REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DU LA RECHERCHE SCIENTIFIQUE

UNIVERSITE YAHIA FARES DE MEDEA

Faculté de Technologie

Département de Génie des Procédés et Environnement

Mémoire de fin d'études pour l'obtention du Diplôme de Master

Filière : **Génie des Procédés**

Spécialité : **Génie des Procédés pharmaceutique**

Modélisation et prédiction de l'activité anticancéreuse des dérivés de benzofuranes à l'aide de l'intelligence artificielle

Présenté par :

- M^{lle}Mendes Ihcene
- M^{me}Yahiaoui Hanane

Encadré par :

Dr. KOUIDER AMAR Mohamed

Dr. ALOUAOUI Redha

Promotion : Juin 2025

Remerciements

Tout d'abord, mes plus sincères remerciements vont à Allah, le Tout-Puissant, qui m'a guidé et soutenu tout au long de mon parcours académique, me permettant d'atteindre cette étape précieuse de ma vie.

Un profond merci à nos parents, source de mon succès et de ma fierté. Leur amour, leur patience et leur soutien inconditionnel ont été la lumière qui m'a guidé à travers les défis et les réussites.

Je tiens à exprimer ma gratitude envers mon directeur de mémoire, Dr. KOUIDER AMAR Mohamed, ainsi que Dr. ALOUAOUI Redha, pour leur encadrement précieux, leur soutien constant et leurs conseils avisés. Leur engagement et leur bienveillance ont grandement contribué à la réussite de ce travail.

À tous ceux qui, d'une manière ou d'une autre, m'ont apporté leur aide et leur encouragement, je vous remercie du fond du cœur.

Enfin, que les membres du jury, qui ont accepté d'évaluer mon travail, trouvent ici l'expression de ma profonde reconnaissance et de ma gratitude anticipée.

Merci infiniment

Dédicace 1

Avant tous nous remercions DIEU le tout puissant car sans la foi qu'on a en lui en ne Serait pas arrivé là ou est aujourd'hui.

À mes très chers parents, dont le sacrifice, la tendresse, l'amour, la Patience, Le soutien, l'aide et les encouragements sont l'essence de ma réussite.

Sans Eux, je ne serais pas ce que je suis aujourd'hui

Mon père : MENDES GUELLIL

Ma mère : BOUTICHE FATIHA

DOCTORANT : ISMAIL RAHAL

Mes très chers frères : ABDELHAK, MOHAMED, ABDELNNOUR

Mes très chères sœurs : NESRINE, NASSIRA

et Ma Chère grand-mère et ma tante NAIMA

A mes amies : SAIDANI ROUMAISSA, ADJLI OUAHIBA, MOUSSAOUI ROUMAISSA. Je dédie cet humble travail.

Tous les enseignants et les profs qui m'ont appris durant toutes mes Années d'étude et Tous ceux qui m'ont aidé de près ou de loin.

IHCENE.M

Dédicace 2

Avant tous nous remercions DIEU le tout puissant qui m'a guidé et facilité mon chemin vers la connaissance et l'apprentissage, et m'a accordé la force et la patience nécessaires pour mener à bien cet humble travail.

Après ce long parcours, qui n'a pas été facile, je ne peux qu'exprimer ma profonde gratitude et ma reconnaissance à tous ceux qui m'ont soutenu et encouragé, que ce soit par leurs paroles, leurs conseils ou même leurs prières sincères.

Je ne peux m'empêcher d'exprimer ma profonde gratitude à ma chère famille, à mes généreux parents, ma principale source d'amour, de soutien et de force dans la vie. Ils n'ont jamais manqué de me prodiguer tout l'amour, les prières, la patience et les encouragements dont j'avais besoin.

À mon père bien-aimé Mohammed

qui a été mon principal soutien et mon maître. Tu as insufflé en moi l'ambition et le goût d'apprendre. Tes paroles ont toujours été une source de motivation, et tes prières pures une forteresse qui m'a protégé pendant mon absence. Merci pour ta compassion, ta sagesse, ta patience et tous tes innombrables sacrifices.

Et à ma chère mère, Hafida,

qui était une étreinte chaleureuse et une véritable source de réconfort. Tu as veillé toute la nuit pour moi, prié en secret et enduré chaque épreuve avec patience et contentement. Je te dédie cet accomplissement, car chaque réussite est le reflet de ton don inconditionnel.



À mes chers frères : Anouar, Sid Ali et Bouchra,

Vous êtes le pouls paisible de mon monde, votre sécurité et votre joie.

Je vous dédie ce message en guise de remerciement pour votre amour inconditionnel et sincère.

Merci pour vos prières, votre soutien et vos encouragements. Votre présence dans ma vie est une bénédiction inestimable.

Et à mon cher époux Moussa,

mon compagnon de vie et ma moitié, dont la présence a été un soutien indéfectible, dont les encouragements m'ont motivée à persévérer, et dont la patience et la compréhension ont illuminé mes journées dans les moments les plus difficiles. Merci pour chaque instant passé à mes côtés, pour votre patience pendant mon absence et mes occupations, et pour vos paroles qui m'ont donné de la force lorsque j'étais faible. Cette réussite est la vôtre autant que la mienne.

À ma petite fleur,

À ma fille qui grandit encore en silence au creux de moi,

Je te dédie le fruit de longues années de lutte, de patience et de nuits blanches.

Je te dédie cette réussite comme le tout premier cadeau de ta maman,

En espérant qu'un jour, tu seras fière d'une mère qui n'a jamais cessé de rêver pour toi.

Avec toi, mon rêve s'est accompli...

Et avec toi, d'autres rêves commencent à naître.





Je remercie également tout particulièrement mes amis et collègues qui ont partagé mon parcours d'études et ont été une incarnation de l'amour, de la fraternité et de la coopération. Ensemble, nous avons vécu des moments inoubliables, remplis d'efforts et de joie, d'anxiété et d'espoir.

Enfin, à tous ceux qui m'ont soutenu par un mot, une prière ou un moment sincère, à tous ceux qui ont laissé une impression positive dans mon cœur, je dis : Merci du fond du cœur, et que la loyauté entre nous dure aussi longtemps que nous vivrons.

HANANE.Y



الملخص

في ظل الحاجة الماسة لاكتشاف جزيئات جديدة فعّالة مضادة للسرطان، يلقى اهتمام متزايد مشتقات البنز وفوران لما تتمتع به من خصائص دوائية واعدة. ومع ذلك، لا تزال أنشطتها المضادة للسرطان غير مدروسة بشكل منهجي كافٍ. تكمن الإشكالية الأساسية لهذا العمل في ضرورة التنبؤ، انطلاقًا من البنية الكيميائية، بالنشاط البيولوجي لتلك المشتقات ضد خلايا سرطان الثدي من نوع MCF-7. الهدف الرئيسي من هذه الدراسة هو تطوير نموذج تنبؤي فعال يمكنه تقدير النشاط المضاد للسرطان ($\log IC_{50}$) لمشتقات البنز وفوران باستخدام تقنيات الذكاء الاصطناعي. ولتحقيق ذلك، تم بناء قاعدة بيانات تحتوي على 198 مركبًا مأخوذًا من 13 مصدرًا علميًا. تم حساب الواصفات الجزيئية وبصمات الجزيئات باستخدام برنامج PaDEL-Descriptor. بعد المعالجة المسبقة للبيانات، تم اختيار المتغيرات باستخدام خوارزمية Boruta، بينما تم بناء النموذج باستخدام طريقة الغابة العشوائية (Random Forest) مع تحسين معلماتها باستخدام خوارزمية أسد النمل (ALO). أظهرت نتائج النموذج النهائي، التي تم تقييمها باستخدام التحقق المتقاطع، دقة تنبؤيه عالية، حيث بلغت قيمة R^2 = 0.92، و$RMSE = 0.64$، و$MAE = 0.36$، و$MAPE = 31.26\%$ على مجموعة الاختبار، مما يدل على قوة النموذج في التنبؤ.

كلمات مفتاحية: سرطان الثدي، مشتقات البنز وفوران، النمذجة الكمية للعلاقة بين البنية والنشاط، والذكاء الاصطناعي.

Abstract

In the ongoing search for effective new anticancer agents, benzofuran derivatives have attracted increasing interest due to their promising pharmacological properties. However, their anticancer activity remains relatively underexplored in a systematic way. The main issue addressed in this work lies in the need to predict, based on their chemical structure, the biological activity of these derivatives against the MCF-7 breast cancer cell line. The primary objective of this study is to develop a robust predictive model capable of estimating the anticancer activity ($logIC_{50}$) of benzofuran derivatives using artificial intelligence techniques. To this end, a dataset of 198 compounds was compiled from 13 bibliographic sources. Molecular descriptors and fingerprints were computed using PaDEL-Descriptor. After data preprocessing, feature selection was performed using the Boruta algorithm, and the modeling was carried out using a Random Forest model whose hyperparameters were optimized via the Ant Lion Optimizer (ALO) algorithm. The performance of the final model, evaluated through cross-validation, demonstrated excellent predictive accuracy with an R^2 of 0.92 on the test set, RMSE of 0.64, MAE of 0.36, and MAPE of 31.26%, indicating the strong predictive capacity of the model.

Keywords: Breast cancer, Benzofuran derivatives, QSAR and Artificial Intelligence.

Résumé

Face à la nécessité de découvrir de nouvelles molécules anticancéreuses efficaces, les dérivés de benzofurane suscitent un intérêt croissant en raison de leurs propriétés pharmacologiques prometteuses. Toutefois, leur activité anticancéreuse reste encore peu étudiée de manière systématique. La problématique principale de ce travail réside dans le besoin de prédire, à partir de leur structure chimique, l'activité biologique de ces dérivés contre la lignée cellulaire MCF-7 du cancer du sein. L'objectif principal de cette étude est de développer un modèle prédictif performant capable d'estimer l'activité anticancéreuse ($logIC_{50}$) des dérivés de benzofurane à l'aide de techniques d'intelligence artificielle. Pour ce faire, une base de données de 198 composés a été constituée à partir de 13 sources bibliographiques. Les descripteurs et empreintes moléculaires ont été calculés à l'aide de PaDEL-Descriptor. Après prétraitement des données, la sélection des variables a été effectuée par l'algorithme Boruta, et la modélisation a été réalisée à l'aide du modèle Random Forest, dont les hyperparamètres ont été optimisés par l'algorithme Ant Lion Optimizer (ALO). Les performances du modèle final, évaluées par validation croisée, ont révélé une excellente précision avec un $R^2 = 0.92$ sur l'ensemble de test, un RMSE = 0.64, une MAE = 0.36 et une MAPE = 31.26%, indiquant une très bonne capacité prédictive du modèle.

Mots clés : Cancer du sein, Dérivés de benzofurane, QSAR et Intelligence artificielle.

Liste de figures

Figure I.1. Resume des principaux signes distinctifs d'un cancer 4

figure I.2. Soustraction sagittale t1 post-contraste. 5

figure I.3. Structure typique associee au carcinome canalaire 5

figure I.4. Schema des etapes de detection du cancer du sein a travers les motifs de methylation de l'adn libre circulant dans le sang 6

figure I.5. Des gouttes de liquide aspire des mamelons provenant de quatre ouvertures canalaire affichent des colorations variees 7

figure I.6. Radiotherapie adjuvante des cancers du sein 8

figure I. 7. Strategies du traitement hormonal du cancer du sein 10

figure I.8. Schema montrant la cascade de metastases du cancer du sein 11

figure I.9. Analyse de la relation structure-activite dans les resultats anticancereux 12

figure I.10. Structure chimique du compose etudie 13

figure I.11. Apoptose par les derives de benzofuranes 14

figure I.12. Activite variable du benzofuran-3-carboxaldehyde contre le cancer du sein 15

figure I.13. Representation schematique du mecanisme d'action des inhibiteurs de la tyrosine kinase 16

figure I.14. Mecanisme d'inhibition des topoisomerases par des molecules therapeutiques ...18

figure I.15. Structure, polymerisation et depolymerisation des microtubules 19

figure II.1. L'ia et l'homme : main dans la main pour la relation client 20

figure II. 2. Schema de decomposition du domaine de l'intelligence artificielle et de ces sous-domaines 21

figure II. 3. Definition de l'apprentissage automatique par arthur samuel 22

figure II. 4. Les quatre types de systeme d'apprentissage automatique 23

figure II.5. Apprentissage supervise : schema d'une unite logistique 24

figure II. 6. L'apprentissage supervise. 24

figure II. 7. La difference entre la classification et la regression 25

Liste de figures

figure II. 8. L'apprentissage non supervise 26

figure II. 9. Avant et apres un apprentissage non supervise (clustering) 26

figure II.10. Arbre de decision. (a) structure generale d'un arbre de decision. (b) exemple en montrant un arbre de decision illustratif utilise pour decider si une photo represente une scene d'interieur ou exterieure 28

figure II. 11. Arbre de decision 29

figure II. 12. Illustration du random forest 32

figure II. 13. Une comparaison entre l'arbre de decision (dt) et la foret aleatoire (rf) [49]. 32

figure III.1. Gui for padel-descriptor 37

figure III.2. Organigramme de selection des fonctionnalites de boruta 40

figure III. 3. Representation de l'algorithme d'apprentissage forets aleatoires. 42

figure III. 4. Diagramme de flux de l'optimisation alo-fa. 44

figure IV.1. Matrice de correlation des descripteurs moleculaires selectionnes par l'algorithme boruta. 48

figure IV.3. Convergence de la fonction fitness et du coefficient de determination (r^2) au cours des iterations de l'algorithme alo. 50

figure IV.4. Évolution de l'exploration et de l'exploitation au cours des iterations de l'algorithme alo. 51

figure IV.5. Graphe de regression montrant les activites predites par rapport aux activites reelles pour les ensembles d'entrainement et de test. 54

figure IV.6. Le graphe de williams visualisant le domaine d'applicabilite du modele random forest. 56

figure IV.7. Interface graphique pour la prediction de l'activite anticancereuse des derives de benzofurane. 57

Liste des tableaux

Tableau II.1: Modèles QSAR pour l'activité inhibitrice de différentes classes d'échafaudages……...34

Tableau III.1: Caractéristiques de la base de données de dérivés de benzofurane utilisés pour la modélisation de l'activité anticancéreuse contre les cellules MCF-7.....38

Tableau IV.1: Métriques de performance du modèle QSAR sur les ensembles d'entraînement et de test..... 54

Sommaire

remerciements i
dedicace 1 II
dedicace 2 III
الملخص VII
abstract VI
resume VIII
liste de figures X
liste des tableaux XII
sommaire XIII
introduction generale 16
introduction generale 1
generalites sur le cancer du sein et les derives benzofuranes 3
i.1. le cancer de sein 3
I.1.1 HISTORIQUE 3
I.1.2 DÉFINITION 3
I.1.3 LES TYPES DE CANCER DE SEIN 4
i.1. 3.1 le cancer du sein invasif *4*
i.1.3.2 le cancer du sein non invasif *5*
I.1.4 LES TRAITEMENTS DE CANCER 6
i.1.4.1 la chirurgie *6*
i.1.4.2 la radiothérapie *7*
i.1.4.3 la chimiothérapie *8*
i.1.4.4 l'hormonothérapie *9*
I.1.5 LE CANCER DU SEIN MÉTASTATIQUE 10
i.ii les derives de benzofuranes 12
I.II.1 HISTORIQUE 12
I.II.2 L'ACTIVITÉ BIOLOGIQUE 12
I.II.3 LES MÉCANISMES D'ACTIONS 13
i.ii.3.1 induction de l'apoptose *14*
i.ii.3.2 inhibition de la prolifération cellulaire *14*
i.ii.3.3 inhibition des tyrosine kinases *16*
i.ii.3.4 inhibition des topoisomérases *17*
i.ii.3.5 inhibition de la polymérisation de la tubuline *18*
I.II.4 REVUE DE LA LITTÉRATURE 19
la modelisation par l'intelligence artificielle 21

Sommaire

ii.1. généralite 20

ii.2. l'apprentissage automatique 21

II.2.1. DÉMARCHE DES ALGORITHMES D'APPRENTISSAGE AUTOMATIQUE 22
II.2.2. OBJECTIFS DE L'APPRENTISSAGE AUTOMATIQUE 22
II.2.3. TYPE DE SYSTÈME D'APPRENTISSAGE AUTOMATIQUE 23
ii.2.3.1. apprentissage supervisé 23
ii.2.3.2. apprentissage non supervisé 25
ii.2.3.3. autres types d'apprentissage automatique 27

ii.3. algorithmes d'apprentissage automatique 27

II.3.2. ARBRE DE DÉCISION 28
II.3.3. L'ALGORITHME DE RANDOM FOREST 29
ii.3.3.1. avantages de random forest 32
ii.3.3.2. les inconvénients de la méthode random forest 33

ii.4. utilisation de l'intelligence artificielle pour les modeles qsar 33

methodologie 38

iii. 1. constitution de la base de donnees de derives de benzofurane 36

iii. 2. calcul des descripteurs et empreintes moleculaires 36

iii. 3. nettoyage et preparation des donnees 39

iii. 4. selection des caracteristiques avec l'algorithme boruta pour la regression 39

iii. 5. augmentation des donnees avec l'algorithme smogn 40

iii. 6. modelisation de l'activite anticancereuse des benzofuranes par forets aleatoires 41

iii. 7. optimisation des hyperparametres du modele forets aleatoires avec l'algorithme alo (ant lion optimizer) 42

iii. 8. criteres d'evaluation 44

resultats et discussion 47

iv.1. nettoyage et preparation des donnees 47

iv.2. selection des caracteristiques par l'algorithme boruta 47

iv.3. augmentation des donnees et impact sur la distribution des activites 48

iv. 4. optimisation des hyperparametres par l'algorithme alo 50

iv. 5. évaluation de la performance du modele qsar 52

iv.6. Évaluation du Domaine d'Applicabilité (DA) 54

iv.7. interface graphique pour la prediction de l'activite anticancereuse 56

conclusion generale 58

conclusion generale 59

Sommaire

references bibliographiques 61
annexe 73

Introduction Générale

Introduction générale

Selon l'Organisation mondiale de la santé (OMS), le cancer du sein a causé 670 000 décès dans le monde en 2022. Il est le cancer le plus fréquent chez la femme dans 157 pays sur 185, et survient dans tous les pays du monde, sans distinction géographique. Environ la moitié des cas de cancer du sein touchent des femmes sans facteurs de risque particuliers, si ce n'est le sexe et l'âge, et entre 0,5 et 1 % des cas concernent les hommes. Ces chiffres témoignent de l'ampleur et de la gravité de ce cancer, qui reste aujourd'hui une cause majeure de mortalité, malgré les avancées thérapeutiques et diagnostiques [1].

Dans ce contexte, la recherche scientifique s'oriente de plus en plus vers des approches innovantes pour découvrir de nouvelles molécules actives. Les modèles de relations quantitatives structure-activité (QSAR) combinés à l'intelligence artificielle (IA) se sont imposés comme des outils prometteurs, permettant de prédire efficacement l'activité biologique de composés chimiques avant même leur synthèse [2]. Ces approches in silico contribuent à réduire les coûts et le temps nécessaire au développement de nouveaux traitements, tout en ciblant de manière plus précise les mécanismes biologiques impliqués dans la maladie.

Parmi les structures chimiques d'intérêt, les dérivés de benzofurane ont attiré l'attention en raison de leurs multiples propriétés pharmacologiques, notamment leur capacité à inhiber la prolifération des cellules cancéreuses [3]. Ces composés peuvent agir à plusieurs niveaux : induction de l'apoptose, inhibition de la polymérisation de la tubuline, des tyrosine kinases ou encore des topo isomérases [4], [5], [6], [7]. Certains dérivés de benzofurane ont déjà démontré une activité significative contre la lignée cellulaire MCF-7.

Ce travail s'inscrit dans cette dynamique et vise à modéliser et prédire l'activité anticancéreuse de dérivés de benzofurane à l'aide de techniques d'intelligence artificielle. La problématique principale réside dans la capacité à construire un modèle robuste et généralisable permettant de prédire l'efficacité anticancéreuse de nouvelles structures. Pour cela, l'approche adoptée repose sur la constitution d'une base de données regroupant une large gamme de composés issus de la littérature, le calcul de leurs descripteurs et empreintes moléculaires, la sélection des variables pertinentes, et l'application d'un modèle d'apprentissage supervisé, notamment les forêts aléatoires optimisés par un algorithme bio-inspirés.

Ce mémoire est organisé en quatre chapitres principaux :

- *Chapitre I* : Généralités sur le cancer du sein et les dérivés de benzofurane. Ce chapitre aborde l'historique, les types, les mécanismes de développement et les traitements du cancer du sein, avant de détailler les propriétés pharmacologiques des dérivés de benzofurane et leurs mécanismes d'action anticancéreux.
- *Chapitre II* : Modélisation par l'intelligence artificielle. Il présente les concepts fondamentaux de l'intelligence artificielle et du machine learning, l'algorithme utilisé (forêts aléatoires), et son application dans la modélisation QSAR.
- *Chapitre III* : Méthodologie. Ce chapitre décrit les étapes suivies pour construire le modèle prédictif : la création de la base de données, le calcul des descripteurs via PaDEL-Descriptor, la normalisation des données, la sélection de variables, le choix des algorithmes, et l'optimisation des hyperparamètres.
- *Chapitre IV* : Résultats et discussions. Les performances des modèles sont évaluées à travers des métriques de validation croisée, des graphiques de prédiction et l'analyse du domaine d'applicabilité. L'interprétation des résultats est discutée en lien avec les mécanismes biologiques potentiels.

CHAPITRE I

Généralités sur le cancer du sein et les dérivés Benzofuranes

I.1. Le cancer de sein

I.1.1 Historique

Le cancer du sein est une maladie redoutée depuis longtemps, connue depuis 1600 la maladie a été découverte pour la première fois par les anciens Égyptiens il y a plus de 3 500 ans avant J.-C. Cette situation a été décrite avec une grande précision dans un papyrus écrit par Edwin Smith et George il y a quelques années. L'une des descriptions d'Ebers était liée aux tumeurs mammaires ronflantes, qui n'avaient pas remède. En 460 avant J.-C., Hippocrate, le père de la médecine occidentale, décrivait le cancer comme les maladies humorales, il a proposé que le cancer était causé par un excès de bile noire. Alors Galien a fourni une explication alternative au cancer du sein basé sur le modèle hippocratique croyait qu'un excès de bile noire était la cause principale, mais contrairement à Hippocrate, il y'a certaines tumeurs sont plus dangereuses que d'autres, et les médicaments Galien a suggéré que nous ayons trouvé de l'opium, de l'huile de ricin, de la réglisse, du soufre, ...etc. Pour le traitement médicamenteux du cancer du sein, Galien et les théories d'Hippocrate et d'autres n'ont pas été largement acceptées avant le XVIIe siècle. Mais en 1680, le médecin français François de la Boë Sylvius commença à remettre en question la théorie humorale du cancer. Il croyait que le cancer n'était pas causé par un excès de bile noire, mais plutôt par un processus chimique qui transformait le liquide lymphatique en acides irritants. En 1730, le médecin parisien Claude-Deshais Gendron rejeta également le système de Galien, affirmant que le cancer survenait lorsque le tissu nerveux et glandulaire se mélangeait aux vaisseaux lymphatiques. D'autres théories ont également proposé de bonnes explications à la maladie, mais ce n'est qu'au XIXe siècle qu'il a été prouvé que le cancer n'est pas seulement une maladie de l'organisme, une maladie du tissu, mais aussi une maladie de la cellule, une maladie du noyau cellulaire [8].

I.1.2 Définition

Le cancer du sein est le cancer le plus fréquent chez la femme et la deuxième cause de décès parmi tous les cancers. Il se développe à partir des cellules initialement normales qui constituent la glande mammaire. Ces cellules se transforment et se multiplient de façon anarchique et excessive par des mutations ou des instabilités génétique (anomalie cytogénétique) pour former une masse cellulaire appelée tumeur maligne [9].

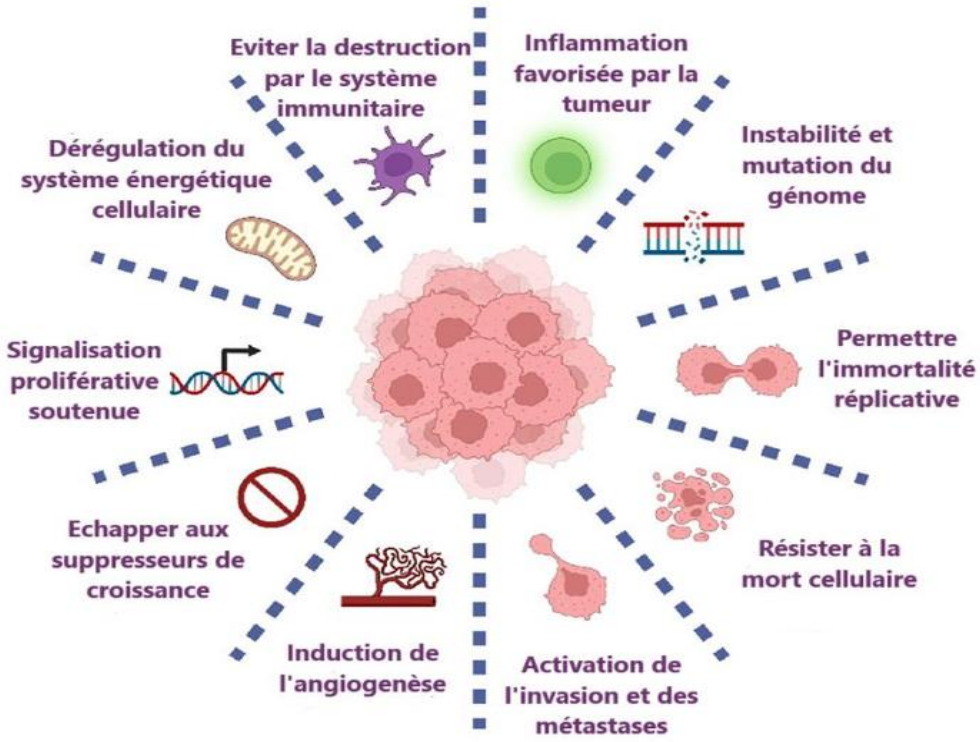


Figure I.1. Résumé des principaux signes distinctifs d'un cancer [10].

I.1.3 Les types de cancer de sein

Le cancer du sein est l'un des cancers les plus courants chez la femme dans le monde. Il existe plusieurs types de cancers mammaires, chacun ayant des caractéristiques et des pronostics différents. Le plus fréquent est le carcinome ductal invasif, qui prend naissance dans les canaux mammaires puis invade les tissus environnants. Un autre type courant est le carcinome lobulaire invasif, se développant dans les lobules du sein [11].

I.1. 3.1 Le cancer du sein invasif

Aussi appelé carcinome invasif, est un type de cancer qui a dépassé la couche interne du tissu mammaire et s'est étendu dans les tissus environnants du sein. Il peut se propager localement aux structures adjacentes ou, dans certains cas, à d'autres parties du corps [12].

La figure I. 2 représente un carcinome canalaire infiltrant dans le quadrant supéro-externe du sein droit (flèche verte) chez une femme de 80 ans, se présentant comme une lésion rehaussée avec des marges irrégulières spiculées [13].

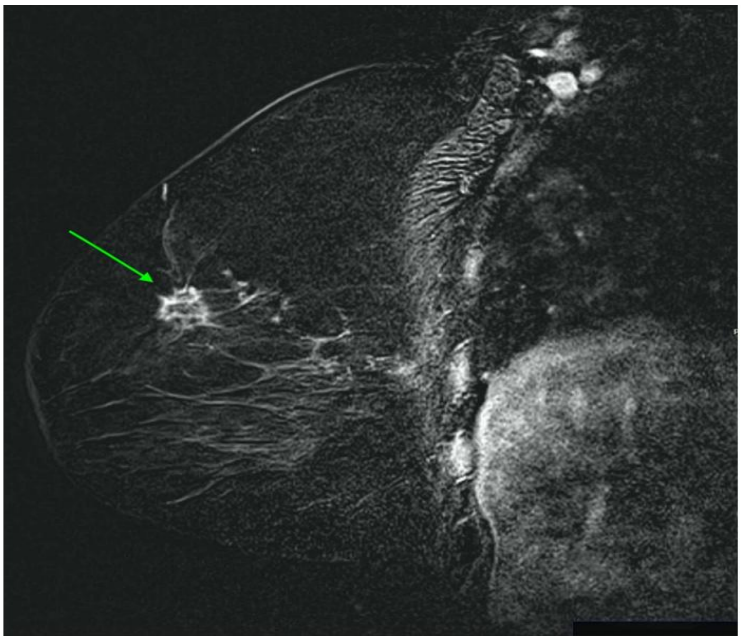


Figure I.2. Soustraction sagittale T1 post-contraste.

I.1.3.2 Le cancer du sein non invasif

Le cancer du sein non invasif, également connu sous le nom de carcinome canalaire in situ (CCIS), se développe dans les canaux lactifères du sein sans envahir les tissus environnants. Ce type de cancer est souvent détecté lors de mammographies de routine et est considéré comme un stade précoce de la maladie, ce qui permet un traitement efficace et augmente les chances de guérison.

