., & Zou, J. (2024). Diffusion models in de novo drug design. *Journal of Chemical Information and Modeling*, 64(10), 2938-2956. https ://doi.org/10.1021/acs.jcim.4c01107

[266] Liu, K., Meng, Z., & Jin, X. (2024). Equivariant score-based generative diffusion framework for 3D molecules. *BMC Bioinformatics*, 25, 156. https ://doi.org/10.1186/s12859-024-05810-w

[267] Kurochkin, I., Khrulkov, V., & Oseledets, I. (2023). The coming of age of AI/ML in drug discovery, development, clinical testing, and manufacturing : The FDA perspectives. *Pharmaceuticals*, 16(9), 1284. https ://doi.org/10.3390/ph16091284

[268] Rana, S., & Monnappa, A. K. (2024). Generative artificial intelligence in drug discovery : basic framework, recent advances, challenges, and opportunities. *Frontiers in Artificial Intelligence*, 7, 1370134. https ://doi.org/10.3389/frai.2024.1370134

[269] Esteva, A., Choudhary, A., & Koller, D. (2023). Multimodal biomedical AI. *Nature Medicine*, 29, 806-817. https ://doi.org/10.1038/s41591-022-01981-2

[270] Sheikh, M., Chavan, S., & Vyas, R. (2023). Advances in artificial intelligence (AI)-assisted approaches in drug screening. *Current Research in Pharmacology and Drug Discovery*, 4, 100398. https ://doi.org/10.1016/j.crphar.2023.100398

[271] Chen, C., Chen, X., Morehead, A., Wu, T., & Cheng, J. (2023). 3D-equivariant graph neural networks for protein model quality assessment. *Bioinformatics*, 39(1), btad030. https ://doi.org/10.1093/bioinformatics/btad030

[272] Duan, Y., Alati, T., & Tang, J. (2023). Artificial intelligence in multi-objective drug design. *Current Opinion in Structural Biology*, 79, 102572. https ://doi.org/10.1016/j.sbi.2023.102518

[273] Crichton, G., Christopeit, M., & Rodriguez-Galindo, M. (2024). Data-driven federated learning in drug discovery with knowledge distillation. *Nature Machine Intelligence*, 6, 454-463. https ://doi.org/10.1038/s42256-025-00991-2

[274] Kadan, A., Godsi, O., & Popov, P. (2024). Guided multi-objective generative AI to enhance structure-based drug design. *arXiv preprint arXiv :2405.11785*. https ://arxiv.org/abs/2405.11785

[275] Karim, A., Singh, A., Fang, X., & Kinghorn, A. D. (2023). Artificial intelligence in natural product drug discovery : Current applications and future perspectives. *Journal of Medicinal Chemistry*, 66(19), 13356–13374. https ://doi.org/10.1021/acs.jmedchem.4c01257

[276] Sheridan, R. P., Williams, J. D., & Clark, R. D. (2023). Federated learning for molecular discovery. *Current Opinion in Structural Biology*, 79, 102489. https ://doi.org/10.1016/j.sbi.2023.102489

[277] Filella-Merce, I., & Gutierrez, B. (2023). Optimizing drug design by merging generative AI with active learning frameworks. *arXiv preprint arXiv :2305.06334*. https ://arxiv.org/abs/2305.06334

[278] Sun, X., & Wang, Z. (2023). De novo drug design by iterative multiobjective deep reinforcement learning with graph-based molecular quality assessment. *Bioinformatics*, 39(4), btad157. https ://doi.org/10.1093/bioinformatics/btad157

[279] Jiang, B., Li, C., Baek, M., & Kim, J. (2023). Graph neural networks and equivariant networks for 3D molecular generation : progress, challenges, and opportunities. *Journal of Medicinal Chemistry*, 66(15), 10324-10337. https ://doi.org/10.1021/acs.jmedchem.3c00306

