Université Mohamed Khider de Biskra
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Département des Sciences de la Matière

MÉMOIRE DE MASTER

Domaine des Sciences de la Matière

Filière de Chimie

Spécialité Chimie Pharmaceutique

Présenté et soutenu par :

BARI Zoulikha **DJAHRA Khadidja**

Le : Samedi 19 Septembre2020

Etude et développement des techniques QSAR des molécules Anti Alzheimer

---

Jury:

Dr. BELAIDI Salah	Prof.	Université Med Khider –Biskra	Président
Dr. KHAMOULI Saida	MCB	Université Med Khider –Biskra	Rapporteur
Dr. MELLAOUI Malika	MCB	Université Med Khider –Biskra	Examinateur

Année universitaire : 2019 -2020

Remerciements :

*Avant tous, Nous remercions "**ALLAH**" le tout puissant d'avoir données la force , la volonté et la courage pour réaliser ce travail .*

*Nous remercions, d'abord, nos encadreur M^me^ **Khamouli Saida** pour ses efforts, ses conseils successifs, son indulgence ; pour son suivi, l'intérêt qu'elle a portée a ce travail.*

Nous tenons aussi à remercier les membres de jury le président

*Mr **Belaidi Salah** et L'examinateur M^me^ **Mellaoui Malika** d'accepter de juger notre travail .*

Nous avons , tiens à remercier tous nos enseignants du département de chimie sans exception qui ont assurés notre formation.

Que soient, enfin, ALHAMDLI ALLAH, nous terminons ce travail , qui est le résultat de nos efforts.

A Tous , merci

Dédicace :

Je dédie ce modeste travail, Aux deux personnes les plus chéres au monde pour moi, mon père et ma mère .

Que dieu mes les gardes.

A mes chères frères : Idris, Nacer et Tarek.

A mes adorables sœurs : Naima, Samia, Imane et Souad.

A nous petites anges : Khadidja, Amina, Nesrine, Zineb et Abd-elraheman

A tentes tous mes oncles et mes tentes.

A ma belle binôme : Khadidja.

A tous mes chères amies : Samiha , Hanane , Bouchra , Samra

Zoulikha

Dédicace :

Je dédie ce modeste travail, Aux deux personnes les plus chéres au monde pour moi, mon père et ma mère .

Que dieu mes les gardes.

A mes chères frères : Omar, Massaoud, Ossama et Mohamed

A mes adorables sœurs : Siham, Rim et Souad

A nous petites anges : Ayham , Jana , Taj elddin ,Yasmin et Amir

A tentes tous mes oncles et mes tentes.

A ma belle binôme : Zoulikha

A tous mes chères amies :Ikram, Fatiha, Soumia , Chaima et Zaza

Khadidja

Sommaire

Liste de tableaux…… I

Liste des figures……..... II

Liste des principales abréviation……... III

Glossaire de biochimie…… V

Introduction Génerale …… 1

Chapitre I: la maladie d'Alzheimer

I.1 Introduction…… 3

I.2 Géneralite……... 4

I.2.1 Definiton……. 4

I.2.2 Historique……... 4

I.3 Les symptomes……... 5

I.3.1 Pertes de mémoire…….. 5

I.3.2 Difficultés à accomplir les tâches quotidiennes……. 5

I.3.3 Problèmes de langage……. 5

I.3.4Désorientation dans le temps et dans l'espace…… 5

I.3.5 Difficultés dans les raisonnements abstraits……... 5

I.3.6 Perte d'objets…….. 5

I.3.7 Altération du jugement……... 5

I.3.8 Modification du comportement…….. 5

I.3.9 Changement de personnalité…….. 6

I.4 Les phases de la maladie d'Alzheimer……... 6

I.5 Les facteurs de risque de l'Alzheimer…… 7

I.5.1 L'âge……... 7

I.5.2 Les antécédents familiaux et la génétique…….. 7

I.5.3 Le sexe…… 8

I.5.4 Les maladies cardiovasculaires…….... 8

I.5.5 Le diabète.....…… 8

I.5.6 Le syndrome de Down..... 9

I.5.7 La déficience cognitive légère (DCL).....…….. 9

I.5.8 Les lésions à la tête.....……. 9

I.6 Physiopathologie de la maladie d'Alzheimer.....……..... 10

I.6.1 Hypothèse de la cascade β-amyloïde..... 10

I.6.2 Hypothèse de la dégénérescence du cytosquelette neuronal..... 10

I.7 Les kinases.....……. 11

I.7.1 Définition.....…… 11

I.7.2 Classification……..…….. 11

I.8 DYRK1A.....…… 13

I.8.1 Définition.....…… 13

I.8.2 Fonctions.....……. 13

I.8.3 Influence de la protéine kinase DYRK1A sur la maladie d'Alzheimer……. 15

I.9 Diagnostic…… 16

I.10 Le traitement de la maladie d'Alzheimer..……. 17

Chapitre II: QSAR : principe et méthodologie

II.1 Historique..... 18

II.2 Définition.....…….. 19

II.3 Principe..... 19

II.4 Outils et Méthodologie de QSAR……... 20

II.4.1 Paramètres biologiques..…… 20

II.4.2 Descripteurs moléculaires……... 21

II.4.2.1 Types de descripteurs…….. 22

II.4.2.1.a Descripteurs constitutionnels…….. 22

II.4.2.1.b Descripteurs topologiques……. 22

II.4.2.1.c Descripteurs géométriques…….. 23

II.4.2.1.d Descripteurs quantiques/électroniques……... 24

II.4.2.1.e Descripteurs physico-chimiques……. 25

II.4.2.1.f Descripteurs thermodynamiques……. 28

II.4.3 Sélection des descripteurs……... 28

II.4.3.1 La sélection objective…….. 29

II.4.3.2 La sélection subjective…… 29

II.5 La théorie de la fonctionnelle de la densité (DFT)……. 30

II.6 Méthodes statistiques……. 30

II.6.1 Régression linéaire simple (SLR)……... 31

II.6.2 La régression linéaire multiple (MLR)……... 31

II.6.3 La régression non linéaire multiple (MNLR)…… 33

II.6.4 Analyse par composantes principales(ACP)……... 33

II.6.5 La méthode de régression des moindres carrés partiels (PLS)……... 34

II.7 Coefficients et tests statistiques standards…….. 34

II.8 Validation du modèle…….. 40

II.8.1 Validation interne…… 41

II.8.1.1 La procédure leave-n-out…… 41

II.8.1.2 La procédure leave-one-out……. 42

II.8.1.3 Test de randomisation……. 43

II.8.2 Validation externe……... 43

II.9 Domaines d'application…….. 44

II.10 Application de QSAR……... 45

Chapitre III:Etude quantitative des propriétés QSAR d'une série de derivés de 6-arylquinazolin-4-amine

III.1 Introduction……... 47

III.2 Matériels et méthodes……... 48

III.2.1 Ensembles des données……. 48

III.2.2 Sélection des descripteurs et méthodes de calcul…….. 52

III.2.3 Analyse statistique……. 53

III.2.4 Méthodologie générale d'une étude QSAR……... 54

III.3 Résultats et discussions……. 56

III.3.1 Analyse des composants principaux……. 56

III.3.2 Les modèles QSAR……... 59

III.3.2.1 La méthode de régression multilinéaire (MLR)…… 59

III.3.2.2 Régression non linéaire multiple (MNLR)…… 66

III.3.3 Validation des modèles…….. 67

III.3.3.1 Validation interne…….. 67

III.3.3.2 Validation externe…….. 68

III.3.3.3 Y-Randomisation……... 69

III.3.4 Domaine d'applicabilité du modèle QSAR……..... 72

Conclusion générale…….... 78

Références bibliographiques…….. 80

Annexes

Liste des Tableaux :

Tableau I.1: Classification des kinases 12

Tableau II. 1: Types de données biologiques utilisées dans l'analyse QSAR.....21

Tableau II. 2: Les descripteurs géométriques calculés dans étude QSAR 23

Tableau II. 3: Table d'analyse de variance. 38

Tableau III.1: Structure chimique des dérivés de 6-arylquinazolin-4-amine 48

Tableau III.2 : Descripteurs moléculaires utilisées dans l'étude QSAR 55

Tableau III.3: La matrice de corrélation (Pearson (n)) entre les différents descripteurs obtenus 57

Tableau III.4: Modèles sélectionnés et paramètres statistiques des corrélations entre les propriétés moléculaires et l'activité biologique. 60

Tableau III.5: Matrice de corrélation 61

Tableau III.6: Analyse de la variance ANOVA: 63

Tableau III.7: Tableau des coefficients 64

Tableau III.8: Valeurs des critères VIF et TF pour les descripteurs significatifs 65

Tableau III.9: Les valeurs de validation croisée..... 67

Tableau III.10: Critères de Tropsha 68

Tableau III.11: les 50 premières itérations de Y-randomisation 70

Tableau III.12: Random Models Parameters..... 71

Tableau III.13 : Valeurs résiduelles normalisées et valeurs de levier MLR et MNLR : 74

Tableau III.14: Les valeurs expérimentales, prédites et résiduelles de pIC50 75

Liste des Figures :

Figure I. 1: Présentation de la maladie d'Alzheimer……6
Figure I. 2: Progression des symptômes de la maladie d'Alzheimer ……..…7
Figure I. 3: Représentation graphique montrant le pourcentage des cas d'Alzheimer en fonction de l'âge. 7
Figure I. 4: Réaction de phospho-transfertpar une protéine kinase. 11
Figure I.5: Les protéines cibles de DYRK1A et ses fonctions (adapté d'après) 15
Figure II. 1: La laison d'Hydrogène…… …….....27

Figure II. 2: principe de la validation croisée leave-n-out. 42
Figure III.1:Structure générale de 6-arylquinazolin-4-amine……..48
Figure III.2: Schéma de La méthodologie générale d'une étude QSAR 54
Figure III.3: Les principales composantes et leurs variances 58
Figure III.4: Cercle de corrélation pour les axes F1, F2 58
Figure III.5 : Diagramme de la contribution en pourcentage de chaque descripteur dans le modèle pIC50 développé expliquant la variation de l'activité 62
Figure III.6: La courbe de Williams pour le modèle MLR 72
FigureIII.7: La courbe de Williams pour le modèle MNLR 73
Figure III.8: Les courbes des valeurs prédictives en fonction des valeurs expérimentales de pIC50 pour MLR etMNLR 76
Figure III.9: La courbe des valeurs résiduelles par rapport à l'expérimentale par MLR 77
Figure III.10 : La courbe des valeurs résiduelles par rapport à l' expérimentale par MNLR 77

Liste des abréviations :

aPKs : protéines kinases eucaryotes atypiques

APP : Amyloid Precursor Protein

ATP : Adénosine TriPhosphate

AVC : Accident Vasculaire Cérébral

Aβ : Amyloïde β

B3LYP: Becke 3-Parameter Lee-Yang-Parr

BACE1 : β-site APP-Cleaving Enzyme 1

d : La densité

DFT : La Théorie de la Fonctionnelle de la Densité.

DNF : Dégénérescences Neurofibrillaires

DSK: Dual-Specificity Kinase

DYRK: Dual-Specificity-Tyrosine phosphorylation-Regulated Kinase

DYRK1A: Dual-specificity-Tyrosinephosphorylation-Regulated Kinase 1A

ePKs: protéines kinases eucaryotes typiques

E_T : L'énergie totale

HOMO : Highest Occupied Molecular Orbital

LOO: Leave One Out

MA : Maladie d'Alzheimer

MLR: Multiple Linear Regression

MM : Mécanique Moléculaire.

MNLR : La régression non linéaire multiple

Liste des abréviations :

MR : La réfractivité moléculaire

MV : Le volume moléculaire

n : L'indice de réfraction

PCA : Analyse en composantes principales

PHF : Paires de filaments hélicoïdaux

PLS : Régression aux moindres carrés partiels

QSAR : La relation quantitative structure activité

SLR : Régression linéaire simple

SSE: Sum of Squares Error

SSR : Sum of Squares Regression

αe : La polarisabilité

μ : Le moment dipolaire

Glossaire de biochimie

Aβ : Amyloïde β ; peptide impliqué dans la maladie d'Alzheimer.

APP : Amyloid precursor protein ; protéine transmembranaire responsable de la formation despeptides amyloïdes β.

Kinase : Enzyme de la famille des transférases catalysant le transfert d'un groupe phosphoryle d'unsubstrat à un autre..