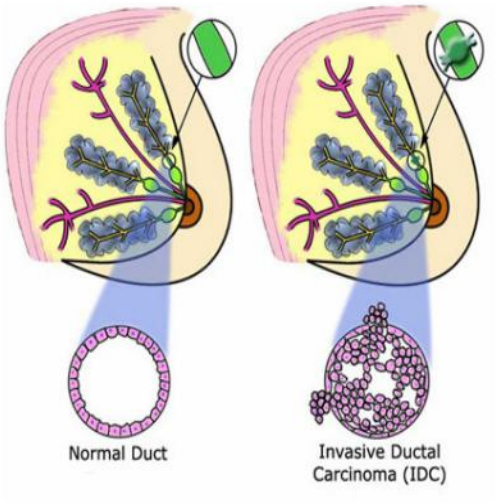
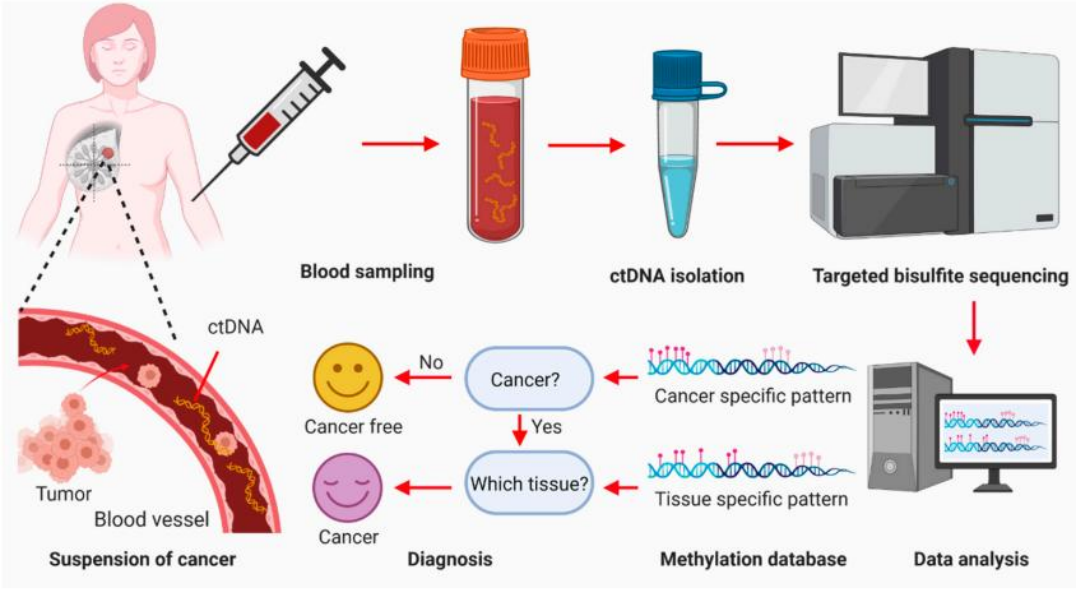


Figure I.3. Structure typique associée au carcinome canalaire [14].

Le CCIS peut être asymptomatique et ne provoque généralement pas de signes visibles, mais il peut parfois se manifester par des modifications dans l'apparence du sein ou des écoulements. Bien que le CCIS ne se propage pas au-delà des canaux, il peut augmenter le risque de développer un cancer du sein invasif à l'avenir, ce qui nécessite une surveillance et



un traitement appropriés[15].

Figure I.4. Schéma des étapes de détection du cancer du sein à travers les motifs de méthylation de l'ADN libre circulant dans le sang [15].

I.1.4 Les traitements de cancer

I.1.4.1 La chirurgie

Est une intervention clé dans le traitement de cette maladie. Elle peut inclure des procédures telles que la tumorectomie, qui consiste à retirer la tumeur et une marge de tissu sain, ou la mastectomie, qui implique l'ablation totale ou partielle du sein. Le choix de la technique chirurgicale dépend de divers facteurs, notamment la taille et le stade de la tumeur, ainsi que les préférences de la patiente.



Figure I.5. Des gouttes de liquide aspiré des mamelons provenant de quatre ouvertures canalaire affichent des colorations variées, suggérant que les systèmes canalaires du sein sont anatomiquement distincts et que certaines propriétés des liquides mammaires présentent des variations entre les canaux [9].

Après la chirurgie, les patientes peuvent nécessiter des traitements complémentaires, tels que la radiothérapie ou la chimiothérapie, pour réduire le risque de récidive. La chirurgie peut également entraîner des complications, comme des infections ou des problèmes de cicatrisation, et il est essentiel de discuter des options de reconstruction mammaire avec les patientes qui le souhaitent. Un suivi régulier est crucial pour surveiller la santé post-opératoire et détecter toute récidive éventuelle [16].

I.1.4.2 La radiothérapie

La radiothérapie du cancer du sein constitue une composante essentielle du traitement multidisciplinaire, visant à réduire les risques de récidive locale et à améliorer la survie globale. Après une chirurgie conservatrice ou une mastectomie, la radiothérapie ciblée est généralement appliquée pour éliminer les cellules cancéreuses restantes. Elle peut inclure une irradiation du sein entier, du lit tumoral, ainsi que des régions ganglionnaires régionales, comme les lymphes axillaires, sous-claviculaires et mammaires internes. L'utilisation de doses fractionnées permet de maximiser l'efficacité tout en minimisant les effets secondaires, et la techniques modernes, comme la radiothérapie conformationnelle ou l'IMRT, contribuent à une meilleure précision du traitement, réduisant ainsi les toxicités pour les tissus sains.

De plus, différentes stratégies de radiothérapie ont été développées en fonction du risque associé au cancer du sein. La radiothérapie partielle du sein, par exemple, est recommandée pour les patients présentant un faible risque de récidive, permettant d'administrer un traitement ciblé du lit tumoral tout en limitant l'exposition aux tissus sains, notamment la peau et les muscles environnants. La décision d'inclure un boost sur le lit tumoral ou non dépend aussi de facteurs tels que l'âge, la taille de la tumeur, et la présence de ganglions envahis. Les techniques modernes visant à réduire la morbidité cardiaque et pulmonaire, notamment pour les cancers du sein situés proches du cœur, ont aussi permis de rendre la radiothérapie plus sûre et mieux adaptée aux besoins de chaque patient [17].

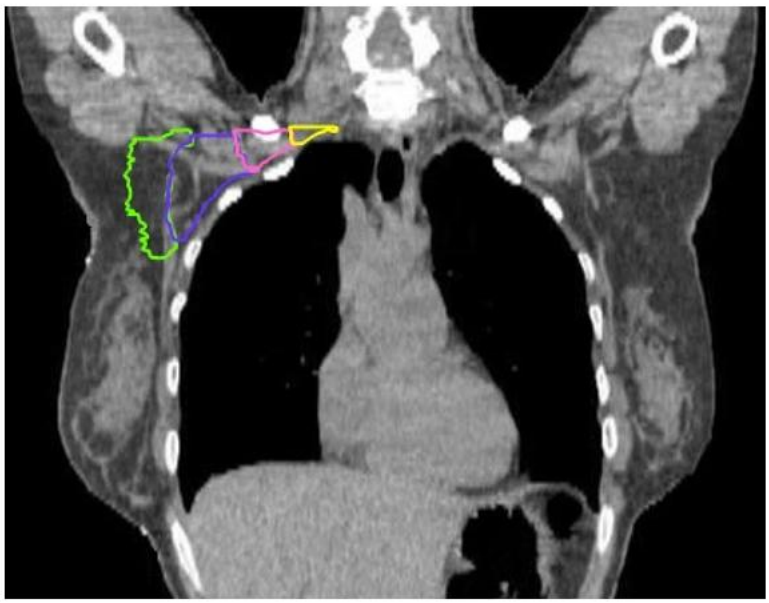


Figure I.6. Radiothérapie adjuvante des cancers du sein : volumes cibles anatomocliniques des niveaux axillaires I à IV sur une coupe coronale conformément aux recommandations de l'European Society for Radiotherapy and Oncology (ESTRO). Vert : niveau I ; bleu : niveau II ; rose : niveau III ; jaune : niveau IV[18].

I.1.4.3 La chimiothérapie

Le traitement par chimiothérapie du cancer du sein consiste à utiliser des médicaments cytotoxiques pour détruire les cellules cancéreuses, que ce soit dans le cas d'un cancer avancé ou en situation adjuvante (après chirurgie pour réduire le risque de récidive). La chimiothérapie est souvent recommandée dans l'objectif de traiter les micrométastases, c'est-à-dire des cellules cancéreuses microscopiques résiduelles pouvant provoquer une rechute à distance. L'approche thérapeutique est généralement adaptée en fonction de caractéristiques

moléculaires spécifiques de la tumeur, permettant un traitement personnalisé. Depuis plusieurs années, la progression dans la compréhension de la biologie du cancer du sein a permis le développement de nouveaux agents et de combinaisons de traitements plus efficaces, tout en évitant certains traitements inutiles chez des patientes peu susceptibles d'en bénéficier.

Les protocoles de chimiothérapie incluent généralement l'administration de plusieurs types de médicaments, parfois en succession ou en association, pour maximiser l'efficacité. Parmi les agents récents couramment utilisés, on trouve des médicaments ciblant des voies moléculaires spécifiques, comme le trastuzumab, qui cible la protéine HER-2/neu, ou des agents anti-angiogéniques comme le bevacizumab. Ces innovations, combinées à une meilleure sélection des patientes grâce à l'analyse de biomarqueurs, offrent aujourd'hui des perspectives améliorées pour le contrôle de la maladie et l'augmentation des taux de survie à long terme[19].

I.1.4.4 L'Hormonothérapie

L'hormonothérapie du cancer du sein repose principalement sur l'utilisation de médicaments qui ciblent les récepteurs aux hormones, notamment les récepteurs aux œstrogènes. Ces médicaments agissent en bloquant ou en modulant l'action des œstrogènes, qui jouent un rôle clé dans la croissance de nombreux cancers du sein. Parmi ces agents, le tamoxifène est le plus couramment utilisé ; il agit en tant que modulateur sélectif des récepteurs aux œstrogènes (SERM), en se liant à ces récepteurs et empêchant l'œstrogène d'activer la croissance tumorale. L'hormonothérapie est souvent administrée pendant plusieurs années, généralement jusqu'à cinq ans ou plus, en fonction du stade et du profil du cancer, pour réduire le risque de récidive.

Cependant, l'efficacité de cette approche peut être limitée par l'apparition de résistance hormonale. Certains cancers du sein deviennent insensibles ou développent des mécanismes permettant d'esquiver l'effet des agents hormonaux, ce qui complique leur traitement. La résistance peut résulter de mutations génétiques dans les récepteurs eux-mêmes ou de changements dans l'expression de ces récepteurs, notamment la perte d'expression des récepteurs œstrogéniques. Pour pallier ces difficultés, de nouvelles molécules telles que les analogues de rapamycine ou les inhibiteurs de l'aromatase sont en cours d'évaluation. Ces agents prometteurs pourraient, à l'avenir, remplacer ou compléter le tamoxifène, offrant de nouvelles options plus efficaces contre le cancer du sein résistants à l'hormonothérapie conventionnelle[20].

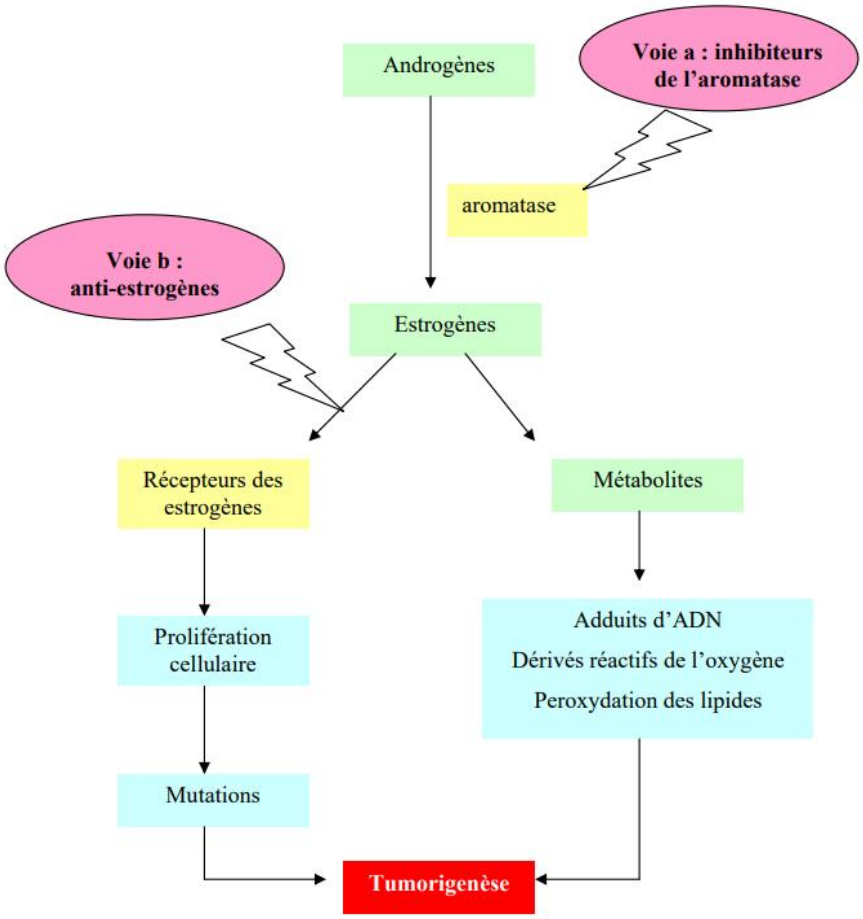
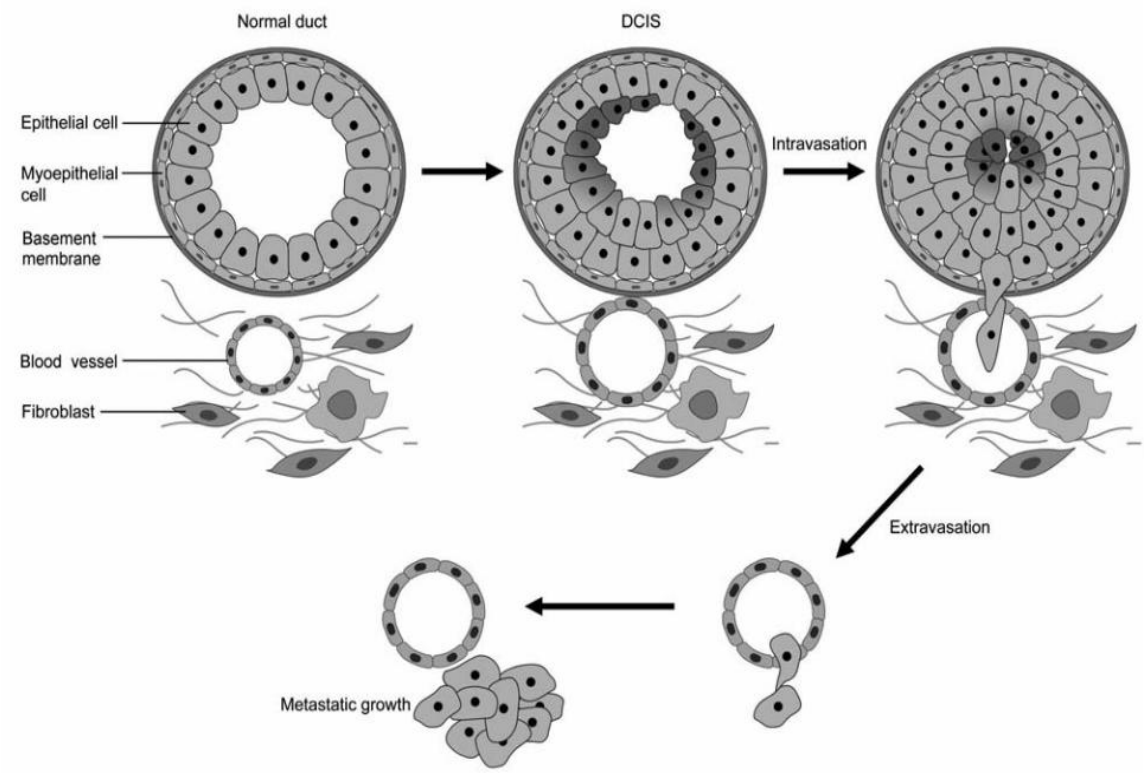


Figure I. 7. Stratégies du traitement hormonal du cancer du sein. Voie a : inhibition de la biosynthèse des estrogènes par les anti-aromatase. Voie b : blocage de l'action des estrogènes au niveau de leurs récepteurs [21].

I.1.5 Le cancer du sein métastatique

Le cancer du sein métastatique se produit lorsque des cellules cancéreuses se détachent de la tumeur primaire et se propagent à d'autres parties du corps via le sang ou les vaisseaux lymphatiques. Ce processus, connu sous le nom de métastase, implique plusieurs étapes, y compris l'invasion locale des tissus environnants, l'intravasation dans les vaisseaux sanguins ou lymphatiques, et l'adhésion aux organes cibles où les cellules cancéreuses peuvent proliférer et former de nouvelles tumeurs.

La détection des métastases du cancer du sein repose sur des manifestations cliniques, des biopsies, des évaluations radiologiques et des marqueurs tumoraux sériques. Bien que des méthodes comme la mammographie aient amélioré le diagnostic précoce, il existe encore des lacunes dans la détection des métastases à un stade précoce, ce qui souligne l'importance de nouvelles approches, telles que l'analyse des cellules tumorales circulantes (CTCs), pour



mieux prédire et gérer la maladie [22].

Figure I.8. Schéma montrant la cascade de métastases du cancer du sein [16].

Les dérivés de benzofuranes suscitent un intérêt croissant dans la recherche contre le cancer du sein en raison de leur capacité à cibler divers mécanismes clés de la maladie. Certains agissent comme des modulateurs des récepteurs d'œstrogènes, bloquant ainsi la croissance des tumeurs hormonodépendantes. D'autres peuvent directement inhiber la prolifération des cellules cancéreuses et induire leur mort, perturber leur cycle de division, ou encore freiner la formation de nouveaux vaisseaux sanguins nécessaires à la croissance tumorale et aux métastases. Bien que les études précliniques soient encourageantes, des recherches cliniques approfondies sont indispensables pour valider leur potentiel thérapeutique chez les patientes atteintes de cancer du sein.

I.II Les dérivés de benzofuranes

I.II.1 Historique

Les dérivés de benzofurane ont suscité un intérêt croissant dans la recherche sur le cancer du sein, notamment en raison de leurs propriétés anticancéreuses et de leur capacité à moduler certaines voies biologiques impliquées dans la prolifération tumorale.

Historiquement, ces composés ont été étudiés pour leur activité inhibitrice sur les kinases (des enzymes essentielles à la croissance des cellules cancéreuses). Des recherches ont montré que certains dérivés de benzofurane pouvaient induire l'apoptose (mort cellulaire programmée) et réduire la viabilité des cellules tumorales, ce qui en fait des candidats prometteurs pour le développement de nouveaux traitements [23].

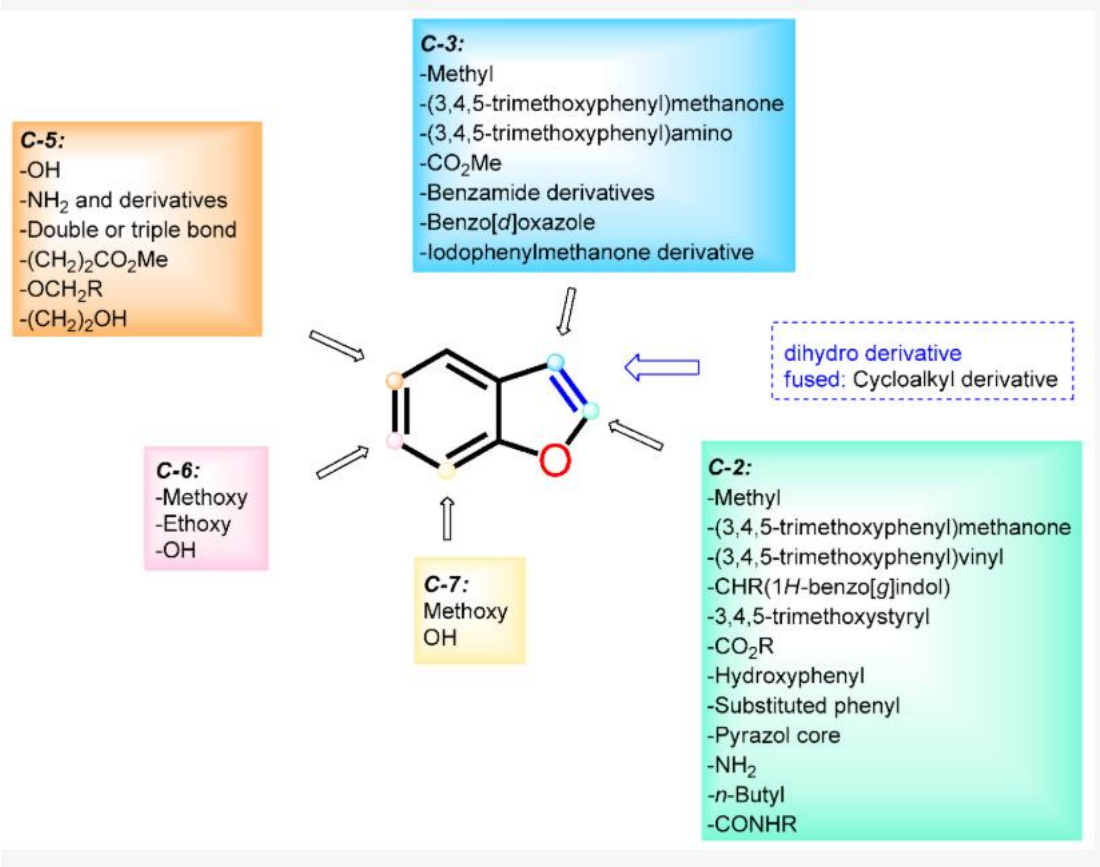


Figure I.9. Analyse de la relation structure-activité dans les résultats anticancéreux [7].

I.II.2 L'activité biologique

Il est suggéré que l'introduction de substituants à des positions spécifiées dans le noyau du

benzofurane aboutit à de nouveaux dérivés avec des caractéristiques structurelles uniques qui peuvent posséder une excellente valeur thérapeutique [24]. Par conséquent, ces dernières années, les dérivés des benzofuranes ont été fréquemment utilisés dans le développement de nouveaux médicaments [25].

Les applications pharmacologiques des dérivés de benzofurane comprenaient des propriétés antituberculeuses, anticonvulsives, antitumorales, antiplaquettaires, antioxydantes, antipaludiques, antidépressives, antimicrobiennes, anti-inflammatoires, immuno modulatrices et contre la maladie d‟Alzheimer et la maladie de Parkinson. Il existe plus de trente médicaments contenant un échafaudage de benzofurane sur le marché, qui ont été approuvés par la Food and Drug Administration (FDA) [26].

La recherche en oncologie investit des efforts considérables pour identifier de nouvelles thérapies sûres et efficaces afin de minimiser les effets secondaires critiques causés par les traitements conventionnels. Les systèmes de noyaux hétérocycliques fusionnés sont devenus des échafaudages importants avec de nombreuses propriétés biologiques [27].

Une série de nouveaux dérivés de Benzofurane substituées à la fois aux positions R1, NR2 et R3, X et Y ont été synthétisés dans le but d'améliorer l'activité spécifique (voir Figure I. 10).

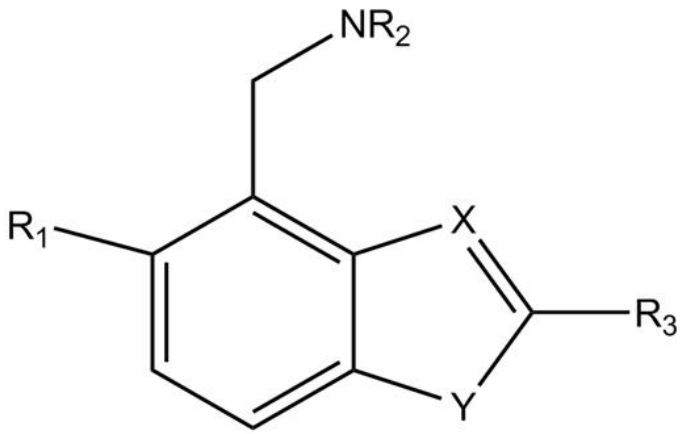


Figure I.10. Structure chimique du composé étudié [27].

I.II.3 Les mécanismes d’actions

Les dérivés de benzofurane sont des composés organiques qui ont montré un potentiel intéressant dans le traitement du cancer du sein. Leur effet repose principalement sur leurs

propriétés antioxydantes et cytotoxiques, qui peuvent inhiber la prolifération des cellules cancéreuses et induire leur apoptose (mort cellulaire programmée) [3].

I.II.3.1 Induction de l'apoptose

L'induction de l'apoptose par les dérivés de benzofuranes dans le cancer du sein repose sur leur capacité à activer des voies moléculaires spécifiques qui conduisent à la mort programmée des cellules tumorales. Ces composés interagissent avec des protéines clés comme les caspases, qui orchestrent la fragmentation de l'ADN et la dégradation des composants cellulaires. Ils peuvent également moduler l'équilibre entre les protéines pro-apoptotiques (comme Bax) et anti-apoptotiques (comme Bcl-2), favorisant ainsi l'activation de la voie mitochondriale de l'apoptose. En ciblant ces mécanismes, les dérivés de benzofuranes offrent une approche prometteuse pour limiter la prolifération des cellules cancéreuses et améliorer l'efficacité des traitements contre le cancer du sein [4].

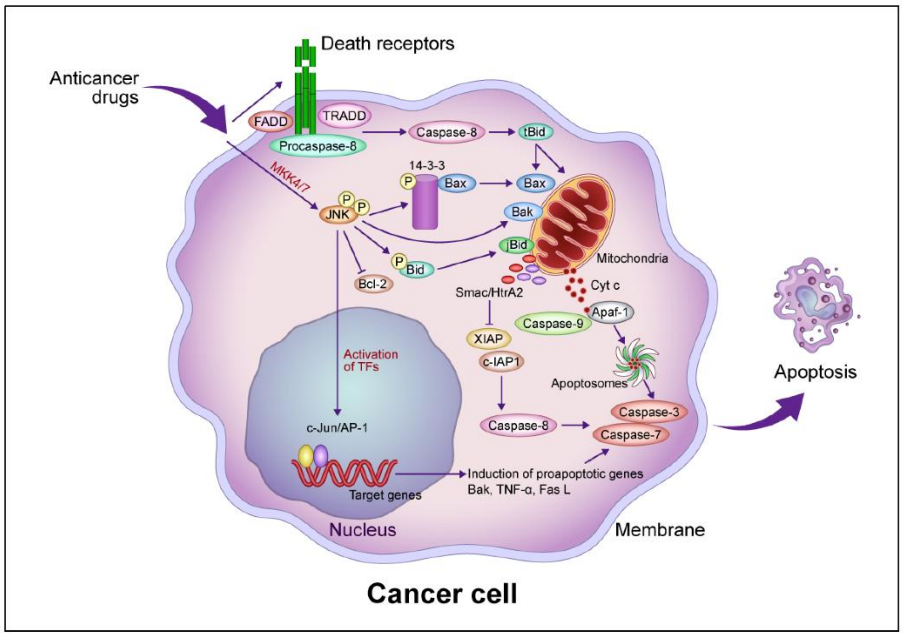


Figure I.11. Apoptose par les dérivés de benzofuranes [28].

I.II.3.2 Inhibition de la prolifération cellulaire

Le cycle cellulaire est un processus fondamental qui orchestre la vie et la reproduction des cellules. L'interphase est une période cruciale de préparation. Elle s'étend de la fin d'une division cellulaire jusqu'au début de la suivante, et sa durée varie grandement selon le type de

cellule et son environnement, car les cellules interagissent constamment avec leur milieu en libérant des substances et en répondant à des signaux chimiques. Cette interphase se décompose elle-même en trois phases clés : la phase **G1**, où la cellule grandit, synthétise les enzymes nécessaires et accumule des réserves pour la division ; la phase **S** (pour "synthèse"), moment essentiel où la cellule réplique son ADN et duplique son centrosome, garantissant que les futures cellules filles recevront un patrimoine génétique identique à celui de la cellule mère ; enfin, la phase G2 est une période de croissance finale et de préparation intensive avant que la cellule n'entre en mitose[5].

La mitose est un processus de division cellulaire qui permet de créer deux cellules filles identiques à partir d'une cellule mère. Elle se déroule en plusieurs étapes successives : la prophase, où les chromosomes se condensent et le noyau commence à disparaître ; la métaphase, où les chromosomes s'alignent au centre de la cellule ; l'anaphase, où les chromatides se séparent et migrent vers les pôles opposés ; et enfin la télophase, où les noyaux des nouvelles cellules se reforment. La cytocinèse achève la séparation en divisant le cytoplasme. Ce mécanisme est essentiel pour la croissance et la régénération des cellules [29].

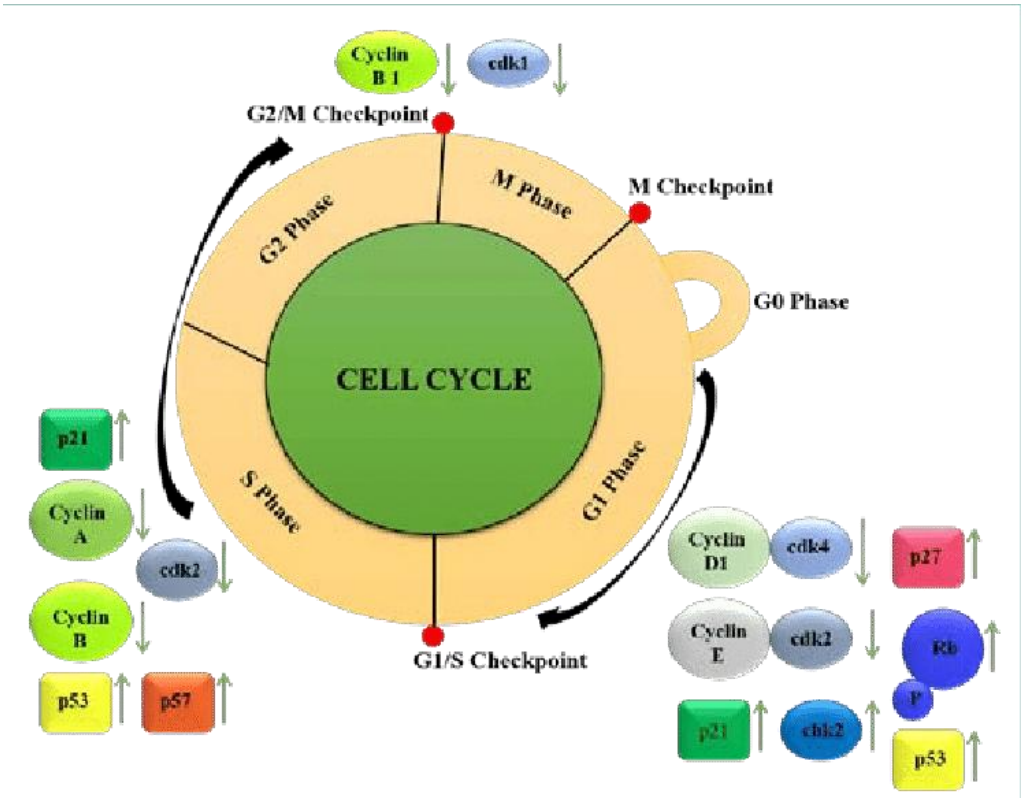


Figure I.12. Activité Variable du Benzofuran-3-carboxaldéhyde contre le Cancer du Sein.