[280] Rodriguez-Martinez, M., Gammie, T., & Di Pietro, A. (2022). Transforming drug discovery with machine learning and new commercially available quantum computers. *Expert Opinion on Drug Discovery*, 17(9), 873-886. https ://doi.org/10.1080/17460441.2022.2096728

[281] Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2, 573-584. https ://doi.org/10.1038/s42256-020-00236-4

[282] Skafte, F., Delaney, A., & Warner, S. L. (2024). Glass box and black box machine learning approaches to exploit compositional descriptors of molecules in drug discovery and aid the medicinal chemist. *Journal of Medicinal Chemistry*, 67(10), 7783-7798. https ://doi.org/10.1021/acs.jmedchem.4c00649

[283] Schärli, P., Gapany, R., Helmy, M., Simonetta, G., Bianchi, F. M., Magni, P., & Cesa-Bianchi, N. (2024). Shaping the future of healthcare : Ethical clinical challenges and pathways to trustworthy AI. *Journal of Personalized Medicine*, 14(3), 170. https ://doi.org/10.3390/jpm14030170

[284] Hoogeboom, E., Satorras, V. G., Vignac, C., & Welling, M. (2022). Equivariant diffusion for molecule generation in 3D. *International Conference on Machine Learning*, 8867-8887.

[285] Liu, J., Tang, G., & Baldi, P. (2022). Interpretable enhanced drug-target interaction prediction with molecule graph and protein structure. *Journal of Chemical Information and Modeling*, 62(22), 5499-5511. https ://doi.org/10.1021/acs.jcim.2c00935

[286] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., & others. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589. https ://doi.org/10.1038/s41586-021-03819-2

[287] Schneider, P., Walters, W.P., Plowright, A.T., Sieroka, N., Listgarten, J., Goodnow, R.A., Fisher, J., Jansen, J.M., Duca, J.S., Rush, T.S., & others. (2020). Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery*, 19(5), 353-364. https ://doi.org/10.1038/s41573-019-0050-3

[288] Yang, X., Wang, Y., Byrne, R., Schneider, G., & Yang, S. (2019). Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews*, 119(18), 10520-10594. https ://doi.org/10.1021/acs.chemrev.8b00728

[289] Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Science Advances*, 4(7), eaap7885. https ://doi.org/10.1126/sciadv.aap7885

[290] Cao, Y., Romero, J., & Aspuru-Guzik, A. (2023). A hybrid quantum computing pipeline for real world drug discovery. *Scientific Reports*, 13, 19376. https ://doi.org/10.1038/s41598-024-67897-8

[291] Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S. C., Endo, S., Fujii, K., McClean, J. R., Mitarai, K., Yuan, X., Cincio, L., & Coles, P. J. (2022). Variational quantum algorithms for drug discovery. *Nature Reviews Chemistry*, 6, 526-549. https ://doi.org/10.1038/s41570-022-00407-4

[292] Meyer, J. G., & Liu, S. (2024). Computer-aided drug discovery : From traditional simulation methods to language models and quantum computing. *Chemical Science*, 15(32), 8633-8645. https ://doi.org/10.1039/D4SC02584J

[293] Anisimov, V. M. (2021). Quantum-inspired algorithms in drug discovery and development. *Future Medicinal Chemistry*, 13(17), 1503-1513. https ://doi.org/10.4155/fmc-2021-0136

[294] Chaudhury, S., Madgwick, J. D., Roy, F., & Govind, N. (2023). Perspective on the current state-of-the-art of quantum computing for drug discovery applications. *Journal of Chemical Theory and Computation*, 19(3), 745-759. https ://doi.org/10.1021/acs.jctc.2c00574

[295] Yu, S., Jafke, J., & Preskill, D. (2023). Towards using quantum computing to speed up drug development. *Nature Computational Science*, 3, 1018-1030. https ://doi.org/10.1038/s43588-023-00472-x

