

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique

École Normale Supérieure de l'Enseignement Technologique de
Skikda



Département de : Physique et chimie

THÈSE

Présentée en vue de l'obtention du diplôme de

Doctorat

Domaine : Sciences de la matière

Filière : Chimie

Spécialité : Chimie théorique et computationnelle

Présentée par

M^r : Mohamed Oussama MOUSSER

Thème :

**Conception et identification de molécules inhibitrices des
kinases : voie de développement de médicaments anti-cancéreux**

Soutenue publiquement le : 13/05/2026

devant le jury composé de :

Nom	Grade	Établissement	Qualité
Mr. Laidi BABOURI	Professeur	ENSET de Skikda	Président
Mr. Mohamed Abdesselem DEMS	Directeur de Recherche	CRBt de Constantine	Directeur de thèse
Mr. Djamel BOUDJAADAR	MCA	ENSET de Skikda	Examineur
Mr. Samir MESKALDJI	MCA	ENSET de Skikda	Examineur
Mme. Amel ALOUI	MCA	ENSET de Skikda	Examinatrice
Mr. Khalid BOUHEDJAR	MRA	CRBt de Constantine	Examineur

Année Universitaire : 2025/2026

Remerciements

Ce travail de thèse a été réalisé au Laboratoire de Chimie Physique et Biologie des Matériaux (LCPBM) de l'École Normale Supérieure de l'Enseignement Technologique (ENSET) de Skikda, dont je tiens à remercier la direction, ainsi que l'ensemble du personnel administratif et pédagogique.

*Je tiens tout d'abord à exprimer ma profonde gratitude à Monsieur **Mohamed Abdesselem DEMS**, Directeur de Recherche au Centre de Recherches en Biotechnologie (CRBt) de Constantine, pour avoir accepté de diriger cette thèse. Ses conseils avisés, sa disponibilité, sa rigueur scientifique et son accompagnement constant ont été déterminants pour la réalisation et l'aboutissement de ce travail.*

*J'adresse mes sincères remerciements à Monsieur **Laidi BABOURI**, Professeur à l'ENSET de Skikda, pour l'honneur qu'il me fait en acceptant de présider le jury de cette thèse, ainsi que pour l'intérêt porté à ce travail.*

*Mes remerciements vont également à Monsieur **Djamel BOUDJAADAR**, Monsieur **Samir MESKALDI** et Madame **Amel ALOUI**, Maîtres de Conférences A à l'ENSET de Skikda, ainsi qu'à Monsieur **Khalid BOUHEDJAR**, Maître de Recherche A au CRBt de Constantine, pour avoir accepté de faire partie du jury et pour l'attention qu'ils ont accordée à l'évaluation de ce travail.*

*Je tiens également à exprimer ma vive reconnaissance à Monsieur **Brahim MATOUGUI**, Maître de Recherche B au CRBt de Constantine, pour l'aide précieuse qu'il m'a apportée tout au long de la réalisation de mon article de thèse. Son engagement, ses conseils pertinents, sa disponibilité et ses remarques constructives ont grandement contribué à la qualité scientifique de mon article et ont été d'un apport déterminant dans sa finalisation.*

*Je souhaite adresser des remerciements tout particuliers à Professeure **Francesca INGROSSO**, de l'Université de Lorraine (France), pour m'avoir chaleureusement accueilli en stage au sein de son équipe. Je la remercie sincèrement pour l'encadrement scientifique de grande qualité qu'elle m'a apporté, notamment dans le domaine de la dynamique moléculaire, ainsi que pour sa disponibilité, ses conseils précieux et sa bienveillance.*

Cette recherche a également bénéficié de l'appui, d'une part, du Laboratoire de Physique et Chimie Théoriques (LPCT) de l'Université de Lorraine (France), dont je remercie chaleureusement l'ensemble de l'équipe, doctorants, chercheurs et personnels administratifs, pour leur bienveillance ainsi que pour les précieux souvenirs que je garde de mes périodes de stage,

et, d'autre part, du CRBt de Constantine, dont les contributions ont permis l'aboutissement de certains aspects de ce travail. À cette occasion, j'exprime tout particulièrement ma gratitude aux membres de l'équipe simulation, modélisation et gestion des données du CRBt de Constantine pour leur accueil chaleureux et la sympathie dont ils ont fait preuve.

*Je remercie Docteur **Adel KRID** pour ses conseils précieux en docking moléculaire, ainsi que pour son soutien moral.*

Enfin, je tiens à remercier toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de cette thèse, tant sur le plan scientifique que moral.

Abstract

ملخص

من بين الاستراتيجيات الحديثة لعلاج السرطان، تلعب العلاجات الموجهة دوراً محورياً، إذ تهدف إلى إيقاف تطور الأورام من خلال تثبيط بروتينات الكيناز ذات النشاط الغير منتظم و التي يرتبط فرط تعبيرها مباشرة بتكاثر الخلايا السرطانية. في هذه الأطروحة، اعتمدنا على عدة طرق حاسوبية لتعزيز الدقة التنبؤية في نمذجة العلاقات الكمية بين البنية والنشاط (*QSARs*) المرتبطة بتثبيط الكيناز، وقد تم اقتراح منهجية جديدة تعتمد على بنية هجينة تجمع بين خوارزمية التعلم الآلي *XGBoost* والشبكات العصبية العميقة (*DNN*) بهدف تحسين جودة التنبؤ. كما تم استخدام تقنيات الفرز الافتراضي (*LBVS*) إلى جانب أدوات دقيقة في تصميم الأدوية المعتمد على بنية الإنزيم (*SBDD*) للتعرف على جزيئات جديدة قد تمتلك قدرة تثبيطية للكينازات، وتبين أن هذه المركبات تحمل احتمالاً لنشاط يتراوح بين المعتدل إلى العالي وفقاً لتحليلات الالتحام الجزيئي ومحاكاة الديناميك الجزيئية، وتدل نتائج هذا العمل من جهة على تقدم مهم في مجال النمذجة النظرية، ومن جهة أخرى، تكشف عن إمكانيات مضادة للسرطان لدى بعض الجزيئات المكتشفة.

الكلمات المفتاحية: *XGBoost* ، *DNN* ، IC_{50} ، *CADD* ، الالتحام الجزيئي، الديناميك الجزيئي.

Abstract

Among recent cancer treatment strategies, targeted therapies play a central role. These aim to halt tumor progression by inhibiting dysregulated protein kinases whose overexpression is directly linked to cancer cell proliferation. In this thesis work, we employed several computational methods to improve predictive accuracy in modeling Quantitative Structure–Activity Relationships (QSARs) related to kinase inhibition. A new approach based on a hybrid eXtreme Gradient Boosting (XGBoost) and Deep Neural Network (DNN) architecture was introduced to enhance QSAR prediction accuracy. Furthermore, we used Ligand-Based Virtual Screening

(LBVS) and rigorous Structure-Based Drug Design (SBDD) tools to identify novel, potentially kinase-inhibiting molecules. These compounds show promising moderate to high potency based on molecular docking and molecular dynamics simulation analyses. Our results suggest, on one hand, a significant advance in theoretical modeling, and on the other hand, reveal the anticancer potential of certain molecules.

Keywords : XGBoost, DNN, IC_{50} , CADD, Molecular docking, Molecular dynamics.

Résumé

Parmi les stratégies récentes de traitement du cancer, les thérapies ciblées jouent un rôle central. Elles visent à stopper la progression tumorale en inhibant des protéine-kinases dérégulées, dont la surexpression est directement liée à la prolifération des cellules cancéreuses. Dans ce travail de thèse, nous avons utilisé plusieurs méthodes computationnelles pour améliorer la précision prédictive dans la modélisation des relations quantitatives structure-activité (QSARs) liées à l'inhibition des kinases. Une nouvelle approche basée sur une architecture hybride combinant l'algorithme d'apprentissage automatique eXtreme Gradient Boosting (XGBoost) aux réseaux neuronaux profonds (DNN) a été introduite afin d'améliorer la qualité des prédictions QSAR. De plus, nous avons utilisé le criblage virtuel basé sur les ligands (LBVS) ainsi que des outils rigoureux de conception de médicaments basée sur la structure (SBDD) pour identifier de nouvelles molécules potentiellement inhibitrices de kinases. Ces composés présentent un potentiel d'activité modérée à élevée selon les analyses de docking moléculaire et les simulations de dynamique moléculaire. Nos résultats suggèrent, d'une part, une avancée majeure dans la modélisation théorique et, d'autre part, révèlent le potentiel anticancéreux de certaines molécules.

Mots-clés : XGBoost, DNN, IC_{50} , CADD, Docking moléculaire, Dynamique moléculaire.

Table des matières

Abréviations	XII
Liste des figures	XIV
Liste des tableaux	XV
Liste des symboles	XVII
Introduction Générale	1
Partie I : Contexte et méthodes	4
1 Les Proteines Kinases	5
1.1 Introduction	5
1.2 Définition et mécanisme d'action des protéines kinases	5
1.2.1 Qu'est-ce qu'une protéine kinase ?	5
1.2.2 Le processus de phosphorylation : Un interrupteur moléculaire	6
1.2.3 La réversibilité : Rôle des phosphatases	6
1.2.4 Fonctions physiologiques des protéines kinases	9
1.2.4.1 Transduction du signal cellulaire	9
1.2.4.2 Régulation de la croissance et différenciation cellulaire	9
1.2.4.3 Rôle dans le métabolisme et l'homéostasie	9
1.3 Role dans le cancer	10
1.3.0.1 Dérégulation des kinases : Oncogènes et suppresseurs de tumeurs	10
1.3.0.2 Mécanismes d'activation oncogénique des kinases	11
1.3.0.3 Exemples de kinases dérégulées et leur impact en cancérogenèse	12
1.3.1 L'importance cruciale de la découverte de nouveaux inhibiteurs de kinases	15
1.3.1.1 Impact clinique et défis actuels	15
1.3.1.2 Perspectives et stratégies futures	16
1.4 Conclusion	19
2 Méthodes Computationnelles de Conception des Médicaments	20
2.1 Introduction	20
2.2 Approches basées sur le ligand (LBDD)	20
2.2.1 Modélisation par groupements pharmacophoriques	21
2.2.1.0.1 Approche basée sur les fragments	21
2.2.2 Relations Quantitatives Structure-Activité (QSARs)	21

2.2.2.1	Modelisation classique	21
2.2.2.1.1	Principales méthodes utilisées	22
2.2.2.1.2	Critères de Validation de Golbraikh et Tropsha . . .	23
2.2.2.1.3	Domaine d'applicabilité (Diagramme de Williams) . .	23
2.2.2.2	Machine learning (supervisé)	24
2.2.2.2.1	Regressions	24
2.2.2.2.2	Classifications	25
2.2.3	Apprentissage non-supervisé	26
2.2.3.0.1	Réduction de la Dimensionnalité	26
2.2.3.0.2	Regroupement (Clustering)	26
2.3	Approches basées sur la structure du récepteur (SBDD)	27
2.3.1	Docking moléculaire	27
2.3.1.1	Fonctions de score en docking moléculaire	28
2.3.1.1.1	Fonctions de score basées sur les champs de force .	28
2.3.1.1.2	Fonctions de score empiriques	29
2.3.1.1.3	Fonctions de score basées sur la connaissance . . .	29
2.3.1.1.4	Fonctions de score basées sur l'apprentissage automatique	30
2.3.1.1.5	Défis et considérations générales pour les fonctions de score	30
2.3.1.2	Algorithmes de recherche en docking moléculaire	32
2.3.1.2.1	Vue d'ensemble des stratégies de recherche	32
2.3.1.2.2	Algorithmes de docking rigide	32
2.3.1.2.3	Algorithmes de docking flexible	33
2.3.1.3	Logiciels et plateformes reconnus	34
2.3.2	Simulations de dynamique moléculaire	36
2.3.2.1	Applications clés en biologie structurale et en conception de médicaments	37
2.3.2.2	Principes fondamentaux d'une simulation de DM	38
2.3.2.3	Champs de force couramment utilisés en DM protéine-ligand	39
2.3.2.3.1	Concept et rôle des champs de force	39
2.3.2.3.2	Champs de force additifs majeurs	39
2.3.2.3.3	Famille AMBER	39
2.3.2.3.4	Famille CHARMM	40
2.3.2.3.5	Famille OPLS	41
2.3.2.3.6	Champs de force polarisables (Drude, AMOEBA) .	41
2.3.2.3.6.1	AMOEBA Force Field	42
2.3.2.4	Critères de sélection et précision des champs de force	42
2.3.2.5	Logiciels pour simulation de dynamique moléculaire protéine-ligand	45
2.3.2.5.1	Principaux logiciels de simulation	45
2.3.2.6	Outils spécialisés et plateformes intégrées	47
2.3.2.7	Analyse des trajectoires et calculs d'énergie libre	48
2.3.2.7.1	Techniques d'analyse courantes	48
2.3.2.7.2	Calculs d'énergie libre avancés	49
2.4	Conclusion	49

3	<i>Modelisation des Relations Structure-Activité des Inhibiteurs de Kinases</i>	53
3.1	Introduction	53
3.2	Méthodologie	56
3.2.1	Bases de données	56
3.2.2	Traitement mathématique et mise à l'échelle	56
3.2.3	Intégration de XGBoost avec les Reseaux de Neurones Profonds (DNN)	57
3.2.3.1	XGBoost	58
3.2.3.2	Ingénierie des variables	59
3.2.3.3	Réseaux de Neurones Profonds (DNNs)	60
3.2.4	Métriques d'évaluation	61
3.3	Resultats et discussion	62
3.3.1	Analyse et prétraitement des données	62
3.3.2	Modélisation	63
3.3.2.1	Hyperparamètres et benchmark	63
3.3.2.2	Hyperparamètres du DNN	65
3.3.3	Comparaison des performances des modèles hybrides XGBoost/DNN avec les modèles XGBoost seuls	65
3.3.3.1	Comparaison des performances de XGBoost/DNN avec d'autres algorithmes standards	68
3.3.3.2	Capacité prédictive des modèles	69
3.3.3.3	Analyse de l'importance des caractéristiques	73
3.3.3.4	Analyse de l'amélioration au niveau structural	75
3.4	Conclusion	77
4	<i>Conception et Identification d'Inhibiteurs de ERK2</i>	78
4.1	Introduction	78
4.2	Méthodologie	80
4.2.1	Criblage basé sur le docking et clustering	80
4.2.2	Simulations de dynamique moléculaire	82
4.2.3	ADMET	83
4.3	Resultats et discussion	83
4.3.1	Analyse des données et apprentissage automatique	83
4.3.2	Criblage docking et clustering	85
4.3.2.1	Analyse de docking	90
4.3.2.2	Dynamique moléculaire	93
4.3.3	ADME-Tox	95
4.4	Conclusion	98
	Conclusions et perspectives	100
	<i>Bibliographie</i>	120

Abréviations

ADMET : Absorption, Distribution, Métabolisme, Excrétion et Toxicité

ADP : Adenosine Diphosphate

AGC : Protein Kinase A, G and C family

AMBER : Assisted Model Building with Energy Refinement

AMOEBA : Atomic Multipole Optimized Energetics for Biomolecular Applications (champ de force polarisable)

ARN : Acide Ribonucleique

ATP : Adenosine Triphosphate

AUC : Area Under the Curve

BCR-ABL1 : Protéine kinase de fusion impliquée dans la LMC

BRAF : B-Rapidly Accelerated Fibrosarcoma (ou B-type Raf kinase)

CADD : Conception de médicaments assistée par ordinateur (Computer-Aided Drug Design)

CaMK : Calcium/Calmodulin-dependent Protein Kinase

CHARMM : Chemistry at HARvard Macromolecular Mechanics

CK1 α : Caséine Kinase 1 alpha

CK1/CK2 : Casein Kinase 1 / Casein Kinase 2

CMGC : CDK/MAPK/GSK3/CLK family

CREB : cAMP Response Element-Binding protein

DAG : Diacylglycérol

DBSCAN : Density-Based Spatial Clustering of Applications with Noise

DM : Dynamique Moléculaire

DNN/DNNs : Réseaux de Neurones Profonds (Deep Neural Networks)

DSK : kinases à double spécificité

EC50 : Concentration Efficace à 50%

FAK : Focal Adhesion Kinase

FDA : Food and drug administration

FN : Faux Négatif (False Negative)

FP : Faux Positif (False Positive)

GA : Algorithmes Génétiques

GIST : Tumeurs Stromales Gastro-Intestinales (Gastrointestinal Stromal Tumor)

GROMACS : GRoningen MACHine for Chemical Simulations

GSK3 : Glycogen Synthase Kinase 3

HER : Human Epidermal growth factor Receptor

IC50 : Concentration Inhibitrice Médiane

IK : Inhibiteurs de kinases

ITK : Inhibiteurs de tyrosine kinases

k-NN : k-Nearest-Neighbors

LBDD(VS) : Ligand-Based Drug Design (Virtual Screening)

LMC : Leucémie Myéloïde Chronique

LOF : Perte de fonction (Loss Of Function)

MAPK : Mitogen-Activated Protein Kinase

MC : Matrice de confusion

MKK4 : Mitogen-Activated Protein Kinase Kinase 4

ML : Machine Learning (Apprentissage Automatique)

MLR : Régression Linéaire Multiple (Multiple Linear Regression)

NAMD : Nanoscale Molecular Dynamics (logiciel de simulation)

NSCLC : Cancer du poumon non à petites cellules (Non-Small Cell Lung Cancer)

OPLS : Optimized Potentials for Liquid Simulations (champ de force)

PCA : Analyse en Composantes Principales (Principal Component Analysis)

PCR : Regression sur Composantes Principales (Principal Component Regression)

PDB : Protein Data Bank

PKA : Protéine Kinase A

PKB : Protéine Kinase B

PKC : Protéine Kinase C

PLS : Régression par Moindres Carrés Partiels (Partial Least Squares)

QSAR : Quantitative Structure-Activity Relationship

QSPR : Quantitative Structure-Property Relationship

Raf : Rapidly Accelerated Fibrosarcoma kinase

RF : Forêts Aléatoires (Random Forests)

RMSD : Root Mean Square Deviation

RMSE : Root Mean Square Error

ROC : Receiver Operating Characteristic

RTK : Récepteurs Tyrosine Kinases

SHAP : SHapley Additive ExPlanations

STE : Sterile kinase family

STK : Sérine/Thréonine Kinases

SVM : Machines à Vecteurs de Support

TK : Tyrosine Kinases

TKL : Tyrosine Kinase-Like

TN : Vrai Négatif (True Negative)

TP : Vrai Positif (True Positive)

VDW : van der Waals

VEGFR : Vascular Endothelial Growth Factor Receptor

XAI : Explainable Artificial Intelligence

Liste des figures

1.1	L'activation et l'inactivation normales des PTK sont essentielles pour maintenir une fonction cellulaire normale. Les mutations d'activation des PTK comprennent des mutations activatrices (gain de fonction), l'amplification génomique et la surexpression, les réarrangements chromosomiques (fusions géniques), ainsi que les boucles autocrines/paracrines de ligands.	12
3.1	Architecture du DNN.	60
3.2	Architecture pratique du modèle hybride proposé.	61
3.3	Graphique des valeurs de IC_{50} selon l'année de publication des essais dans le jeu de données du récepteur EGF.	63
3.4	Nombre de modèles selon leur précision.	67
3.5	Courbe de densité estimée pour la différence de score entre les ensembles d'apprentissage et de test, pour les modèles XGBoost seuls et hybrides.	68
3.6	Matrices de confusion des modèles EGFR et mTOR.	69
3.7	Matrices de confusion des modèles FLT3 et VEGFR2.	71
3.8	Courbes ROC pour les modèles du récepteur du facteur de croissance analogue à l'insuline de type 1 (IGF1R).	72
3.9	Importance des caractéristiques SHAP pour le modèle XGBoost/DNN appliqué aux données EGFR.	74
3.10	Comparaison de la calibration et du score de Brier entre les modèles XGBoost et XGBoost/DNN appliqués aux données EGFR.	75
3.11	Structures chimiques de dérivés pyrazolopyrimidines (inhibiteurs de mTOR) présentant des "activity-cliffs". L'étiquette d'activité correspond uniquement au seuil utilisé dans la présente étude.	76
4.1	Graphique des valeurs de IC_{50} selon l'année de publication de l'essai dans l'ensemble de données sélectionné.	83
4.2	Carte thermique de la matrice de confusion du classificateur et courbe ROC. . .	85
4.3	Superposition des poses DRX et redockées.	86
4.4	Structures chimiques et identifiants des meilleurs représentants des 5 clusters, leurs scores de docking et leur taux de similarité avec le ligand de référence. . .	88
4.5	Cartes de similarité chimique des 5 composés sélectionnés avec le ligand de référence SCH772984 (38Z). Méthode de coloration : vert : différence positive, gris : aucune variation de similarité, rose : différence négative.	89
4.6	Vues 3D des interactions avec la pose dockée des cinq composés (A-ZN) (l'unité de distance est l'Å).	92
4.7	Écarts quadratiques moyens calculés au cours du temps de simulation. À gauche, les valeurs de RMSD sont rapportées pour le squelette protéique (a), tandis que celles calculées pour les ligands sélectionnés sont présentées à droite (b). . . .	93

4.8	Capture de la disposition des composés (B), (D), (ZN) et Ulixertinib dans la poche de liaison de la protéine après 50 ns de simulation (l'unité de distance est Å).	95
4.9	Les figures radar de biodisponibilité (la zone rose représente l'intervalle optimal pour chaque propriété). LIPO : XLOGP3 entre -0.7 et +5.0, SIZE : poids moléculaire entre 150 et 500 g/mol, POLAR : SPT entre 20 et 130 Å ² , INSOLU : logS ne dépassant pas 6, INSATU : fraction de carbones en hybridation sp ³ supérieur ou égale à 0.25, et FLEX : pas plus de 9 liaison rotatives.	97
4.10	Diagramme pharmacocinétique BOILED-Egg pour les composés sélectionnés.	98

Liste des tableaux

1.1	Classification des protéines kinases.	8
1.2	Kinases dérégulées dans le cancer et leurs mécanismes oncogéniques.	14
1.3	Exemples d'inhibiteurs de kinases approuvés en phase clinique et leurs indications.	18
2.1	Types de fonctions de score en docking moléculaire.	31
2.2	Comparaison des champs de force couramment utilisés en DM protéine-ligand.	44
3.1	Tableau récapitulatif des informations pertinentes sur les modèles construits pour benchmark.	64
3.2	Description et plages d'expérimentation pour chaque hyperparamètre.	64
3.3	Paramètres du DNN.	65
3.4	Détails pertinents sur les modèles réalisés (XGB, Dim et VC désignent respectivement : XGBoost, Dimension et Validation Croisée).	66
3.5	Valeurs de précision de RF, SVM, DNN et du modèle hybride XGBoost/DNN pour divers jeux de données couvrant un éventail illustratif de tailles.	68
3.6	Métriques d'évaluation pertinentes pour certains des modèles de protéines les plus oncogènes (Sensit et Spec désignent respectivement la sensibilité et la spécificité, tandis que XGB désigne XGBoost).	72
3.7	Cinq principales caractéristiques SHAP pour le modèle XGBoost/DNN appliqué aux données EGFR.	74
4.1	Informations pertinentes du modèle.	84
4.2	Rapport du criblage virtuel par modèle QSAR.	85
4.3	Identifiants des composés dans chaque cluster, leur affinité de liaison (prédite en kcal/mol) et les plages de scores dans chaque groupe (le représentant de chaque groupe est en rouge).	87
4.4	Pourcentage de similarité des composés sélectionnés avec les inhibiteurs connus de ERK.	89
4.5	Résultats de la procédure de vérification à l'aide de classifieurs d'apprentissage automatique pour d'autres kinases.	90
4.6	Composantes énergétiques MM-GBSA et énergies libres de liaison (en kcal/mol) pour les cinq composés sélectionnés et l'ulixertinib en complexe avec la protéine.	94
4.7	Propriétés physicochimiques des composés sélectionnés et des ligands de référence (le 38Z et Ulixertinib).	96

Liste des symboles

$\text{\AA}$: Angström

$\Delta Acc_{train/test}$: Différence entre scores (train/test)

ΔG : Énergie libre de liaison

$f(x)$: Modèle

γ : Paramètre de complexité (XGBoost)

λ : Coefficient de régularisation (XGBoost)

$\mathcal{L}$: Perte totale (Loss)

$Obj(\Theta)$: Fonction objective de XGBoost

$\Omega(f_p)$: Terme de régularisation du p -ième arbre (XGBoost)

ω : Score du noeud feuille (XGBoost)

$q(x)$: Noeud feuille correspondant à l'échantillon x (XGBoost)

r : Coefficient de corrélation de Pearson

$\sigma(z)$: Fonction d'activation (DNN)

$\sim$: Similitude ou approximation

T : Nombre total de feuilles de l'arbre (XGBoost)

X : valeur réelle de la variable

X' : valeur normalisée de la variable

X_{max} : valeur maximale dans la colonne du descripteur

X_{min} : valeur minimale dans la colonne du descripteur

y_i : Label ou valeur observable

$\hat{y}_i$: Label ou valeur prédite

z : Transformation linéaire (Pré-activation du neurone en DNN)

Introduction Générale

Dans le contexte de la chimie médicinale, l'un des objectifs principaux est d'identifier de nouvelles molécules à potentiel thérapeutique, ce qui implique la recherche de nouveaux composés chimiques possédant une activité biologique rationnellement prévisible. Les cibles biologiques sont généralement des systèmes macromoléculaires tels qu'un récepteur spécifique, une enzyme ou un gène qui, lorsqu'ils sont modulés, influencent une fonction biologique associée à une maladie [1]. Le besoin de conceptualiser et de développer de nouvelles molécules actives est toujours croissant, en raison du manque de médicaments pour plusieurs maladies et des problèmes de résistance aux médicaments [2].

Les protéines kinases constituent une famille d'enzymes fondamentales dont le rôle est de réguler la quasi-totalité des processus cellulaires vitaux, tels que la croissance, la prolifération, la survie, le métabolisme, la division et la différenciation. Elles catalysent la phosphorylation, une modification post-traductionnelle essentielle qui agit comme un interrupteur moléculaire, orchestrant des réseaux de signalisation intracellulaire complexes. Le génome humain encode plus de 500 protéines kinases, formant un système d'une complexité et d'une sophistication remarquables, caractérisé par des interactions internes et externes intriquées. Cette omniprésence et cette interconnexion confèrent au système une vulnérabilité intrinsèque : une perturbation, même minime, de l'activité kinase peut déclencher une série de changements pathologiques en cascade au sein de la cellule, conduisant à diverses maladies, notamment le cancer. La dérégulation de l'activité des kinases, qu'elle soit due à des mutations, une surexpression, des translocations chromosomiques ou une signalisation aberrante, est en effet une caractéristique distinctive de nombreuses pathologies malignes. Plus de 100 oncogènes et près de 30 gènes suppresseurs de tumeurs sont des protéines kinases, soulignant leur rôle pivot dans la biologie du cancer. Des exemples notables incluent les mutations dans le récepteur du facteur de croissance épidermique (EGFR) et BRAF (du nom du gène qui la code), qui sont des moteurs clés de la prolifération et de la survie incontrôlées des cellules cancéreuses [3].

En raison de leur rôle central dans la pathogenèse du cancer, les protéines kinases sont devenues des cibles thérapeutiques majeures. Cette vulnérabilité du système de signalisation kinase, bien que pathologique, représente une opportunité thérapeutique majeure : si une seule kinase dérégulée peut entraîner une prolifération incontrôlée, son inhibition ciblée peut potentiellement "court-circuiter" l'ensemble du processus oncogénique. Au cours de la dernière décennie, la recherche intensive a conduit à l'approbation par l'agence américaine des produits alimentaires et médicamenteux (FDA, Food and Drug Administration [4]) de plus de 50 inhibiteurs de kinases (IK) en clinique, avec des centaines d'autres en cours de développement et d'essais cliniques. Les IK représentent actuellement 50% à 70% des efforts de découverte de médicaments anticancéreux actifs, marquant un changement de paradigme dans le traitement du cancer. L'impact

thérapeutique des IK est révolutionnaire, transformant des maladies autrefois fatales, comme la leucémie myéloïde chronique (LMC) avec l'Imatinib, en conditions chroniques gérables. Initialement, la recherche a identifié la dérégulation des kinases comme une cause fondamentale du cancer. Les avancées en criblage par ARN interférent et en séquençage génomique ont accéléré l'identification de kinases "essentiels" ou de "mutations pilotes". Cette connaissance approfondie de la biologie des kinases a permis de passer d'une simple observation à une conception rationnelle de médicaments [5].

Cependant, malgré ces avancées, l'efficacité des IK est souvent limitée par le développement de mécanismes de résistance acquise, principalement dus à des mutations dans le domaine kinase ou à l'activation de voies de signalisation alternatives. Cela crée un cycle où de nouveaux inhibiteurs doivent être développés pour contrer les mécanismes de résistance aux générations précédentes. La poursuite de la découverte de nouveaux inhibiteurs n'est donc pas seulement une question d'élargir l'arsenal thérapeutique, mais de maintenir une longueur d'avance sur l'adaptation tumorale. Le développement continu de nouvelles molécules, guidées par des relations structure-activité et des avancées technologiques, promet une spécificité et une efficacité accrues, tout en cherchant à minimiser les effets hors-cible et à surmonter les mécanismes de résistance émergents. L'évolution vers des IK de nouvelle génération (allostériques, covalents) est une conséquence directe de cette compréhension structurelle et mécanistique approfondie, permettant de cibler des isoformes spécifiques et de surmonter la résistance. Cette "course aux armements" thérapeutique souligne la nécessité d'une recherche continue et d'une approche dynamique dans le développement de médicaments, intégrant des stratégies comme les thérapies combinées, le suivi en temps réel des mutations et l'utilisation de l'intelligence artificielle pour prédire et contrer la résistance [6].

Le processus de découverte de médicaments est une tâche coûteuse et longue. L'identification d'une substance réellement active et ses essais cliniques peuvent prendre environ 12 ans et coûter aux alentours de 1,8 milliard de dollars américains [7]. Afin d'optimiser temps et argent, et faire face aux défis thérapeutiques et économiques de façon efficace, l'utilisation de nouvelles techniques de conception *in silico* devient cruciale [7].

De nos jours, avec la croissance de la disponibilité des données biologiques et de la puissance de calcul, la conception de médicaments assistée par ordinateur (CADD, Computer-Aided Drug Design) est devenue indispensable pour raccourcir le processus de découverte de médicaments [7]. L'intervention des chimistes théoriciens dans ce contexte est alors d'une extrême nécessité. De nombreux médicaments approuvés par la FDA ont été complètement ou partiellement identifiés grâce à l'utilisation des techniques CADD. Le Captopril, un médicament ciblant l'enzyme de conversion de l'angiotensine, a été identifié en utilisant à la fois la conception de médicaments basée sur le ligand (LBDD, Ligand-Based Drug Design) et sur la structure (SBDD, Structure-Based Drug Design). Le Norfloxacin (inhibiteur de la topoisomérase II et IV), le Flurbiprofène (ciblant la cyclooxygénase-2) et l'Imatinib (inhibiteur de tyrosine kinase) sont également des exemples de médicaments découverts en utilisant des approches *in silico* [8].

Comme évoqué précédemment, dans le cadre de la découverte de médicaments anti-cancéreux, il y a un besoin urgent de développer de nouveaux IK. Certaines kinases n'ont pas encore d'inhibiteur approuvé par la FDA, d'autres ont des inhibiteurs avec une toxicité élevée, et les cliniciens soulignent également le besoin de différents inhibiteurs pour une même cible

thérapeutique, en raison du problème de résistance aux médicaments [2].

Les deux principales approches utilisées sont le LBDD et le SBDD. Le principal objectif de ce travail de thèse est de concevoir et d'identifier de nouveaux candidats inhibiteurs de protéines kinases, en s'inscrivant simultanément dans une dynamique de développement et d'affinement de méthodes computationnelles dédiées à cette finalité. Afin d'offrir une vision claire et structurée des contributions apportées, les travaux réalisés sont présentés selon l'articulation suivante. Les chapitres 1 et 2 constituent le socle théorique de cette thèse, à travers une revue bibliographique et un état de l'art critique. Le premier chapitre expose le contexte biologique et pathologique des protéines kinases, en détaillant leurs rôles fondamentaux dans la transduction du signal et leurs implications majeures dans les processus oncogéniques. Le second chapitre présente les deux grandes approches méthodologiques mentionnées en introduction, à savoir le Ligand-Based Drug Design (LBDD) et le Structure-Based Drug Design (SBDD), en discutant leurs principes, leurs champs d'application respectifs, ainsi que leurs complémentarités dans le processus de découverte de candidats médicaments.

Le chapitre 3 propose une contribution méthodologique originale, à travers le développement d'une nouvelle approche visant à améliorer le pouvoir prédictif de la modélisation QSAR contemporaine, en réponse aux limitations actuelles de cette dernière en matière de robustesse et de généralisabilité. Le chapitre 4 présente les travaux computationnels menés au cours de ces quatre années de recherche, consacrés à l'identification de nouvelles substances potentiellement inhibitrices de la kinase ERK2, une cible thérapeutique d'intérêt en oncologie. Enfin, une conclusion générale dresse le bilan des résultats obtenus, discute les limites rencontrées, et ouvre sur les perspectives envisagées pour prolonger et enrichir ces travaux, tant sur le plan méthodologique que sur le plan applicatif.

Partie I : Contexte et méthodes

Chapitre 1

Les Proteines Kinases

1.1 Introduction

La phosphorylation des protéines représente l'une des modifications post-traductionnelles les plus répandues et les plus importantes dans les systèmes biologiques [9]. Ce mécanisme réversible est fondamental pour la communication et la régulation cellulaires, agissant comme un interrupteur dynamique qui contrôle l'activité et la fonction des protéines [10]. Il permet aux cellules de répondre rapidement et avec précision aux signaux internes et externes, orchestrant ainsi une multitude de processus biologiques complexes.

Au cœur de ce système de régulation se trouvent les protéines kinases. Ces enzymes, membres d'une vaste famille de kinases, sont spécifiquement responsables de la catalyse de l'événement de phosphorylation [9]. Leur présence ubiquitaire et leur implication dans la quasi-totalité des voies cellulaires soulignent leur rôle indispensable dans le maintien de l'homéostasie cellulaire et la coordination des diverses réponses biologiques.

1.2 Définition et mécanisme d'action des protéines kinases

1.2.1 Qu'est-ce qu'une protéine kinase ?

Les protéines kinases sont des enzymes qui appartiennent à la grande famille des kinases et sont spécifiquement chargées du mécanisme de phosphorylation des protéines. Elles accomplissent cette tâche en transférant un groupe phosphate (PO_4) d'une molécule d'ATP, qui sert de donneur de phosphate, vers un résidu d'acide aminé spécifique au sein d'une protéine cible [9].

Le génome humain code pour 518 protéines kinases identifiées, formant un ensemble connu sous le nom de "kinome" [9]. Ces enzymes sont classées en fonction du résidu d'acide aminé qu'elles phosphorylent : les sérine/thréonine kinases (STKs), les tyrosine kinases (TKs) et les kinases à double spécificité (DSKs) qui peuvent phosphoryler les trois types d'acides aminés [9]. Au-delà de cette classification, elles sont subdivisées en sous-familles telles que AGC, CaMK, CK1, CMGC, STE, TK et TKL [9]. La présence de 518 protéines kinases humaines et la désignation du "kinome" indiquent que la phosphorylation n'est pas un événement isolé, mais une couche de régulation profondément intégrée et omniprésente dans presque tous les processus cellulaires. Le fait que jusqu'à 30% de toutes les protéines humaines puissent être

modifiées par l'activité des kinases [9] souligne la nature envahissante de ce mécanisme de contrôle. Cela signifie que pour comprendre les protéines kinases, il ne suffit pas d'étudier des enzymes individuelles, mais d'apprécier les réseaux complexes et interconnectés qu'elles forment. Cette compréhension est essentielle pour déchiffrer la complexité cellulaire et identifier de vastes opportunités thérapeutiques.

1.2.2 Le processus de phosphorylation : Un interrupteur moléculaire

Le mécanisme central de la phosphorylation implique le transfert d'un groupe phosphate vers le groupe hydroxyle (-OH) situé sur la chaîne latérale d'acides aminés spécifiques : la sérine, la thréonine ou la tyrosine [9]. Cette addition modifie les propriétés de la protéine, la transformant généralement d'une nature hydrophobe à hydrophile en raison de la charge négative du groupe phosphate [9].

Ce changement de charge et d'hydrophilie entraîne des altérations significatives de la conformation tridimensionnelle de la protéine [9]. Ces modifications conformationnelles agissent comme un interrupteur moléculaire "marche/arrêt", capable d'activer ou de désactiver l'activité de la protéine, de modifier sa stabilité, d'altérer sa localisation subcellulaire ou d'influencer son activité enzymatique [10]. Par exemple, un acide aminé phosphorylé peut se lier à des molécules capables d'interagir avec d'autres protéines, facilitant ainsi l'assemblage et le détachement de complexes protéiques [9].

Les protéines kinases possèdent un domaine catalytique, généralement composé de sous-domaines N-terminal et C-terminal reliés par un brin peptidique, formant un site actif. L'activation de ce domaine se produit souvent par phosphorylation d'une boucle d'activation ou par des mécanismes allostériques [9]. Les kinases disposent également de domaines non catalytiques qui permettent la fixation de substrats et le recrutement d'autres protéines de signalisation [9]. L'observation qu'un "bon équilibre énergétique favorise également la phosphorylation" et qu'il existe un "équilibre constant entre les événements de phosphorylation et de déphosphorylation médiés par les kinases, les phosphatases, l'ATP et/ou l'ADP" [9] met en évidence que ce processus n'est pas seulement une réaction chimique. C'est un mécanisme piloté par l'énergie. L'utilisation de l'ATP comme donneur de phosphate souligne que la phosphorylation est un mécanisme énergétiquement favorable et efficace. Cette contribution d'énergie élevée permet un contrôle rapide, précis et réversible de la fonction protéique, ce qui en fait un système idéal pour la signalisation cellulaire dynamique qui exige des réponses rapides et une gestion efficace de l'énergie.

1.2.3 La réversibilité : Rôle des phosphatases

La phosphorylation des protéines est un processus réversible, essentiel pour maintenir l'équilibre cellulaire et assurer une signalisation transitoire [9]. La réaction inverse, appelée déphosphorylation, est catalysée par des enzymes nommées protéines phosphatases [10]. Ces enzymes éliminent le groupe phosphate de la protéine, lui permettant de retrouver sa conformation et son activité d'origine [10].

Ce mécanisme de déphosphorylation est vital pour "éteindre" les voies de transduction du signal une fois qu'un signal a été reçu et traité [10]. Il garantit également que les protéines kinases sont disponibles pour les événements de signalisation ultérieurs, prévenant ainsi une activation

incontrôlée ou prolongée. L'interaction constante entre les kinases et les phosphatases [9] établit un équilibre dynamique crucial. Sans l'action précise et opportune des phosphatases, les signaux persisteraient indéfiniment, entraînant un dysfonctionnement cellulaire et potentiellement des processus incontrôlés. Cela souligne que la déphosphorylation n'est pas seulement un mécanisme de "nettoyage", mais un composant actif et indispensable de la fidélité du signal et de la régulation cellulaire, garantissant que les réponses cellulaires sont proportionnées, transitoires et adaptables aux conditions changeantes.

TABLEAU 1.1 – Classification des protéines kinases.

Classe de Kinase	Acides Aminés Ciblés	Caractéristiques Générales
Sérine/Thréonine Kinases (STKs)	Ser, Thr	Majorité des kinases humaines ; impliquées dans des changements conformationnels importants.
Tyrosine Kinases (TKs)	Tyr	Associées à la modification de la localisation cellulaire des protéines ; souvent des récepteurs membranaires.
Kinases à Double Spécificité (DSKs)	Ser, Thr, Tyr	Peuvent phosphoryler les trois types d'acides aminés.

1.2.4 Fonctions physiologiques des protéines kinases

1.2.4.1 Transduction du signal cellulaire

Les protéines kinases régulent la grande majorité des voies cellulaires, jouant un rôle particulièrement crucial dans la transduction du signal [9]. Elles sont des composants clés des cascades de signalisation, où une kinase phosphoryle et active une autre, entraînant une série d'événements qui transmettent et amplifient un signal extracellulaire en une réponse intracellulaire [9]. Les kinases agissent comme des interrupteurs "marche/arrêt" pour de nombreux récepteurs de surface cellulaire et leurs voies de signalisation en aval [10]. Cela permet un contrôle précis des réponses cellulaires à divers stimuli, allant des hormones et neurotransmetteurs aux facteurs de croissance [11]. Au-delà de la simple transmission de signaux, les kinases sont également connues pour contrôler l'amplitude du signal, garantissant des réponses cellulaires appropriées [9]. Le concept de "cascades d'événements" [12] et la capacité des kinases à "contrôler l'amplitude du signal" [13] mettent en évidence leur rôle au-delà de la simple transmission de signaux. Chaque étape d'une cascade de kinases offre une opportunité d'amplification du signal, permettant à un stimulus externe subtil de déclencher une réponse cellulaire interne robuste. De plus, l'interconnexion de diverses voies de signalisation [13] implique que les kinases agissent comme des intégrateurs sophistiqués d'informations. Elles traitent simultanément plusieurs signaux entrants, permettant aux cellules de prendre des décisions complexes et dépendantes du contexte (par exemple, proliférer, se différencier ou subir l'apoptose) plutôt que de simplement réagir à une seule entrée. Cela les rend centrales à la capacité "computationnelle" de la cellule.

1.2.4.2 Régulation de la croissance et différenciation cellulaire

Les protéines kinases sont profondément impliquées dans des processus cellulaires fondamentaux tels que la prolifération cellulaire, la différenciation, l'apoptose (mort cellulaire programmée) et la réponse à divers stress cellulaires [10].

L'implication profonde des kinases dans des processus critiques comme la croissance cellulaire, la prolifération et la survie [10] révèle un principe fondamental : les voies mêmes qui sont essentielles à la vie normale peuvent, lorsqu'elles sont dérégulées, devenir les moteurs de maladies. Par exemple, le rôle d'Akt1 dans l'inhibition de l'apoptose et la promotion de la croissance [13] est vital pour le développement et la réparation des tissus, mais s'il est incontrôlé, il contribue directement à la croissance cancéreuse en empêchant la mort cellulaire et en encourageant une division incontrôlée. Cela met en évidence que le cancer exploite et détourne souvent la machinerie fondamentale de la cellule, faisant de ces kinases essentielles des cibles privilégiées à la fois pour la régulation physiologique normale et pour l'intervention pathologique.

1.2.4.3 Rôle dans le métabolisme et l'homéostasie

Les protéines kinases font partie intégrante de la régulation du métabolisme cellulaire et du maintien de l'homéostasie physiologique globale [13].

Des exemples spécifiques incluent :

- *Protéine Kinase A (PKA)* : Une famille d'enzymes activées par l'AMP cyclique (AMPC). La PKA régule les métabolismes du glycogène, du sucre et des lipides [13]. Elle régule également la transcription génique via des effecteurs comme CREB et phosphoryle d'autres kinases telles que Raf, GSK3 et FAK [13].

- *Protéine Kinase B (PKB) / Akt* : Une protéine cruciale dans la signalisation cellulaire des mammifères. Akt1 inhibe l'apoptose et favorise la biosynthèse des protéines et la croissance tissulaire [13]. Akt2 est impliquée dans la signalisation de l'insuline et le transport du glucose [14].

- *Protéine Kinase C (PKC)* : Une famille de sérine/thréonine kinases activées par le calcium et le diacylglycérol (DAG). La PKC transmet les signaux extracellulaires et active d'autres protéines kinases, notamment les MAP kinases [13].

- *Mitogen-Activated Protein Kinase (MAPK)* : Protéines intracellulaires qui catalysent la phosphorylation des protéines MAP. Les MAPK sont activées par une cascade de kinases (MAPKKK, MAPKK, MAPK) et sont impliquées dans des événements de la vie cellulaire tels que la mitose, l'apoptose, la différenciation et la survie cellulaire en réponse à divers signaux externes [13].

- *Récepteurs Tyrosine Kinases (RTKs)* : Une famille de près de 60 récepteurs transmembranaires qui reçoivent des signaux externes et les transmettent à l'intérieur de la cellule. Ils jouent un rôle vital dans le contrôle de processus tels que le cycle cellulaire, la migration cellulaire, le métabolisme, la prolifération et la différenciation. Des exemples incluent les récepteurs de l'insuline (IR), des facteurs de croissance épidermiques (EGFR) et des facteurs de croissance de l'endothélium vasculaire (VEGFR) [13].

- Plus généralement, les catégories de voies de signalisation cellulaire, y compris celles impliquant les protéines kinases, sont connues pour affecter divers aspects du métabolisme cellulaire [13].

1.3 Role dans le cancer

1.3.0.1 Dérégulation des kinases : Oncogènes et suppresseurs de tumeurs

Les protéines kinases sont fréquemment dérégulées dans le cancer, jouant un rôle central dans son développement et sa progression [13]. Cette dérégulation peut impliquer l'activation de proto-oncogènes (gènes qui favorisent la croissance cellulaire) en oncogènes, ou l'inactivation de gènes suppresseurs de tumeurs (gènes qui inhibent la croissance cellulaire et favorisent l'apoptose) [15].

Dans les cellules saines, l'activité des kinases est étroitement régulée. Cependant, dans le cancer, les mécanismes normaux d'auto-inhibition communs à de nombreuses kinases sont souvent perturbés, conduisant à des protéines constitutivement actives [15]. Cette activité incontrôlée favorise la prolifération cellulaire, la survie et d'autres caractéristiques du cancer. Cette activation persistante peut entraîner un phénomène connu sous le nom de "dépendance oncogénique", où les cellules cancéreuses deviennent dépendantes de la signalisation continue d'une kinase dérégulée spécifique pour leur survie et leur croissance [12]. Cette dépendance

rend ces kinases des cibles attrayantes pour l'intervention thérapeutique. Le concept de "dépendance oncogénique" [15] est une pierre angulaire de la thérapie ciblée moderne contre le cancer. Il implique que, malgré le paysage génétique complexe du cancer, certaines tumeurs deviennent fortement dépendantes de l'activité aberrante d'une seule voie oncogénique, souvent pilotée par une kinase dérégulée. Cette dépendance transforme une maladie apparemment intraitable en une maladie présentant une vulnérabilité spécifique et exploitable. En identifiant et en ciblant ces "mutations pilotes" [16], les cliniciens peuvent obtenir des réponses cliniques spectaculaires, faisant des IK un outil puissant contre des types de cancer spécifiques.