I.II.3.3 Inhibition des tyrosine kinases

Les tyrosine kinases sont des enzymes qui catalysent la phosphorylation des résidus tyrosine des protéines, activant ainsi des voies de signalisation impliquées dans la croissance et la survie des cellules tumorales. L'inhibition de ces enzymes empêche la transmission du signal de croissance, ce qui peut induire l'apoptose des cellules cancéreuses et limiter leur prolifération.

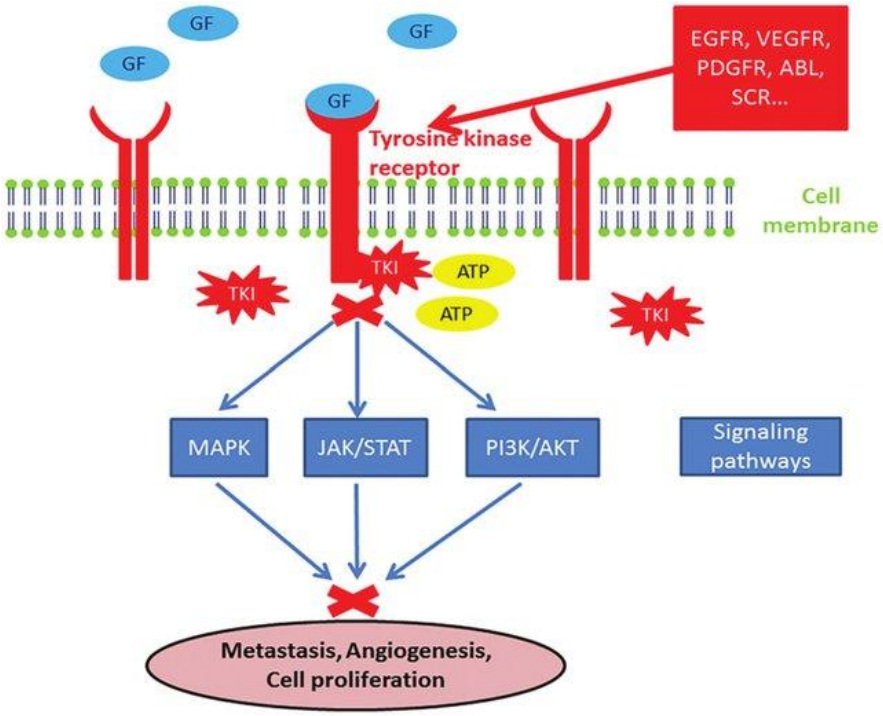


Figure I .13. Représentation schématique du mécanisme d'action des inhibiteurs de la tyrosine kinase (TKI), où GF : facteur de croissance ; EGDR : récepteur du facteur de croissance endothélial ; VEGFR : récepteur du facteur de croissance endothélial vasculaire ; PDGFR : récepteur du facteur de croissance dérivé des plaquettes ; ABL : famille des tyrosine kinases Abelson ; SCR : famille des tyrosine kinases SCR ; ATP : adénosine triphosphate ; MAPK : protéines kinases activées par les mitogènes ; JAK/STAT : Janus kinases/transducteur de signal et activation des protéines de transcription et PI3K/AKT : phosphoinositide-3-kinase/protéine kinase B [6].

Les cellules communiquent entre elles via des facteurs de croissance (GF), lesquels se lient à des récepteurs tyrosine kinase (RTK) présents sur la membrane cellulaire. Ces RTK incluent, entre autres, l'EGFR, le VEGFR et le PDGFR. Lorsque le GF se lie au RTK, celui-ci s'active, initiant une cascade de signalisation intracellulaire. Cette cascade implique des

protéines kinases comme les MAPK, les JAK/STAT, et les PI3K/AKT. L'activation de ces voies de signalisation mène, en fin de compte, à des processus cellulaires tels que la métastase (propagation du cancer), l'angiogenèse (formation de nouveaux vaisseaux sanguins), et la prolifération cellulaire. L'activité de ces kinases dépend de l'ATP (adénosine triphosphate) comme source d'énergie pour la phosphorylation (l'ajout d'un groupe phosphate qui active la protéine) [6].

Les TKIs agissent en bloquant l'activité de ces RTK, soit en empêchant la liaison du GF au RTK, soit en inhibant la phosphorylation de la protéine, empêchant ainsi l'utilisation de l'ATP. En bloquant ces RTK, les TKIs interrompent les cascades de signalisation intracellulaire qui conduisent à la métastase, à l'angiogenèse et à la prolifération cellulaire. Par conséquent, ils peuvent être utilisés comme thérapie ciblée pour inhiber la croissance tumorale et la propagation du cancer [6].

Les dérivés de benzofurane peuvent agir comme inhibiteurs des tyrosine kinases en se fixant sur le site actif de l'enzyme, bloquant ainsi son activité. Certains composés ont été étudiés pour leur capacité à inhiber la voie PI3K/AKT, qui est souvent dérégulée dans le cancer du sein. En bloquant cette voie, ces molécules peuvent réduire la prolifération des cellules tumorales et favoriser leur mort programmée [30].

I.II.3.4 Inhibition des topoisomérases

Les dérivés de benzofurane sont étudiés pour leur potentiel anticancéreux, notamment dans l'inhibition des topoisomérases, des enzymes essentielles à la régulation de la topologie de l'ADN. Les topoisomérases jouent un rôle clé dans la réplication et la transcription de l'ADN, et leur inhibition peut entraîner des dommages irréversibles à l'ADN des cellules cancéreuses, conduisant à leur apoptose.

Les inhibiteurs des topoisomérases agissent en stabilisant le complexe enzyme-ADN, empêchant ainsi la réparation des cassures temporaires de l'ADN. Cela entraîne une accumulation de lésions qui perturbent la division cellulaire et favorisent la mort des cellules tumorales. Certains dérivés de benzofurane ont montré une affinité particulière pour ces enzymes, ce qui en fait des candidats prometteurs pour le développement de nouveaux traitements contre le cancer du sein [7].

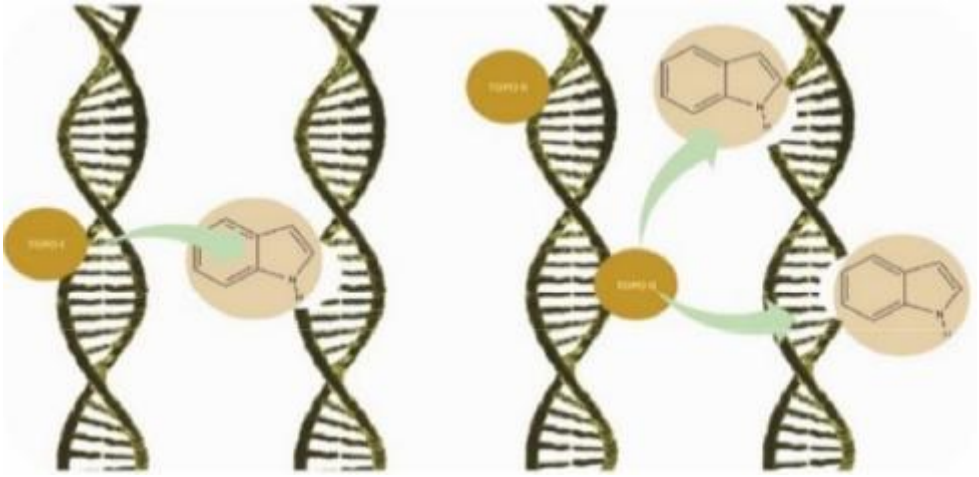


Figure I.14. Mécanisme d'inhibition des Topoisomérases par des molécules thérapeutiques [31].

I.II.3.5 Inhibition de la polymérisation de la tubuline

L'inhibition de la polymérisation de la tubuline par les dérivés de benzofuranes est un mécanisme clé dans le traitement du cancer du sein. La tubuline est une protéine essentielle à la formation des microtubules, qui jouent un rôle crucial dans la division cellulaire. En bloquant la polymérisation de la tubuline, ces composés empêchent la formation correcte du fuseau mitotique, ce qui entraîne l'arrêt du cycle cellulaire et la mort des cellules cancéreuses.

Les dérivés de benzofuranes agissent en se liant à la tubuline et en perturbant son assemblage, ce qui entraîne une instabilité des microtubules et empêche la prolifération des cellules tumorales. Ce mécanisme est similaire à celui des alcaloïdes de la pervenche et des taxanes, qui sont des agents chimiothérapeutiques bien établis [32].

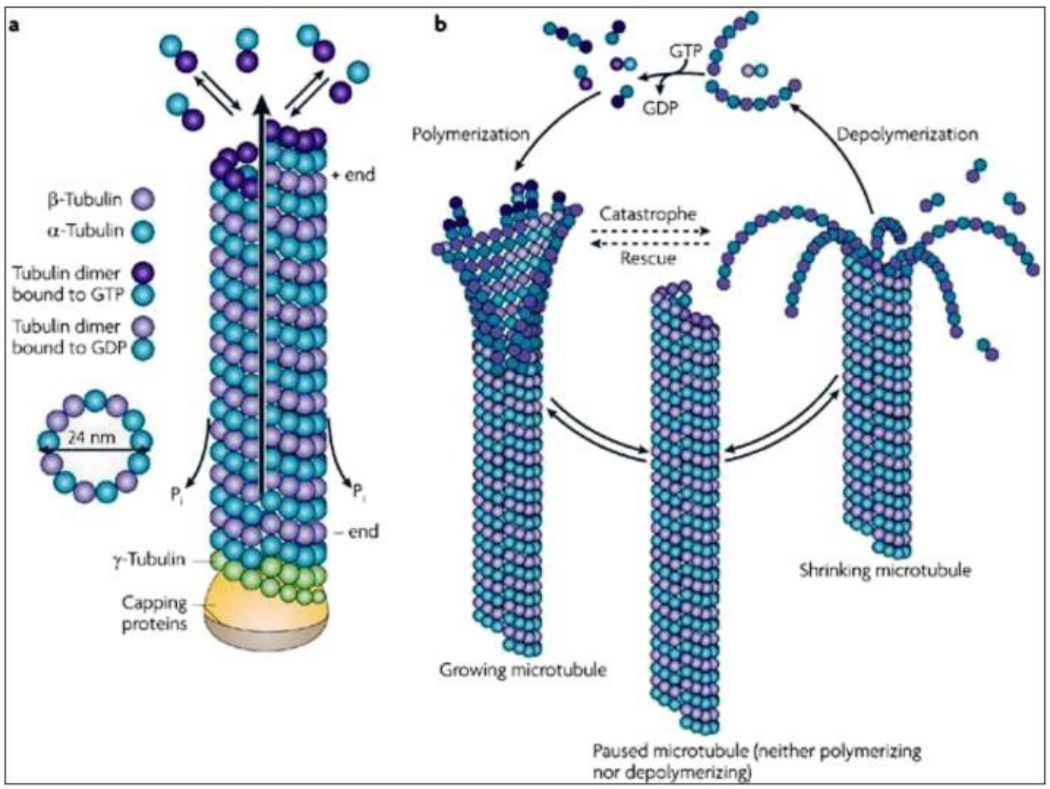


Figure I.15. Structure, polymérisation et dépolymérisation des microtubules [33].

I.II.4 Revue de la littérature

La méthode utilisée pour évaluer l'activité anticancéreuse des dérivés de la benzofuran sur le cancer du sein (cellules MCF-7) est une étude in vitro utilisant le test de coloration au sulforhodamine B (SRB assay). Cette technique permet de mesurer la cytotoxicité des composés en déterminant leur concentration inhibitrice 50 (IC50), c'est-à-dire la concentration nécessaire pour inhiber la croissance cellulaire de 50%. Les composés ont été exposés aux lignées cellulaires concernées, puis leur effet sur la prolifération tumorale a été quantifié par cette méthode colorimétrique

La méthode colorimétrique utilisée pour évaluer la cytotoxicité des dérivés consiste en l'utilisation du test de coloration au sulforhodamine B (SRB assay). Cette technique repose sur la capacité du colorant SRB à se lier aux protéines cellulaires, ce qui permet de mesurer la densité optique correspondant au nombre de cellules vivantes après traitement.

Concrètement :

1. Les cellules (MCF-7) sont cultivées en culture dans des plaques de microtitration.

2. Après traitement avec les différentes concentrations de composés, les cellules sont fixées en utilisant une solution d'acide trichloroacétique (TCA).
3. Le colorant SRB est ajouté aux puits, où il se lie aux protéines cellulaires par des interactions électrostatiques.
4. Après un lavage pour éliminer l'excès de colorant, la liaison est solubilisée dans un solvant (habituellement du Tris) pour permettre la lecture.
5. La densité optique (DO) est mesurée à l'aide d'un spectrophotomètre ou d'une microplaquelector à une longueur d'onde spécifique (habituellement 510 nm).

L'intensité de la coloration est proportionnelle au nombre de cellules vivantes, ce qui permet de déterminer la concentration inhibitrice 50 (IC50)[34].

CHAPITRE II

La modélisation par l'intelligence artificielle

II.1. Généralité

L'origine du mot « intelligence » remonte aux termes latins intellegere ou intelligere, qui signifient « choisir entre ». L'Intelligence Artificielle (IA) est une discipline interdisciplinaire qui se situe à la croisée de l'informatique et des mathématiques, englobant diverses méthodes algorithmiques et théories dans le but de créer des machines aptes à reproduire l'intelligence humaine [35][36].

L'objectif de l'intelligence artificielle est de doter une machine automatisée de la capacité de reproduire les aptitudes cognitives en perpétuel développement. Cela nécessite l'élaboration de systèmes capables d'apprendre, de réfléchir, de décider et d'interagir avec leur environnement à la manière d'un être humain. Cette réalisation scientifique est une régulation stimulante qui regroupe diverses sciences, théories et techniques dans le but de faire refléter les capacités cognitives évolutives d'une machine automatisée (Figure II.1) [37].



Figure II.1. L'IA et l'Homme : main dans la main pour la relation client [36].

L'intelligence artificielle englobe divers sous-domaines interconnectés, parmi lesquels l'apprentissage automatique (machine learning) occupe une place centrale. L'apprentissage automatique consiste à acquérir, à partir de l'expérience ou de bases de données, des règles implicites permettant de résoudre des problèmes spécifiques. Ce domaine repose fortement sur l'analyse statistique des données d'apprentissage, et les algorithmes utilisés s'appuient sur une grande variété de modèles mathématiques. Parmi ces algorithmes, les réseaux de neurones sont particulièrement importants (Figure II. 2). Ils sont à la base de l'apprentissage profond

(deep learning), un ensemble de techniques qui exploitent les réseaux de neurones complexes pour résoudre des problèmes particulièrement difficiles [38].

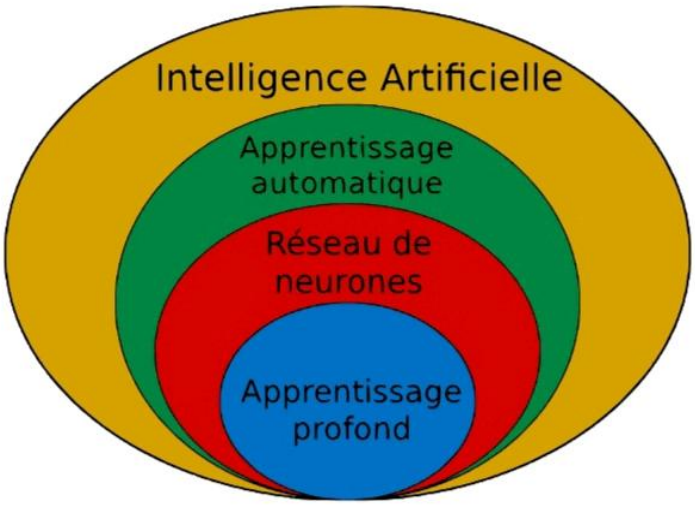


Figure II. 2. Schéma de décomposition du domaine de l'intelligence artificielle et de ces sous-domaines [39].

II.2. L'apprentissage automatique

L'apprentissage automatique (machine learning) n'est pas un concept récent. En 1958, le psychologue américain Frank Rosenblatt inventait le premier réseau de neurones artificiels, le « perceptron » [40]. Plusieurs définitions de l'apprentissage automatique existent. En 1959, Arthur Samuel le définit comme la capacité des ordinateurs à apprendre sans être explicitement programmés [39] (voir Figure II.3). Plus tard, en 1997, Tom Mitchell a formalisé cette notion d'apprentissage : « Un programme informatique apprend de l'expérience E relativement à une classe de tâches T et une mesure de performance P, si sa performance sur les tâches T, mesurée par P, s'améliore avec l'expérience E » [41]. En 2020, un sous-groupe de l'Union européenne a défini l'apprentissage automatique comme une méthode informatique visant à élaborer des modèles capables de résoudre des problèmes complexes à partir de données.

Ces définitions soulignent que l'apprentissage automatique, sous-domaine de l'intelligence artificielle, est une approche multidisciplinaire qui vise à concevoir des systèmes informatiques capables d'acquérir des connaissances de manière autonome à partir de données et de les exploiter pour résoudre des problèmes complexes. Aujourd'hui, les algorithmes d'apprentissage automatique sont largement utilisés dans de nombreux domaines, tels que la

médecine, le filtrage des courriels, la reconnaissance vocale, la vision artificielle, la protection des données, la sécurité, la santé, le marketing, la détection de fraudes, la recherche en ligne, les véhicules autonomes et les systèmes de recommandation, entre autres [39], [42].

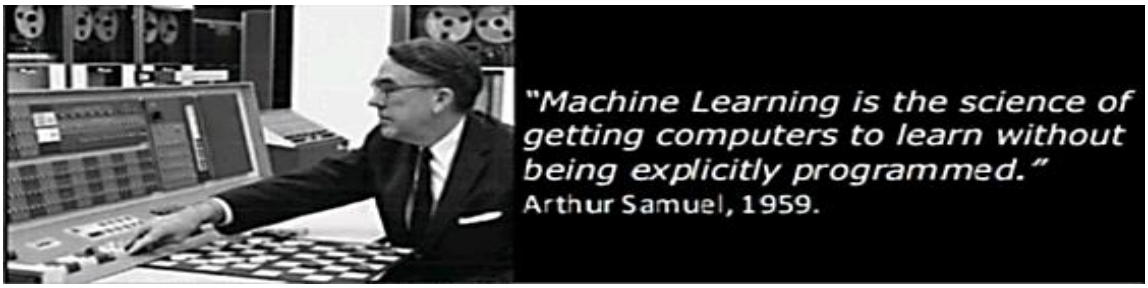


Figure II. 3. Définition de l'apprentissage automatique par Arthur Samuel [39].

II.2.1. Démarche des algorithmes d'apprentissage automatique

Les algorithmes d'apprentissage automatique nécessitent généralement deux ensembles de données distincts :

- Ensemble d'entraînement : Cet ensemble constitue la base de connaissances utilisée pour entraîner l'algorithme. Durant cette phase d'entraînement, les paramètres du modèle sont ajustés itérativement afin d'optimiser ses performances.
- Ensemble de test : Cet ensemble sert à évaluer les performances du modèle sur des données qu'il n'a jamais vues auparavant, permettant ainsi de mesurer sa capacité à généraliser ses connaissances.

La théorie de l'apprentissage automatique s'appuie sur des outils mathématiques issus de la théorie des probabilités et de la théorie de l'information. Ces outils permettent d'évaluer et de comparer l'optimalité de différentes méthodes d'apprentissage [43].

II.2.2. Objectifs de l'apprentissage automatique

L'objectif principal de l'apprentissage automatique est de développer des modèles capables d'améliorer leurs performances, d'identifier des structures complexes, et de résoudre de nouveaux problèmes en se basant sur les données issues de modèles antérieurs [35]. En d'autres termes, le machine learning vise à former des algorithmes capables de traiter efficacement et pertinemment diverses données [40].

Utilisé couramment dans la science des données et l'analyse de données, l'apprentissage automatique offre la possibilité de créer, tester et implémenter des algorithmes

d'analyse prédictive pour anticiper les tendances futures. L'objectif est de doter les algorithmes de la capacité d'identifier des "patterns" – des motifs récurrents – au sein des ensembles de données [35].

En somme, l'apprentissage automatique cherche à équiper les machines de la capacité d'apprendre à partir des données, d'identifier des motifs et de les généraliser, permettant ainsi de prendre des décisions ou de prédire de nouvelles informations. Ces algorithmes, qu'ils soient supervisés ou non supervisés, exploitent de vastes ensembles de données pour créer des modèles capables de réaliser des prédictions précises ou de fournir des analyses significatives.

II.2.3. Type de système d'apprentissage automatique

Quatre types principaux des algorithmes d'apprentissage automatique peuvent être identifiés dans la figure II. 4.

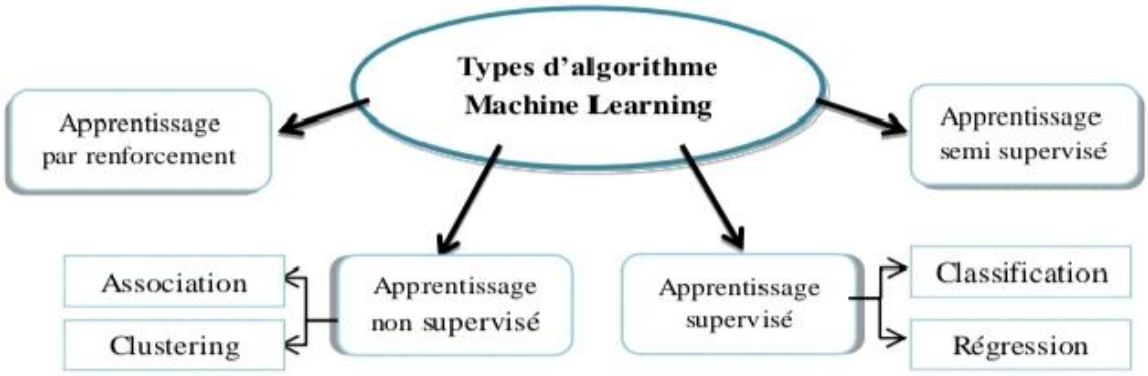


Figure II. 4. les quatre Types de système d'apprentissage automatique [35].

II.2.3.1. Apprentissage supervisé

L'apprentissage supervisé est une technique de machine learning où un algorithme est entraîné sur un ensemble de données étiquetées. L'objectif est d'apprendre une fonction de correspondance (mapping) entre les variables d'entrée (x) et une variable de sortie (Y), représentée par l'équation : $Y = f(x)$. Ainsi, l'algorithme apprend à prédire la sortie (Y) pour de nouvelles données d'entrée (x) inédites. Pendant l'entraînement, l'algorithme compare ses prédictions aux étiquettes réelles des données d'entraînement et ajuste son modèle en conséquence, améliorant ainsi sa capacité de prédiction [39]. La Figure II. 5 illustre ce processus.

L'apprentissage supervisé est couramment utilisé dans la science des données, l'intelligence artificielle et le machine learning, et il est appliqué à diverses tâches, notamment la classification de données, la régression linéaire, l'identification d'images et la reconnaissance vocale [39]. Les deux principales catégories de problèmes résolus par l'apprentissage supervisé sont la classification et la régression. La Figure II. 6 présente des exemples de ces deux catégories [40].

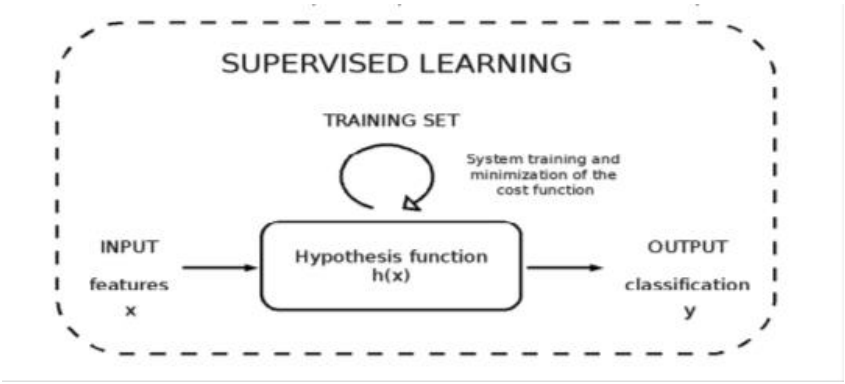


Figure II.5. Apprentissage supervisé : schéma d'une unité logistique [36].

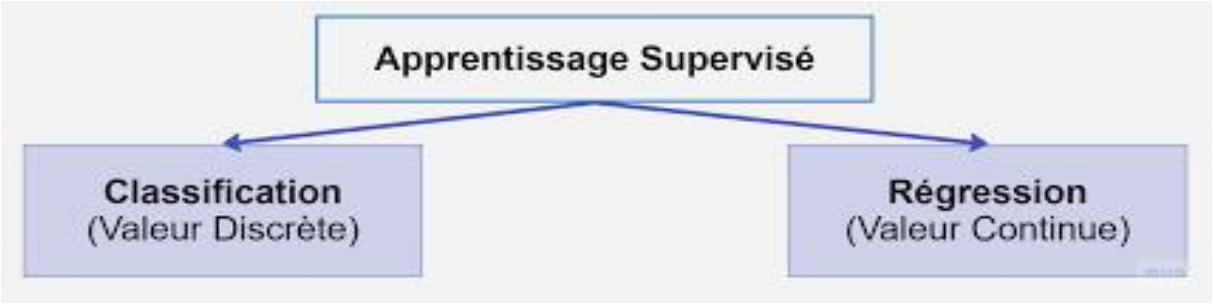


Figure II. 6. L'apprentissage supervisé.

- Classification : Un problème de classification se pose lorsque la variable cible est catégorique, c'est-à-dire qu'elle appartient à un ensemble de catégories distinctes (par exemple, "rouge", "bleu", "malade", "non malade") [40]. Un exemple en médecine est le diagnostic d'une maladie à partir d'un test sanguin ou PCR : le résultat est soit "positif" (malade), soit "négatif" (non malade).
- Régression : Un problème de régression se pose lorsque la variable de sortie est une valeur continue (par exemple, "dollars", "poids") [40]. Un exemple est un modèle qui permet de compter les cellules dans une image microscopique [44]. En médecine, on

pourrait aussi parler de régression de la taille d'une tumeur cancéreuse après un traitement ou d'une activité anticancéreuse comme adressé dans ce travail.

La Figure II.7 illustre la différence entre la classification (prédiction de catégories, comme "malade" vs. "sain", séparées par une limite de décision) et la régression (prédiction de valeurs continues, comme la durée de survie, modélisée par une ligne), deux types fondamentaux d'apprentissage supervisé.

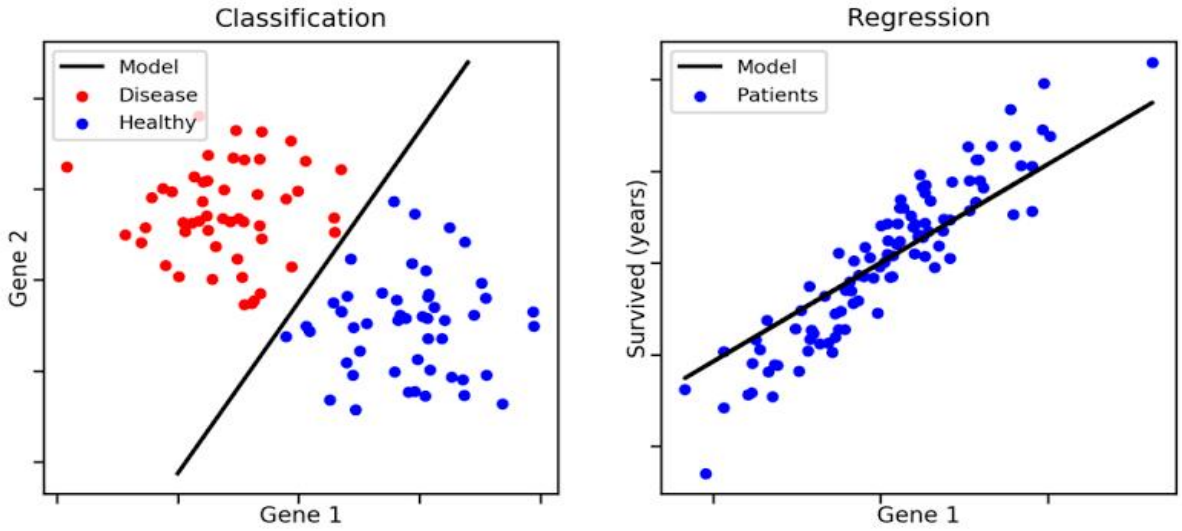


Figure II. 7. La différence entre la classification et la régression [40].

II.2.3.2. Apprentissage non supervisé

L'apprentissage non supervisé (Unsupervised Learning) est une technique d'apprentissage automatique où un algorithme est entraîné avec des données non étiquetées, afin de découvrir des structures et des relations au sein de ces données. Dans un système d'apprentissage non supervisé, l'algorithme a la liberté de détecter des motifs et des schémas dans les données sans orientation explicite. Les algorithmes peuvent exploiter des méthodes comme la réduction de dimensionnalité, le regroupement (clustering) et l'élaboration de modèles pour révéler des structures dissimulées au sein des données (voir Figure II.8) [39].