[296] Pasqal Technologies. (2024). How quantum computing is changing molecular drug development. *World Economic Forum*. https ://www.weforum.org/stories/2025/01/quantum-computing-drug-development/

[297] Emani, P. S., Warrell, J., Anticevic, A., Bekiranov, S., Gandal, M., McConnell, M. J., Moore, G., Penhoet, E., Song, J., Utz, P., White, C., Huang, S. C., Rothberg, B., Katz, Y., & Gerstein, M. (2021). Quantum computing at the frontiers of biological sciences. *Nature Methods*, 18(7), 701-709. https ://doi.org/10.1038/s41592-021-01233-0

[298] Fingerhuth, M., Babej, T., & Wittek, P. (2022). The next revolution in computational simulations : Harnessing AI and quantum computing in molecular dynamics. *Current Opinion in Structural Biology*, 89, 102942. https ://doi.org/10.1016/j.sbi.2024.102942

[299] Mukherjee, R., & Wu, Q. (2024). Quantum computing in drug discovery : Quantum molecular simulations. *Quantum Zeitgeist*. https ://quantumzeitgeist.com/quantum-computing-in-drug-discovery-quantum-molecular-simulations/

[300] Vella, D., Nehmé, R., & Marcou, G. (2023). Few-shot learning for low-data drug discovery. *Journal of Chemical Information and Modeling*, 63(1), 27-42. https ://doi.org/10.1021/acs.jcim.2c00779

[301] Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Central Science*, 3(4), 283-293. https ://doi.org/10.1021/acscentsci.6b00367

[302] Vieira, M., Lemos, J., & Pinho, A. (2023). A systematic review of few-shot learning in medical imaging. *Computers in Biology and Medicine*, 169, 107914. https ://doi.org/10.1016/j.compbiomed.2024.107914

[303] DiCenzo, R. G., Ferracane, S., & Rajasekaran, P. (2024). Few-shot meta-learning applied to whole brain activity maps improves systems neuropharmacology and drug discovery. *Pharmacological Research*, 203, 107032. https ://doi.org/10.1016/j.phrs.2024.107032

[304] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks : A review. *Neural Networks*, 113, 54-71. https ://doi.org/10.1016/j.neunet.2019.01.012

[305] Li, Z., & Hoiem, D. (2018). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12), 2935-2947. https ://doi.org/10.1109/TPAMI.2017.2773081

[306] Stark, H., Gopalakrishnan, K., & Ahmadi, S. (2024). Deep learning for low-data drug discovery : Hurdles and opportunities. *Current Opinion in Structural Biology*, 84, 102599. https ://doi.org/10.1016/j.sbi.2024.102599

[307] Wang, S., Guo, Y., Wang, Y., Sun, H., & Huang, J. (2022). Self-supervised graph learning for recommendation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2666-2676. https ://doi.org/10.1145/3477495.3532077

[308] Magar, R., Xu, P., & Sarkar, S. (2024). Supervised machine learning in drug discovery and development : Algorithms, applications, challenges, and prospects. *Artificial Intelligence in Life Sciences*, 5, 100525. https ://doi.org/10.1016/j.ailsci.2024.100525

[309] Breger, F., Sutherland, J. J., & Barzilay, R. (2023). Integrating generative AI with expert knowledge for drug discovery. *Drug Discovery Today*, 28(9), 103556. https ://doi.org/10.1016/j.drudis.2023.103556

[310] Agrawal, P., Obayemi, J., & Chen, B. (2023). Exploring new horizons : Empowering computer-assisted drug design with few-shot learning. *Artificial Intelligence in Life Sciences*, 3, 100302. https ://doi.org/10.1016/j.ailsci.2023.100302

[311] Mohamad Zobir, S. Z., Dinesh, N. S., & Yusoff, Z. M. (2023). A cloud-based platform for collaborative drug discovery. *Digital Health*, 9, 1-12. https ://doi.org/10.1177/20552076231158988

[312] Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., & Wu, Z. (2019). Deep learning for the life sciences : Applying deep learning to genomics, microscopy, drug discovery, and more. O'Reilly Media.