1.3.0.2 Mécanismes d'activation oncogénique des kinases

La dérégulation des protéines kinases dans le cancer résulte généralement de diverses altérations génétiques et épigénétiques :

- *Mutations ponctuelles à gain de fonction* : Des modifications spécifiques de la séquence d'ADN qui entraînent une kinase hyperactive [17].
- *Amplification génique et surexpression* : Une augmentation du nombre de copies d'un gène de kinase, entraînant une surproduction de la protéine kinase et son activation constitutive. Un exemple notable est ERBB2/neu, fréquemment surexprimé dans les cancers du sein, de l'estomac et du poumon [17].
- *Réarrangements chromosomiques et protéines de fusion* : Des translocations ou d'autres changements chromosomiques qui conduisent à la création de nouvelles protéines de fusion avec une activité kinase aberrante. L'exemple classique est la protéine de fusion BCR-ABL1 dans la LMC [17].
- *Délétions partielles* : Des délétions au sein de la séquence codante ou des régions régulatrices du récepteur qui peuvent également contribuer à la dérégulation [17].
- *Mutations à perte de fonction dans les suppresseurs de tumeurs* : Alors que de nombreuses kinases oncogéniques sont activées, certaines kinases agissent comme des suppresseurs de tumeurs. Des mutations à perte de fonction (LOF) ou des délétions dans ces kinases (par exemple, MKK4) peuvent supprimer leur effet inhibiteur sur la croissance cellulaire, favorisant ainsi la tumorigenèse [17].

La figure 1.1 [2] illustre les voies normales et anormales d'activation des protéines tyrosine kinases (PTK) et montre ce que provoque une signalisation aberrante en faveur de la tumorigenèse.

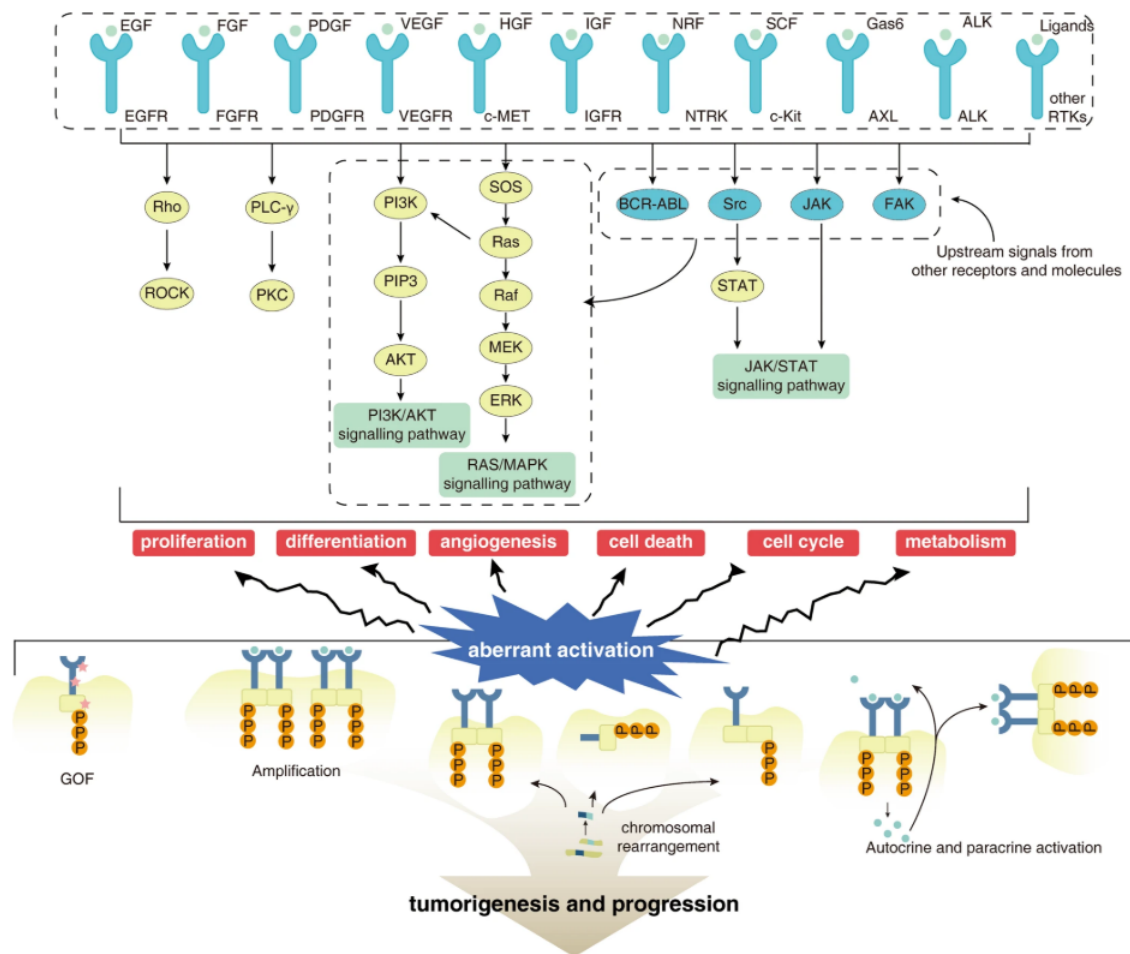


FIGURE 1.1 – L'activation et l'inactivation normales des PTK sont essentielles pour maintenir une fonction cellulaire normale. Les mutations d'activation des PTK comprennent des mutations activatrices (gain de fonction), l'amplification génomique et la surexpression, les réarrangements chromosomiques (fusions géniques), ainsi que les boucles autocrines/paracrines de ligands.

La diversité des mécanismes (mutations ponctuelles, amplifications, fusions, délétions) [13] par lesquels les kinases deviennent oncogéniques met en évidence l'hétérogénéité génétique extrême du cancer. Cette complexité rend directement nécessaire le passage à la médecine de précision, où l'empreinte moléculaire spécifique d'une tumeur (c'est-à-dire son ensemble unique de kinases dérégulées) guide les décisions thérapeutiques. Comprendre ces mécanismes variés est crucial pour concevoir des thérapies ciblées capables de contrecarrer précisément le moteur oncogénique spécifique chez un patient donné, allant au-delà des traitements à large spectre.

1.3.0.3 Exemples de kinases dérégulées et leur impact en cancérogénèse

- *BCR-ABL1* : C'est un exemple classique de protéine de fusion résultant d'une translocation chromosomique (chromosome Philadelphie) dans la LMC. BCR-ABL1 est une tyrosine kinase constitutivement active qui entraîne la prolifération incontrôlée des cellules myéloïdes [18].

- *EGFR (Récepteur du Facteur de Croissance Épidermique)* : Un récepteur tyrosine kinase (RTK) fréquemment surexprimé ou muté dans divers cancers, notamment le cancer du poumon

non à petites cellules (NSCLC). Sa dérégulation conduit à une prolifération et une survie cellulaires accrues [18].

- *HER2 (Récepteur 2 du Facteur de Croissance Épidermique Humain)* : Un autre RTK, souvent amplifié dans le cancer du sein, entraînant une activation constitutive et une croissance tumorale agressive [18].

- *CK1 α (Caséine Kinase 1 alpha)* : Cette sérine/thréonine kinase est hyperactive dans diverses hémopathies malignes, y compris les leucémies myéloïdes et lymphocytaires, le lymphome diffus à grandes cellules B (DLBCL) et le myélome multiple (MM) [12]. Elle module des voies critiques telles que Wnt/ β -caténine, NF- κ B, PI3K/AKT, la réponse aux dommages à l'ADN et l'autophagie, favorisant souvent la survie et la prolifération cellulaire [17].

- *CK2 (Caséine Kinase 2)* : Constitutivement active et surexprimée dans les leucémies aiguës et chroniques, les lymphomes non hodgkiniens (LNH) et le myélome multiple (MM) [18]. CK2 agit en synergie avec les oncogènes, activant des voies clés comme PI3K/AKT, NF- κ B, Wnt/ β -caténine et JAK/STAT, renforçant ainsi les caractéristiques du cancer, contrecarrant les suppresseurs de tumeurs et protégeant le "onco-kinome" [19]. Son comportement correspond souvent au concept de "dépendance non oncogénique" [17].

- *MKK4 (Mitogen-Activated Protein Kinase Kinase 4)* : Une kinase à double spécificité initialement identifiée comme un suppresseur de tumeur, avec des mutations à perte de fonction (LOF) associées à divers cancers. Elle peut favoriser la sénescence, un mécanisme suppresseur de tumeur [17]. Cependant, son rôle est controversé, car certaines études ont rapporté une fonction oncogénique dans des contextes spécifiques, tels que certains cancers du sein et du pancréas [17].

Les exemples de CK1 α et CK2 [18] ayant des rôles complexes, parfois doubles, et MKK4 [17] présentant des fonctions à la fois suppresseurs de tumeurs et oncogéniques selon le contexte, révèlent une nuance critique : les kinases ne sont pas simplement "bonnes" ou "mauvaises". Leur impact dépend fortement de l'environnement cellulaire, du contexte génétique spécifique de la tumeur et du réseau de signalisation précis dans lequel elles opèrent. Cette nature pléiotropique explique pourquoi les thérapies ciblées peuvent avoir une efficacité et des effets secondaires variables, soulignant la nécessité d'une compréhension approfondie du contexte oncogénique spécifique et du potentiel d'effets hors cible ou de voies compensatoires.

TABLEAU 1.2 – Kinases dérégulées dans le cancer et leurs mécanismes oncogéniques.

Kinase Dérégulée	Type de Kinase	Cancer Associé	Mécanisme de Dérégulation	Rôle en Cancérogenèse
BCR-ABL1	TK	Leucémie Myéloïde Chronique (LMC)	Protéine de fusion (translocation chromosomique)	Prolifération cellulaire incontrôlée, survie accrue.
EGFR	RTK	Cancer du poulmon non à petites cellules, autres cancers	Surexpression, mutations activatrices	Croissance, survie et prolifération cellulaire.
HER2	RTK	Cancer du sein, estomac, poulmon	Amplification génique, surexpression	Croissance tumorale agressive.
CK1 α	Sér/Thr Kinase	Leucémies myéloïdes et lymphocytaires, DLBCL, myélome multiple	Surexpression, hyperactivité	Promotion de la survie cellulaire, régulation de la réponse au stress et de l'autophagie.
CK2	Sér/Thr Kinase	Leucémies aiguës et chroniques, lymphomes non hodgkiniens, myélome multiple	Surexpression, hyperactivité constitutive	Synergie avec les oncogènes, anti-apoptose, soutien de la survie cellulaire, protection de l'onco-kinome.
MKK4	Kinase à double spécificité	Cancers pancréatiques, pulmonaires (rôle controversé)	Mutations LOF (suppresseur de tumeur) ; parfois oncogène (contexte spécifique)	Suppression tumorale (par sénescence) ; rôle oncogénique dans certains contextes.

1.3.1 L'importance cruciale de la découverte de nouveaux inhibiteurs de kinases

Les IK, en particulier les inhibiteurs de tyrosine kinases (ITK), constituent une classe de thérapies ciblées qui bloquent spécifiquement l'activité des enzymes kinases [18]. Ils sont conçus pour interférer avec des protéines de signalisation mutantes spécifiques qui favorisent la croissance et la survie du cancer [18].

Un avantage majeur des IK, par rapport à la chimiothérapie conventionnelle, est leur capacité à cibler et à tuer sélectivement les cellules cancéreuses tout en minimisant les dommages aux cellules saines [18]. Cela se traduit par moins d'effets secondaires et moins de gravité pour les patients. Les IK ont révolutionné le traitement du cancer, transformant des maladies potentiellement mortelles comme la LMC en maladies chroniques gérables [18]. Ils peuvent induire une rémission à long terme, prolonger la survie des patients et servir d'alternatives efficaces lorsque d'autres traitements anticancéreux ont échoué [18]. En novembre 2022, 50 ITK étaient approuvés par la FDA pour divers cancers solides et hématologiques [18]. L'avènement des IK marque un changement fondamental dans la thérapie du cancer, passant des agents cytotoxiques à large spectre à des interventions moléculaires très précises. Ce changement est enraciné dans la compréhension de la "dépendance oncogénique" [18], où l'identification d'une kinase aberrante spécifique permet une perturbation ciblée de la survie des cellules cancéreuses. Le succès de l'imatinib dans la LMC [18] en est un excellent exemple, démontrant comment la compréhension moléculaire peut transformer une maladie mortelle en une condition chronique, validant ainsi l'ensemble du domaine de la thérapie ciblée et de l'oncologie de précision.

1.3.1.1 Impact clinique et défis actuels

Les IK sont approuvés pour un nombre croissant de cancers, notamment le cancer du sein HER2-positif, la LMC, les tumeurs stromales gastro-intestinales (GIST), le cancer du rein, le cancer du poumon non à petites cellules (NSCLC), le mélanome et divers cancers de la thyroïde [17]. Parmi les exemples spécifiques figurent l'Imatinib (LMC, GIST), le Gefitinib (NSCLC), le Sorafenib, le Lenvatinib, le Vandétanib, le Cabozantinib, le Selpercatinib, le Pralsetinib, le Larotrectinib, l'Entrectinib, le Dabrafenib et le Tramétinib (divers cancers de la thyroïde) [20].

Malgré des réponses initiales spectaculaires, la résistance aux IK est une limitation inévitable et majeure [20]. Cette résistance peut être primaire (de novo) ou acquise [20]. Les mécanismes clés incluent :

- Modifications du gène cible : Des mutations au sein du domaine kinase de la protéine cible (par exemple, mutations BCR-ABL1, EGFR T790M) qui empêchent la liaison de l'inhibiteur [20].
- Surexpression de BCR-ABL1 : Une production accrue de la kinase cible, qui submerge l'inhibiteur [20].
- Activation de "voies de contournement" : Des voies de signalisation compensatoires qui deviennent actives et permettent aux cellules cancéreuses de croître et de survivre même lorsque la cible primaire est inhibée [20].

- Changements dans l'activité des transporteurs de médicaments : Une augmentation de l'efflux du médicament hors des cellules cancéreuses [20].
- Transformation histologique : Les cellules cancéreuses changent de type, devenant moins dépendantes de la voie oncogénique initiale [17].

Bien qu'une sélectivité élevée soit souvent souhaitée, de nombreux IK actuellement utilisés en clinique ne sont pas entièrement sélectifs, ciblant plusieurs kinases [20]. Ce multi-ciblage peut parfois offrir une efficacité plus large (par exemple, l'efficacité de l'Imatinib dans la LMC et les GIST grâce à l'inhibition de c-Kit, le Dasatinib, le Nilotinib, le Sunitinib, le Sorafenib, le Lapatinib) [20]. Cependant, cela augmente potentiellement le risque d'effets secondaires indésirables et de toxicité en inhibant les kinases dans les cellules saines [12]. De plus, les inhibiteurs doivent être très puissants pour concurrencer les concentrations élevées d'ATP intracellulaire [20].

Malgré leur nature ciblée, les IK peuvent encore provoquer des effets secondaires importants. Les problèmes courants incluent des problèmes cutanés (éruptions cutanées, sécheresse, desquamation, acné, rougeurs, démangeaisons), des changements au niveau des ongles, de la fatigue, de la fièvre et des changements de poids [20]. Certains IK peuvent entraîner des problèmes cardiovasculaires plus graves comme l'hypertension et la protéinurie [20], ou des saignements/caillots sanguins [20]. Ils peuvent également ralentir la cicatrisation des plaies [20]. Le problème omniprésent de la résistance aux médicaments [20] souligne que le cancer n'est pas une maladie statique mais une entité évolutive. Les thérapies ciblées exercent une forte pression sélective, conduisant à l'émergence de clones résistants par divers mécanismes (mutations, voies de contournement, changements de transporteurs) [20]. Cela crée une "course aux armements" dans le traitement du cancer, nécessitant une recherche et un développement continus de nouveaux inhibiteurs et des stratégies de traitement dynamiques (par exemple, thérapies combinées, traitements séquentiels) pour surmonter ou retarder la résistance. La discussion nuancée sur la sélectivité [20] souligne en outre que la conception des médicaments est un équilibre complexe entre spécificité et multi-ciblage pour un bénéfice clinique optimal.

1.3.1.2 Perspectives et stratégies futures

Un objectif principal de la recherche future est le développement d' "inhibiteurs de kinases de nouvelle génération" spécifiquement conçus pour surmonter la résistance aux médicaments, en particulier les mutations au sein du domaine kinase [12]. Des exemples incluent LOXO-195 pour les cancers positifs à la fusion TRK avec des mutations de résistance, l'Osimertinib pour la mutation EGFR T790M dans le NSCLC, l'Avapritinib pour les GIST avec la mutation PDGFRA D842V, et l'Asciminib pour la LMC résistante aux inhibiteurs ATP-compétitifs [12].

De nouvelles stratégies de conception d'inhibiteurs sont explorées :

- **Ciblage des kinases mutantes** : Des composés spécifiquement conçus pour inhiber les formes mutantes de kinases résistantes aux médicaments [12].
- **Inhibiteurs allostériques** : Des molécules qui se lient à un site autre que le site actif, induisant un changement conformationnel qui inhibe l'activité kinase [12]. Cela offre un

mécanisme différent pour contourner les mutations du site actif.

- **Inhibiteurs covalents** : Des médicaments qui forment une liaison permanente avec la kinase, ce qui peut conduire à une inhibition plus soutenue [12].

De plus en plus, les stratégies impliquent la combinaison de plusieurs inhibiteurs pour cibler différents nœuds au sein d'une voie de signalisation ou à travers des voies interconnectées [12]. Cela inclut :

Double inhibition : Par exemple, la combinaison d'inhibiteurs de MEK et de BRAF pour retarder la résistance dans le mélanome malin [12].

Combinaisons de médicaments à faible dose : Une nouvelle approche axée sur le ciblage de réseaux clés de signalisation kinase plutôt que de kinases ou de voies de signalisation individuelles, visant à réaliser le plein potentiel des traitements combinés [21].

La recherche explore également des systèmes de délivrance avancés, tels que les systèmes de nanoparticules pour les inhibiteurs de tyrosine kinases multi-cibles, afin d'améliorer l'administration et l'efficacité des médicaments dans la thérapie du cancer [22]. L'avenir met l'accent sur une approche réfléchie et individualisée de la gestion du cancer, tirant parti d'une compréhension plus approfondie des mécanismes moléculaires pour adapter les traitements [23]. Le passage à une "thérapie combinée multidroge à faible dose qui cible les réseaux clés de signalisation kinase plutôt que les kinases ou les voies de signalisation individuelles" [21] et le développement d'inhibiteurs allostériques/covalents [12] représente une évolution cruciale dans le développement de médicaments. Cela reconnaît que l'inhibition à cible unique, bien qu'efficace, est souvent transitoire en raison de la plasticité des cellules cancéreuses. L'objectif futur est de perturber des réseaux pathologiques entiers, d'anticiper les mécanismes de résistance et d'adapter dynamiquement les stratégies thérapeutiques, reflétant une compréhension plus sophistiquée et systémique de la biologie du cancer.

TABLEAU 1.3 – Exemples d'inhibiteurs de kinases approuvés en phase clinique et leurs indications.

Inhibiteur de Kinase	Type de Kinase Ciblée	Indications Cliniques Notables
Imatinib (Gleevec)	BCR-ABL1, KIT, PDGFR	LMC, GIST
Gefitinib (Iressa)	EGFR	NSCLC
Sorafenib (Nexavar)	Multi-TK (VEGFR, PDGFR, RAF)	Cancer du rein, carcinome hépatocellulaire, cancer de la thyroïde
Lenvatinib (Lenvima)	Multi-TK (VEGFR, FGFR, RET, KIT, PDGFR)	Cancer de la thyroïde, carcinome hépatocellulaire, cancer du rein
Vandétanib (Caprelsa)	RET, EGFR, VEGFR	Cancer médullaire de la thyroïde
Cabozantinib (Cabometyx)	MET, VEGFR2, RET, KIT	Cancer médullaire de la thyroïde, carcinome hépatocellulaire, cancer du rein
Selpercatinib (Retevmo)	RET	Cancers de la thyroïde (médullaire, anaplasique, papillaire, folliculaire) avec mutations RET
Pralsetinib (Gavreto)	RET	Cancer médullaire de la thyroïde
Larotrectinib (Vitrakvi)	NTRK	Cancers avec fusions géniques NTRK (thyroïde, etc.)
Entrectinib (Rozlytrek)	NTRK, ROS1, ALK	Cancers avec fusions géniques NTRK (thyroïde, etc.), NSCLC
Dabrafénib (Tafinlar)	BRAF	Mélanome, cancer du poumon, cancer de la thyroïde (en combinaison avec Tramétinib)
Tramétinib (Mekinist)	MEK	Mélanome, cancer du poumon, cancer de la thyroïde (en combinaison avec Dabrafénib)
Osimertinib	EGFR (mutant T790M)	NSCLC (résistant à d'autres ITK)
Avapritinib	KIT, PDGFR	GIST (avec mutation PDGFRA D842V)
Asciminib	BCR-ABL1	LMC (résistant aux ITK ATP-compétitifs)

1.4 Conclusion

Les protéines kinases sont des enzymes indispensables qui orchestrent la quasi-totalité des fonctions cellulaires par le processus réversible de phosphorylation des protéines. Leur régulation précise est vitale pour des processus allant de la transduction du signal et du métabolisme à la croissance, la différenciation et la survie cellulaire. Cependant, leur dérégulation fréquente par des mutations, des amplifications ou des fusions est un moteur de la prolifération et de la survie incontrôlées caractéristiques du cancer, conduisant à une "dépendance oncogénique".

Le développement d'IK ciblés a révolutionné la thérapie du cancer, offrant des traitements plus précis et moins toxiques pour des populations de patients spécifiques. Pourtant, le défi persistant de la résistance aux médicaments, induite par l'adaptabilité inhérente des cellules cancéreuses, nécessite une innovation continue. Les stratégies futures se concentreront sur le développement d'inhibiteurs de nouvelle génération capables de surmonter les mutations de résistance, l'exploration de nouveaux mécanismes de liaison (par exemple, les inhibiteurs allostériques, covalents) et la mise en œuvre de thérapies combinées sophistiquées qui ciblent des réseaux de signalisation entiers plutôt que des kinases isolées. Cette compréhension évolutive promet de libérer davantage le potentiel thérapeutique de l'inhibition des kinases, conduisant à des réponses plus durables et à de meilleurs résultats pour les patients atteints de cancer.

Chapitre 2

Méthodes Computationnelles de Conception des Médicaments

2.1 Introduction

Les méthodes computationnelles de conception de médicaments se divisent principalement en deux grandes approches : celles basées sur le ligand et celles basées sur la structure. Les méthodes "ligand-based" s'appuient sur l'hypothèse que des composés ayant des structures similaires interagissent avec les mêmes cibles biologiques, avec un mécanisme de reconnaissance similaire. Parmi ces méthodes, la modélisation QSAR permet de prédire l'activité biologique de nouvelles molécules en analysant la relation entre la structure chimique de ligands connus et leur activité, aidée par des descripteurs structuraux et physicochimiques.

Ces approches sont particulièrement utiles lorsque la structure 3D de la cible biologique n'est pas connue, en exploitant les données expérimentales liées aux ligands pour concevoir ou optimiser de nouveaux composés actifs.

Les méthodes "structure-based" tirent parti de la connaissance tridimensionnelle de la cible biologique, obtenue par des techniques de biologie structurale telles que la cristallographie aux rayons X ou la RMN. Elles consistent à modéliser l'interaction entre la cible et des ligands potentiels au niveau atomique afin de concevoir des molécules capables de se lier spécifiquement à la cible. Le docking moléculaire est une technique phare dans cette catégorie, permettant de prédire la meilleure conformation d'un ligand dans le site de liaison de la protéine cible et d'évaluer l'affinité de liaison. Cette approche favorise une conception rationnelle des médicaments, surtout lorsque la structure de la cible est bien caractérisée, et elle complète souvent les données obtenues par les méthodes "ligand-based" pour améliorer la découverte et l'optimisation des candidats médicaments.

2.2 Approches basées sur le ligand (LBDD)

Les approches basées sur le ligand reposent sur les connaissances extraites des substances actives actuellement connues, telles que leurs caractéristiques structurales, physiques et chimiques. Ces éléments sont considérés comme des descripteurs importants permettant de caractériser les molécules déjà identifiées, à partir desquelles de nouveaux composés à potentiel thérapeutique seront prédits avec certains aspects similaires et des profils améliorés

sur le plan de la bioactivité et de la toxicité. Le principe est que les mêmes caractéristiques structurales entraînent des propriétés similaires, et il existe plusieurs approches qui se sont révélées très efficaces dans la CADD. Nous introduirons brièvement dans ce qui suit, les méthodes basées sur le ligand les plus utilisées.

2.2.1 Modélisation par groupements pharmacophoriques

La modélisation basée sur le pharmacophore est une approche essentielle en conception de médicaments. Un pharmacophore représente l'ensemble des caractéristiques stéréoélectroniques nécessaires pour assurer les interactions optimales avec une cible biologique spécifique, et provoquer son effet biologique désiré [24]. En d'autres termes, il s'agit d'un modèle qui identifie et définit les éléments structuraux communs à des molécules actives, tels que les groupes donneurs ou accepteurs de liaisons hydrogène, les régions hydrophobes, et les charges électriques [24]. Cette méthode permet de filtrer rapidement de grandes bases de données de composés pour identifier des candidats potentiels, facilitant ainsi la découverte de nouveaux médicaments avec des profils d'activité et de sélectivité améliorés. Ce type de modélisation est particulièrement utile lorsqu'il existe peu d'informations structurales sur la cible biologique, permettant de guider la synthèse et l'optimisation des nouvelles molécules [25].

2.2.1.0.1 Approche basée sur les fragments C'est une approche basée sur de très petits fragments organiques, ayant une interaction souhaitée avec les résidus du site actif ou allostérique du récepteur. En partant de cette petite molécule, il est possible d'ajouter d'autres fragments améliorant potentiellement son activité, et résultant ainsi en la conception de nouveaux composés avec un grand potentiel d'attachement au récepteur macromoléculaire [25].

2.2.2 Relations Quantitatives Structure-Activité (QSARs)

2.2.2.1 Modélisation classique

La modélisation classique des QSARs est une approche fondamentale en chimie médicinale et en toxicologie, dont les racines modernes remontent au début des années 1960 avec les travaux pionniers de Corwin Hansch et de ses collaborateurs [26]. Le principe sous-jacent est que les variations de l'activité biologique des composés sont directement liées à leurs différences de propriétés structurales [27, 28]. Avant l'avènement des techniques de machine learning (ML), les études QSAR classiques établissaient des corrélations entre des points finaux biologiques (tels que les affinités de liaison, les constantes d'inhibition ou les constantes de vitesse) et des propriétés physico-chimiques ou structurales des molécules, notamment la lipophilie, la polarisabilité, et les effets électroniques et stériques [27, 29]. Les modèles initiaux étaient principalement linéaires, avec des approches clés comme l'analyse de Hansch et l'analyse de Free-Wilson [27, 29, 30].

Le processus typique avant toute modélisation QSAR par ML passe d'abord par la représentation adéquate des Molécules. Chaque molécule est convertie en un vecteur numérique de descripteurs moléculaires. Ceux-ci peuvent être :

- Descripteurs 1D-2D : Propriétés calculées à partir de la formule brute ou de la topologie 2D (ex. : poids moléculaire, nombre d'atomes lourds, nombre de liaisons rotatives, polarité, surface polaire topologique (SPT), indices de connectivité ... etc.) [31].
- Empreintes Digitales Moléculaires (Molecular Fingerprints) : Vecteurs binaires (ou de comptes) qui encodent la présence/absence de sous-structures spécifiques ou de chemins atomiques (ex. : les Morgan fingerprints ou "empreintes circulaires") [32].
- Descripteurs 3D : Basés sur la conformation 3D de la molécule (ex. : champs de potentiel électrostatique, moments dipolaires, volumes moléculaires ... etc.) [31].
- Descripteurs basés sur les graphes : Représentations de graphes qui capturent la structure atomique et les liaisons, souvent utilisées avec des réseaux neuronaux graphiques (GNN) pour apprendre directement des structures [33].

2.2.2.1.1 Principales méthodes utilisées Historiquement, plusieurs méthodes de régression ont été employées pour construire ces modèles QSAR :

1. **Régression Linéaire Multiple (MLR, Multiple Linear Regression)** : C'est l'une des méthodes les plus anciennes et les plus couramment utilisées en QSAR [34]. Sa simplicité et l'interprétabilité directe de ses expressions mathématiques sont ses principaux avantages [34]. Cependant, la MLR est sensible à la multicolinéarité entre les descripteurs, ce qui peut rendre difficile l'identification des variables les plus significatives [34]. Pour pallier ces limitations, des améliorations ont été développées, telles que la Best Multiple Linear Regression (BMLR), la Heuristic Method (HM) et la Genetic Algorithm based MLR (GA-MLR) [34].
2. **Régression par les Moindres Carrés Partiels (PLS, Partial Least Squares)** : Développée par S. Wold, la PLS est devenue une méthodologie très populaire et pragmatique en QSAR/QSPR [35, 36]. Elle est particulièrement efficace lorsque le nombre de descripteurs est comparable ou supérieur au nombre de composés, ou en présence de corrélations entre les variables, situations où les méthodes classiques de moindres carrés peuvent échouer ou produire des résultats instables [37]. La PLS permet de construire des modèles stables, corrects et hautement prédictifs même avec des descripteurs corrélés [37]. Les variables latentes de la PLS sont choisies pour maximiser la corrélation avec la variable dépendante, assurant ainsi un nombre minimal de facteurs nécessaires [37]. Elle peut également révéler des relations entre l'activité et les descripteurs même si ces derniers contribuent peu aux premières composantes principales [37].
3. **Régression sur Composantes Principales (PCR, Principal Component Regression)** : La PCR est une technique de régression basée sur la PCA [38]. Elle est utilisée pour estimer les coefficients de régression en utilisant les composantes principales des variables explicatives comme régresseurs [38]. La PCR est efficace pour surmonter le problème de la multicolinéarité en excluant les composantes principales à faible variance [38]. Le processus implique d'abord une PCA sur la matrice des données pour obtenir les composantes principales, puis de sélectionner un sous-ensemble de ces compo-

santes pour la régression, ce qui permet une réduction de dimensionnalité [38]. La PCR est souvent mentionnée aux côtés de la MLR et de la PLS dans les études QSAR [34].

2.2.2.1.2 Critères de Validation de Golbraikh et Tropsha La validation est un aspect essentiel de toute modélisation QSAR [39]. Alexander Golbraikh et Alexander Tropsha ont souligné en 2002 que la validation externe est la seule méthode fiable pour établir un modèle QSAR robuste [39]. Ils ont démontré qu'une valeur élevée du R^2 de validation croisée (leave-one-out, LOO ou q^2), souvent supérieure à 0.5, est une condition nécessaire mais non suffisante pour garantir une forte capacité prédictive du modèle [39]. Pour qu'un modèle QSAR soit considéré comme prédictif, Golbraikh et Tropsha ont formulé un ensemble de critères statistiques [39,40] :

- Pour l'ensemble d'entraînement, le coefficient de détermination R^2 entre les activités prédites et observées doit être supérieur à 0.6 [40].
- Pour l'ensemble de test, le q^2 doit être supérieur à 0.5 [41].
- Un modèle QSAR fiable doit présenter une bonne concordance entre les activités prédites et expérimentales, idéalement s'approchant de la ligne $y = x$ [40].
- Des critères supplémentaires incluent :
 - Les coefficients de corrélation estimés entre les valeurs expérimentales et prédites avec (r^2) et sans (r_0^2) ordonnée à l'origine. La valeur de $[(r^2 - r_0^2)/r^2]$ doit être inférieure à 0.1 [41].
 - Les valeurs des pentes k (régression des prédites sur les expérimentales) et k' (régression des expérimentales sur les prédites) doivent être comprises entre 0.85 et 1.15 [41].
 - Le R_{pred}^2 (coefficient de corrélation au carré prédictif pour la validation externe) est également un indicateur important [40].
 - Le test de randomisation Y (ou Y-randomization test) est utilisé pour s'assurer que le modèle n'est pas le fruit du hasard; les valeurs de q^2 pour l'ensemble de données réel doivent être significativement plus élevées que celles obtenues avec des propriétés cibles randomisées [40].

Ces critères sont essentiels pour garantir la robustesse et la capacité prédictive des modèles QSAR avant leur application pratique.

2.2.2.1.3 Domaine d'applicabilité (Diagramme de Williams) La définition du Domaine d'Applicabilité (DA) est un aspect crucial de la validation des modèles QSAR, car la fiabilité des prédictions est généralement limitée aux composés chimiquement similaires à ceux utilisés pour construire le modèle [42]. Le diagramme de Williams (ou plot de levier) est un outil graphique couramment utilisé pour établir ce domaine d'applicabilité [41]. Il représente les résidus standardisés (axe des y) en fonction des valeurs de levier (valeurs de "hat"; axe des x)

pour chaque composé de l'ensemble d'entraînement [42]. Le DA est typiquement défini dans une zone carrée délimitée par ± 2 écarts-types pour les résidus et un seuil de levier h^* [42]. Le levier d'un composé est proportionnel à sa distance de Mahalanobis par rapport au centroïde de l'ensemble d'entraînement [42]. Un seuil d'alerte h^* est généralement fixé à trois fois le levier moyen (p/n , où p est le nombre de paramètres du modèle et n le nombre de composés d'entraînement) [42]. Les composés dont le levier dépasse h^* sont considérés comme étant en dehors du DA et leurs prédictions sont jugées peu fiables [42].

2.2.2.2 Machine learning (supervisé)

Avec l'avènement des mégadonnées et des avancées algorithmiques, l'apprentissage automatique a émergé comme un paradigme puissant, transformant la manière dont ces relations sont établies et exploitées. Le ML permet de modéliser avec une précision et une complexité sans précédent les interactions entre descripteurs moléculaires, d'accélérer le criblage virtuel de bibliothèques chimiques massives et de guider la conception rationnelle de nouveaux composés [43].

L'intégration du ML en QSAR s'articule principalement autour de deux grandes catégories d'apprentissage : l'apprentissage supervisé (pour la prédiction) et l'apprentissage non supervisé (pour l'inspection et l'organisation des données). L'apprentissage supervisé est la pierre angulaire des applications prédictives en QSAR. Il repose sur l'utilisation de données "étiquetées", où les entrées (descripteurs moléculaires) sont associées à des sorties connues (propriétés cibles). Dans ce qui suit, nous introduisons brièvement les méthodes couramment utilisées pour les régressions et les classifications, les deux principaux axes du ML supervisé.

2.2.2.2.1 Régressions Les modèles de régression en ML sont conçus pour prédire des propriétés continues des molécules. Ces propriétés incluent l'activité biologique (ex. : la concentration inhibitrice médiane, IC50; la constante d'inhibition, Ki; ou la concentration effective médiane, EC50), des caractéristiques de toxicité (ex. : la dose létale 50%, LD50), des propriétés physico-chimiques (ex. : la solubilité, le coefficient de partage octanol-eau logP), ou encore des propriétés ADME (Absorption, Distribution, Métabolisme, Excrétion).

Les algorithmes de régression courants en ML sont ceux qui ont une efficacité approuvée au sein de la communauté des chercheurs. Avec l'évolution rapide et continue de l'informatique, chaque année, les théoriciens publient de nouvelles études rapportant l'efficacité de nouveaux algorithmes dans les QSARs. Les algorithmes les plus utilisés sont :

Les Régressions Linéaires Multiples (MLR, Multiple Linear Regression) : Modèles simples et interprétables [44].

Les Machines à Vecteurs de Support (SVM, Support Vector Machines) pour la Régression (SVR) : Efficaces pour des données de haute dimensionnalité [44].

Les Forêts Aléatoires (RF, Random Forests) : Ensembles d'arbres de décision robustes et moins sujets au surapprentissage [44].

Les Gradient Boosting Machines (GBM) (ex. : XGBoost, LightGBM) : Algorithmes performants qui construisent un modèle séquentiellement en corrigeant les erreurs des modèles précédents [44].

Réseaux Neuraux (NN, Neural Networks) : Capables d'apprendre des relations non linéaires complexes et d'extraire des caractéristiques latentes. Les DNNs sont particulièrement puissants [44].

Validation du Modèle : La performance est évaluée à l'aide de métriques telles que le coefficient de détermination (R^2), l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE), et la validation croisée pour assurer la robustesse et la généralisabilité du modèle [44].

2.2.2.2.2 Classifications Les modèles de classification ML sont utilisés pour attribuer des molécules à des catégories discrètes. Il peut s'agir de déterminer si une molécule est "active" ou "inactive" (classification binaire) ou de la classer dans plusieurs catégories (classification multi-classes), par exemple "non-toxique", "faiblement toxique", "modérément toxique", "hautement toxique". Ces modèles sont cruciaux pour le criblage virtuel, permettant de filtrer rapidement de grandes bibliothèques de composés [45]. Les algorithmes de classification courants sont :

La Régression Logistique : Très utilisée en classification binaire pour sa simplicité, interprétabilité et faible coût computationnel [46].

Les Machines à Vecteurs de Support (SVM) : Robustes pour les problèmes de classification [45].

Forêts Aléatoires (RF) : Offrent une bonne interprétabilité et une robustesse contre le surapprentissage [47].

Les k-plus proches voisins (k-NN, k-Nearest Neighbors) : Classe les points de données en fonction de la majorité de leurs voisins [45].

Les Naïve Bayes : Un classifieur probabiliste basé sur le théorème de Bayes [45].

Les Réseaux Neuraux de Classification : DNNs avec une couche de sortie adaptée à la classification (ex. : fonction softmax pour la classification multi-classes) [44].

Afin de valider les modèles obtenus, la performance est mesurée par l'exactitude (accuracy), la précision (precision), le rappel (recall), le score F1, la matrice de confusion, et surtout l'aire sous la courbe ROC (AUC-ROC), qui évalue la capacité du modèle à distinguer les classes [48].

Il est important de noter que les méthodes de régression utilisées dans la modélisation QSAR classique visent principalement à établir une relation explicite entre les variables et la propriété cible. Cependant, ce n'est pas automatiquement le cas pour les nouvelles approches de régression (ou de classification) en ML, qui utilisent le même principe mathématique de base (pour corrélérer la propriété aux variables), mais sans accorder d'importance particulière à la contribution individuelle de chaque variable.

Lorsque l'on souhaite connaître la contribution de chaque variable dans la décision finale d'un modèle de ML, plusieurs méthodes existent, telles que l'analyse SHAP ou la feature importance [49]. Ces méthodes sont notamment utilisées dans le contexte des récents QSARs basés sur l'intelligence artificielle explicable (XAI, Explainable Artificial Intelligence) [50].

2.2.3 Apprentissage non-supervisé

L'apprentissage non supervisé est appliqué pour découvrir des structures, des motifs ou des relations cachées dans des données non étiquetées. Il est essentiel pour l'exploration de l'espace chimique, la réduction de la complexité des données, et l'identification de groupes de molécules sans connaissance a priori de leurs propriétés.

2.2.3.0.1 Réduction de la Dimensionnalité :

- *Analyse par Composantes Principales (PCA, Principal Component Analysis)* : Projette les données de haute dimensionnalité dans un espace de dimension inférieure tout en conservant la variance maximale. Utile pour visualiser des jeux de données complexes de descripteurs moléculaires et identifier les principales sources de variation [38].

- *UMAP (Uniform Manifold Approximation and Projection) et t-SNE (t-Distributed Stochastic Neighbor Embedding)* : Ces techniques non linéaires sont particulièrement efficaces pour visualiser des structures complexes et des clusters de données de haute dimensionnalité, révélant des regroupements naturels de molécules en fonction de leur similarité structurale ou de propriétés [51,52].

Leur utilité se résume dans la facilitation de la visualisation de l'espace chimique, l'identification de régions densément peuplées de molécules actives, ou la détection des "outliers" (molécules atypiques) qui pourraient représenter de nouveaux scaffolds intéressants.

2.2.3.0.2 Regroupement (Clustering) :

- *k-means Clustering* : Regroupe les molécules en k clusters, où chaque molécule appartient au cluster dont le centre (centroïde) est le plus proche [45].

- *Clustering Hiérarchique* : Construit une hiérarchie de clusters, représentée par un dendrogramme, utile pour explorer les relations de similarité à différents niveaux [53].

- *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)* : Identifie les clusters basés sur la densité des points de données, capable de détecter des clusters de formes arbitraires et de gérer le bruit [54].

Elles sont utiles pour le regroupement et fondamentales pour identifier des "chemotypes" (groupes de molécules ayant des caractéristiques structurales et/ou des activités similaires), pour organiser de grandes bibliothèques de composés, pour sélectionner des sous-ensembles diversifiés pour le criblage ou pour identifier des doublons dans les bases de données. Elles aident également à comprendre la diversité d'une bibliothèque chimique ou à cibler des régions

spécifiques de l'espace chimique pour la synthèse.

2.3 Approches basées sur la structure du récepteur (SBDD)

2.3.1 Docking moléculaire

Le docking moléculaire représente une technique computationnelle fondamentale dans le domaine de la biologie structurale moléculaire et de la conception de médicaments assistée par ordinateur [55–57]. Son objectif principal est de prédire les modes de liaison prédominants d'un ligand, typiquement une petite molécule, avec une protéine cible dont la structure tridimensionnelle est déjà connue [55, 56]. Cette approche modélise l'interaction entre une petite molécule et une protéine au niveau atomique, évaluant leur compatibilité spatiale et énergétique au sein du site actif d'un récepteur [56, 57]. L'objectif ultime est d'atteindre une conformation optimisée pour le récepteur et le ligand, ainsi qu'une orientation relative, afin de minimiser l'énergie libre du système global [57]. L'approche trouve aussi des applications importantes pour étudier les interactions entre différentes macromolécules biologiques.

Historiquement, le concept du docking a évolué depuis le modèle initial de "clé-serrure", qui postulait une rigidité stricte des partenaires de liaison, vers le modèle plus nuancé d'ajustement induit [56]. Ce dernier reconnaît la flexibilité conformationnelle des molécules, où la liaison peut entraîner des changements structuraux dans le ligand, le récepteur, ou les deux, pour optimiser l'interaction [56]. Ces modèles théoriques sous-tendent la capacité du docking à prédire non seulement l'orientation optimale, mais aussi l'affinité de liaison et la forme d'interaction entre les molécules [56].

L'importance du docking moléculaire s'étend à diverses étapes cruciales du développement de médicaments. Il est un outil pivot dans la conception de médicaments basée sur la structure par le criblage virtuel et l'optimisation des leads [55, 56]. Dans le criblage virtuel, il permet d'évaluer rapidement de vastes bibliothèques de composés chimiques, accélérant ainsi l'identification de hits potentiels [55, 56]. Pour l'optimisation des leads, le docking est inestimable car il propose des hypothèses structurales détaillées sur la manière dont les ligands interagissent avec et inhibent leur cible, permettant d'affiner les composés existants pour améliorer leur attachement et leur sélectivité [55, 57]. Au-delà de la découverte de médicaments, le docking contribue à caractériser le comportement des petites molécules (substrats naturels) dans le site de liaison des protéines cibles et à élucider des processus biochimiques fondamentaux, offrant une compréhension atomique des interactions moléculaires [57, 58].

Malgré ses avancées, la tâche de docking est intrinsèquement complexe. Le grand nombre de façons dont les molécules peuvent s'associer, combiné à une dépendance exponentielle à la taille des molécules, conduit à une explosion du nombre de configurations possibles à explorer [57]. Pour résoudre ce problème, le docking repose sur deux composants critiques et interdépendants : une procédure de recherche efficace et une fonction de score précise [57]. La procédure de recherche est chargée d'explorer l'immense espace conformationnel pour identifier les poses pertinentes, tandis que la fonction de score évalue et classe ces poses en fonction de leur affinité de liaison prédite [55–57, 59, 60]. Une performance médiocre dans l'une de ces composantes peut compromettre l'ensemble du processus de docking. Par exemple, si un algorithme de recherche ne parvient pas à trouver la pose correcte, même la

fonction de score la plus précise ne pourra pas la classer correctement. Inversement, si la fonction de score est incapable de distinguer les poses natives des non-natives ou d'estimer avec précision les affinités de liaison, les résultats du criblage virtuel seront peu fiables, indépendamment de la qualité de la recherche des poses [57, 59]. Cela souligne que les avancées dans le domaine du docking nécessitent des améliorations simultanées des deux aspects, ou l'adoption de stratégies qui compensent les faiblesses d'un composant par la force de l'autre, comme un criblage initial rapide suivi d'un raffinement avec des méthodes plus sophistiquées.

2.3.1.1 Fonctions de score en docking moléculaire

Les fonctions de score sont des composants essentiels du docking moléculaire, jouant un rôle central dans l'évaluation et le classement des différentes conformations ligand-récepteur générées par les algorithmes de recherche [56, 57, 59, 60]. Leur fiabilité est directement liée à la précision des prédictions de liaison, car elles guident la détermination des poses de ligands et la prévision de l'affinité de liaison [59]. Malgré des recherches approfondies menées au fil des années, la prédiction précise et rapide des interactions protéine-ligand demeure un défi significatif [57, 59]. Les fonctions de score peuvent être classées en plusieurs catégories principales selon leur dérivation et les principes sous-jacents qu'elles utilisent pour estimer l'énergie de liaison [59, 60]. Nous présentons brièvement dans les paragraphes suivants les types de fonctions de score existants.

2.3.1.1.1 Fonctions de score basées sur les champs de force Ces fonctions sont développées à partir des principes des interactions atomiques physiques, incluant les interactions de van der Waals (VDW), les interactions électrostatiques, et les forces de tension/flexion/torsion des liaisons [60]. Les paramètres de ces champs de force sont généralement dérivés à la fois de données expérimentales et de calculs de mécanique quantique, ancrant ainsi la prédiction dans les lois fondamentales de la physique [60]. Un exemple typique est la fonction de score utilisée par le logiciel DOCK, dont les paramètres énergétiques sont tirés des champs de force Amber. Cette fonction est composée de deux termes énergétiques principaux : un terme VDW de Lennard-Jones et un terme électrostatique [57, 60].

Cependant, un défi majeur pour les fonctions de score basées sur les champs de force réside dans le traitement du solvant lors de la liaison du ligand [60]. Bien que des méthodes rigoureuses comme la perturbation de l'énergie libre (FEP) et l'intégration thermodynamique (TI) traitent explicitement les molécules d'eau, elles sont extrêmement coûteuses en calcul, les rendant impraticables pour le criblage virtuel à grande échelle (vastes bibliothèques de composés) [60]. Pour réduire la charge computationnelle, des modèles accélérés, tels que les modèles Poisson-Boltzmann/surface area (PB/SA) et Generalized-Born/surface area (GB/SA), sont souvent utilisés en post-scoring [60]. Néanmoins, ces modèles de solvation implicite ne peuvent pas toujours rendre compte de manière adéquate de l'effet de désolvatation, où les groupes chargés et non polaires ont des préférences distinctes pour les environnements aqueux ou non aqueux. Ignorer cet effet peut biaiser la fonction de score vers des ligands fortement chargés [60]. De plus, la combinaison des termes énergétiques individuels dans une seule fonction de score est difficile et nécessite souvent des coefficients de pondération empiriques, obtenus par ajustement à des données de liaison expérimentales. Trouver un ensemble universel

de coefficients applicable à une grande diversité de complexes protéine-ligand reste un défi, car les termes d'énergie libre individuels peuvent ne pas être simplement additifs [60].