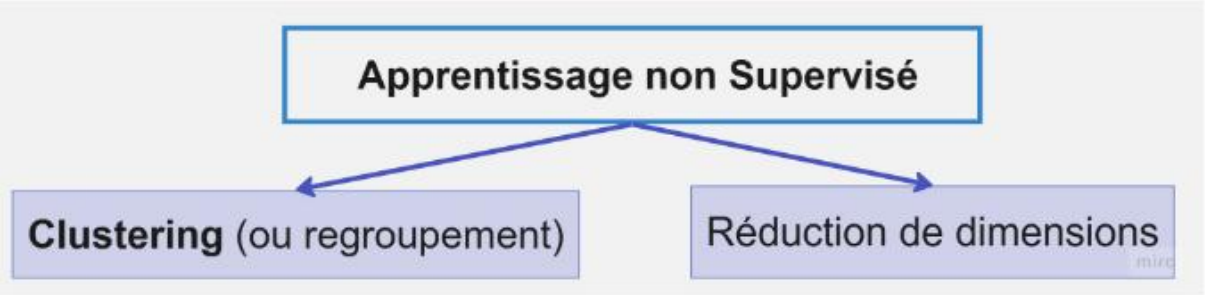


Figure II. 8. L'apprentissage non supervisé [39].

Le clustering (ou regroupement) est une méthode d'apprentissage non supervisé qui repose sur une mesure de similarité, généralement évaluée en termes de distances [44]. Une autre technique courante est la réduction de dimensionnalité, qui facilite la visualisation des données pour une meilleure compréhension.

La Figure II.9 illustre l'effet de l'apprentissage non supervisé, en particulier du clustering, sur un ensemble de données. Le graphique "Avant clustering" (Original unclustered data) montre des points de données sans structure apparente, tandis que le graphique "Après clustering" (Clustered data) montre les mêmes données après avoir été regroupées en trois clusters distincts, représentés par différentes couleurs. Cela illustre comment les algorithmes d'apprentissage non supervisé peuvent identifier des motifs et des structures inhérents dans des données non étiquetées, révélant des regroupements basés sur la similarité.

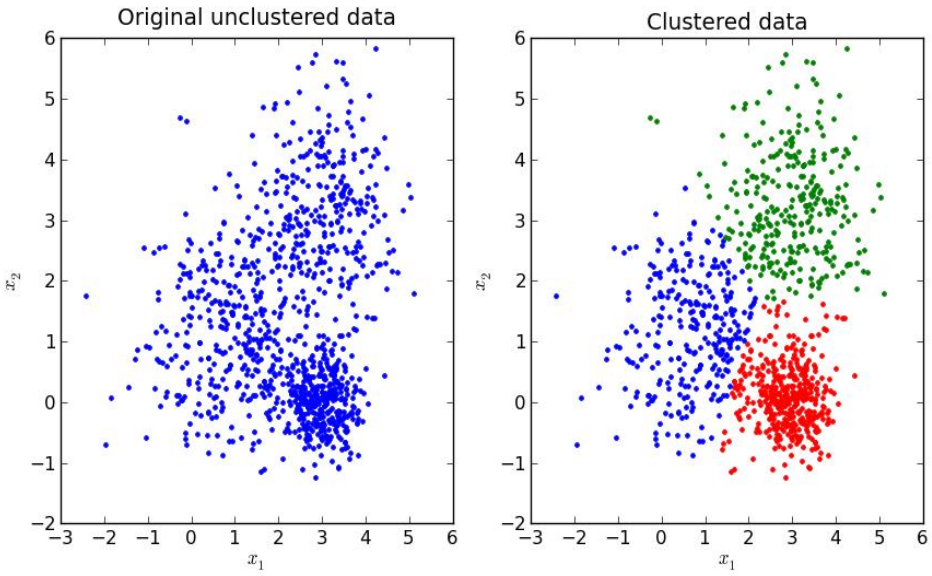


Figure II. 9. Avant et Après un apprentissage non supervisé (clustering) [45].

II.2.3.3. Autres types d'apprentissage automatique

Bien que l'apprentissage supervisé et non supervisé soient les approches les plus courantes, d'autres types d'apprentissage automatique existent, offrant des solutions pour des scénarios spécifiques. Parmi ces approches, on trouve l'apprentissage semi-supervisé et l'apprentissage par renforcement.

- Apprentissage semi-supervisé : Se situant entre l'apprentissage supervisé et non supervisé, l'apprentissage semi-supervisé utilise un ensemble de données d'entraînement partiellement étiqueté. L'instructeur fournit un signal d'apprentissage partiel, avec une ou plusieurs sorties cibles manquantes. Une variante de ce concept est la transduction, où toutes les instances du problème sont utilisées pendant la période d'apprentissage, à l'exception du fait que certaines cibles sont absentes [46].
- Apprentissage par renforcement : L'apprentissage par renforcement est une approche de l'IA axée sur l'apprentissage d'un système par le biais de ses interactions avec son environnement. L'algorithme, appelé "agent", interagit avec l'"environnement" pour trouver une solution optimale. L'agent essaie différentes solutions (exploration), observe la réaction de l'environnement, puis adapte son comportement (exploitation) pour trouver la meilleure stratégie. L'équilibre entre l'exploration et l'exploitation est un concept clé dans ce type de problème. La Figure II.10 illustre ce processus [43]. Cette méthode est particulièrement adaptée aux problèmes nécessitant un compromis entre la quête de récompenses à court terme et celle de récompenses à long terme [36].

II.3. Algorithmes d'apprentissage automatique

Il existe une grande variété d'algorithmes d'apprentissage automatique, chacun adapté à différents types de problèmes et de données. Quelques exemples courants incluent :

- k Nearest Neighbors (k-NN),
- K-Means,
- Arbres de décision,
- Naïve Bayes,
- Support Vector Machines (SVM),
- Réseaux de neurones artificiel (RNA).

Le choix de l'algorithme approprié est essentiel et dépend du type d'apprentissage souhaité (supervisé, non supervisé ou par renforcement), des exigences de la tâche et des caractéristiques des données disponibles.

Dans le cadre de ce travail, nous nous concentrerons plus particulièrement sur deux algorithmes : la "Forêt aléatoires" (Random Forest) et "l'Arbre de Décision" [35].

II.3.2. Arbre de Décision

Le modèle d'arbres de décision (Figure II.10.) s'inscrit parmi les modèles de l'apprentissage automatique statistique, qui repose sur une méthode géométrique pour diviser l'espace des caractéristiques. L'algorithme se base sur l'élaboration d'une collection de classificateurs, dénommés classificateurs faibles, qui sont progressivement intégrés dans le but de créer une division de l'espace des caractéristiques facilitant la classification des données. Cette série de classeurs est par la suite visualisée comme un ensemble de nœuds reliés et disposés selon une hiérarchie. Le produit final est un arbre qui illustre la séquence des classificateurs en positionnant en haut le nœud associé au classificateur ayant la plus grande capacité de classification parmi tous, et en descendant vers les autres nœuds, représentant ainsi des classificateurs moins performants dans un ordre décroissant d'importance. Le résultat final, est un arbre renversé comme celui représenté dans la figure II. 10 [47].

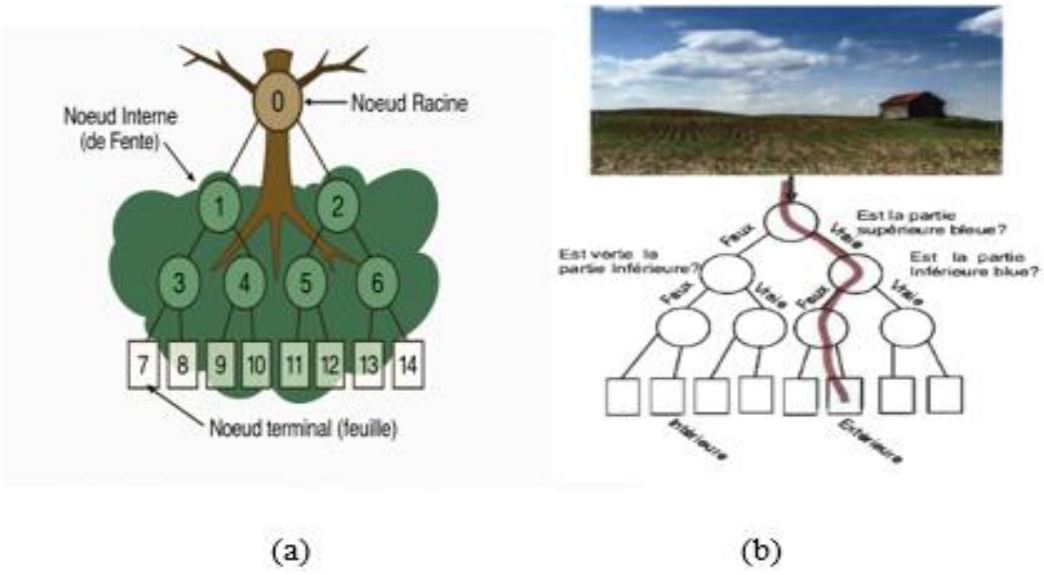


Figure II.10. Arbre de décision. (a) Structure générale d'un arbre de décision. (b) Exemple en montrant un arbre de décision illustratif utilisé pour décider si une photo représente une scène d'intérieur ou extérieure [47].

Comme son nom l'indique, un arbre de décision est un modèle basé sur une structure arborescente de règles "si-alors-sinon". En général, plus la profondeur de l'arbre et le nombre de nœuds sont importants, meilleures sont les performances du modèle [35].

Dans un arbre de décision (Figure II.11) :

- Une feuille indique une classe (la prédiction).
- Un nœud spécifie un test portant sur un certain attribut, chaque branche issue de ce nœud correspondant à une valeur possible de cet attribut [46].

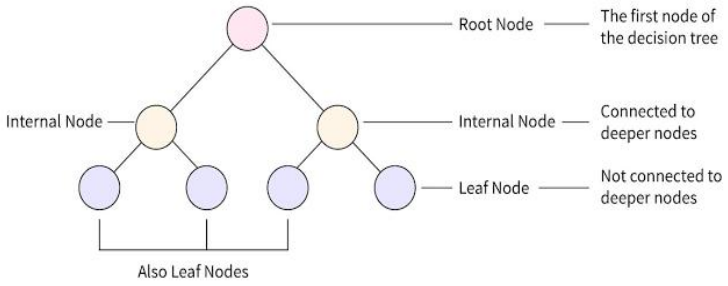


Figure II. 11. Arbre de décision [48].

II.3.3. L'algorithme de Random Forest

Le random forest (Forêts aléatoires) est une méthode ensembliste, introduit par Leo Breiman et Adèle Cutler en 2001 [41].

La régression par forêt aléatoire est une méthode d'apprentissage automatique qui combine plusieurs arbres de décision pour produire une prédiction plus précise et plus robuste. Voici une explication détaillée de son fondement mathématique. Chaque forêt aléatoire est constituée de multiples arbres de décision. Pour un problème de régression, un arbre de décision partitionne récursivement l'espace des caractéristiques X en régions R_j et attribue une valeur constante c_j à chaque région (Équation II. 1) [49].

$$f(x) = \sum_{j=1}^{J} c_j \quad \mathbf{1}_{x \in R_j} \qquad \text{(II. 1)}$$

Où $\mathbf{1}_{x\in R_j}$ est une fonction indicatrice égale à 1 si x appartient à la région R_j et 0 sinon.

La construction de l'arbre minimise l'erreur quadratique sur l'ensemble d'apprentissage (Équation II. 2).

$$\min_{R_j,c_j} \sum_{j=1}^{J} \sum_{i:x_i\in R_j} \left(y_i - c_j\right)^2 \quad \text{(II. 2)}$$

Pour chaque région R_j , la valeur optimale c_j est simplement la moyenne des observations y_i dans cette région on applique l'Equation II. 3.

$$c_j = \frac{1}{|i:x_i\in R_j|} \sum_{i:x_i\in R_j} y_i \quad \text{(II. 3)}$$

La forêt aléatoire utilise une technique appelée "bagging" (Bootstrap Aggregating) pour créer des arbres diversifiés. Pour B arbres et pour $b = 1, 2, ..., B$: échantillonnage avec remplacement d'un ensemble de données D_b à partir des données d'origine et Construction d'un arbre de décision T_b sur D_b. La prédiction finale est la moyenne des prédictions de tous les arbres (Équation II. 4).

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \quad \text{(II. 4)}$$

En plus du bagging, la forêt aléatoire introduit une randomisation supplémentaire dans la construction des arbres. Lors de la recherche de la meilleure division à chaque nœud, seul un sous-ensemble aléatoire de m caractéristiques (typiquement $m \approx \sqrt{p}$ pour un problème avec p caractéristiques) est considéré.

Si on considère que les arbres sont identiquement distribués avec une variance σ^2 et une corrélation moyenne ρ entre les arbres, alors la variance de la moyenne des B arbres est donnée par l'Équation II. 5.

$$Var\left(\hat{f}(x)\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \quad \text{(II. 5)}$$

Quand B devient grand, le second terme tend vers zéro, laissant $\rho\sigma^2$ comme variance incompressible. La randomisation supplémentaire des forêts aléatoires réduit ρ , diminuant ainsi la variance globale.

La puissance de la régression par forêt aléatoire réside dans sa capacité à combiner la simplicité des arbres de décision avec la robustesse du bagging et de la randomisation supplémentaire. Ce cadre mathématique explique pourquoi les forêts aléatoires produisent généralement des modèles précis, résistants au surapprentissage et capables de gérer efficacement des données de haute dimension.

La forêt aléatoire est un ensemble d'algorithmes d'apprentissage supervisé pour la classification et la régression, utilisés dans la modélisation prédictive et les techniques d'apprentissage automatique. La forêt aléatoire a attiré l'attention des universitaires en raison de sa rapidité et de sa précision de catégorisation. Elle rassemble les résultats et les prédictions de plusieurs arbres de décision afin de choisir le meilleur résultat : le mode des classes (la valeur qui apparaît le plus souvent dans les résultats de l'arbre de décision) ou la prédiction moyenne. La forêt aléatoire divise l'ensemble de données en deux sections : l'ensemble d'apprentissage et l'ensemble de test. Elle sélectionne ensuite aléatoirement plusieurs échantillons de l'ensemble d'apprentissage.

Le chercheur utilise ensuite l'arbre de décision pour chaque échantillon, qui divise chaque sélection en deux filles selon la meilleure division. Les utilisateurs doivent ensuite répéter la dernière étape pour voter pour chaque résultat de prédiction et sélectionner la prédiction la plus votée comme résultat final. Les principaux hyperparamètres de la forêt aléatoire servent soit à accroître la puissance prédictive du modèle, soit à le rendre plus rapide. Dans ce contexte, un nombre plus élevé d'arbres peut améliorer les performances et stabiliser les prédictions, mais il augmente également le temps de traitement. L'utilisation d'un nombre maximal de caractéristiques, en plus d'un nombre minimal de feuilles, peut améliorer les performances de l'algorithme. Une fois l'étape d'apprentissage terminée, le modèle peut être appliqué à un jeu de données de test. Cette procédure permet d'estimer les prédictions, puis de comparer les résultats aux valeurs attendues. La figure II. 13 illustre comment chaque arbre produit une sortie distincte après avoir été alimenté par un vecteur d'échantillons aléatoires indépendant. La forêt aléatoire est utilisée pour sa technique de généralisation des erreurs, et sa précision s'améliore à mesure que la forêt grandit. Après avoir sélectionné aléatoirement les caractéristiques pour le taux d'erreur, la précision dépend entièrement de la corrélation entre les arbres. Les caractéristiques de la forêt aléatoire peuvent être créées en suivant l'erreur et la

corrélation entre les nœuds. Par conséquent, la pertinence d'une variable peut être mesurée [50].

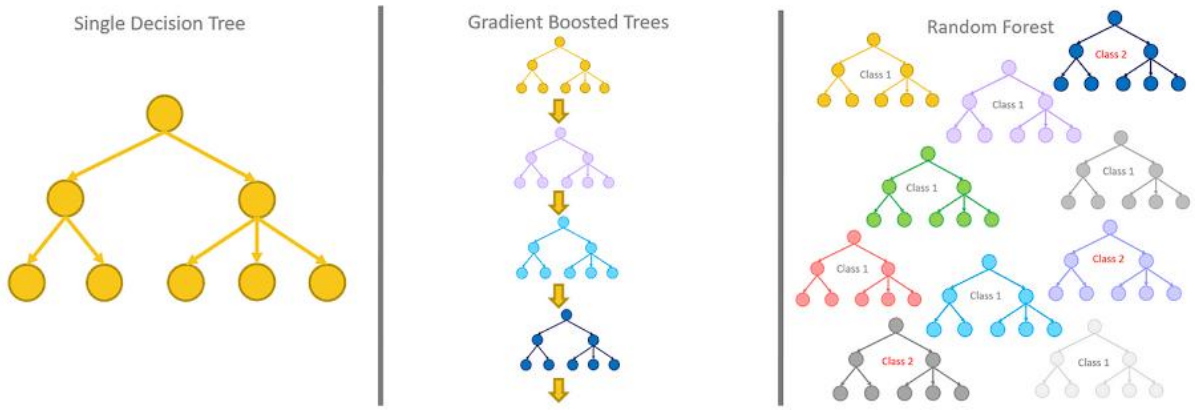


Figure II. 12. Illustration du Random forest [41].

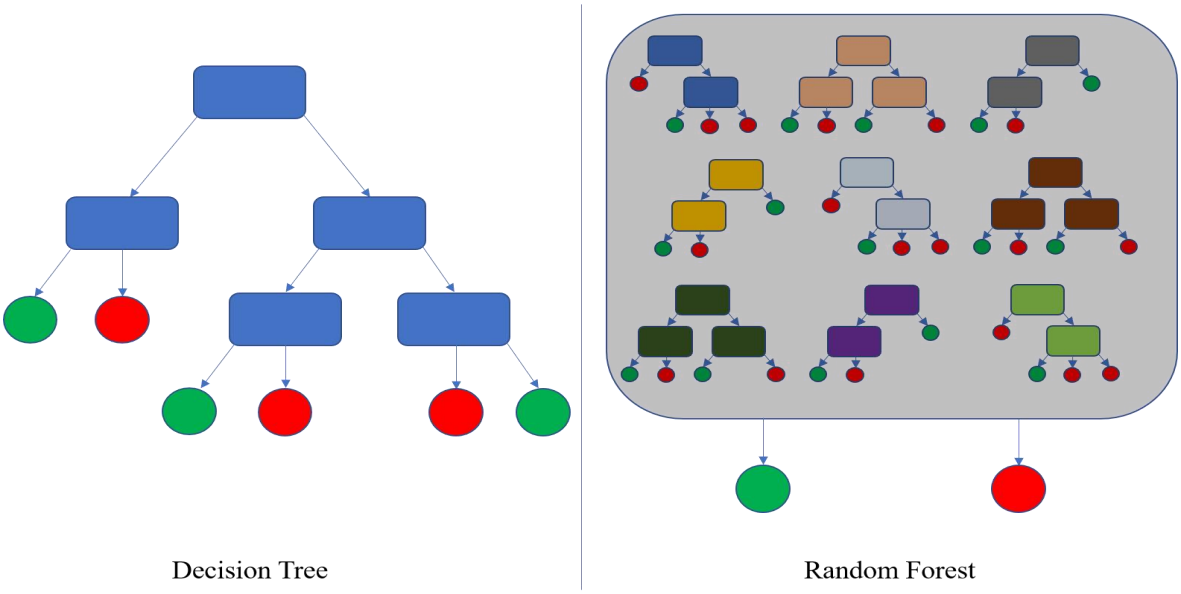


Figure II. 13. Une comparaison entre l'arbre de décision (DT) et la forêt aléatoire (RF) [50].

II.3.3.1. Avantages de Random Forest

L'algorithme Random Forest est très prisé pour ses nombreux avantages : facilité d'utilisation, efficacité, précision et polyvalence, puisqu'il peut être utilisé aussi bien pour la classification que pour la régression. Un avantage majeur est sa capacité à corriger l'overfitting (surapprentissage), un problème courant des arbres de décision, grâce à l'utilisation de la méthode de bagging et de la sélection aléatoire des fonctionnalités. De plus, Random Forest

conserve généralement sa précision même en présence de données manquantes et se révèle plus efficace qu'un seul arbre de décision lors de l'analyse de grandes bases de données [47], [51], [52]. La bibliothèque scikit-learn offre une implémentation simple et efficace de Random Forest en Python, facilitant son utilisation par les scientifiques des données.

II.3.3.2. Les inconvénients de la méthode random forest

Malgré ses nombreux atouts, Random Forest n'est pas sans inconvénients, car "*il n'y a pas de déjeuner gratuit en statistique*" [52], c'est-à-dire qu'aucune méthode ne surpasse toutes les autres pour tous les ensembles de données possibles. En raison du grand nombre d'arbres de décision qu'il utilise, Random Forest peut nécessiter une quantité importante de mémoire pour les projets de grande envergure, ce qui peut potentiellement le rendre moins rapide que d'autres algorithmes. Bien que Random Forest atténue le problème d'overfitting en utilisant des sous-ensembles aléatoires de caractéristiques et en construisant des arbres plus petits, il reste sensible à la configuration des hyperparamètres. Augmenter excessivement le nombre et la profondeur des arbres peut devenir contre-productif, entraînant une perte d'efficacité du modèle face à de nouvelles données, car il devient trop similaire aux données d'entraînement [41], [51].

II.4. Utilisation de l'intelligence artificielle pour les modelés QSAR

L'intégration des modèles d'intelligence artificielle dans les études de relations quantitatives structure-activité (QSAR) a révolutionné la découverte de médicaments et la chimie computationnelle. Traditionnellement, les modèles QSAR s'appuyaient sur des méthodes statistiques linéaires ou non linéaires pour établir des corrélations entre les descripteurs moléculaires et l'activité biologique. Cependant, l'IA, avec ses algorithmes d'apprentissage automatique sophistiqués, offre une capacité accrue à modéliser des relations complexes et non linéaires, à gérer des ensembles de données volumineux et diversifiés, et à identifier des motifs subtils qui échappent aux méthodes conventionnelles [53].

La lignée cellulaire MCF-7, dérivée d'un carcinome mammaire humain, est un modèle in vitro largement utilisé dans la recherche sur le cancer du sein en raison de sa sensibilité aux œstrogènes et de son profil biologique bien caractérisé. Modéliser l'activité anticancéreuse de différents dérivés sur cette lignée est essentiel pour comprendre les mécanismes d'action moléculaires, identifier des cibles thérapeutiques potentielles et concevoir des médicaments plus efficaces. En outre, ces modèles permettent de prédire l'activité de nouvelles molécules

avant leur synthèse, réduisant ainsi les coûts et accélérant le processus de découverte de médicaments [54], [55], [56], [57], [58].

L'analyse des valeurs des coefficients de corrélations (R^2) obtenues pour les différentes études QSAR axées sur l'activité anticancéreuse des dérivés sur la lignée MCF-7 révèle des niveaux de succès variés (Tableau II. 1). Les valeurs de R^2 les plus élevées, avoisinant 0.9, indiquent une forte corrélation entre les descripteurs moléculaires utilisés et l'activité biologique observée, suggérant que le modèle capture efficacement les relations structure-activité pertinentes. Ces modèles précis peuvent être utilisés avec confiance pour prédire l'activité de nouvelles molécules et guider la conception de dérivés plus puissants. En revanche, des valeurs de R^2 plus faibles pourraient indiquer que le modèle ne parvient pas à capturer pleinement la complexité des interactions moléculaires ou que des descripteurs plus pertinents sont nécessaires pour améliorer la capacité de prédiction du modèle.

Parallèlement aux études QSAR basées sur des données de lignées cellulaires, il est important de souligner l'intérêt des approches pharmacophores pour concevoir des inhibiteurs de cibles spécifiques. L'étude axée sur les benzimidazoles et l'inhibition de l'EGFR est un excellent exemple [59]. En construisant un modèle 3D-QSAR pharmacophore, les chercheurs ont identifié des caractéristiques structurales clés qui sont importantes pour l'activité inhibitrice, et en utilisant cette information, ils ont conçu de nouveaux composés potentiels plus rapidement que le simple criblage de nouveaux composés. Avec un coefficient de corrélation de 0.8412, elle soulige le potentiel des analyses pharmacophores combinées aux criblages virtuels pour identifier des composés actifs et conduire à la conception rationnelle de nouveaux médicaments.

Tableau II.1. Modèles QSAR pour l'activité inhibitrice de différentes classes d'échafaudages.

Échafaudage des composés	**R^2**	**Réf.**
(E)-5-(2-Arylvinyl)-1,3,4-oxadiazol-2-yl)benzenesulfonamides	0.93	[54]
2-[(4-Amino-6-N-substituted-1,3,5-triazin-2-yl)methylthio]-4-chloro-5-methyl-N-(1Hbenzo[*d*]imidazol-2(3H)ylidene)Benzenesulfonamides	0.86	[55]
Arylsulfonylhydrazones	0.796	[56]
6-bromo-coumarin-ethylidene-hydrazonylthiazolyl and 6-bromo-coumarin-thiazolyl-based derivatives	0.92	[57]

2,6,9-Trisubstituted Purine derivatives	0.97	[58]
Benzimidazoles	0.84	[59]

Dans le paysage actuel de la recherche sur le cancer du sein, une attention considérable est accordée au développement d'inhibiteurs ciblant des voies moléculaires spécifiques pour contrer la résistance aux traitements conventionnels. Toutefois, certaines classes de composés, bien que potentiellement intéressantes pour leurs propriétés pharmacologiques, restent relativement inexplorées. C'est le cas des dérivés de benzofuranes, une famille de molécules hétérocycliques présentes dans de nombreux composés bioactifs, mais dont l'activité anticancéreuse potentielle demeure sous-étudiée. Face à cette lacune, et en tenant compte des résultats de certaines études (comme celle sur les benzimidazoles présentée précédemment) où les modèles QSAR initiaux peuvent ne pas atteindre une précision optimale, il devient impératif de consacrer des efforts à la construction de modèles QSAR robustes et prédictifs spécifiquement adaptés aux dérivés de benzofuranes. En palliant ce manque de données et en optimisant la modélisation in silico, il est possible d'ouvrir de nouvelles perspectives pour le développement d'agents thérapeutiques innovants et plus efficaces contre le cancer du sein.

CHAPITRE III

Méthodologie

III. 1. Constitution de la base de données de dérivés de benzofurane

Dans cette étude, nous allons mettre en place une approche de modélisation par intelligence artificielle pour prédire l'activité anticancéreuse de dérivés de benzofurane contre les cellules MCF-7 (cancer du sein). La base de données sera construite à partir de 13 publications scientifiques (Tableau III. 1), totalisant 198 points de données de dérivés divers. Ces dérivés comprennent : des pyrazolines et pyridines portant un fragment benzofurane, des nitriles contenant un échafaudage benzofurane, de l'acide 2-(1-benzofuran-2-yl) quinoléine-4-carboxylique substitué et ses esters, des dérivés de benzofurane et d'indole, des 2-pyrazolyl benzofuranes, des hybrides de benzofurane, des chalcones et thiopyrimidines benzofurane, des 3-(pipérazinylméthyl)benzofuranes, des benzo[*c*]furan-chalcones, des benzofuranes et fonctions hétérocycliques liées au 5*H*-furo[3,2-*g*]chromone, des benzofurane-thiazoles, -azos, -hydrazos et -pipérazines, des benzofuranes et furochromones, ainsi que des benzofurane-2-carboxamides. Cette diversité de structures moléculaires assure une exploration large de l'espace chimique et permet d'établir des corrélations robustes entre la structure et l'activité anticancéreuse. La gamme d'activité de ces composés, mesurée sur les cellules MCF-7, s'étend de 0.000000367 µM à 127.41 µM, ce qui offre une base de données hétérogène et pertinente pour l'entraînement d'un modèle prédictif.

III. 2. Calcul des descripteurs et empreintes moléculaires

Afin de quantifier les propriétés physico-chimiques et structurales des 198 dérivés de benzofurane, un ensemble complet de descripteurs moléculaires et d'empreintes moléculaires a été calculé. Ce calcul a été réalisé en utilisant le logiciel PaDEL-Descriptor v2.20 (Figure III.1) [60].

Les descripteurs moléculaires sont des représentations numériques des caractéristiques d'une molécule. Ils incluent divers types de paramètres, tels que :

- Les descripteurs constitutionnels représentant le nombre d'atomes, nombre de liaisons, poids moléculaire, etc. Ils décrivent la composition de la molécule.
- Les descripteurs topologiques représentant les indices de connectivité (Wiener, Randić), indices de Balaban, etc. Ils caractérisent la connectivité des atomes dans la molécule.
- Les descripteurs géométriques représentant la surface moléculaire, volume moléculaire, etc. Ils rendent compte de la forme et de la taille de la molécule.

- Les descripteurs électroniques représentant les charges atomiques, polarisabilité, moments dipolaires, etc. Ils décrivent la distribution des électrons dans la molécule.

Les descripteurs physico-chimiques comme le LogP (coefficient de partage octanol/eau), solubilité, etc. Ils sont liés au comportement de la molécule dans différents environnements.

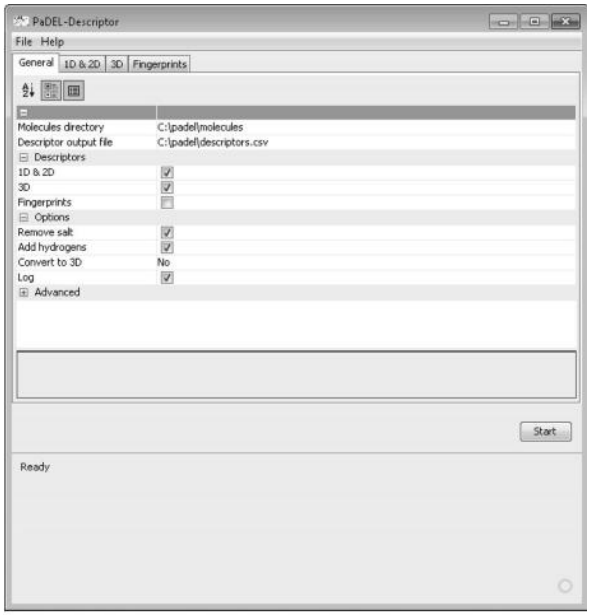


Figure III.1. GUI for PaDEL-Descriptor [61].

Les empreintes moléculaires sont des représentations binaires (chaînes de bits) qui codent la présence ou l'absence de certains fragments structuraux, motifs ou caractéristiques dans une molécule. Dans cette étude, nous avons utilisé les empreintes moléculaires PubChem Fingerprints, qui sont constituées de 881 bits. Chaque bit représente une caractéristique ou un fragment spécifique de la molécule. La présence (bit à 1) ou l'absence (bit à 0) de chaque caractéristique est déterminée par un algorithme pré-défini. PubChem Fingerprints couvrent un large éventail de fragments et de propriétés, ce qui en fait un choix polyvalent pour la modélisation Relations Quantitatives Structure-Activité (QSAR). L'utilisation combinée de descripteurs moléculaires et d'empreintes moléculaires permet de capturer une information riche et variée sur la structure des molécules, indispensable pour construire un modèle de prédiction robuste [31].