Le kinome est l'ensemble des kinases exprimées dans une cellule, une partie d'une cellule (membranes, organites) ou un groupe de cellules (organe, organisme, groupe d'organismes) dans des conditions données et à un moment donné. Il est à ce titre un sous-protéome, un sous-ensemble d'un protéome.

le terme eucaryote désigne l'ensemble des organismes unicellulaires ou multicellulaires dont les cellules sont dites « eucaryotes ». Elles possèdent un noyau et des organites (réticulum endoplasmique, appareil de Golgi, plastes divers, mitochondries, etc.) délimités par des membranes.

PS1 ou 2 : Préséniline 1 ou 2 ; protéine impliquée dans le clivage de la protéine APP en intervenant dans l'activité de la γ-sécrétase.

Tau : Protéine appartenant à la famille des microtubules associated proteins (MAP), responsable de l'assemblage et de la stabilisation des microtubules.

α/γ sécrétase : Enzyme mise en jeu dans le clivage de la protéine APP.

Les dégénérescences neurofibrillaires (DNF): également appelées enchevêtrements neurofibrillaires sont des agrégats de protéines tau hyperphosphorylées, retrouvés au sein des neurones.

Les plaques séniles: également appelées plaques amyloïdes correspondent à des agrégats extracellulaires de protéines organisées en fibrilles. Le constituant majoritaire de ces plaques est le peptide amyloïde β (Aβ).

La phosphorylation des protéines: est une modification post-traductionnelle impliquée dans un très grand nombre de phénomènes cellulaires.

Introduction Générale

Introduction générale :

La maladie d'Alzheimer est une des maladies neurodégénératives les plus répandues dans le monde et est la cause principale de démence. En effet, 60 à 70% des cas seraient dus à cette maladie, ce qui représente 46,8 millions de personnes environ en 2015 (World Alzheimer Report World 2015). Moins de 2% des personnes âgées de moins de 65 ans sont touchées contre 15% pour celles âgées de 80 ans et plus (Inserm). Ainsi, avec l'augmentation et le vieillissement de la population, ce fléau ne va tendre qu'à augmenter au cours des prochaines années. Le nombre de personnes touchées en 2050 pourrait atteindre 131,5 millions (World Alzheimer Report World 2015). De plus, la prise en charge des patients est extrêmement coûteuse ; le coût mondial de la maladie a été estimé en 2015 à environ 726 milliards d'euros [1]. La maladie d'Alzheimer apparait donc véritablement comme une priorité de santé publique.

La phosphorylation par les kinases est le mécanisme le plus universellement utilisé par les cellules pour contrôler les fonctions de leurs protéines. Tous les principaux phénomènes physiologiques sont régulés par phosphorylation et de nombreuses maladies sont associées à une phosphorylation anormale [2]. Parmi les 518 kinases humaines, la kinase DYRK1A (et les kinases proches) constitue bien une cible thérapeutique d'intérêt majeur puisqu'elle entre en jeu dans le diabète et certains cancers aussi plusieurs maladies neurologiques graves comme la maladie d'Alzheimer [2].

Les études théoriques s'orientent actuellement vers la conception rationnelle "Rational design" qui signifie que la connaissance des relations entre les propriétés physico-chimiques et la structure moléculaire des molécules connues permet aux scientifiques de développer des nouvelles molécules, avec une assez bonne anticipation [3].

Parmi les techniques de chimie computationnelle, nous pouvons citer les techniques de QSAR (Quantitative Structure-Activity Relationships) qui consiste à trouver une corrélation entre une activité biologique mesurée pour un panel de composés et certains descripteurs moléculaires.

Elle permet également de guider les processus de développement de nouvelles molécules, sans avoir à les synthétiser, ou à analyser des familles entières de composés. Les

relations entre les structures des molécules et leurs propriétés ou activités sont généralement établies à l'aide de méthodes de modélisation par apprentissage statistique. Les techniques usuelles reposent sur la caractérisation des molécules par un ensemble de descripteurs ; nombres réels mesurés ou calculés traduisant des propriétés des structures moléculaires. Actuellement, il est possible de générer des milliers de descripteurs, cependant un seul petit nombre peut se révéler pertinent pour modéliser une activité biologique donnée. Pour sélectionner les descripteurs les plus efficaces pour les composés, les méthodes de sélection de descripteur appropriées sont nécessaires [4-5].

Dans ce contexte, l'objectif de ce travail est d'élaborer des modèles QSAR robustes et fiables, en respectant toutes les étapes d'une étude QSAR pour une série de la 6-arylquinazoline-4-amine pour l'inhibition de DYRK1A, qui peuvent être utilisés pour concevoir des dérivés nouveaux et puissants comme médicaments potentiels pour le traitement de MA .

Ce mémoire se divise en trois chapitres :

- le premier chapitre présente un aperçu général sur la maladie d'Alzheimer : les symptômes, les stades et la physiopathologie de la maladie, la fonction de la protéine kinase, en particulier, de la fonction de la DYRK1A kinase et de son importance dans le développement des maladies d'Alzheimer, le diagnostic et du traitement de cette maladie.
- Le deuxième chapitre présent une généralité sur QSAR, son principe, son application et les différentes étapes à respecter pour la réaliser.
- Le troisième chapitre est consacré à la partie pratique "Étude quantitative des propriétés QSAR d'une série de dérivés de 6- arylquinazolin-4-amine"et nous présentons les résultats obtenus et leurs discussions.
- Enfin, nous terminerons ce manuscrit par une conclusion générale.

Chapitre 01:

la maladie d'Alzheimer

I.1 Introduction :

La maladie d'Alzheimer constitue la cause la plus fréquente de démence ; elle représente un véritable problème de santé publique.

La démence est une affection cérébrale qui entraine une altération de la mémoire et autres fonctions cognitives avec une répercussion fonctionnelle et sociale. Cette pathologie se caractérise par l'affaiblissement de la sensibilité, de l'intelligence et de la volonté de l'individu. Elle est caractérisée par une évolution irrémédiablement progressive de ce déficit. La démence est généralement due à une atteinte cérébrale organique plus ou moins diffuse, un processus de dégénérescence des neurones, vasculaire, infectieuse, traumatique, toxique ou tumorale. On distingue 2 types de démences dégénératives en fonction de la localisation:

- Les démences corticales comprenant la maladie d'Alzheimer, les démences frontotemporales et les démences à corps de Lewy.
- Les démences sous corticales en rapport avec la Chorée de Huntington, la maladie de Creutzfeld-Jacob, la maladie de Parkinson et la maladie de Steele Richardson [6].

La définition anatomo-clinique de la maladie d'Alzheimer se caractérise par l'association d'un syndrome démentiel d'évolution progressive (avec des troubles mnésiques (qui se rapporte à la mémoire) importants au premier plan) et des lésions cérébrales caractéristiques [7].

I.2 Généralité

I.2.1 Définition

La maladie d'Alzheimer est une affection du cerveau dite « neurodégénérative » qui entraîne une disparition progressive des neurones. Elle provoque une altération des facultés cognitives: mémoire, langage, raisonnement, etc. L'extension des lésions cérébrales cause d'autres troubles qui réduisent progressivement l'autonomie de la personne [8]. Elle apparaît plus souvent chez les personnes âgées, mais elle n'est pas une conséquence normale du vieillissement.

I.2.2 Historique

L'histoire de la maladie d'Alzheimer commence en 1906 en Allemagne. Auguste D. Tübingen est adressée à Aloïs Alzheimer, neuropsychiatre. Cette patiente de 53 ans, admise dans le secteur psychiatrique, présente une symptomatologie variée associant une dégradation cognitive progressive, des comportements incohérents et imprévisibles, des hallucinations, de la confusion mentale et une inaptitude psycho-sociale. Le premier symptôme de la maladie a été un fort sentiment de jalousie envers son mari. Très vite, Auguste D montre des signes de dégradation mnésique. Elle est désorientée, place des objets n'importe où dans son appartement et les cache. Ses hallucinations la font hurler de peur. Aloïs Alzheimer décrit les troubles suivants : trouble de l'écriture d'origine mnésique, incompréhension du langage oral, oublis à mesure, apraxie, discours spontané entravé par des persévérations et des désordres paraphasiques. Auguste D. meurt d'une septicémie quatre ans et demi plus tard. Aloïs Alzheimer décide de pratiquer une autopsie sur son cerveau. Cet examen révèle la présence de plaques séniles au niveau cortical et d'amas anormaux de fibrilles dans les neurones. Ces deux types de lésions cérébrales seront dès lors considérés comme caractéristiques de la maladie. La maladie d'Alzheimer sera ainsi nommée par le maître d'Alzheimer, Kraepelin, dans son ouvrage publié en 1910. Selon lui, il s'agit d'une démence du sujet jeune, rare, dégénérative et incurable [9]. En 1978, avec Terry, que la démence sénile et la maladie d'Alzheimer furent considérées comme une entité unique classée parmi les maladies neurodégénératives; survenant exceptionnellement dans le presenium et habituellement après 65 ans. Les lésions associent une atrophie corticale progressive avec perte neuronale coexistant avec une dégénérescence neurofibrillaire et des plaques séniles [9].

I.3 Les symptômes

I.3.1 Pertes de mémoire:

La personne oublie de plus en plus souvent des événements récents touchant sa vie personnelle et son entourage mais garde une très bonne mémoire des souvenirs anciens [10].

I.3.2 Difficultés à accomplir les tâches quotidiennes:

La personne rencontre des difficultés pour effectuer des travaux pourtant familiers comme par exemple les étapes de préparation d'un repas, faire ses courses, gérer les dates de péremption des aliments dans le frigidaire... [10]

I.3.3 Problèmes de langage:

La personne ne retrouve plus des mots simples, usuels et en utilise d'autres plus ou moins appropriés [10].

I.3.4 Désorientation dans le temps et dans l'espace:

Le sens de l'orientation de la personne diminue. Elle peut se perdre, même dans des endroits pourtant familiers, et confondre les saisons [10].

I.3.5 Difficultés dans les raisonnements abstraits:

La personne rencontre des difficultés pour effectuer les formalités administratives, pour gérer ses finances, pour rédiger un chèque, pour appeler quel qu'un au téléphone[10].

I.3.6 Perte d'objets:

La personne a tendance à placer des objets dans des endroits insolites (une montre dans le four) sans jamais les retrouver [10].

I.3.7 Altération du jugement:

La personne n'arrive plus à évaluer les situations : elle porte des vêtements d'hiver en été, fait des achats démesurés de nourriture…[10]

I.3.8 Modification du comportement:

L'entourage constate l'apparition d'une tendance dépressive chez la personne ou de manifestations d'anxiété, d'irritabilité, d'agitation…[10]

I.3.9 Changement de personnalité:

La personne devient tout à fait différente de ce qu'elle était et perd son caractère propre : jalousie, idées obsessionnelles de préjudice, exubérance excessive... [10]

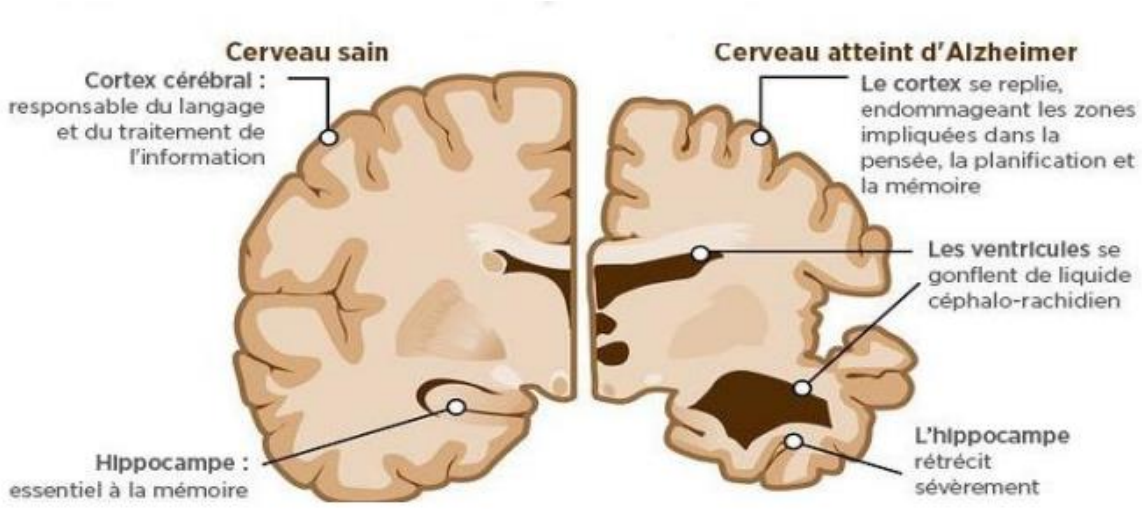


Figure I.1 : Présentation de la maladie d'Alzheimer

I.4 Les phases de la maladie d'Alzheimer

La maladie d'Alzheimer est une pathologie dégénérative qui entraine le déclin progressif des fonctions cognitives. Elle se caractérise par trois grandes phases : la phase prodromale ou MCI (de l'anglais, Mild Cognitive Impairement) qui correspond au début de la phase symptomatique, la phase de démence (légère à modérée) et la phase de démence très sévère. Selon le stade de la maladie, différents symptômes apparaissent tels que des troubles de la mémoire, des problèmes de langage, d'orientation, de motricité, d'agressivité et de personnalité (Figure I.2).