2.3.1.1.2 Fonctions de score empiriques Les fonctions de score empiriques estiment l'affinité de liaison en se basant sur un ensemble de termes énergétiques pondérés [60]. L'affinité de liaison (ΔG , bien que sa dénomination en tant que telle est controversée dans le contexte du docking [61]) est calculée comme une somme pondérée de différents termes énergétiques, tels que l'énergie VDW, l'électrostatique, les liaisons hydrogène, la désolvatation, l'entropie et l'hydrophobicité. Les coefficients de pondération sont déterminés en ajustant la fonction aux données d'affinité de liaison d'un ensemble d'entraînement de complexes protéine-ligand dont les structures et les affinités sont connues [60].

L'avantage principal de ces fonctions est leur rapidité computationnelle, les rendant beaucoup plus rapides que les fonctions basées sur les champs de force en raison de leurs termes énergétiques plus simples [57, 58, 60]. Plusieurs exemples notables de développement incluent SCORE1 de Böhm, ChemScore d'Eldridge et al., et X-Score de Wang et al., chacun intégrant différents termes énergétiques et étant basé sur des ensembles de données d'entraînement croissants pour affiner leur précision [60].

Cependant, les fonctions empiriques présentent des défis. Il est difficile d'éviter les problèmes de double-comptage lorsque de nombreux termes énergétiques sont inclus dans la fonction [60]. De plus, leur applicabilité générale peut dépendre fortement de la composition de l'ensemble d'entraînement utilisé pour le calibrage. Si l'ensemble d'entraînement n'est pas suffisamment diversifié, la fonction pourrait ne pas être performante pour des systèmes qui s'écartent significativement des exemples sur lesquels elle a été entraînée [58, 60]. Néanmoins, l'augmentation continue du nombre de complexes protéine-ligand avec des structures 3D et des affinités de liaison connues rend possible le développement de fonctions empiriques plus générales et robustes [60].

2.3.1.1.3 Fonctions de score basées sur la connaissance Également désignées sous le nom de fonctions basées sur le potentiel statistique, ces approches utilisent des potentiels énergétiques qui sont dérivés directement des informations structurales contenues dans les structures atomiques déterminées expérimentalement [60]. Les potentiels de paires sont obtenus à partir de la fréquence d'occurrence des paires d'atomes dans une base de données de structures, en appliquant la relation de Boltzmann inverse [60]. Cette méthode permet d'extraire des informations sur les interactions favorables directement à partir de l'observation de complexes réels [60].

Ces fonctions offrent un bon équilibre entre précision et vitesse, ce qui les rend attrayantes pour le criblage de grandes bases de données [58, 60]. Elles sont considérées comme robustes et relativement insensibles à l'ensemble d'entraînement, car les potentiels sont extraits des structures plutôt que d'être ajustés aux affinités. La base de données d'entraînement peut ainsi être très grande et diversifiée [60]. Leur caractéristique de paire permet un scoring rapide, facilitant leur application à de vastes ensembles de composés [60].

Le principal défi persistant pour les fonctions basées sur la connaissance est l'obtention d'un état de référence précis [60]. La plupart des fonctions actuelles approximent l'état de référence par un état atomique aléatoire, ignorant des effets importants tels que le volume exclu et la connectivité atomique [60]. Pour surmonter cette limitation, des méthodes itératives, comme ITScore, ont été développées. Ces méthodes visent à contourner le besoin d'un calcul précis de l'état de référence en ajustant les potentiels de paires jusqu'à ce qu'ils reproduisent les fonctions de distribution de paires déterminées expérimentalement, en tenant compte de l'ensemble du paysage énergétique de liaison [60].

2.3.1.1.4 Fonctions de score basées sur l'apprentissage automatique Inspirées par les fonctions de score basées sur la connaissance, ces approches représentent une avancée significative en utilisant des méthodes d'apprentissage automatique (ML) pour ajuster les affinités de liaison d'un ensemble d'entraînement [59, 60]. Elles exploitent l'occurrence de paires d'atomes et des caractéristiques de paires d'atomes dépendantes de la distance pour générer des potentiels d'interaction [60].

Un avantage clé de ces fonctions est leur capacité à gérer et à ajuster les affinités de liaison à partir de très grands ensembles d'entraînement, grâce à l'inclusion d'un nombre plus élevé de paramètres via des caractéristiques de paires d'atomes basées sur la connaissance [60]. Par exemple, le RF-Score de Ballester et Mitchell, développé à l'aide d'une méthode d'apprentissage automatique, a démontré une corrélation élevée ($R=0.953$) sur un vaste ensemble d'entraînement de 1105 complexes protéine-ligand, soulignant leur potentiel prédictif [60].

L'évolution des fonctions de score, des modèles physiques simples (champs de force) aux approches empiriques, puis basées sur la connaissance, et enfin basées sur l'apprentissage automatique, reflète une quête continue d'amélioration de la précision et/ou de la vitesse [59, 60]. Les fonctions basées sur les champs de force, bien qu'ancrées dans la physique, sont lentes et rencontrent des difficultés avec la solvation et la combinaison des termes énergétiques [58, 60]. Les fonctions empiriques sont plus rapides mais leur généralisabilité dépend fortement de l'ensemble d'entraînement et peuvent souffrir de problèmes de double-comptage [58, 60]. Les fonctions basées sur la connaissance offrent un bon compromis entre vitesse et précision, mais sont confrontées à des défis liés à la définition de l'état de référence [60]. Les fonctions basées sur l'apprentissage automatique, plus récentes, ont la capacité de traiter de très grands ensembles de données et de capturer des relations complexes non linéaires, ce qui peut conduire à une performance prédictive supérieure. Cependant, leur nature de "boîte noire" peut rendre l'interprétation des interactions spécifiques plus difficile [60]. Cette trajectoire d'évolution suggère une reconnaissance croissante de la complexité inhérente aux interactions biologiques, qui ne peuvent pas toujours être entièrement capturées par des modèles physiques simples ou des ajustements linéaires. Il est probable que l'avenir des fonctions de score résidera dans des approches hybrides, combinant la rigueur physique avec la puissance prédictive de l'apprentissage automatique pour optimiser les performances.

2.3.1.1.5 Défis et considérations générales pour les fonctions de score Un défi persistant dans le développement et l'application des fonctions de score est d'atteindre simultanément une prédiction précise et rapide des interactions protéine-ligand [59]. Les fonctions de score doivent être suffisamment rapides pour être appliquées à un grand nombre de dispositions po-

tentielles générées par les algorithmes de recherche, tout en étant suffisamment discriminantes pour distinguer efficacement les conformations de liaison natives des non-natives [57]. Le problème principal souvent rapporté n'est pas tant la capacité des programmes de docking à reproduire la pose correcte (bien que cela reste un défi), mais plutôt l'estimation précise des affinités de liaison relatives des poses de ligands [57].

De nombreuses fonctions de score existantes ne fonctionnent bien que pour une ou deux des trois applications principales du docking (criblage virtuel, prédiction de pose, prédiction d'affinité), ce qui souligne la difficulté de développer une fonction universellement applicable et performante dans tous les contextes [60]. Pour améliorer la fiabilité des prédictions, une technique de plus en plus utilisée est le scoring par consensus. Cette approche combine plusieurs scores différents pour évaluer une conformation de docking, ce qui peut potentiellement atténuer les faiblesses individuelles des fonctions de score et fournir une évaluation plus robuste [58].

TABLEAU 2.1 – Types de fonctions de score en docking moléculaire [60].

Type de Fonction	Principes Fondamentaux	Avantages Clés	Limitations/Défis
Basée sur les Champs de Force	Interactions atomiques physiques (VDW, électrostatique, liaisons)	Ancrée dans la physique, potentiellement plus rigoureuse	Lenteur, traitement complexe du solvant, difficulté à combiner les termes
Empirique	Termes énergétiques pondérés ajustés aux données expérimentales	Rapidité, simplicité des termes énergétiques	Dépendance à l'ensemble d'entraînement, problèmes de double-comptage
Basée sur la Connaissance	Potentiels statistiques dérivés des fréquences d'occurrence de paires d'atomes dans les structures	Bon équilibre précision/vitesse, robustesse, insensibilité à l'ensemble d'entraînement	Difficulté à définir un état de référence précis
Basée sur l'Apprentissage Automatique	Utilise des méthodes ML pour ajuster les affinités de liaison à partir de caractéristiques de paires d'atomes	Capacité à gérer de grands ensembles d'entraînement, potentiellement haute précision	"Boîte noire", interprétabilité réduite, dépendance à la qualité des données d'entraînement

Synthèse des caractéristiques principales des quatre types fondamentaux de fonctions de score, facilitant une compréhension rapide de leurs fondements, forces et faiblesses, ce qui est crucial pour les chercheurs choisissant une méthode de docking.

Le domaine du docking moléculaire est confronté à un compromis fondamental entre la vitesse computationnelle requise pour le criblage de grandes bibliothèques de composés et la précision nécessaire pour prédire des affinités de liaison fiables. Comme le montre l'analyse des différents types de fonctions de score, aucune fonction unique ne semble exceller dans tous les aspects [60]. Les fonctions basées sur les champs de force sont rigoureuses mais lentes, tandis que les fonctions empiriques et basées sur la connaissance sont plus rapides mais peuvent être moins ancrées dans la physique fondamentale ou souffrir de problèmes de généralisabilité [60]. Les fonctions basées sur l'apprentissage automatique, bien que prometteuses en termes de précision, peuvent manquer d'interprétabilité directe [60]. Le

scoring par consensus est une tentative de combiner plusieurs fonctions pour compenser les faiblesses individuelles [58]. Cette situation implique que l'avenir de la prédiction d'affinité de liaison résidera probablement dans des approches hybrides ou des pipelines multi-étapes. Par exemple, des fonctions rapides pourraient être utilisées pour un pré-criblage à haut débit, permettant de réduire considérablement la taille des bibliothèques de composés. Ensuite, des fonctions plus précises, bien que plus lentes, seraient appliquées pour le raffinement des hits les plus prometteurs. Cette stratégie permet d'optimiser l'efficacité globale du processus de découverte de médicaments en tirant parti des forces de différentes fonctions de score à différents stades du processus de criblage.

2.3.1.2 Algorithmes de recherche en docking moléculaire

Les algorithmes de recherche constituent le second pilier fondamental du docking moléculaire. Leur rôle est crucial : ils sont chargés d'explorer l'immense espace conformationnel des interactions ligand-récepteur afin d'identifier les poses de liaison les plus favorables [58,62]. L'efficacité de cette recherche est d'une importance primordiale, car la complexité du problème augmente de manière exponentielle avec la taille des molécules impliquées [57]. Les algorithmes de docking sont principalement classés en fonction du degré de flexibilité qu'ils sont capables de gérer : le docking rigide, le docking semi-flexible et le docking flexible [56,58,62].

2.3.1.2.1 Vue d'ensemble des stratégies de recherche Les algorithmes de docking comportent généralement deux étapes fondamentales. La première consiste à prédire la conformation du ligand ainsi que sa position et son orientation au sein des sites de liaison du récepteur, ce que l'on appelle couramment la "pose" [56,58]. La seconde étape est l'évaluation de l'affinité de liaison de cette pose à l'aide d'une fonction de score [56,58].

Les méthodes de recherche employées peuvent être catégorisées en trois types principaux : déterministes, systématiques et stochastiques [56]. Il est également courant que certaines techniques combinent des éléments de ces différents groupes pour optimiser leur performance [56]. La vitesse et la précision sont des caractéristiques clés pour l'obtention de résultats réussis dans les simulations de docking [62]. Un algorithme doit être rapide pour explorer un grand nombre de configurations, tout en étant suffisamment précis pour identifier les poses biologiquement pertinentes [57,62].

2.3.1.2.2 Algorithmes de docking rigide Le docking rigide est un modèle relativement simple qui considère le récepteur (protéine cible) et le ligand (petite molécule) comme des corps solides et immobiles tout au long du processus de docking [56,58,62]. Cette méthode ne tient pas compte de la flexibilité conformationnelle ni du ligand ni du récepteur, se concentrant essentiellement sur les complémentarités géométriques entre les deux molécules [62].

Bien que cette simplification limite la spécificité et la précision des résultats par rapport aux méthodes plus avancées, le docking rigide est capable d'identifier des sites de liaison potentiels pour diverses protéines [62]. Il est couramment utilisé pour le criblage virtuel initial de grandes bases de données de petites molécules, principalement en raison de sa rapidité. Il offre un moyen rapide de filtrer un grand nombre de composés, en éliminant ceux qui sont

géométriquement inappropriés par une recherche de complémentarité de forme [62]. Parmi les logiciels qui implémentent le docking rigide, ZDOCK est une application notable qui combine la complémentarité de forme, la désolvation et les paramètres électrostatiques à l'aide d'un algorithme de Transformée de Fourier Rapide (FFT) [62]. MS-DOCK est un autre outil de docking rigide basé sur l'algorithme DOCK [62].

2.3.1.2.3 Algorithmes de docking flexible Les méthodes de docking flexible sont conçues pour surmonter les limitations du docking rigide en permettant de prendre en compte la flexibilité du ligand, du récepteur, ou des deux simultanément pendant le processus de docking [56, 62]. Cette capacité conduit à des prédictions plus précises des modes de liaison, mais elle s'accompagne d'un coût computationnel significativement plus élevé en raison de l'augmentation de l'espace de recherche conformationnel [62]. Nous citons les exemples de méthodes suivants :

Construction Incrémentale (IC) : Cette méthode décompose le ligand en fragments qui sont ensuite dockés séparément dans le site du récepteur. Une fois les fragments ancrés, ils sont fusionnés en tenant compte des parties flexibles du ligand [62]. Cette fragmentation permet à l'algorithme de gérer la flexibilité du ligand [62]. Des programmes comme DOCK, FLEXX, FLOG et Surflex utilisent des algorithmes de construction incrémentale [62].

Algorithmes Génétiques (GA) : Les algorithmes génétiques sont une classe d'algorithmes d'optimisation inspirés des processus d'évolution naturelle. Ils visent à trouver la conformation exacte ou la plus proche du minimum global d'énergie en simulant la reproduction, la mutation et la sélection [62]. Des opérateurs génétiques comme le "crossover" (croisement) combinent les caractéristiques de "chromosomes parents" pour produire de nouveaux "descendants" potentiellement améliorés [62]. GOLD (Genetic Optimization for Ligand Docking) est un exemple de logiciel qui utilise un algorithme génétique pour explorer la flexibilité conformationnelle du ligand avec une flexibilité partielle de la protéine [58, 59, 63]. AutoDock, un programme largement reconnu, utilise une variante appelée algorithme génétique Lamarckien (LGA), qui alterne entre l'espace génotypique (mutation et crossover) et l'espace phénotypique (optimisation de la fonction d'énergie) [62, 63].

Monte Carlo (MC) : Initialement utilisées pour la minimisation en dynamique moléculaire (DM), les simulations Monte Carlo ont été adaptées pour le docking flexible. Cette méthode tente de docker le ligand au sein du site récepteur par de nombreuses positions et rotations aléatoires, réduisant ainsi la probabilité de se retrouver piégé dans des minima locaux [58, 62]. Après chaque mouvement aléatoire, une minimisation d'énergie basée sur un champ de force est appliquée pour créer des complexes binaires plus fiables [62]. RosettaDock est principalement basé sur un algorithme de Monte Carlo multi-start et multi-échelle [58]. D'autres programmes comme MCDOCK et ICM implémentent également les simulations MC [62].

Recuit Simulé (SA) : Dans le recuit simulé, un système biomoléculaire est simulé en diminuant progressivement la "température" pendant des intervalles réguliers à chaque cycle de simulation [58, 62]. Cette méthode peut potentiellement atteindre une plus grande précision que le Monte Carlo, car elle prend en compte l'état conformationnel et la flexibilité de la protéine et du ligand dans différents états thermodynamiques au fil du temps. Cependant, le recuit simulé peut être plus coûteux en temps de calcul [62]. Pour minimiser ce coût, des

procédures de filtrage peuvent être appliquées entre le placement du ligand et les cycles du SA. Le recuit simulé peut également être combiné avec d'autres méthodes, comme le Monte Carlo simulé (MCSA) utilisé dans AutoDock, ce qui permet d'explorer une plus grande gamme de conformations et d'améliorer la précision [62].

Autres Algorithmes : D'autres algorithmes de recherche incluent la Géométrie des Distances (DG), qui utilise des informations exprimées par des distances intra- et intermoléculaires pour calculer des structures [62]. Les méthodes de Fast Shape Matching (SM) se concentrent sur le chevauchement géométrique entre deux molécules et sont utilisées dans des programmes comme SYSDOC, EUDOC et DOCK [62]. La Programmation Évolutionnaire (EP) est une classe d'algorithmes qui simulent l'évolution des espèces pour résoudre des problèmes d'optimisation [62]. Enfin, la Recherche Tabou (TS) est une procédure méta-heuristique itérative qui utilise une mémoire pour éviter de revisiter des solutions déjà explorées et favoriser la recherche de nouvelles configurations, démontrant une haute précision en évitant les minima locaux [62].

Le choix de l'algorithme de recherche est un compromis direct entre la fidélité de la modélisation de la flexibilité moléculaire et les ressources computationnelles disponibles. Le docking rigide, bien que rapide, ne tient pas compte de la flexibilité des molécules, ce qui peut limiter sa précision pour des systèmes où les changements conformationnels sont importants [56, 62]. En revanche, les méthodes de docking flexible, qui capturent mieux la dynamique de liaison et les changements conformationnels induits, sont intrinsèquement plus coûteuses en termes de temps de calcul et de puissance CPU [62]. Cela signifie que pour des applications telles que le criblage à haut débit de très grandes bibliothèques de composés, des méthodes plus rapides et moins précises (comme le docking rigide ou semi-flexible) sont souvent utilisées comme première passe. Cette étape initiale permet de filtrer rapidement les composés peu prometteurs, réduisant ainsi la taille de la bibliothèque à un ensemble plus gérable. Les composés les plus prometteurs sont ensuite soumis à des méthodes de docking flexible plus rigoureuses et plus coûteuses pour un raffinement et une analyse plus approfondis. Cette approche multi-étapes est essentielle pour rendre le criblage virtuel à grande échelle pratiquement réalisable et efficace dans le processus de découverte de médicaments.

2.3.1.3 Logiciels et plateformes reconnus

Le paysage du docking moléculaire est riche en outils logiciels et plateformes web, chacun offrant des capacités, des algorithmes et des fonctions de score distincts, adaptés à diverses applications et niveaux de complexité [56]. Les programmes les plus populaires sont les suivants :

AutoDock et AutoDock Vina : Ces logiciels sont parmi les plus reconnus et largement utilisés dans la communauté scientifique, notamment en raison de leur disponibilité gratuite et de leur flexibilité. AutoDock Vina, en particulier, est apprécié pour sa vitesse et sa précision améliorées par rapport à son prédécesseur, AutoDock 4, ce qui en fait un choix privilégié pour le criblage virtuel [63]. AutoDock 4 utilise l'algorithme génétique Lamarckien (LGA) pour la recherche conformationnelle du ligand, tandis qu'AutoDock Vina intègre un algorithme de recherche optimisé [62, 63]. Dans une étude de benchmarking, AutoDock (avec 50 exécutions) a montré une précision de 82,35% pour la prédiction de pose ($\text{RMSD} \leq 2$ Å) et une valeur AUC de 0,8926 pour le criblage virtuel, démontrant sa robustesse [63].

PyRx est une interface graphique (GUI) conviviale qui intègre AutoDock et AutoDock Vina, simplifiant la préparation des ligands, le docking par lots et la visualisation. Cependant, PyRx présente des limitations pour la gestion de grands complexes protéine-ligand et le docking flexible, et il lui manque des outils d'analyse post-docking avancés, nécessitant souvent l'utilisation de logiciels externes comme PyMOL ou Discovery Studio pour ces tâches. Malgré ces limitations, sa simplicité et son efficacité en font un choix populaire pour les chercheurs.

Glide : Développé par Schrödinger, Glide est un logiciel payant réputé pour ses performances élevées. Il s'est distingué dans une étude de benchmarking en surpassant tous les autres programmes testés (GOLD, AutoDock, FlexX, MVD) en prédisant correctement 100% des poses de liaison ($\text{RMSD} \leq 2 \text{ \AA}$) pour les enzymes cyclooxygénases [63]. Pour le criblage virtuel, Glide a obtenu une valeur AUC impressionnante de 0,9275 pour le jeu de données COX2 [63]. Il opère généralement en mode "standard-precision (sp)" et utilise le champ de force OPLS3e pour ses calculs [63].

GOLD (Genetic Optimization for Ligand Docking) : Ce logiciel payant, développé par le Cambridge Crystallographic Data Centre (CCDC), utilise l'algorithme génétique pour explorer les modes de liaison des ligands [58, 63]. GOLD propose plusieurs fonctions de score, dont ChemPLP, GoldScore, ChemScore et ASP. Parmi elles, ChemPLP s'est avérée la plus performante pour la prédiction de pose, atteignant une précision de 76,47% ($\text{RMSD} \leq 2 \text{ \AA}$) [63]. Pour le criblage virtuel combiné, GOLD a affiché une valeur AUC de 0,7609 [63].

FlexX (LeadIT) : Faisant partie de la plateforme LeadIT de BioSolveIT, FlexX est un autre logiciel de docking payant [63]. Dans l'étude de benchmarking mentionnée, FlexX a montré les résultats les plus faibles en prédiction de pose, avec seulement 60,78% de précision ($\text{RMSD} \leq 2 \text{ \AA}$), et une valeur AUC de 0,7379 pour le criblage virtuel combiné [63].

Molegro Virtual Docker (MVD) : Également un logiciel payant, MVD utilise l'algorithme de recherche MolDock SE [63]. Ses performances en prédiction de pose ont été évaluées à 70,58% de précision ($\text{RMSD} \leq 2 \text{ \AA}$) [63].

DOCK : Historiquement, DOCK est l'un des premiers programmes de docking développés. Il est basé sur le principe de la complémentarité de forme et utilise des paramètres dérivés du champ de force Amber pour son scoring [56, 57, 60, 62].

RosettaDock : Ce programme est métaphoriquement décrit comme un "véhicule tout-terrain" capable de gérer des terrains complexes, ce qui implique sa capacité à traiter des systèmes plus difficiles ou des flexibilités plus importantes. Son algorithme principal est basé sur une approche de Monte Carlo multi-start et multi-échelle [58].

Quant aux plateformes web, elles offrent des services de docking moléculaire, rendant la technique plus accessible sans nécessiter de ressources computationnelles locales importantes. Nous citons les exemples de SwissDock et GalaxyDock qui sont des plateformes fournissant des résultats cohérents pour les petits complexes protéine-ligand. HADDOCK est une autre alternative web, souvent recommandée pour des études de docking plus complexes, notamment celles impliquant des informations expérimentales. D'autres plateformes et serveurs web incluent Docking Server, Blaster, Id Target, Flex Pep Dock, GPCRautomode, et Pose & Rank [56].

La diversité des outils de docking moléculaire disponibles, qu'ils soient logiciels ou plateformes web, est considérable, chacun intégrant ses propres algorithmes et fonctions de score [56, 63]. Les performances de ces outils varient significativement en fonction de critères spécifiques tels que la prédiction de pose ou le criblage virtuel, et peuvent même différer selon les systèmes biologiques étudiés, comme le montrent les résultats entre les enzymes COX1 et COX2 [63]. Certains outils sont gratuits et offrent une grande convivialité, comme PyRx qui facilite l'utilisation d'AutoDock et Vina, les rendant populaires pour les chercheurs disposant de ressources limitées. D'autres sont des solutions payantes, comme Glide ou GOLD, qui promettent des performances supérieures pour des tâches spécifiques, souvent au prix de ressources computationnelles plus importantes. Cette situation, souvent décrite par des métaphores comme celle de la "berline fiable" (AutoDock Vina), du "véhicule tout-terrain" (RosettaDock) ou de la "voiture de sport" (Glide et GOLD), illustre parfaitement la variété des capacités et des coûts associés à chaque outil en fonction du "terrain" (la complexité du système biologique) et du "carburant" (les ressources computationnelles et le budget). Il n'existe donc pas de solution "taille unique" en matière de logiciel de docking. Le choix d'un outil doit être guidé par le contexte spécifique de l'étude : la complexité du système (par exemple, la flexibilité du récepteur ou du ligand), les ressources computationnelles disponibles, le budget alloué, et l'objectif précis de l'expérience (qu'il s'agisse de criblage à haut débit, d'optimisation de lead ou de prédiction de pose très précise). Une évaluation comparative (benchmarking) est souvent indispensable pour identifier l'outil le plus adapté à une application donnée, permettant aux chercheurs d'optimiser leurs efforts et d'obtenir les résultats les plus pertinents.

2.3.2 Simulations de dynamique moléculaire

Les simulations de DM constituent une technique largement employée pour étudier la dynamique et les interactions des biomolécules, y compris les protéines et les ligands [64]. Le principe fondamental des simulations de DM repose sur la résolution numérique des équations de mouvement de Newton pour un système de particules, typiquement des atomes ou des groupes d'atomes. [64] Cette approche permet de prédire le mouvement de chaque atome au sein d'un système moléculaire au fil du temps, en se basant sur un modèle général des lois physiques régissant les interactions interatomiques [65].

Contrairement aux méthodes expérimentales statiques, telles que la cristallographie aux rayons X ou la cryo-microscopie électronique (cryo-EM), qui fournissent généralement une structure moyenne ou un cliché statique des interactions protéine-ligand [65–67], les simulations de DM capturent les détails atomiques complets et la résolution temporelle du comportement moléculaire. [65] Cette capacité à révéler les positions et les mouvements de tous les atomes à l'échelle de la femtoseconde est essentielle. Les données expérimentales structurales, bien que cruciales, offrent souvent une image figée des interactions. Les simulations de DM, en résolvant les équations de mouvement de Newton, saisissent intrinsèquement l'évolution temporelle et le comportement dynamique de ces systèmes. Cela permet de combler la lacune d'information dynamique qui est souvent nécessaire pour une compréhension plus approfondie et pour évaluer l'effet des mutations [66]. Ainsi, la DM n'est pas seulement un outil complémentaire, mais un moyen transformateur, offrant des informations dynamiques cruciales (par exemple, les changements conformationnels, la flexibilité) qui sont souvent

requis pour une compréhension plus profonde des processus biologiques et de l'action des médicaments. Cette perspective dynamique est particulièrement vitale pour des domaines comme la conception de médicaments, où la compréhension de la manière dont un ligand interagit avec une protéine flexible au fil du temps peut être plus prédictive de l'efficacité qu'une simple pose de liaison statique.

Un complexe protéine-ligand désigne l'association d'une protéine et d'un ligand formée par leur interaction [68]. Dans le contexte de la DM, la définition d'un ligand est large et englobe diverses molécules telles que les sous-unités protéiques, les molécules médicamenteuses, les oligonucléotides, les glucides et les lipides [69]. La compréhension de ces interactions au niveau atomique est d'une importance capitale pour de nombreux domaines de la bioscience, notamment la conception de médicaments et l'ingénierie enzymatique [66]. Les simulations de DM sont particulièrement précieuses pour l'étude des systèmes protéine-ligand car elles peuvent révéler la composition, l'hétérogénéité et la dynamique de ces interactions, et dans certains cas, leurs arrangements structuraux et leur liaison [69].

2.3.2.1 Applications clés en biologie structurale et en conception de médicaments

L'impact des simulations de DM sur la biologie moléculaire et la découverte de médicaments s'est considérablement accru. Elles se sont avérées précieuses pour décrypter les mécanismes fonctionnels des protéines, élucider les bases structurales des maladies, et concevoir ou optimiser de petites molécules, des peptides et des protéines [65]. Les applications spécifiques incluent :

1. **Identification et Optimisation des Molécules Candidates (Leads) :** La simulation moléculaire peut identifier des composés candidats potentiels et optimiser leur affinité de liaison et leur spécificité pour une protéine cible [64]. Cela implique d'affiner les poses du ligand et d'identifier les interactions clés de liaison ligand-protéine [65].
2. **Prédiction de l'Affinité et de la Cinétique de Liaison :** Les simulations de DM offrent des estimations plus précises des affinités de liaison des ligands (énergies libres) que d'autres approches computationnelles comme le docking [65]. Des techniques telles que la perturbation de l'énergie libre (FEP) sont reconnues pour fournir des estimations précises, en particulier pour les énergies de liaison relatives entre des ligands similaires [65, 70]. La technologie FEP+ de Schrödinger est particulièrement mise en avant pour sa précision prédictive, approchant les méthodes expérimentales (1 kcal/mol) sur un large éventail de ligands et de classes de protéines [71]. Elle est largement adoptée dans la découverte de médicaments pour optimiser la puissance, la sélectivité et la solubilité [71]. La FEP peut également être utilisée pour l'ingénierie des protéines et la prédiction de la thermostabilité [71].

Il est important de noter un changement de paradigme dans la conception de médicaments : un intérêt croissant se manifeste pour la détermination des propriétés cinétiques, telles que le temps de résidence médicament-cible, car elles sont mieux corrélées à l'efficacité du ligand "in vivo" que les propriétés thermodynamiques comme la constante de dissociation à l'équilibre [72]. Ce changement signifie que la DM ne se limite plus à prédire si un ligand se lie, mais aussi combien de temps il

reste lié et comment il interagit dynamiquement. Cette avancée est cruciale pour la conception rationnelle de médicaments, car elle permet le développement de molécules avec une efficacité "in vivo" et une durée d'action améliorées, allant au-delà des simples métriques de puissance. Des approches basées sur la DM, comme la DM par titration thermique (TTMD), émergent pour l'estimation qualitative de la stabilité des complexes protéine-ligand en évaluant la conservation du mode de liaison natif à des températures progressivement croissantes [72].

3. **Compréhension du Comportement Dynamique :** La DM peut évaluer la mobilité et la flexibilité des biomolécules, révélant l'étendue des mouvements de diverses régions à l'équilibre et les types de fluctuations structurales qu'elles subissent [65]. Cela s'étend également au comportement dynamique des molécules d'eau et des ions salins, qui sont souvent cruciaux pour la fonction des protéines et la liaison des ligands [65].
4. **Affinement et Validation Structurale :** Les simulations peuvent tester la précision des structures modélisées ou les affiner, corrigeant les artefacts dus à l'empilement du réseau cristallin ou à l'absence d'une bicouche lipidique [65]. Une pose de ligand stable en simulation est plus susceptible d'être précise [65].
5. **Exploration des Processus Biomoléculaires :** La DM peut observer des processus fonctionnels tels que la liaison des ligands, les changements conformationnels, le repliement des protéines ou le transport membranaire, fournissant des informations mécanistiques au niveau atomique difficiles à obtenir expérimentalement [65, 72]. Cela inclut la compréhension de l'allostérie et la conception de ligands biaisés [65].
Les simulations de DM peuvent également prédire comment les biomolécules réagissent aux perturbations (mutations, changements de ligand, protonation) et peuvent générer de nouvelles hypothèses testables expérimentalement, influençant ainsi la poursuite des travaux expérimentaux [65]. Cela signifie que les simulations de DM ne sont pas seulement un outil d'analyse "a posteriori", mais une partie active du cycle de découverte scientifique. En fournissant des informations mécanistiques au niveau atomique et en prédisant les résultats des perturbations, la DM peut directement éclairer et guider la conception expérimentale, rendant le processus de recherche plus efficace et axé sur les hypothèses. Cela élève la DM d'un outil descriptif à un outil prédictif et prescriptif dans la recherche biomoléculaire.

2.3.2.2 Principes fondamentaux d'une simulation de DM

Le protocole de simulation de DM implique généralement plusieurs étapes clés [64] :

1. **Initialisation :** Le système est initialisé avec un ensemble de coordonnées et de vitesses pour toutes les particules (atomes ou groupes d'atomes) [64]. Cela implique souvent la préparation de la structure protéique (par exemple, à partir de la PDB, ajout de boucles manquantes, protonation) et sa solvation dans un environnement approprié (par exemple, une boîte d'eau, une bicouche lipidique) avec des contre-ions [73].

2. **Équilibration** : Le système est équilibré à une température et une pression désirées en utilisant des techniques telles que le thermostating et le barostating [64]. Cette étape permet au système de se détendre à partir de sa configuration initiale et d'atteindre un état stable cohérent avec l'ensemble thermodynamique choisi.
3. **Production** : La simulation est exécutée pendant une période spécifiée, au cours de laquelle les coordonnées et les vitesses des particules sont enregistrées à intervalles réguliers, générant ainsi une trajectoire [64]. La durée de cette phase de production est cruciale, des simulations de plus de 100 ns sont nécessaires pour les protéines membranaires ou très flexibles, tandis que des simulations courtes (< 5 ns) peuvent être insuffisantes pour une prédiction structurale précise [67].

2.3.2.3 Champs de force couramment utilisés en DM protéine-ligand

2.3.2.3.1 Concept et rôle des champs de force Dans les simulations de DM, le complexe protéine-ligand est représenté comme un système de particules interagissant via une fonction d'énergie potentielle, appelée champ de force (FF) [64]. Le champ de force décrit les interactions entre les atomes, y compris les interactions liées (par exemple, liaisons, angles, dièdres) et non liées (par exemple, forces de van der Waals et électrostatiques) [64]. La précision des simulations de DM dépend intrinsèquement des paramètres du champ de force employés, car une description précise de l'énergie est nécessaire pour que le système puisse peupler les états de plus basse énergie [74,75]. Le gradient de cette fonction d'énergie fournit les forces agissant sur les atomes individuels [74].

2.3.2.3.2 Champs de force additifs majeurs Les champs de force additifs ont atteint un état de maturité après des décennies de développement et sont largement utilisés pour étudier la dynamique des protéines, les interactions protéine-protéine et les applications pharmacologiques [74]. Ils sont basés sur des charges atomiques fixes et ne tiennent pas explicitement compte des effets de polarisation induite. Les champs de force additifs les plus couramment utilisés pour les simulations protéine-ligand sont CHARMM, AMBER et OPLS [64]. De nos jours, certaines équipes de recherche développent par elles-mêmes des champs de force directement dérivés des calculs par des méthodes quantiques (Quantum Mechanical Derived Force Fields QMD), ce qui pousse la fiabilité et la précision des DM à un degré inégalé [76].

2.3.2.3.3 Famille AMBER Les champs de force AMBER (Assisted Model Building with Energy Refinement) sont un choix populaire pour les simulations biomoléculaires, développés par le groupe Kollman [64]. AMBER se compose à la fois d'un ensemble de champs de force mécaniques moléculaires (MM) pour la simulation des biomolécules et d'une suite de simulation qui implémente ces champs de force [77]. Les champs de force AMBER ont fait l'objet d'améliorations continues, se concentrant particulièrement sur les angles dièdres importants [74]. Les versions notables incluent :

1. **ff99SB** : Introduit des modifications au potentiel de la chaîne principale en s'ajustant à des données de niveau quantique [74].
2. **ff99SB-ILDN** : A modifié les potentiels de torsion des chaînes latérales pour une précision améliorée [74].

3. **ff14SB** : Une version largement utilisée pour les simulations de protéines [78].

Les champs de force AMBER prennent en charge la plupart des biomolécules courantes, y compris l'ADN, l'ARN, les glucides et les phospholipides [74]. La suite d'AMBER comprend des outils comme Antechamber et MCPB pour créer des champs de force pour les molécules organiques générales et les centres métalliques, respectivement [77].

General AMBER Force Field (GAFF) pour les ligands : GAFF est spécifiquement conçu pour fournir des résultats raisonnables pour un large éventail de petites molécules organiques, ce qui le rend hautement compatible avec les champs de force AMBER existants pour les protéines et les acides nucléiques [66, 78, 79]. Il dispose de paramètres pour la plupart des molécules organiques et pharmaceutiques composées de H, C, N, O, S, P et halogènes [78]. GAFF utilise une forme fonctionnelle simple et un nombre limité de types d'atomes, intégrant des modèles empiriques et heuristiques pour les constantes de force et les charges atomiques partielles [78].

L'exactitude d'une simulation de DM protéine-ligand ne dépend pas uniquement du champ de force de la protéine ou du ligand isolément, mais de manière critique de leur compatibilité et de la cohérence de leur paramétrisation. Un désaccord peut entraîner des interactions imprécises, soulignant l'importance du développement intégré des champs de force et des outils comme Antechamber [66, 78] qui facilitent cette compatibilité pour diverses chimies de ligands.

- *Attribution des charges* : Les charges atomiques partielles dans GAFF ont été initialement développées en utilisant des charges de potentiel électrostatique restreint (RESP) dérivées de la mécanique quantique (QM), mais en pratique, la méthode semi-empirique Austin Model 1-Bond Charge Corrections (AM1-BCC) est souvent adoptée pour son efficacité [79].
- *GAFF2* : Une deuxième génération, GAFF2, a mis à jour les paramètres liés et non liés pour mieux reproduire les géométries moléculaires, les spectres vibrationnels et les surfaces d'énergie potentielle à partir de calculs QM de plus haut niveau, ainsi que des données ab-initio. GAFF2 montre une corrélation statistiquement significative, meilleure avec les données expérimentales pour l'énergie libre de liaison et l'enthalpie, et un excellent accord pour l'entropie de liaison par rapport à GAFF [79].

Applications : GAFF est largement appliqué dans la CADD, en particulier pour prédire avec précision l'affinité et la sélectivité des interactions récepteur-ligand [78, 79]. Il est utilisé conjointement avec des outils comme Antechamber pour la paramétrisation des ligands dans les simulations de DM [66, 78].

2.3.2.3.4 Famille CHARMM Le champ de force CHARMM (Chemistry at HARvard Macromolecular Mechanics) est un champ de force largement utilisé pour les simulations biomoléculaires, développé par le groupe Karplus [64]. Il est complet, couvrant les protéines, les acides nucléaires, les lipides et les glucides, et s'étend aux composés courants en chimie médicinale grâce au CHARMM General FF (CGenFF) [74].

- **CHARMM36 (C36)** : Cette version représente une amélioration significative de la

surface d'énergie potentielle, conservant la même forme fonctionnelle que son prédécesseur. Les améliorations incluent un nouveau potentiel CMAP pour la chaîne principale, des paramètres dièdres optimisés pour les chaînes latérales et des paramètres de Lennard-Jones affinés [74]. C36 a montré des améliorations significatives dans l'échantillonnage conformationnel (par exemple, une diminution de l'échantillonnage des conformations α R et une augmentation de l'échantillonnage PPII), s'alignant mieux avec les données expérimentales par rapport à d'autres champs de force [74].

CHARMM est un programme polyvalent de modélisation et de simulation moléculaire qui comprend un ensemble complet de champs de force et de vastes capacités pour simuler des systèmes à plusieurs particules [80, 81].

2.3.2.3.5 Famille OPLS OPLS (Optimized Potentials for Liquid Simulations) est un champ de force développé par le groupe Jorgensen, optimisé pour les systèmes organiques et biologiques [64]. Bien que moins d'informations détaillées soient fournies dans les extraits par rapport à AMBER et CHARMM, il est constamment répertorié comme l'un des champs de force les plus couramment utilisés pour les simulations protéine-ligand [64].

Le développement des champs de force est un processus itératif continu, stimulé par la disponibilité de données expérimentales et de mécanique quantique plus précises. Cela signifie que les champs de force ne sont pas des entités statiques, mais qu'ils évoluent continuellement pour mieux représenter la réalité moléculaire. Les chercheurs doivent être conscients des dernières versions et de leurs améliorations spécifiques, car l'utilisation de champs de force obsolètes pourrait compromettre la précision et la fiabilité de leurs simulations. Cela suggère également une boucle de rétroaction où les techniques expérimentales avancées informent les modèles computationnels, et vice-versa.

2.3.2.3.6 Champs de force polarisables (Drude, AMOEBA) Alors que les champs de force additifs sont robustes, la prochaine étape majeure dans l'amélioration de la précision des champs de force des protéines implique l'inclusion des effets de polarisation des charges [74]. Les champs induits par les ions, le solvant, d'autres macromolécules et la protéine elle-même affectent considérablement les interactions électrostatiques [74]. Les champs de force polarisables visent à capturer ces réponses électroniques et dynamiques.

Drude Polarizable Force Field : Développé dans le cadre de CHARMM, le champ de force polarisable de Drude modélise la polarisation en attachant une particule chargée fictive (particule de Drude) à chaque atome, connectée par un ressort harmonique [74].

- *Développement* : Commencé en 2001, il est maintenant pris en charge par NAMD, ChemShell QM/MM et OpenMM [74].
- *Modèles d'eau* : Les modèles d'eau initiaux (SWM4-DP) ont été réoptimisés en SWM4-NDP (charge négative sur la particule de Drude), calibrés pour reproduire les propriétés expérimentales comme l'enthalpie de vaporisation, la densité et la constante diélectrique [74].
- *Paramétrisation* : Les paramètres électrostatiques sont déterminés sur la base d'un ajustement contraint aux cartes de potentiel électrostatique (ESP) QM perturbées [74]. Les défis liés au transfert des paramètres de petits composés modèles aux polypeptides

ont conduit à des protocoles d'optimisation sophistiqués, y compris le recuit simulé (SA) incorporant des observables QM pour les polypeptides plus grands [74].

- *Performance* : Les modèles de Drude nécessitent généralement un pas de temps d'intégration de 1 fs et entraînent un surcoût d'environ 2 fois pour les calculs de polarisation, ce qui les rend environ 4 fois plus lents que les simulations additives avec un pas de temps de 2 fs [74]. Cependant, l'implémentation NAMD est hautement parallélisable [74].
- *Considérations* : Plus sensibles aux conditions initiales et peuvent subir des "catastrophes de polarisation", suggérant une équilibration initiale avec un champ de force additif avant la conversion [74].

Les chercheurs sont confrontés à un compromis fondamental entre la précision de la simulation et le coût computationnel lors du choix d'un champ de force. Alors que les champs de force polarisables offrent une description plus réaliste physiquement des électrostatiques (cruciale pour la liaison), leurs exigences computationnelles plus élevées signifient que les champs de force additifs, malgré leurs limites, restent très pertinents pour les systèmes plus grands ou les échelles de temps plus longues où les ressources computationnelles sont une contrainte. La complexité et la sensibilité des champs de force polarisables nécessitent des protocoles de configuration minutieux. Cela représente un piège potentiel pour les utilisateurs moins expérimentés et souligne que l'obtention d'une plus grande précision s'accompagne souvent d'une rigueur méthodologique accrue.

2.3.2.3.6.1 AMOEBA Force Field AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) est un autre champ de force polarisable classique, dont le développement est en cours depuis 1995 [74].

- *Modèle électrostatique* : Modélise l'énergie électrostatique en utilisant des contributions permanentes (interactions multipôle-multipôle atomiques jusqu'au quadrupôle) et induites (dipôles induits générés itérativement) [74].
- *Couverture* : Initialement développé pour l'eau, il a été étendu aux ions et à diverses molécules organiques, montrant de bonnes performances dans les calculs d'énergie libre d'hydratation [74].
- *Paramétrisation des protéines* : La dérivation des paramètres électrostatiques implique la détermination des multipôles atomiques permanents pour les résidus basés sur des dipeptides coiffés, en les optimisant par rapport au potentiel électrostatique QM de haut niveau [74].

2.3.2.4 Critères de sélection et précision des champs de force

Le choix du champ de force a un impact sur la précision des simulations de DM. Alors que les champs de force additifs sont robustes, les champs de force polarisables offrent une représentation plus précise des interactions électrostatiques, qui sont cruciales dans les systèmes biologiques [74]. Les modèles polarisables montrent des moments dipolaires systématiquement plus grands pour les liaisons peptidiques par rapport aux modèles additifs, ce qui suggère que

les modèles additifs pourraient ne pas surestimer suffisamment les moments dipolaires moléculaires pour l'environnement protéique [74]. La variabilité des dipôles dans le modèle de Drude suggère des forces physiques sous-jacentes différentes par rapport aux modèles additifs [74]. Le raffinement continu des champs de force additifs (par exemple, CHARMM36, AMBER ff99SB-ILDN) et polarisables vise à améliorer l'accord avec les données expérimentales (par exemple, les couplages J RMN) et les calculs de mécanique quantique, améliorant ainsi leur pouvoir prédictif pour les modes de liaison et les propriétés spectroscopiques [74, 75].

TABLEAU 2.2 – Comparaison des champs de force couramment utilisés en DM protéine-ligand.

Champ de Force	Type	Caractéristiques Clés	Applications et Notes
AMBER (ff14SB, etc.)	Additif	Développé par le groupe Kollman. Améliorations continues sur les angles dièdres.	Simulations de protéines, acides nucléiques. Largement utilisé en biophysique computationnelle [64, 74, 77].
GAFF (General AMBER FF)	Additif	Compatible avec AMBER pour protéines. Paramètres pour molécules organiques et pharmaceutiques (H, C, N, O, S, P, halogènes). Utilise RESP/AM1-BCC pour les charges. GAFF2 est une version améliorée.	Essentiel pour la paramétrisation des ligands dans les systèmes protéine-ligand. Utilisé en conception de médicaments [66, 78, 79].
CHARMM (CHARMM36)	Additif	Développé par le groupe Karplus. Complet pour protéines, acides nucléiques, lipides, glucides. CGenFF pour la chimie médicinale. CMAP amélioré pour la chaîne principale.	Large application en biomolécules. C36 améliore le sampling conformationnel [64, 74, 80].
OPLS	Additif	Développé par le groupe Jorgensen. Optimisé pour les systèmes organiques et biologiques.	Couramment utilisé pour les systèmes organiques et biologiques [64].
Drude	Polarisable	Modélise la polarisation avec des particules Drude. Supporté par NAMD, OpenMM. Nécessite une étape d'équilibration additive initiale.	Représentation plus précise des interactions électrostatiques. Coût computationnel plus élevé (environ 4x plus lent que l'additif) [74].
AMOEBA	Polarisable	Modélise l'énergie électrostatique avec des multipôles atomiques permanents et des dipôles induits.	Amélioration de la précision des interactions électrostatiques. Développement continu depuis 1995 [74].

Une vue d'ensemble rapide et comparative des principaux champs de force utilisés en dynamique moléculaire protéine-ligand. Elle synthétise des informations clés sur le type, les caractéristiques et les applications de chaque champ de force, facilitant ainsi la comparaison des options disponibles. Pour un chercheur, elle permet d'identifier rapidement les spécificités de chaque modèle, comme l'importance de GAFF pour les ligands, et d'orienter les choix méthodologiques en fonction des besoins spécifiques de la simulation, par exemple en privilégiant la précision électrostatique avec les champs de force polarisables ou la vitesse avec les champs de force additifs. Elle sert de référence rapide pour la planification et l'exécution de nouvelles simulations.