[313] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015 : A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1), D1045-D1053. https ://doi.org/10.1093/nar/gkv1072

[314] He, X., Zhao, K., & Chu, X. (2021). AutoML : A survey of the state-of-the-art. *Knowledge-Based Systems*, 212, 106622. https ://doi.org/10.1016/j.knosys.2020.106622

[315] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning : Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1-19. https ://doi.org/10.1145/3298981

[316] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119. https ://doi.org/10.1038/s41746-020-00323-1

[317] Sheridan, R. P., Williams, J. D., & Clark, R. D. (2023). Federated learning for molecular discovery. *Current Opinion in Structural Biology*, 79, 102489. https ://doi.org/10.1016/j.sbi.2023.102489

[318] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets : Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv :1704.04861*.

[319] Kuchaiev, O., Ginsburg, B., Wu, J., Nguyen, P., Busche, M., & Mikayelyan, M. (2023). Low-resource deep learning for drug discovery on edge devices. *Frontiers in Drug Discovery*, 3, 1168599. https ://doi.org/10.3389/fddsv.2023.1168599

[320] Vatansever, S., Schlessinger, A., Wacker, D., Kaniskan, H. Ü., Jin, J., & Zhou, M. (2023). Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases : State-of-the-arts and future directions. Medicinal Research Reviews, 43(2), 786-823.

[321] Blanco-González, A., Cabezón, A., Seco-González, A., Conde-Torres, D., Antelo-Riveiro, P., & Piñeiro, Á. (2023). The Role of AI in Drug Discovery : Challenges, Opportunities, and Strategies. Pharmaceuticals, 16(6), 891.

[322] Pasqal & Qubit Pharmaceuticals. (2025). Quantum computing for drug development : Optimizing molecular interactions through quantum modeling. World Economic Forum.

[323] Madushanka, A., Laird, E., Clark, C., & Kraka, E. (2024). SmartCADD : AI-QM Empowered Drug Discovery Platform with Explainability. Journal of Chemical Information and Modeling, 64(17), 6799-6817.

[324] Madushanka, A., Laird, E., Clark, C., & Kraka, E. (2024). AI and quantum mechanics team up to accelerate drug discovery. ScienceDaily. Retrieved May 9, 2025.

[325] Singh, S., Kumar, R., Payra, S., & Singh, S. K. (2023). Artificial Intelligence and Machine Learning in Pharmacological Research : Bridging the Gap Between Data and Drug Discovery. Cureus, 15, e44359.

[326] FDA. (2025). Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. Draft Guidance for Industry, January 2025.

[327] World Economic Forum. (2025). How 2025 can be a pivotal year of progress for Biopharma. Retrieved May 9, 2025.

[328] Dhudum, R., Ganeshpurkar, A., & Pawar, A. (2024). Revolutionizing Drug Discovery : A Comprehensive Review of AI Applications. Drug Discovery and Computation, 3, 148-171.

[329] Vora, L. K., Gholap, A. D., Jetha, K., Thakur, R. R. S., Solanki, H. K., & Chavda, V. P. (2023). Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. Pharmaceutics, 15(7), 1916.

[330] Lloyd, S., & Weedbrook, C. (2018). Quantum Generative Adversarial Networks for learning and loading random distributions. Quantum AI : Harnessing Quantum Computing for AI. Post Quantum, October 31, 2024.

[331] Nishan, M. D. N. H. (2025). AI-powered drug discovery for neglected diseases : accelerating public health solutions in the developing world. Journal of Global Health, 15, 03002.

[332] The Pharmaceutical Journal. (2024). How AI is transforming drug discovery. July 3, 2024.

[333] Gangwal, A., Ansari, A., Ahmad, I., Azad, A. K., & Wan Sulaiman, W. M. A. (2024). Current strategies to address data scarcity in artificial intelligence-based drug discovery : A comprehensive review. Computers in Biology and Medicine, 179, 108734.