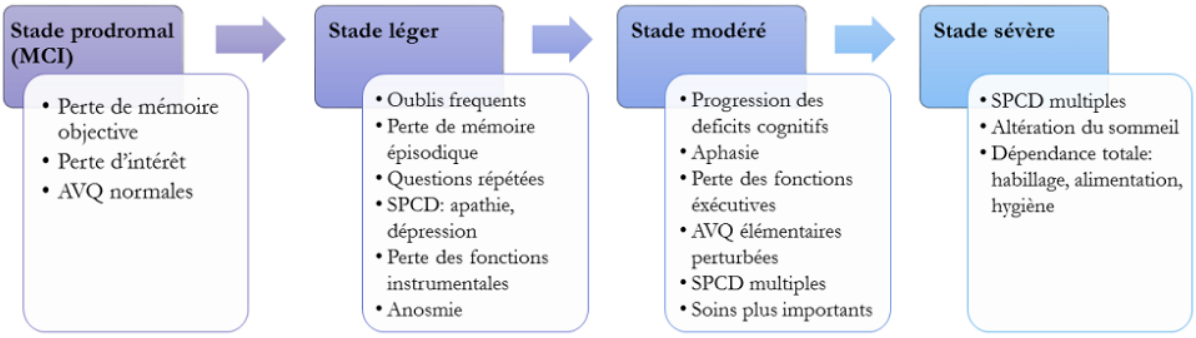


Figure I.2 : Progression des symptômes de la maladie d'Alzheimer (AVQ : Activités de la vie quotidienne ; SPCD : Symptômes psychologiques et comportementaux des démences). Figure adaptée [11].

I.5 Les facteurs de risque de l'Alzheimer:

I.5.1 L'âge :

En vieillissant, les mécanismes naturels de réparation de l'organisme sont moins efficaces. Ce changement se produit dans le cerveau à différents rythmes selon les personnes et ces différences contribuent à la susceptibilité d'une personne de développer la maladie d'Alzheimer avec l'âge [12] (voir la figure I.3).

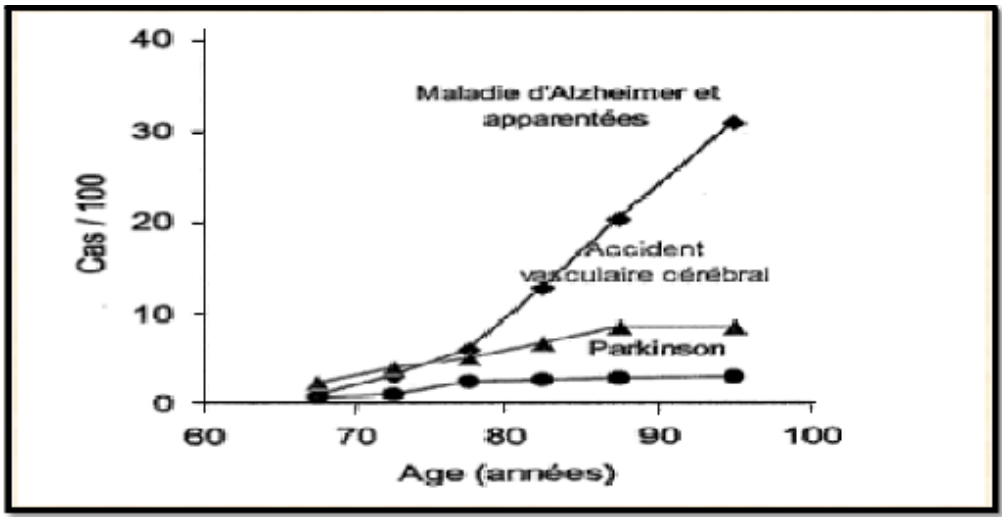


Figure I.3 : représentation graphique montrant le pourcentage des cas d'Alzheimer en fonction de l'âge.

I.5.2 Les antécédents familiaux et la génétique :

Un faible pourcentage (5 à 7 %) des personnes atteintes de la maladie d'Alzheimer ont la forme « familiale » de cette maladie (FFMA) (autrefois appelée « forme précoce de la maladie d'Alzheimer »). À un moment donné de l'histoire familiale, certains gènes ont subi

une mutation et ont développé les caractéristiques anormales qui causent la FFMA. Ces gènes héréditaires ont une grande influence: si l'un des parents a la FFMA, chacun des enfants aura une probabilité de 50% d'hériter de la maladie. Si les deux parents ont la FFMA, tous leurs enfants développeront la maladie d'Alzheimer à l'âge adulte [13].

I.5.3 Le sexe :

Deux fois plus de femmes que d'hommes ont la maladie d'Alzheimer. Certains pensent que ceci est attribuable en grande partie aux changements hormonaux qui surviennent à la ménopause, particulièrement la diminution de l'œstrogène, une hormone importante. Auparavant, on prescrivait souvent de l'œstrogène pour soulager les symptômes de la ménopause et réduire le risque de développer la maladie d'Alzheimer. Toutefois, une étude clinique à grande échelle assez récente recommandait d'interrompre le traitement hormonal substitutif (THS) en raison de ses effets secondaires potentiellement dangereux. Plusieurs chercheurs cliniques considèrent cependant que le THS devrait faire l'objet d'études plus poussées dans le contexte de la maladie d'Alzheimer. Toute décision quant à son usage devrait être faite en consultation avec un médecin [13].

I.5.4 Les maladies cardiovasculaires :

Tous les facteurs de risque de maladies cardiovasculaires (comme l'hypertension et l'hypercholestérolémie) sont aussi des facteurs de risque de la maladie d'Alzheimer et de la maladie cérébro-vasculaire. Les AVC et les « mini-AVC », généralement détectés lors d'examens ultérieurs, sont aussi des facteurs de risque bien établis pour la maladie d'Alzheimer et la maladie cérébro-vasculaire [13].

I.5.5 Le diabète :

On sait depuis plusieurs années que le diabète de type 2 (diabète adulte) est un facteur de risque de la maladie d'Alzheimer. On croyait généralement que ces deux maladies étaient liées aux problèmes cardiaques, qui sont associés au diabète et constituent des facteurs de risque de la maladie d'Alzheimer. On savait aussi que le glucose était moins bien assimilé dans le cerveau des personnes atteintes de la maladie d'Alzheimer, une situation similaire aux diabétiques de type 2, dont l'organisme assimile mal le glucose [13]. Les chercheurs ont en effet découvert que chez les personnes atteintes de la maladie d'Alzheimer, la production d'insuline dans le cerveau est réduite et que les neurones y sont moins sensibles (la production

de l'insuline dans le cerveau est indépendante de la production de l'insuline dans le pancréas, le principal organe de production d'insuline).

I.5.6 Le syndrome de Down :

Le cerveau de la plupart des adultes atteints du syndrome de Down qui atteignent la quarantaine développera les changements anormaux caractéristiques de la maladie d'Alzheimer (plaques et écheveaux) [13].

I.5.7 La déficience cognitive légère (DCL) :

Dans la DCL, le niveau de détérioration cognitive et/ou des troubles de la mémoire est supérieur à celui enregistré dans le processus normal de vieillissement, mais il n'est pas suffisamment avancé pour qu'on puisse parler d'Alzheimer ou de maladie apparentée. On estime que 85 % des personnes qui ont reçu un diagnostic de DCL, généralement au début de la quarantaine ou de la cinquantaine, développeront la maladie d'Alzheimer dans les dix années suivantes, ce qui fait de la DCL un facteur de risque important. Les chercheurs savent maintenant que les changements cérébraux anormaux qui caractérisent la maladie d'Alzheimer peuvent commencer à apparaître chez les personnes qui ont reçu un diagnostic de DCL au moins 20 ans avant tout signe visible d'Alzheimer ou de maladie apparentée. L'imagerie cérébrale pourrait permettre de repérer les personnes atteintes de DCL les plus à risque de développer la maladie [13].

I.5.8 Les lésions à la tête :

Les lésions à la tête, notamment les commotions à répétition, sont considérées par la plupart des cliniciens comme des facteurs de risque de développement ultérieur de la maladie d'Alzheimer. En plus des facteurs de risque précités, tous les éléments suivants ont été recensés comme facteurs de risque de la maladie d'Alzheimer : les inflammations chroniques (indiquant possiblement une défaillance du système immunitaire), des épisodes passés de dépression clinique, le stress et le manque d'exercice du cerveau. D'autres facteurs de risque comme le tabagisme, la consommation excessive d'alcool et les toxicomanies restent moins fondés [13].

I.6 Physiopathologie de la maladie d'alzheimer

I.6.1 Hypothèse de la cascade β-amyloïde:

Dans l'hypothèse de la cascade β-amyloïde, le clivage de l'APP interviendrait avant l'hyperphosphorylation de tau et entraînerait une surproduction de peptides amyloïdes$A\beta_{42/43}$. Selon cette hypothèse, le clivage de l'APP induit la surproduction toxique de peptides $A\beta_{42/43}$, provoquant une mort neuronale, une perte synaptique et l'apparition de plaques séniles. Les plaques séniles seraient une réponse adaptative de la cellule nerveuse et permettraient de séquestrer les peptides toxiques $A\beta_{42/43}$ (42acides aminés)sous-forme d'agrégats insolubles, les rendant inactifs. La présence de DNFs serait un effet périphérique de la surproduction de peptides β-amyloïde. Lorsque la cinétique de formation des plaques séniles, séquestrant les peptides Aβ est insuffisante pour la cellule, la mort neuronale et la perte synaptique interviendraient et généreraient une démence[14].

I.6.2 Hypothèse de la dégénérescence du cytosquelette neuronal:

Dans l'hypothèse de la dégénérescence du cytosquelette neuronal, l'inducteur initial serait une perturbation de l'équilibre entre kinases et phosphatases de tau (activation des kinases et/ou inhibition des phosphatases), ce qui génère une hyperphosphorylation des protéines tau, ce qui conduit à une déstabilisation du cytosquelette neuronal. La déstabilisation du cytosquelette provoquerait une perte synaptique, la survenue de DNFs et de plaques séniles. Les DNFs seraient une réponse adaptative de la cellule nerveuse qui permettrait de séquestrer les protéines tau anormalement phosphorylées, ce qui les empêcheraient ainsi de réguler la dynamique du cytosquelette neuronal. Lorsque la cinétique de formation des DNFs, séquestrantles protéines tau hyperphosphorylées, est insuffisante pour la cellule, la mort neuronale et la perte synaptique interviendraient et conduiraient à une démence [14].

Selon ces hypothèses, Des anomalies de phosphorylation sont également impliquées dans la maladie d'Alzheimer. Ainsi, la formation des plaques séniles et celle des dégénérescences neurofibrillaires sont-elles toutes deux la conséquence de l'activité anormale de certaines protéines kinases.

I.7 Les kinasese:

I.7.1 Définition:

Une protéine kinase catalyse le transfert d'un γ-phosphate, prélevé d'une molécule d'ATP (Adénosine TriPhosphate) ou plus rarement de GTP (Guanosine TriPhosphate), sur le site de phosphorylation d'une protéine réceptrice ou cible (Figure I.4). Ce site de phosphorylation est un acide aminé contenant un groupement hydroxyl, habituellement Sérine (Ser), Thréonine (Thr) ou Tyrosine (Tyr) chez les cellules eucaryotes.