2.3.2.5 Logiciels pour simulation de dynamique moléculaire protéine-ligand

2.3.2.5.1 Principaux logiciels de simulation Une variété de progiciels puissants et polyvalents sont disponibles pour effectuer des simulations de DM de systèmes protéine-ligand. Ces outils exploitent souvent les ressources de calcul haute performance, y compris les GPU, pour accélérer les calculs.

AMBER : AMBER est une suite complète de programmes de simulation biomoléculaire largement utilisée en chimie computationnelle et en biophysique [82]. Il fournit des outils pour les simulations de DM haute performance, la minimisation d'énergie et les calculs d'énergie libre [82].

- **Fonctionnalités** : Inclut "pmemd" pour les simulations de DM (avec support GPU pour les champs de force avancés comme AMOEBA et les calculs d'énergie libre), "sander" pour la fonctionnalité QM/MM, et Antechamber pour la création de champs de force pour les molécules organiques générales [77, 82]. La suite prend en charge l'accélération GPU et le calcul parallèle [77, 82].
- **Applications** : Largement utilisé dans l'industrie pharmaceutique pour l'identification de cibles, la découverte de molécules candidates et l'optimisation de molécules candidates [77].

GROMACS : GROMACS est un progiciel de DM polyvalent et haute performance principalement conçu pour les simulations de protéines, de lipides et d'acides nucléiques [83]. Il est reconnu pour ses performances exceptionnelles grâce à des optimisations algorithmiques et à l'utilisation intensive des intrinsèques SIMD pour les CPU et de CUDA, OpenCL et SYCL pour les GPU [83].

- **Fonctionnalités** : GROMACS peut utiliser simultanément les CPU et les GPU disponibles dans un système [83]. Il est convivial, avec des fichiers de topologie et de paramètres en texte clair, et offre de nombreux contrôles de cohérence et des messages d'erreur clairs [83]. Il n'y a pas de script requis, tous les programmes utilisant une interface simple avec des options de ligne de commande [83]. Il prend en charge la compression des fichiers et propose une large sélection d'outils flexibles pour l'analyse des trajectoires [83].
- **Parallélisation** : Peut être exécuté en parallèle, en utilisant le protocole de communication MPI standard ou via sa propre bibliothèque "Thread MPI" pour les postes de travail à nœud unique [83].
- **Constructeur de topologie** : Le progiciel comprend un constructeur de topologie entièrement automatisé pour les protéines, même les structures multimériques, avec des blocs de construction disponibles pour les 20 résidus d'acides aminés standard, les nucléotides, les sucres et les lipides, ainsi que des groupes spéciaux [83].

CHARMM : CHARMM est un programme polyvalent de modélisation et de simulation moléculaire conçu pour offrir de vastes capacités de simulation de systèmes à plusieurs particules [80]. Il inclut un ensemble complet de champs de force pour simuler les biomolécules,

telles que les protéines, les acides nucléiques, les glucides, les lipides, leurs assemblages et les petites molécules [80].

- **Plateformes et Accélération GPU** : Conçu pour fonctionner sur une variété de plateformes compatibles UNIX, des ordinateurs parallèles haute performance avec GPU intégrés aux stations de travail plus petites basées sur GPU [80]. CHARMM prend en charge trois moteurs accélérés par GPU : BLaDE et DOMDEC (développés en interne) et OpenMM [80]. Tous prennent en charge la DM avec différents ensembles de contraintes et d'intégrateurs, et BLaDE et DOMDEC prennent en charge les simulations de dynamique λ pour les calculs d'énergie libre [80].
- **Applications** : Cible principalement les systèmes biologiques et trouve également de larges applications pour les matériaux inorganiques dans la conception de matériaux [81]. Il contient un ensemble complet d'outils d'analyse et de construction de modèles [81].

NAMD : NAMD est un code de DM parallèle conçu pour la simulation haute performance de grands systèmes biomoléculaires [84, 85]. Basé sur les objets parallèles Charm++, NAMD s'adapte à des centaines de cœurs pour les simulations typiques et à plus de 500 000 cœurs pour les plus grandes simulations [84, 85].

- **Intégration et Compatibilité** : NAMD utilise le populaire programme de graphiques moléculaires VMD pour la configuration de la simulation et l'analyse des trajectoires, mais est également compatible avec les fichiers AMBER, CHARMM et X-PLOR [84, 85].
- **Support GPU** : Les versions récentes (NAMD 3.0.1) offrent un nouveau mode résident GPU pour NVIDIA et les GPU AMD compatibles, ainsi qu'un support d'offload GPU pour les séries Intel GPU Max [84, 85].
- **Distribution** : Distribué gratuitement avec le code source [84, 85].

Desmond : Desmond est un moteur de DM haute performance alimenté par GPU pour simuler des systèmes biologiques tels que de petites protéines, des capsides virales, des complexes protéine-ligand, de petites molécules dans des solvants mixtes, des solides organiques et des complexes macromoléculaires synthétiques [86, 87].

- **Performance et Précision** : Offre un débit exceptionnel sur les clusters Linux avec des réseaux typiques et haut de gamme, améliorant la vitesse de calcul de 100 fois sur GPU par rapport à un seul CPU [86, 87]. Il est construit avec un accent sur la précision numérique, la stabilité et la rigueur, permettant la simulation de caractéristiques à grande échelle (nanomètres à microns) sur des échelles de temps de picosecondes à microsecondes [86, 87].
- **Fonctionnalités** : Fournit un cadre robuste pour le calcul des énergies et des forces pour les modèles de champ de force atomistiques, compatible avec les chimies couramment utilisées dans la recherche biomoléculaire [86, 87]. Prend en charge la configuration automatisée de simulations, y compris les simulations de DM multi-étapes avec des

protocoles intégrés [86].

- **Interface et Analyse :** Une interface intuitive permet une configuration rapide des expériences computationnelles [86]. Les résultats peuvent être visualisés et examinés dans le même environnement de modélisation Maestro, qui se connecte à une suite complète d'outils de modélisation [86, 87].

OpenMM : OpenMM est une boîte à outils logicielle moderne et open-source conçue pour les simulations de DM [73]. Elle se distingue par son accent unique sur l'extensibilité, permettant aux utilisateurs d'ajouter facilement de nouvelles fonctionnalités telles que des forces avec de nouvelles formes fonctionnelles, de nouveaux algorithmes d'intégration et de nouveaux protocoles de simulation [73].

- **Extensibilité et Accélération GPU :** Ces fonctionnalités sont conçues pour fonctionner automatiquement sur tous les types de matériel pris en charge (CPU et GPU) avec de bonnes performances, souvent avec un codage minimal [73]. OpenMM lui-même n'a pas besoin de modification pour ces nouvelles fonctionnalités, et elles peuvent être distribuées indépendamment [73].
- **Limitation et Solution :** Historiquement, OpenMM manquait d'un module dédié à la construction de systèmes récepteur-ligand, ce qui limitait son utilisation dans la découverte de médicaments [73]. Cette lacune est comblée par OpenMMDL, une boîte à outils open-source qui permet la préparation et la simulation de complexes protéine-ligand dans OpenMM, ainsi que l'analyse ultérieure des interactions protéine-ligand [73].

2.3.2.6 Outils spécialisés et plateformes intégrées

Au-delà des progiciels de simulation généraux, des outils spécialisés et des plateformes intégrées ont été développés pour simplifier et automatiser des aspects spécifiques des simulations de DM protéine-ligand, rendant ces techniques plus accessibles et efficaces.

Enlighten2 : Enlighten2 est un outil conçu pour rendre les simulations de dynamique moléculaire de systèmes protéine-ligand accessibles aux utilisateurs non-experts [66]. Il se présente sous la forme d'un package Python et d'un plugin convivial pour le programme de visualisation populaire PyMOL [66].

- **Fonctionnalités Clés :** Enlighten2 offre la paramétrisation automatique des ligands (via Antechamber, supportant le champ de force GAFF), la relaxation du système pour éliminer les chocs initiaux (artefacts), et l'exécution de simulations de DM de longueur désirée [66]. Il permet également d'ajouter une sphère de solvation de taille et de position arbitraires et de définir une valeur de pH spécifique pour la simulation [66].
- **Accessibilité :** Il n'est pas nécessaire d'apprendre de nouveaux programmes, et tous les outils sous-jacents sont gratuits et open source [66]. Il peut automatiquement récupérer une image Docker incluant Enlighten2 et toutes les utilités requises, évitant ainsi à l'utilisateur d'installer manuellement un progiciel de DM [66]. Cela permet aux expérimentateurs d'effectuer facilement des simulations de DM de base sans nécessiter une connaissance approfondie des logiciels de DM, ce qui est particulièrement utile

pour l'enseignement en classe ou les ateliers en ligne [66].

- **Limitations et Potentiel Avancé** : Enlighten2 ne permet pas l'étude des grands changements conformationnels ou de la dissociation complète du ligand, et sa paramétrisation des ligands repose sur Antechamber, limitant le support aux chimies de ligands couvertes par GAFF [66]. Cependant, les utilisateurs expérimentés peuvent modifier les fichiers intermédiaires pour des configurations plus avancées (par exemple, simulations avec conditions aux limites périodiques) [66].

Moira : Moira (molecular dynamics trajectory analysis) est un cadre qui automatise l'ensemble du processus, allant du docking aux simulations de DM et à diverses analyses et visualisations [67]. Il vise à résoudre le problème du manque d'exploration systématique de l'évolution des trajectoires de DM et des informations potentiellement omises [67].

- **Capacités** : Moira intègre le docking moléculaire, la DM et diverses méthodes d'analyse, permettant le traitement par lots d'un grand nombre de systèmes [67]. Il permet d'analyser les simulations de DM en termes de caractéristiques géométriques (RMSD, Protein-Ligand Interaction Profiler) et énergétiques (MM-PBSA) [67].

MDFit : MDFit est un flux de travail (framework) de simulations moléculaires automatisé qui permet une évaluation à haut débit de la dynamique ligand-protéine [88].

OpenMMDL : OpenMMDL est une boîte à outils open-source qui étend OpenMM en se concentrant sur la préparation efficace des complexes protéine-ligand pour la simulation [73].

- **Composants** : Il se compose de trois parties principales : OpenMMDL Setup (une interface utilisateur graphique pour préparer les paramètres de protéine et de simulation), OpenMMDL Simulation (pour effectuer les simulations de DM et le post-traitement des trajectoires), et OpenMMDL Analysis (pour analyser les résultats de simulation) [73].
- **Analyse Avancée** : OpenMMDL est un outil polyvalent pour analyser les interactions protéine-ligand et générer des modes de liaison des ligands tout au long des simulations. Il suit et regroupe également les molécules d'eau, en particulier celles qui présentent un déplacement minimal, fournissant des informations sur la dynamique du solvant [73].

CHAPERONg : CHAPERONg est un outil pour les simulations de DM automatisées basées sur GROMACS et les analyses de trajectoires, particulièrement utile pour la technique d'échantillonnage parapluie (Umbrella Sampling) [89]. Il peut exécuter des pipelines de simulation GROMACS complets sans dépendances supplémentaires installées [89].

2.3.2.7 Analyse des trajectoires et calculs d'énergie libre

L'analyse des simulations de DM implique le calcul de diverses propriétés pour comprendre les interactions protéine-ligand [64].

2.3.2.7.1 Techniques d'analyse courantes

1. **Déviati on Quadrati que Moyenne (RMSD) :** Une mesure de la déviati on du complexe protéine-ligand par rapport à sa structure initiale [64]. Le RMSD est une caractéristique typique capable d'estimer le mouvement de chaque atome [67].
2. **Fluctuation Quadrati que Moyenne (RMSF) :** Une mesure de la flexibilité de la protéine et du ligand [64].
3. **Énergie Libre de Liaison :** Une mesure de l'énergie associée à la liaison du ligand à la protéine [64].

2.3.2.7.2 Calculs d'énergie libre avancés Plusieurs techniques d'analyse peuvent être utilisées pour comprendre les interactions protéine-ligand à partir des simulations de DM :

1. **Calculs d'énergie libre de liaison (MM-PBSA/MM-GBSA) :** Des techniques telles que l'aire de surface de Poisson-Boltzmann de mécanique moléculaire (MM-PBSA) et l'aire de surface de Born généralisée de mécanique moléculaire (MM-GBSA) peuvent être utilisées pour estimer l'énergie libre de liaison du ligand à la protéine [64].
2. **Perturbation de l'Énergie Libre (FEP) :** La FEP est une technique qui peut être utilisée pour estimer l'énergie libre de liaison des complexes protéine-ligand [64, 70]. Elle est reconnue pour sa précision, en particulier pour les énergies de liaison relatives entre des ligands similaires [65]. Les protocoles FEP avancés peuvent combiner la DM tout-atome, l'échantillonnage amélioré (REST) et des contraintes sphériques pour calculer l'énergie libre de liaison de ligands se liant de manière diffusive [70]. FEP+ est une technologie propriétaire de Schrödinger qui prédit la liaison protéine-ligand avec une précision comparable aux méthodes expérimentales, permettant d'explorer un vaste espace chimique et d'optimiser plusieurs propriétés simultanément [71].
3. **Échantillonnage Parapluie (Umbrella Sampling) :** L'échantillonnage parapluie est une technique qui peut être utilisée pour étudier la cinétique de liaison et de déliaison [64]. C'est une méthode puissante pour calculer le potentiel de force moyenne (PMF) le long d'une coordonnée de réaction [89]. La génération d'une topologie moléculaire pour une simulation d'échantillonnage parapluie est similaire à toute autre simulation de DM [89].
4. **Analyse des liaisons hydrogène :** L'analyse des motifs de liaisons hydrogène entre la protéine et le ligand peut fournir des informations sur le mécanisme de liaison [64].
5. **PCA :** Peut être utilisée pour identifier les modes de mouvement dominants dans le complexe protéine-ligand [64].

2.4 Conclusion

En conclusion, le ML est un moteur d'innovation en QSAR, offrant des capacités de prédiction sans précédent via la régression et la classification, et une compréhension approfondie des

données moléculaires grâce à l'apprentissage non supervisé. Ces applications combinées accélèrent considérablement le processus itératif de découverte et de développement de médicaments, menant à des traitements plus efficaces et plus sûrs.

Concernant le docking moléculaire, malgré les avancées significatives, des défis persistants demeurent. L'estimation précise de l'affinité de liaison, la modélisation réaliste de la flexibilité des molécules (ligand et récepteur) et le traitement adéquat du solvant restent des domaines de recherche actifs. Un compromis inhérent existe entre la vitesse computationnelle, essentielle pour le criblage de grandes bibliothèques, et la précision requise pour des prédictions fiables. Aucune fonction de score ou algorithme de recherche unique n'excelle dans tous les aspects, ce qui conduit à la nécessité d'approches nuancées [58,60].

L'avenir du docking moléculaire s'oriente vers des approches hybrides et des pipelines multi-étapes qui exploitent les forces de différentes méthodes à des stades variés du processus de criblage. L'intégration croissante de l'apprentissage automatique dans le développement des fonctions de score offre des perspectives prometteuses pour améliorer la précision prédictive, même si des défis subsistent en termes d'interprétabilité. La diversité des logiciels et plateformes disponibles souligne l'importance pour les chercheurs de comprendre les forces et les faiblesses de chaque outil afin de faire un choix éclairé, adapté au contexte spécifique de leur étude et aux ressources disponibles.

En dépit de ces défis, le docking moléculaire demeure un pilier essentiel pour accélérer la découverte de médicaments en permettant le criblage virtuel, l'optimisation des leads et l'élucidation des mécanismes biologiques au niveau atomique. Son rôle continuera de croître à mesure que les capacités computationnelles augmentent et que les méthodologies s'affinent, promettant des avancées encore plus significatives dans la conception rationnelle de nouvelles thérapies.

Enfin, pour les simulations de DM protéine-ligand, elles se sont imposées comme un pilier indispensable de la biophysique computationnelle et de la découverte de médicaments. En offrant une vue atomique et dynamique des interactions, elles comblent la lacune des données expérimentales statiques, permettant une compréhension approfondie des mécanismes biologiques et des processus de liaison. L'évolution des champs de force, des modèles additifs robustes comme AMBER et CHARMM aux champs polarisables plus précis comme Drude et AMOEBA, reflète une quête continue de fidélité à la réalité physique, bien que cela s'accompagne de compromis en termes de coût computationnel. La compatibilité entre les champs de force protéiques et ligandiques, illustrée par l'importance de GAFF, est cruciale pour la fiabilité des simulations.

Parallèlement, le paysage logiciel a connu une innovation remarquable, avec des plateformes polyvalentes telles qu'AMBER, GROMACS, CHARMM, NAMD, Desmond et OpenMM, complétées par des outils spécialisés comme Enlighten2, Moira, MDFit, OpenMMDL et CHAPERONg. Ces avancées ont démocratisé l'accès aux simulations de DM, tout en permettant des analyses sophistiquées des trajectoires et des calculs d'énergie libre, essentiels pour la prédiction de l'affinité et de la cinétique de liaison. La transition d'un accent purement thermodynamique vers une considération accrue de la cinétique de liaison, comme le temps de résidence, souligne l'évolution de la DM vers des prédictions plus corrélées à l'efficacité *in vivo*.

En définitive, les simulations de DM ne sont pas de simples outils d'analyse, mais des générateurs d'hypothèses et des guides pour la conception expérimentale, accélérant le cycle de découverte scientifique. L'intégration continue de nouvelles méthodologies, l'amélioration des algorithmes et l'exploitation des architectures de calcul haute performance promettent de repousser encore les limites de ce que nous pouvons comprendre et concevoir au niveau moléculaire.

Partie II : Contributions et résultats

Chapitre 3

Modelisation des Relations Structure-Activité des Inhibiteurs de Kinases

3.1 Introduction

L'activité anormale de nombreuses tyrosine kinases joue un rôle majeur dans le cancer [90]. Elle est impliquée dans les caractéristiques fondamentales du cancer, notamment le cycle cellulaire, la survie, le métabolisme, la motilité et l'instabilité génomique [91]. Le développement de nouveaux médicaments contre le cancer est un processus complexe et chronophage [92], et même lorsqu'un traitement potentiel est identifié, il peut falloir des années pour le mettre sur le marché [93], d'autant plus que les méthodes expérimentales en conception de médicaments sont coûteuses et ne garantissent pas toujours des résultats positifs [94]. Dans ce contexte, se concentrer sur l'utilisation de techniques modernes de modélisation, avec les meilleures pratiques, constitue une voie prometteuse pour relever à la fois les défis thérapeutiques et économiques. Cela est devenu d'une grande importance dans la recherche et le développement des traitements contre le cancer [94–96].

Bien que la majorité des études de relations quantitatives/qualitatives structure-activité (QSAR) soient menées en utilisant un petit nombre de descripteurs, de nombreuses études récentes se sont basées sur un grand nombre [44, 97]. Un ensemble réduit de descripteurs permet généralement une modélisation orientée vers un criblage virtuel basé sur un pharmacophore, ou simplement de disséquer les descripteurs les plus corrélés à l'activité [98, 99]. Cependant, lors de la construction de modèles pour un criblage de grandes bases de données, un grand nombre de variables est un meilleur choix. À cette fin, il existe plusieurs approches. L'utilisation d'empreintes digitales chimiques est une stratégie bien connue [100], en plus de l'usage de différents types de descripteurs moléculaires [101]. L'approche à grand nombre permet un meilleur apprentissage, notamment sur les plus petits détails moléculaires. Une réduction de dimensionnalité prononcée, via différentes méthodes (par exemple, comme rapporté par Almkadi et al. [102, 103]), est en effet une approche efficace, mais elle peut entraîner la suppression de détails moléculaires, qui peuvent ne pas sembler pertinents mathématiquement au premier abord, mais qui régissent potentiellement des phénomènes biologiques complexes. De plus, lors du travail sur des jeux de données larges et diversifiés, avec un grand nombre de descripteurs, il est possible de diminuer le besoin en domaines d'applicabilité [98], et donc de pouvoir explorer différentes possibilités structurelles.

De nombreuses difficultés peuvent être rencontrées lors de la modélisation sur de grandes quantités de données. Les problèmes de sur/sous-apprentissage sont fréquemment rencontrés lors de la construction de ce type de modèles QSAR [44]. En 2020, X. Yu et al. [104] et Q. Liu et al. [105] ont construit des modèles QSAR avec peu de descripteurs moléculaires. Les deux études ont rapporté qu'un plus grand ensemble de descripteurs pourrait compliquer la modélisation et diminuer sa qualité statistique. Pour faire face à ce problème, le choix de l'algorithme d'apprentissage adéquat et l'utilisation de bonnes pratiques de paramétrage pourraient constituer une solution. Cela permettrait ainsi de diminuer le besoin en domaines d'applicabilité, et de faire passer au crible de grandes quantités de composés chimiques.

La Forêt Aléatoire (Random Forest, RF) et la Machine à Vecteurs de Support (Support Vector Machine, SVM) comptent parmi les méthodes d'arbres de décision les plus populaires, avec des caractéristiques et des applications distinctes. Ces algorithmes sont fréquemment utilisés en QSAR, avec des résultats relativement satisfaisants [44]. Le RF est une technique d'apprentissage par ensemble qui construit plusieurs arbres de décision et combine leurs résultats afin d'améliorer la précision des prédictions et de limiter le surapprentissage. En revanche, le SVM est une méthode d'apprentissage supervisé qui cherche à trouver l'hyperplan optimal séparant les différentes classes, ce qui la rend particulièrement efficace dans les problèmes de classification à forte dimensionnalité.

De son côté, XGBoost (de son nom en anglais : Extreme Gradient Boosting) est un algorithme puissant pour l'apprentissage supervisé en ML [106]. Il présente divers avantages en modélisation QSAR, notamment le fait qu'il ne demande pas beaucoup en ressources computationnelles [107]. XGBoost capture efficacement les interactions entre les variables et peut gérer à la fois les relations linéaires et non linéaires dans les données [108]. Malgré sa puissance, l'algorithme a été timidement utilisé pour les études QSAR quantitatives et rarement appliqué aux études QSAR qualitatives pour la prédiction de l'inhibition des kinases [47, 91, 93, 109, 110]. Cependant, le classifieur souffre toujours du principal inconvénient des modèles basés sur des arbres de décision, à savoir la difficulté à résoudre des relations hautement non linéaires entre les variables [111].

Dans le contexte de l'apprentissage profond, les réseaux de neurones (NNs) sont conçus pour identifier des relations complexes et non linéaires en utilisant plusieurs couches de neurones qui apprennent progressivement à partir des données [112]. Cela permet aux NNs de découvrir des variables latentes et de représenter des caractéristiques très abstraites [113]. Cependant, sans un réglage minutieux et un travail important en ingénierie des variables, les NNs seuls ont souvent des difficultés à apprendre à partir de données tabulaires structurées [114]. Pour ce type de données, les approches basées sur des arbres de décision ont constamment donné de meilleurs résultats que les DNNs [115].

L'apprentissage par ensemble (de son nom en anglais : ensemble learning) est une technique bien connue qui améliore la performance en entraînant plusieurs modèles et en agrégeant leurs prédictions [116]. La classification par ensemble peut tirer profit de la diversité des classifieurs pour surmonter les limitations de chacun. De plus, il a été démontré empiriquement qu'un ensemble XGBoost/DNN performe mieux que tout modèle seul ou qu'un ensemble classique (sans apprentissage profond) [114]. En 2016, Heaton et al. [117] ont clarifié pourquoi les modèles en ensemble surpassent généralement les modèles individuels. Cela est dû à la

différence dans les variables synthétisées par les modèles basés sur les arbres de décision et les modèles NN. Ainsi, les ensembles combinant ces deux approches peuvent atteindre des performances très élevées.

Les stacked ensembles sont des modèles construits sur le principe de transmettre l'information d'un (ou de plusieurs) généraliseur(s) à un autre, jusqu'à l'obtention d'une estimation finale [118]. Une pratique fondamentale et largement utilisée des stack ensembles consiste à combiner les sorties de plusieurs modèles comme une entrée unique d'un modèle final [118]. Cette approche est principalement utilisée pour corriger les erreurs (biais) générées par un modèle unique ou par un ensemble de modèles [118]. L'approche est similaire au principe de vote [119] (à travers un ensemble de sorties de modèles), mais atteint généralement de meilleures performances [119]. La méthode a rencontré un succès considérable dans la prédiction de bioactivité et la conception de médicaments [120, 121]. Toutefois, dans certaines applications, il a été empiriquement démontré que les classifieurs empilés peuvent fonctionner aussi bien que la simple sélection du meilleur classifieur de l'ensemble [119]. Dans ce contexte, la force collaborative de XGBoost et des DNNs s'est révélée précieuse dans une variété de domaines d'utilisation.

Fait intéressant, XGBoost s'est avéré être l'algorithme de sélection de caractéristiques le plus performant dans une étude récente [122]. Il a été sélectionné parmi cinq autres algorithmes pour traiter les données en vue d'une modélisation hybride supplémentaire par DNN. En outre, XGBoost a été utilisé comme un outil efficace de sélection de caractéristiques via le score d'importance des variables (feature importance scoring) dans un travail pertinent présenté en 2020 [123]. Dans cette dernière étude, les caractéristiques présentant les scores les plus élevés selon le classifieur XGBoost ont été utilisées comme entrées d'un DNN, atteignant une précision de 97%. En 2022, Zivkovic *et al.* ont appliqué un réseau de neurones convolutionnel (CNN) à quatre couches pour extraire des caractéristiques à partir d'images [124]. Ces caractéristiques ont été utilisées comme entrées pour XGBoost, qui a joué le rôle de classifieur intermédiaire dans un modèle hybride à trois composants. Bien que le succès de l'approche dépende du troisième algorithme, la haute précision a été principalement obtenue par l'association CNN/XGBoost, conduisant à une catégorisation efficace des points de données. La même année, Zhao *et al.* ont employé avec succès une approche hybride XGBoost/DNN pour une application liée à la sécurité routière [125]. Les auteurs ont appliqué une transformation des caractéristiques sur l'ensemble des nœuds terminaux du modèle de gradient boosting, et ont utilisé l'ensemble affiné comme entrée d'un DNN. Le modèle hybride résultant a atteint un score de 91.03%.

En apprentissage automatique, y compris en apprentissage profond, l'ingénierie des variables est une tâche essentielle. La performance d'un modèle dépend souvent de la qualité de l'ingénierie des variables [126]. L'ingénierie des variables consiste à prétraiter les données d'entrée en créant de nouvelles variables, telles que des différences, des rapports ou d'autres transformations mathématiques sur les variables existantes [117]. Compte tenu de la force collaborative de XGBoost et des DNNs, ainsi que de l'importance de l'apprentissage par ensemble, l'idée d'une approche combinant les deux pour obtenir des QSARs plus efficaces est très attrayante.

Dans ce chapitre, nous présentons une architecture hybride spécifique d'apprentissage automatique qui exploite l'ingénierie des variables avec XGBoost et les DNNs, afin d'améliorer les

relations structure-activité. Nous présentons d'abord une étude comparative effectuée sur des données de PIM kinases, entre les trois algorithmes d'apprentissage précédemment cités, qui nous a guidée vers le choix de XGBoost comme algorithme de base pour l'apprentissage par ensemble. Ensuite, nous énonçons l'approche en question. Cette approche repose en premier lieu sur la capacité de XGBoost à capturer des motifs et des interactions complexes dans les données. Cette capacité est utilisée pour générer des probabilités prédictives. Celles-ci sont adoptées pour construire un ensemble de variables raffinées. Ces variables sont intégrées comme entrées supplémentaires dans un DNN pour la classification finale. Le réseau de neurones, grâce à sa capacité à apprendre des représentations hiérarchiques, affine davantage les relations au sein du nouveau jeu de données. L'approche est évaluée sur 40 grands jeux de données d'IK (40 protéines kinases), montrant une précision et une robustesse supérieures par rapport aux modèles XGBoost seuls ainsi qu'aux modèles standards hautement performants. Nos résultats suggèrent que la combinaison des sorties probabilistes de XGBoost avec un DNN améliore la performance des modèles QSAR.

3.2 Méthodologie

3.2.1 Bases de données

Les données brutes sur les bioactivités expérimentales relatives à chaque cible protéique ont été collectées à partir de la base de données ChEMBL [127]. Il s'agit d'une grande base de données en libre accès de composés chimiques bioactifs, incluant des améliorations importantes telles qu'un enregistrement robuste et une représentation détaillée des essais expérimentaux, avec un nouveau système de dépôt de données permettant la mise à jour continue des jeux de données. De plus, elle contient des structures 2D, des propriétés calculées (par exemple, logP, Poids Moléculaire, Paramètres de Lipinski, etc.) et des paramètres de bioactivité condensés (par exemple, constantes de liaison, pharmacologie et données ADMET) [127].

Les données collectées consistent en des fichiers CSV contenant des informations détaillées sur chaque composé étudié, incluant, sans s'y limiter : un identifiant, un nom de molécule (commercial) si connu, le poids moléculaire, les violations des règles de Lipinski, le coefficient de partage et le code SMILES.

3.2.2 Traitement mathématique et mise à l'échelle

Le traitement des données et l'analyse mathématique ont été réalisés en Python, en utilisant pandas [128] et NumPy [129], et les descripteurs moléculaires ont été calculés à l'aide du logiciel PaDEL [130].

Le processus de nettoyage et de préparation implique d'abord la suppression des doublons/conformères et une vérification de validité structurale. Ensuite, l'élimination de tous les composés pour lesquels la valeur de IC_{50} (concentration inhibitrice médiane) est inexacte (incertaine). Ces composés ne possèdent pas le symbole "=" dans la colonne 'Standard Relation' (relation entre le IC_{50} rapporté et le réel). Afin de considérer davantage de composés similaires aux inactifs, seuls les composés pour lesquels la valeur inexacte de IC_{50} est considérablement supérieure à 10000 sont conservés.

Pour éviter toute perte d'information dans les étapes suivantes, les composés présentant des valeurs manquantes dans les colonnes suivantes : 'Molecule ID', 'SMILES', 'Standard Relation', 'Standard Value' (IC_{50}), 'Standard Units' (l'unité de IC_{50}), 'Document Journal' (l'article rapportant la bioactivité) et 'Document Year' (année de publication) ont également été exclus. Tous les IC_{50} exprimés en μM et $\mu g/mL$ ont été convertis en nM. Les composés actifs ont ainsi été regroupés dans la classe positive (1), et les inactifs dans la classe négative (0).

Les descripteurs moléculaires unidimensionnels et bidimensionnels (1D & 2D) ont été générés, et les colonnes contenant des valeurs constantes ont été supprimées. Le coefficient de corrélation de Pearson (r) [131] a été calculé entre chaque paire de colonnes (équation 3.1), sur la base duquel une seule colonne est conservée si la corrélation $> 90\%$.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.1)$$

Où :

x_i = valeurs de la variable x dans un échantillon

$\bar{x}$ = moyenne des valeurs de la variable x

y_i = valeurs de la variable y dans un échantillon

$\bar{y}$ = moyenne des valeurs de la variable y

Une procédure de mise à l'échelle a été appliquée à chaque jeu de données, afin d'améliorer la performance des classifications, comme rapporté par Singh et al. [132]. Pour chaque descripteur, nous avons pris la valeur maximale et minimale, et remplacé chaque valeur X par une valeur X' mise à l'échelle, calculée à l'aide de l'équation 3.2.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3.2)$$

Où :

X' = valeur normalisée de la variable

X = valeur réelle de la variable

X_{min} = valeur minimale dans la colonne du descripteur

X_{max} = valeur maximale dans la colonne du descripteur

3.2.3 Intégration de XGBoost avec les Reseaux de Neurones Profonds (DNN)

La structure du modèle hybride proposé est illustrée dans la figure 3.2. Le travail peut être divisé en trois phases. (1) Dans le modèle final proposé, XGBoost [108] agit comme apprenant de base, produisant une prédiction initiale en améliorant de manière itérative ses performances sur des données structurées. Bien que ces prédictions soient efficaces pour capturer les relations globales entre les variables, elles peuvent laisser des résidus, représentant des erreurs ou des complexités non traitées. (2) Les probabilités prédites par XGBoost sont combinées avec l'ensemble de variables d'origine, formant un ensemble augmenté. (3) Le DNN est ensuite appliqué en tant que méta-apprenant pour analyser ces variables enrichies et découvrir des motifs cachés. Cette approche en trois étapes garantit que la performance prédictive du modèle repose à la fois sur l'interaction des variables structurées (gérée par XGBoost) et l'extraction hiérarchique de variables complexes (fournie par le DNN).

3.2.3.1 XGBoost

L'algorithme XGBoost [108] a été utilisé pour entraîner les modèles d'apprentissage automatique de base, tel qu'implémenté dans la bibliothèque scikit-learn de Python [133]. Ce choix a été fait après un travail de benchmark réalisé entre l'algorithme en question, Random Forest (RF) [134] et Support Vector Machines (SVM) [135]. Il s'agit d'une amélioration de l'algorithme de boost du gradient de Chen Tianqi. Il conserve la haute précision des méthodes de boost classiques tout en offrant un support flexible pour le calcul parallèle [136]. Malgré sa puissance, XGBoost est rarement utilisé dans les études QSAR (quantitatives/qualitatives) pour la prédiction de l'inhibition des kinases. Il est adopté dans ce travail comme classifieur de base pour le modèle hybride (équation 3.3).

Étant donné un vecteur de variables d'entraînement $x_i \in R^d$, le label de classe correspondante est $y_i \in \{0, 1\}$, ($i = 1, \dots, n$). Où x_i est un vecteur de descripteurs, y_i désigne la classe cible, et n est le nombre d'échantillons dans l'ensemble d'entraînement. La formule suivante décrit le modèle de prédiction XGBoost :

$$\hat{y}_i = F(x_i) = \sum_{p=0}^P f_p(x_i) \quad (3.3)$$

Où P est le nombre total d'arbres, $f_p(x_i)$ est le p -ième arbre, et $\hat{y}_i$ est la prédiction pour l'échantillon x_i . La fonction f_p est entraînée en optimisant la fonction objectif suivante :

$$Obj(\Theta) = \sum_{i=0}^n l(y_i, \hat{y}_i) + \sum_{p=1}^K \Omega(f_p) \quad (3.4)$$

$\Omega(f_p)$ représente le terme de régularisation du p -ième arbre, et $l(y_i, \hat{y}_i)$ est l'erreur d'entraînement de l'échantillon x_i . Pour chaque arbre de classification, le modèle peut s'écrire comme suit :

$$f(x) = \omega_{q(x)}, \omega \in R^T, q \in R^d \rightarrow \{1, 2, \dots, T\} \quad (3.5)$$

Où $q(x)$ est le noeud feuille correspondant à l'échantillon x et ω est le score du noeud feuille. T représente le nombre total de feuilles de l'arbre. Par conséquent, la complexité de l'arbre peut s'exprimer ainsi :

$$\Omega_f = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega^2(j) \quad (3.6)$$

Où λ est le coefficient de régularisation et γ est le paramètre de complexité.

XGBoost utilise le développement en série de Taylor d'ordre deux pour approximer la fonction objectif originale. Cette dernière peut être approximée comme indiqué dans la équation 3.7 :

$$Obj = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \left[G_j \omega_j + \frac{1}{2} (H_j + \lambda) \omega_j^2 \right] + \gamma T \quad (3.7)$$

La fonction objectif est exprimée comme T fonctions quadratiques indépendantes de variables uniques. Par conséquent, la équation 3.8 représente le score optimal $\omega^*(j)$ pour chaque feuille dans XGBoost, tandis que la équation 3.9 est la solution de la fonction objectif :

$$\omega^*(j) = \frac{G_j}{H_j + \lambda} \quad (3.8)$$

$$Obj = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (3.9)$$

Le score du noeud feuille n'a aucun impact sur la équation 3.9, qui ne dépend que de la structure $q(x)$ de l'arbre. Ainsi, une fois la structure de l'arbre déterminée, la fonction objectif correspondante peut être calculée.

XGBoost utilise une méthode heuristique pour déterminer la structure optimale de l'arbre, en employant un algorithme glouton exact et une formule de gain. La équation 3.10 définit la fonction de gain pour un arbre I pouvant être scindé en deux sous-arbres I_L et I_R , avec $I = I_L \cup I_R$:

$$Gain = -\frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (3.10)$$

γ représente le coût de complexité associé à l'ajout de nouveaux noeuds feuilles.

3.2.3.2 Ingénierie des variables

L'ingénierie des variables est l'aspect le plus crucial de l'apprentissage automatique, y compris l'apprentissage profond. Une ingénierie efficace des variables peut améliorer la performance du modèle.

Étant donné un vecteur de variables d'entraînement $x_i \in R^d$, $i = 1, \dots, n$, où x_i est un vecteur de descripteurs, et n est le nombre d'échantillons dans l'ensemble d'apprentissage. Soit $F_{XGBoost}(x_i)$ le modèle XGBoost entraîné qui produit des probabilités prédictives p pour x_i :

$$F_{XGBoost}(x_i) = p_i^{(0)}, p_i^{(1)} \quad (3.11)$$

Où $p_i^{(0)} + p_i^{(1)} = 1$

Le nouvel ensemble de variables enrichies $s_i \in R^{d+2}$ peut être exprimé en concaténant le vecteur original x_i avec les probabilités prédictives de XGBoost $F_{XGBoost}(x_i)$. Pour chaque échantillon i :

$$s_i = [x_i, F_{XGBoost}(x_i)] = [x_{i1}, x_{i2}, \dots, x_{id}, p_i^{(0)}, p_i^{(1)}] \quad (3.12)$$

3.2.3.3 Réseaux de Neurones Profonds (DNNs)

Un réseau de neurones artificiel comportant plus de trois couches est appelé réseau de neurones profond, également connu sous le nom de perceptron multicouche profond (MLP). Il peut exprimer la représentation distribuée des données d'entrée, approximer des fonctions complexes et présente une grande capacité à extraire des variables clés à partir d'un petit ensemble d'échantillons, grâce à sa structure profonde non linéaire [137]. De nos jours, l'architecture de l'apprentissage profond a démontré des performances exceptionnelles en classification dans un large éventail de problèmes de prise de décision. Un DNN typique se compose de trois types de couches principales : la couche d'entrée, les couches cachées et la couche de sortie. Dans ce chapitre de thèse, nous proposons une nouvelle architecture DNN, comme illustré dans la figure 3.1. Les couches sont entièrement connectées, ce qui signifie que chaque neurone de la i -ème couche est connecté à tous les neurones de la couche $(i + 1)$.

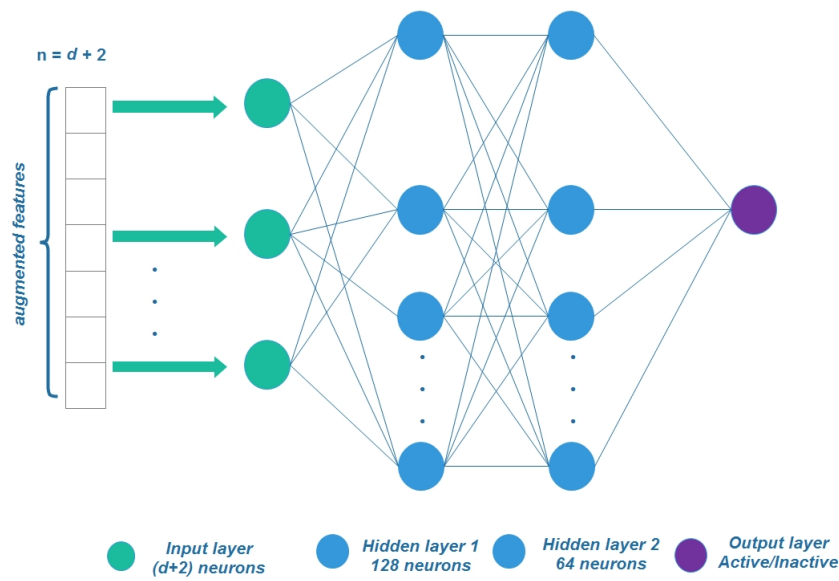


FIGURE 3.1 – Architecture du DNN.

Dans un modèle de DNN, chaque couche est composée d'une transformation linéaire suivie d'une fonction d'activation $\sigma(z)$. La transformation linéaire est définie comme suit :

$$z^l = \sum \omega_i^{(l)} x_i^{(l-1)} + b^{(l)} \quad (3.13)$$

où x_i décrit la valeur du i -ème neurone provenant de la couche précédente $(l - 1)$, ω_i représente le poids du i -ème neurone dans la couche actuelle l , et b est le biais.

Pour la première couche cachée, x_i représente les variables d'entrée dérivées de l'étape précédente. Ainsi, l'équation 3.13 peut être réécrite comme suit :

$$z^l = \sum \omega_i^{(l)} s_i^{(l-1)} + b^{(l)} \quad (3.14)$$

La rectification linéaire ReLU ($\sigma(z) = \max(0, z)$) a été utilisée comme fonction d'activation dans cette étude. Elle laisse z inchangé s'il est supérieur ou égal à 0, et le fixe à 0 s'il est négatif. Étant donné que la couche $(l - 1)$ comporte k neurones, et que la couche l contient p neurones, le poids ω de la couche l est une matrice $p \times k$ notée W^l , tandis que le biais b

de la couche l est un vecteur $p \times 1$ noté b^l . La sortie de la couche $(l - 1)$ est un vecteur a de dimension $k \times 1 : a^{(l-1)}$. Ainsi, la sortie de la couche l est :

$$a^l = \sigma(z^l) = \sigma(W^l a^{(l-1)} + b^l) \quad (3.15)$$

L'entropie croisée binaire a été utilisée comme fonction de perte dans nos DNNs. Elle mesure la divergence entre les valeurs prédites et les valeurs réelles de l'activité des composés selon l'équation 3.16, où $\mathcal{L}$ est la perte totale (loss) calculée sur le jeu de données, N est le nombre total d'échantillons, y_i est la valeur réelle de la classe pour l'échantillon i (elle vaut 0 ou 1) et $\hat{y}_i$ est la probabilité prédite par le modèle que l'échantillon i appartienne à la classe 1.

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (3.16)$$

L'algorithme d'optimisation Adam (Adaptive Moment Estimation) a été adopté pour améliorer l'efficacité de l'entraînement, grâce à ses performances de calcul [138], afin de mettre à jour de manière adaptative les paramètres d'apprentissage du DNN. La figure 3.2 présente l'architecture pratique du modèle hybride proposé.

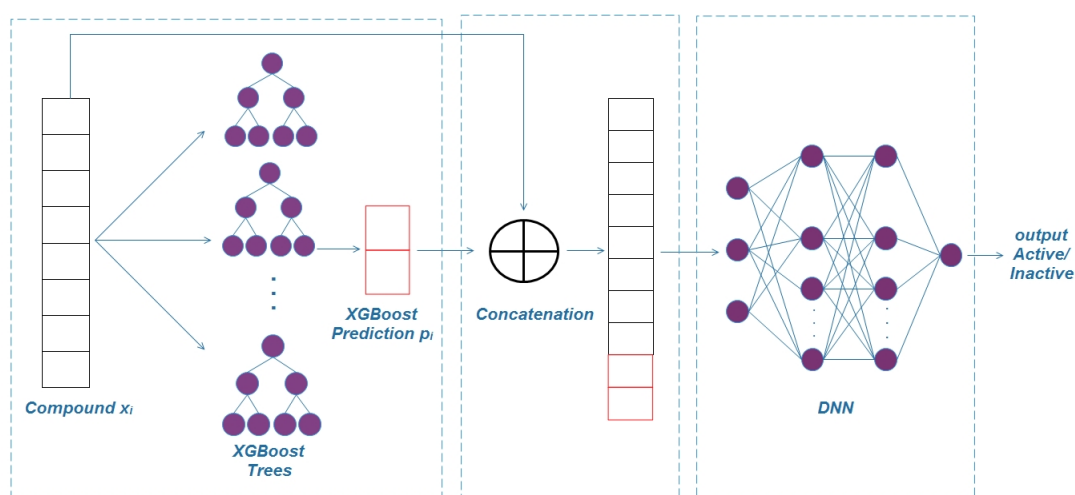


FIGURE 3.2 – Architecture pratique du modèle hybride proposé.

3.2.4 Métriques d'évaluation

Pour évaluer les modèles de classification, nous avons calculé la précision, la sensibilité, la spécificité et l'aire sous la courbe (AUC). Les vrais positifs (TP) sont les composés actifs correctement identifiés, les vrais négatifs (TN) sont les composés inactifs correctement identifiés, les faux positifs (FP) sont les composés actifs incorrectement classifiés, et les faux négatifs (FN) sont les composés inactifs incorrectement prédits. L'AUC évalue la performance du classifieur dans la distinction des classes (taux de vrais positifs (TPR) contre taux de faux positifs (FPR)). Ces métriques ont été calculées à l'aide de l'équation 3.17 pour la précision, l'équation 3.18 pour la sensibilité, et l'équation 3.19 pour la spécificité.

$$Score = \frac{TP + TN}{TP + FP + FN + TN} \quad (3.17)$$

$$Sens = \frac{TP}{TP + FN} \quad (3.18)$$

$$Spec = \frac{TN}{TN + FP} \quad (3.19)$$

TP=Vrai Positif, FP=Faux Positif, TN=Vrai Négatif, FN=Faux Négatif

3.3 Resultats et discussion

3.3.1 Analyse et prétraitement des données

L'étiquetage de l'activité est réalisé comme suit : les ligands ayant un IC_{50} inférieur à 200 nM sont considérés comme actifs, et ceux au-dessus de 1000 nM comme inactifs. Afin de construire une frontière d'activité bien séparée, les composés intermédiaires sont également considérés comme inactifs. Cela peut améliorer la précision lorsque les modèles seront utilisés pour le criblage virtuel.

La figure 3.3 montre la variation temporelle de la bioactivité dans les essais rapportés, pour les données EGFR. Le graphique illustre clairement la cohérence et la disponibilité des données, en particulier pour les composés de type inactif. La distribution de densité indique une tendance continue et régulière à rapporter des composés fortement et faiblement actifs. En examinant l'aspect global des données, on peut observer que pour le classifieur EGFR, l'algorithme peut mieux apprendre sur la classe des composés inactifs, ce qui permet une identification significative de composés de type inactif, dans un processus de criblage virtuel [139]. La distribution de densité indique également une augmentation du nombre de composés actifs rapportés au cours de la dernière décennie.