Tableau III.1. Caractéristiques de la base de données de dérivés de benzofurane utilisés pour la modélisation de l'activité anticancéreuse contre les cellules MCF-7.

Dérivés	**Nombre de points**	**Etendue de l'activité (µM)**	**Écart-type**	**Réf**
Pyrazolines et pyridines portant un fragment benzofurane.	5	10.43 – 12.22	0.89	[62]
Nitriles contenant un échafaudage benzofurane.	11	8.36 – 81.58	23.64	[63]
Acide 2-(1-benzofuran-2-yl) quinoléine-4-carboxylique substitué et de ses esters.	17	0.52 – 25.97	6.54	[64]
Benzofurane et d'indole.	22	5.8 – 11.4	1.55	[65]
2-pyrazolyl benzofurane.	11	3.96 – 83.68	25.35	[66]
Hybrides de benzofurane.	7	11.58 – 76.91	25.53	[67]
chalcone et thiopyrimidine benzofurane.	21	0.000000367 – 0.00856	0.003	[68]
3-(pipérazinylméthyl)benzofurane.	20	2.92 – 127.41	45.89	[69]
benzo[*c*]furan-chalcones.	23	0.000355 – 69.09	20.78	[70]
Benzofuranes et fonctions hétérocycliques diverses liées au 5*H*-furo[3,2-*g*]chromone.	14	0.004 – 0.87	0.22	[71]
Benzofurane-thiazole, benzofurane-azo, benzofurane-hydrazo et benzofurane-pipérazine.	17	4.8 – 88.46	27.02	[30]
Benzofuranes et des furochromones.	16	5.56 – 20.8	3.89	[72]
Benzofurane-2-carboxamides.	14	4.9 – 82.8	25.42	[73]
Global	198	0.000000367 – 127.41	28.08	

III. 3. Nettoyage et préparation des données

La qualité des données est primordiale pour construire un modèle QSAR performant. C'est pourquoi une étape rigoureuse de nettoyage et de préparation des données a été effectuée. Cette étape a visé à réduire la dimensionalité de l'ensemble de descripteurs et empreintes moléculaires, en éliminant les colonnes redondantes, non informatives ou susceptibles de nuire à la robustesse du modèle [74].

Le processus s'est déroulé comme suit :

- L'ensemble de données initial comportait un certain nombre de composés et un ensemble de descripteurs et empreintes moléculaires.
- Les colonnes contenant des valeurs manquantes (NaN) ont été supprimées.
- Les colonnes avec variance nulle ont été éliminées.
- Les colonnes avec variance très faible ont également été supprimées.
- Les descripteurs présentant des valeurs extrêmes ont été éliminés.
- Les descripteurs fortement corrélés ont été éliminés, conservant le plus pertinent.

III. 4. Sélection des caractéristiques avec l'algorithme Boruta pour la régression

Afin d'identifier les descripteurs moléculaires et empreintes moléculaires les plus pertinents pour la prédiction de l'activité anticancéreuse, un processus de sélection des caractéristiques basé sur l'algorithme Boruta pour la régression a été mis en œuvre [75].

L'algorithme Boruta est un algorithme de sélection de caractéristiques de type “wrapper” qui s'appuie sur les forêts aléatoires (Figure III. 2). Il vise à identifier toutes les caractéristiques pertinentes dans un ensemble de données, en capturant à la fois les dépendances fortes et faibles avec la variable cible. Contrairement à d'autres méthodes qui sélectionnent un sous-ensemble fixe de caractéristiques, Boruta tente de déterminer quelles caractéristiques sont véritablement importantes, plutôt que simplement celles qui sont “les plus importantes”.

Pour chaque descripteur ou empreinte moléculaire réel, une "caractéristique fantôme" est créée en permutant aléatoirement les valeurs de ce descripteur à travers tous les composés. Les caractéristiques fantômes représentent ainsi des descripteurs aléatoires, non corrélés avec

l'activité biologique. Ensuite, une forêt aléatoire est entraînée en utilisant l'ensemble de données combinant les descripteurs réels et les descripteurs fantômes, avec l'activité anticancéreuse comme variable cible. La forêt aléatoire calcule l'importance de chaque descripteur (réel et fantôme) en fonction de sa capacité à réduire l'erreur du modèle.

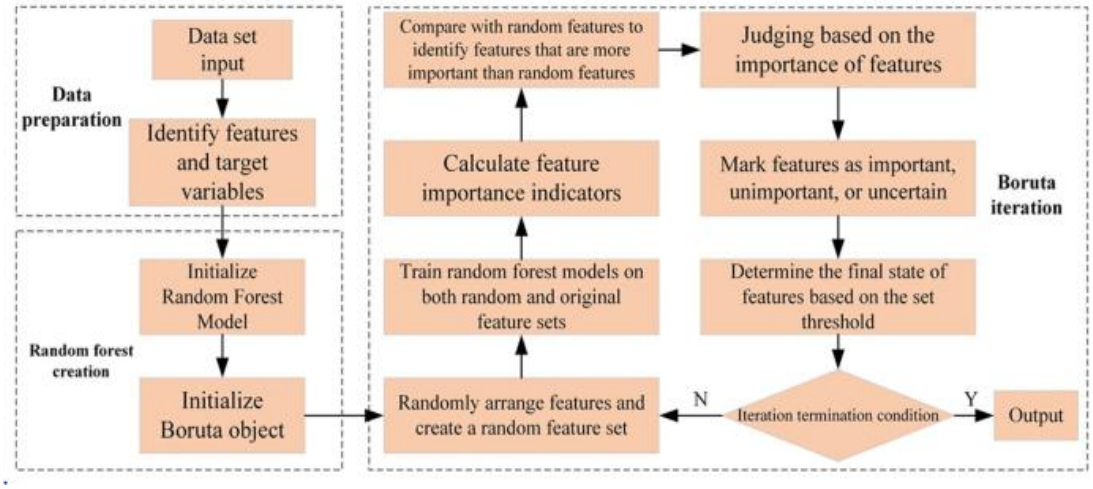


Figure III.2. Organigramme de sélection des fonctionnalités de Boruta [76].

Pour chaque descripteur réel, son importance est comparée à l'importance du meilleur descripteur fantôme (celui ayant l'importance la plus élevée). Si l'importance du descripteur réel est significativement supérieure à celle du meilleur descripteur fantôme (selon un test statistique), le descripteur est considéré comme important et est conservé. Si l'importance du descripteur réel est significativement inférieure à celle du meilleur descripteur fantôme, le descripteur est considéré comme non important et est éliminé. Si l'importance du descripteur réel n'est pas significativement différente de celle du meilleur descripteur fantôme, le descripteur est considéré comme tentatif et est maintenu pour l'itération suivante. À chaque itération, les caractéristiques fantômes sont recréées et une nouvelle forêt aléatoire est entraînée. Les caractéristiques tentatives sont réévaluées jusqu'à ce qu'elles soient confirmées ou rejetées. À la fin des itérations, l'algorithme Boruta fournit un ensemble de caractéristiques confirmées, qui sont considérées comme les plus pertinentes pour la prédiction de l'activité anticancéreuse [77].

III. 5. Augmentation des données avec l'algorithme SMOGN

Afin d'améliorer la robustesse et la performance du modèle QSAR, une technique d'augmentation des données basée sur l'algorithme SMOGN (Synthetic Minority Oversampling Technique for Regression with Gaussian Noise) a été mise en œuvre [78].

Dans les problèmes de modélisation, en particulier lorsque la taille de l'ensemble de données est limitée, l'augmentation des données joue un rôle crucial pour plusieurs raisons. Un ensemble de données restreint peut conduire à un surapprentissage, où le modèle apprend par cœur les données d'entraînement et ne parvient pas à généraliser correctement à de nouvelles données. L'augmentation des données permet d'accroître la diversité de l'ensemble d'entraînement, réduisant ainsi le risque de surapprentissage. En entraînant le modèle sur un ensemble de données plus vaste et diversifié, on améliore sa robustesse et sa capacité à gérer des variations dans les données réelles [79]. L'augmentation des données permet donc d'explorer plus en profondeur l'espace des descripteurs moléculaires, identifiant potentiellement des relations structure-activité qui n'auraient pas été détectées avec l'ensemble de données initial.

III. 6. Modélisation de l'activité anticancéreuse des benzofuranes par Forêts Aléatoires

Dans cette étude, nous avons utilisé l'algorithme de machine learning des Forêts Aléatoires (Random Forest) pour modéliser l'activité anticancéreuse des dérivés de benzofurane contre les cellules MCF-7 [80].

Les Forêts Aléatoires (Figure III. 3) sont un algorithme d'apprentissage supervisé de type ensemble, c'est-à-dire qu'il combine les prédictions de plusieurs modèles individuels pour obtenir une prédiction plus précise et robuste. Dans le cas des Forêts Aléatoires, les modèles individuels sont des arbres de décision. L'ensemble de données d'entraînement initial est échantillonné de manière aléatoire avec remplacement, créant ainsi plusieurs sous-ensembles de données (appelés “bootstrap samples”). Chaque arbre de décision est entraîné sur un sous-ensemble différent. Cela permet d'introduire une certaine diversité dans l'ensemble d'arbres. Lors de la construction de chaque arbre de décision, à chaque nœud de l'arbre, un sous-ensemble aléatoire des descripteurs (ou caractéristiques) est sélectionné. Seuls ces descripteurs sont pris en compte pour déterminer la meilleure séparation des données à ce nœud. Cela contribue également à la diversité des arbres et réduit la corrélation entre eux.

Chaque arbre de décision est construit en divisant récursivement l'ensemble de données en sous-ensembles de plus en plus homogènes, en fonction des valeurs des descripteurs sélectionnés aléatoirement. Le critère de division est la minimisation de l'erreur quadratique moyenne (pour la régression). Les arbres sont construits jusqu'à une certaine

profondeur ou jusqu'à ce qu'un critère d'arrêt soit atteint (par exemple, un nombre minimal d'échantillons par feuille). Pour effectuer une prédiction pour un nouveau composé, chaque arbre de décision de la forêt prédit une valeur d'activité. Les prédictions des différents arbres sont ensuite agrégées pour obtenir la prédiction finale. Dans le cas de la régression, la prédiction finale est la moyenne des prédictions des arbres [31].

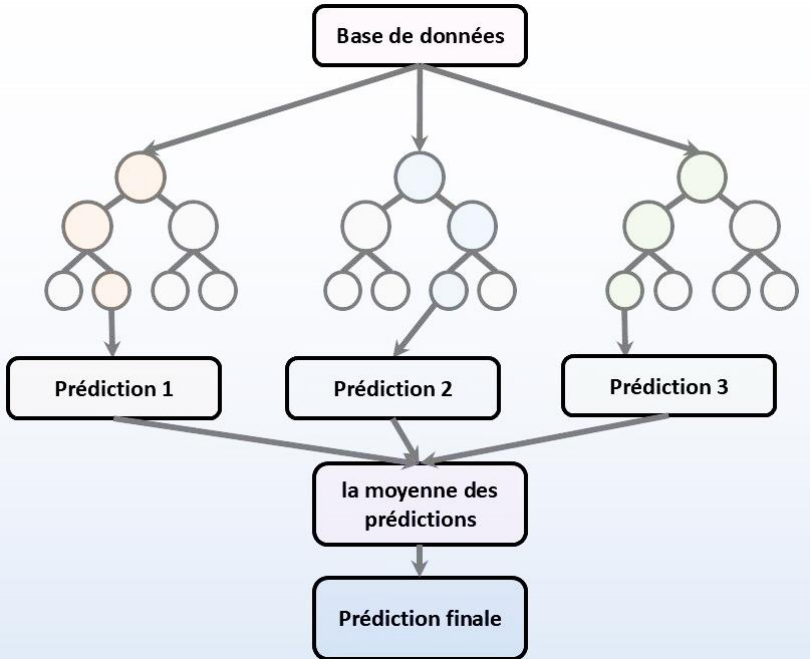


Figure III. 3. Représentation de l'algorithme d'apprentissage Forêts Aléatoires.

III. 7. Optimisation des hyperparamètres du modèle Forêts Aléatoires avec l'algorithme ALO (Ant Lion Optimizer)

Afin d'optimiser les performances du modèle Forêts Aléatoires pour la prédiction de l'activité anticancéreuse, une optimisation des hyperparamètres a été réalisée en utilisant l'algorithme ALO [31].

L'algorithme ALO (Figure III. 4) est un algorithme d'optimisation métaheuristique inspiré par le mécanisme de chasse des fourmilions (Ant Lions). Il s'agit d'une technique d'optimisation qui imite le comportement des fourmilions et des fourmis pour résoudre des problèmes complexes [81]. Dans le contexte de l'optimisation des hyperparamètres, ALO recherche la combinaison optimale de valeurs pour des paramètres tels que le nombre d'arbres dans la forêt (n_estimators), la profondeur maximale des arbres (max_depth), le nombre

minimum d'échantillons requis pour diviser un nœud (min_samples_split), et le nombre minimum d'échantillons requis dans une feuille (min_samples_leaf).

Les étapes de l'optimisation sont :

- Une population initiale de solutions potentielles (fourmis) est générée de manière aléatoire dans l'espace de recherche des hyperparamètres. Chaque solution représente une combinaison spécifique de valeurs pour les hyperparamètres du modèle Forêts Aléatoires.
- Chaque solution (fourmi) est évaluée en utilisant une fonction objective qui mesure la performance du modèle Forêts Aléatoires entraîné avec les hyperparamètres correspondants. Dans cette étude, la fonction objective est la négation du score R^2 moyen obtenu par validation croisée (5 K-fold cross-validation). L'objectif est de minimiser cette fonction, ce qui équivaut à maximiser le score R^2.
- Pour chaque combinaison d'hyperparamètres proposée par ALO, un modèle Forêts Aléatoires est créé et entraîné. La performance du modèle est évaluée à l'aide d'une validation croisée (5 K-fold cross-validation). Cette approche consiste à diviser les données d'entraînement en k (5) sous-ensembles, à entraîner le modèle sur k-1 sous-ensembles et à évaluer sa performance sur le sous-ensemble restant. Ce processus est répété k fois, chaque sous-ensemble étant utilisé une fois comme ensemble de validation. Le score R^2 moyen sur les k validations est utilisé comme mesure de performance. La négation de ce score est utilisée pour la minimisation.
- Les fourmilions représentent les meilleures solutions trouvées jusqu'à présent. Les fourmilions sont sélectionnés en fonction de leur performance (fitness) : les meilleures solutions deviennent des fourmilions.
- L'algorithme ALO itère à travers les étapes suivantes jusqu'à ce qu'un critère d'arrêt soit atteint (un nombre maximal d'itérations) :
    - Déplacement des fourmis : Les fourmis (solutions potentielles) se déplacent aléatoirement dans l'espace de recherche, influencées par les fourmilions. Le déplacement des fourmis est modélisé par une marche aléatoire (random walk).
    - Piégeage des fourmis : Les fourmis sont piégées dans les "pièges" des fourmilions. La taille du piège est proportionnelle à la fitness du fourmilion. Cela signifie que les fourmilions les plus performants attirent davantage de fourmis.

- Jet des fourmis dans le sable : Lorsqu'une fourmi atteint le centre du piège (c'est-à-dire, améliore la solution du fourmilion), le fourmilion jette du sable et change de position. Ce processus simule l'amélioration de la solution du fourmilion en exploitant l'information apportée par la fourmi.
- Mise à jour du fourmilion élite : Le meilleur fourmilion de chaque itération (le fourmilion élite) est conservé et utilisé pour guider la recherche.

- L'algorithme s'arrête lorsqu'un certain nombre d'itérations est atteint ou lorsqu'une solution satisfaisante est trouvée.
- À la fin de l'exécution de l'algorithme, la solution correspondant au meilleur fourmilion (ou au fourmilion élite) est sélectionnée comme la combinaison optimale d'hyperparamètres pour le modèle Forêts Aléatoires.

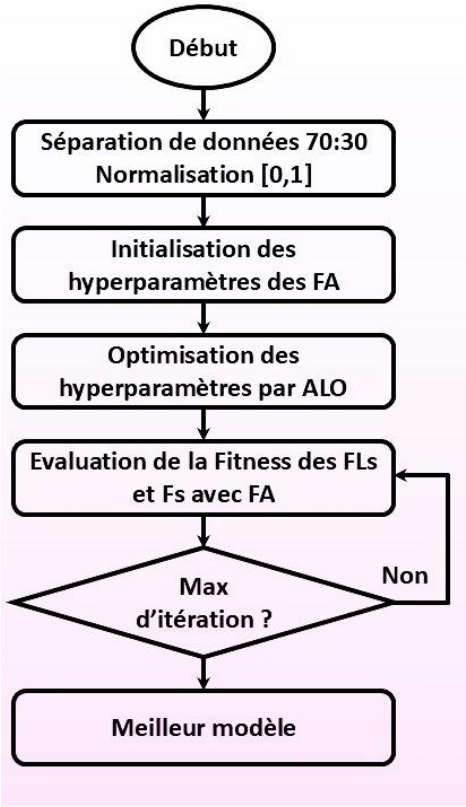


Figure III. 4. Diagramme de flux de l'optimisation ALO-FA.

III. 8. Critères d'évaluation

L'évaluation des performances du modèle de forêt aléatoire (FA) développé pour prédire l'activité anticancéreuse des dérivés benzofuraniques contre le cancer du sein a été réalisée à

l'aide de plusieurs indicateurs statistiques classiques [82]. Le coefficient de détermination R^2, qui mesure la proportion de variance expliquée par le modèle, est défini par l'équation III. 1.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad \text{(III. 1)}$$

Où y_i représente les valeurs expérimentales, $\hat{y}_i$ les valeurs prédites, et $\bar{y}_i$ la moyenne des valeurs expérimentales. Une valeur de R^2 proche de 1 indique une forte capacité prédictive.

L'erreur absolue moyenne (MAE) quantifie l'écart moyen entre les prédictions et les valeurs réelles, selon l'équation III. 2.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{(III. 2)}$$

L'erreur quadratique moyenne (MSE), qui pénalise davantage les erreurs importantes, est donnée par l'équation III. 3 et sa racine carrée (RMSE), qui ramène l'unité de l'erreur à celle de la variable de sortie, est définie par l'équation III. 4.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{(III. 3)}$$

$$RMSE = \sqrt{MSE} \quad \text{(III. 4)}$$

Enfin, le pourcentage d'erreur absolue moyenne (MAPE) a été utilisé pour fournir une mesure relative de l'erreur (équation III. 5).

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad \text{(III. 5)}$$

Toutes ces métriques ont été calculées à la fois sur les ensembles d'apprentissage et de test afin d'évaluer la qualité d'ajustement du modèle, ainsi que sa capacité de généralisation.

En complément de l'évaluation statistique, l'analyse du domaine d'applicabilité (AD) a été réalisée afin d'identifier les composés pour lesquels les prédictions peuvent être considérées comme fiables [83]. Cette analyse s'appuie sur la méthode du levier (leverage), qui évalue l'influence de chaque composé en fonction de sa position dans l'espace des descripteurs. Le levier h_i pour chaque échantillon est calculé à partir de la matrice X des descripteurs normalisés, selon l'équation III. 6.

$$h_i = x_i^T (X^T X)^{-1} x_i \quad \text{(III. 6)}$$

Où x_i est le vecteur descripteur de l'échantillon i. Le seuil critique de levier (h^*) a été défini comme étant trois fois la moyenne des leviers de l'ensemble d'entraînement (équation III. 7).

$$h^* = 3 \cdot \frac{1}{n} \sum_{i=1}^{n} h^* \qquad \text{(III. 7)}$$

Les composés pour lesquels $h_i \leq h^*$ sont considérés comme étant à l'intérieur du domaine d'applicabilité, tandis que ceux pour lesquels $h_i > h^*$ sont en dehors du domaine, et leurs prédictions doivent être interprétées avec prudence. Cette approche renforce la fiabilité et la validité des résultats obtenus avec le modèle FA.

CHAPITRE IV

Résultats et Discussion

IV.1. Nettoyage et préparation des données

La phase initiale de modélisation QSAR a consisté en un nettoyage et une préparation des données, des étapes essentielles pour assurer la qualité et la fiabilité du modèle final. L'ensemble de données initial, composé de 198 dérivés de benzofurane, a été caractérisé par un ensemble étendu de 2325 descripteurs moléculaires 1D/2D et empreintes moléculaires PubChem, capturant ainsi une diversité d'informations structurelles et physico-chimiques.

Cependant, cet ensemble de données initial contenait des imperfections susceptibles de nuire à la performance du modèle. Tout d'abord, l'identification et la suppression des colonnes contenant des valeurs manquantes (NaN) ont réduit le nombre de descripteurs à 1503. Ces valeurs manquantes, potentiellement dues à des erreurs lors du calcul des descripteurs ou à des limitations du logiciel, auraient pu introduire des biais et perturber l'entraînement du modèle.

Ensuite, l'élimination des descripteurs constants et quasi-constants a permis de supprimer les variables non informatives qui n'apportent aucune discrimination entre les composés. Cette étape a conduit à une réduction significative du nombre de descripteurs, passant de 1503 à 989, tout en conservant l'information pertinente pour la modélisation.

La suppression des colonnes contenant des valeurs extrêmes (outliers) a contribué à améliorer la robustesse du modèle en éliminant les descripteurs potentiellement erronés ou spécifiques à des composés atypiques. Cette étape a conduit à une légère réduction du nombre de descripteurs à 979.

Enfin, la réduction de la colinéarité entre les descripteurs a permis de minimiser la redondance de l'information et d'améliorer la stabilité du modèle. L'élimination des descripteurs fortement corrélés a conduit à une réduction drastique du nombre de descripteurs, passant de 979 à seulement 281. Ce sous-ensemble final de descripteurs, non redondant et informatif, a servi de base pour la construction du modèle RQSA.

IV.2. Sélection des caractéristiques par l'algorithme Boruta

L'algorithme Boruta a permis d'identifier les descripteurs moléculaires les plus pertinents pour la modélisation de l'activité anticancéreuse des dérivés de benzofurane, optimisant ainsi la sélection des caractéristiques afin d'améliorer la robustesse du modèle QSAR. Parmi les descripteurs retenus, (ATSC5c, ATSC4e, ATSC5e, ATSC8i, MATS3c, MATS5c, MDEO-12, n5Ring et PubchemFP562) jouent un rôle clé dans la caractérisation des composés étudiés.

Les descripteurs ATSC et MATS permettent d'analyser la connectivité des atomes ainsi que la distribution de leurs propriétés atomiques, notamment la charge, la masse et la polarisabilité, tandis que MDEO-12 met en évidence l'importance des liaisons au sein des structures moléculaires. Par ailleurs, n5Ring identifie la présence de cycles à 5 atomes, et PubchemFP562 représente un fragment structural spécifique (Br-C-C=O), qui pourrait être déterminant dans l'interaction avec la cible biologique.

L'analyse de la matrice de corrélation a révélé des relations positives et négatives entre ces descripteurs, suggérant des interactions complexes pouvant influencer l'activité anticancéreuse des dérivés de benzofurane. De plus, l'absence de forte colinéarité garantit que chaque descripteur apporte une information unique et complémentaire, renforçant ainsi la fiabilité du modèle prédictif. Cette approche combinée, alliant sélection optimisée des descripteurs et étude approfondie des corrélations, contribue à une meilleure compréhension des facteurs déterminants dans l'efficacité des composés et à l'amélioration de la précision des prédictions du modèle QSAR**.**

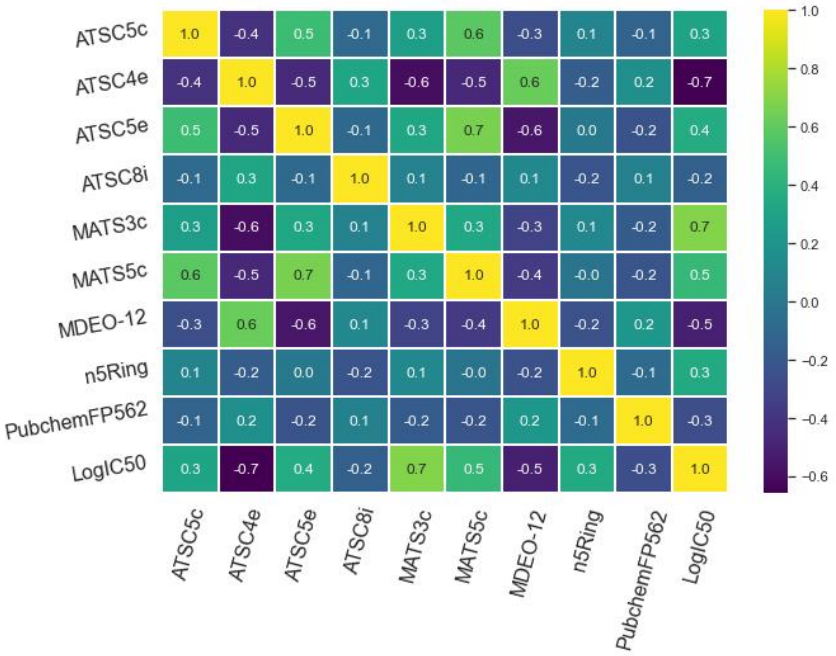


Figure IV.1. Matrice de corrélation des descripteurs moléculaires sélectionnés par l'algorithme Boruta.

IV.3. Augmentation des données et impact sur la distribution des activités

Afin d'améliorer la robustesse du modèle et d'atténuer le risque de surapprentissage, une stratégie d'augmentation des données a été mise en œuvre, conduisant à un ensemble de données étendu de 612 points de données. L'algorithme SMOGN a été utilisé pour générer des

exemples synthétiques tout en préservant les caractéristiques statistiques de l'ensemble de données initial. La Figure IV. 2 illustre l'impact de l'augmentation des données sur la distribution des valeurs de $LogIC_{50}$, qui représentent l'activité anticancéreuse des dérivés de benzofurane.

L'augmentation des données a permis d'élargir la gamme des valeurs de $LogIC_{50}$ présentes dans l'ensemble de données. On observe une distribution plus étendue vers les valeurs les plus faibles de $LogIC_{50}$, correspondant aux composés les plus actifs. Cela est particulièrement bénéfique pour améliorer la capacité du modèle à prédire l'activité des composés très actifs.

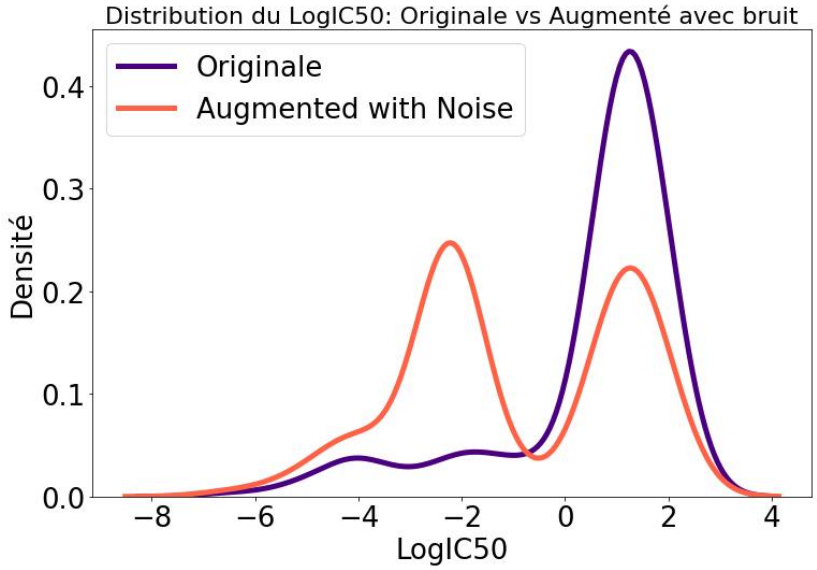


Figure IV. 2. Distribution des $LogIC_{50}$ avant et après l'augmentation des données avec l'algorithme SMOGN.

L'algorithme SMOGN a contribué à lisser la distribution des $LogIC_{50}$, réduisant ainsi les pics et les vallées observés dans la distribution originale. Cela est dû à la génération de nouveaux exemples synthétiques dans les régions de l'espace des descripteurs où les données étaient rares. Un lissage de la distribution peut améliorer la stabilité du modèle et sa capacité à généraliser à de nouvelles données.

Bien que la distribution ait été lissée et la gamme d'activités élargie, les caractéristiques générales de la distribution originale ont été préservées. On observe toujours deux pics principaux, indiquant la présence de deux groupes de composés avec des activités différentes. Cela suggère que l'algorithme SMOGN a réussi à générer des exemples synthétiques qui reflètent les tendances générales de l'ensemble de données original.

IV. 4. Optimisation des hyperparamètres par l'algorithme ALO

L'optimisation des hyperparamètres est une étape essentielle pour maximiser la performance d'un modèle de machine learning. Dans cette étude, l'algorithme ALO (Ant Lion Optimizer) a été utilisé pour rechercher les hyperparamètres optimaux de l'algorithme des Forêts Aléatoires (Random Forest) pour la modélisation de l'activité anticancéreuse des dérivés de benzofurane contre les cellules MCF-7. L'algorithme ALO a permis de déterminer les valeurs optimales suivantes pour les hyperparamètres du modèle Random Forest : n_estimators=299, max_depth=20, min_samples_split=2, et min_samples_leaf=1.

La Figure IV. 3 illustre l'évolution de la fonction fitness (exprimée ici comme une valeur négative) et du coefficient de détermination (R^2) au cours des 50 itérations de l'algorithme ALO. La fonction fitness représente la performance du modèle Random Forest sur l'ensemble de validation, avec des valeurs plus négatives indiquant une meilleure performance. Le R^2 quantifie la proportion de la variance de l'activité anticancéreuse expliquée par le modèle, avec des valeurs proches de 1 indiquant un bon ajustement.