[334] Zinner, M., Dahlhausen, F., Boehme, P., Ehlers, J., Bieske, L., & Fehring, L. (2022). Toward the institutionalization of quantum computing in pharmaceutical research. Drug Discovery Today, 27(2), 378-383.

[335] Drug Discovery and Development. (2024). 2024 : The year AI drug discovery and protein structure prediction took center stage—2025 set to amplify growth. November 25, 2024.

[336] Recursion Pharmaceuticals. (2024). Industrial-scale drug discovery through AI and high-dimensional biology. The State of AI in Drug Discovery 2024. Genetic Engineering & Biotechnology News, December 13, 2024.

[337] FDA. (2025). FDA Proposes Framework to Advance Credibility of AI Models Used for Drug and Biological Product Submissions. FDA News Release, January 2025.

[338] MIT News. (2024). A smarter way to streamline drug discovery. June 17, 2024.

[339] Fromer, J., & Coley, C. (2024). SPARROW : An algorithmic framework to automatically identify optimal molecular candidates for drug discovery. Nature Computational Science, June 2024.

[340] Sifted. (2024). Drug discovery in 2024 : Poised for an AI revolution ? Startup Europe.

[341] Mohammadi, S., Balador, A., Sinaei, S., & Flammini, F. (2024). Balancing privacy and performance in federated learning : A systematic literature review on methods and metrics. Journal of Parallel and Distributed Computing, 192, 104918.

[342] World Economic Forum. (2024). Quantum computing, AI and drug discovery : 7 key guardrails. Centre for Trustworthy Technology, February 2024.

[343] The Pharmaceutical Journal. (2024). How AI is transforming drug discovery. July 3, 2024.

[344] FDA. (2025). Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. Federal Register, January 7, 2025.

[345] FDA. (2024). Artificial Intelligence for Drug Development. Center for Drug Evaluation and Research.

[346] Foley & Lardner LLP. (2025). AI Drug Development : FDA Releases Draft Guidance. January 15, 2025.

[347] STAT News. (2025). FDA's new guidance on AI in drug discovery centers the risk introduced by the technology. January 6, 2025.

[348] Harishbhai Tilala, M., Kumar Chenchala, P., Choppadandi, A., Kaur, J., Naguri, S., & Saoji, R. (2024). Ethical Considerations in the Use of Artificial Intelligence and Machine Learning in Health Care : A Comprehensive Review. BMC Medical Ethics, 25, 55.

[349] World Economic Forum. (2024). Quantum computing, AI and drug discovery : 7 key guardrails. February 2024.

[350] STAT News. (2025). FDA's new guidance on AI in drug discovery centers the risk introduced by the technology. January 6, 2025.

[351] Foley & Lardner LLP. (2024). AI in Drug Discovery : 2025 Outlook. December 18, 2024.

[352] The Pharmaceutical Journal. (2024). How AI is transforming drug discovery. July 3, 2024.

[353] Front Line Genomics. (2024). AI in Drug Discovery 2024 : Where are we now ? April 4, 2024.

[354] ASDEvents. (2024). AI in Drug Discovery 2024 Conference, March 11-12, 2024, London, United Kingdom.

[355] World Economic Forum. (2025). How we can future-proof AI in health with a focus on equity. April 2025.

[356] O'Neil, S., Taylor, S., & Sivasankaran, A. (2021). Data Equity to Advance Health and Health Equity in Low- and Middle-Income Countries : A Scoping Review. Digital Health, 7, 20552076211061922.

[357] Nishan, M. D. N. H. (2025). AI-powered drug discovery for neglected diseases : accelerating public health solutions in the developing world. Journal of Global Health, 15, 03002.