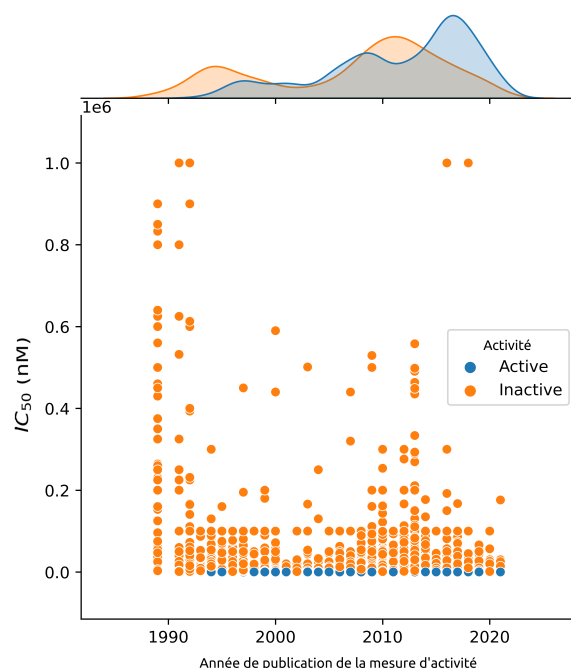


FIGURE 3.3 – Graphique des valeurs de IC_{50} selon l'année de publication des essais dans le jeu de données du récepteur EGF.

Après l'application des filtres de faible variance et de forte corrélation, nous supposons que les descripteurs restants contiennent la quantité maximale d'informations utiles. Par exemple, dans l'ensemble d'entraînement mTOR, les corrélations entre le nombre de donneurs de liaisons hydrogène et les descripteurs suivants sont : nombre d'atomes aromatiques : 0.6, nombre d'atomes d'azote : 0.41, nombre d'atomes d'oxygène : 0.03.

Pour évaluer la performance d'un modèle proposé sur un ensemble de données indépendant, les données ont été divisées en sous-ensembles d'entraînement et de test. Nous avons choisi une répartition aléatoire selon un ratio 75 :25. Les dimensions de chaque jeu de données sont présentées dans le tableau 3.4. Celles-ci sont exprimées au format $(M_{\text{Lignes}}, N_{\text{Colonnes}})$, où M représente le nombre de molécules et N le nombre de descripteurs du jeu de données.

3.3.2 Modélisation

3.3.2.1 Hyperparamètres et benchmark

Le benchmark a été effectué sur la base de combinaisons des plages d'hyperparamètres fréquemment utilisés pour les 3 algorithmes (XGBoost, RF et SVM) [47]. Le tableau 3.1 présente les résultats obtenus sur les données des 3 PIM kinases. Ces résultats suggèrent la supériorité de XGBoost sur ce type de données structurés.

TABLEAU 3.1 – Tableau récapitulatif des informations pertinentes sur les modèles construits pour benchmark.

Modèles	Algorithme	Dimensions train	Dimensions test	Sensibilité	Spécificité	AUC	Score
PIM1	RF	(1011, 501)	(336, 501)	0.79	0.9	0.85	0.85
	SVM			0.02	1	0.51	0.57
	XGBoost			0.75	0.9	0.83	0.84
PIM2	RF	(570, 478)	(190, 478)	0.78	0.86	0.82	0.83
	SVM			0.68	0.81	0.75	0.75
	XGBoost			0.79	0.89	0.84	0.85
PIM3	RF	(274, 461)	(91, 461)	0.38	0.93	0.66	0.78
	SVM			0	1	0.5	0.71
	XGBoost			0.58	0.92	0.75	0.82

XGBoost possède de nombreux paramètres ajustables permettant de l'adapter à des besoins spécifiques. Le taux d'apprentissage (*eta*) et la profondeur maximale des arbres (*max_depth*) sont des paramètres cruciaux qui affectent la précision et la complexité du modèle. D'autres paramètres importants incluent, sans s'y limiter, le poids minimum des ramifications (*min_child_weight*) et la fraction de colonnes utilisées par arbre (*colsample_bytree*). Le nombre d'arbres de décision (*n_estimators*) et la réduction minimale de perte (*gamma*) sont également des paramètres importants, pouvant améliorer les performances de l'algorithme. L'ajustement de ces paramètres est essentiel pour éviter le surapprentissage ou le sous-apprentissage.

La validation croisée est une technique utilisée en apprentissage automatique pour évaluer les performances d'un modèle et estimer son erreur de généralisation. Elle consiste à diviser les données en plusieurs sous-ensembles, et à entraîner et évaluer le modèle sur différentes combinaisons de ces blocs. Une validation croisée sur 10 est utilisée, les données sont alors divisées en 10 sous-ensembles de taille égale. Nous combinons cette technique avec différents hyperparamètres XGBoost, en utilisant la fonction *GridSearchCV* (de *scikit-learn*), qui permet d'expérimenter toutes les combinaisons possibles.

Chaque valeur possible d'un paramètre est testée en combinaison avec toutes les autres valeurs possibles (des paramètres restants). Le tableau 3.2 présente en détails la procédure de choix des paramètres optimaux d'XGBoost. En plus de l'intervalle dans lequel les expériences ont été menées pour chaque hyperparamètre XGBoost [140], une brève description de son implication est mentionnée.

TABLEAU 3.2 – Description et plages d'expérimentation pour chaque hyperparamètre.

Hyperparamètre	Description	Interval
<i>n_estimators</i>	Nombre de tours de boosting.	[10, 2000]
<i>max_depth</i>	Profondeur maximale de chaque arbre (complexité de l'arbre).	[3, 15]
<i>learning_rate</i>	Pas d'apprentissage de l'algorithme de boosting.	[0.01, 0.3]
<i>colsample_bytree</i>	Fraction de colonnes sélectionnées pour chaque arbre.	[0.2, 1]
<i>gamma</i>	Réduction de perte minimale requise pour effectuer une nouvelle division dans un nœud feuille.	[0, 2]
<i>min_child_weight</i>	Somme minimale des poids (hessien) nécessaires dans une ramification.	[1, 10]

La meilleure combinaison est sélectionnée en fonction du score de validation croisée, et de la précision sur l'ensemble externe.

D'après les expériences menées sur tous les ensembles de données étudiés, les hyperparamètres permettent d'atteindre une performance prédictive optimale à des valeurs spécifiques. L'alignement optimal est le suivant : max_depth est fixé à 12, learning_rate à 0.15, colsample_bytree à 0.4, gamma à 0.3, et min_child_weight à 1.

Étant donné que le nombre d'arbres de décision exprime la complexité de l'arbre requis pour un apprentissage efficace sur les données, sa meilleure valeur dépend de chaque ensemble de données. La taille, ainsi que la structure et la répartition spécifiques des descripteurs déterminent le nombre d'estimateurs requis.

Les observations indiquent que les valeurs par défaut sont optimales pour les paramètres subsample, reg_alpha (régularisation L1) et reg_lambda (régularisation L2).

3.3.2.2 Hyperparamètres du DNN

Pour développer le modèle DNN, de nombreuses expériences de configuration ont également été menées. En termes de performance et de complexité, la configuration optimale a été choisie comme décrit dans le tableau 3.3.

TABLEAU 3.3 – Paramètres du DNN.

Hyperparamètre	Description	Valeurs
Nombre de couches	Nombre total de couches d'entrée et de sortie dans un DNN.	4
Nombre de neurones par couche	Nombre de neurones dans chaque couche cachée.	[128,64]
Fonction d'activation	Non-linéarité appliquée au modèle.	ReLU
Optimiseur	Méthode d'optimisation utilisée pour ajuster les poids.	Adam
Taux d'apprentissage	Pas d'ajustement des poids lors de l'apprentissage.	0.001
Taille des lots	Nombre d'exemples d'apprentissage dans chaque lot.	32
Fonction de perte	Fonction de mesure de l'erreur de prédiction.	binary_crossentropy
Époques	Nombre de passages complets sur l'ensemble d'apprentissage.	100
Taux de dropout	Régularisation	[0.4,0.4]

3.3.3 Comparaison des performances des modèles hybrides XGBoost/DNN avec les modèles XGBoost seuls

Les modèles hybrides XGBoost/DNN et les modèles XGBoost seuls ont été évalués sur 40 ensembles de données d'inhibition de kinases distincts, chacun générant un modèle QSAR unique. Cette approche permet une évaluation complète à travers 40 tâches de classification différentes. Les résultats de la comparaison sont résumés dans le tableau 3.4, qui inclut les cibles thérapeutiques, les dimensions des ensembles de données, les scores de validation croisée, ainsi que les scores sur les ensembles de test. Les résultats montrent une amélioration significative des performances du modèle hybride par rapport au classifieur XGBoost seul, pour 39 kinases. À l'exception de JNK2, SYK et ROCK1, les précisions en test du modèle hybride dépassent 90% pour toutes les cibles. En comparaison, le modèle XGBoost seul n'atteint 90% que pour les cas INSR et IGF1R. Notamment, INSR est le seul ensemble de données pour lequel les deux modèles atteignent une précision équivalente. Malgré une amélioration de 5% sur la validation croisée, les deux atteignent 90% de bonnes prédictions en test.

Les deux seuls modèles XGBoost avec une précision inférieure à 80% (ROCK1 et Trk-A, affichant respectivement 75% et 78%) sont améliorés de 11% et 14%. Pour les 9 modèles XGBoost avec des scores compris entre 80% et 82%, l'amélioration est de 10% pour 3 modèles (CHK1, JAK1 et ROCK2), de 12% pour JAK3, de 13% pour les deux isozymes VEGFR, et de 14% pour BRAF et JAK2. Pour les 5 classifieurs à 83% : LCK est amélioré de 8%, AGMX1 et FMS de 9%, Aurora-A et SCFR de 11%. Les 17 modèles XGBoost avec une précision entre 84 et 87% ont été portés entre 89 et 96%. FAK et PLK1 sont passés de 88 à 97%, IKK- β et ERK2 de 89 à 95%, avec une amélioration notable du modèle IGF1R (de 90 à 96%). La figure 3.4 présente le nombre de modèles construits selon leur précision obtenue, pour les modèles XGBoost seuls et hybrides.

Cette amélioration est due à l'architecture hybride, qui combine la capacité de XGBoost à traiter des données structurées avec celle du DNN à apprendre des relations non linéaires entre les caractéristiques.

TABLEAU 3.4 – Détails pertinents sur les modèles réalisés (XGB, Dim et VC désignent respectivement : XGBoost, Dimension et Validation Croisée).

Cible	Dim Train	Dim Test	XGB VC	XGB/DNN VC	Score XGB	Score XGB/DNN
LCK	(1513, 476)	(504, 476)	0.82	0.94	0.83	0.91
AGMX1	(1312, 454)	(437, 454)	0.82	0.94	0.83	0.92
AKT1	(1344, 463)	(448, 463)	0.84	0.94	0.85	0.94
ALK	(989, 461)	(329, 461)	0.83	0.95	0.87	0.93
Aurora-A	(1588, 492)	(529, 492)	0.8	0.95	0.83	0.94
BRAF	(1764, 469)	(588, 469)	0.8	0.94	0.8	0.94
CDK2	(1347, 509)	(448, 509)	0.85	0.95	0.86	0.94
CHK1	(1543, 477)	(514, 477)	0.81	0.94	0.82	0.92
DYRK1	(750, 490)	(249, 490)	0.79	0.92	0.84	0.92
EGFR	(5675, 510)	(1891, 510)	0.84	0.96	0.86	0.96
ERK2	(622, 434)	(207, 434)	0.89	0.92	0.89	0.95
FAK	(646, 468)	(215, 468)	0.87	0.92	0.88	0.97
FGFR1	(1197, 449)	(399, 449)	0.85	0.95	0.87	0.95
FLT3	(1631, 453)	(543, 453)	0.8	0.95	0.85	0.95
FMS	(708, 453)	(236, 453)	0.83	0.91	0.83	0.92
GSK3 β	(2231, 523)	(743, 523)	0.81	0.95	0.84	0.95
HGFR	(2436, 472)	(812, 472)	0.85	0.96	0.85	0.96
IGF1R	(1106, 477)	(368, 477)	0.89	0.95	0.9	0.96
IKK- β	(900, 470)	(299, 470)	0.86	0.95	0.89	0.95
INSR	(737, 477)	(245, 477)	0.88	0.93	0.9	0.9
JAK1	(1017, 473)	(338, 473)	0.76	0.93	0.82	0.92
JAK2	(1786, 480)	(595, 480)	0.77	0.93	0.8	0.94
JAK3	(1487, 469)	(495, 469)	0.78	0.93	0.81	0.93
MAPK8	(969, 466)	(322, 466)	0.85	0.94	0.84	0.93
MK2	(742, 454)	(247, 454)	0.89	0.94	0.85	0.9
mTOR	(1416, 458)	(472, 458)	0.85	0.96	0.86	0.94
PDGFR β	(997, 453)	(332, 453)	0.82	0.94	0.87	0.94
PI3K α	(2774, 468)	(924, 468)	0.83	0.94	0.84	0.96
PIK3C δ	(1981, 476)	(660, 476)	0.81	0.94	0.81	0.94
PIK3C γ	(1932, 483)	(644, 483)	0.82	0.94	0.85	0.96
PLK1	(904, 473)	(301, 473)	0.88	0.95	0.88	0.97
ROCK1	(626, 441)	(208, 441)	0.72	0.89	0.75	0.86
ROCK2	(910, 435)	(303, 435)	0.75	0.9	0.81	0.91
SCFR	(892, 440)	(297, 440)	0.78	0.92	0.83	0.94
SRC	(2466, 473)	(821, 473)	0.86	0.95	0.87	0.96
SYK	(703, 446)	(234, 446)	0.81	0.93	0.86	0.89
Trk- α	(674, 454)	(224, 454)	0.73	0.9	0.78	0.92
VEGFR1	(731, 436)	(243, 436)	0.84	0.91	0.81	0.94
VEGFR2	(5088, 471)	(1696, 471)	0.78	0.94	0.81	0.94

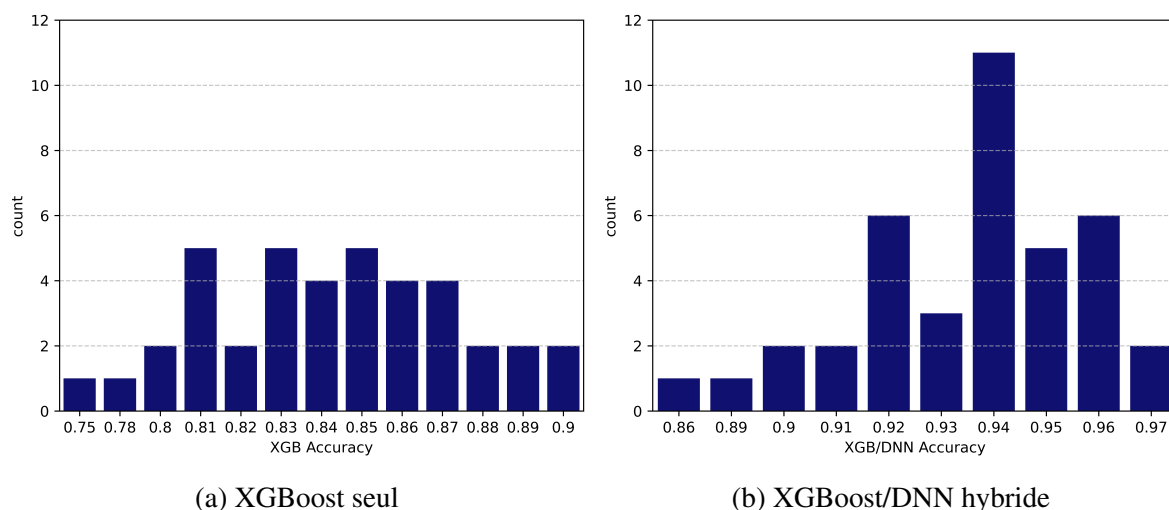


FIGURE 3.4 – Nombre de modèles selon leur précision.

En apprentissage automatique, obtenir une précision en test proche de celle en apprentissage est un résultat positif. Cela démontre la robustesse du modèle, et prouve qu'il ne se contente pas de bien fonctionner sur les données d'entraînement, mais aussi sur des données nouvelles et inédites (données de test). Une grande différence indiquerait un surapprentissage (overfitting), rendant le modèle moins généralisable. Le surapprentissage nuit fréquemment à la pertinence des QSAR. Réduire l'écart entre la précision train/test est donc essentiel pour concevoir des modèles non biaisés.

La figure 3.5 présente les courbes de densité des $\Delta Acc_{train/test}$ (densité en fonction de la différence de score). Bien que les modèles XGBoost seuls ne soient pas considérés comme surentraînés, la courbe rouge (XGBoost) montre un $\Delta Acc_{train/test}$ plus important que celui des modèles hybrides (courbe verte), qui présentent clairement des écarts réduits. Cela prouve que l'approche hybride parvient à résoudre les problèmes de surapprentissage dans la construction des QSARs.

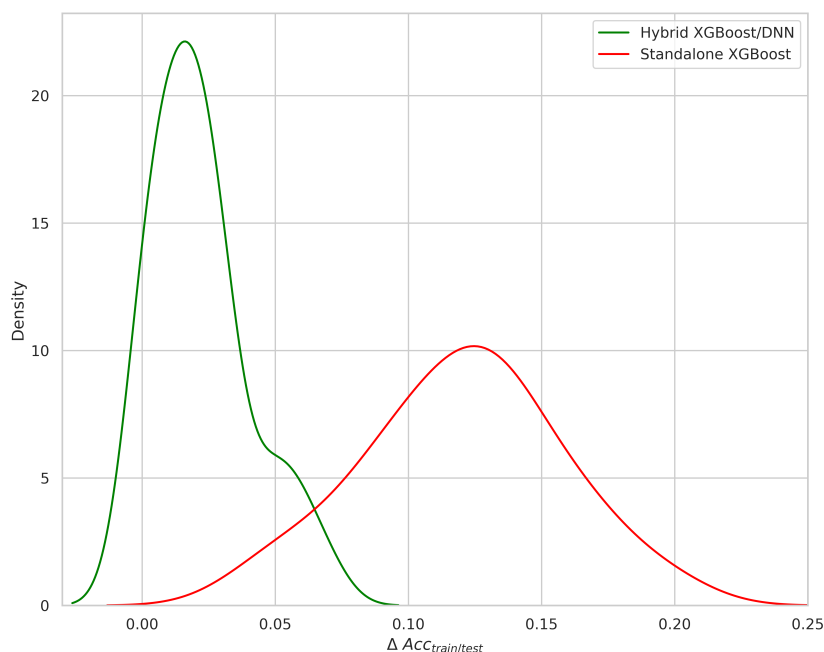


FIGURE 3.5 – Courbe de densité estimée pour la différence de score entre les ensembles d'apprentissage et de test, pour les modèles XGBoost seuls et hybrides.

3.3.3.1 Comparaison des performances de XGBoost/DNN avec d'autres algorithmes standards

Pour étudier la supériorité de l'architecture proposée, nous comparons ses performances à celles d'autres méthodes de pointe, notamment la Forêt Aléatoire (RF) [134], les Machines à Vecteurs de Support (SVM) [135] et les DNNs [137]. Les scores reportés dans le tableau 3.5 montrent que l'approche hybride surpasse efficacement les méthodes standards susmentionnées, sur une gamme étendue de tailles de données. De FAK et PLK1 (646 et 904 molécules dans l'ensemble d'entraînement) à PI3K α et EGFR (2774 et 5675 molécules), la méthode XGBoost/DNN dépasse les autres algorithmes connus et performants.

TABEAU 3.5 – Valeurs de précision de RF, SVM, DNN et du modèle hybride XGBoost/DNN pour divers jeux de données couvrant un éventail illustratif de tailles.

Modèle	Score RF	Score SVM	Score DNN	Score XGB/DNN
FAK	0.87	0.88	0.87	0.97
PLK1	0.89	0.87	0.9	0.97
PDGFR β	0.86	0.86	0.82	0.94
AKT1	0.84	0.84	0.83	0.94
mTOR	0.86	0.87	0.85	0.93
SRC	0.85	0.86	0.83	0.96
PI3K α	0.85	0.85	0.86	0.96
EGFR	0.86	0.86	0.85	0.95

3.3.3.2 Capacité prédictive des modèles

Les matrices de confusion (MC) sont des tableaux utilisés pour évaluer la performance d'un modèle de classification sur un ensemble de données de test. Elles sont importantes pour évaluer la fiabilité des prédictions d'un modèle [141]. Une matrice unique compare les valeurs réelles d'une variable avec les prédictions produites par le classificateur. Elle fournit des informations sur le nombre de vrais positifs, faux positifs, vrais négatifs et faux négatifs. L'importance des MC réside dans le fait que toutes les métriques d'évaluation en découlent. Les matrices de confusion des modèles XGBoost et hybrides EGFR, mTOR, VEGFR2 et FLT3 sont présentées dans les figs. 3.6 and 3.7.

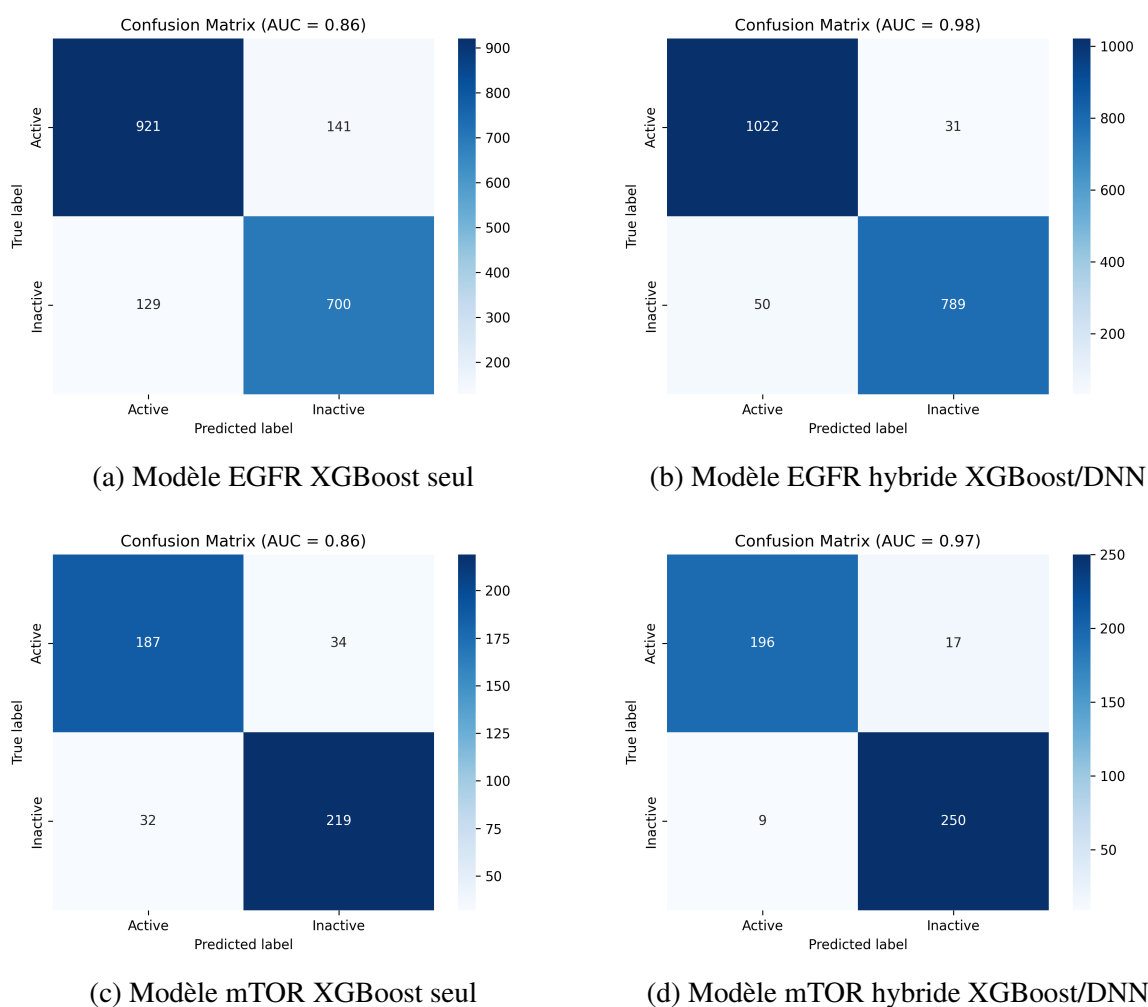


FIGURE 3.6 – Matrices de confusion des modèles EGFR et mTOR.

Les classifications binaires dans la figure 3.6 (EGFR et mTOR) atteignent une précision de 86% pour le modèle XGBoost seul. Les matrices montrent la capacité du modèle à classer correctement près de 86% des instances dans les deux classes ($\frac{1621}{1891}$ pour EGFR et $\frac{406}{472}$ pour mTOR). Bien que seulement 14% des instances soient mal prédites, le modèle hybride élève la précision à 96% pour EGFR et à 94% pour mTOR. Les classifications incorrectes globales sont ainsi réduites à 4% et 6% respectivement pour les données des deux kinases.

Un pourcentage de 87 et 85 de la classe positive est correctement évalué respectivement pour

les modèles EGFR et mTOR ($\frac{921}{1062}$ pour EGFR et $\frac{187}{221}$ pour mTOR). Il s'agit du caractère sensible de la classification par apprentissage automatique (sensibilité ou TPR). Cela représente la proportion de substances actives correctement détectées. La sensibilité de l'approche hybride atteint 97% pour l'EGFR ($\frac{1022}{1053}$) et 92% pour mTOR ($\frac{196}{213}$). Cela est important dans les applications où éviter les faux négatifs est crucial. Ainsi, la sensibilité plus élevée de l'approche hybride indique que les modèles sont plus efficaces pour identifier les composés actifs.

Il est également observé dans la MC du modèle mTOR (XGBoost) que les cases (TN) et (FP) contiennent respectivement 219 et 32 valeurs, menant à une spécificité de $\frac{219}{251}$ (87%). Cette dernière est encore améliorée par l'approche proposée pour atteindre un ratio de $\frac{250}{259}$ (97%).

Ces prédictions sont généralement appuyées par d'autres techniques In Silico, telles que les investigations basées sur la structure et les simulations de DM, ou bien directement vérifiées par des essais in vitro. De plus, les 4% de prédictions incorrectes du modèle hybride EGFR (6% pour mTOR et CDK2) donnent beaucoup plus de pertinence aux classifieurs, comparativement aux 14% des modèles XGBoost seuls. L'amélioration fournit des fonctions coût plus acceptables, en particulier lorsqu'on travaille sur de grandes quantités de composés chimiques (criblage virtuel).

Sur la base des connaissances de l'algorithme (XGBoost) seul sur les données, on observe dans le jeu de test EGFR qu'un pourcentage de 13 (15% pour mTOR) représente la faiblesse du modèle à identifier les composés réellement inactifs. A travers l'approche hybride, cette perte est réduite à seulement 6% (3% pour mTOR). Les 94% restants (97% pour mTOR) sont effectivement inactifs.



FIGURE 3.7 – Matrices de confusion des modèles FLT3 et VEGFR2.

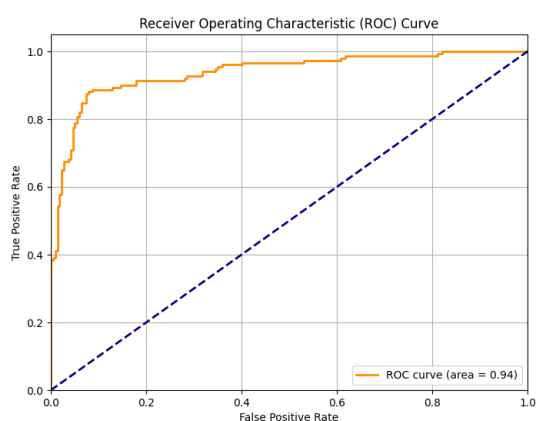
Selon les informations issues des modèles XGBoost, les classifications pour FLT3 et VEGFR2 ne permettent de prédire correctement que 85% et 81% respectivement d'un espace chimique externe donné. La performance de ces modèles est portée à 95% et 94% (soit une amélioration de 13% pour VEGFR2). Bien que la spécificité obtenue pour FLT3 et VEGFR2 (respectivement 90% et 80%) à l'aide de l'approche XGBoost soit acceptable, l'approche hybride améliore significativement ces résultats pour atteindre 95% et 93%, ce qui suggère que la majorité des composés prédits comme actifs pourraient effectivement présenter une cytotoxicité pertinente. Lorsqu'on utilise les deux modèles améliorés basés sur leur caractère sensible, seuls 5% des composés identifiés (hits) peuvent ne présenter en réalité aucune activité significative, contre 10% pour le modèle FLT3 seul (20% pour VEGFR2). Une forte sensibilité est observée pour plusieurs modèles testés. Le tableau 3.6 présente le score, la sensibilité, la spécificité et l'aire sous la courbe (AUC) pour 8 modèles de protéines hautement oncogènes.

TABLEAU 3.6 – Métriques d'évaluation pertinentes pour certains des modèles de protéines les plus oncogènes (Sensit et Spec désignent respectivement la sensibilité et la spécificité, tandis que XGB désigne XGBoost).

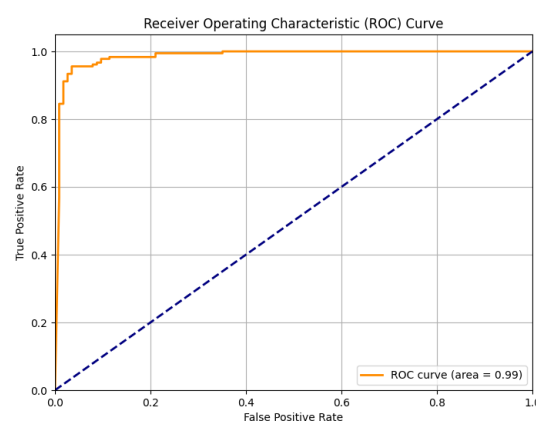
Cible	Score XGB	Score XGB/DNN	XGB Sensit	XGB/DNN Sensit	XGB Spec	XGB/DNN Spec	XGB AUC	XGB/DNN AUC
EGFR	0.86	0.96	0.87	0.97	0.85	0.94	0.86	0.98
AKT1	0.85	0.94	0.83	0.95	0.88	0.94	0.85	0.97
mTOR	0.86	0.94	0.85	0.92	0.87	0.97	0.86	0.97
PDGFR β	0.87	0.94	0.94	0.95	0.76	0.9	0.85	0.97
SRC	0.87	0.96	0.94	0.96	0.72	0.93	0.83	0.98
VEGFR1	0.81	0.94	0.95	0.96	0.48	0.87	0.71	0.97
VEGFR2	0.81	0.95	0.82	0.95	0.8	0.93	0.81	0.97
FLT3	0.85	0.94	0.77	0.95	0.9	0.95	0.84	0.97

La courbe ROC (Receiver Operating Characteristic), qui est une métrique d'évaluation pour les classifications binaires, est une courbe de probabilité qui trace le TPR en fonction du FPR, pour différentes valeurs de seuil, montrant la performance du classifieur à chaque seuil [142, 143]. L'AUC est utilisée comme indicateur sur les courbes ROC. La figure 3.8 présente les courbes ROC des classifieurs IGF1R XGBoost et hybrides.

Lorsqu'on utilise les courbes ROC pour évaluer nos modèles QSAR, on recherche la plus grande surface sous la courbe TPR(FPR), ce qui est proportionnel à un TPR maximal correspondant à un FPR minimal, au point d'inflexion. Une classification binaire idéale aurait une courbe ROC passant par TPR = 1 et FPR = 1 [143].



(a) XGBoost seul



(b) XGBoost/DNN hybride

FIGURE 3.8 – Courbes ROC pour les modèles du récepteur du facteur de croissance analogue à l'insuline de type 1 (IGF1R).

L'utilisation de l'AUC est une bonne pratique pour mesurer la capacité de décision dichotomique [143, 144]. Lorsque l'AUC = 1, le classifieur peut distinguer correctement tous les points des classes positive et négative. En revanche, si elle est proche de zéro, cela reviendrait à inverser toutes les classes, et prédire tout à tort. Ainsi, une AUC comprise entre 0.5 et 1 correspond à la plage dans laquelle le modèle peut raisonnablement distinguer les actifs des inactifs, en raison de sa capacité à détecter plus de (TN) et (TP) que de (FN) et (FP). Bien que les valeurs d'AUC des modèles seuls soient observées au-dessus de 0.7, celles observées pour l'approche hybride sont toutes supérieures à 0.9, la plupart avoisinant la région de 0.98. L'amélioration observée pour le modèle IGF1R via l'approche hybride exprime clairement la

fiabilité de l'approche proposée (voir figure 3.8).

En général, un bon modèle de classification devrait présenter à la fois une sensibilité et une spécificité élevées. Toutefois, dans certains cas, un compromis peut exister entre les deux paramètres. Si l'augmentation de la sensibilité de nos modèles peut entraîner davantage de faux positifs, une augmentation de la spécificité peut également engendrer davantage de faux négatifs. Ainsi, l'équilibre approprié entre sensibilité et spécificité dépend de l'application spécifique. L'approche hybride proposée permet d'atteindre des valeurs élevées pour les deux.

3.3.3.3 Analyse de l'importance des caractéristiques

Pour mieux comprendre le rôle de chaque entrée dans le modèle XGBoost/DNN, nous avons employé la méthode SHAP (SHapley Additive exPlanations) [145] afin de calculer la contribution de chaque descripteur. La figure 3.9 montre le graphique récapitulatif SHAP pour le modèle hybride EGFR, tandis que le tableau 3.7 présente les cinq caractéristiques principales classées selon leurs valeurs SHAP absolues moyennes.

Une analyse détaillée des descripteurs moléculaires a montré que l'activité biologique dépend d'un équilibre fin entre la compacité géométrique, la polarisation électronique et la rigidité structurale. Le *PetitjeanNumber* et les indices JGI décrivent la compacité et la distribution de charge, mettant en évidence le rôle des effets électrostatiques et de polarisation [146, 147]. Les descripteurs électroniques tels que *SpMin5_Bhm*, *BCUTw-1l* et *GATS8p* capturent la polarisabilité, la flexibilité conformationnelle et les interactions $\pi - \pi$ [146]. Les caractéristiques structurales comme *MDEC-33* et *naasC* indiquent la complexité topologique et la lipophilicité, tandis que *nF7Ring* et *nF10Ring* reflètent la rigidité et la taille moléculaire [147].

Dans nos modèles de kinases, *MDEC-33* était un facteur clé pour SRC, son analogue "azote" pour EGFR, et le *max/minHBint10* s'est révélé pertinent dans les modèles SRC et mTOR. Ce dernier est une caractéristique clé pour évaluer la force des liaisons hydrogène [147]. Le nombre d'atomes de fluor attachés à des cycles carbonés (*nF(n)*) et les descripteurs de régions hydrophobes (*minsCH3*, *minHCsatu*) ont également contribué à la stabilité moléculaire et à l'affinité de liaison. Globalement, les résultats soulignent que l'affinité optimale ligand-récepteur découle de la synergie entre compacité, polarisation, rigidité et interactions hydrophobes/liaisons hydrogène.

D'autre part, l'analyse montre également que les valeurs SHAP associées aux probabilités XGBoost étaient de loin les plus élevées, tandis que celles liées aux descripteurs moléculaires initiaux étaient beaucoup plus faibles. Cela suggère que le DNN n'apprend pas de nouvelles règles décisionnelles à partir de zéro, mais exploite plutôt les sorties de probabilité de XGBoost. En d'autres termes, le réseau neuronal agit principalement comme une couche de calibration.

Pour déterminer le rôle exact du DNN dans le modèle hybride, nous avons comparé la calibration des deux modèles. Nous avons tracé les courbes de calibration et calculé les scores de Brier correspondants (figure 3.10). Le modèle hybride a obtenu un score de Brier bien plus faible que le XGBoost seul, et sa courbe est plus proche de la ligne de calibration parfaite. Ces résultats montrent que combiner XGBoost avec un DNN va au-delà de l'amélioration des

scores de prédiction. Le modèle hybride produit des estimations de probabilité mieux alignées aux résultats réels, ce qui met en évidence le rôle du DNN comme calibrateur pratique et efficace.

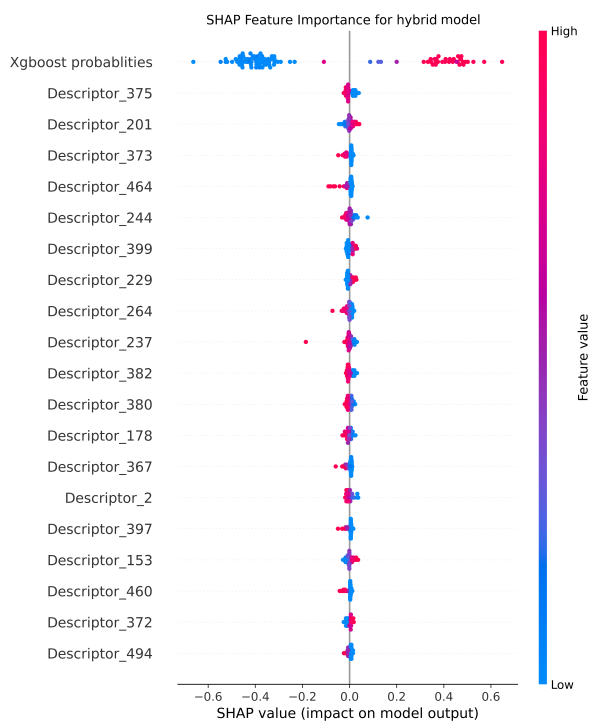


FIGURE 3.9 – Importance des caractéristiques SHAP pour le modèle XGBoost/DNN appliqué aux données EGFR.

TABEAU 3.7 – Cinq principales caractéristiques SHAP pour le modèle XGBoost/DNN appliqué aux données EGFR.

Feature	Mean SHAP
probabilités Xgboost	0.398812
minsCH3	0.014609
GATS8p	0.011788
minHCsat	0.011037
MDEN-33	0.010521

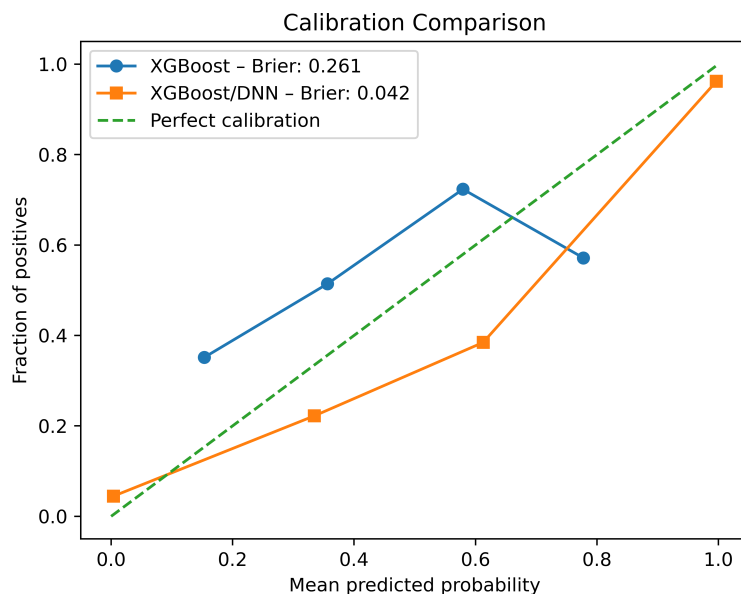


FIGURE 3.10 – Comparaison de la calibration et du score de Brier entre les modèles XGBoost et XGBoost/DNN appliqués aux données EGFR.

Considérant qu'après un criblage QSAR, des investigations supplémentaires doivent être envisagées, une haute précision peut être l'indicateur le plus important en conception de médicaments. Les 40 modèles hybrides présentés atteignent des précisions extrêmement élevées, en plus de résultats de sensibilité/spécificité élevés, avec un équilibre optimal entre les deux. De plus, la robustesse contre les problèmes de surentraînement est remarquable.

3.3.3.4 Analyse de l'amélioration au niveau structural

Pour comprendre l'aspect chimique de l'amélioration engendrée par l'approche présentée, il est nécessaire d'analyser les caractéristiques structurales des composés incorrectement prédits par le modèle XGBoost seul, et correctement prédits par le modèle hybride. Les modèles d'apprentissage automatique éprouvent souvent des difficultés à prédire correctement l'activité des "cliff pairs". Celles-ci sont définies comme des molécules très similaires présentant de grandes différences en activité [148]. Dans la figure 3.11 sont présentés quatre composés "cliffs" mal classés par le modèle XGBoost mTOR, et correctement classés par le modèle XGBoost/DNN.

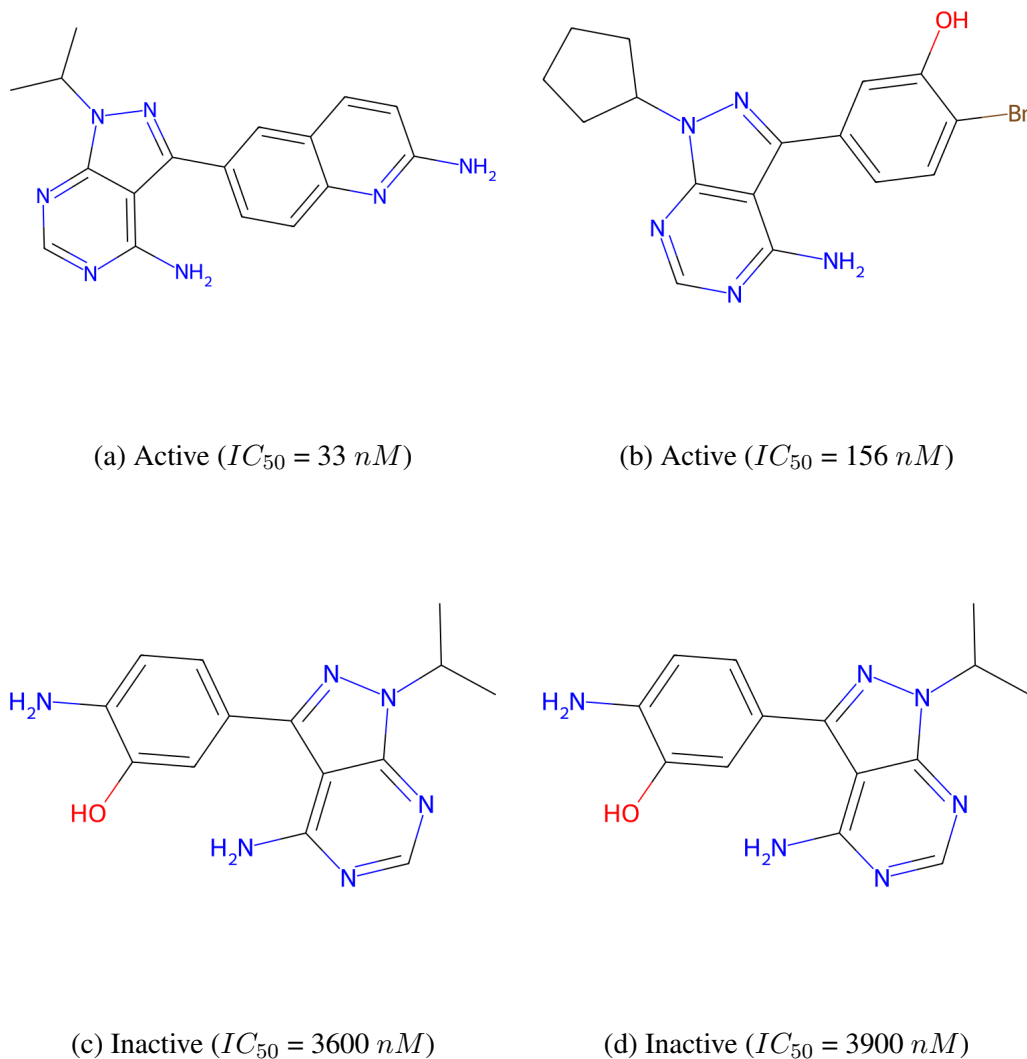


FIGURE 3.11 – Structures chimiques de dérivés pyrazolopyrimidines (inhibiteurs de mTOR) présentant des "activity-cliffs". L'étiquette d'activité correspond uniquement au seuil utilisé dans la présente étude.

Le modèle classique, lorsqu'il a été entraîné sur des données où certains dérivés pyrazolopyrimidines étaient actifs et d'autres ne l'étaient pas [149], n'a pas réussi à prédire correctement leur activité. Le scaffold comportant le fragment quinolin-2-amine (figure 3.11a) a été classé comme inactif malgré sa forte activité réelle. Bien que le motif amino-phénol (figure 3.11c) entraîne un impact négatif sur la l'activité réelle, le modèle a classé le composé comme actif. Il en va de même pour les deux structures portant des substituants méthylcyclopentane (figure 3.11b) et pyrrolidine (figure 3.11d). Malgré leurs différences substantielles d'activité, les quatre composés présentent de fortes similarités structurales. Cet exemple d'erreurs spécifiques a été corrigé grâce à la calibration des probabilités fournie par le composant DNN de l'architecture hybride.

3.4 Conclusion

Dans cette étude, nous avons collecté des données expérimentales de bioactivité relatives à 40 protéines kinases oncogènes. Les essais d'inhibition enzymatique *in vitro*, ainsi que les informations de bioactivité des médicaments approuvés, ont été analysés. Les données ont été nettoyées et organisées, avant un étiquetage de la bioactivité selon la valeur d' IC_{50} . Des descripteurs moléculaires ont été générés avant d'appliquer un traitement mathématique et une procédure de sélection de variables. Ensuite, les jeux de données préparés ont été soumis à XGBoost pour une classification binaire. Afin de traiter les problèmes de précision, notamment rencontrés dans les recherches sur l'inhibition des tyrosine kinases, ainsi que dans le LBDD de façon générale, ce chapitre de thèse a introduit une nouvelle approche hybride, combinant XGBoost à des réseaux neuronaux profonds, pour construire des modèles QSAR fiables.

Nous avons construit des modèles prédictifs performants, parmi lesquels : la tyrosine kinase de la rate (SYK), le récepteur à l'insuline (INSR), la kinase dépendante des cyclines 2 (CDK2), le récepteur du facteur de croissance des fibroblastes 1 (FGFR1), la tyrosine-protéine kinase proto-oncogène (SRC), la kinase d'adhésion focale (FAK), ont respectivement atteint : 89%, 90%, 94%, 95%, 96%, 97% de précision. De plus, nous avons montré que la méthode hybride surpasse la méthode classique, en atteignant une fiabilité supérieure. Une plage de précision allant de 86% à 97%, bien supérieure à celle obtenue avec XGBoost seul (75% à 90%), a été démontrée. Par ailleurs, la remarquable robustesse face aux problèmes de surapprentissage a été illustrée.

Les résultats suggèrent que les estimations de probabilité produites par XGBoost jouent un rôle dominant dans l'orientation du DNN. Plutôt que de découvrir un grand nombre de nouvelles interactions non linéaires entre les descripteurs moléculaires, le réseau neuronal affine et calibre principalement la sortie de XGBoost. Le cadre hybride fonctionne de manière à combiner la forte capacité prédictive du "gradient boosting" sur les données tabulaires avec l'aptitude de l'apprentissage profond à améliorer la calibration des probabilités. Nous recommandons donc cette approche comme une stratégie pratique et efficace pour construire des classifieurs Structure-Activité fiables pour les cibles étudiées ou pour d'autres jeux de données présentant des caractéristiques similaires.