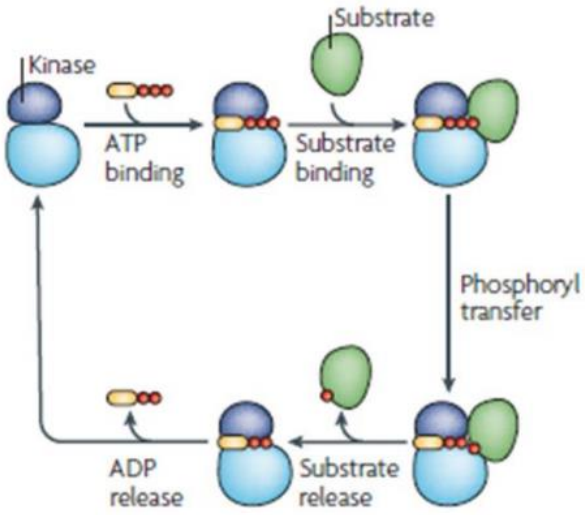


Figure I.4: Réaction de phospho-transfertpar une protéine kinase.

Le cycle catalytique de la phosphorylation d'une protéine cible par une protéine kinase: En haut à gauche, l'ATP puis le substrat se lient au niveau du site actif de la kinase. Une fois cette liaison établie, le γ-phosphate de l'ATP (rouge) est transféré sur un résidu Ser, Thr ou Tyr de la protéine cible. Après la phosphorylation, la protéine réceptrice et l'ADP sont libérés du site actif de la kinase [15].

I.7.2 Classification :

Les protéines kinases ont été identifiées par le projet « kinome humain » à partir des entrées présentes dans les bases de données génomiques et protéiques [16]. En analysant les similarités des séquences et des structures de leurs domaines catalytiques, elles ont été divisées en deux super-familles et nommées respectivement« ePKs» pour protéines kinases eucaryotes typiques et « aPKs » pour protéines kinases eucaryotes atypiques. Au nombre de 478, les ePKS représentent la grande majorité des kinases. Celles-ci présentent une forte homologie de séquence au niveau de leur domaine catalytique [17]. Au contraire, les aPKs sont moins nombreuses en nombre (40) et ne présentent pas de similarité avec le domaine

kinase des ePKs. En fonction de leurs homologies de séquence, de leurs domaines catalytiques, de la structure de leurs régions N- et C-terminales, de leurs fonctions biologiques et de l'existence d'une classification précédemment établie chez les protéines kinases de levure, de *C.elegans* et de drosophile, les protéines kinases humaines ont d'abord été classées en 9 groupes, 134 familles et 201 sous-familles, comme présentées dans le Tableau I.1.

Tableau I.1. Classification des kinases

Super-familles	Groupes	Origine du nom	Familles	Sous-familles	Kinases de levures	Kinases de C.elegans	Kinases de drosophile	Kinases humaines	Exemples de kinases humaines
ePK	AGC	PKA/PKG/PKC kinases group	14	21	17	30	30	63	PKB, PKA, PKC, PKG
	CAMKIIKI	Calcium/Calmodulin regulated kinases group	17	33	21	46	32	74	CAMKL, CAMKII, PDK
	CKI	Casein kinase 1 group	3	5	4	85	10	12	CKI, TTBK, VRK,
	CMGC	CDK/MAPK/GSK3/CLK kinases group	8	24	21	49	33	61	CDK5, ERK1, DYRK1A, GSK3
	STE	Yeast STE-like MAPK kinases group	3	13	14	25	18	47	MAP2K, MAP3K, MAP4K
	TK	Tyrosine kinases group	30	30	0	90	32	90	TrkA, FGFR, SRC
	TKL	Tyrosine kinases-like group	7	13	0	15	17	43	LRRK, MLK, STRK
	RGC	Receptor Guanylate Cyclase group	1	1	0	27	6	5	GCA, GC2D, GC2F
	Other	Other ePK kinases	37	39	38	67	45	83	CAMKK, PLK, AUR
aPK	Atypical	Atypical protein kinases	14	22	15	20	17	40	HistK, BCR, PDHK
Total			134	201	130	454	240	518	

En plus de l'homologie du domaine catalytique utilisée dans le projet « human kinome » la nature de l'acide aminé récepteur du γ-phosphate d'ATP constitue un second critère de classification. Trois grands groupes sont généralement évoqués :

- ✓ **Les protéines sérine/thréonine kinases (PSTKs):**

  Alors que ces protéines ciblent des résidus sérine ou thréonine, elles représentent 80 % du kinome humain et sont activées par d'autres kinases situées en amont dans la cascade de signalisation via la phosphorylation de résidus sérine, thréonine ou tyrosine localisés au niveau de leur boucle d'activation [15]. Les membres de la famille des MAPKs (ERK1/2), des CAMKs (CAMKII) ou des GSK3 ne sont que des exemples la plus connus.
- ✓ **Les protéines tyrosine kinases (PTKs):**

La nature de leurs acides aminés cibles sont des tyrosines. Elles représentent 20 % des protéines kinases et sont présentes sous deux formes : **récepteurs** à activité tyrosine kinase (comme les facteurs neurotrophiques NGF ou BDNF) ou protéines tyrosines kinases **solubles**. Pouvant être cytosoliques ou nucléaires, ces dernières ont la capacité lorsqu'elles sont actives de recruter des protéines d'ancrage, des facteurs de transcription (comme par exemple STAT3) ou d'autres protéines kinases [18]-[19]

- ✓ **Les protéines kinases à double spécificité (DSKs):**

  Découverte dans les années 1990, ces protéines kinases ont la capacité de phosphoryler leurs cibles protéiques à la fois sur des résidus sérine/thréonine et les résidus tyrosine (double activité kinase : sérine/thréonine et tyrosine) [20]. Il est à noter qu'il y a cependant de **vrais DSKs** et des **DSKs activées par phosphorylation**. Ces dernières regroupent la famille des DYRKs, dont DYRK1A pour « Dual-specificity Tyrosine (Y) phosphorylation Regulated Kinase 1A», qui une fois activées par autophosphorylation de leurs propres résidus Tyrosine situés dans leurs boucles d'activation, phosphorylent des résidus Ser/Thr (activité Ser/Thr kinase « simple ») [21].

I.8 DYRK1A:

I.8.1 Définition:

La protéine DYRK1A fait partie des DSKs ou protéines kinases à double spécificité. Elle est cependant à différencier des « vrai » DSKs, qui phosphorylent à la fois des résidus Tyr et des Ser/Thr, puisque DYRK1A va d'abord activer par autophosphorylation ses propres résidus Tyrosines pour catalyser ensuite des réactions de phospho-transferts sur des résidus Serine ou Thréonine comme une « simple » Ser/Thr kinase [22].

I.8.2 Fonctions:

Les protéines phosphorylées par DYRK1A interviennent dans diverses fonctions et font parties intégrantes de voies de transduction du signal . En raison de la multiplicité des fonctions exercées par ses cibles protéiques, DYRK1A est potentiellement associé, plus ou moins directement, à de nombreuses fonctions physiologiques et physiopathologiques :

- Morphogenèse du cerveau : contrôle du volume du cerveau et de la densité cellulaire,mise en place des neurones et des cellules gliales, mise en place des neurites.

- Régulation de la différenciation, de la prolifération et de la croissance cellulaire.DYRK1A participe au contrôle du cycle cellulaire, de l'apoptose et de la surviecellulaire en phosphorylant des protéines comme la caspase 9 ou la P53.
- Contrôle de l'endocytose, de la fusion des vésicules synaptiques et de la libération des neurotransmetteurs dans la fente synaptique. DYRK1A phosphoryle des protéines synaptiques, ce qui lui confère un rôle de régulation de la plasticité synaptique.
- Transcription de gènes via la phosphorylation de facteurs de transcription comme CREB, NFAT, STAT3.
- Communications entre les cellules et métabolisme : glycogène, néoglucogenèse…DYRK1A phosphoryle des protéines intervenant dans ces mécanismes (Notch,Glycogène synthase …).
- Protection de l'hypertrophie cardiaque.
- Certaines protéines (Cycline L2, ASF …) intervenant dans l'épissage alternatif sont phosphorylées par DYRK1A.
- Des protéines associées à la mise en place de certaines maladies neurodégénératives sont phosphorylées par DYRK1A ou voient leurs phosphorylation modifiée par l'intervention de DYRK1A sur d'autres protéines : les protéines Tau et APP, la Setpin4 et l'α-Synucléine, la prenisilin1.

Globalement, DYRK1A joue un rôle dans le maintien de l'activité neuronale chez l'adulteainsi que dans la consolidation de la mémoire et l'apprentissage [23].

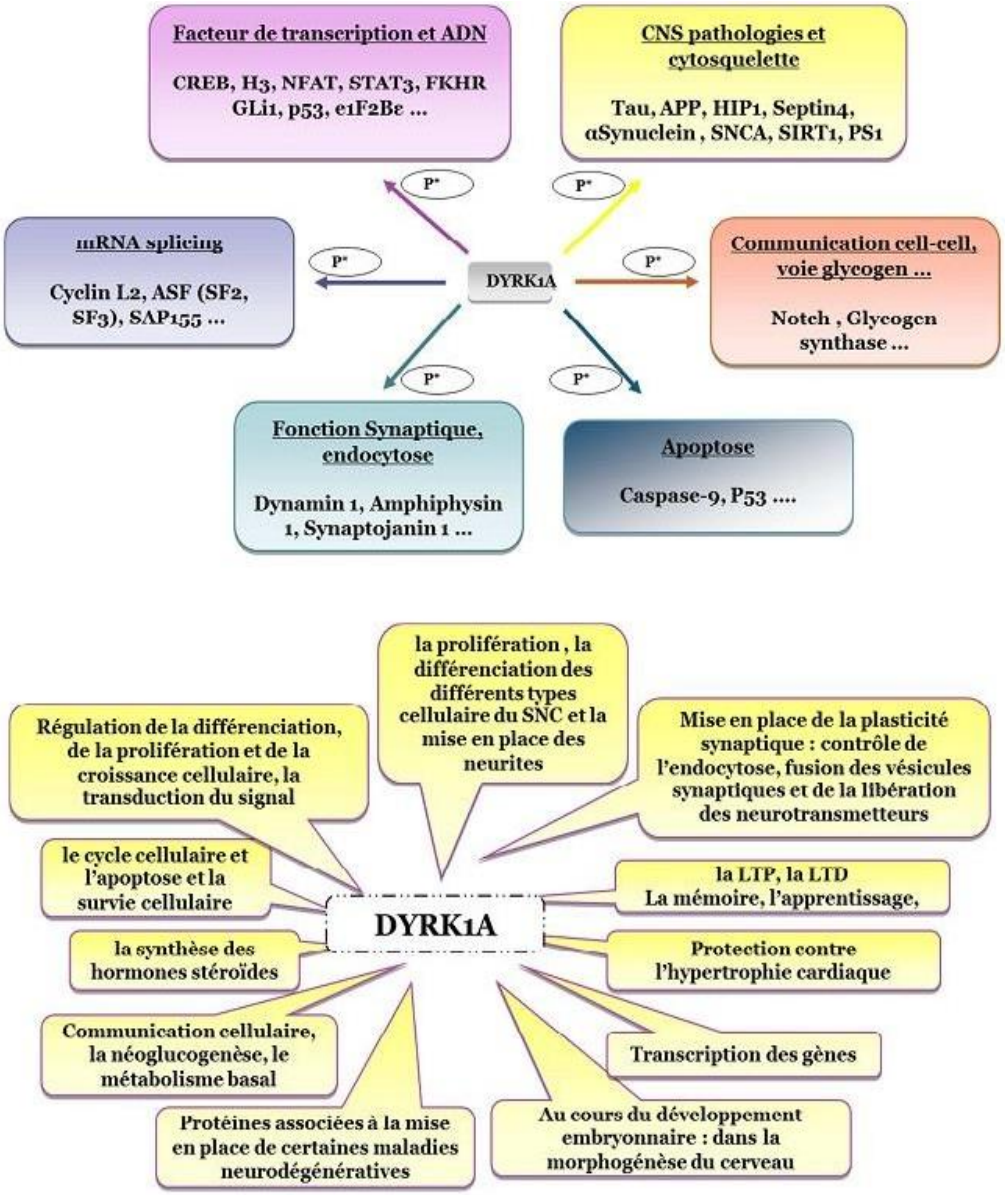


Figure I.5: Les protéines cibles de DYRK1A et ses fonctions **(adapté d'après** [24]-[26]

I.8.3 Influence de la protéine kinase DYRK1A sur la maladie d'Alzheimer

La kinase-1A régulée par la tyrosine à phosphorylation à double spécificité (DYRK1A) est un membre de la protéine kinase de la famille DYRK , qui comprend cinq kinases . Les niveaux de DYRK1A sont étroitement réglementés, car la surexpression et le manque de L'activité DYRK1A est associée à un retard mental. DYRK1A peut directement phosphoryler le tau sur de nombreuses sérines et résidus de thréonine.