[358] Bhutta, Z. A., Sommerfeld, J., Lassi, Z. S., Salam, R. A., & Das, J. K. (2014). Global burden, distribution, and interventions for infectious diseases of poverty. Infectious Diseases of Poverty, 3, 21.

[359] Ibeneme, S., Karamagi, H., Muneene, D., Goswami, K., Chisaka, N., & Okeibunor, J. (2022). Strengthening Health Systems Using Innovative Digital Health Technologies in Africa. Frontiers in Digital Health, 4, 854339.

[360] Life. (2024). AI approaches to pandemic preparedness and response. 14, 233.

[361] Gangwal, A., Ansari, A., Ahmad, I., Azad, A. K., & Wan Sulaiman, W. M. A. (2024). Current strategies to address data scarcity in artificial intelligence-based drug discovery : A comprehensive review. Computers in Biology and Medicine, 179, 108734.

[362] Artificial Intelligence and Quantum Computing as the Next Pharma Disruptors. (2021). PubMed, PMC. Retrieved from National Library of Medicine.

[363] World Economic Forum. (2024). AI and quantum revolution will transform drug development.

[364] World Economic Forum. (2025). How quantum computing is changing molecular drug development. January 2025.

[365] World Economic Forum. (2025). How quantum computing is changing molecular drug development. January 2025.

[366] Winkler, D. A. (2022). The impact of machine learning on future tuberculosis drug discovery. Expert Opinion on Drug Discovery, 17, 925-927.

[367] Insilico Medicine. (2023). Study combines quantum computing and generative AI for drug discovery. May 19, 2023.

[368] World Economic Forum. (2025). How 2025 can be a pivotal year of progress for Biopharma. January 2025.

[369] SnoQap. (2024). The Role of Quantum Computing in Drug Discovery. August 10, 2024.

[370] Bielska, W., Jaszczyszyn, I., Dudzic, P., Janusz, B., Chomicz, D., Wrobel, S., ... & Krawczyk, K. (2025). Applying computational protein design to therapeutic antibody discovery–current state and perspectives. arXiv preprint arXiv :2503.00913.

[371] Deng, J., Yang, Z., Ojima, I., Samaras, D., & Wang, F. (2022). Artificial intelligence in drug discovery : applications and techniques. Briefings in Bioinformatics, 23(1), bbab430.

[372] Mehrjou, A., Soleymani, A., Jesson, A., Notin, P., Gal, Y., Bauer, S., & Schwab, P. (2021). Genedisco : A benchmark for experimental design in drug discovery. arXiv preprint arXiv :2110.11875.

[373] Schapin, N., Majewski, M., Varela-Rial, A., Arroniz, C., & De Fabritiis, G. (2023). Machine learning small molecule properties in drug discovery. Artificial Intelligence Chemistry, 1(2), 100020.

[374] Li, Linwei, et al. "An updated review on developing small molecule kinase inhibitors using computer-aided drug design approaches." International Journal of Molecular Sciences 24.18 (2023) : 13953.

[375] Sanna, Vanna, and Mario Sechi. "Therapeutic potential of targeted nanoparticles and perspective on nanotherapies." ACS Medicinal Chemistry Letters 11.6 (2020) : 1069-1073.

[376] Saqib, Uzma, et al. "The fate of drug discovery in academia ; dumping in the publication landfill ?." Oncotarget 15 (2024) : 31.

[377] Fu, Ying, et al. "Combination of virtual screening protocol by in silico toward the discovery of novel 4-hydroxyphenylpyruvate dioxygenase inhibitors." Frontiers in Chemistry 6 (2018) : 14.

[378] Aldewachi, Hasan, et al. "High-throughput screening platforms in the discovery of novel drugs for neurodegenerative diseases." Bioengineering 8.2 (2021) : 30.

[379] Wang, Beilei, et al. "An overview of kinase downregulators and recent advances in discovery approaches." Signal Transduction and Targeted Therapy 6.1 (2021) : 423.