Nos travaux et notre analyse peuvent fournir une voie directe pour le criblage virtuel de grandes bibliothèques chimiques externes, et servir aussi bien dans le cadre des thérapies anti-angiogéniques que pour la conception générale de médicaments.

Chapitre 4

Conception et Identification d’Inhibiteurs de ERK2

4.1 Introduction

Les "Mitogen-Activated Protein Kinases" (MAPKs) ont été initialement identifiées à la fin des années 1980 comme des phosphoprotéines stimulées par les mitogènes [150, 151]. Leur mécanisme de signalisation implique une cascade de phosphorylation, connue sous le nom de voie RAS–RAF–MEK–ERK, nommée d’après les protéines qui la composent. Divers membres de la famille des MAPK ont évolué pour répondre à des stimuli spécifiques [152]. Parmi eux, ERK1 et ERK2 (Extracellular signal-Regulated Kinase 1 et 2) sont les effecteurs finaux de cette voie. Ces deux isoformes partagent 84% d’identité de séquence [153]. La dérégulation de ERK1/2 a été impliquée dans de nombreuses maladies, notamment le diabète [154], les troubles neurodégénératifs et du développement [155, 156], ainsi que diverses formes de cancer [153].

Dans des conditions physiologiques normales, l’activation de la voie permet à ERK1/2 de phosphoryler des substrats distincts en fonction de leur localisation subcellulaire, soit dans le cytoplasme, soit dans le noyau, régulant ainsi la prolifération et la survie cellulaires [157]. Dans le cytoplasme, ils ciblent des protéines cytosquelettiques, influençant la motilité cellulaire, l’adhésion et la régulation d’autres cascades de signalisation [158]. En revanche, les altérations génétiques et épigénétiques qui affectent le comportement cellulaire, en particulier dans le cancer, conduisent souvent à une activation aberrante de cette cascade et à une signalisation ERK soutenue.

Malgré leur importance, les mécanismes sous-jacents à la mort cellulaire médiée par ERK1/2 restent mal compris [159]. Leur activation est connue pour favoriser à la fois la survie et la prolifération cellulaires. Dans le noyau, ERK1/2 activent des facteurs de transcription, renforçant leur pertinence en tant que cibles thérapeutiques dans le cancer [160–162]. ERK2, en particulier, a démontré qu’elle favorisait la progression tumorale via la stimulation de la transition épithélio-mésenchymateuse [163], et est plus abondamment exprimé que ERK1 dans les tissus de mammifères [164]. Alors que l’activité de ERK1/2 confère une protection contre la tumorigenèse dans les cellules normales dotées de mécanismes suppresseurs de tumeurs intacts, des mutations et altérations de ces systèmes peuvent convertir leur fonction en un rôle pro-oncogénique [165]. De manière plus générale, les kinases sont des enzymes multifonctionnelles essentielles au traitement de l’information biologique critique [166]. Notamment, ERK2

phosphoryle et régule négativement CIITA, une protéine clé dans l'initiation des réponses immunitaires [167].

Étant donné le rôle crucial de la voie MAPK dans le cancer, un intérêt croissant a été porté à son inhibition thérapeutique via le développement d'inhibiteurs sélectifs [168]. Les premières stratégies se concentraient principalement sur les kinases en amont. Cependant, la résistance due à des mutations génétiques a recentré l'attention sur les composants en aval, en particulier ERK [169].

Deschenes-Simard et al. (2014) ont souligné que l'activité de ERK1/2 offre une protection contre la tumorigenèse dans les cellules normales, mais favorise la tumorigenèse lorsque les modules suppresseurs de tumeurs sont altérés [165]. Dans le domaine de la conception de médicaments, les méthodes rationnelles *in silico* ont considérablement fait progresser la découverte d'inhibiteurs de ERK. Parmi les cibles en aval, ERK2 est particulièrement priorisé en raison de son rôle central dans la signalisation oncogénique, contrairement à ERK1 [170]. ERK2 a également été spécifiquement associé à la transition épithélio-mésenchymateuse [163] et est plus fortement exprimé dans les tissus de mammifères [164].

Plusieurs inhibiteurs de ERK ont été proposés, et beaucoup font actuellement l'objet d'évaluations précliniques et cliniques [171, 172]. Parmi eux, Ulixertinib est le seul à avoir montré des avancées cliniques significatives à ce jour [173, 174].

De nombreux inhibiteurs potentiels n'atteignent pas le stade clinique en raison d'une faible sélectivité, ce qui peut entraîner des effets secondaires, une spécificité pharmacodynamique réduite, et des échecs précoces dans le développement de médicaments. Par conséquent, la conception et l'identification de nouvelles classes d'inhibiteurs à sélectivité améliorée constituent une étape cruciale vers un traitement anticancéreux plus efficace [175].

Le SBVS a permis la synthèse de dérivés de tétrahydropyridopyrimidine, bien que certains aient montré des limitations *in vivo* [176]. Des améliorations ultérieures ont consisté à remplacer le lien urée-pipéridine par un noyau pyridone, conduisant à des inhibiteurs plus sélectifs adaptés à l'évaluation préclinique [177]. Les optimisations finales – telles que le remplacement du groupement terminal tétrahydropyrane par un groupe 1-méthyl-1H-pyrazole ont conduit au développement de Ravoxertinib, actuellement en investigation clinique [178]. Malgré les progrès, aucun inhibiteur de ERK n'a encore reçu l'approbation de la FDA [179]. L'émergence de résistances médicamenteuses souligne la nécessité d'identifier plusieurs molécules à potentiel thérapeutique ciblant un ou plusieurs composants de cet axe de signalisation [180]. De plus, l'obtention d'une perméabilité à la barrière hémato-encéphalique reste un défi pour les inhibiteurs de ERK ciblant le cerveau [181].

Les avancées récentes ont élargi le paysage des inhibiteurs de ERK1/2 au-delà des agents de première génération tels que SCH772984, Ulixertinib et Ravoxertinib [178, 182, 183]. De nouveaux composés comme LY3214996 [172] et MK-8353 [184] ont montré des résultats prometteurs en phase préclinique et clinique précoce, offrant une inhibition sélective de ERK1/2 avec une toxicité gérable. Ces inhibiteurs de nouvelle génération visent à surmonter la résistance associée aux inhibiteurs en amont de RAF et MEK, en particulier dans les tumeurs mutées RAS/BRAF. Les recherches en cours se concentrent sur l'amélioration de la sélectivité, la réduction des effets hors cible, et le contournement de la résistance adaptative dans les

cellules tumorales.

La découverte moderne de médicaments intègre diverses stratégies computationnelles pour l'identification et l'optimisation d'agents thérapeutiques. Le LBDD exploite l'apprentissage automatique supervisé - ML (Machine Learning), via des algorithmes adaptés ou des cadres d'ensemble, pour modéliser les profils de bioactivité [185]. Ces modèles sont particulièrement utiles pour le criblage virtuel de grandes bibliothèques de composés. De plus, des techniques non supervisées telles que la PCA et l'analyse par composantes indépendantes sont utilisées pour la réduction de dimensionnalité et le regroupement de composés selon des caractéristiques structurales ou physico-chimiques [185]. Toutefois, les approches ML non supervisées n'ont pas encore été appliquées à l'identification de nouveaux inhibiteurs de ERK2. De même, la combinaison de ML supervisé et non supervisé pour découvrir de nouvelles familles de composés actifs reste une stratégie inexplorée. Les combinaisons précédentes ont été principalement utilisées pour l'étiquetage de composés lorsque des données annotées suffisantes étaient disponibles [186]. Par conséquent, notre cadre proposé, combinant ML supervisé et non supervisé pour identifier de nouvelles familles de composés actifs, représente une approche innovante, en particulier pour cibler une voie de signalisation activée de manière aberrante dans plus de 30% des cancers humains [187].

Dans ce chapitre, nous présentons un travail visant à identifier de nouveaux noyaux chimiques avec un fort potentiel inhibiteur de ERK2 en tant que candidats d'agents anticancéreux. Nous commençons par utiliser le ML supervisé pour modéliser les relations structure-activité, suivi d'un criblage virtuel basé sur les ligands et la structure (QSAR et docking moléculaire). Ensuite, nous réalisons une analyse de regroupement afin de classer les composés identifiés en familles chimiques distinctes. Sur la base des caractéristiques moléculaires, nous identifions cinq nouveaux composés, chacun représentant un "chemotype" unique avec un potentiel thérapeutique prometteur. Nous évaluons également leur similarité structurale avec les inhibiteurs connus de ERK2. Pour évaluer la sélectivité, nous testons les composés contre d'autres kinases périphériques à l'aide de classifieurs ML internes (présentés dans le chapitre 3). Les interactions de liaison sont ensuite explorées via le docking moléculaire et les simulations de DM, et leurs propriétés pharmacocinétiques sont analysées par profilage ADMET. Cette étude introduit un cadre ML supervisé/non supervisé pour la conception d'inhibiteurs de ERK2 et présente cinq types de composés inédits aux propriétés pharmacologiques prometteuses, identifiés et caractérisés ici pour la première fois. Ces résultats constituent une base solide pour l'optimisation future de candidats médicaments anticancéreux.

4.2 Méthodologie

4.2.1 Criblage basé sur le docking et clustering

Le criblage virtuel basé sur le docking a été effectué en utilisant Autodock Vina [188, 189], et le programme UCSF Chimera a été utilisé pour la préparation des molécules cibles (protéine) et des ligands [190]. Les structures cristallines de la protéine humaine ERK2 formant des complexes avec SCH772984 et avec ulixertinib ont été récupérées de la base de données RCSB Protein Data Bank [191] (références PDB : 4qta et 6gdq, respectivement [182, 183]). En plus de ulixertinib, le seul inhibiteur de ERK1/2 utilisé dans les essais cliniques à ce jour, nous avons choisi SCH772984 en raison de sa forte puissance inhibitrice [182] et de sa forme moléculaire

distinctive. Il est connu pour induire un changement conformationnel significatif dans le site allostérique de la MAPK1. Pour cette raison, nous avons sélectionné la structure PDB 4QTA afin de permettre un meilleur classement des molécules candidates pendant le criblage virtuel basé sur le docking. Cependant, étant donné que la forme et la taille moléculaire des composés finaux sélectionnés ressemblent davantage à celles d'ulixertinib, la structure PDB 6GDQ a été utilisée pour les analyses suivantes (docking et dynamique moléculaires).

Tous les résidus manquants ont été complétés par modélisation homologue en utilisant Modeller [192]. Ensuite, tous les résidus non standard ont été supprimés (sauf les ligands co-cristallisés), les atomes d'hydrogène ont été ajoutés et les charges ont été attribuées à tous les atomes conservés dans les structures. Les ligands de référence (SCH772984 et ulixertinib) ont été isolés des structures protéiques pour la validation du protocole de docking. L'espace de recherche a été défini avec UCSF Chimera, dans une boîte parallélépipédique entourant le ligand, dont la taille était de (20.73 Å, 25.46 Å, 19.68 Å) pour le complexe avec SCH772984, et de (19 Å, 20 Å, 14 Å) pour le complexe avec ulixertinib. Pour assurer une procédure de docking optimale, le nombre de modes de liaison a été fixé à cinq, l'exhaustivité de la recherche à 8 (équilibre entre vitesse et précision), et la différence d'énergie maximale à 3 kcal/mol (seules les poses thermodynamiquement raisonnables).

Le clustering a été réalisé en utilisant l'algorithme K-Means [193] tel qu'implémenté dans Scikit-learn [133]. Il s'agit d'un algorithme de clustering basé sur les centroides qui partitionne les données en k clusters en minimisant la distance entre les points et leurs centroides assignés [193]. Pour déterminer le nombre optimal de clusters, la méthode Silhouette a été appliquée [194]. C'est une technique basée sur la comparaison de la compacité et de la séparation de chaque cluster. La méthode montre quels points se trouvent bien à l'intérieur de leurs clusters, et lesquels sont simplement quelque part entre les clusters, en évaluant la largeur de la silhouette $s(i)$ pour l'entité $i \in I$ définie par l'équation 4.1 [194] :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4.1)$$

Où $a(i)$ est la distance moyenne entre i et toutes les autres entités du cluster auquel i appartient et $b(i)$ est le minimum des distances moyennes entre i et toutes les entités de chaque autre cluster. Les valeurs de $s(i)$ varient dans l'intervalle (-1, 1). Si la valeur de $s(i)$ d'une entité est proche de zéro, cela signifie qu'elle pourrait être assignée à un autre cluster. Si la largeur de la silhouette est proche de -1, cela signifie que l'entité est mal classée. Si toutes les valeurs de la largeur de la silhouette sont proches de 1, cela signifie que l'ensemble I est bien groupé. Un clustering peut être caractérisé par la largeur moyenne de la silhouette $\bar{s}$ (ou score) des entités individuelles. La plus grande valeur de $\bar{s}$, pour différents k , indique le meilleur nombre de clusters.

Les structures chimiques ont été représentées en utilisant RDKit [195]. Pour fournir une comparaison de similarité structurale, le taux de similarité a été calculé sur la base des empreintes circulaires Morgan de 1024 bits [196, 197] avec un rayon de 2 ($r = 2$ liaisons). Celles-ci sont basées sur le principe des Empreintes de Connectivité Étendues [198] telles qu'implémentées dans la bibliothèque RDKit [195]. Le logiciel PyMOL [199] a été utilisé pour l'inspection visuelle en 3D des résidus interactifs.

4.2.2 Simulations de dynamique moléculaire

Les poses dockées présentant les scores de docking les plus favorables ont été sélectionnées pour les simulations de DM, réalisées à l'aide de la suite de programmes AMBER16 [200]. Les états de protonation de la protéine ont été attribués pour un pH neutre à l'aide du serveur web H++ [201]. Le champ de force ff14SB [202] a été utilisé pour modéliser la protéine, tandis que les topologies des ligands et les fichiers de coordonnées ont été générés avec le module Antechamber [203]. Cet outil permet une conversion rapide des formats moléculaires, l'attribution des types d'atomes et le calcul des charges partielles. Les fichiers obtenus sont compatibles avec le champ de force GAFF [204].

Chaque complexe protéine-ligand a été solvaté dans une boîte octaédrique tronquée contenant des molécules d'eau, avec une distance de tampon de 10 Å entre la protéine et les bords de la boîte. Le modèle Optimal Point Charge (OPC) a été utilisé pour représenter les molécules d'eau [205], car il a été démontré qu'il reproduit plus fidèlement diverses propriétés du liquide en vrac par rapport à d'autres modèles d'eau à trois ou quatre points [206]. La charge du système a été neutralisée par l'ajout de quatre ions Na^+ . Pour reproduire une salinité physiologique de 0.15 M, 41 ions Na^+ et Cl^- supplémentaires ont été ajoutés. Les paramètres du champ de force Li-Merz 12-6 ont été utilisés pour modéliser les anions [206]. Le système ainsi construit a été relaxé par une minimisation d'énergie impliquant les molécules d'eau et les ions (10 000 cycles), suivie d'un chauffage progressif de l'ensemble du système à température et pression constantes (ensemble NPT, $P = 1$ bar), jusqu'à convergence du volume de la boîte de simulation à 310 K. Les minimisations ont été effectuées à l'aide de l'algorithme de descente de gradient le plus raide durant les 5000 premiers cycles, puis en utilisant l'algorithme du gradient conjugué pour les cycles restants [207]. Le thermostat de Langevin [208] et le barostat de Monte Carlo [209] ont été utilisés respectivement pour le contrôle de la température et de la pression. L'algorithme SHAKE [210] a été utilisé pour contraindre les liaisons impliquant des atomes d'hydrogène. Après réduction progressive des contraintes sur la protéine et le ligand, le système a été équilibré pendant 1 ns à pression et température constantes, puis à nouveau minimisé et relaxé avec des contraintes appliquées uniquement au squelette protéique, jusqu'à relâchement progressif de toutes les contraintes.

Pour chaque système, des simulations de production MD ont été effectuées sur une échelle de temps de 100 ns à 310 K et 1 bar, en utilisant un pas de temps de 2 fs. Des conditions aux limites périodiques ont été appliquées avec un couplage de pression (ensemble NPT), et les interactions électrostatiques à longue portée ont été traitées avec la méthode Particle Mesh Ewald [211], avec une coupure de 9 Å. Les coordonnées ont été sauvegardées toutes les 125 ps pour un traitement ultérieur.

Les calculs MM-GBSA (Molecular Mechanics Generalized-Born Surface Area) ont été réalisés à l'aide du code MMPBSA.py [212] afin d'obtenir l'énergie libre de liaison ($\Delta G_{\text{binding}}$), par sommation des termes d'énergie reflétant la différence (Δ) entre les états lié et non lié. Ce code implémente la relation suivante [213] :

$$\Delta G_{\text{binding}} = \Delta E_{\text{MM}} + \Delta G_{\text{GB}} + \Delta G_{\text{SA}} \quad (4.2)$$

où :

- ΔE_{MM} est la somme des deux termes standards MM électrostatique et de van der Waals ($\Delta E_{\text{MM}} = \Delta E_{\text{vdW}} + \Delta E_{\text{elec}}$);

- ΔG_{GB} , appelée composante polaire, est une quantité obtenue à l'aide du modèle GB [214]. Elle approxime le changement d'énergie de solvation électrostatique lors de la liaison ($\Delta G_{GB} = G_{GB}^{complex} - (G_{GB}^{receptor} + G_{GB}^{ligand})$);
- ΔG_{SA} est la contribution non polaire à l'énergie libre de solvation, estimée à partir d'une relation linéaire avec la surface accessible au solvant (SASA) [215].

4.2.3 ADMET

Le profile ADMET a été réalisé à l'aide du web-server SwissADME [216]. Ce serveur fournit une prédiction de plusieurs paramètres d'évaluation liés à la pharmacocinétique et à la toxicité des petites molécules, principalement basés sur des modèles d'apprentissage automatique intégrés.

4.3 Resultats et discussion

4.3.1 Analyse des données et apprentissage automatique

Comme décrit dans le chapitre précédent, afin d'assurer une forte capacité de distinction entre les deux classes, les composés chimiques avec une activité intermédiaire sont également considérés comme inactifs. La figure 4.1 montre la tendance des mesures IC_{50} au fil des années (pour la ERK2), ce qui exprime la disponibilité des données. La distribution de densité indique également une tendance continue à rapporter à la fois des composés bioactifs forts et faibles au fil du temps.

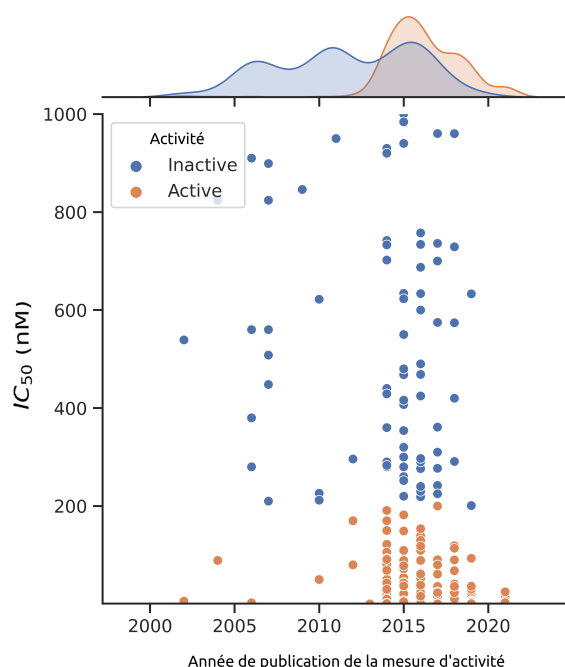


FIGURE 4.1 – Graphique des valeurs de IC_{50} selon l'année de publication de l'essai dans l'ensemble de données sélectionné.

Le modèle construit atteint une précision de 89% à la fois en validation croisée (10-fold CV) et en validation externe. Le tableau 4.1 résume les métriques d'évaluation du modèle et la

figure 4.2 montre la matrice de confusion et la courbe ROC du classifieur utilisé. La matrice de confusion révèle sa capacité à identifier correctement 89% des composés dans les deux classes ($\frac{185}{207}$). Ainsi, seulement 11% des instances de l'ensemble de test sont prédites de manière incorrecte. Tandis que 91% de la classe inactive est correctement prédite ($\frac{112}{123}$). Cela représente le caractère spécifique de la classification par apprentissage automatique (spécificité ou FPR). Ainsi, une bonne spécificité est observée [217]. De plus, 11% de prédictions incorrectes représentent une fonction coût acceptable, en particulier lors du traitement de grandes quantités de composés chimiques.

Parallèlement, la sensibilité mesure la proportion de composés actifs correctement identifiés. La sensibilité observée (87%) indique que le modèle est efficace pour identifier les composés actifs.

TABLEAU 4.1 – Informations pertinentes du modèle.

Dimensions des données d'entraînement	(622, 434)
Dimensions des données de test	(207, 434)
Score sur l'entraînement	0.98
Validation croisée	0.89
Score	0.89
Sensibilité	0.87
Spécificité	0.91
AUC	0.89

La courbe ROC (receiver operating characteristic) est une courbe de probabilité qui trace le taux de vrais positifs (TP) contre le taux de faux positifs (FP) à différentes valeurs de seuil. Elle montre la performance du classifieur à chaque seuil [142, 143]. L'utilisation de l'AUC comme résumé des courbes ROC est une pratique courante pour mesurer la puissance des processus de décision dichotomiques [143, 144]. La valeur AUC observée (0.89) indique une forte capacité à distinguer les actifs des inactifs (capacité à détecter plus de TN et TP que de FN et FP). Comme on peut le voir, il y a une surface substantielle sous la courbe TPR(FPR). La fiabilité du modèle est indiquée à la fois par les métriques d'évaluation observées et par sa résilience aux problèmes de sur/sous-apprentissage.

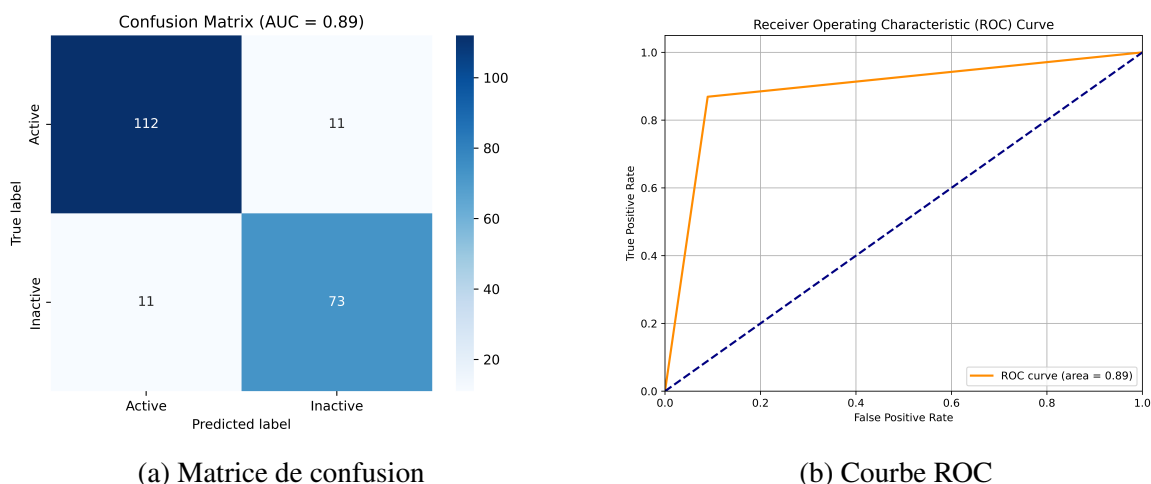


FIGURE 4.2 – Carte thermique de la matrice de confusion du classificateur et courbe ROC.

Le LBVS a été effectué en considérant trois bases de données : Le National Cancer Institute (NCI) [218], ZINC natural products [219], et Coconut [220]. tableau 4.2 présente le nombre de composés criblés et le nombre de substances prédites comme actives.

TABLEAU 4.2 – Rapport du criblage virtuel par modèle QSAR.

Base de données	Composés	Actifs
Coconut	103692	3055
NCI	58333	88
ZINC Prod. Nat.	10000	469
Total	172025	3612

Comme il est nécessaire de prendre en compte la validité des structures chimiques et la similitude avec des pistes médicamenteuses (lead-likeness), ces bases de données étaient plus grandes au départ, et ont été réduites au nombre de composés cité grâce à une étape de préparation. Cette dernière comprend un filtre basé sur la règle de Lipinski (Ro5 de Pfizer) [221] appliqué selon une masse moléculaire (MW) inférieure à 500 Da, un nombre d'accepteurs de liaisons hydrogène (HBA) ≤ 10 , un nombre de donneurs de liaisons hydrogène (HBD) ≤ 5 , et un coefficient octanol-eau (LogP) ≤ 5 . Ce filtre est appliqué dans la découverte de médicaments pour évaluer si un composé chimique a les propriétés nécessaires pour être administré par voie orale [222, 223].

4.3.2 Criblage docking et clustering

Le re-docking des ligands de référence (SCH772984 et Ulixertinib) dans la ERK2 a été réalisé pour valider le protocole de docking. Le calcul du RMSD (Root Mean Square Deviation) à partir d'un composé de référence (ligand co-cristallisé) permet de mesurer la qualité géométrique de la pose de liaison supposée par l'algorithme. Leurs scores de liaison à la protéine cible ont été évalués à -14.8 kcal/mol pour SCH772984 et -9.34 kcal/mol pour Ulixertinib. Le RMSD calculée entre les poses cristallographiques et les meilleures poses redockées a été évalué par le web-server Dock-RMSD [224] et a été de 0.508 Å pour SCH772984 et 0.542 Å

pour Ulixertinib, ce qui est considéré comme excellent (≤ 2 Å), et prouve que le protocole est fiable pour l'évaluation des poses et affinités de liaison de nos composés [225]. La figure 4.3 présente la superposition des poses DRX et redockées pour les deux références.

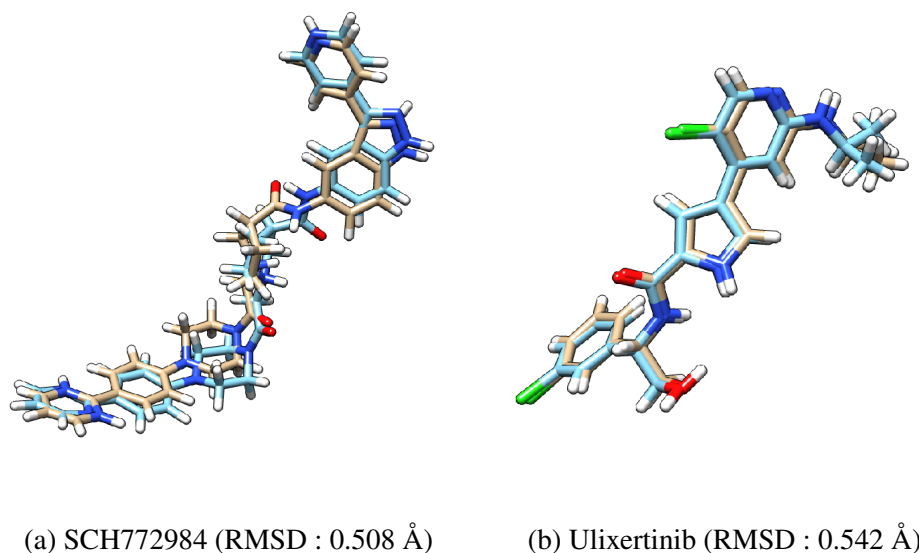


FIGURE 4.3 – Superposition des poses DRX et redockées.

Les 3612 composés prédits actifs ont tous été criblés dans le site de liaison de ERK2 (4qta), et les scores d'énergie libre (ΔG) ont révélé de bons résultats. Aucun composé n'a montré une meilleure affinité avec la kinase que le ligand SCH772984, mais de nombreuses structures ont été prédites avec une affinité inférieure à -12 kcal/mol. Les 53 meilleures molécules (basées sur le score ΔG) ont été sélectionnées et soumises à une procédure d'apprentissage automatique non supervisée (clustering), afin de choisir un candidat représentatif pour chaque type de structure.

Pour déterminer un nombre optimal de clusters, la méthode d'Elbow [226] a été testée pour les deux approches : K-Means [193] et le clustering spectral [227], sans convergence vers un nombre réduit de groupes pour les données étudiées. Par conséquent, le nombre de clusters a été défini à l'aide de la méthode silhouette [228].

Le nombre optimal de clusters est observé à $k = 6$ [229]. Ainsi, six clusters différents ont été créés à partir des descripteurs moléculaires. Les coordonnées des centroides dans chaque cluster ont été calculées, et les distances entre chaque point du cluster et son centroïde respectif ont été évaluées. Sur cette base, le composé représentatif le plus proche du centroïde a été sélectionné pour chaque groupe. Le ligand de référence a été retrouvé seul dans le sixième cluster, ce qui en fait le représentant unique de son type structural. Le tableau 4.3 présente les identifiants des composés dans chaque cluster, leur affinité de liaison (score prédit en kcal/mol) et les plages d'affinités dans chaque groupe.

Le SCH772984 a été retrouvé seul dans le dernier cluster (tableau 4.3), ce qui en fait le seul représentant. Les empreintes chimiques de Morgan (Morgan chemical fingerprints), également connues sous le nom d'empreintes circulaires, sont construites en appliquant l'algorithme

TABLEAU 4.3 – Identifiants des composés dans chaque cluster, leur affinité de liaison (prédite en kcal/mol) et les plages de scores dans chaque groupe (le représentant de chaque groupe est en rouge).

Cluster (A)		Cluster (B)		Cluster (C)		Cluster (D)		Cluster (ZN)		Cluster (38Z)	
ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score
CNP0012288	-13.8	CNP0310180	-13.7	ZINC20714006	-13.93	CNP0405698	-13.01	ZINC8792314	-12.58	REF 38Z	-14.8
CNP0012265	-13.74	CNP0372341	-13.31	CNP0154402	-13.7	CNP0409246	-12.63	-	-	-	-
CNP0045919	-13.6	CNP0402309	-13.28	CNP0256306	-13.23	CNP0151771	-12.56	-	-	-	-
ZINC15675938	-13.06	CNP0384303	-13.19	ZINC8792343	-12.77	ZINC96114512	-12.52	-	-	-	-
ZINC8878844	-13.01	CNP0235282	-13.08	CNP0342600	-12.72	CNP0190721	-12.51	-	-	-	-
CNP0295146	-13	CNP0414552	-12.97	CNP0139356	-12.64	CNP0322014	-12.46	-	-	-	-
CNP0046863	-12.85	CNP0334703	-12.91	CNP0392645	-12.63	ZINC85876885	-12.43	-	-	-	-
CNP0408092	-12.8	CNP0375843	-12.86	ZINC20610294	-12.61	CNP0271422	-12.4	-	-	-	-
CNP0137997	-12.75	CNP0198351	-12.81	CNP0284570	-12.54	CNP0114172	-12.39	-	-	-	-
CNP0012408	-12.68	CNP0402323	-12.65	CNP0327175	-12.53	-	-	-	-	-	-
CNP0408196	-12.65	CNP0261203	-12.54	CNP0358119	-12.51	-	-	-	-	-	-
CNP0114477	-12.57	CNP0264653	-12.5	CNP0395215	-12.48	-	-	-	-	-	-
CNP0012287	-12.47	CNP0411233	-12.5	-	-	-	-	-	-	-	-
CNP0012260	-12.47	CNP0403590	-12.48	-	-	-	-	-	-	-	-
CNP0277877	-12.46	CNP0120550	-12.43	-	-	-	-	-	-	-	-
CNP0012294	-12.4	-	-	-	-	-	-	-	-	-	-
	[-12.4, -13.8]		[-12.43, -13.7]		[-12.48, -13.93]		[-12.39, -13.01]	Plage des scores	[-12.58]		[-14.8]

de Morgan à un ensemble d'invariants atomiques fournis par l'utilisateur [195, 196]. Nous les utilisons pour évaluer la similarité de nos composés avec le ligand SCH772984 et les inhibiteurs connus de ERK.

La figure 4.4 montre le représentant sélectionné de chaque type structural, leur affinité de liaison à la ERK2, ainsi que le pourcentage de similarité avec le ligand de référence SCH772984. La figure 4.5 affiche les cartes de similarité par rapport à ce même ligand de référence, et dans le tableau 4.4 sont présentées les mesures de similarité par rapport aux composés actifs connus.

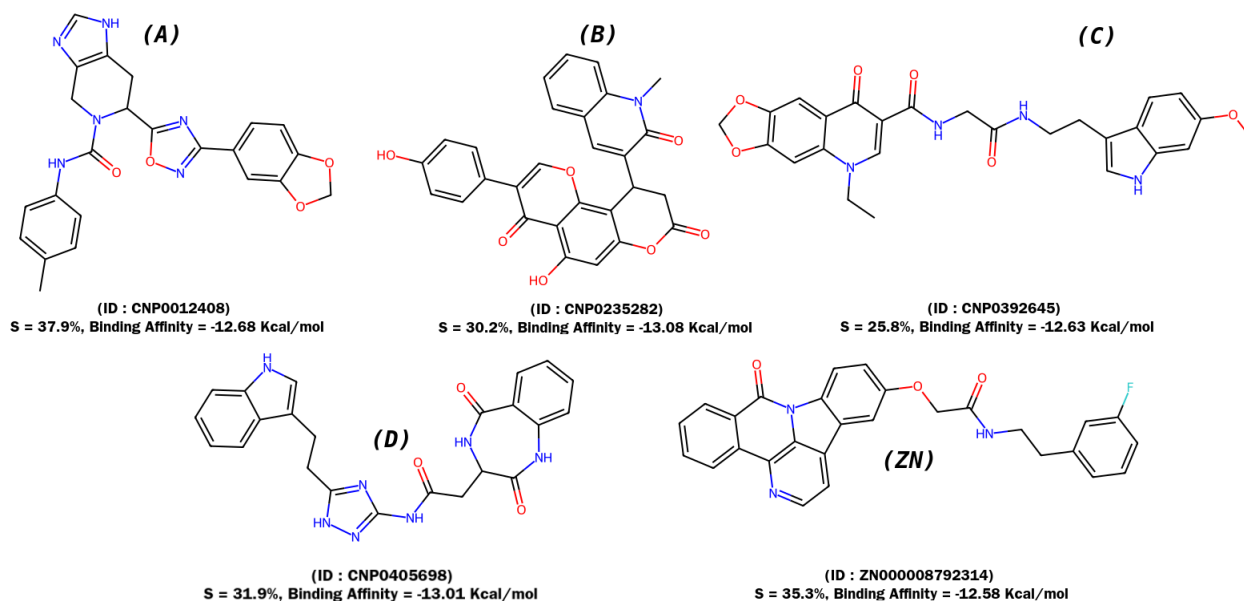


FIGURE 4.4 – Structures chimiques et identifiants des meilleurs représentants des 5 clusters, leurs scores de docking et leur taux de similarité avec le ligand de référence.

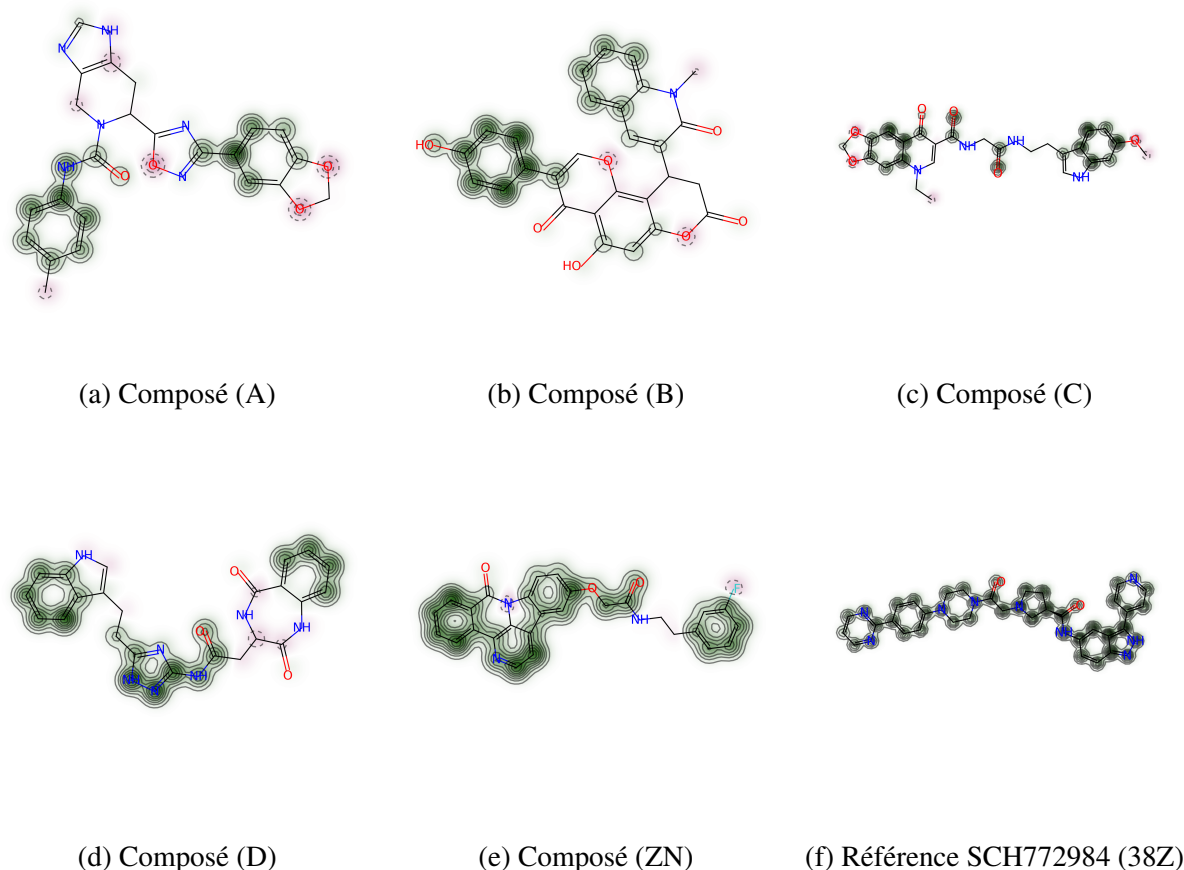


FIGURE 4.5 – Cartes de similarité chimique des 5 composés sélectionnés avec le ligand de référence SCH772984 (38Z). Méthode de coloration : vert : différence positive, gris : aucune variation de similarité, rose : différence négative.

TABEAU 4.4 – Pourcentage de similarité des composés sélectionnés avec les inhibiteurs connus de ERK.

Composé	Inhibiteurs connus de ERK					
	FR180204 [230]	TCS ERK 11e [231]	Selumetinib [232]	LRRK2-IN-1 [233]	AMG PERK 44 [234]	ERK5-IN-1 [235]
(A)	30.41	36.27	27.91	33.18	32.08	30.9
(B)	34.09	31.31	25.99	31.53	41.47	29.41
(C)	24.86	33.17	33.71	28.7	26.61	26.78
(D)	38.82	36.46	25.73	35.19	35.07	33.62
(ZN)	35.23	39.39	35.03	32.43	41.47	31.93
REF (38Z)	32.35	30.09	23.41	38.4	36.73	36.84

Les structures chimiques des composés sélectionnés sont différentes de celles des inhibiteurs connus de ERK ainsi que du ligand SCH772984. Le taux de similarité avec les inhibiteurs connus ne dépasse pas 41.5%, tandis qu'avec le ligand de référence, seules des similarités de 37.9%, 30.2%, 25.8%, 31.9% et 35.3% sont respectivement observées pour (A), (B), (C), (D) et (ZN).

Une différence positive significative peut être observée sur la carte de similarité du composé (ZN), en raison de la présence de deux cycles benzéniques, d'un cycle pyridine et d'un groupement carbonyle amide. Ces éléments interagissent également via différents types de contacts dans la pose de docking. Moins significativement, on observe le même comportement

pour les autres ligands (présence de cycles benzéniques, de carbonyles amides et de nombreux hétérocycles azotés). En revanche, une différence négative est notée avec la présence d'un cycle dioxolane dans (A) et (C), ainsi qu'un radical fluor dans (ZN). Les fragments cités jouent un rôle significatif dans la stabilité des complexes et sont souvent responsables de l'effet médicamenteux des molécules organiques.

Certains dérivés carboxamides ont déjà été identifiés pour leur fort potentiel anticancéreux [236]. Par ailleurs, certains polyamides bioactifs à base de pyrrole-imidazole sont connus pour leurs effets hors cible multiples [237]. Les motifs pyrrole, imidazole (et les hétérocycles azotés en général) sont présents dans notre pentade. Compte tenu des interactions possibles avec d'autres cibles kinases, nous soulignons l'importance de la sélectivité de cible. L'activité envers d'autres kinases a été vérifiée à l'aide de certains modèles de ML présentés dans le chapitre précédent. Les kinases concernées sont des enzymes périphériques à ERK2 (EGFR, IGF1R, BRAF, PI3K α , Akt1 et mTOR).

Les cascades de réactions enzymatiques montrent des relations complexes entre plusieurs kinases [236], et de nombreuses réactions catalytiques intermédiaires ont lieu depuis les récepteurs EGF/IGF (entrée cellulaire) jusqu'aux ERKs [238]. La procédure de vérification permet d'estimer la sélectivité de nos ligands. Le tableau 4.5 présente les résultats de cette procédure.

TABLEAU 4.5 – Résultats de la procédure de vérification à l'aide de classifieurs d'apprentissage automatique pour d'autres kinases.

Classificateur ML	EGFR	IGF1R	BRAF	Akt1	PI3K α	mTOR
(A)	Inactif ×	Inactif ×	Inactif ×	Inactif ×	Inactif ×	Inactif ×
(B)	Inactif ×	Inactif ×	Inactif ×	Inactif ×	Actif ✓	Actif ✓
(C)	Inactif ×	Inactif ×	Inactif ×	Inactif ×	Inactif ×	Inactif ×
(D)	Inactif ×	Inactif ×	Actif ✓	Inactif ×	Inactif ×	Actif ✓
(ZN)	Inactif ×	Inactif ×	Actif ✓	Actif ✓	Inactif ×	Inactif ×

Une attention particulière est accordée à l'inactivité prédite de tous les composés vis-à-vis des récepteurs transmembranaires EGF/IGF. Il s'agit d'une hypothèse positive concernant leur potentiel de pénétration cellulaire. De plus, (A) et (C) sont prédits inactifs pour toutes les kinases mentionnées ci-dessus, indiquant un potentiel caractère fortement sélectif. Le composé (ZN) est prédit actif vis-à-vis des enzymes BRAF et Akt1, (B) vis-à-vis de PI3K α et mTOR, tandis que (D) est estimé actif envers BRAF et mTOR. Ces résultats peuvent orienter de futures investigations expérimentales concernant les effets médicinaux anticancéreux des molécules étudiées.

4.3.2.1 Analyse de docking

Les poses de liaison des cinq ligands ont révélé des profils d'interaction prometteurs. La figure 4.6 présente les vues des interactions des cinq ligands dans leur conformation de liaison. (A) forme une liaison hydrogène (HB) à travers l'un des oxygènes de l'anneau dioxolane et l'atome d'hydrogène du groupe amine de la Tyr36 (2.92 Å). Une interaction π -cation se

produit à 3.87 Å entre le cycle benzénique adjacent et le NH_3^+ de la Lys54, tandis qu'une interaction hydrophobe est observée à 3.14 Å entre ce même groupe et la Val39. De plus, un empilement $\pi - \pi$ en forme de T peut être observé avec le groupe phényle de la Tyr36 (4.61 Å). Ces interactions peuvent être observées dans la figure 4.6a.

Le groupe amine de la Lys151 forme une HB à 2.33 Å avec le groupe carbonyle sur le tétrahydropyran du composé (B), et le groupe phénolique O-H du composé (B) joue à la fois un rôle de donneur et d'accepteur de HB par rapport aux groupes amine et carbonyle du résidu Met108 (~ 2.03 Å) (voir la figure 4.6b). Pour le composé (C), deux HB pertinentes se forment avec Gln105 et Ser153, respectivement à 2.72 et 2.65 Å, comme le montre la figure 4.6c, en plus de contacts hydrophobes avec plusieurs résidus, tels que la Met108.

Le composé (D) forme une HB à 3.33 Å avec l'Asp111 et un empilement $\pi - \pi$ pertinent à 3.73 Å entre le fragment indole et le groupe phényle de la Tyr36 (figure 4.6d). Ce dernier fragment donne naissance au même type d'interaction avec (A) et (ZN), pour lesquels un autre empilement π en forme de T (4.7 Å) peut être observé avec le cycle fluorobenzène. Cet anneau se trouve à 3.9 Å du groupe NH_3^+ de la Lys54, ce qui suggère un contact stabilisant π -cation (similairement observé pour le ligand (A)). De plus, deux HBs notables avec Ala35 et Asp167 sont observées respectivement à 3.2 et 2.3 Å, comme le montre la figure 4.6e.

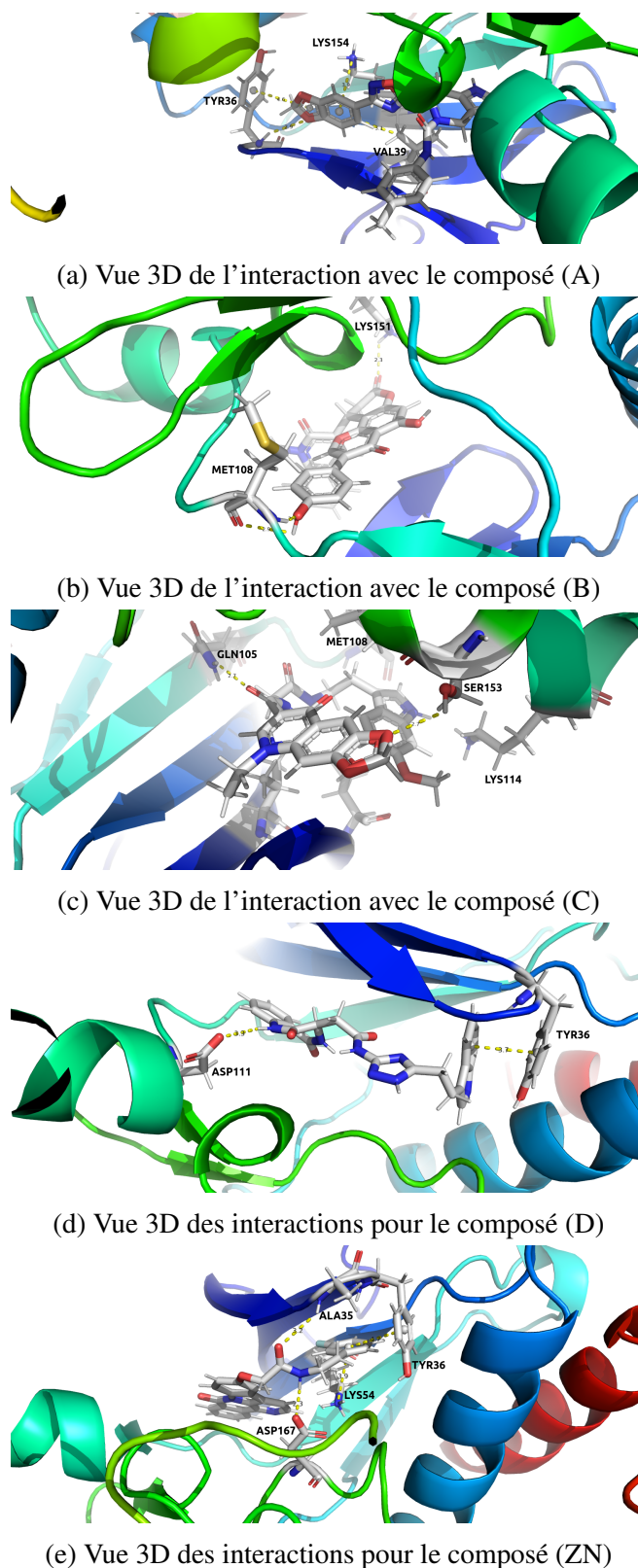


FIGURE 4.6 – Vues 3D des interactions avec la pose dockée des cinq composés (A-ZN) (l'unité de distance est l'Å).