À cette fin, la surexpression de DYRK1A contribue à l'accumulation de DNF dans Syndrome de Down et la maladie d'Alzheimer en plus de ses effets sur le tau pathologie,

DYRK1A semble être lié au métabolisme APP / Aβ.

Par exemple, dans les neurones corticaux primaires de rat, l'inhibition de DYRK1A réduit le tau phosphorylation à plusieurs épitopes de manière dose-dépendante. Le même inhibiteur a réduit la production d'Aβ Cellules HEK293 sur exprimant APP. C'est conforme aux résultats antérieurs montrant que DYRK1A phosphoryle APP et améliore son affinité pour BACE1 et le complexe γ-sécrétase, augmentant ainsi les niveaux globaux d'Aβ et le dépôt de plaque. Les premières preuves suggèrent également que l'accumulation d'Aβ peut faciliter Activité DYRK1A, créant un cercle vicieux. Cohérent avec ces observations, les niveaux de DYRK1A sont augmentés en post-mortem cerveaux MA humains . Ensemble, ces données suggèrent que la réduction de l'activité de DYRK1A pourrait être une stratégie thérapeutique valable pour ciblant à la fois Aβ et tau dans la MA [27].

I.9 Diagnostic :

Le diagnostic de la maladie d'Alzheimer repose sur différents examens :

- Un entretien approfondi avec le patient et sa famille, afin de retracer l'historique d'apparition des troubles. Il peut s'appuyer sur des questionnaires dans lesquels la famille évalue l'intensité des troubles et leur retentissement sur la vie quotidienne (faire sa toilette, prendre ses repas, faire les courses, prendre les transports en commun…).

- Un examen clinique approfondi. Il élimine la présence d'autres causes pouvant expliquer les troubles neurologiques.

- Des tests psychométriques. Ils permettent d'explorer la mémoire, le langage, les fonctions exécutives, l'orientation dans l'espace et dans le temps… Ils sont réalisés par un psychologue ou un médecin.

- Des examens biologiques. Ils permettent de vérifier que ce ne sont pas des facteurs somatiques qui sont à l'origine des troubles cognitifs.

- Un scanner ou un IRM viennent en général compléter ces examens.

Le diagnostic est parfois difficile à établir en début de maladie, car les symptômes sont encore peu nombreux [28].

I.10 Le traitement de la maladie d'Alzheimer

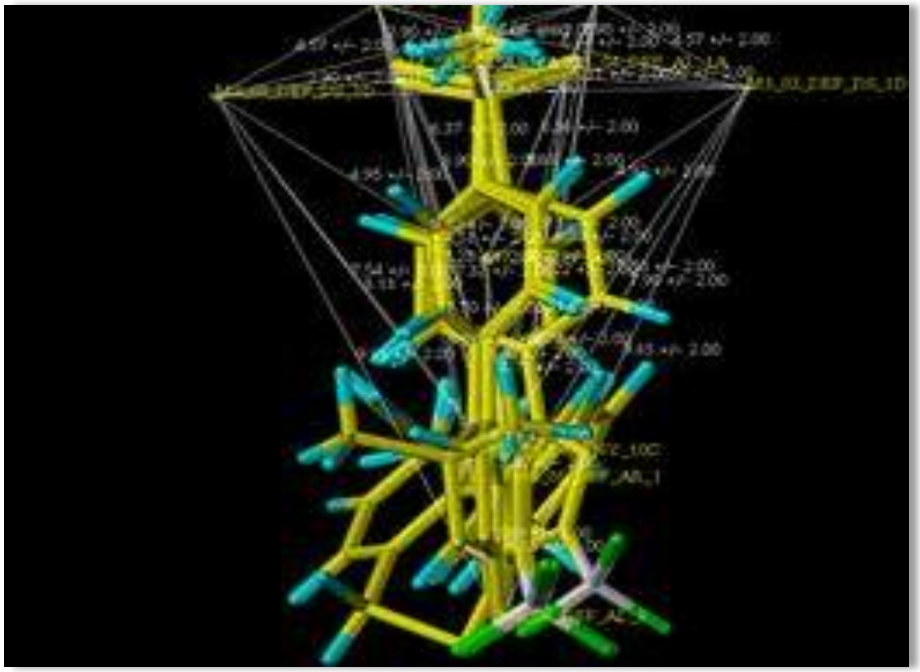
Actuellement, il n'existe aucun traitement guérissant la maladie d'Alzheimer, ni même permettant d'arrêter son évolution, mais il existe quelques médicaments susceptibles de retarder son évolution. Ils permettent d'atténuer les pertes de mémoires, les problèmes de langage et de raisonnement, ou bien tous simplement de ralentir au moins en apparence la progression de la maladie. La cause exacte est encore inconnue, mais il est supposé que des facteurs environnementaux et génétiques y contribuent.

Les plus grandes méthodes dans le traitement de la MA se focalisent sur deux aspects fondamentaux : traitement au niveau des plaques séniles ou plaque amyloïdes, et traitement de la transmission cholinergique. Ce sont les eux approches dans le traitement de cette maladie. Dans la première approche, Il s'agit donc de lésions extracellulaires de la maladie d'Alzheimer. Ces plaques correspondent à l'accumulation d'un peptide anormal et neurotoxique de 42 acides aminés, le peptide Aβ.Ce peptide, provient d'un mauvais clivage de la protéine APP (Amyloïd Protein Precursor). Ceci participerait à l'entrée massive de calcium dans le neurone et activerait la microglie (réaction inflammatoire), ce qui entraîne la mort du neurone par nécrose ou par apoptose. Donc il est préalable de prévenir la formation de ce peptide ou au mois diminuer sa génération ou déposition. De ce fait il est demandé de bloquer ou inhiber l'APP qui est responsable de la formation des Aβ peptides [29].

La plus grande partie des recherches scientifiques réalisées sur le traitement de la maladie d'Alzheimer (MA) a été guidée par un constat : l'existence d'une relation dans cette maladie, entre le déficit des neurotransmissions cholinergiques et les différents symptômes, en particulier cognitifs. Les recherches en donc eu pour but d'augmenter l'activité cholinergique centrale. [30] ce qui définit la deuxième approche dans le traitement de la maladie d'Alzheimer. Les plus étudiés, les inhibiteurs des cholinestérases, sont les seuls à avoir obtenu jusqu'à présent une application clinique [30].

Chapitre 02:

QSAR: Principe et Méthodologie



II.1 Historique :

Il y a plus d'un siècle et demi, en 1863, Cros a observé que le point d'ébullition et le point de fusion des alcanes augmente avec le nombre d'atomes de carbone et la masse moléculaire. Il a observé également une diminution de la solubilité dans l'eau des alcools avec l'augmentation du nombre d'atomes de carbone et la masse moléculaire, cela est considéré depuis comme la première formulation générale en RQSP [108].

Cinq ans après, en 1868, Crum-Brown et Fraser postulèrent que « l'activité biologique d'une molécule est une fonction de sa constitution chimique ». Quelques décennies plus tard, en 1893, Richet a montré que la cytotoxicité de certains composés organiques était inversement proportionnelle à leur solubilité dans l'eau.

A la fin du 19ème siècle, Meyer en 1899 et Overtonen 1901, ont indépendamment observé « une relation linéaire entre l'activité des narcotiques et leur coefficient de partage huile-eau »[108].

Six ans après, en 1907, Fühneret Neubauer [31]ont montré pour une série de narcotiques homologues, que l'activité augmentait en fonction de la progression géométrique de la série de composés, ceci montrant l'importance de la contribution d'additivité de groupements fonctionnels pour l'activité biologique.

En 1962, Hansen [32]a montré l'existence d'une corrélation entre la toxicité des acides benzoïques substitués et les constantes électroniques « σ » des substituants.

L'année 1964 est considérée comme le début des méthodes RQSA modernes. Hanschet Fujitaont établi les premières corrélations entre les propriétés physico-chimiques (log P, pKa, paramètres stériques et électroniques) et l'activité biologique (activité enzymatique, pharmacologique), Ces méthodes seront appelées par la suite l'analyse de Hanschet l'analyse de Free Wilson [33-34]). Sept ans plus tard, Hanschet Lien ont réalisé une étude RQSA sur différentes familles d'antifongiques : benzoquinones, sels d'alkylpyridinium, imidazoles et phénols. Ils ont observé que quels que soient la famille et le champignon utilisé, l'activité antifongique dépend du coefficient de partage Eau-Octanol, expérimental ou calculé.

Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physico-chimiques des analystes avec les temps de rétention obtenus expérimentalement : c'est l'étude quantitative des relations structure temps de rétention noté RQSR [108].

II.2 Définition :

La relation quantitative structure activité (QSAR) est le processus par lequel la structure chimique est corrélée quantitativement avec un processus bien défini, tel que l'activité biologique ou la réactivité chimique.

L'activité biologique peut être exprimée quantitativement par la concentration d'une substance requise pour donner une certaine réponse biologique. L'expression mathématique peut être alors utilisée pour prédire la réponse biologique d'autres structures chimiques **[35]**.

La réactivité chimique peut être exprimée par des propriétés telles que la lipophilicite, la solubilité et la perméabilité. [36]

La forme mathématique la plus générale de QSAR est : [34]

Activité = f (x)

X : propriétés physico-chimiques et / ou propriétés structurelles.

L'objectif de ces méthodes est alors d'analyser les données structurales afin de déterminer les facteurs influençant l'activité mesurée [37]. Pour ce faire, différents types de méthodes statistiques peuvent être employées.

La relation mathématique obtenue peut alors être utilisée comme moyen prédictif de l'activité biologique étudiée de nouvelles molécules ou des molécules pour lesquelles les données expérimentales ne sont pas encore disponibles. Ils peuvent être également utilisés pour mieux comprendre les mécanismes et les modes d'action [105].

II.3 Principe :

La méthodologie générale d'une étude QSAR est la suivante :

1. Constituer la base de données structure – activité à partir de mesures quantitatives, fiables pour chaque composé.
2. Sélectionner des descripteurs moléculaires en relation avec l'activité cible afin de traduire de manière numérique la structure des molécules.
3. Diviser ce jeu de données aléatoirement en un d'apprentissage et un autre de test .
4. Établir le modèle mathématique en utilisant la série d'apprentissage à l'aide des méthodes statistiques.

5. Caractériser ce modèle par leurs indices statistiques et par une validation interne .

6. Valider les modèles avec la série de test et calculer leur indice de corrélation externe .

7. Élaborer le domaine d'applicabilité du modèle proposé .

8. Explorer et exploiter les modèles validés pour comprendre les mécanismes possibles et faire des prévisions d'activité biologique de nouvelles molécules [37].

Il y a plusieurs raisons pratiques qui justifient l'utilisation des études QSAR

- ✓ Prédire les propriétés de l'activité biologique par des moyens rationnels.
- ✓ Économiser le coût de développement de médicament.
- ✓ Les prévisions pourraient réduire l'exigence de tests longs et coûteux pratiqués sur les animaux [105].

II.4 Outils et Méthodologie de QSAR :

II.4.1 Paramètres biologiques :

Les modèles QSAR sont dépendants des données expérimentales utilisées pour leur construction. Le modélisateur doit tenir compte des données à modéliser. Le choix de la base de données est donc une étape très importante dans le développement des modèles QSAR.
Ces données devraient, idéalement, être de grande qualité, ce qui signifie qu'elles devraient être fiables et cohérentes. Il est donc important de les choisir parmi celles présentant des incertitudes faibles afin de limiter les barres d'erreurs expérimentales. De plus, le modélisateur doit s'assurer que les données expérimentales utilisées ont été obtenues selon le même protocole. En effet les conditions expérimentales ont, généralement, une forteinfluence sur les valeurs obtenues [38].