[380] Ferreira, Leonardo G., et al. "Molecular docking and structure-based drug design strategies." Molecules 20.7 (2015) : 13384-13421.

[381] Li, Qingxin. "Application of fragment-based drug discovery to versatile targets." Frontiers in molecular biosciences 7 (2020) : 180.

[382] Zhang, Yue, et al. "Application of computational biology and artificial intelligence in drug design." International journal of molecular sciences 23.21 (2022) : 13568.

[383] Fellmann, C., et al., Cornerstones of CRISPR-Cas in drug development and therapy. Nat Rev Drug Discov (2017) 16(2) :89-100. DOI : 10.1038/nrd.2016.238

[384] Gallego, Víctor, et al. "AI in drug development : a multidisciplinary perspective." Molecular Diversity 25 (2021) : 1461-1479.

[385] Mamoshina, P., Vieira, A., Putin, E., & Zhavoronkov, A. (2016). Applications of Deep Learning in Biomedicine. Molecular pharmaceutics, 13(5), 1445–1454. https ://doi.org/10.1021/acs.molpharmaceut.5b00982

[386] Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. Drug discovery today, 26(1), 80–93. https ://doi.org/10.1016/j.drudis.2020.10.010

[387] Koutroumpa, N. M., Papavasileiou, K. D., Papadiamantis, A. G., Melagraki, G., & Afantitis, A. (2023). A Systematic Review of Deep Learning Methodologies Used in the Drug Discovery Process with Emphasis on In Vivo Validation. International journal of molecular sciences, 24(7), 6573. https ://doi.org/10.3390/ijms24076573

[388] Bian, Y., & Xie, X.S. (2020). Generative chemistry : drug discovery with deep learning generative models. Journal of Molecular Modeling, 27.

[389] Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. Science advances, 4(7), eaap7885. https ://doi.org/10.1126/sciadv.aap7885

[390] Schneider G. (2018). Automating drug discovery. Nature reviews. Drug discovery, 17(2), 97–113. https ://doi.org/10.1038/nrd.2017.232

[391] Li, L., Zeng, L., Gao, Z., Yuan, S., Bian, Y., Wu, B., ... & Heng, P. A. (2022). Imdrug : A benchmark for deep imbalanced learning in ai-aided drug discovery. arXiv preprint arXiv :2209.07921.

[392] Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. Journal of cheminformatics, 9, 1-14.

[393] Lu, X., Xie, L., Xu, L., Mao, R., Chang, S., & Xu, X. (2023). Integrating Chemical Language and Molecular Graph in Multimodal Fused Deep Learning for Drug Property Prediction. arXiv preprint arXiv :2312.17495.

[394] Preuer, K., Klambauer, G., Rippmann, F., Hochreiter, S., & Unterthiner, T. (2019). Interpretable deep learning in drug discovery. Explainable AI : interpreting, explaining and visualizing deep learning, 331-345.

[395] Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. Science advances, 4(7), eaap7885.

[396] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature reviews. Drug discovery, 18(6), 463–477. https ://doi.org/10.1038/s41573-019-0024-5

[397] Dobbelaere, M.R., Lengyel, I., Stevens, C.V. et al. Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. J Cheminform 16, 99 (2024). https ://doi.org/10.1186/s13321-024-00895-0

[398] Kroll, A., Ranjan, S., & Lercher, M. J. (2024). A multimodal Transformer Network for protein-small molecule interactions enhances predictions of kinase inhibition and enzyme-substrate relationships. PLoS computational biology, 20(5), e1012100. https ://doi.org/10.1371/journal.pcbi.1012100

[399] Zhu, J., Xia, Y., Wu, L., Xie, S., Qin, T., Zhou, W., ... & Liu, T. Y. (2022, August). Unified 2d and 3d pre-training of molecular representations. In Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining (pp. 2626-2636).