Les interactions décrites ci-dessus favorisent une disposition stable des ligands présentés dans la protéine étudiée. Cela est renforcé par le fait que les résidus interagissant intensivement avec les cinq ligands sont les mêmes acides aminés clés connus du schéma d'inhibition de la

cible [239].

4.3.2.2 Dynamique moléculaire

Pour démontrer l'efficacité de notre criblage basé sur l'apprentissage automatique, nous commençons notre analyse par l'examen des valeurs de l'écart quadratique moyen (RMSD) des complexes protéine–ligand sélectionnés. Cela constitue un premier indicateur de la stabilité structurale des ligands prédits. Plus précisément, nous comparons les profils de RMSD des complexes formés par les composés (A)–(ZN) à celui du complexe de référence Ulixertinib. Dans la figure 4.7, nous rapportons les valeurs de RMSD pour le squelette protéique ainsi que pour les atomes lourds de la protéine et du ligand dans tous les systèmes.

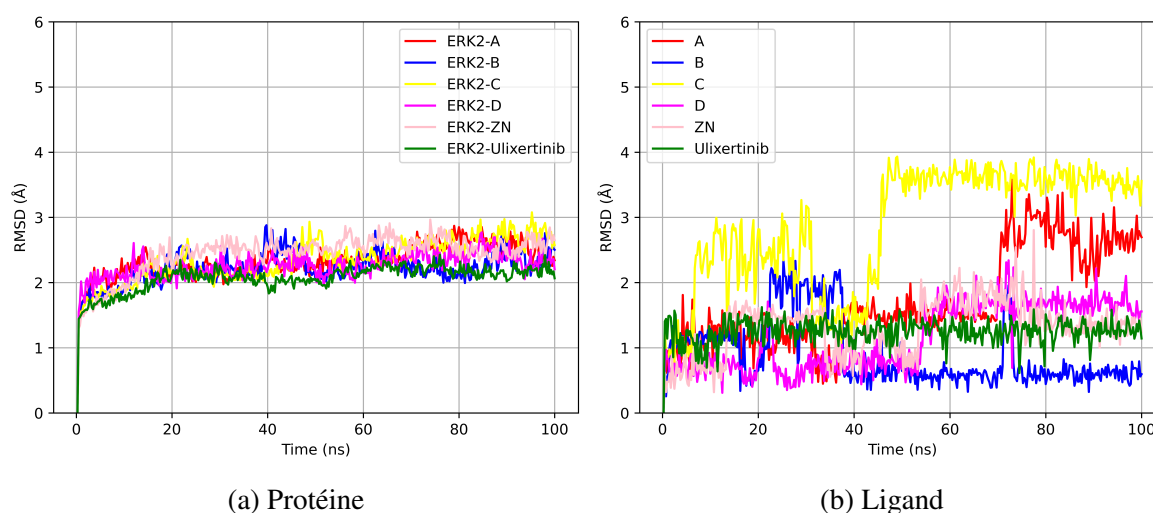


FIGURE 4.7 – Écarts quadratiques moyens calculés au cours du temps de simulation. À gauche, les valeurs de RMSD sont rapportées pour le squelette protéique (a), tandis que celles calculées pour les ligands sélectionnés sont présentées à droite (b).

Tout au long des 100 ns de simulation, les structures protéiques présentent des fluctuations limitées, atteignant rapidement des plages d'équilibre similaires dans tous les cas. Le RMSD moyen du squelette n'excède pas 2.44 Å (observé pour le complexe P–(ZN)), une valeur comparable à celle du complexe P–Ulixertinib (2.07 Å). Notamment, les complexes P–(B) et P–(D) présentent des RMSD moyens de 2.22 et 2.25 Å, respectivement, tandis que les complexes P–(A) (2.33 Å) et P–(C) (2.34 Å) restent également proches de la référence. Ces amplitudes de RMSD constamment faibles dans les cinq systèmes liés aux ligands sont cohérentes avec une stabilité structurale marquée du squelette protéique pour tous les complexes.

Cette observation est renforcée par l'analyse des valeurs de RMSD des ligands. Bien que les composés (A), (B) et (C) subissent des changements conformationnels plus importants, chacun d'eux présente des périodes prolongées de stabilité relative. Le ligand (A) conserve un RMSD proche de 1.5 Å par rapport à sa conformation initiale jusqu'à 70 ns, après quoi il subit un réarrangement plus marqué. Globalement, le ligand (B) présente moins de fluctuations, avec deux changements conformationnels modérés d'environ 2 Å, survenant entre 20–40 ns et autour de 73 ns. Une conformation stable est atteinte au cours des 25 dernières ns (environ 0.6 Å). Le ligand (C) se stabilise autour de 3.6 Å de sa pose initiale après 50 ns et jusqu'à 100 ns.

Les ligands (D) et (ZN) présentent un degré plus élevé de stabilité conformationnelle durant toute la phase de production. Le ligand (D) reste proche de 0.5 Å jusqu'à 55 ns, après quoi il se stabilise autour de 1.7 Å. Le ligand (ZN) fluctue jusqu'à 1.5 Å entre 15–35 ns, mais atteint ensuite un état d'équilibre autour de 1.7 Å à partir de 55 ns, à l'exception de deux déviations brèves de 2.7 et 2.8 Å vers 75 ns.

Pour résumer, comparés à l'ulixertinib, qui se stabilise autour de 1.3 Å (RMSD moyen de 1.25 Å), les cinq ligands montrent une stabilité conformationnelle modérée à bonne. Leurs valeurs moyennes de RMSD sont : 1.70 Å pour (A), 0.91 Å pour (B), 2.86 Å pour (C), 1.16 Å pour (D) et 1.30 Å pour (ZN). Ces résultats montrent un comportement comparable à celui du médicament de référence dans la plupart des cas, avec une stabilité presque identique pour (ZN) et des performances légèrement meilleures pour (B) et (D). En particulier, les ligands (D) et (ZN) se distinguent par leur remarquable stabilité conformationnelle.

L'analyse des énergies de liaison, réalisée au moyen des calculs MM-GBSA effectués sur les trajectoires de 100 ns, offre de nouvelles perspectives sur la stabilité des complexes protéine–ligand (P–L). Les composantes énergétiques détaillées ainsi que les énergies libres de liaison totales calculées sont présentées dans le tableau 4.6. Nous ferons référence aux termes définis dans l'équation 4.2.

TABLEAU 4.6 – Composantes énergétiques MM-GBSA et énergies libres de liaison (en kcal/mol) pour les cinq composés sélectionnés et l'ulixertinib en complexe avec la protéine. La définition des différents termes est fournie dans la équation 4.2.

Composé	(A)	(B)	(C)	(D)	(ZN)	Ulixertinib
ΔE_{vdW}	-49.57	-37.51	-44.71	-52.47	-49.58	-51.78
ΔE_{elec}	-13.06	-26.51	-9.68	-55.88	-20.84	-38.96
ΔE_{GB}	37.42	41.57	33.69	75.64	41.59	51.12
ΔE_{SA}	-5.86	-5.10	-5.40	-6.88	41.5	-6.63
$\Delta G_{Binding}$	-31.07	-27.55	-26.09	-39.59	-34.90	-46.25

Comme attendu, la référence médicamenteuse (ulixertinib) présente la plus forte affinité de liaison, avec une énergie libre de liaison de -46.25 kcal/mol. Parmi les ligands candidats, le composé (D) montre l'affinité la plus proche (-39.59 kcal/mol), suggérant une interaction tout aussi forte avec ERK2. Le complexe avec le composé (ZN) présente également une affinité de liaison importante (-34.90 kcal/mol). Les composés (A), (B) et (C) suivent avec des énergies libres de liaison de -31.07 , -27.55 et -26.09 kcal/mol respectivement, indiquant une stabilité plus faible mais néanmoins significative en comparaison avec la référence.

D'après les analyses conformationnelles et de liaison, le profil des ligands (B), (D) et (ZN) est prometteur, notamment en comparaison avec la molécule de référence (ulixertinib), qui interagit avec les mêmes acides aminés selon un schéma de connexion similaire. Cette dernière observation est renforcée par l'énergie libre de liaison de (D) et (ZN). La figure 4.8 montre la disposition des 4 composés au niveau du site de liaison de l'ERK2 après 50 ns, avec une vue sur les résidus principaux environnants.

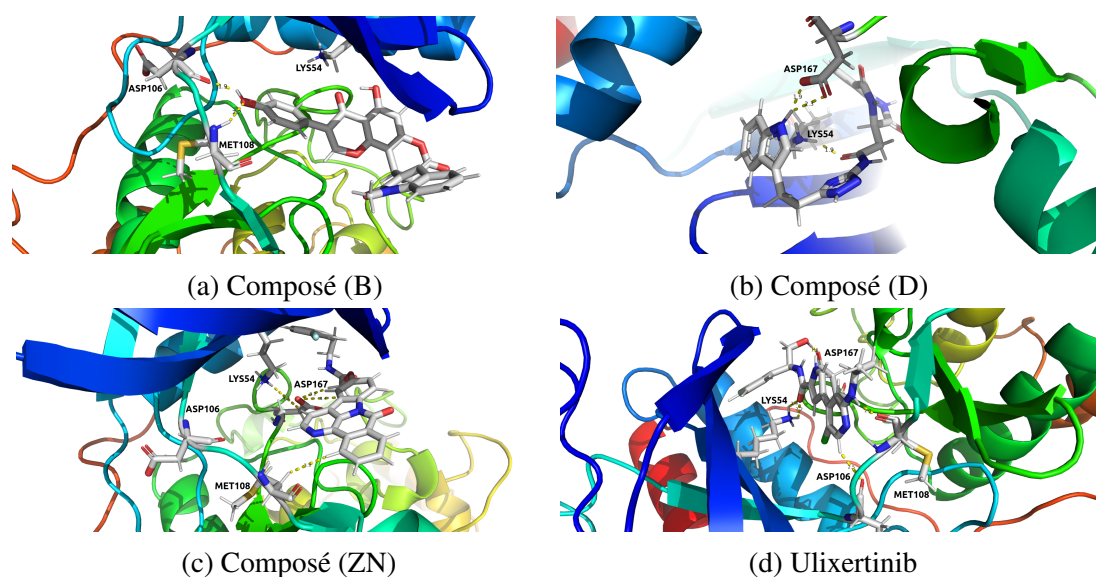


FIGURE 4.8 – Capture de la disposition des composés (B), (D), (ZN) et Ulixertinib dans la poche de liaison de la protéine après 50 ns de simulation (l'unité de distance est Å).

4.3.3 ADME-Tox

La pré-estimation du profil ADMET des composés bioactifs peut considérablement réduire les échecs liés à la pharmacocinétique en phase clinique [240, 241]. Le profilage des 5 composés fournit une estimation fiable de leurs propriétés pharmacocinétiques, notamment en comparaison avec les ligands de référence.

Étant donné qu'un filtre de Lipinski a été appliqué avant le criblage virtuel, aucun composé (à l'exception du ligand de référence) ne viole la règle des cinq (Ro5) de Lipinski. De plus, les règles de Ghose, Veber, Egan et Muegge [242–245] ne sont violées que par de légers dépassements en : poids moléculaire (PM) (490.51 et 481.45 g/mol pour (C) et (B)), réfractivité molaire (RM) (133.53 (ZN), 133.59 (C) et 133.83 (B)), et Surface Polaire Totale (SPT) (144.66 pour le (D)). Certains médicaments oraux approuvés par la FDA respectent toutes ces règles, tandis que d'autres ne les respectent pas, citons les exemples suivants : Selumetinib ne viole aucune règle de similarité à des médicaments oraux, alors que Fostemsavir et Lurbinectedin les violent toutes, et Selpercatinib viole les filtres de Lipinski et de Ghose [246]. Le composé (A) ne viole aucune règle de similarité aux médicaments oraux.

Trois composés sur cinq ne sont pas prédits comme étant des substrats de la p-glycoprotéine (p-gp). L'effet d'efflux de ce transporteur membranaire peut influencer l'absorption, la distribution, le métabolisme, l'excrétion et la toxicité des médicaments [247], puisqu'il est exprimé à des niveaux élevés dans de nombreux tissus (BHE, tractus gastro-intestinal (GIT), placenta, rein et foie [248–250, 250, 251]). Comme pour Ulixertinib, les modèles de machine learning du serveur prédisent que les molécules (B), (D) et (ZN) ne sont pas des substrats de la p-gp. En outre, (D) et (ZN) sont prédits comme étant des inhibiteurs du CYP3A4, ce qui n'est pas le cas pour (B).

Seuls (A) et (C) (contenant un cycle dioxolane) sont des substrats de la protéine de résistance aux médicaments multiples 1 (p-gp). Certains médicaments approuvés (comme ulixerti-

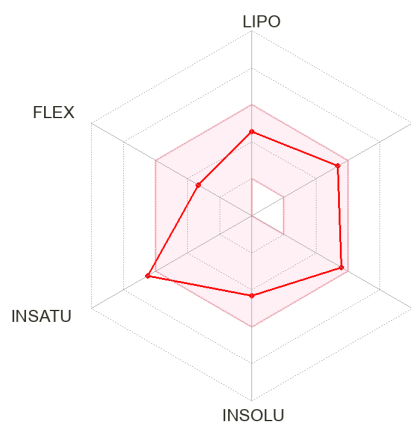
nib et regorafenib) ne sont pas prédits comme substrats de la p-gp, bien que d'autres le soient [97, 252]. Un score de biodisponibilité de 0.55 pour tous les composés présentés indique également de bonnes caractéristiques pharmacocinétiques [253].

Les valeurs de pourcentage d'absorption ont été calculées par : $ABS = 109 - [0.345 \times SPT]$, selon Zhao et al. [254, 255], et sont rapportées dans le tableau 4.7 avec d'autres propriétés physicochimiques. Le composé (ZN) présente un fort potentiel d'absorption (83.92%). De plus, 68.15%, 67.95%, 66.33% sont observés pour (A), (B) et (C) respectivement. La molécule (D) possède le plus faible potentiel d'absorption gastro-intestinale (GI). Ces valeurs expriment une estimation globale de la perméabilité à travers la membrane plasmique [253]. Comparés à la référence (38Z), les composés (A) et (B) montrent un meilleur profil d'absorption GI. Tandis que (ZN) est meilleur que les deux molécules de référence.

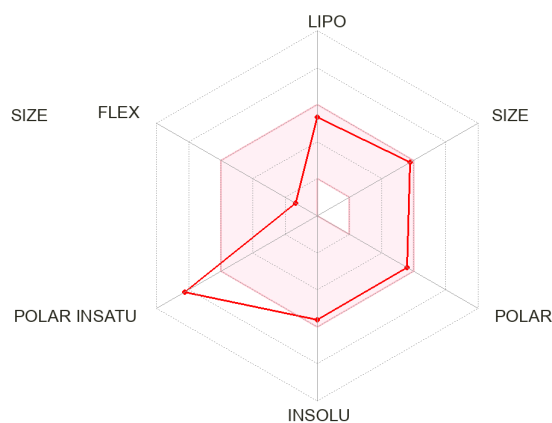
TABLEAU 4.7 – Propriétés physicochimiques des composés sélectionnés et des ligands de référence (le 38Z et Ulixertinib).

Composé	(A)	(B)	(C)	(D)	(ZN)	Réf (38Z)	Ulixertinib
PM (g/mol)	444.44	481.45	490.51	443.46	465.48	587.67	433.33
Atomes lourds	33	36	36	33	35	44	29
Atomes lourds aromatiques	22	26	19	20	26	27	17
Fraction Csp3	0.22	0.11	0.27	0.17	0.11	0.27	0.24
Liaisons rotatives	5	2	10	7	7	9	8
Accepteurs de liaison H	7	7	6	5	5	7	3
Donneurs de liaison H	2	2	3	5	1	2	4
Réfractivité molaire	121.05	133.83	133.59	128.38	133.53	179.38	116.42
SPT (Å ²)	118.4	118.97	123.68	144.66	72.70	123.24	90.04
%ABS	68.15%	67.95%	66.33%	59.09%	83.92%	66.48%	77,93%

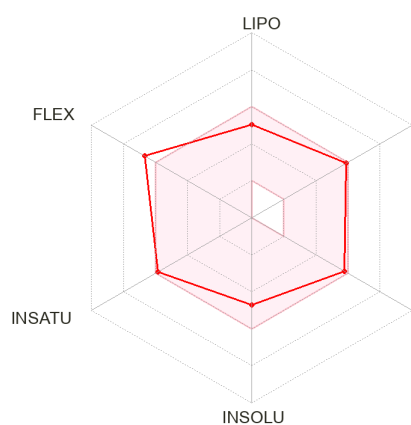
Les diagrammes radar des propriétés physicochimiques présentés dans la figure 4.9 montrent la probabilité de biodisponibilité orale de ces molécules. D'après les dépassements mineurs de la zone colorée, la potentiel d'utilisation comme médicament de la pentade est notable, en particulier pour (A) et (C). Le premier présente seulement un léger excès dans la direction de l'insaturation, causé par la valeur de la fraction Csp3 de 0.22 (< 0.25), et le second, par son taux de flexibilité dépassant 9 (10 liaisons rotatives). La fraction d'insaturation des molécules (ZN) et (B) (0.11), ainsi que celle du (D) (0.17), sont également inférieures à 0.25, ce qui déplace légèrement la ligne rouge du tracé au-delà de la zone colorée.



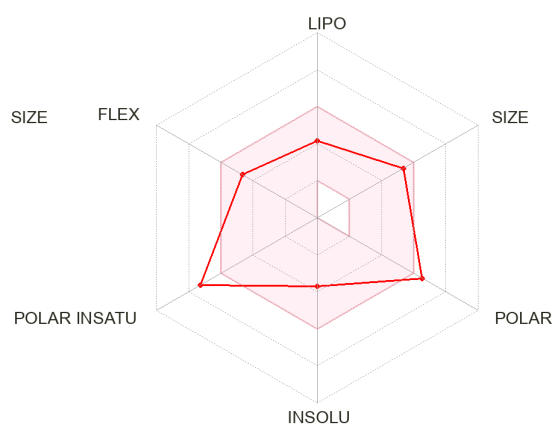
(a) Composé (A)



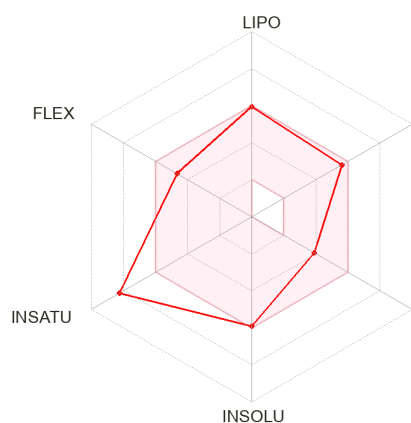
(b) Composé (B)



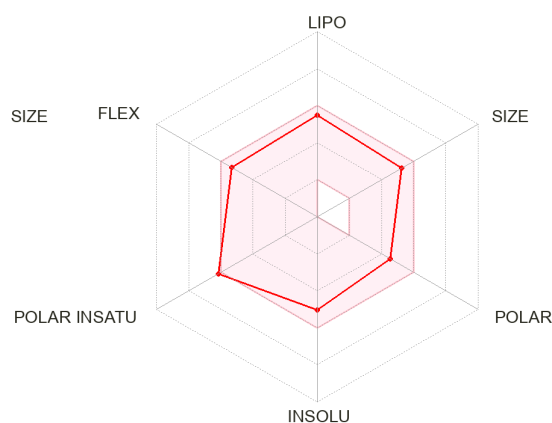
(c) Composé (C)



(d) Composé (D)



(e) Composé (ZN)



(f) Ulixertinib

FIGURE 4.9 – Les figures radar de biodisponibilité (la zone rose représente l'intervalle optimal pour chaque propriété). LIPO : XLOGP3 entre -0.7 et +5.0, SIZE : poids moléculaire entre 150 et 500 g/mol, POLAR : SPT entre 20 et 130 Å², INSOLU : logS ne dépassant pas 6, INSATU : fraction de carbones en hybridation sp³ supérieur ou égale à 0.25, et FLEX : pas plus de 9 liaison rotatives.

Les prédictions montrent également qu'aucune des cinq molécules ne peut traverser la barrière hémato-encéphalique (BHE). Dans le diagramme en oeuf bouilli présenté dans la figure 4.10, la valeur relativement élevée de la SPT du composé (D) (un excès de $4.66 \text{ \AA}^2$ [243]) entraîne une faible probabilité prédite de son absorption gastro-intestinale (GI). Ainsi, il est prédit en dehors de la zone d'absorption GI. Le profil ADMET des cinq ligands révèle leur fort potentiel de similarité médicamenteuse, en comparaison avec le SCH772984 et ulixertinib.

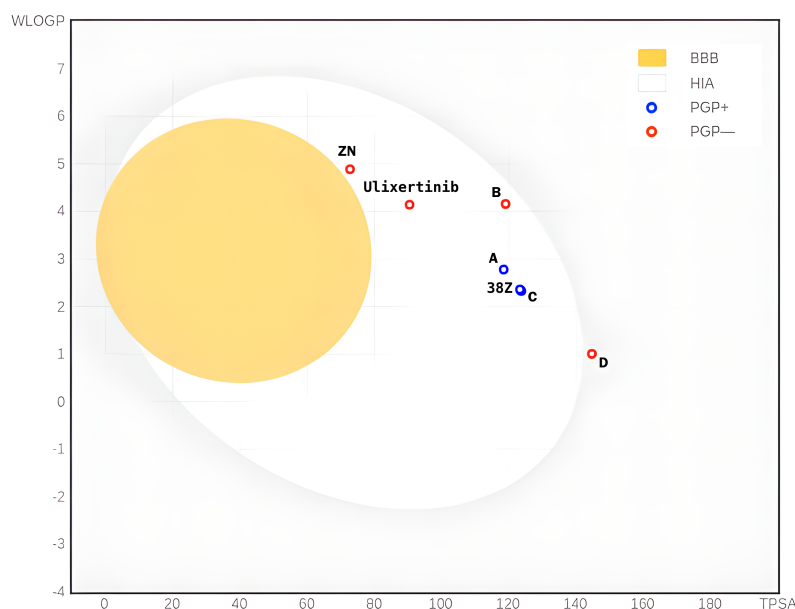


FIGURE 4.10 – Diagramme pharmacocinétique BOILED-Egg pour les composés sélectionnés.

Les résultats ci-dessus peuvent fournir des connaissances significatives sur le mécanisme d'inhibition de l'ERK2, les possibilités de régulation du signal, ainsi que sur la conception de petites substances de type médicamenteux pour une utilisation en thérapies ciblées anticancéreuses.

4.4 Conclusion

Dans ce chapitre, nous avons présenté la partie du travail dans laquelle nous avons collecté des données expérimentales de bioactivité liées à l'inhibition de la ERK2 protéine kinase. Les données brutes ont été nettoyées et organisées avant l'étiquetage de l'activité basé sur les valeurs de IC_{50} . Nous avons extrait des informations biochimiques à l'aide de l'apprentissage supervisé, afin de modéliser les relations qualitatives structure-activité et de cribler virtuellement trois bibliothèques chimiques. Un criblage basé sur le docking a également été effectué, aboutissant à la sélection des 53 meilleures molécules. Ces composés ont été regroupés en 5 clusters, et le meilleur représentant a été sélectionné pour chaque groupe. Le profil d'interaction avec la poche de liaison de la kinase a été analysé, et les propriétés médicamenteuses ont été estimées par un profil ADMET.

Les molécules (A) et (C) représentent un type structural avec une forte probabilité d'inhibition sélective, selon plusieurs modèles d'apprentissage automatique sur kinases. Le composé (B) représente une variété de squelette chimique, avec une stabilité conformationnelle significative au cours du temps de simulation. Enfin, les composés (D) et (ZN) constituent les meilleurs

candidats de notre étude, avec une activité prédite élevée et des propriétés médicamenteuses favorables. Ces deux composés ont montré une description énergétique similaire à celle du médicament de référence (ulixertinib), avec une toxicité plus faible.

Les 5 dérivés carboxamides étudiés ainsi que les 48 autres composés restants (voir tableau 4.3) constituent une piste importante pour le processus complexe d'inhibition de la ERK2, ainsi que pour la régulation potentielle d'autres activités tyrosine kinase. Nous espérons que des investigations expérimentales complémentaires des agents présentés permettront de confirmer leur activité biologique et leur potentiel en tant que traitements anticancéreux.

Conclusions et perspectives

Dans ce travail de recherche, nous avons utilisé des méthodes informatiques avancées pour approfondir la conception et l'identification de nouvelles substances chimiques inhibitrices des protéines kinases, dont l'inhibition joue un rôle central dans le cadre des nouvelles approches thérapeutiques anticancéreuses appelées "thérapies ciblées".

Nous avons collecté des données expérimentales sur différents inhibiteurs de 43 protéines kinases, que nous avons rigoureusement traitées et préparées en vue d'une procédure de modélisation des relations qualitatives structure-activité (QSARs).

Afin d'améliorer la modélisation de ces QSARs, nous avons proposé une nouvelle approche visant à augmenter la force prédictive. Cette approche est basée sur un ensemble d'apprentissage utilisant XGBoost comme algorithme primaire, et des réseaux de neurones profonds comme algorithme complémentaire. Les descripteurs initiaux, combinés aux prédictions issues de XGBoost, ont constitué, grâce à une technique d'ingénierie des probabilités, un nouveau jeu de données utilisé comme entrées pour les réseaux de neurones profonds.

Nous avons présenté une étude comparative des résultats obtenus par cette nouvelle approche et ceux issus des approches classiques. Cette étude a démontré le succès de la méthode proposée, montrant sa robustesse et son potentiel pour diverses applications en drug design.

Par ailleurs, nous avons utilisé l'un des modèles construits via l'approche classique pour le criblage virtuel de grandes bases de données chimiques, en vue de l'identification de nouvelles molécules inhibitrices de la ERK2 (également appelée MAPK1). Les hits obtenus ont été soumis à un second criblage virtuel basé sur la structure de la protéine. Les meilleurs candidats de ce dernier criblage ont ensuite été regroupés en cinq nouvelles classes d'inhibiteurs, chacune présentant un type structural prometteur, doté d'un fort potentiel inhibiteur et de propriétés pharmacocinétiques intéressantes.

Les cinq meilleurs représentants de chaque cluster ont fait l'objet d'une validation computationnelle, incluant d'abord une analyse de docking des interactions avec la cible, puis des calculs de trajectoires de DM visant à simuler leur comportement avec la protéine sur une période de temps.

Les candidats ont montré une stabilité comparable à celle des inhibiteurs connus, avec une toxicité légèrement améliorée, ce qui suggère un fort potentiel pharmacologique. Toutefois, ce potentiel ne pourra être pleinement attesté et optimisé que par des essais expérimentaux et des études complémentaires utilisant des techniques de bioisostérie.

La valeur ajoutée de ce travail de thèse se divise en deux parties. La première est le développe-

ment d'une nouvelle approche en machine learning hybride visant à améliorer la modélisation QSAR. Celle-ci permet d'atteindre une précision prédictive et un pouvoir de généralisabilité supérieurs à ceux des approches classiques. Quant à la deuxième partie, elle réside dans l'utilisation d'outils de CADD pour identifier de nouvelles classes d'inhibiteurs de la kinase ERK2.

Les principales conclusions sont les suivantes : concernant les QSARs, le potentiel remarquable des méthodes d'intelligence artificielle est encore renforcé par l'approche proposée, ouvrant ainsi de nouvelles perspectives et des avancées prometteuses si cette méthode est adoptée dans les études futures. Du côté des méthodes basées sur la structure de la protéine, nous concluons que le docking moléculaire a connu des avancées notables, facilitant la prédiction de l'affinité avec un haut niveau de précision, cependant, cette technique reste insuffisante pour valider la stabilité potentielle des complexes. Les simulations de DM se révèlent alors d'une efficacité redoutable pour la validation effective du potentiel de stabilité des candidats médicamenteux.

L'une des perspectives de ce travail est l'amélioration de l'approche hybride proposée, de façon à ce que les réseaux de neurones ne jouent plus uniquement le rôle d'un calibrateur sophistiqué des prédictions de XGBoost, mais participent plutôt à la décision finale de manière à ce que les probabilités issues de XGBoost aient un poids similaire à celui du reste des descripteurs moléculaires. Par ailleurs, l'intégration de plusieurs types de représentations moléculaires (au-delà des descripteurs physico-chimiques) pourrait également contribuer à améliorer la fiabilité des prédictions.

D'autre part, les ligands proposés pour l'inhibition de la ERK2 présentent un profil de stabilité moyen d'après les simulations de DM. Ces molécules pourraient faire l'objet de modifications structurales visant à améliorer les mécanismes de reconnaissance par la cible et, par conséquent, la stabilité des complexes.

Enfin, nous espérons voir l'application de l'approche proposée dans cette étude pour les QSARs, ainsi que la validation du potentiel thérapeutique des molécules identifiées par des essais expérimentaux et des études de bioisostérie, afin de marquer une avancée significative dans le domaine des nouvelles thérapies anticancéreuses ciblées.

Bibliographie

- [1] Dančík V. Wagner B. et al. Schenone, M. Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology*, 2013. DOI: 10.1038/nchembio.1199.
- [2] Li S. Wang Y. et al. Yang, Y. Protein tyrosine kinase inhibitor resistance in malignant tumors : molecular mechanisms and future perspective. *Signal Transduction and Targeted Therapy*, 2022. DOI: 10.1038/s41392-022-01168-8.

- [3] John Brognard and Tony Hunter. Protein kinase signaling networks in cancer. *Current Opinion in Genetics & Development*, 21(1) :4–11, 2011. DOI: 10.1016/j.gde.2010.10.012.
- [4] Asher Mullard. 2023 fda approvals. *Nature Reviews Drug Discovery*, 2024. DOI: d41573-024-00001-x.
- [5] R. Kannaiyan and D. Mahadevan. A comprehensive review of protein kinase inhibitors for cancer therapy. *Expert Review of Anticancer Therapy*, 2018. DOI: 10.1080/14737140.2018.1527688.
- [6] Dharmendra Kumar Yadav. Kinase inhibitors in cancer therapy. *Frontiers in Cell and Developmental Biology*, 10 :1020297, 2022. DOI: 10.3389/fcell.2022.1020297.
- [7] Jingyu Lee Chanjin Jung Bilal Shaker, Sajjad Ahmad and Dokyun Na. In silico methods and tools for drug discovery. *Computers in Biology and Medicine*, 2021. DOI: 10.1016/j.compbiomed.2021.104851.
- [8] Victor T. Sabe, Thandokuhle Ntombela, Lindiwe A. Jhamba, Glenn E.M. Maguire, Thavendran Govender, Tricia Naicker, and Hendrik G. Kruger. Current trends in computer aided drug design. *European Journal of Medicinal Chemistry*, 2021. DOI: 10.1016/j.ejmech.2021.113705.
- [9] Fatima Ardito, Michele Giuliani, Donatella Perrone, Giuseppe Troiano, and Lorenzo Lo Muzio. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *International Journal of Molecular Medicine*, 40(2) :271–280, 2017. <https://doi.org/10.3892/ijmm.2017.3036>.
- [10] Emma Cordover, Audrey Minden, S Lehman, and O Zhao. Signaling pathways downstream to receptor tyrosine kinases : Targets for cancer treatment. *Journal of Cancer Metastasis and Treatment*, 6 :45, 2020. <http://dx.doi.org/10.20517/2394-4722.2020.101>.
- [11] SM Keyse. Signal transduction, by gomperts, bd, kramer, im and tatham, per. *Journal of Cell Science*, 115(23) :4391–4391, 2002. 10.1016/B978-0-12-369441-6.X0001-3.
- [12] Lynette A Smyth and Ian Collins. Unlocking the potential of next-gen kinase inhibitors in clinical therapy. White paper, Sino Biological Inc., 2025. URL: <https://www.news-medical.net>.
- [13] Elvira An and John Brognard. Orange is the new black : Kinases are the new master regulators of tumor suppression. *IUBMB Life*, 2018. <https://doi.org/10.1002/iub.1981>.
- [14] Katla Kristjansdottir et al. A partial loss-of-function variant in akt2 is associated with reduced insulin-mediated glucose uptake. *Journal of Allergy and Clinical Immunology*, 2025. DOI: 10.1016/j.jaci.2024.10.002.
- [15] Sophie Deltour, Valerie Chopin, and Dominique Leprince. Modifications épigénétiques et cancer. *M/S : médecine sciences*, 21(4) :405–411, 2005. <https://id.erudit.org/iderudit/010775ar>.
- [16] Christine M. Lovly and Alice T. Shaw. Molecular pathways : Resistance to kinase inhibitors and implications for therapeutic strategies. *Clinical Cancer Research*, 20(9) :2249–2256, 2014. DOI: 10.1158/1078-0432.CCR-13-1610.
- [17] Zaira Spinello, Anna Fregnani, Laura Quotti Tubi, Livio Trentin, Francesco Piazza, and Sabrina Manni. Targeting protein kinases in blood cancer : Focusing on ck1a and ck2. *International Journal of Molecular Sciences*, 22(3716), 2021. DOI: 10.3390/ijms22073716.

- [18] Christine M Lovly and Alice T Shaw. Molecular pathways : Resistance to kinase inhibitors and implications for therapeutic strategies. *Clinical Cancer Research*, 20(9) :2249–2256, 2014. DOI: 10.1158/1078-0432.CCR-13-1610.
- [19] & Pinna L. A. Ruzzene, M. Addiction to protein kinase ck2 : a common denominator of diverse cancer cells ? *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 2010. DOI: 10.1016/j.bbapap.2009.07.018.
- [20] Lynette A Smyth and Ian Collins. Measuring and interpreting the selectivity of protein kinase inhibitors. *Journal of Chemical Biology*, 2(3) :131–151, 2009. DOI: 10.1007/s12154-009-0023-9.
- [21] Ali E Yesilkanal, Gary L Johnson, Alexandre F Ramos, and Marsha Rich Rosner. New strategies for targeting kinase networks in cancer. *Journal of Biological Chemistry*, 297(4) :101128, 2021. DOI: 10.1016/j.jbc.2021.101128.
- [22] Francesca Musumeci and Silvia Schenone. Unlocking potential and limits of kinase inhibitors : The highway to enhanced cancer targeted therapy, 2024. <https://doi.org/10.3390/pharmaceutics16050625>.
- [23] Quentin Devin. Nouvelles perspectives dans le traitement des gist avancées. *Centre Leon Berard*, 2025. <https://centreonberard.fr/institution/actualites/nouvelles-perspectives-dans-le-traitement-des-gist-avancees>.
- [24] M. L. Barreiro and A. K. Kümmerle. Applications of the pharmacophore concept in natural product inspired drug design. *Molecular Informatics*, 2020. DOI: 10.1002/minf.202000059.
- [25] Jens Meiler Yidan Tang, Rocco Moretti. Recent advances in automated structure-based de novo drug design. *Journal of Chemical Information and Modeling*, 2024. DOI: 10.1021/acs.jcim.4c00247.
- [26] John C. Dearden. The history and development of quantitative structure-activity relationships (qsars). *International Journal of Quantitative Structure–Property Relationships*, 2017. DOI: 10.4018/ijqspr.2017070104.
- [27] Kunal Roy and Supratik Kar. 3d-qsar in drug design-a review. *Current Topics in Medicinal Chemistry*, 2002. DOI: 10.2174/156802610790232260.
- [28] CD Selassie, SB Mekapati, and RP Verma. Qsar : then and now. *Current Topics in Medicinal Chemistry*, 2(12) :1357–1379, 2002. DOI: 10.2174/1568026023392823.
- [29] Sivaramakrishnan M. & Rabiya M. Prema, V. A concise review on role of qsar in drug design. *Asian Journal of Research in Chemistry*, 2023. DOI: 10.52711/0974-4150.2023.00076.
- [30] Bernard Testa. Qsar-hansch analysis and related approaches in drug design. *Trends in Pharmacological Sciences*, 1995. DOI: 10.1016/s0165-6147(00)89046-x.
- [31] Tian Y. S. Kawashita N. Moriwaki, H. and T. Takagi. Mordred : a molecular descriptor calculator. *Journal of Cheminformatics*, 2018. DOI: 10.1186/s13321-018-0258-y.
- [32] Yi Ding, Minchun Chen, Chao Guo, Peng Zhang, and Jingwen Wang. Molecular fingerprint-based machine learning assisted qsar model development for prediction of ionic liquid properties. *Journal of Molecular Liquids*, 326 :115212, 2021. DOI: 10.1016/j.molliq.2020.115212.
- [33] Dejun Jiang et al. Could graph neural networks learn better molecular representation for drug discovery ? *Research Square*, 2020. DOI: 10.21203/rs.3.rs-79416/v1.

- [34] W. Liu, P.; Long. Current mathematical methods used in qsar/qspr studies. *International Journal of Molecular Sciences*, 2009. DOI: 10.3390/ijms10051978.
- [35] Sild S. & Maran U. Ruusmann, V. Qsar databank repository : open and linked qualitative and quantitative structure–activity relationship models. *Journal of Cheminformatics*, 2015. DOI: 10.1186/s13321-015-0082-6.
- [36] Svante Wold, Erik Johansson, Marina Cocchi, et al. Pls : partial least squares projections to latent structures. pages 523–550, 1993. DOI: 10.1002/0470845015.cpa012.
- [37] Gabriele Cruciani Manuel Pastor and Sergio Clementi. Smart region definition : a new way to improve the predictive ability of 3d-qsar. *Journal of Medicinal Chemistry*, 1997. DOI: 10.1021/jm9608016.
- [38] Rolf Ergon. *Principal component regression (PCR) and partial least squares regression (PLSR)*, chapter 8, pages 121–142. John Wiley & Sons, Ltd, 2014. DOI: 10.1002/9781118434635.ch08.
- [39] Alexander Tropsha Alexander Golbraikh. Beware of q^2 ! *Journal of Molecular Graphics and Modelling*, 2002. DOI: 10.1016/s1093-3263(01)00123-1.
- [40] Alexander Tropsha. Best practices for qsar model development, validation, and exploitation. *Molecular Informatics*, 29(6-7) :476–488, 2010. DOI: 10.1002/minf.201000061.
- [41] M Abdullahi, GA Shallangwa, and A Uzairu. In silico qsar and molecular docking simulation of some novel aryl sulfonamide derivatives as inhibitors of h5n1 influenza a virus subtype. *Beni-Suef University Journal of Basic and Applied Sciences*, 2020. DOI: 10.1186/s43088-019-0023-y.
- [42] Kunal Roy Supratik Kar and Jerzy Leszczynski. Applicability domain : a step toward confident predictions and decidability for qsar modeling. *Protocol*, 2018. DOI: 10.1007/978-1-4939-7899-1_6.
- [43] Alexander Tropsha, Olexandr Isayev, Alexandre Varnek, Gisbert Schneider, and Artem Cherkasov. Integrating qsar modelling and deep learning in drug discovery : the emergence of deep qsar. *Nature Reviews Drug Discovery*, 23(2) :141–155, 2024. DOI: 10.1038/s41573-023-00832-0.
- [44] Zhenxing Wu, Minfeng Zhu, Yu Kang, Elaine Lai-Han Leung, Tailong Lei, Chao Shen, Dejun Jiang, Zhe Wang, Dongsheng Cao, and Tingjun Hou. Do we need different machine learning algorithms for qsar modeling? a comprehensive assessment of 16 machine learning algorithms on 14 qsar data sets. *Briefings in Bioinformatics*, 22(4) :bbaa321, 2021. DOI: 10.1093/bib/bbaa321.
- [45] Tripathi-G. Singh D. B. Jain P. & Kumar A. Priya, S. Machine learning approaches and their applications in drug discovery and design. *Chemical Biology & Drug Design*, 2022. DOI: 10.1111/cbdd.14057.
- [46] Abdulhamit Subasi and Ergun Ercelebi. Classification of eeg signals using neural network and logistic regression. *Computer Methods and Programs in Biomedicine*, 78(2) :87–99, 2005. DOI: 10.1016/j.cmpb.2004.10.009.
- [47] Mohamed Oussama Mousser, Khairedine Kraim, Fouad Chafaa, and Mohamed Brahimi. Machine learning-based qsar classifications for pim kinases inhibition prediction : Towards the neoplastic in silico drug design. In *International Conference on Disruptive Technologies : Innovations & Interdisciplinary Considerations*, pages 99–109. Springer, 2022. DOI: 10.1007/978-3-031-59329-1_8.

- [48] Željko Vujović et al. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6) :599–606, 2021. DOI: 10.14569/i-jacsa.2021.0120670.
- [49] Mohamed Oussama Mousser, Brahim Matougui, Fouad Chafaa, and Mohamed Abdesselem Dems. Enhancing predictive modeling with xgboost-engineered probabilities and deep neural networks : A hybrid approach for building reliable kinase inhibition qsar models. *Journal of Molecular Graphics and Modelling*, 2025. DOI: 10.1016/j.jmgm.2025.109216.
- [50] Kevser Kübra Kırboğa and Mesut Işık. Explainable artificial intelligence in the design of selective carbonic anhydrase i-ii inhibitors via molecular fingerprinting. *Journal of Computational Chemistry*, 45(18) :1530–1539, 2024. DOI: 10.1002/jcc.27335.
- [51] McInnes et al. Uniform manifold approximation and projection (umap) and its variants. *Journal of Open Source Software*, 2018. DOI: 10.21105/joss.00861.
- [52] Anna C Belkina, Carly O Ciccolella, Rene Anno, Rachel Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications*, 10(1) :5415, 2019. DOI: 10.1038/s41467-019-13055-y.
- [53] Pedro Contreras and Fionn Murtagh. Hierarchical clustering. *Handbook of cluster analysis*, pages 103–123, 2015. DOI: 10.1201/b19706.
- [54] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. Density-based spatial clustering of applications with noise. 240(6), 1996. DOI: <https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf>.
- [55] Garrett M Morris and Marguerita Lim-Wilby. Molecular docking. In *Methods in Molecular Biology*, volume 443, pages 365–382. Humana Press, 2008. DOI: 10.1007/978-1-59745-177-2_19.
- [56] N. D. Kulkarni, S. J. Momin, L. P. Jain, R. B. More, S. R. Jadhav, and M. B. Lungase. A comprehensive review on molecular docking in drug discovery. *International Journal of Scientific Research and Technology*, 02 :386–395, 2025. DOI: 10.5281/zenodo.15632400.
- [57] Rahul Haque, Mukti Mohammad, and Md Maidul Islam. Molecular docking an important tool for drug designing. *Modern Approaches in Drug Designing*, 1 :000518, 2018. DOI: 10.31031/MADD.2018.01.000518.
- [58] Fernando D Prieto-Martínez, Marcelino Arciniega, and José L Medina-Franco. Molecular docking : current advances and challenges. *TIP revista especializada en ciencias químico-biológicas*, 2019. DOI: 10.22201/fesz.23958723e.2018.0.143.
- [59] Jin Li, Ailing Fu, and Le Zhang. An overview of scoring functions used for protein-ligand interactions in molecular docking. *Interdisciplinary Sciences : Computational Life Sciences*, 11 :320–328, 2019. DOI: 10.1007/s12539-019-00327-w.
- [60] Sheng-You Huang, Sam Z Grinter, and Xiaoqin Zou. Scoring functions and their evaluation methods for protein-ligand docking : recent advances and future directions. *Physical Chemistry Chemical Physics*, 12 :12899–12908, 2010. DOI: 10.1039/c0cp00151a.
- [61] A. Panssar, T.; Poso. Binding affinity via docking : fact and fiction. *Molecules*, 2018. DOI: 10.3390/molecules23081899.
- [62] Raquel Dias and Walter Filgueira de Azevedo Jr. Molecular docking algorithms. *Current Drug Targets*, 9 :1040–1047, 2008. DOI: 10.2174/138945008786949432.

- [63] Sara Shamsian, Babak Sokouti, and Siavoush Dastmalchi. Benchmarking different docking protocols for predicting the binding poses of ligands complexed with cyclooxygenase enzymes and screening chemical libraries. *Bioimpacts*, 14 :29955, 2023. DOI: 10.34172/bi.2023.29955.
- [64] Number Analytics. Molecular simulation of protein-ligand interactions : Methods and applications.
- [65] Scott A. Hollingsworth et al. Molecular dynamics simulation for all. *Neuron*, 99(6) :1129–1142, 2018. DOI: 10.1016/j.neuron.2018.08.011.
- [66] Kirill Zinovjev, Birgit Schiøtt, and Joseph O’Connell. Enlighten2 : molecular dynamics simulations of protein–ligand systems made accessible. *Bioinformatics*, 36(20) :5104–5106, 2020. DOI: 10.1093/bioinformatics/btaa643.
- [67] Nian Wu, Rui Zhang, Xinyu Peng, Lu Fang, Kuan Chen, and Juho S Jestilä. Elucidation of protein–ligand interactions by multiple trajectory analysis methods. *Physical Chemistry Chemical Physics*, 26(8) :6903–6912, 2024. DOI: 10.1039/D3CP03492E.
- [68] Flashcards World. Flashcards for "dynamique moléculaire". <https://flashcards.world/flashcards/sets/505402eb-30db-4ce2-8346-ed269fff3b8f/>.
- [69] Justin LP Benesch and Brian T Ruotolo. Mass spectrometry of protein-ligand interactions : new insights from the gas phase. *Current Opinion in Structural Biology*, 28 :11–19, 2014. DOI: 10.1016/j.sbi.2014.07.001.
- [70] Vytautas Gapsys, Servaas Michielssens, Daniel Seeliger, and Bert L de Groot. Applying absolute free energy perturbation molecular dynamics to diffusively binding ligands. *Journal of Chemical Theory and Computation*, 11(8) :3463–3471, 2015. DOI: 10.1021/acs.jctc.5c00121.
- [71] Schrödinger. Fep+. <https://www.schrodinger.com/platform/products/fep/>.
- [72] Andrea Cuzzolin, Sara Pavan, Marco Furlan, and Massimo Sgobba. Qualitative estimation of protein–ligand complex stability through thermal titration molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 62(23) :5945–5958, 2022. DOI: 10.1021/acs.jcim.2c00995.
- [73] Talagayev et al. Openmmdl—simplifying the complex : Building, simulating, and analyzing protein–ligand systems in openmm. *Journal of Chemical Information and Modeling*, 2024. DOI: 10.1021/acs.jcim.4c02158.
- [74] Pedro EM Lopes, Olgun Guvench, and Alexander D MacKerell Jr. Current status of protein force fields for molecular dynamics simulations. *Methods in Molecular Biology*, 1215 :47–71, 2015. DOI: 10.1007/978-1-4939-1465-4_3.
- [75] Yogesh Tadarwal. *Accurate Force Fields for Spectroscopic Studies of Protein–Ligand Interactions and Self-Assembly Structures*. PhD thesis, KTH Royal Institute of Technology, 2024.
- [76] Ingrosso F. Segalina A. Caramori S. Foggi P. & Pastore M. Prampolini, G. Dynamical and environmental effects on the optical properties of an heteroleptic ru(ii)–polypyridine complex. *Journal of Chemical Theory and Computation*, 2018. DOI: 10.1021/acs.jctc.8b01031.
- [77] Advanced HPC. Amber. <https://www.advancedhpc.com/pages/amber>.