Il faut également que la distribution des données soit la plus homogène et normale que possible, car la plupart des méthodes statistiques sont basées sur ce type de distribution**[39].**
L'efficacité d'un modèle QSAR dépend également du type de molécules qui y sont incluses, plus le modèle présentera des composés de structures proches et similaires, plus il aura de chance d'être performant.

Les données biologiques sont généralement exprimées en logarithmes inverses (log1/C) afin d'obtenir des valeurs mathématiques plus élevées lorsque les structures sont biologiquement très efficaces [40,41]. Quelque exemples de données biologiques, utilisées dans l'analyse QSAR, sont décrits dans le tableau II.1

Tableau II.1 : Types de données biologiques utilisées dans l'analyse QSAR.

Source d'activité	Paramètres biologiques
1. Récepteurs isolés	
Constante de vitesse	Log k
Constante de Michaelis-Menten	Log $1/K_m$
Constante d'inhibition	Log $1/K_i$
2. Systèmes cellulaires	
Constante d'inhibition	Log $1/IC_{50}$
Résistance croisée	Log CR
Données biologiques *in vitro*	Log $1/C$
Mutation de gène	Log TA_{98}
3. Systèmes *in vivo*	
Facteur de bioconcentration	Log BCF
Vitesses de la réaction *in vivo*	Log I (induction)
Vitesses pharmacodynamiques	Log T (clairance totale)

II.4.2 Descripteurs moléculaires :

Un descripteur moléculaire est un paramètre (une valeur numérique) propre à une structure chimique donnée. Ces valeurs peuvent être obtenues expérimentalement ou calculées à partir de la structure de la molécule. Les descripteurs calculés, permettent d'effectuer des prédictions sans avoir à synthétiser les molécules, ce qui est l'un des objectifs de la modélisation moléculaire.

Les descripteurs moléculaires jouent un rôle fondamental dans les études de la relation quantitative structure activité/propriété. Ils sont utilisés en tant que variables indépendantes pour prédire une variable dépendante (activité ou propriété).

L'utilisation des descripteurs moléculaires dans le développement de modèles RQSA/RQSP n'est pas une tâche aisée. Tout d'abord, un très grand nombre de descripteurs moléculaires, de différentes complexités et de conceptions diverses a été introduit au cours des dernières années. Ensuite, pendant ce temps, aucune règle stricte n'a été établie pour la sélection de descripteurs adaptés parmi le grand nombre de descripteurs disponibles. Ce choix a souvent été basé sur l'intuition chimique des chercheurs, ou en se pliant à la tradition. [42]

II.4.2.1 Types de descripteurs :

Les descripteurs moléculaires sont des représentations mathématiques formelles d'une molécule, obtenues par un algorithme bien défini et appliquées à une représentation moléculaire définie ou à une procédure expérimentale bien définie : le descripteur moléculaire est le résultat final d'une procédure logique et mathématique qui transforme l'information chimique codée dans une représentation symbolique d'une molécule en un nombre utile ou le résultat d'une expérience standardisée. [43]

II.4.2.1.a. Descripteurs constitutionnels :

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

_ La masse molaire

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

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

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

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

Ces descripteurs sont très utilisés du fait de leur extrême simplicité non seulement d'un point de vue conceptuel mais surtout calculatoire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution. c-a-d, si on développe des modèles avec ce type de descripteurs seulement, alors que ces derniers peuvent poser problème pour l'interprétation des mécanismes d'interaction mis en jeu pour la propriété étudiée [106].

II.4.2.1.b. Descripteurs topologiques :

Les descripteurs topologiques sont des indices obtenus à partir d'une structure 2D de la molécule, à savoir une simple table de connectivité des atomes dans la molécule. Ils contiennent en leur sein des informations sur la taille globale du système, sa forme globale et ses ramifications [42].

Ces descripteurs s'inspirent de la théorie des graphes appliquée à la table de connectivité qui n'est autre qu'une représentation compacte de la connectivité interatomique au sein de la molécule. Les indices topologiques les plus fréquemment utilisés sont l'indice de Wiener [44], L'indice de Randic [45],L'indice de connectivité de valence de Kier-Hall [46] et l'indice de Balaban [47].

Tableau II.2 : Les descripteurs topologiques calculés dans étude QSAR [48].

Descripteur	Définition
L'indice de Wiener (ω)	$\omega = \frac{1}{2}\sum_{ij} d_{ij} = \frac{1}{2}\sum_{i} s_i$ Est simplement la demi-somme de tous les éléments d'une telle matrice où Si est la somme des distances
L'index J de Balaban (J)	$J = C.\sum_{k} (s_i . s_j)_k^{-0.5}$ Où C = m (1 + u) ; m est le nombre d'arêtes et u est le nombre cyclomatique.

En général, ce type de descripteurs simplifie grandement la représentation de la connectivité chimique au sein de la molécule puisqu'ils ne prennent pas en compte les différences de distances, d'angles et d'ordres de liaison ni même la nature des atomes dans la molécule. Si certains indices ont été développés pour intégrer de manière très approximative ce genre d'informations, ils restent souvent insuffisants pour caractériser l'intégralité des propriétés moléculaires. Finalement, les indices topologiques sont souvent considérés comme des descripteurs convenables d'un point de vue numériques. Cela dit, dans la plupart des cas, l'interprétation des équations QSAR/QSPR qui en résultent n'est pas aisée, puisqu'il est difficile de les relier aux mécanismes sous-jacents [107].

II.4.2.1.c. Descripteurs géométriques :

Ils sont évalués à partir des positions relatives des atomes d'une molécule dans l'espace, ainsi que des rayons et masses atomiques. Celles-ci peuvent être obtenues expérimentalement bien entendu mais le plus souvent par modélisation moléculaire, empirique ou ab initio. Ils sont basés sur l'arrangement spatial des atomes constituant la molécule et sont définis par les coordonnées des noyaux atomiques de la molécule représentée. Ces descripteurs incluent des informations sur la surface moléculaire obtenue par les aires de Van Der Waals et leur superposition. Les volumes moléculaires peuvent être obtenus parles volumes de Van Der Waals [49].

Parmi ces descripteurs, on retrouve le volume, la surface moléculaire, le moment d'inertie ou encore des distances, angles ou angles dièdres particuliers entre atomes dans la molécule, le nombre de liaisons, la surface de Van Der Waals et le volume de Van Der Waals,…ect.[105]

Le volume moléculaire : noté MV, en cm^3, est défini par la formule suivante :

$$MV = \frac{MW}{d} \qquad (1)$$

Avec : MW est le poids moléculaire et d la densité [108].

II.4.2.1.d. Descripteurs quantiques/électroniques :

Afin d'aller plus loin dans la description des structures moléculaires, des caractéristiques supplémentaires de la structure moléculaire peuvent encore être calculées. Ces descripteurs électroniques [50-51] permettent de quantifier différents types d'interactions inter et intramoléculaires de grande influence sur l'activité biologique. Ces descripteurs nécessitent des calculs plus sophistiqués pour la recherche de la géométrie pour laquelle l'énergie est minimale, et fait souvent appel à la chimie quantique qui nous donne accès à des informations supplémentaires telles que des données énergétiques, vibrationnelles et orbitalaires du système [52-53].

Les structures étudiées dans ce travail ont été optimisées en utilisant la base 6-31G(d) de la fonctionnelle B3LYP qui est une sorte de la méthode de théorie de la fonctionnelle de la densité DFT. Le calcul des descripteurs commence par le dessin des molécules dans le logiciel GaussView 5 [54]. Puis l'ouverture de ces structures dans le programme Gaussian 09[55] et ensuite l'exécution de l'optimisation (les calculs). A la fin de ces calculs, des propriétés électroniques seront obtenues.

Parmi ces propriétés, que nous avons utilisées dans nos travaux, on trouve :

-L'énergie totale : Pour une molécule isolée à l'état fondamental, l'énergie totale calculée, notée Et, mesurée en eV, peut être utilisée comme descripteur moléculaire quantique. Cette énergie approximative a été calculée pour une conformation optimisée de la géométrie la plus stable dont la structure d'énergie est minimale [108].

- Le moment dipolaire : noté μ, mesuré en debye (D), mesure la polarité nette moléculaire, et décrit la séparation de charge dans une molécule où la densité d'électrons est partagée inégalement entre les atomes. L'existence d'un moment dipolaire dans une molécule a son origine dans la différence d'électronégativité entre les atomes. La densité électronique est plus élevée au voisinage de l'atome le plus électronégatif. Ceci entraîne une dissymétrie dans la répartition des électrons de liaison. Ainsi, plus le moment dipolaire d'une molécule est élevé, plus la dissymétrie dans la molécule est importante [108].

-L'énergie HOMO : notée E_{HOMO}, mesurée en eV, est le niveau d'énergie le plus élevé dans la molécule qui contient des électrons, il est directement lié au potentiel d'ionisation. Lorsqu'une molécule agit comme une base de Lewis (un doublet d'électrons donneur) dans la formation d'une liaison, les électrons sont alimentés à partir de cette orbite. Il mesure la nucléophilie d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par des électrophiles [56].

-L'énergie LUMO : notée E_{LUMO}, mesurée en eV, est le niveau d'énergie le plus bas dans la molécule qui ne contient pas d'électrons, il est directement lié à l'affinité d'électron. Lorsqu'une molécule agit comme un acide de Lewis (un doublet d'électrons accepteur) dans la formation de liaisons, des doublets d'électrons entrants sont reçus dans cette orbite. Il mesure l'électrophilicité d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par les nucléophiles [56].

- L'électronégativité :

$$\chi = -\mu = -\left(\frac{\partial E}{\partial N}\right)_{v(r)} = -\frac{(E_{LUMO} + E_{HOMO})}{2} \qquad (2)$$

II.4.2.1.e. Descripteurs physico-chimiques :

Les descripteurs physico-chimiques, certains d'entre eux reflètent la composition moléculaire du composé, d'autres représentent le caractère hydrophile ou lipophile de la molécule généralement évalué à partir du coefficient de partage Octanol/eau représenté par le log P. Parmi ces descripteurs, le log P, la réfractivité moléculaire, l'indice de réfraction, la

polarisabilité, la densité, le parachor, la surface de tension, le nombre de donneurs de liaisons hydrogène, le nombre d'accepteurs de liaisons hydrogène,..ect [57].

-Le coefficient de partage Octanol/Eau :

Le transport, le passage à travers les membranes et l'activité pharmacologique d'une molécule peuvent être conditionnés par son partage entre une phase lipidique et une phase aqueuse, c'est-à-dire son caractère hydrophile. Celui-ci peut être quantifié par le coefficient de partage Octanol Eau, noté (log P), qui mesure la solubilité différentielle d'un soluté dans ces deux solvants non miscibles [58]. C'est une mesure importante pour l'identification de la similarité médicamenteuse, selon la règle de Lipinski, les médicaments délivrés par voie orale doivent avoir des valeurs de log P supérieures ou égales à -2 et inférieures ou égales à 5) [59]. Il est défini par la formule suivante :

$$\log \mathrm{P} = \log \frac{[\mathrm{Octanol}]}{[\mathrm{H_2O}]} \qquad (3)$$

[Octanol] et **[H_2O]** sont les concentrations du soluté dans l'Octanol et l'eau.

Les composés qui ont les valeurs de log P > 0 sont dites lipophiles, et les composés qui ont les valeurs de log P <0 sont dites hydrophiles. Si le Log P est positif et très élevé, cela exprime le fait que la molécule est plus soluble dans l'Octanol que dans l'eau, ce qui reflète son caractère lipophile, et inversement, si le Log P est négatif cela signifie que la molécule est hydrophile. Un Log P nul signifie que la molécule est aussi soluble dans un solvant que dans l'autre.

- La réfractivité moléculaire : notée (MR), en m3/mol, est le volume de la substance absorbée par mole de cette substance. Elle est définie par Lorentz-Lorenz [60] par la formule suivante :

$$\mathrm{MR} = \frac{\mathrm{n}^2 - 1}{\mathrm{n}^2 + 2}\frac{\mathrm{MW}}{\mathrm{d}} = \frac{\mathrm{n}^2 - 1}{\mathrm{n}^2 + 2}\,\mathrm{MV} \qquad (4)$$

Où : **MW** est le poids moléculaire ; d est la densité ; n est l'indice de réfraction ; MV est le volume molaire.

La réfractivité moléculaire est également proportionnelle à la polarisabilité αe, par la relation suivante [61]: MR =4/3π N_A αe

Où : $\mathbf{N_A}$ est le nombre d'Avogadro qui est, le nombre de molécules dans une mole de substance, $N_A = 6.022\times 10^{23}$.