On observe que la fonction fitness diminue rapidement au cours des premières itérations, indiquant une amélioration rapide de la performance du modèle. Après quelques itérations, la fonction fitness se stabilise, suggérant que l'algorithme ALO a convergé vers une région de l'espace des hyperparamètres où les améliorations sont plus marginales. Simultanément, le R^2 augmente rapidement au début des itérations, puis se stabilise également, confirmant l'amélioration de l'ajustement du modèle aux données.

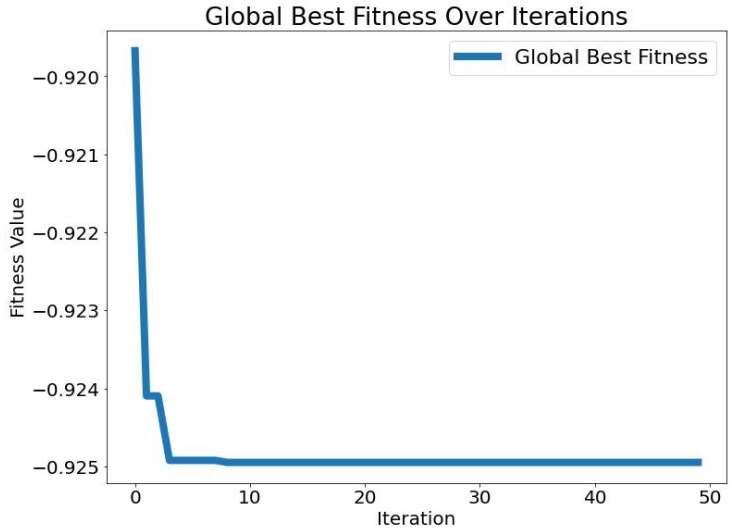


Figure IV.3. Convergence de la fonction fitness et du coefficient de détermination (R^2) au Cours des itérations de l'algorithme ALO.

La Figure IV. 4 présente l'évolution des phases d'exploration et d'exploitation de l'algorithme ALO au cours des 50 itérations. L'exploration fait référence à la capacité de l'algorithme à explorer de nouvelles régions de l'espace des hyperparamètres, tandis que l'exploitation se concentre sur l'amélioration des solutions existantes dans les régions prometteuses.

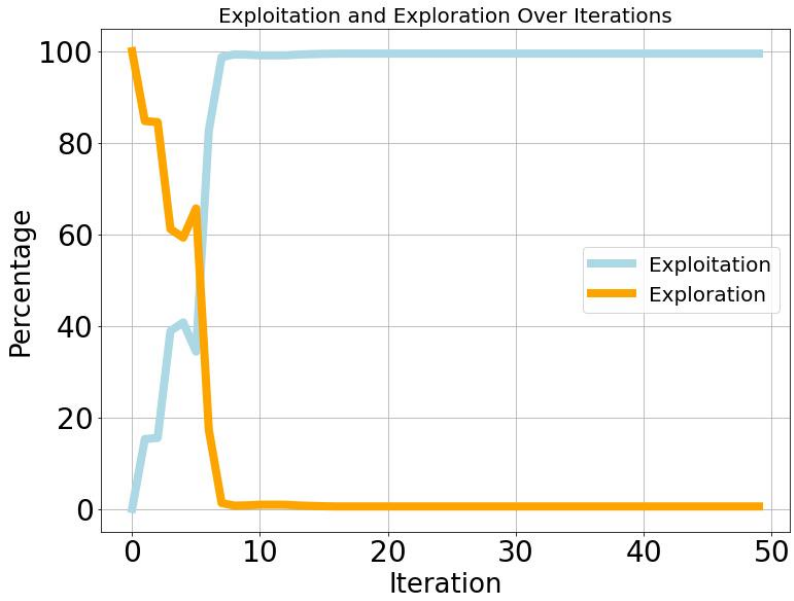


Figure IV.4. Évolution de l'exploration et de l'exploitation au cours des itérations de l'algorithme ALO.

Au début des itérations, l'algorithme ALO se concentre principalement sur l'exploration, avec un pourcentage élevé d'exploration et un faible pourcentage d'exploitation. Cela permet à l'algorithme de découvrir des régions potentiellement intéressantes de l'espace des hyperparamètres. Au fur et à mesure que l'algorithme progresse, la phase d'exploitation prend de l'importance, tandis que l'exploration diminue. Cela indique que l'algorithme se concentre sur l'amélioration des solutions trouvées dans les régions prometteuses.

L'équilibre entre l'exploration et l'exploitation est essentiel pour la performance de l'algorithme ALO. Une exploration excessive peut conduire à une perte de temps dans des régions non prometteuses, tandis qu'une exploitation excessive peut empêcher l'algorithme de découvrir des solutions optimales dans d'autres régions de l'espace des hyperparamètres. La Figure IV. 4 montre que l'algorithme ALO a réussi à trouver un bon équilibre entre l'exploration et l'exploitation, convergeant vers une solution optimale en un nombre raisonnable d'itérations.

IV. 5. Évaluation de la performance du modèle QSAR

L'évaluation de la performance du modèle QSAR pour la prédiction de l'activité anticancéreuse, exprimée en $LogIC_{50}$, a été menée en utilisant des ensembles d'entraînement et de test, et quantifiée par plusieurs métriques statistiques.

La Figure IV. 5 présente un graphe de régression visualisant la relation entre les valeurs prédites et réelles de $LogIC_{50}$. La ligne diagonale (en cyan) indique la prédiction idéale (où la prédiction égale la valeur réelle), et la bande rose représente un intervalle de confiance.

L'observation du graphique révèle que :

- Une forte corrélation positive existe entre les valeurs réelles et prédites de $LogIC_{50}$ pour les deux ensembles de données.
- Les points d'entraînement (bleus) sont très proches de la ligne de régression idéale, indiquant une bonne performance du modèle sur les données sur lesquelles il a été calibré.
- Les points de test (oranges) suivent également la tendance générale et sont relativement proches de la ligne, bien qu'avec un peu plus de dispersion par rapport aux données d'entraînement.
- Quelques points de test (oranges) s'écartent plus significativement de la ligne, suggérant des erreurs de prédiction plus importantes pour ces cas.

Les performances quantitatives du modèle sont les suivantes :

Performance sur l'ensemble d'entraînement :

- R^2 = 0,9902 : Cette valeur très élevée indique que le modèle explique 99,02% de la variance des données d'entraînement. C'est une excellente performance, suggérant un très bon ajustement du modèle à ces données.
- MAE = 0,1223 : L'erreur absolue moyenne est très faible, ce qui signifie que les prédictions s'écartent en moyenne de seulement 0,1223 unité $LogIC_{50}$ des valeurs réelles.
- MSE = 0,0543 : L'erreur quadratique moyenne est également très faible, indiquant de petites erreurs de prédiction et une absence de grandes erreurs aberrantes.

- RMSE = 0,2330 : La racine carrée de l'erreur quadratique moyenne, qui est dans la même unité que la variable prédite, confirme la faible ampleur des erreurs.
- MAPE = 12,0062% : Le pourcentage moyen d'erreur absolue est relativement bas, soulignant des prédictions précises et cohérentes.

Performance sur l'ensemble de test :

- R^2 = 0,9239 : Bien que légèrement inférieur à celui de l'entraînement, un R^2 de 0,9239 reste excellent. Il indique que le modèle explique 92,39% de la variance des données de test, démontrant une très bonne capacité à généraliser.
- MAE = 0,3678 : L'erreur absolue moyenne est plus élevée que pour l'entraînement, mais reste raisonnable. Les prédictions s'écartent en moyenne de 0,3678 unité $LogIC_{50}$ des valeurs réelles sur les données non vues.
- MSE = 0,4101 : L'erreur quadratique moyenne a augmenté par rapport à l'entraînement, mais reste acceptable, suggérant que les erreurs sont encore gérables.
- RMSE = 0,6404 : L'erreur quadratique moyenne racine est plus élevée, mais dans l'ensemble des valeurs de $LogIC_{50}$, elle indique que les erreurs de prédiction sont raisonnables.
- MAPE = 31,2620% : Le pourcentage moyen d'erreur absolue est plus élevé pour les données de test, suggérant que, bien que les prédictions soient globalement bonnes, certaines peuvent avoir une erreur relative plus importante.

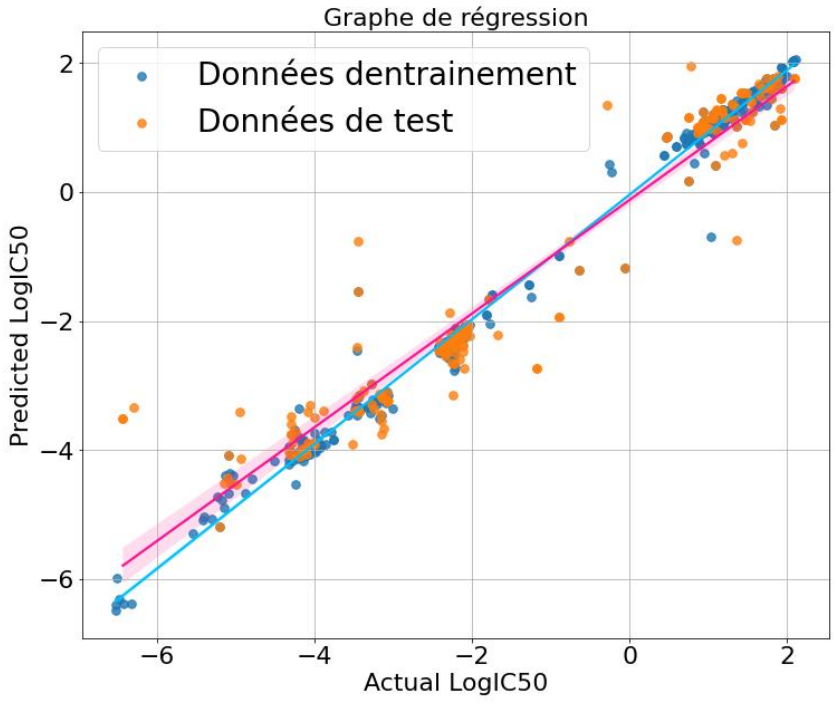


Figure IV.5. Graphe de régression montrant les activités prédites par rapport aux activités réelles pour les ensembles d'entraînement et de test.

Tableau IV. 1. Métriques de performance du modèle QSAR sur les ensembles d'entraînement et de test.

Métrique	**Ensemble d'Entraînement**	**Ensemble de Test**
R^2 (coefficient de détermination)	0.9902	0.9239
MAE (erreur absolue moyenne)	0.1223	0.3678
MSE (erreur quadratique moyenne)	0.0543	0.4101
RMSE (racine carrée de la MSE)	0.233	0.6404
MAPE (erreur absolue moyenne en %)	12.00	31.26

Les résultats démontrent que le modèle QSAR est très performant pour prédire les valeurs de $LogIC_{50}$.

La performance exceptionnelle sur l'ensemble d'entraînement indique que le modèle a très bien appris les motifs présents dans les données. Plus important encore, la solide performance sur l'ensemble de test (R^2 = 0,9239) prouve que le modèle a une excellente capacité à généraliser et à faire des prédictions fiables sur de nouvelles molécules qu'il n'a jamais rencontrées.

Bien qu'il y ait une légère baisse de précision sur les données de test par rapport à l'entraînement (ce qui est normal et souhaitable pour éviter le surapprentissage), l'écart reste modéré. Le graphe de régression confirme visuellement cette bonne concordance entre les valeurs prédites et réelles, pour les deux ensembles de données.

En conclusion, le modèle QSAR est un outil efficace et robuste pour la prédiction de l'activité anticancéreuse.

IV.6. Évaluation du Domaine d'Applicabilité (DA)

L'évaluation du domaine d'applicabilité est une étape cruciale pour juger de la fiabilité des prédictions d'un modèle QSAR. La méthode du Williams Plot, basée sur l'analyse des leverages et des résidus standardisés, permet de visualiser et de quantifier le DA du modèle développé. Le Williams Plot (Figure IV. 6) représente les résidus standardisés en fonction des valeurs de leverage pour chaque composé de l'ensemble d'entraînement. Le leverage, une mesure de l'influence d'un composé sur le modèle, est comparé à un seuil de leverage, calculé ici à 0.0631. Les composés situés à droite de cette ligne verticale sont considérés comme ayant un leverage élevé et pourraient influencer de manière disproportionnée les prédictions du modèle. Les lignes horizontales, situées à ±2 résidus standardisés, délimitent les valeurs aberrantes dans les prédictions du modèle. Les composés situés au-dessus ou en dessous de ces lignes sont considérés comme mal prédits par le modèle.

L'analyse des résultats révèle que sur les 416 composés de l'ensemble d'entraînement, 404 se situent à l'intérieur du DA, tandis que 12 sont considérés comme étant à l'extérieur. Sur les 184 composés de l'ensemble de test, 158 se situent à l'intérieur du DA, tandis que 26 sont considérés comme étant à l'extérieur. Ces résultats indiquent que la majorité des composés, tant dans l'ensemble d'entraînement que dans l'ensemble de test, se situent à l'intérieur du domaine d'applicabilité du modèle. Cela suggère que le modèle est globalement fiable et capable de fournir des prédictions précises pour la plupart des composés étudiés.

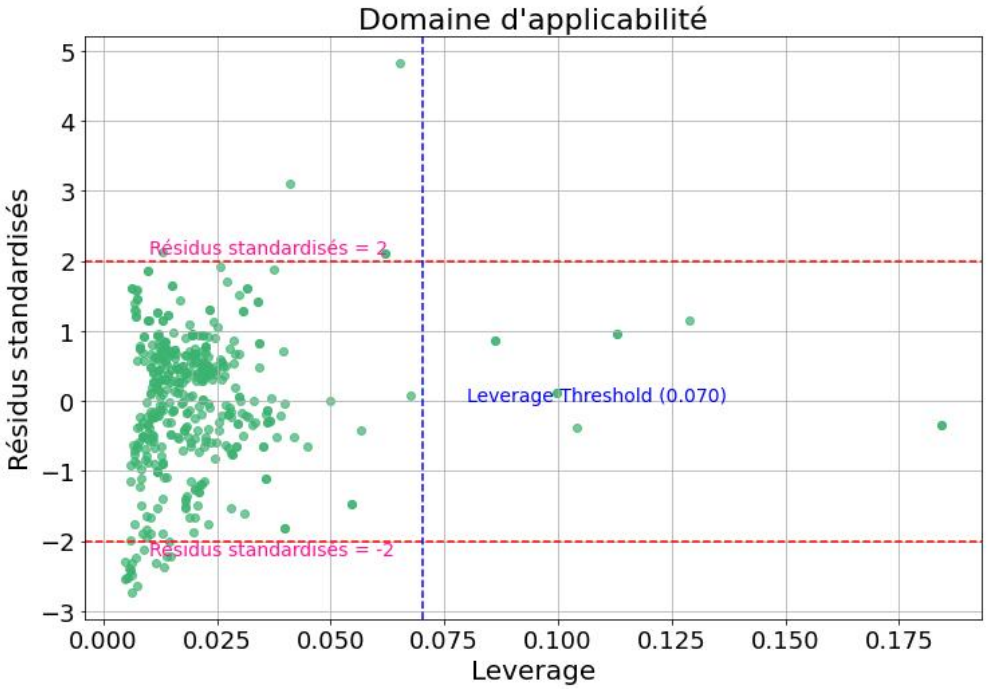


Figure IV.6. Le graphe de williams visualisant le domaine d'applicabilité du modèle random forest.

IV.7. Interface graphique pour la prédiction de l'activité anticancéreuse

L'interface graphique (GUI) développée pour la prédiction de l'activité anticancéreuse des dérivés de benzofurane (Figure IV. 7) représente un outil précieux pour faciliter la recherche et la conception de nouvelles molécules potentiellement actives. En fournissant une interface conviviale et intuitive, elle permet aux chercheurs, même sans expertise approfondie en modélisation moléculaire, d'exploiter la puissance du modèle QSAR pour orienter leurs efforts de synthèse et d'évaluation.

L'interface propose des champs de saisie pour les neuf descripteurs moléculaires sélectionnés par l'algorithme Boruta comme étant les plus pertinents pour la prédiction de l'activité anticancéreuse : ATSC5c, ATSC4e, ATSC5e, ATSC8i, MATS3c, MATS5c, MDEO-12, n5Ring et PubchemFP562. L'utilisateur peut entrer les valeurs de ces descripteurs pour une molécule donnée. Le résultat de la prédiction, exprimé en IC_{50} (µM), est affiché dans une boîte de dialogue. L'utilisation d'une boîte de dialogue permet de présenter clairement le résultat à l'utilisateur.

En permettant de prédire l'activité anticancéreuse des molécules avant leur synthèse et leur évaluation expérimentale, l'interface permet de réduire les coûts et le temps de développement de nouveaux médicaments.

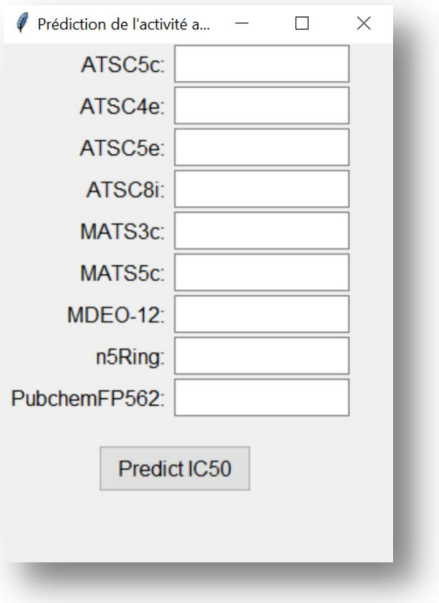


Figure IV.7. Interface Graphique pour la prédiction de l'activité anticancéreuse des dérivés de benzofurane.

Conclusion Générale

Conclusion générale

Le présent travail s'inscrit dans une démarche scientifique visant à contribuer à la lutte contre le cancer du sein, maladie responsable de centaines de milliers de décès dans le monde chaque année. Face à l'émergence de résistances aux traitements conventionnels et à la nécessité d'identifier de nouvelles cibles thérapeutiques, l'étude des dérivés de benzofurane s'est révélée particulièrement pertinente en raison de leurs propriétés pharmacologiques diversifiées et prometteuses. Ces composés hétérocycliques ont démontré, à travers différentes études, leur capacité à inhiber plusieurs mécanismes moléculaires impliqués dans la prolifération tumorale, tels que les tyrosine kinases, les topoisomérases, ou encore la polymérisation de la tubuline.

Dans cette optique, une base de données composée de 198 dérivés de benzofurane a été constituée à partir de 13 publications scientifiques. L'activité anticancéreuse de ces composés, mesurée en termes de $\log IC_{50}$ sur la lignée cellulaire MCF-7, a servi de variable cible pour la modélisation. Les descripteurs moléculaires et empreintes numériques ont été calculés à l'aide du logiciel PaDEL-Descriptor, générant un ensemble riche en information structurale et physicochimique. Un prétraitement rigoureux a permis de nettoyer et normaliser les données, garantissant la qualité des entrées utilisées pour la modélisation.

Le processus de sélection des variables a été mené à l'aide de l'algorithme Boruta, permettant de retenir uniquement les descripteurs les plus pertinents, réduisant ainsi la dimensionnalité du jeu de données et limitant le surapprentissage. L'algorithme Boruta a permis de sélectionner les descripteurs suivants : ATSC5c, ATSC4e, ATSC5e, ATSC8i, MATS3c, MATS5c, MDEO-12, n5Ring et PubchemFP562. Ces descripteurs représentent une combinaison de différentes classes d'informations moléculaires, mettant en évidence l'importance des interactions internes à la molécule et rôle de la géométrie globale pour l'activité biologique. Pour la phase de modélisation, plusieurs algorithmes ont été testés, avec une attention particulière portée au modèle Random Forest, reconnu pour sa robustesse et sa capacité à gérer des données complexes. L'optimisation des hyperparamètres de ce modèle a été effectuée à l'aide de l'algorithme ALO, un algorithme bio-inspiré efficace dans la recherche globale de solutions optimales.

Afin d'améliorer la robustesse du modèle et de corriger le déséquilibre présent dans la distribution des valeurs de $\log IC_{50}$, une phase d'augmentation des données a été mise en

œuvre à l'aide de l'algorithme SMOGN. Cet algorithme est spécifiquement conçu pour les problèmes de régression comportant des valeurs rares ou extrêmes, souvent sous-représentées dans les jeux de données expérimentaux. En générant artificiellement de nouveaux points de données autour de ces zones peu denses, SMOGN a permis d'augmenter la base initiale de 198 composés à 612 instances, tout en préservant la variabilité et la cohérence des propriétés structurales. Cette extension a significativement renforcé la capacité d'apprentissage du modèle, en particulier dans les zones critiques de l'espace de réponse.

Le modèle final, construit à l'aide de la méthode Random Forest et optimisé par l'algorithme bio-inspiré ALO, a été validé par des techniques de validation croisée. Il a affiché des performances élevées : $R^2 = 0.92$, RMSE = 0.64, MAE = 0.36 et MAPE = 31.26% sur l'ensemble de test. Le modèle Random Forest obtenu repose sur les hyperparamètres optimaux suivants : n_estimators=299, max_depth=20, min_samples_split=2, et min_samples_leaf=1. Cette configuration permet une combinaison optimale entre complexité et généralisation du modèle. Ces résultats traduisent une excellente capacité prédictive et une généralisation satisfaisante. Par ailleurs, l'analyse du domaine d'applicabilité a confirmé que le modèle couvre efficacement la majorité des composés étudiés. Une interface utilisateur intuitive a également été développée, facilitant la prédiction de l'activité pour de nouveaux dérivés.

En conclusion, ce mémoire confirme l'intérêt des approches in silico basées sur l'intelligence artificielle pour la modélisation QSAR de composés bioactifs. Il ouvre la voie à la découverte rationnelle de nouveaux dérivés de benzofurane actifs contre le cancer du sein, tout en fournissant un cadre méthodologique reproductible, combinant sélection de variables, apprentissage automatique et visualisation des résultats. Ce travail peut servir de base à des recherches futures intégrant davantage de données biologiques, de simulations moléculaires ou d'évaluations expérimentales pour le développement de nouvelles thérapies ciblées.

Références bibliographiques

[1] World Health Organization (WHO), “Breast cancer.”

[2] S. Bahrami, F. Shafiei, A. Marjani, and T. M. Isfahani, “Comparative QSAR modeling of 2-phenylindol derivatives for predicting the anticancer activity using genetic algorithm multiple linear regression and back-propagation-artificial neural network techniques,” *Int. J. Quantum Chem.*, vol. 123, no. 6, p. e27057, Mar. 2023, doi: https://doi.org/10.1002/qua.27057.

[3] J. Farhat, L. Alzyoud, M. Alwahsh, and B. Al-omari, “Structure – Activity Relationship of Benzofuran Derivatives with Potential Anticancer Activity,” pp. 1–22, 2022.

[4] H. Makhloufi, “Agents apoptotiques issus de la flore endolichénique pour la lutte contre les cancers chimio-résistants To cite this version : HAL Id : tel-04810629 Thèse de doctorat,” 2024.

[5] A. Rambur, I. De, R. A. S. Mapk, and P. I. K. Akt, “Importance de la co-dérégulation des voies RAS / MAPK et PI3K / AKT / mTOR dans la transformation épithéliale prostatique . Approche in vivo à l ’ aide d ’ un modèle dans les glandes accessoires de la Drosophile Amandine Rambur To cite this version : HAL ,” 2020.

[6] K. Gabora, P. Andra, B. Iulian Claudiu, L. Maria Iulia, S. Ioan-Adrian, and D. and Piciu, “Current evidence on thyroid related adverse events in patients treated with protein tyrosine kinase inhibitors,” *Drug Metab. Rev.*, vol. 51, no. 4, pp. 562–569, Oct. 2019, doi: 10.1080/03602532.2019.1687512.

[7] L. Arce-ramos and J. Castillo, “Synthesis and Biological Studies of Benzo [b] furan Derivatives : A Review from 2011 to 2022,” 2023.

[8] N. ZERIGUI, M. Lounes, and S. Madjid, “Rôle de l’éducation thérapeutique dans l’amélioration de la qualité de vie chez les personnes atteintes de maladies chroniques « cas de cancer du sein unité de BELLOUA, CHU Tizi-Ouzou » Réalisé,” 2020.

[9] N. Radia, K. Nedjma, B. Oumaima, and H. Sabah, “La prédisposition Génétique du gène BRCA1 au Cancer du sein Présenté,” 2022.

[10] R. Cabrillac, L. O. Présentation, and O. Indien, “L ’ OHADA : Présentation générale To cite this version : HAL Id : hal-02544455 L ’ O H A D A : p r é s e n t a t i o n g é n é r a l e 1,” pp. 9–15, 2020.

[11] P. J. Van Diest, E. Van Der Wall, and J. P. A. Baak, “Prognostic value of proliferation in invasive breast cancer: A review,” *J. Clin. Pathol.*, vol. 57, no. 7, pp. 675–681, 2004, doi: 10.1136/jcp.2003.010777.

[12] R. W. Carlson *et al.*, “Invasive Breast Cancer Clinical Practice Guidelines in Oncology,” vol. 9, no. 2, pp. 136–222, 2011.

[13] A. Alaref, A. Hassan, R. S. Kandel, R. Mishra, and J. Gautam, “Magnetic Resonance Imaging Features in Different Types of Invasive Breast Cancer : A Systematic Review of the Literature,” vol. 13, no. 3, pp. 1–10, 2021, doi: 10.7759/cureus.13854.

[14] J. Adv, “March 2010 Revised: 19,” vol. 1, no. 2, pp. 109–126, 2010.

[15] W. Cao and D. Liu, “Non-Invasive Biomarkers for Early Detection of Breast Cancer,” pp. 1–28, 2020.

[16] M. Bazin, “anti-infectieuse et antiproliférative To cite this version : HAL Id : tel-03732173 Habilitation à Diriger des Recherches Marc-Antoine Bazin Conception de biomolécules hétérocycliques à visées,” 2022.

[17] J. Haussmann *et al.*, “Recent advances in radiotherapy of breast cancer,” *Radiat. Oncol.*, vol. 15, no. 1, p. 71, 2020, doi: 10.1186/s13014-020-01501-x.

[18] S. Rivera, G. Louvel, E. R. Del Campo, A. Boros, H. Oueslati, and É. Deutsch, “Radiothérapie axillaire prophylactique des cancers du sein,” *Cancer / Radiother.*, pp. 1–8, 2015, doi: 10.1016/j.canrad.2015.05.001.

[19] M. S. U. Hassan, J. Ansari, D. Spooner, and S. A. Hussain, “Chemotherapy for breast cancer (Review),” pp. 1121–1131, 2010, doi: 10.3892/or.

[20] I. H. Abdulkareem and I. B. Zurmi, “Review of hormonal treatment of breast cancer,” vol. 15, no. 1, pp. 9–14, 2012, doi: 10.4103/1119-3077.94088.

[21] U. D. E. Limoges, “Thèse,” 2005.

[22] O. J. Scully, B.-H. Bay, G. Yip, and Y. Yu, “Breast cancer metastasis,” *Cancer Genomics Proteomics*, vol. 9, no. 5, pp. 311–320, 2012.

[23] F. Rossi and F. Rossi, “Vers une valorisation industrielle d ’ un remède traditionnel pour le traitement des intoxications ciguatériques To cite this version : Vers une valorisation industrielle d ’ un remède traditionnel pour le traitement des intoxications ciguatériques,” 2014.

[24] C. Alleman, “pour la synthèse d ’ analogues simplifiés en vue d ’ étudier les interactions protéine-protéine To cite this version : HAL Id : tel-04531886,” 2024.

[25] J. Xu, Z.; Zhao, S.; Lv, Z.; Feng, L.; Wang, Y.; Zhang, F.; Bai, L.; Deng, “Benzofuran Derivatives and Their Anti-Tubercular, Anti-Bacterial Activities,” vol. 162, pp. 266–276, 2019.

[26] K. M. Dawood, “An update on benzofuran inhibitors: a patent review. Expert Opinion on Therapeutic Patents,” vol. 29, no. 11, pp. 841–870, 2019.

[27] P. V. . Martins, P.; Jesus, J.; Santos, S.; Raposo, L.R.; Roma-Rodrigues, C.; Baptista and A. . Fernandes, “Heterocyclic Anticancer Compounds: Recent Advances and the Paradigm Shift towards the Use of Nanomedicine“s Tool Box,” vol. 20, pp. 16852–16891, 2015.

[28] R. Kim, T. Kin, and W. T. Beck, “Impact of Complex Apoptotic Signaling Pathways on Cancer Cell Sensitivity to Therapy,” 2024. doi: 10.3390/cancers16050984.

[29] F. Uhlmann, F. Lottspeich, and K. Nasmyth, “Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesin subunit Scc1,” pp. 37–42, 1999.

[30] O. A. El-khouly, M. A. Henen, M. A. El-sayed, M. I. Shabaan, and S. M. El-messery, “Bioorganic & Medicinal Chemistry Synthesis , anticancer and antimicrobial evaluation of new benzofuran based derivatives : PI3K inhibition , quorum sensing and

molecular modeling study," *Bioorg. Med. Chem.*, vol. 31, no. May 2019, p. 115976, 2021, doi: 10.1016/j.bmc.2020.115976.

[31] M. Kouider Amar, H. Moussa, and M. Hentabli, "Predicting the anticancer activity of indole derivatives: A novel GP-tree-based QSAR model optimized by ALO with insights from molecular docking and decision-making methods," *Comput. Biol. Med.*, vol. 189, p. 109988, 2025, doi: https://doi.org/10.1016/j.compbiomed.2025.109988.

[32] Q. Li *et al.*, "Synthesis and biological evaluation of benzofuran-based 3 , 4 , 5-trimethoxybenzamide derivatives as novel tubulin polymerization," *Bioorg. Chem.*, p. 104076, 2020, doi: 10.1016/j.bioorg.2020.104076.

[33] R. Kaur, G. Kaur, R. K. Gill, R. Soni, and J. Bariwal, "Recent developments in tubulin polymerization inhibitors: An overview," *Eur. J. Med. Chem.*, vol. 87, pp. 89–124, 2014, doi: https://doi.org/10.1016/j.ejmech.2014.09.051.