- [78] Junmei Wang, Robert M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9) :1157–1174, 2004. DOI: 10.1002/jcc.20035.
- [79] Wei Yang Tai-Sung Lee Junmei Wang Xibing He, Viet H. Man. A fast and high-quality charge model for the next generation general amber force field. *The Journal of Chemical Physics*, 2020. DOI: 10.1063/5.0019056.
- [80] Brooks Research Group. Charmm. <https://brooks.chem.lsa.umich.edu/register/>.
- [81] Academic CHARMM. Academic charmm. <https://academiccharmm.org/>.
- [82] Biosoft. Amber. <https://www.biosoft.com/software/amber>.
- [83] GROMACS. About gromacs. <https://www.gromacs.org/about.html>.
- [84] Theoretical and Computational Biophysics Group (TCBG), University of Illinois at Urbana-Champaign. Namd. (<https://www.ks.uiuc.edu/Research/namd/>).
- [85] Theoretical and Computational Biophysics Group (TCBG), University of Illinois at Urbana-Champaign. Namd. (<https://tcbg.illinois.edu/Research/namd/>).
- [86] Schrödinger. Desmond. <https://www.schrodinger.com/platform/products/ms-desmond/>.
- [87] Schrödinger. Desmond. <https://www.schrodinger.com/platform/products/desmond/>.
- [88] Alexander C Brueckner, Benjamin Shields, Palani Kirubakaran, Alexander Suponya, Manoranjan Panda, Shana L Posy, Stephen Johnson, and Sirish Kaushik Lakkaraju. Mdfit : automated molecular simulations workflow enables high throughput assessment of ligands-protein dynamics. *Journal of Computer-Aided Molecular Design*, 38(1) :24, 2024. DOI: 10.1007/s10822-024-00564-2.
- [89] Justin A Lemkul. Umbrella sampling. http://www.mdtutorials.com/gmx/umbrella/01_pdb2gmx.html.
- [90] Yosef Yarden Soma Ghosh, Ilaria Marrocco. Roles for receptor tyrosine kinases in tumor progression. *Advances in Cancer Research*, 2020. DOI: 10.1016/bs.acr.2020.04.002.
- [91] Laura Rinaldi, Francesco Chiuso, Emanuela Senatore, Domenica Borzacchiello, Luca Lignitto, Rosa Iannucci, Rossella Delle Donne, Mariano Fuggi, Carla Reale, Filomena Russo, et al. Downregulation of praja2 restrains endocytosis and boosts tyrosine kinase receptors in kidney cancer. *Communications Biology*, 7(1) :208, 2024. DOI: 10.1038/s42003-024-05823-4.
- [92] Vinay Prasad and Sham Mailankody. Research and development spending to bring a single cancer drug to market and revenues after approval. *JAMA Internal Medicine*, 177(11) :1569–1575, 2017. DOI: 10.1001/jamainternmed.2017.3601.
- [93] Katerina Trajanoska, Claude Bhérier, Daniel Taliun, Sirui Zhou, J Brent Richards, and Vincent Mooser. From target discovery to clinical drug development with human genetics. *Nature*, 620(7975) :737–745, 2023. DOI: 10.1038/s41586-023-06388-8.
- [94] Laurence SJ Roope. The economic challenges of new drug development. *Journal of Controlled Release*, 345 :275–277, 2022. DOI: 10.1016/j.jconrel.2022.03.026.

- [95] Antonio Lavecchia. Machine-learning approaches in drug discovery : methods and applications. *Drug Discovery Today*, 20(3) :318–331, 2015. DOI: 10.1016/j.drudis.2014.10.012.
- [96] Abdul Wadood, Nasir Ahmed, Lubna Shah, Ayaz Ahmad, Hamad Hassan, Sulaiman Shams, et al. In-silico drug design : An approach which revolutionarised the drug discovery process. *OA Drug Design & Delivery*, 1(1) :3, 2013. URL: Scispace.
- [97] Abbas Salimi, Jong Hyeon Lim, Jee Hwan Jang, and Jin Yong Lee. The use of machine learning modeling, virtual screening, molecular docking, and molecular dynamics simulations to identify potential vegfr2 kinase inhibitors. *Scientific Reports*, 12(1) :18825, 2022. DOI: 10.1038/s41598-022-22992-6.
- [98] Nannan Zhou, Yuan Xu, Xian Liu, Yulan Wang, Jianlong Peng, Xiaomin Luo, Mingyue Zheng, Kaixian Chen, and Hualiang Jiang. Combinatorial pharmacophore-based 3d-qsar analysis and virtual screening of fgfr1 inhibitors. *International Journal of Molecular Sciences*, 16(6) :13407–13426, 2015. DOI: 10.3390/ijms160613407.
- [99] Wang H. & Yan A. Qin, R. Classification and qsar models of leukotriene a4 hydrolase (lta4h) inhibitors. *SAR and QSAR in Environmental Research*, 2021. DOI: 10.1080/1062936x.2021.1910862.
- [100] Miao Yu, Qiong Gu, and Jun Xu. Discovering new pi3k α inhibitors with a strategy of combining ligand-based and structure-based virtual screening. *Journal of computer-aided molecular design*, 32(2) :347–361, 2018. DOI: 10.1007/s10822-017-0092-8.
- [101] X.; Zhao D.; Zou Y.; Huang X.; Wang J.; Zhang C Li, B.; Kang. Machine learning models combined with virtual screening and molecular docking to predict human topoisomerase i inhibitors. *molecules*, 2019. DOI: 10.3390/molecules24112107.
- [102] Aleksander Jankowski Miron B. Kursa and Witold R. Rudnicki. Boruta—a system for feature selection. *Fundamenta Informaticae*, 2010. DOI: 10.3233/fi-2010-288.
- [103] Hassan Almukadi et al. Combining machine learning and structure-based approaches to develop oncogene pim kinase inhibitors. *Frontiers in Chemistry*, 11 :1137444, 2023. DOI: 10.3389/fchem.2023.1137444.
- [104] Xinliang Yu. Prediction of chemical toxicity to tetrahymena pyriformis with four-descriptor models. *Ecotoxicology and Environmental Safety*, 190 :110146, 2020. DOI: 10.1016/j.ecoenv.2019.110146.
- [105] Deng J. & Liu M Liu, Q. Classification models for predicting the antimalarial activity against plasmodium falciparum. *SAR and QSAR in environmental research*, 2020. DOI: 10.1080/1062936x.2020.1740890.
- [106] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system, 2016. arXiv:1603.02754.
- [107] Robert P Sheridan, Wei Min Wang, Andy Liaw, Junshui Ma, and Eric M Gifford. Extreme gradient boosting as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 56(12) :2353–2360, 2016. DOI: 10.1021/acs.jcim.6b00591.
- [108] Tianqi Chen and Carlos Guestrin. XGBoost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. DOI: 10.1145/2939672.2939785.
- [109] Irsan Hardi T. R. Novandy, Ghalieb Mutig Idroes. Machine learning approach to predict axl kinase inhibitor activity for cancer drug discovery using xgboost. *Journal of Soft Computing and Data Mining*, 2024. DOI: 10.30880/jscdm.2024.05.01.004.

- [110] Yujia Tian Yunyang Zhao Xiaoyang Pang Guo Li, Jiaxuan Li and Aixia Yan. Machine learning-based classification models for non-covalent bruton's tyrosine kinase inhibitors. *Molecular Diversity*, 2023. DOI: 10.1007/s11030-023-10696-6.
- [111] Gede Surya Mahendra Airlangga et al. A hybrid gradient boosting and neural network model for predicting urban happiness. *Machine Learning and Knowledge Extraction*, 7(1) :4, 2025. DOI: 10.3390/make7010004.
- [112] Mahmoud G Abdolrasol, S M Suhail Hussain, Taha Selim Ustun, Md Ruhul Amin Sarker, Mahammad A Hannan, Pin Jern Ker, Shafiqur Rahman, Mohd Mansor, Kashem M Muttaqi, and Zhao Yang Dong. Artificial neural networks based optimization techniques : A review. *Electronics*, 10(21) :2689, 2021. DOI: 10.3390/electronics10212689.
- [113] Manoj Kumar Mishra Inakoti Ramesh Raja Surya Kiran Chebrolu B Swarna R Anand Babu, Vishwa Priya V. TabPFN : A transformer that solves small tabular classification problems in a second. *International Journal of Engineering, Science, and Information Technology*, 2025. DOI: 10.52088/ijesty.v5i3.1146.
- [114] Ravid Shwartz-Ziv and Amitai Armon. Tabular data : Deep learning is not all you need. *Information fusion*, 81 :84–90, 2022. DOI: 10.1016/j.inffus.2021.11.011.
- [115] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35 :507–520, 2022. NeurIPS 2022 Proceedings.
- [116] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. 2004. DOI: 10.1145/1015330.1015432.
- [117] Jeff Heaton. An empirical analysis of feature engineering for predictive modeling. In *SoutheastCon 2016*, pages 1–6. IEEE, 2016. DOI: 10.1109/secon.2016.7506650.
- [118] David H Wolpert. Stacked generalization. *Neural Networks*, 5(2) :241–259, 1992. DOI: 10.1016/S0893-6080(05)80023-1.
- [119] Saso Džeroski and Bernard Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54(3) :255–273, 2004. DOI: 10.1023/b:mach.0000015881.36452.6e.
- [120] Olutomilayo Olayemi Petinrin and Faisal Saeed. Stacked ensemble for bioactive molecule prediction. *IEEE Access*, 7 :153952–153957, 2019. DOI: 10.1109/access.2019.2945422.
- [121] Nalini Schaduangrat, Nutta Homdee, and Watshara Shoombuatong. Stacker : a novel smiles-based stacked approach for the accelerated and efficient discovery of $\text{er}\alpha$ and $\text{er}\beta$ antagonists. *Scientific Reports*, 13(1) :22994, 2023. DOI: 10.1038/s41598-023-50393-w.
- [122] Ahmad Zainudin, Love Allen Chijioke Ahakonye, Rubina Akter, Dong-Seong Kim, and Jae-Min Lee. An efficient hybrid-dnn for ddos detection and classification in software-defined iiot networks. *IEEE Internet of Things Journal*, 10(10) :8491–8504, 2022. DOI: 10.1109/jiot.2022.3196942.
- [123] Preethi Devan and Neelu Khare. An efficient xgboost–dnn-based classification model for network intrusion detection system. *Neural Computing and Applications*, 32(16) :12499–12514, 2020. DOI: 10.1007/s00521-020-04708-x.
- [124] Miodrag Zivkovic, Nebojsa Bacanin, Milos Antonijevic, Bosko Nikolic, Goran Kvascev, Marina Marjanovic, and Nikola Savanovic. Hybrid cnn and xgboost model tuned by

- modified arithmetic optimization algorithm for covid-19 early diagnostics from x-ray images. *Electronics*, 11(22) :3798, 2022. DOI: 10.3390/electronics11223798.
- [125] Chen Zhao, Xia Zhao, Zhao Li, and Qiong Zhang. Xgboost-dnn mixed model for predicting driver's estimation on the relative motion states during lane-changing decisions : a real driving study on the highway. *Sustainability*, 14(11) :6829, 2022. DOI: 10.3390/su14116829.
- [126] Xiaowei He, Siqi Li, Xin Tian He, Wenqiang Wang, Xiang Zhang, and Bin Wang. A novel ensemble learning model combined xgboost with deep neural network for credit scoring. *Journal of Information Technology Research (JITR)*, 15(1) :1–18, 2022. DOI: 10.4018/jitr.299924.
- [127] David Mendez, Anna Gaulton, A Patr'icia Bento, Jon Chambers, Marleen De Veij, Eloy F'elix, Mar'ia Paula Magari nos, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. ChEMBL : towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1) :D930–D940, 2019. DOI: 10.1093/nar/gky1075.
- [128] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010. DOI: 10.25080/Majora-92bf1922-00a.
- [129] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585 :357–362, 2020. DOI: 10.1038/s41586-020-2649-2.
- [130] Chun Wei Yap. Padel-descriptor : An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7) :1466–1474, 2011. DOI: 10.1002/jcc.21707.
- [131] Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients : appropriate use and interpretation. *Anesthesia & analgesia*, 126(5) :1763–1768, 2018. DOI: 10.1213/ane.0000000000002864.
- [132] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97 :105524, 2020. DOI: 10.1016/j.asoc.2019.105524.
- [133] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011. DOI: 10.5555/1953048.2078195.
- [134] Leo Breiman. Random forests. *Machine Learning*, 45(1) :5–32, 2001. DOI: 10.1023/A:1010933404324.
- [135] William S Noble. What is a support vector machine? *Nature Biotechnology*, 24(12) :1565–1567, 2006. DOI: 10.1038/nbt1206-1565.
- [136] Mesut Gumus and Mustafa S Kiran. Crude oil price forecasting using xgboost. In *2017 International conference on computer science and engineering (UBMK)*, pages 1100–1103. IEEE, 2017. DOI: 10.1109/ubmk.2017.8093500.

- [137] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786) :504–507, 2006. DOI: 10.1126/science.1127647.
- [138] Diederik P Kingma and Jimmy Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014. arXiv:1412.6980.
- [139] Peter Ripphausen, Britta Nisius, and Jürgen Bajorath. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today*, 16(9-10) :372–376, 2011. DOI: 10.1016/j.drudis.2011.02.011.
- [140] Wei Dong, Yimiao Huang, Barry Lehane, and Guowei Ma. Xgboost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. *Automation in Construction*, 114 :103155, 2020. DOI: 10.1016/j.autcon.2020.103155.
- [141] Deepak Kumar Sharma, Mayukh Chatterjee, Gurmehak Kaur, and Suchitra Vavilala. Deep learning applications for disease diagnosis. In *Deep learning for medical applications with unique data*, pages 31–51. Elsevier, 2022. DOI: 10.1016/b978-0-12-824145-5.00005-8.
- [142] Guoping Zeng. On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics-Theory and Methods*, 49(9) :2080–2093, 2020. DOI: 10.1080/03610926.2019.1568485.
- [143] Christopher D Brown and Herbert T Davis. Receiver operating characteristics curves and related decision measures : A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80(1) :24–38, 2006. DOI: 10.1016/j.chemolab.2005.05.004.
- [144] Charles X Ling, Jin Huang, and Harry Zhang. Auc : a better measure than accuracy in comparing learning algorithms. In *Advances in Artificial Intelligence : 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11–13, 2003, Proceedings 16*, pages 329–341. Springer, 2003. DOI: 10.1007/3-540-44886-1_25.
- [145] Ziqi Li. Extracting spatial effects from machine learning model using local interpretation method : An example of shap and xgboost. *Computers, Environment and Urban Systems*, 96 :101845, 2022. DOI: 10.1016/j.compenvurbsys.2022.101845.
- [146] Muhammad Hassam, Jawwad A Shamsi, Ajmal Khan, Ahmed Al-Harrasi, and Reaz Uddin. Prediction of inhibitory activities of small molecules against pantothenate synthetase from mycobacterium tuberculosis using machine learning models. *Computers in Biology and Medicine*, 145 :105453, 2022. DOI: 10.1016/j.compbimed.2022.105453.
- [147] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*. John Wiley & Sons, 2008. DOI: 10.1002/9783527613106.
- [148] Derek Van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23) :5938–5951, 2022. DOI: 10.1021/acs.jcim.2c01073.
- [149] Beth Apsel, Jimmy A Blair, Beatriz Gonzalez, Tamim M Nazif, Morri E Feldman, Brian Aizenstein, Randy Hoffman, Roger L Williams, Kevan M Shokat, and Zachary A Knight. Targeted polypharmacology : discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nature Chemical Biology*, 4(11) :691, 2008. DOI: 10.1038/nchem-bio.117.

- [150] L Bryan Ray and Thomas W Sturgill. Rapid stimulation by insulin of a serine/threonine kinase in 3t3-l1 adipocytes that phosphorylates microtubule-associated protein 2 in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 84(6) :1502–1506, 1987. DOI: 10.1073/pnas.84.6.1502.
- [151] L Bryan Ray and TW Sturgill. Characterization of insulin-stimulated microtubule-associated protein kinase. rapid isolation and stabilization of a novel serine/threonine kinase from 3t3-l1 cells. *Journal of Biological Chemistry*, 263(25) :12721–12727, 1988. DOI: 10.1021/bi00215a038.
- [152] Hugo Lavoie, Jessica Gagnon, and Marc Therrien. Erk signalling : a master regulator of cell behaviour, life and fate. *Nature reviews Molecular cell biology*, 21(10) :607–632, 2020. DOI: 10.1038/s41580-020-0255-7.
- [153] Amardeep Singh Dhillon, Suzanne Hagan, O Rath, and W Kolch. Map kinase signalling pathways in cancer. *Oncogene*, 26(22) :3279–3290, 2007. DOI: 10.1038/sj.onc.1210421.
- [154] Jean-François Tanti and Jennifer Jager. Cellular mechanisms of insulin resistance : role of stress-regulated serine kinases and insulin receptor substrates (irs) serine phosphorylation. *Current Opinion in Pharmacology*, 9(6) :753–762, 2009. DOI: 10.1172/jci10583.
- [155] Eun Kyung Kim and Eui-Ju Choi. Pathological roles of mapk signaling pathways in human diseases. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1802(4) :396–405, 2010. DOI: 10.1016/j.bbadis.2009.12.009.
- [156] William E Tidyman and Katherine A Rauen. The rasopathies : developmental syndromes of ras/mapk pathway dysregulation. *Current opinion in genetics & development*, 19(3) :230–236, 2009. DOI: 10.1016/j.gde.2009.04.001.
- [157] Mauricio Burotto, Victoria L Chiou, Jung-Min Lee, and Elise C Kohn. The mapk pathway across different malignancies : a new perspective. *Cancer*, 120(22) :3446–3456, 2014. DOI: 10.1002/cncr.28864.
- [158] Li Ma, Zhenbang Chen, Hediye Erdjument-Bromage, Paul Tempst, and Pier Paolo Pandolfi. Phosphorylation and functional inactivation of tsc2 by erk : implications for tuberous sclerosis and cancer pathogenesis. *Cell*, 121(2) :179–193, 2005. DOI: 10.1016/j.cell.2005.02.031.
- [159] Reiko Sugiura, Ryosuke Satoh, and Teruaki Takasaki. Erk : a double-edged sword in cancer. erk-dependent apoptosis as a potential therapeutic strategy for cancer. *Cells*, 10(10) :2509, 2021. DOI: 10.3390/cells10102509.
- [160] Libero Santarpia, Scott M Lippman, and Adel K El-Naggar. Targeting the mapk–ras–raf signaling pathway in cancer therapy. *Expert opinion on therapeutic targets*, 16(1) :103–119, 2012. DOI: 10.1517/14728222.2011.645805.
- [161] Muralidhar Beeram, Amita Patnaik, and Eric K Rowinsky. Raf : a strategic target for therapeutic development against cancer. *Journal of Clinical Oncology*, 23(27) :6771–6790, 2005. DOI: 10.1200/jco.2005.08.036.
- [162] Cornelia Braicu, Mihail Buse, Constantin Busuioc, Rares Drula, Diana Gulei, Lajos Raduly, Alexandru Rusu, Alexandru Irimie, Atanas G Atanasov, Ondrej Slaby, et al. A comprehensive review on mapk : a promising therapeutic target in cancer. *Cancers*, 11(10) :1618, 2019. DOI: 10.3390/cancers11101618.
- [163] Sejeong Shin, Christopher A Dimitri, Sang-Oh Yoon, William Dowdle, and John Blenis. Erk2 but not erk1 induces epithelial-to-mesenchymal transformation via def

- motif-dependent signaling events. *Molecular Cell*, 38(1) :114–127, 2010. DOI: 10.1016/j.molcel.2010.02.020.
- [164] Roser Buscà, Jacques Pouyssegur, and Philippe Lenormand. Erk1 and erk2 map kinases : specific roles or functional redundancy ? *Frontiers in Cell and Developmental Biology*, 4 :53, 2016. DOI: 10.3389/fcell.2016.00053.
- [165] Xavier Deschênes-Simard, Filippas Kottakis, Sylvain Meloche, and Gerardo Ferbeyre. Erks in cancer : friends or foes ? *Cancer Research*, 74(2) :412–419, 2014. DOI: 10.1158/0008-5472.can-13-2381.
- [166] Aroon S. Karra and Melanie H. Cobb. Map kinase pathways : Functions and modulation. *Tocris Scientific Review Series*, 2016. DOI: 10.1101/cshperspect.a011254.
- [167] Lilien N Voong, Allison R Slater, Sebila Kratovac, and Drew E Cressman. Mitogen-activated protein kinase erk1/2 regulates the class ii transactivator. *Journal of Biological Chemistry*, 283(14) :9031–9039, 2008. DOI: 10.1074/jbc.m706487200.
- [168] Jessie M. English and Melanie H. Cobb. Pharmacological inhibitors of mapk pathways. *Trends in Pharmacological Sciences*, 23(1) :40–45, 2002. DOI: 10.1016/S0165-6147(00)01865-4.
- [169] Leilei Fu, Siwei Chen, Gu He, Yi Chen, and Bo Liu. Targeting extracellular signal-regulated protein kinase 1/2 (erk1/2) in cancer : an update on pharmacological small-molecule inhibitors. *Journal of Medicinal Chemistry*, 65(20) :13561–13573, 2022. DOI: 10.1016/b978-0-12-803505-4.00013-8.
- [170] Ciccarelli C. & Zani B. M. Marampon, F. Biological rationale for targeting mek/erk pathways in anti-cancer therapy. *International journal of molecular sciences*, 2019. DOI: 10.3390/ijms20102530.
- [171] Lei Li, Min Huang, Leping Li, Yi Chen, Shuai Tang, Yi Su, Ruiping Dong, Jian Ding, and Meiyu Geng. Hh2710, a highly potent and selective erk1/2 inhibitor for the treatment of mapk mutant tumors. *Cancer Research*, 80(16 Supplement) :3760–3760, 2020. DOI: 10.1158/1538-7445.am2020-3760.
- [172] Shripad V Bhagwat, William T McMillen, Shufen Cai, Baohui Zhao, Matthew Whitesell, Weihua Shen, Lisa Kindler, Robert S Flack, Wenjuan Wu, Bryan Anderson, et al. Erk inhibitor ly3214996 targets erk pathway–driven cancers : A therapeutic approach toward precision medicine. *Molecular cancer therapeutics*, 19(2) :325–336, 2020. DOI: 10.1158/1535-7163.mct-19-0183.
- [173] Ryan J Sullivan, Jeffrey R Infante, Filip Janku, Deborah Jean Lee Wong, Jeffrey A Sosman, Vicki Keedy, Manish R Patel, Geoffrey I Shapiro, James W Mier, Anthony W Tolcher, et al. First-in-class erk1/2 inhibitor ulixertinib (bvd-523) in patients with mapk mutant advanced solid tumors : results of a phase i dose-escalation and expansion study. *Cancer Discovery*, 8(2) :184–195, 2018. DOI: 10.1158/2159-8290.cd-17-1119.
- [174] Andrew H Ko, Mark Zalupski, Raed Moh’d Taiseer Al-Rajabi, Khalid Matin, Deirdre Jill Cohen, Smitha S Krishnamurthi, Brent Kreider, Jessica A Box, Caroline Emery, Martin Teresk, et al. A phase 2 basket trial of ulixertinib (bvd-523) in combination with hydroxychloroquine in patients with advanced gastrointestinal malignancies harboring mapk pathway mutations (bvd-523-hcq)., 2023.
- [175] Xiaoli Pan, Junping Pei, Aoxue Wang, Wen Shuai, Lu Feng, Faqian Bu, Yumeng Zhu, Lan Zhang, Guan Wang, and Liang Ouyang. Development of small molecule extracellular signal-regulated kinases (erks) inhibitors for cancer therapy. *Acta Pharmaceutica Sinica B*, 12(5) :2171–2192, 2022. DOI: 10.1016/j.apsb.2021.12.022.

- [176] James F. Blake, John J. Gaudino, Jason De Meese, Peter Mohr, Mark Chicarelli, Hongqi Tian, Rustam Garrey, Allen Thomas, Christopher S. Siedem, Michael B. Welch, Gabrielle Kolakowski, Robert Kaus, Michael Burkard, Matthew Martinson, Huifen Chen, Brian Dean, Danette A. Dudley, Stephen E. Gould, Patricia Pacheco, Sheerin Shahidi-Latham, Weiru Wang, Kristina West, Jianping Yin, John Moffat, and Jacob B. Schwarz. Discovery of 5,6,7,8-tetrahydropyrido[3,4-d]pyrimidine inhibitors of erk2. *Bioorganic and Medicinal Chemistry Letters*, 24(12) :2635–2639, 2014. DOI: 10.1016/j.bmcl.2014.04.068.
- [177] Li Ren, Jonas Grina, David Moreno, James F. Blake, John J. Gaudino, Rustam Garrey, Andrew T. Metcalf, Michael Burkard, Matthew Martinson, Kevin Rasor, Huifen Chen, Brian Dean, Stephen E. Gould, Patricia Pacheco, Sheerin Shahidi-Latham, Jianping Yin, Kristina West, Weiru Wang, John G. Moffat, and Jacob B. Schwarz. Discovery of highly potent, selective, and efficacious small molecule inhibitors of erk1/2. *Journal of Medicinal Chemistry*, 58(4) :1976–1991, 2015. DOI: 10.1021/jm501921k.
- [178] Xin Linghu et al. Discovery and development of gdc-0994 : A selective and efficacious small molecule inhibitor of erk1/2. *Complete Accounts of Integrated Drug Discovery and Development : Recent Examples from the Pharmaceutical Industry Volume 1*, 2018. DOI: 10.1021/bk-2018-1307.ch001.
- [179] Wen Shuai, Huan Xiao, Panpan Yang, Yiwen Zhang, Faqian Bu, Yongya Wu, Qiu Sun, Guan Wang, and Liang Ouyang. Structure-guided discovery and pre-clinical assessment of novel (thiophen-3-yl)aminopyrimidine derivatives as potent erk1/2 inhibitors. *Journal of Medicinal Chemistry*, 67(9) :7227–7247, 2024. DOI: 10.1021/acs.jmedchem.3c02392.
- [180] Liqin Wang Haojie Jin and René Bernards. Rational combinations of targeted cancer therapies : background, advances and challenges. *Nature Reviews Drug Discovery*, 2022. DOI: 10.1038/s41573-022-00615-z.
- [181] Yan Wu, Joel R Walker, Michael Westberg, Lin Ning, Michelle Monje, Thomas A Kirkland, Michael Z Lin, and Yichi Su. Kinase-modulated bioluminescent indicators enable noninvasive imaging of drug activity in the brain. *ACS Central Science*, 9(4) :719–732, 2023. DOI: 10.1021/acscentsci.3c00074.
- [182] Apirat Chaikuad, Eliana MC Tacconi, Jutta Zimmer, Yanke Liang, Nathanael S Gray, Madalena Tarsounas, and Stefan Knapp. A unique inhibitor binding site in erk1/2 is associated with slow binding kinetics. *Nature Chemical Biology*, 10(10) :853–860, 2014. DOI: 10.1038/nchembio.1629.
- [183] Tom D Heightman, Valerio Berdini, Hannah Braithwaite, Ildiko M Buck, Megan Cassidy, Juan Castro, Aurélie Courtin, James E H Day, Charlotte East, Lynsey Fazal, et al. Fragment-based discovery of a potent, orally bioavailable inhibitor that modulates the phosphorylation and catalytic activity of erk1/2. *Journal of Medicinal Chemistry*, 61(11) :4978–4992, 2018. DOI: 10.1021/acs.jmedchem.8b00421.
- [184] Stergios J Moschos, Ryan J Sullivan, Wen-Jen Hwu, Ramesh K Ramanathan, Alex A Adjei, Peter C Fong, Ronnie Shapira-Frommer, Hussein A Tawbi, Joseph Rubino, Thomas S Rush III, et al. Development of mk-8353, an orally administered erk1/2 inhibitor, in patients with advanced solid tumors. *JCI Insight*, 3(4) :e92352, 2018. DOI: 10.1172/jci.insight.92352.
- [185] Yu-Chen Lo, Stefano E Rensi, Wen Torng, and Russ B Altman. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8) :1538–1546, 2018. DOI: 10.1016/j.drudis.2018.05.010.

- [186] George Obaido, Ibomoiye Domor Mienye, Oluwaseun F Egbelowo, Ikiomoye Douglas Emmanuel, Adeola Ogunleye, Blessing Ogbuokiri, Pere Mienye, and Kehinde Aruleba. Supervised machine learning in drug discovery and development : Algorithms, applications, challenges, and prospects. *Machine Learning with Applications*, 17 :100576, 2024. DOI: 10.1016/j.mlwa.2024.100576.
- [187] Dezhong Ji, Lingzhi Zhang, Qihua Zhu, Ying Bai, Yaoyao Wu, and Yungen Xu. Discovery of potent, orally bioavailable erk1/2 inhibitors with isoindolin-1-one structure by structure-based drug design. *European Journal of Medicinal Chemistry*, 164 :334–341, 2019. DOI: 10.1016/j.ejmech.2018.12.040.
- [188] Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0 : New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8) :3891–3898, 2021. DOI: 10.26434/chemrxiv.14774223.v1.
- [189] Oleg Trott and Arthur J Olson. Autodock vina : improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2) :455–461, 2010. DOI: 10.1002/jcc.21334.
- [190] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13) :1605–1612, 2004. DOI: 10.1002/jcc.20084.
- [191] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1) :235–242, 2000. DOI: 10.1093/nar/28.1.235.
- [192] Bhusan K Kuntal, Polamarasetty Aparoy, and Pallu Reddanna. Easymodeller : A graphical interface to modeller. *BMC Research Notes*, 3 :1–5, 2010. DOI: 10.1186/1756-0500-3-226.
- [193] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2) :451–461, 2003. DOI: 10.1016/s0031-3203(02)00060-2.
- [194] Peter J Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 :53–65, 1987. DOI: 10.1016/0377-0427(87)90125-7.
- [195] Greg Landrum et al. Rdkit : A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8 :31, 2013. DOI: 10.1186/1758-2946-3-s1-o5.
- [196] Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. Circular fingerprints : flexible molecular descriptors with applications from physical chemistry to adme. *IDrugs*, 9(3) :199, 2006. DOI: 10.1007/978-3-030-63896-2.
- [197] Sereina Riniker and Gregory A Landrum. Similarity maps—a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics*, 5 :1–7, 2013. DOI: 10.1186/1758-2946-5-43.
- [198] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5) :742–754, 2010. DOI: 10.1021/ci100050t.
- [199] Schrödinger, LLC. Pymol molecular graphics system, 2020. Accessed via <http://www.pymol.org/pymol>.

- [200] David A Case, R Betz, D S Cerutti, Thomas E Cheatham III, T A Darden, R E Duke, T J Giese, H Gohlke, A W Goetz, N Homeyer, S Izadi, P Janowski, J Kaus, A Kovalenko, T S Lee, S LeGrand, P Li, C Lin, T Luchko, R Luo, B Madej, Kenneth M Merz, G Monard, P Needham, H Nguyen, H T Nguyen, I Omelyan, A Onufriev, D R Roe, A Roitberg, R Salomon-Ferrer, C L Simmerling, W Smith, J Swails, R C Walker, J Wang, R M Wolf, X Wu, L Xiao, D M York, and P A Kollman. AMBER 2016, 2016. AMBER Reference Manual.
- [201] Ramu Anandakrishnan, Boris Aguilar, and Alexey V. Onufriev. H++ 3.0 : automating pk prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research*, 40(W1) :W537–W541, 07 2012.
- [202] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb : improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8) :3696–3713, 2015. DOI: 10.1021/acs.jctc.5b00255.
- [203] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Antechamber : An accessory software package for molecular mechanical calculations, 2001. Presented at the American Chemical Society National Meeting, Chicago, IL.
- [204] KG Sprenger, Vance W Jaeger, and Jim Pfaendtner. The general amber force field (gaff) can accurately predict thermodynamic and transport properties of many ionic liquids. *The Journal of Physical Chemistry B*, 119(18) :5882–5895, 2015. DOI: 10.1021/acs.jpcb.5b00689.
- [205] Parviz Seifpanahi Shabane, Saeed Izadi, and Alexey V Onufriev. General purpose water model can improve atomistic simulations of intrinsically disordered proteins. *Journal of Chemical Theory and Computation*, 15(4) :2620–2634, 2019. DOI: 10.1021/acs.jctc.8b01123.
- [206] Zhen Li, Lin Frank Song, Pengfei Li, and Kenneth M Merz Jr. Systematic parametrization of divalent metal ions for the opc3, opc, tip3p-fb, and tip4p-fb water models. *Journal of Chemical Theory and Computation*, 16(7) :4429–4442, 2020. DOI: 10.1021/acs.jctc.0c00194.
- [207] Nadine Homeyer, Anselm HC Horn, Harald Lanig, and Heinrich Sticht. Amber force-field parameters for phosphorylated amino acids in different protonation states : phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *Journal of Molecular Modeling*, 12 :281–289, 2006. DOI: 10.1007/s00894-005-0028-4.
- [208] Ruslan L Davidchack, Richard Handel, and MV Tretyakov. Langevin thermostat for rigid body dynamics. *The Journal of Chemical Physics*, 130(23), 2009. DOI: 10.1063/1.3149788.
- [209] Johan Åqvist, Petra Wennerström, Martin Nervall, Sinisa Bjelic, and Bjørn O Brand-sdal. Molecular dynamics simulations of water and biomolecules with a monte carlo constant pressure algorithm. *Chemical Physics Letters*, 384(4-6) :288–294, 2004. DOI: 10.1016/j.cplett.2003.12.039.
- [210] Vincent Kräutler, Wilfred F Van Gunsteren, and Philippe H Hünenberger. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*, 22(5) :501–508, 2001. DOI: 10.1002/1096-987x(20010415)22:5<501::aid-jcc1021>3.0.co;2-v.

- [211] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103(19) :8577–8593, 1995. DOI: 10.1063/1.470117.
- [212] Bill R Miller III, T Dwight McGee Jr, Jason M Swails, Nadine Homeyer, Holger Gohlke, and Adrian E Roitberg. Mmpbsa. py : an efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation*, 8(9) :3314–3321, 2012. DOI: 10.1021/ct300418h.
- [213] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the mm/pbsa and mm/gbsa methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 51(1) :69–82, 2011. DOI: 10.1039/c4cp03179b.
- [214] Christopher J Cramer and Donald G Truhlar. General parameterized scf model for free energies of solvation in aqueous solution. *Journal of the American Chemical Society*, 113(22) :8305–8311, 1991. DOI: 10.1021/ja00022a017.
- [215] Samuel Genheden and Ulf Ryde. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5) :449–461, 2015. DOI: 10.1517/17460441.2015.1032936.
- [216] Antoine Daina, Olivier Michielin, and Vincent Zoete. Swissadme : a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, 7(1) :42717, 2017. DOI: 10.1038/srep42717.
- [217] Matthias Kohl. Performance measures in binary classification. *International Journal of Statistics in Medical Research*, 1(1) :79, 2012. DOI: 10.1016/b978-0-323-95502-7.00033-6.
- [218] Michael R Grever, Saul A Schepartz, and Bruce A Chabner. The national cancer institute : cancer drug discovery and development program. In *Seminars in oncology*, volume 19, pages 622–638, 1992. DOI: 10.1021/bk-1993-0534.ch007.
- [219] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1) :177–182, 2005. DOI: 10.1021/ci049714+.
- [220] Maria Sorokina, Peter Merseburger, Kohulan Rajan, Mehmet Aziz Yirik, and Christoph Steinbeck. Coconut online : collection of open natural products database. *Journal of Cheminformatics*, 13(1) :1–13, 2021. DOI: 10.21203/rs.3.rs-75600/v1.
- [221] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 64 :4–17, 2012. DOI: 10.1016/j.addr.2012.09.019.
- [222] Megan Egbert, Adrian Whitty, Gyorgy M Keseru, and Sandor Vajda. Why some targets benefit from beyond rule of five drugs. *Journal of Medicinal Chemistry*, 62(22) :10005–10025, 2019. DOI: 10.1021/acs.jmedchem.8b01732.s001.
- [223] Bradley Croy Doak, Bjorn Over, Fabrizio Giordanetto, and Jan Kihlberg. Oral druggable space beyond the rule of 5 : insights from drugs and clinical candidates. *Chemistry & biology*, 21(9) :1115–1142, 2014. DOI: 10.1016/j.chembiol.2014.08.013.
- [224] Eric W Bell and Yang Zhang. Dockrmsd : an open-source tool for atom mapping and rmsd calculation of symmetric molecules through graph isomorphism. *Journal of Cheminformatics*, 11(1) :1–9, 2019. DOI: 10.1186/s13321-019-0362-7.

- [225] Alejandro Castro-Alvarez, Anna M Costa, and Jaume Vilarrrasa. The performance of several docking programs at reproducing protein–macrolide-like crystal structures. *Molecules*, 22(1) :136, 2017. DOI: 10.3390/molecules22010136.
- [226] Mengyao Cui et al. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, 1(1) :5–8, 2020. DOI: 10.1063/5.0128945.
- [227] Hongjie Jia, Shifei Ding, Xinzheng Xu, and Ru Nie. The latest research progress on spectral clustering. *Neural Computing and Applications*, 24 :1477–1486, 2014. DOI: 10.1007/s00521-013-1439-2.
- [228] Trupti M Kodinariya, Prashant R Makwana, et al. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6) :90–95, 2013. DOI: 10.1007/978-3-642-34522-7_59.
- [229] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020. DOI: 10.1109/dsaa49011.2020.00096.
- [230] Makoto Otori, Mikio Takeuchi, Riyo Maruki, Hidenori Nakajima, and Hiroshi Miyake. Fr180204, a novel and selective inhibitor of extracellular signal-regulated kinase, ameliorates collagen-induced arthritis in mice. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 374 :311–316, 2007. DOI: 10.1007/s00210-006-0117-7.
- [231] Rikhia Chakraborty, Oliver A Hampton, Harshal Abhyankar, Daniel J Zinn, Amanda Grimes, Brooks Skull, Olive Eckstein, Nadia Mahmood, David A Wheeler, Dolores Lopez-Terrada, et al. Activating mapk1 (erk2) mutation in an aggressive case of disseminated juvenile xanthogranuloma. *Oncotarget*, 8(28) :46065, 2017. DOI: 10.18632/oncotarget.17521.
- [232] Anthony Markham and Susan J Keam. Selumetinib : first approval. *Drugs*, 80 :931–937, 2020. DOI: 10.1007/s40265-020-01331-x.
- [233] Nathaniel Weygant, Dongfeng Qu, William L Berry, Randal May, Parthasarathy Chandrakesan, Daniel B Owen, Sripathi M Sureban, Naushad Ali, Ralf Janknecht, and Courtney W Houchen. Small molecule kinase inhibitor lrrk2-in-1 demonstrates potent activity against colorectal and pancreatic cancer through inhibition of doublecortin-like kinase 1. *Molecular cancer*, 13(1) :1–14, 2014. DOI: 10.1186/1476-4598-13-103.
- [234] Gemma Roest, Evelien Hesemans, Kirsten Welkenhuyzen, Tomas Luyten, Nikolai Engedal, Geert Bultynck, and Jan B Parys. The er stress inducer l-azetidine-2-carboxylic acid elevates the levels of phospho-eif2 α and of lc3-ii in a ca²⁺-dependent manner. *Cells*, 7(12) :239, 2018. DOI: 10.3390/cells7120239.
- [235] Pamela A Lochhead, Julie A Tucker, Natalie J Tatum, Jinhua Wang, David Oxley, Andrew M Kidger, Victoria P Johnson, Megan A Cassidy, Nathanael S Gray, Martin EM Noble, et al. Paradoxical activation of the protein kinase-transcription factor erk5 by erk5 kinase inhibitors. *Nature Communications*, 11(1) :1383, 2020. DOI: 10.1038/s41467-020-15031-3.
- [236] Xingye Chen, Wuchen Xie, Yan Yang, Yi Hua, GuoMeng Xing, Li Liang, Chenglong Deng, Yuchen Wang, Yuanrong Fan, Haichun Liu, et al. Discovery of dual fgfr4 and egfr inhibitors by machine learning and biological evaluation. *Journal of Chemical Information and Modeling*, 60(10) :4640–4652, 2020. DOI: 10.1021/acs.jcim.0c00652.s002.
- [237] Jason Lin and Hiroki Nagase. The road not taken with pyrrole-imidazole polyamides : Off-target effects and genomic binding. *Biomolecules*, 10(4) :544, 2020. DOI: 10.3390/biom10040544.

- [238] Aoife A Nolan, Nourhan K Aboud, Walter Kolch, and David Matallanas. Hidden targets in raf signalling pathways to block oncogenic ras signalling. *Genes*, 12(4) :553, 2021. DOI: 10.3390/genes12040553.
- [239] Sofiene Larif, Chaker Ben Salem, Housseem Hmouda, and Kamel Bouraoui. In silico screening and study of novel erk2 inhibitors using 3d qsar, docking and molecular dynamics. *Journal of Molecular Graphics and Modelling*, 53 :1–12, 2014. DOI: 10.1016/j.jmgm.2014.07.001.
- [240] Michael Hay, David W Thomas, John L Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1) :40–51, 2014. DOI: 10.1038/nbt.2786.
- [241] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8) :711–716, 2004. DOI: 10.1038/nrd1470.
- [242] Arup K Ghose, Vellarkad N Viswanadhan, and John J Wendoloski. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases. *Journal of Combinatorial Chemistry*, 1(1) :55–68, 1999. DOI: 10.1021/cc9800071.
- [243] Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12) :2615–2623, 2002. DOI: 10.1021/jm020017n.
- [244] William J Egan, Kenneth M Merz, and John J Baldwin. Prediction of drug absorption using multivariate statistics. *Journal of Medicinal Chemistry*, 43(21) :3867–3877, 2000. DOI: 10.1021/jm000292e.
- [245] Ingo Muegge, Sarah L Heald, and David Brittelli. Simple selection criteria for drug-like chemical matter. *Journal of Medicinal Chemistry*, 44(12) :1841–1846, 2001. DOI: 10.1021/jm015507e.
- [246] Shelly Pathania and Pankaj Kumar Singh. Analyzing fda-approved drugs for compliance of pharmacokinetic principles : should there be a critical screening parameter in drug designing protocols? *Expert Opinion on Drug Metabolism & Toxicology*, 17(4) :351–354, 2021. DOI: 10.1080/17425255.2021.1865309.
- [247] Harold E Selick, Alan P Beresford, and Michael H Tarbit. The emerging importance of predictive adme simulation in drug discovery. *Drug Discovery Today*, 7(2) :109–116, 2002. DOI: 10.1016/s1359-6446(01)02100-6.
- [248] Franz Thiebaut, Takashi Tsuruo, Hirofumi Hamada, Michael M Gottesman, IRA Pastan, and Mark C Willingham. Cellular localization of the multidrug-resistance gene product p-glycoprotein in normal human tissues. *Proceedings of the National Academy of Sciences of the United States of America*, 84(21) :7735–7738, 1987. DOI: 10.1073/pnas.84.21.7735.
- [249] Richard B Kim, Martin F Fromm, Christoph Wandel, Brenda Leake, AJ Wood, Dan M Roden, Grant R Wilkinson, et al. The drug transporter p-glycoprotein limits oral absorption and brain entry of hiv-1 protease inhibitors. *Journal of Clinical Investigation*, 101(2) :289–294, 1998. DOI: 10.1172/jci1269.
- [250] Carlos Cordon-Cardo, James P O’Brien, Dolores Casals, Lana Rittman-Grauer, June L Biedler, Myron R Melamed, and Joseph R Bertino. Multidrug-resistance gene (p-glycoprotein) is expressed by endothelial cells at blood-brain barrier sites. *Proceedings*

- of the National Academy of Sciences of the United States of America*, 86(2) :695–698, 1989. DOI: 10.1073/pnas.86.2.695.
- [251] Kirsi Vähäkangas and Päivi Myllynen. Drug transporters in the human blood-placental barrier. *British Journal of Pharmacology*, 158(3) :665–678, 2009. DOI: 10.1111/j.1476-5381.2009.00336.x.
- [252] Chun Chen, Ming-Han Lee, Ching-Feng Weng, and Max K Leong. Theoretical prediction of the complex p-glycoprotein substrate efflux based on the novel hierarchical support vector regression scheme. *Molecules*, 23(7) :1820, 2018. DOI: 10.3390/molecules23071820.
- [253] Joanna Bojarska, Milan Remko, Martin Breza, Izabela D Madura, Krzysztof Kaczmarek, Janusz Zabrocki, and Wojciech M Wolf. A supramolecular approach to structure-based design with a focus on synthons hierarchy in ornithine-derived ligands : Review, synthesis, experimental and in silico studies. *Molecules*, 25(5) :1135, 2020. DOI: 10.3390/molecules25051135.
- [254] Yuan H Zhao, Michael H Abraham, Joelle Le, Anne Hersey, Chris N Luscombe, Gordon Beck, Brad Sherborne, and Ian Cooper. Rate-limited steps of human oral absorption and qsar studies. *Pharmaceutical Research*, 19 :1446–1457, 2002. DOI: 10.1023/a:1020444330011.
- [255] Faizul Azam, Arwa M Madi, and Hamed I Ali. Molecular docking and prediction of pharmacokinetic properties of dual mechanism drugs that block mao-b and adenosine a2a receptors for the treatment of parkinson’s disease. *Journal of Young Pharmacists*, 4(3) :184–192, 2012. DOI: 10.4103/0975-1483.100027.