[400] Recanatini, M., & Cabrelle, C. (2020). Drug Research Meets Network Science : Where Are We ?. Journal of medicinal chemistry, 63(16), 8653–8666. https ://doi.org/10.1021/acs.jmedchem.9b01989

[401] Dana, D., Gadhiya, S. V., St Surin, L. G., Li, D., Naaz, F., Ali, Q., Paka, L., Yamin, M. A., Narayan, M., Goldberg, I. D., & Narayan, P. (2018). Deep Learning in Drug Discovery and Medicine ; Scratching the Surface. Molecules (Basel, Switzerland), 23(9), 2384. https ://doi.org/10.3390/molecules23092384

[402] Barrett, J.S., Goyal, R.K., Gobburu, J. et al. An AI Approach to Generating MIDD Assets Across the Drug Development Continuum. AAPS J 25, 70 (2023). https ://doi.org/10.1208/s12248-023-00838-x

Résume :

La découverte d'un médicament fait face à des défis importants, puisque seulement 10% des composés utilisés dans les essais cliniques obtiennent l'approbation réglementaire. Cette thèse intègre la prédiction des propriétés moléculaires (MPP) et la génération moléculaire (MG) en utilisant l'apprentissage profond pour transformer le pipeline de découverte de médicaments.
Nous développons deux nouveaux approches basés sur des graphes : D'abord, le Graph Molecular Property Prediction Neural Network (GMPP-NN) combine les réseaux neuronaux de passage de messages avec des classificateurs perceptron multicouches, démontrant une performance supérieure sur les ensembles de données de référence MoleculeNet (VIH, BACE, BBBP, ClinTox).
Deuxièmement, le cadre ME &PP-MG &RC-DL intègre l'encodage moléculaire, la prédiction des propriétés, la génération et la classification de la réalité, permettant d'obtenir une précision exceptionnelle dans la prévision des propriétés chimiques quantiques tout en générant des structures moléculaires valides, uniques et nouvelles.
Nos approches organisent efficacement les représentations spatiales latentes, facilitant l'exploration chimique ciblée de l'espace et fournissant des informations sur les relations structure-propriété.
L'intégration du MPP et de la MG représente un changement de paradigme dans la découverte de médicaments par ordinateur, offrant des outils pour naviguer dans l'espace chimique et identifier les candidats thérapeutiques avec une efficacité sans précédent, accélérant potentiellement la découverte de médicaments et réduisant les coûts de développement.

Mots-clefs (5) : Découverte de médicaments, intelligence artificielle, apprentissage profond, réseaux neuronaux graphiques, prédiction des propriétés moléculaires, génération moléculaire, réseaux neuronaux de messages passant, encodeur automatique variationnel, chimie computationnelle, recherche pharmaceutique

Abstract :

Drug discovery faces significant challenges, with only 10% of the compounds used in clinical trials receiving regulatory approval. This thesis integrates molecular property prediction (MPP) and molecular generation (MG) by using deep learning to transform the drug discovery pipeline.
We are developing two new frameworks based on graphs: First, the Graph Molecular Property Prediction Neural Network (GMPP-NN) combines neural networks with multilayer perceptron classifiers, demonstrating superior performance on MoleculeNet reference datasets (HIV, BACE, BBBP, ClinTox). Secondly, the ME &PP-MG &RC-DL framework integrates molecular encoding, property prediction, reality generation and classification, providing exceptional precision in predicting quantum chemical properties while generating valid, unique and novel molecular structures.
Our approaches effectively organize latent spatial representations, facilitating targeted chemical exploration of space and providing information on structure-ownership relationships. The integration of MPP and MG represents a paradigm shift in drug discovery by computer, offering tools to navigate chemical space and identify therapeutic candidates with unprecedented efficiency, Potentially accelerating drug discovery and reducing development costs.

Keywords (5) : Drug discovery, artificial intelligence, deep learning, graph neural networks, molecular property prediction, de novo molecular generation, message-passing neural networks, variational autoencoders, computational chemistry, pharmaceutical research