-L'indice de réfraction : noté n, est défini par la formule de Lorentz suivante [60]:

$$n = \sqrt{\frac{2\,MR + MW}{MV\ - MR}} \qquad (5)$$

- La polarisabilité : notée (α_e), en (m^3), est l'aptitude à la déformation du nuage électronique de la molécule sous l'influence d'un champ électrique uniforme. C'est l'un des paramètres qui traduisent les propriétés moléculaires liées à l'hydrophobie et par conséquent aux activités biologiques [62-64]. Elle est calculée à partir de la réfractivité molaire ou du volume molaire comme suit :

$$\alpha_e = 0.3964308\ \times MR = 0.3964308\ \times \frac{n^2 - 1}{n^2 + 2}\ MV \qquad (6)$$

- La densité : notée (d), en (kg/m^3), est liée à la masse et la taille de la molécule. C'est le rapport du poids moléculaire MW au volume moléculaire MV [108]:

$$d = \frac{MW}{MV} \qquad (7)$$

-Energie d'hydratation :

La liaison hydrogène (ou liaison H) est de type électrostatique (charge partielle, dipôle) et stérique entre deux groupements d'une même molécule ou de deux molécules voisines.

La liaison hydrogène joue un rôle primordial dans la solubilité des molécules médicamenteuse et leurs interactions avec les récepteurs biologiques [65].

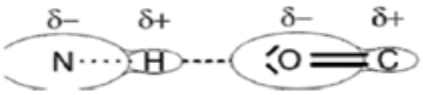


Figure II.1 : La liaison hydrogène

II.4.2.1.f. Descripteurs thermodynamiques :

Ce sont des descripteurs peu utilisés dans les études RQSA/RQSP. Ils peuvent être exprimés par la fonction de partition Q de la molécule utilisée en thermodynamique statistique ainsi que de ses dérivées [66-68]. Cette fonction décrit la façon avec laquelle l'énergie d'un système de molécules est répartie parmi les individus moléculaires. Sa valeur dépend du poids moléculaire, de la température, du volume moléculaire, des distances inter nucléaires, des mouvements moléculaires et des forces intermoléculaires. La fonction de partition est le point le plus commode entre les propriétés microscopiques des molécules indépendantes (niveaux d'énergie, moments d'inertie) avec les propriétés macroscopiques (point de fusion, point d'ébullition, entropie). L'expression de cette fonction s'écrite :

$$Q = Q_{éle} * Q_{trans} * Q_{rot} * Q_{vibr} \qquad (8)$$

Avec :

$Q_{éle} = \Sigma_i g_i \exp(-\varepsilon / kT)$ Fonction de partition électronique ;

$Q_{vibr} = \Pi_i (1 - \exp(-hv_i/kT))^{-1}$ Fonction de partition de vibration ;

$Q_{rot} = (8\pi^2 (8\pi^3 ABC)^{1/2} (kT)^{3/2}) / \sigma h^3$ Fonction de partition de rotation ;

$Q_{trans} = ((2\pi mkT)^{3/2} V) / h^3$ Fonction de partition de translation.

T est la température en °K, **k** est la constante de Boltzmann k=1.38 10^{-23} J/K, gi représente la dégénérescence du niveau d'énergieε_i, h est la constante de Planck

h = 6.62 10^{-34}J.s, $v_i$les fréquences de vibration de la molécule,σ est le degré de symétrie,

A, B et C sont les trois moments d'inertie par rapport aux axes x, y et z, m la masse de la particule, **V** le volume de la molécule.

II.4.3 Sélection des descripteurs :

La recherche de l'ensemble des descripteurs qui forment le bon modèle, exprimant l'activité biologique avec un coût raisonnable de calcul, constitue l'étape déterminante, car le calcul de tous les modèles possibles n'est pas pratique eu égard au nombre élevé des descripteurs (plus de 1600 descripteurs).

Afin d'éviter la formation de modèles dus à la chance un contrôle rigoureux est exigé sur la taille de l'ensemble des descripteurs. Ainsi, l'approche rationnelle de la sélection des

variables permet d'éviter les redondances, de diminuer le coût calculatoire et de trouver les meilleurs sous-ensembles de descripteurs. La procédure de sélection des variables peut être divisée en deux étapes [69].

- ✓ Sélection objective
- ✓ Sélection subjective

II.4.3.1 La sélection objective :

Elle consiste à la sélection des variables en réduisant le nombre de descripteurs sans faire participer la variable dépendante (la réponse biologique) afin de diminuer les corrélations entre les descripteurs.

La suppression d'un descripteur, ayant un pourcentage élevé de valeurs identiques pour l'ensemble des composés, aura lieu au début de la sélection. Il en est de même pour les descripteurs fournissant des informations superflues. Ensuite, le coefficient de corrélation (R) entre les descripteurs est calculé par paires. Un des deux descripteurs est supprimé si leur combinaison possède un coefficient de détermination supérieur au seuil requit (R>0.90). Cette valeur numérique de R est le seuil utilisé pour toutes les applications de sélection en réduisant le nombre de descripteurs sans pour autant perdre de l'information [109].

II.4.3.2 La sélection subjective :

- ➢ **Introduction progressive :**

Cette méthode consiste à incorporer, une à une, les variables au modèle en sélectionnant à chaque étape la variable dont la corrélation partielle avec la grandeur modélisée est la plus élevée [70].

- ➢ **Elimination progressive :**

Cette méthode consiste en l'établissement du modèle avec l'ensemble des descripteurs pour ensuite ne garder que ceux qui permettent l'obtention d'un modèle ayant une bonne corrélation [70].

- ➢ **Sélection pas à pas :**

C'est la combinaison des deux méthodes citées précédemment. Les variables sont incorporées une à une dans le modèle par sélection progressive. Cependant, à chaque étape, on vérifie que les corrélations partielles des variables précédemment introduites sont encore significatives [71].

II.5 La théorie de la fonctionnelle de la densité (DFT) :

La théorie de la fonctionnelle de la densité est basée sur le postulat propose par Thomas et Fermi qui dit que les propriétés électroniques peuvent être décrites en terme de fonctionnelles de la densité électronique, en appliquant localement des relations appropriées à un système électronique homogène [72].

Hohenberg et Kohn, en 1964 [73,74], ont repris la théorie de Thomas-Fermi et ont montré qu'il existe une fonctionnelle de l'énergie E[ρ(r)] associée à un principe variationel, ce qui a permis de jeter les bases de la théorie de la fonctionnelle de la densité .

La théorie de la fonctionnelle de la densité est basée sur le théorème Hohenberg-Kohn [75], qui établit que l'énergie d'un système dans son état fondamental est une fonctionnelle de la densité électronique de ce système, ρ(r), et que toute densité, ρ'(r), autre que la densité réelle conduit nécessairement à une énergie supérieure. Ainsi contrairement aux méthodes précédentes, la théorie de la fonctionnelle de la densité ne consiste pas à chercher une fonction d'onde complexe, ψ, à 3N-dimensions décrivant le système à étudier, mais plutôt une simple fonction à trois dimensions : la densité électronique totale ρ [76]. Il existe trois types de fonctionnelles énergies d'échange-corrélation: les fonctionnelles locales, les fonctionnelles à correction du gradient et les fonctionnelles hybrides.

II.6 Méthodes statistiques :

Par définition, la statistique est « la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation de phénomènes dans lesquels le hasard intervient (phénomène aléatoire) ». Par conséquent, l'objectif principal de la statistique est de maîtriser au mieux l'incertitude pour extraire des informations utiles des données, par l'intermédiaire de l'analyse des variations dans les observations. En outre, l'analyse des données est utilisée pour décrire, comprendre et gérer les phénomènes étudiés, faire des prévisions et prendre des décisions [108].

Les techniques statistiques ou chimio-métriques constituent la base mathématique de la construction d'un modèle QSAR. L'élaboration ce modèle n'est pas une chose facile. La première difficulté réside dans la différence d'échelles existant entre les données à corréler. La structure étant à une échelle moléculaire alors que les activités /propriétés à prédire sont à une échelle macroscopique. Un des problèmes importants réside également dans le traitement

de données. En fait, de nombreux outils existent et il s'agit de trouver le moyen le plus adapté pour obtenir un modèle fiable à partir des données disponibles [110].
Les principaux outils statistiques pour obtenir un modèle sont :

- Régression linéaire simple (SLR)
- Régression linéaire multiple (MLR)
- La régression non linéaire multiple (MNLR)
- Analyse en composantes principales (ACP)
- La régression par les moindres carrés partiels (PLS)

II.6.1 Régression linéaire simple (SLR) :

Cette méthode effectue comme un calcul de régression linéaire standard dans la génération du modèle QSAR sous la forme d'équations qui incluent un seul descripteur indépendant x et y en tant que variable dépendante. Cette technique s'avère très prometteuse pour générer des relations de structure et d'activité en explorant certains des descripteurs les plus importants utilisés pour gouverner l'activité, alors que certaines des interactions de plusieurs descripteurs ont été négligées[111]. La régression linéaire simple peut être exprimée par l'équation II.9:

$$\mathbf{y = a + bx} \qquad \mathbf{(9)}$$

Où **y** est la variable dépendante, **x** est la variable indépendante, **a** est la constante et **b:** est le coefficient de régression (2).

II.6.2 La régression linéaire multiple (MLR) :

La régression linéaire multiple MLR est l'une des méthodes de modélisation les plus populaires grâce à sa simplicité d'utilisation et facilité d'interprétation. L'avantage important de la régression linéaire multiple est qu'elle est très transparente, puisque l'algorithme est disponible, et que les prédictions peu0vent être réalisées facilement [77].

L'analyse de régression linéaire multiple repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante *Y* et une série de *n* variables indépendantes *Xi*. Pour les études de régression multiple, le nombre de variables doit être inférieur ou égal au nombre d'individus (molécules). L'objectif est d'obtenir une équation de la forme suivante [105] :

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n \quad (10)$$

Cette équation est linéaire par rapport aux paramètres (coefficients de régression) $a_0, a_1 \ldots, a_n$.

La détermination de l'équation II.10 à partir d'un ensemble de données de p échantillons revient à résoudre un système de p équations :

$$y_1 = a_0 + a_1x_{1,1} + a_2x_{2,1} + \cdots + a_nx_{n,1} + b_1$$

$$y_2 = a_0 + a_1x_{1,2} + a_2x_{2,2} + \cdots + a_nx_{n,2} + b_2$$

$$\vdots$$

$$y_p = a_0 + a_1x_{1,p} + a_2x_{2,p} + \cdots + a_nx_{n,p} + b_p$$

Ou les résidus bi représentent l'erreur du modèle, constituée par l'incertitude sur la variable dépendante Yi d'une part, sur les variables indépendantes Xi d'autre part, mais aussi par les informations contenues dans les variables indépendantes mais non exprimées via les variables dépendantes.

Ce système d'équation peut être donné sous la forme matricielle suivante :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & \cdots & x_{n,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & \cdots & x_{n,2} \\ \vdots & \vdots & & & & \\ \vdots & \vdots & & & & \\ 1 & x_{1,p} & x_{2,p} & \cdots & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ \vdots \\ a_p \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_p \end{pmatrix}$$

Soit de manière condensée :

$$Y = XA + B$$

Où Y, X, A et B représentent respectivement le vecteur de propriété, la matrice des attributs (descripteurs), la matrice des coefficients et la matrice des erreurs de régression [113].

II.6.3 La régression non linéaire multiple (MNLR):

La régression non linéaire multiple MNLR est une méthode non linéaire (exponentielle, logarithmique, polynomiale, …) qui permet de déterminer le modèle mathématique qui permet d'expliquer non-linéairement au mieux la variabilité d'une propriété ou d'une activité y en fonction des descripteurs moléculaires. Dans l'ensemble de nos travaux nous avons utilisé le modèle polynomial en nous basant sur les descripteurs proposés par le modèle linéaire qui seront élevés à la puissance 2 selon l'équation suivante :

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + b_i x_i^2 \qquad (11)$$

Avec : **y**: est la variable dépendante (à expliquer ou à prédire) ; $\mathbf{x_i}$sont les variables indépendantes (explicatives) ; **i:**est le nombre de variables explicatives ;$\mathbf{a_0}$: est la constante de l'équation du modèle ;$\mathbf{a_i}$et $\mathbf{b_i}$sont les coefficients de descripteurs dans l'équation du modèle[108].