[34] A. S. Al Wasidi, A. S. Hassan, and A. M. Naglah, "In vitro cytotoxicity and druglikeness of pyrazolines and pyridines bearing benzofuran moiety," *J. Appl. Pharm. Sci.*, vol. 10, no. 4, pp. 142–148, 2020, doi: 10.7324/JAPS.2020.104018.

[35] C. Rania, Z. Ikram, A. Ouidad, D. Ouahiba, and C. Rabah, "Développement d'un modèle de machine learning (apprentissage automatique) pour la classification automatique des organismes microscopiques," 2024.

[36] Z. Islem, "Par :L'intelligence artificielle principe, outils et objectifs. Domaine," 2019.

[37] C. Khadija, "L ' usage de l ' intelligence artificielle dans l ' enseignement - apprentissage du FLE L ' usage de l ' intelligence artificielle dans l ' ensei gnement-apprentissage du FLE," 2025.

[38] L. D. E. Matteis *et al.*, "Introduction à l ' apprentissage automatique," pp. 1–18, 2022.

[39] "Chapitre 2," pp. 4–18.

[40] M. Houcine, "Vers une amélioration de la détection d'intrusion par les méthodes de sélection des fonctionnalités à l'aide des arbres de décision," 2021.

[41] HILLAIRET Caroline ;MODRY Jean ;BESSEYRE Fabien;MEYER Laurie, "COMPARAISON DES METHODES CLASSIQUES ET ALTERNATIVES AVEC LE MACHINE LEARNING POUR LA CONSTRUCTION D'UNE TABLE DE MORTALITE D'EXPERIENCE BEST ESTIMATE," vol. 106, 2019.

[42] B. Mohamed, S. Djillali, B. M. Lamine, M. B. Amina, and M. Abderahim, "Extraction des connaissances à partir d ' une base de Données (Application à la Détection de Fraude dans la Consommation d ' E lectricité et du Gaz)," 2020.

[43] S. Yacine, T. djoudi Salim, Charfa, and Z. F. Zohra, "Extraction de motifs basée sur word2vec Réalisé," 2020.

[44] M. GUENDOUZ, M. ROUMEISSA, B. Abderrahmane, and B. Abdelbasset, "Une approche QSAR basée sur Deep Learning pour la sélection des descripteurs Explicatifs.," 2021.

[45] U. Kose, "TECHNIQUES FOR ADVERSARIAL EXAMPLES THREATENING THE

SAFETY OF TECHNIQUES FOR ADVERSARIAL EXAMPLES THREATENING THE SAFETY OF," no. September, 2019.

[46] C. Aida, S.-M. Labiba, and H. BAHI, "Approche hybride pour l ' apprentissage automatique incrémental par," 2020.

[47] J. ALEJANDRO and L. GONZALEZ, "EXPLORATION DES ARBRES DE DÉCISION ET DES SUPPORT VECTOR MACHINES EN VUE D' APPLICATIONS DANS L'ANALYSE DE TEXTE," 2016.

[48] F. Stefani and G. Rosa, "Clasificación de huevos de animales ovíparos mediante algoritmos de machine learning: Una Revisión Sistemática," 2025.

[49] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[50] S. Alnemari and M. Alshammari, "Detecting Phishing Domains Using Machine Learning," 2023.

[51] H. Rania, D. Akila, K. Noha, and B. Sarah, "Thème :Amélioration des forêts aléatoires pour la classification des données médicales Présenté," 2021.

[52] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision Forests for Classification , Regression , Density Estimation , Manifold Learning and Semi-Supervised Learning," 2011.

[53] Z. Dong, H. Chen, Y. Yang, and H. Hao, "Research on the optimization model of anti-breast cancer candidate drugs based on machine learning," *Front. Genet.*, vol. Volume 16, 2025.

[54] K. Szafrański, J. Sławiński, Ł. Tomorowicz, and A. Kawiak, "Synthesis, Anticancer Evaluation and Structure-Activity Analysis of Novel (E)- 5-(2-Arylvinyl)-1,3,4-oxadiazol-2-yl)benzenesulfonamides," 2020. doi: 10.3390/ijms21062235.

[55] Ł. Tomorowicz, J. Sławiński, B. Żołnowska, K. Szafrański, and A. Kawiak, "Synthesis, Antitumor Evaluation, Molecular Modeling and Quantitative Structure–Activity Relationship (QSAR) of Novel 2-[(4-Amino-6-N-substituted-1,3,5-triazin-2-yl)methylthio]-4-chloro-5-methyl-N-(1H-benzo[d]imidazol-2(3H)-ylidene)Benzenesulfonamides," 2020. doi: 10.3390/ijms21082924.

[56] V. T. Angelova *et al.*, "Novel Arylsulfonylhydrazones as Breast Anticancer Agents Discovered by Quantitative Structure-Activity Relationships," 2023. doi: 10.3390/molecules28052058.

[57] M. O. Sarhan, S. S. Abd El-Karim, M. M. Anwar, R. H. Gouda, W. A. Zaghary, and M. A. Khedr, "Discovery of New Coumarin-Based Lead with Potential Anticancer, CDK4 Inhibition and Selective Radiotheranostic Effect: Synthesis, 2D & 3D QSAR, Molecular Dynamics, In Vitro Cytotoxicity, Radioiodination, and Biodistribution Studies," 2021. doi: 10.3390/molecules26082273.

[58] C. O. Salas *et al.*, "Promising 2,6,9-Trisubstituted Purine Derivatives for Anticancer Compounds: Synthesis, 3D-QSAR, and Preliminary Biological Assays," 2020. doi: 10.3390/ijms21010161.

[59] R. Verma, S. Verma, and N. Abbas, "Development of Benzimidazole Anticancer Leads for EGFR Inhibition by 3D-QSAR Based Virtual Screening, Molecular Docking, Molecular Simulation and Drug Likeness Study," *Lett. Drug Des. & Discov.*, vol. 21, no. 19, pp. 4833–4844, 2024, doi: https://doi.org/10.2174/0115701808344738241231072218.

[60] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *J. Comput. Chem.*, vol. 32, no. 7, pp. 1466–1474, May 2011, doi: https://doi.org/10.1002/jcc.21707.

[61] C. W. E. I. Yap, "Software News and Update PaDEL-Descriptor : An Open Source Software to Calculate Molecular Descriptors and Fingerprints," 2010, doi: 10.1002/jcc.

[62] S. Fares *et al.*, "Design , synthesis , and analysis of antiproliferative and apoptosis-inducing activities of nitrile derivatives containing a benzofuran scaffold : EGFR inhibition assay and molecular modelling study," *J. Enzyme Inhib. Med. Chem.*, vol. 36, no. 1, pp. 1488–1499, 2021, doi: 10.1080/14756366.2021.1946044.

[63] S. Fares *et al.*, "Design, synthesis, and analysis of antiproliferative and apoptosis-inducing activities of nitrile derivatives containing a benzofuran scaffold: EGFR inhibition assay and molecular modelling study," *J. Enzyme Inhib. Med. Chem.*, vol. 36, no. 1, pp. 1487–1498, Jan. 2021, doi: 10.1080/14756366.2021.1946044.

[64] R. Anantacharya, M. K., S. N.D., S. S., and M. Y. and Kaviraj, "Antiproliferative, DNA cleavage, and ADMET study of substituted 2-(1-benzofuran-2-yl) quinoline-4-carboxylic acid and its esters," *Cogent Chem.*, vol. 2, no. 1, p. 1158382, Dec. 2016, doi: 10.1080/23312009.2016.1158382.

[65] N. F. El-Sayed, M. El-Hussieny, E. F. Ewies, M. F. El Shehry, H. M. Awad, and M. A. Fouad, "Design, synthesis, biological evaluation, and molecular docking of new benzofuran and indole derivatives as tubulin polymerization inhibitors," *Drug Dev. Res.*, vol. 83, no. 2, pp. 485–500, Apr. 2022, doi: https://doi.org/10.1002/ddr.21880.

[66] M. Sherif, S. Tawfik, F. E.-N. Goda, and L. Abou-Zeid, "Synthesis and molecular modelling studies of some 2-pyrazolyl benzofuran derivatives as anti-tumor agents," *Delta Univ. Sci. J.*, vol. 5, no. 2, pp. 187–200, 2022, doi: 10.21608/dusj.2022.281400.

[67] O. A. El-Khouly, M. A. Henen, M. A.-A. El-Sayed, and S. M. El-Messery, "Design, synthesis and computational study of new benzofuran hybrids as dual PI3K/VEGFR2 inhibitors targeting cancer," *Sci. Rep.*, vol. 12, no. 1, p. 17104, 2022, doi: 10.1038/s41598-022-21277-2.

[68] O. M. Abdelhafez, K. M. Amin, H. I. Ali, M. M. Abdalla, and E. Y. Ahmed, "Design, synthesis and anticancer activity of benzofuran derivatives targeting VEGFR-2 tyrosine kinase," *RSC Adv.*, vol. 4, no. 23, pp. 11569–11579, 2014, doi: 10.1039/C4RA00943F.

[69] W. M. Eldehna *et al.*, "Identification of 3-(piperazinylmethyl)benzofuran derivatives as novel type II CDK2 inhibitors: design, synthesis, biological evaluation, and in silico insights," *J. Enzyme Inhib. Med. Chem.*, vol. 37, no. 1, pp. 1227–1240, Dec. 2022, doi: 10.1080/14756366.2022.2062337.

[70] M. J. Mphahlele, M. M. Maluleka, N. Parbhoo, and S. T. Malindisa, "Synthesis, Evaluation for Cytotoxicity and Molecular Docking Studies of Benzo[c]furan-

Chalcones for Potential to Inhibit Tubulin Polymerization and/or EGFR-Tyrosine Kinase Phosphorylation," 2018. doi: 10.3390/ijms19092552.

[71] K. M. Amin, Y. M. Syam, M. M. Anwar, H. I. Ali, T. M. Abdel-Ghani, and A. M. Serry, "Synthesis and molecular docking study of new benzofuran and furo[3,2-g]chromone-based cytotoxic agents against breast cancer and p38α MAP kinase inhibitors," *Bioorg. Chem.*, vol. 76, pp. 487–500, 2018, doi: https://doi.org/10.1016/j.bioorg.2017.12.029.

[72] S. M. S. Atta, D. S. Farrag, A. M. K. Sweed, and A. H. Abdel-Rahman, "Preparation of new polycyclic compounds derived from benzofurans and furochromones. An approach to novel 1,2,3-thia-, and selena-diazolofurochromones of anticipated antitumor activities," *Eur. J. Med. Chem.*, vol. 45, no. 11, pp. 4920–4927, 2010, doi: https://doi.org/10.1016/j.ejmech.2010.07.065.

[73] M. Hranjec *et al.*, "Antiproliferative potency of novel benzofuran-2-carboxamides on tumour cell lines: Cell death mechanisms and determination of crystal structure," *Eur. J. Med. Chem.*, vol. 59, pp. 111–119, 2013, doi: https://doi.org/10.1016/j.ejmech.2012.11.009.

[74] D. Sharmistha, P. M., S. K. R.R., and H. and Ashoka, "Enhanced in silico QSAR-based screening of butyrylcholinesterase inhibitors using multi-feature selection and machine learning," *SAR QSAR Environ. Res.*, vol. 36, no. 2, pp. 79–99, Feb. 2025, doi: 10.1080/1062936X.2025.2466020.

[75] A. Kryshchyshyn, O. Devinyak, D. Kaminskyy, P. Grellier, and R. Lesyk, "Development of Predictive QSAR Models of 4-Thiazolidinones Antitrypanosomal Activity Using Modern Machine Learning Algorithms," *Mol. Inform.*, vol. 37, no. 5, p. 1700078, May 2018, doi: https://doi.org/10.1002/minf.201700078.

[76] X. Hu, X. Zhou, H. Liu, H. Song, S. Wang, and H. Zhang, "Enhanced predictive modeling of hot rolling work roll wear using TCN - LSTM - Attention," *Int. J. Adv. Manuf. Technol.*, no. February, 2024, doi: 10.1007/s00170-024-13105-w.

[77] T. Öznacar and T. Güler, "Prediction of Early Diagnosis in Ovarian Cancer Patients Using Machine Learning Approaches with Boruta and Advanced Feature Selection," 2025. doi: 10.3390/life15040594.

[78] R. Zhou *et al.*, "JoCoRank: Joint correlation learning with ranking similarity regularization for imbalanced fetal brain age regression," *Comput. Biol. Med.*, vol. 171, p. 108111, 2024, doi: https://doi.org/10.1016/j.compbiomed.2024.108111.

[79] I. Cortes-Ciriano and A. Bender, "Improved Chemical Structure–Activity Modeling Through Data Augmentation," *J. Chem. Inf. Model.*, vol. 55, no. 12, pp. 2682–2692, Dec. 2015, doi: 10.1021/acs.jcim.5b00570.

[80] A. P. Lind and P. C. Anderson, "Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties," *PLoS One*, vol. 14, no. 7, p. e0219774, Jul. 2019.

[81] W. Arbia, M. Kouider Amar, L. Adour, and A. Amrane, "Maximizing chitin and chitosan recovery yields from fusarium verticillioides using a many-factors-at-a-time approach," *Int. J. Biol. Macromol.*, p. 136708, 2024, doi:

https://doi.org/10.1016/j.ijbiomac.2024.136708.

[82] A. Karampuri and S. Perugu, “A breast cancer-specific combinational QSAR model development using machine learning and deep learning approaches,” *Front. Bioinforma.*, vol. Volume 3-, 2024, doi: 10.3389/fbinf.2023.1328262.

[83] A. B. Umar, A. Uzairu, S. Uba, and G. A. Shallangwa, “QSAR and Docking Studies on Some Potential Anti-Cancer Agents to Predict their Effect on M14 Melanoma Cell Line,” *Chem. Africa*, vol. 3, no. 4, pp. 1009–1022, 2020, doi: 10.1007/s42250-020-00185-w.

Annexe

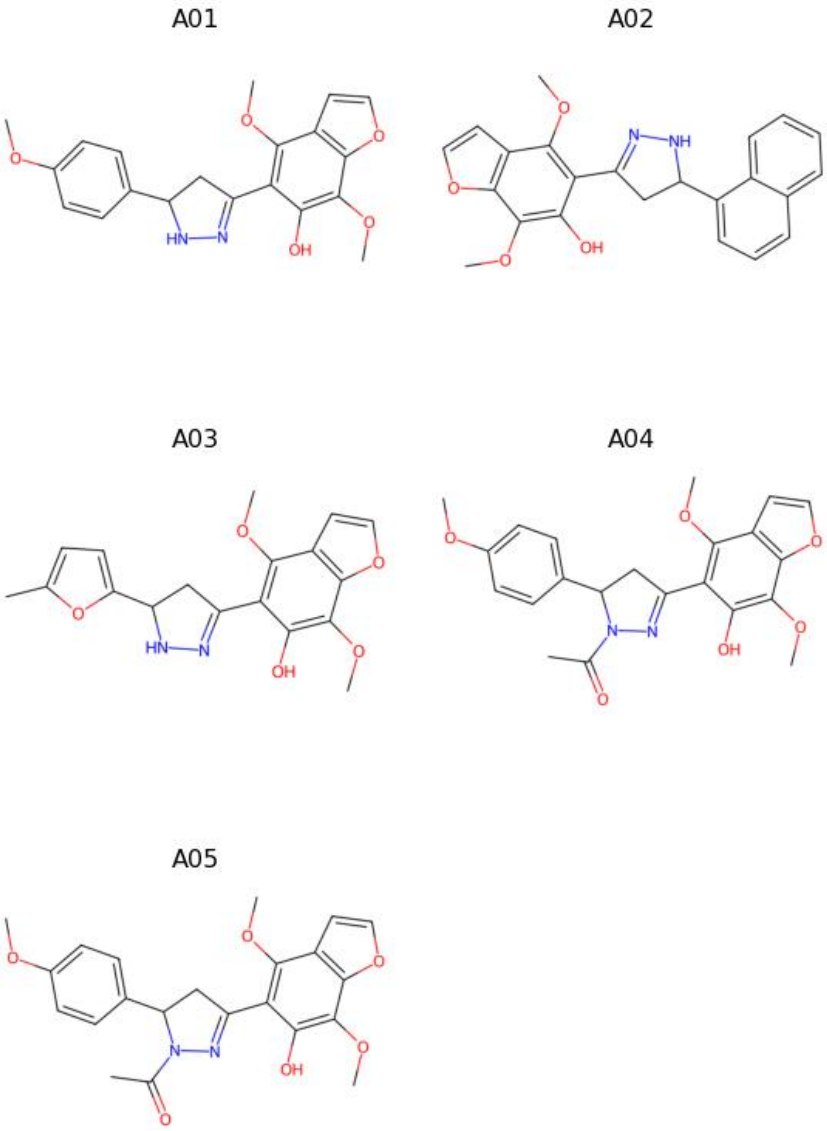


Figure.A1. Les dérivés de pyrazolines et pyridines portant un fragment benzofurane [62].

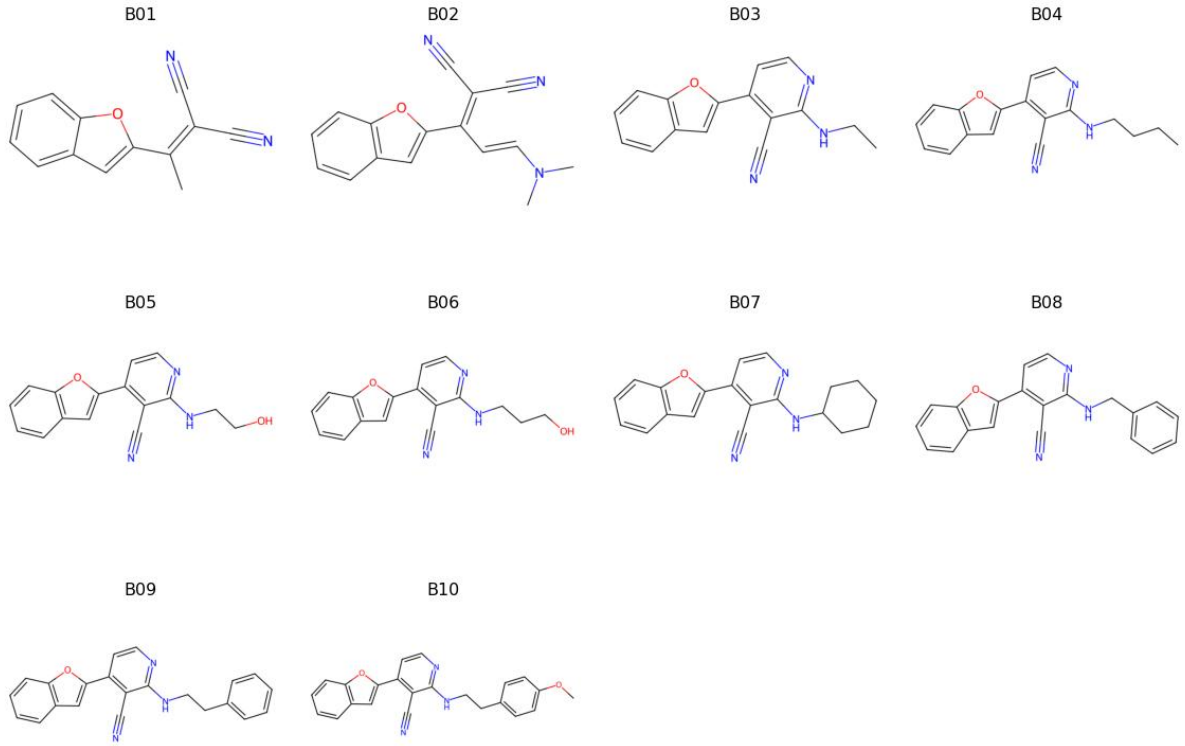


Figure.A2. Les dérivés nitriles contenant un échafaudage benzofurane [63].

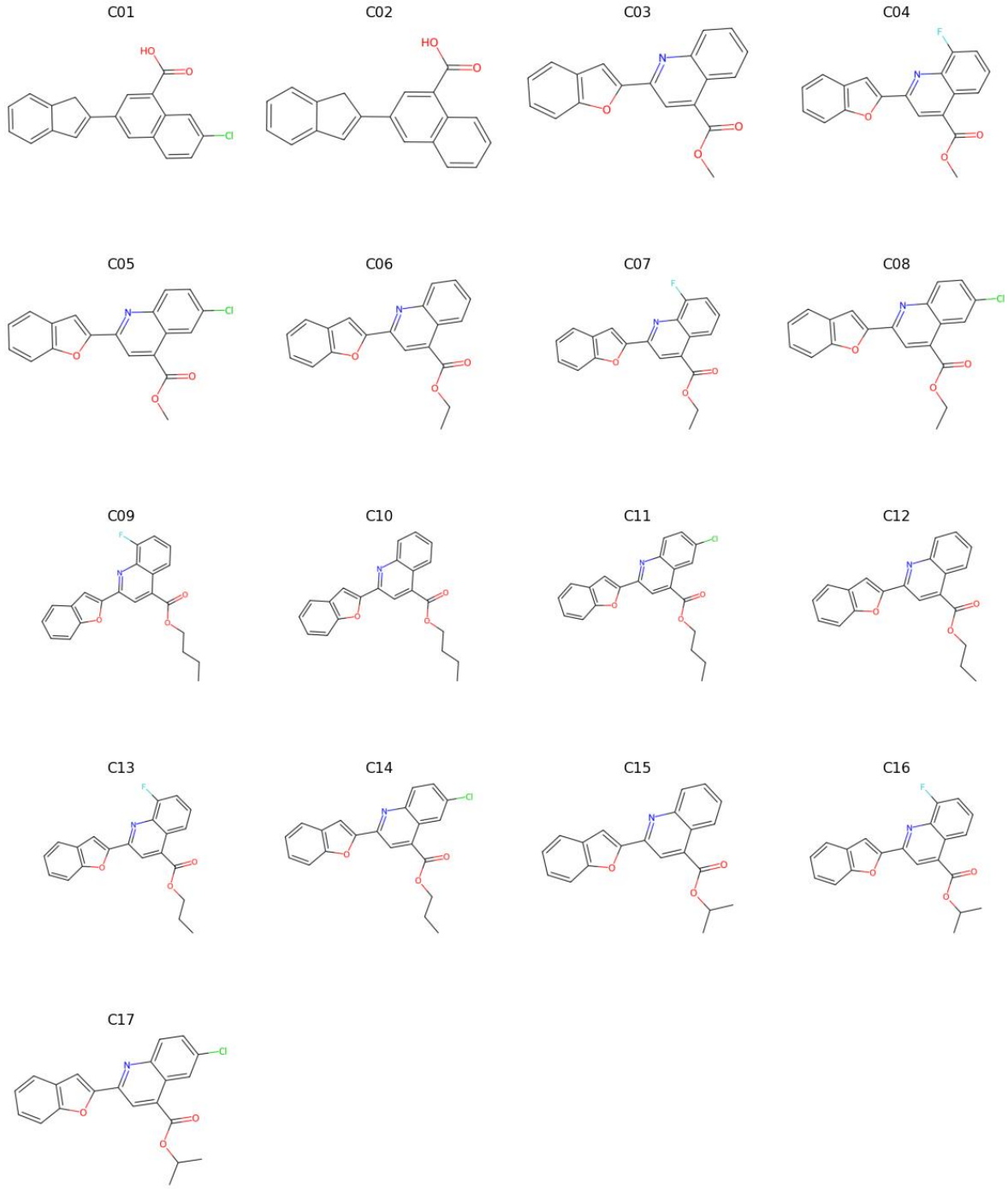


Figure.A3. Les dérivés d'acide 2-(1-benzofuran-2-yl) quinoléine-4-carboxylique substitué et de ses esters [64].

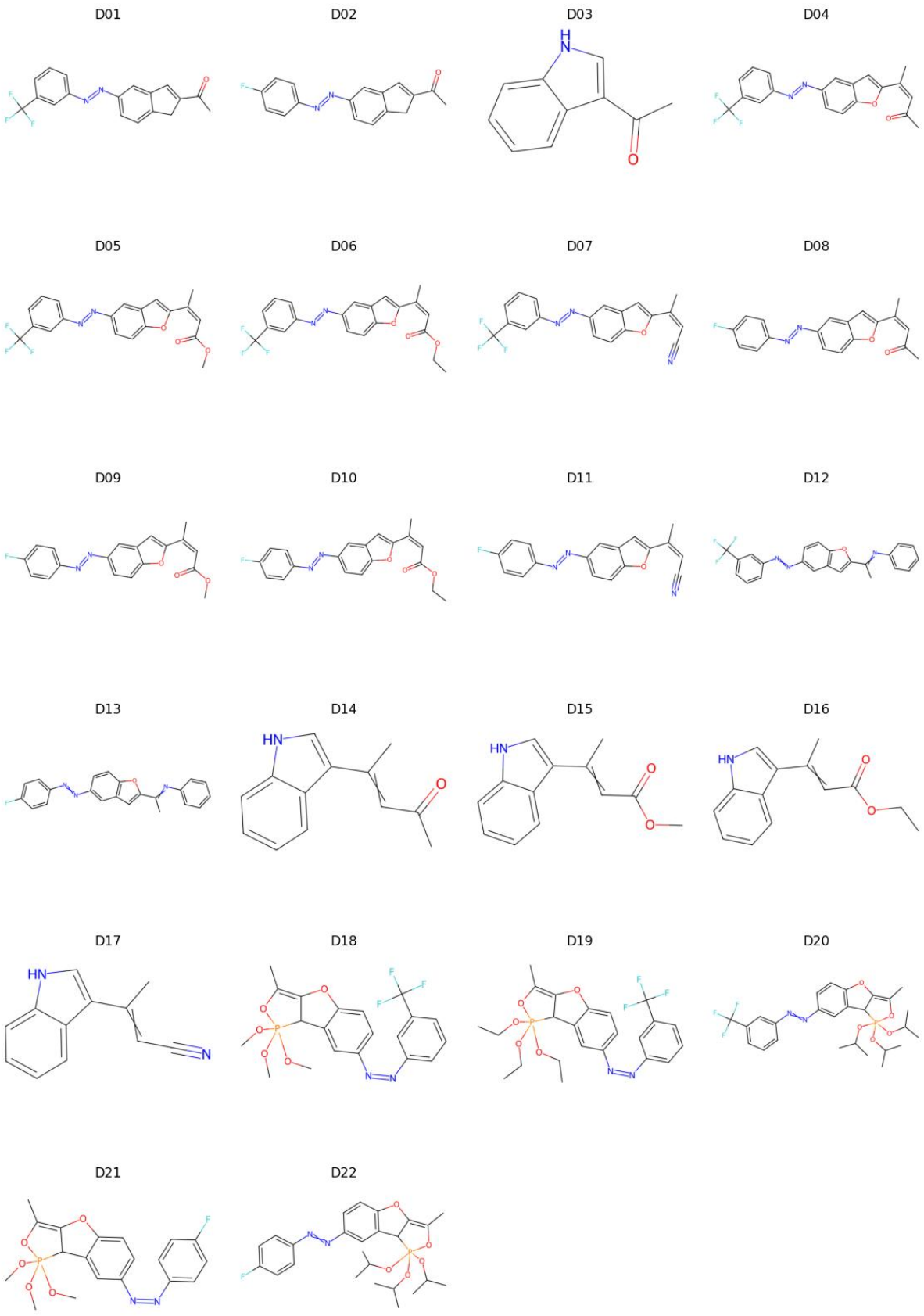


Figure.A4. Dérivés de benzofurane et d'indole [65].

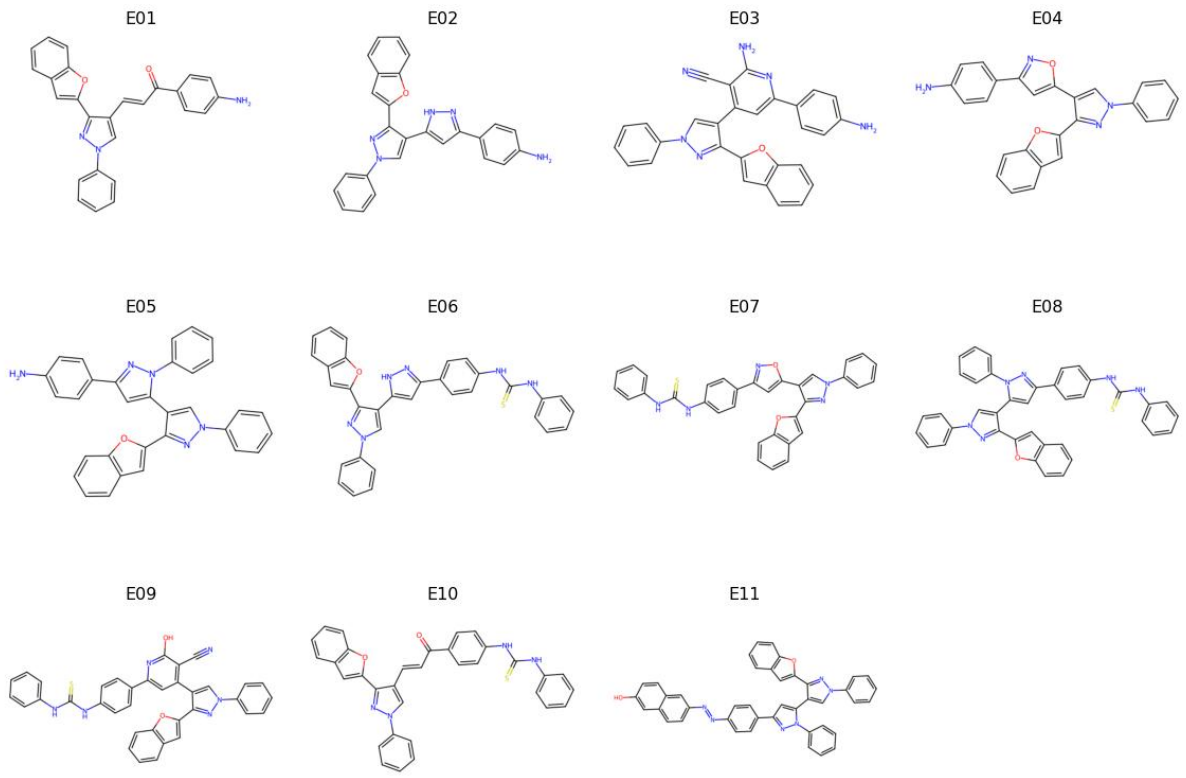


Figure.A5. Les dérivés de 2-pyrazolyl benzofurane [66].

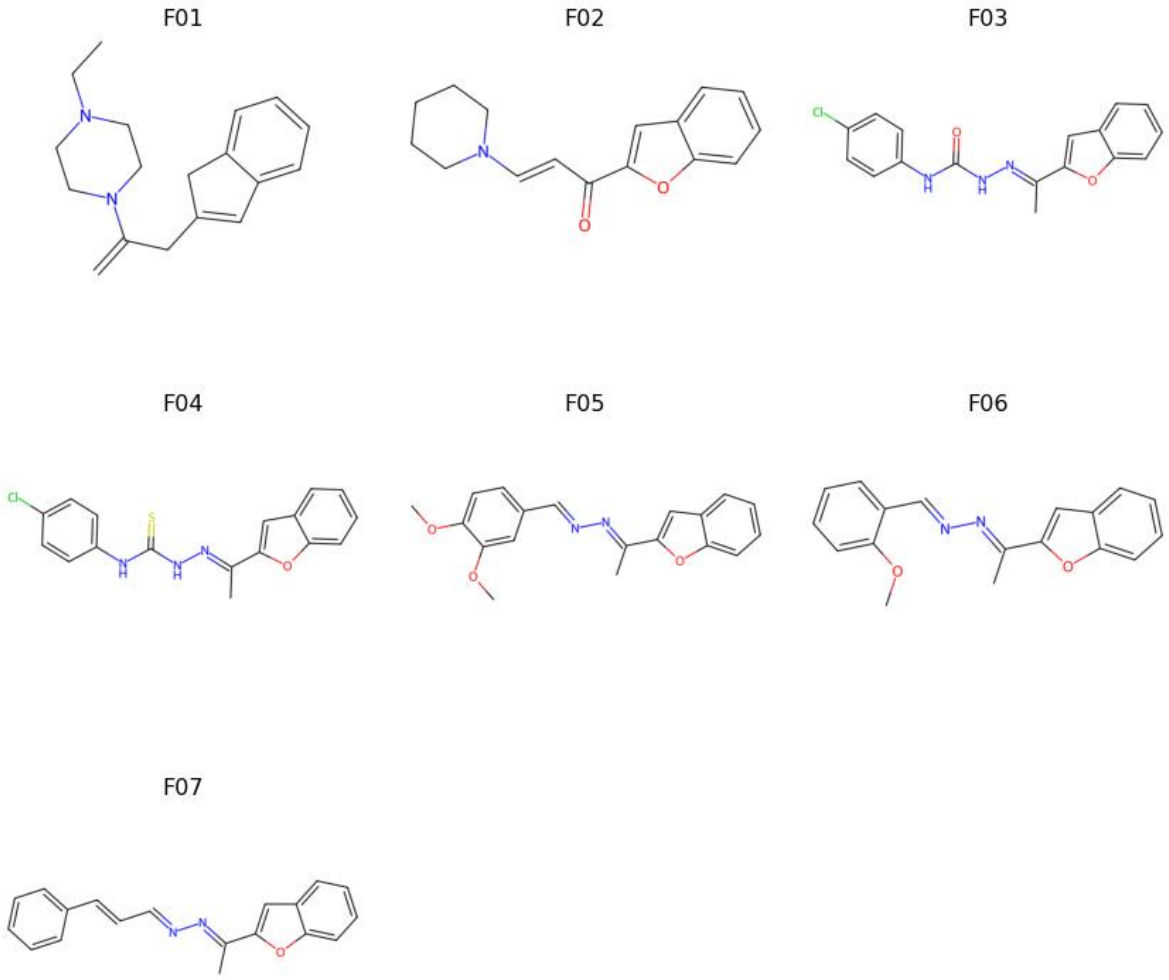


Figure.A6. Hybrides de benzofurane [67].

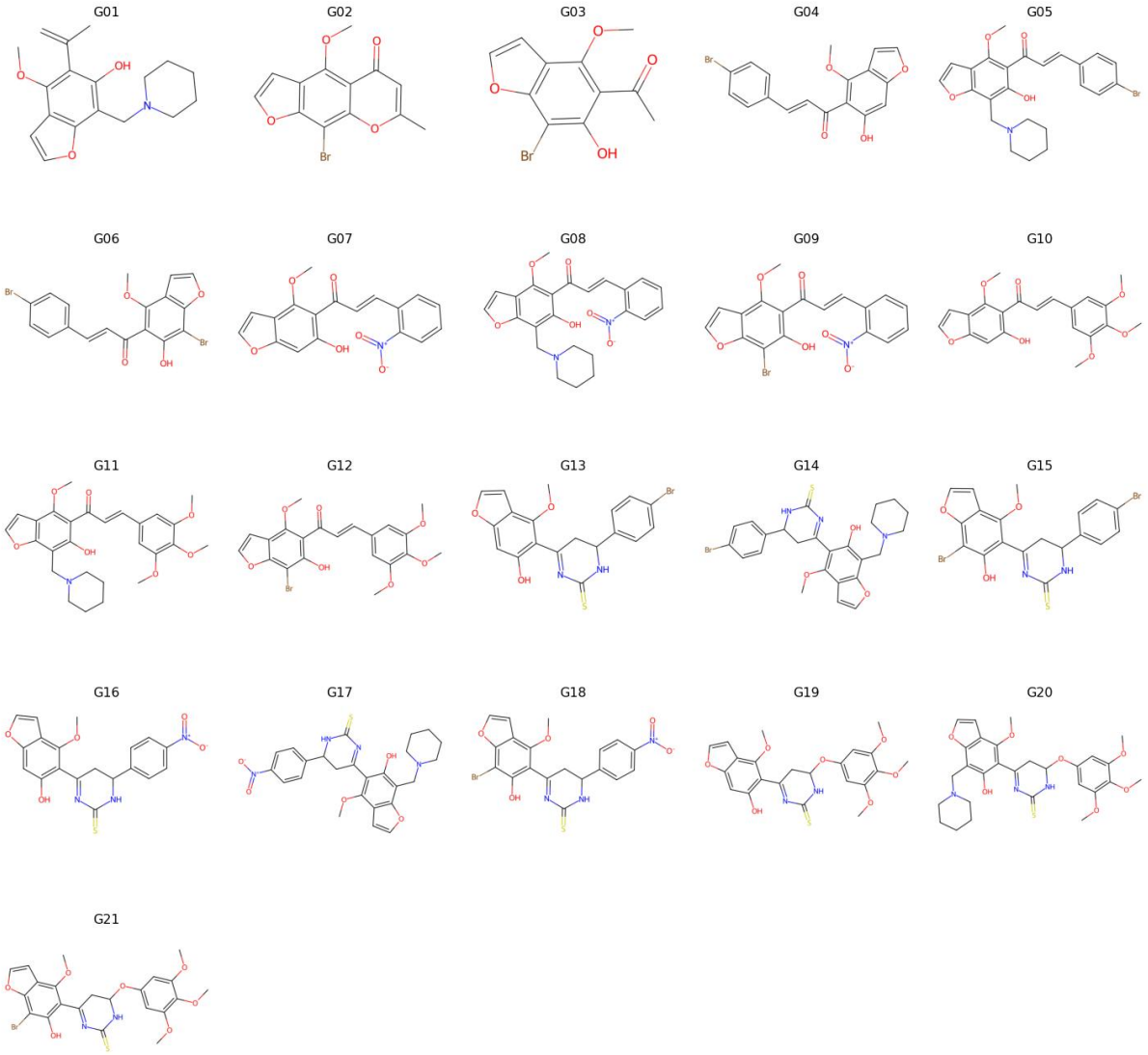


Figure.A7. Dérivés chalcone et thiopyrimidine benzofurane [68].

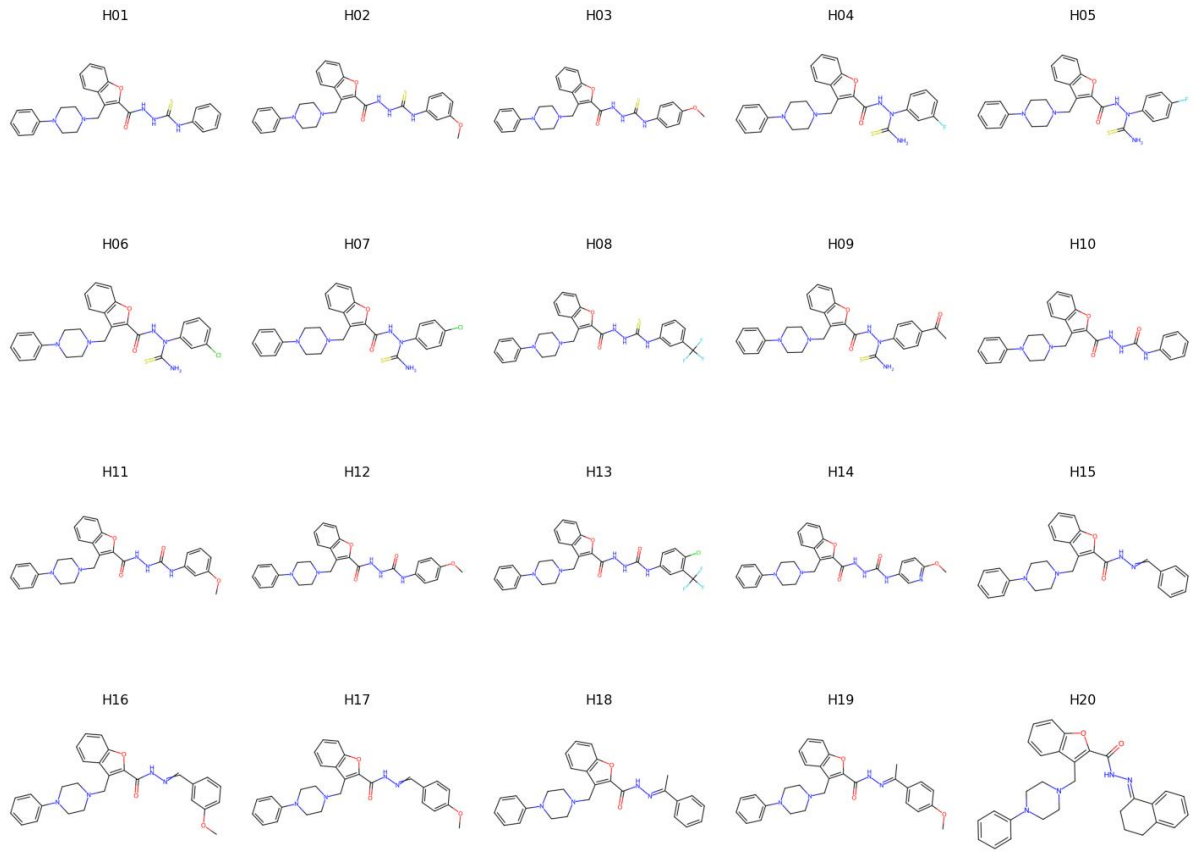


Figure.A8. Des dérivés de 3-(pipérazinylméthyl)benzofurane [69].

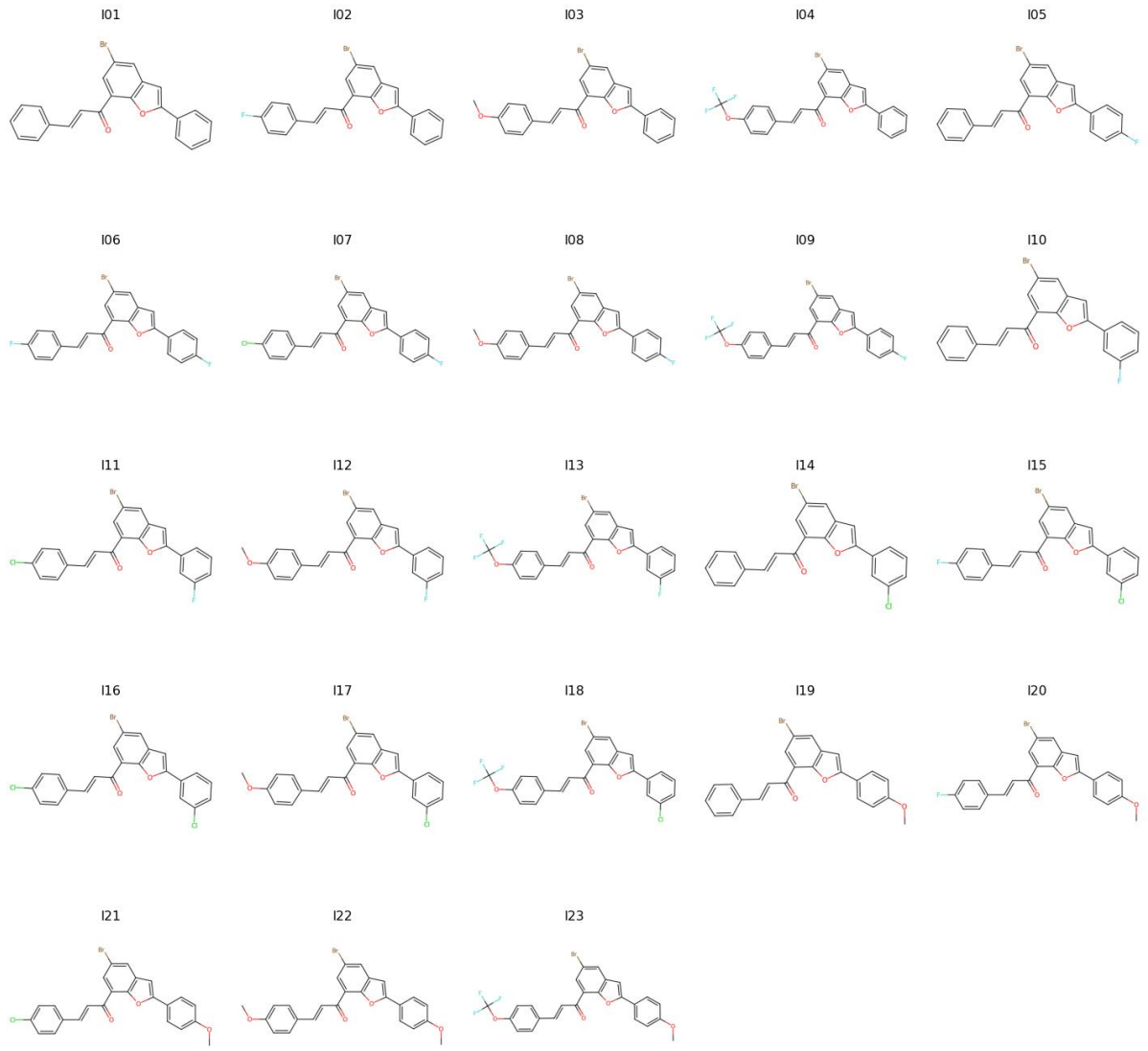


Figure.A9. Dérivés de benzo[*c*]furan-chalcones [70].

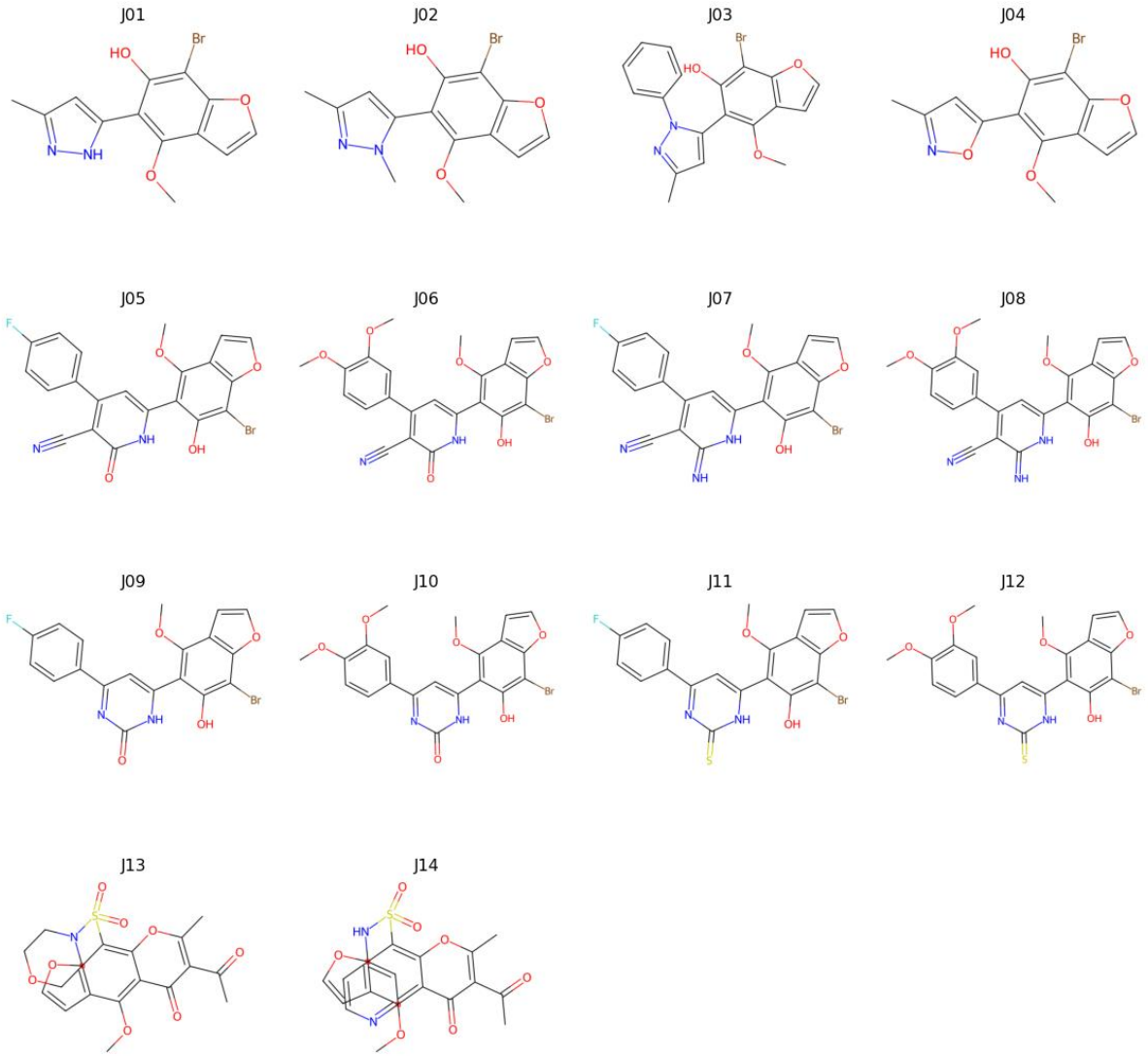


Figure.A10. Benzofuranes et fonctions hétérocycliques diverses liées au 5*H*-furo[3,2-*g*]chromone [71].

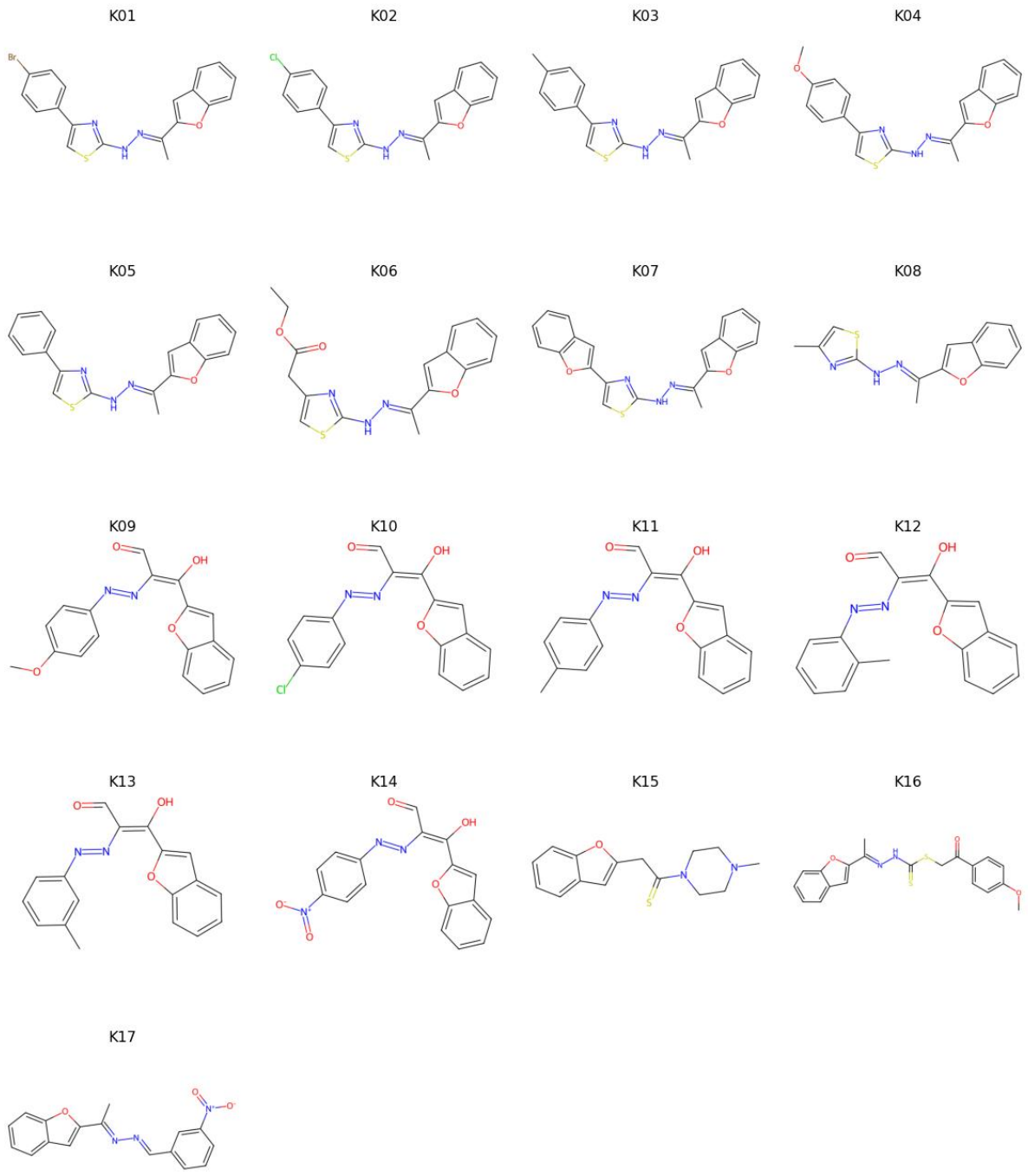


Figure.A11. Hybrides de benzofurane-thiazole, benzofurane-azo, benzofurane-hydrazo et benzofurane-pipérazine [30].

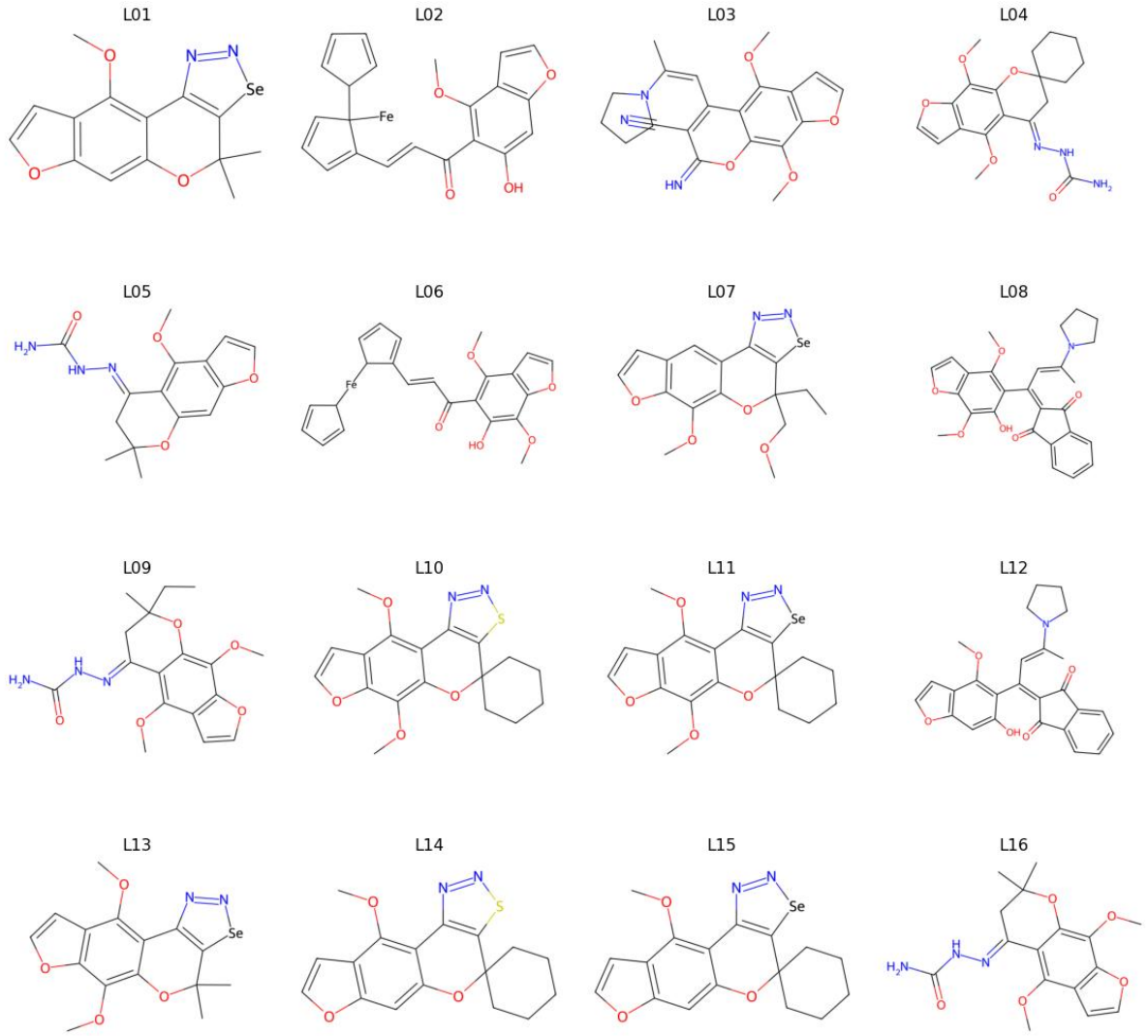


Figure.A12. Composés polycycliques dérivés des benzofuranes et des furochromones [72].

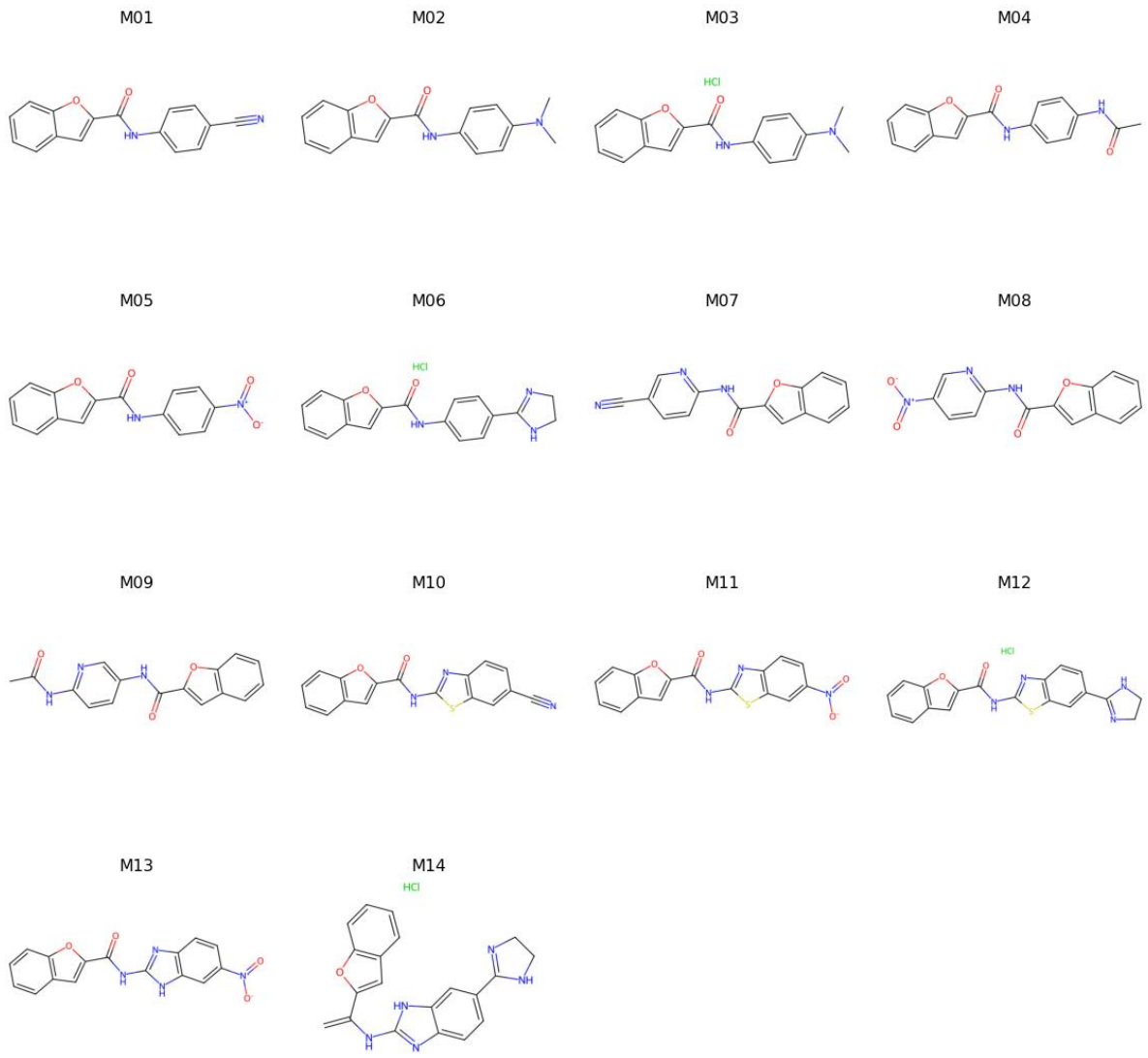


Figure.A13. Dérivés de benzofurane-2-carboxamides [73].