II.6.4 Analyse par composantes principales (ACP) :

L'analyse en composantes principales **(**ACP**),** ou *principal component Analysis*(*PCA*) en anglais, qui est sans doute la technique exploratoire la plus répandue pour décrire les données d'entrée, rechercher d'éventuels aberrations et à connaître les corrélations entre les variables d'entrée. Elle permet d'analyser et de visualiser un jeu de données concernant des individus décrits par plusieurs variables quantitatives. Elle est non supervisée, c'est-à-dire sans phase d'apprentissage [78-80].

C'est une méthode statistique qui permet d'explorer des données dites multivariées (données avec plusieurs variables). Chaque variable pourrait être considérée comme une dimension différente.

Si vous avez plus de 3 variables dans votre jeu de données, il pourrait être très difficile de visualiser les données dans un "hyper-espace" multidimensionnel.

L'analyse en composantes principales est utilisée pour extraire et de visualiser les informations importantes contenues dans une table de données multivariées. L'ACP
Synthétise cette information en seulement quelques nouvelles variables appelées composantes principales. Ces nouvelles variables correspondent à une combinaison linéaire des variables

originales. Le nombre de composantes principales est inférieur ou égal au nombre de variables d'origine. L'information contenue dans un jeu de données correspond à la variance ou l'inertie totale qu'il contient. L'objectif de l'ACP est d'identifier les directions (axes principaux ou composantes principales) le long desquelles la variation des données est maximale.

En d'autres termes, l'ACP réduit les dimensions de données multivariées à deux ou trois composantes principales, qui peuvent être visualisées graphiquement, en perdant le moins possible d'information. La compréhension des détails de l'ACP nécessite des connaissances de l'algèbre linéaire [105].

II.6.5 La méthode de régression des moindres carrés partiels (PLS) :

La régression par les moindres carrés partiels (PLS) est une technique qui diminue le nombre de descripteurs à un plus petit ensemble de composantes non corrélées et qui effectue la régression par les moindres carrés sur ces composantes, plutôt que sur les données initiales. La fonctionnalité PLS est particulièrement utile lorsque les descripteurs sont fortement colinéaires, ou lorsqu'il y a plus de descripteurs que d'observations et que la régression sur les moindres carrés échoue complètement ou produit des coefficients avec des erreurs élevées.

Les moindres carrés partiels (PLS) ne supposent pas que les descripteurs sont fixes, à la différence de la régression multiple. Ainsi, les mesures des descripteurs tolèrent des erreurs, ce qui signifie que l'analyse PLS gère mieux l'incertitude des mesures [81-82].

II.7 Coefficients et tests statistiques standards :

❖ Coefficient de corrélation R (et coefficient de détermination R^2) :

C'est l'indicateur statistique le plus répandu qui évalue la part de la variance de l'activité /la propriété cible expliquée par le modèle.

La qualité du modèle est souvent visualisée sur un diagramme de dispersion, sur lequel sont portées les valeurs calculées de la propriété (activité biologique), en fonction de celles expérimentales. La qualité de la modélisation est meilleure quand les points de ce graphique sont proches de la droite d'ajustement [83]. L'ajustement des points à cette droite peut être évalué par le coefficient de détermination R^2 :

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (12)$$

y: La valeur expérimentale de l'activité

$\hat{y}$: La valeur calculée de l'activité

$\bar{y}$: La valeur moyenne des valeurs calculées de l'activité.

Limitation du coefficient de détermination R^2 :

R^2 qui varie entre 0 et 1, mesure la proportion de variation totale de Y autour de la moyenne expliquée par la régression. Plus la valeur de R^2 sera proche de 1 (cas idéal) et plus les valeurs prédites et observées sont corrélées. Un R^2 faible signifie que le modèle a un faible pouvoir explicatif et les descripteurs (certains d'eux) sont sans effet sur la réponse.

Le jugement sur la valeur de R^2 est très subjectif. Bien que ce coefficient soit très facile à comprendre, il faut se garder d'y attacher trop d'importance car il est loin de fournir un critère suffisant pour juger de la qualité d'une régression. Il n'est pas recommandé d'utiliser R^2 pour comparer des modèles avec un nombre différent de descripteurs, le coefficient R^2 nous dira toujours de choisir le modèle avec le plus grand nombre de descripteurs car son R^2 sera plus important (on projette sur un espace plus grand), même si les variables sont sans effets sur la réponse Y.

La valeur de R^2 dépend de la taille de l'échantillon et le nombre de variables prédictives dans l'équation. Il garde la même valeur ou augmente lors d'une nouvelle variable de prédiction est ajoutée à l'équation de régression, même si la variable ajoutée ne contribue pas à la réduction de la variance inexpliquée. Par conséquent, un autre paramètre statistique peut être utilisé, appeléR^2ajusté (R^2ajusté) [112].

❖ Coefficient de détermination ajuste R^2ajuste :

Ce coefficient est utilise en régression multiple par ce qu'il tient compte du nombre de paramètres (descripteurs) du modèle.

$$R^2_{adj} = \frac{R^2(n-1)-p}{n-p-1} \quad (13)$$

Avec : **n** est le nombre des observations (les molécules) ; **p** est le nombre de variables indépendantes (les descripteurs) ; $\mathbf{R^2}$est le coefficient de détermination du modèle.
Cette formule indique notamment que R^2_{adj} est toujours inferieur à R^2, et ceci d'autant plus que le modèle contient un grand nombre de prédicteurs (descripteurs) [112].

❖ Critère de validation croisée : PRESS

La somme des erreurs quadratiques de prédiction « prédiction sum of squares»(PRESS) est définie par :

$$\mathrm{PRESS} = \sum_{i=1}^{n} \varepsilon_{(i)}^2 \qquad (14)$$

Ce critère permet de sélectionner les modèles ayant un bon pouvoir prédictif. (On cherche toujours le PRESS le plus petit)[112].

❖ Le test de Fisher F :

L'indice de Fisher *F-test* est employé afin de mesurer le niveau de signifiance statistique du modèle à « x% » (le niveau usuel est 95%), c'est-à-dire la qualité du choix du jeu de paramètres. La conclusion obtenue ne doit pas nous faire penser que la corrélation a « x % » de chances d'être vraie mais seulement que la corrélation est vérifiée pour « x% » des composés pris pour référence et qu'une abstraction est faite pour les autres.
Hypothèses :
H_0: les variances des échantillons sont homogènes
H_1: les variances des échantillons ne sont pas homogènes
La valeur à calculer est :
On calcule le F (observé) à partir de la formule :

$$\mathrm{F(observé)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum\left(y_i - \hat{y}_i\right)^2}\frac{n-p-1}{p} \qquad (15)$$

Avec : **F** est l'indice de Fisher ; $\mathbf{y_i}$ et $\hat{\mathbf{y}}_\mathbf{i}$ sont, respectivement, les valeurs observées et calculées de la variable dépendante ;$\hat{\mathbf{y}}$ est la valeur moyenne des valeurs prédites ; est le nombre des observations (les molécules) ;**p** est le nombre de variables indépendantes (les descripteurs).

Après le calcul de F (observé) on le compare avec le F théorique obtenu à partir des tables statistiques usuelles (la table de Fisher).

- Si F observé est plus grand que le F théorique : refus de l'hypothèse nulle H_0 et cela signifie que les variances des échantillons sont trop différentes pour être considérées comme homogènes.
- Si F observé est plus petit que le F théorique : acceptation de l'hypothèse nulle H_1 et cela signifie que les deux variances ont des valeurs suffisamment proches pour qu'on accepte l'idée qu'elles soient homogènes [108].

- résultats permettent d'interpréter les tables d'analyse de variance complètes fournies par tout logiciel mettant en œuvre la régression linéaire.

Représentation de l'analyse de la variance :

Toutes ces quantités sont présentées habituellement sur un tableau appelé table d'analyse de variance ou table d'ANOVA (Analysis Of Variance) faisant apparaitre les sources de variation : le total (en 3ème ligne de la table) qui se décompose en deux parties :
La partie modèle et la partie erreur. A chaque source de variation correspond un nombre de degrés de liberté (DF) respectivement égal à n-1, p, n-p-1 (ou n est le nombre d'observations, p le nombre de variables répresseurs (la variable X_0, constante égale à 1, correspondant au paramètre β_0 , n'est pas comprise). Nous présentons le tableau général de l'analyse de variance, tous les logiciels effectuant des calculs de régression, donnent comme sortie ce tableau suivant [112].

Tableau II.3: Table d'analyse de variance.

Source de Variation	*Degrés de liberté DF*	*Somme des carrés SS*	*Moyenne des carrées MS*
Model	p	$\sum(\hat{Y}_i - \overline{Y})^2$	$\frac{\sum_i^n (\hat{Y}_i - \overline{Y})^2}{p}$
Error	n-1-p	$\sum(Y_i - \hat{Y}_i)^2$	$\frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-p-1}$
Total	n-1	$\sum_i^n (Y_i - \overline{Y})^2$	$\frac{\sum_i^n (\hat{Y}_i - \overline{Y})^2}{p} + \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-p-1}$

DF : degré de liberté

SS : Sum of squares : somme carrée des écarts

MS : Mean square, est le rapport SS/DF

MSE: Mean Square Errer La représentation géométrique de tous ces écarts est donnée

❖ Déviation standard (SD) :

La fiabilité de la prédiction de la variable dépendante peut être évaluée également par la valeur déviation standard (SD)[114].

$$SD = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - p - 1}} = \sqrt{MSE} \qquad (16)$$

Avec: $\mathbf{y_i}$ et $\hat{\mathbf{y}}_\mathbf{i}$ sont, respectivement, les valeurs observées et calculées de la variable dépendante ; **n** est le nombre des observations ;**p** est le nombre de variables indépendantes .

Déviation standard SD (appelé aussi L'estimation de l'erreur-type s) est une mesure de la dispersion des valeurs observées de la variable dépendante sur la droite de régression . Les petites valeurs de SD signifient un bon ajustement statistique du modèle et une forte fiabilité de la prédiction [112].

❖ Le facteur d'inflation de la variance VIF :

C'est un paramètre qui permet de détecter la colinéarité entre les descripteurs utilisés dans un modèle statistique, Il est défini par [108] :

$$\mathrm{VIF}(x_k) = \frac{1}{1 - r_k^2} \qquad (17)$$

Avec : **$VIF(x_k)$** est la valeur du facteur d'inflation de la variance pour le descripteur **x_k** (k=1 ;2 ;… ;5), et **r^2_k** est le coefficient de corrélation carrée résultant de la régression du descripteur x_k sur le reste des descripteurs. Une valeur de VIF ≥ 10 signifie la présence d'une forte colinéarité entre les descripteurs, l'absence de la multicolinéarité est signifiée par une valeur de VIF < 10.

On peut aussi confirmer la présence ou l'absence de la forte multicolinéarité par une autre grandeur, similaire au VIF, appelée facteur de tolérance TF donnée par l'équation (Equ II. 18) :

$$\mathrm{TF}(x_k) = \frac{1}{\mathrm{VIF}(x_k)} = 1 - R^2_{x_k} \qquad (18)$$

Les valeurs de TF varient dans l'intervalle 0< TF < 1. Les valeurs de TF > 0,5 et TF < 0,5 correspondent à la présence et à l'absence de fortes multicolinéarité entre les descripteurs respectivement [109].

❖ Le test de Student :

L'indice de Student (le *t-test* de Student) est employé afin d'évaluer la pertinence des descripteurs dans un modèle. Il s'agit de tester l'hypothèse considérant le descripteur comme non significatif. Pour une régression multilinéaire, cela revient à supposer le coefficient qui lui est associé comme nul.

$$|t_i| = \left|\frac{a_i}{s(a_i)}\right| > t^{n-p-2}_{1-\frac{\alpha}{2}} \qquad (19)$$

Avec: **t_i** est le *t-test* pour le descripteur « i » ; **a_i** est le coefficient associé au descripteur « i » dans le modèle ; **$s(a_i)$** est l'erreur type associé au descripteur « i » ; **α** est l'intervalle de

confiance ;**n** est le nombre des observations (les molécules) ;**p** est le nombre de variables indépendantes (les descripteurs).

Cette hypothèse est rejetée (avec un intervalle de confiance α) si le ratio t_i entre a_i et son erreur type s (a_i) atteint la valeur du fractile d'ordre (1- α/2) de la loi de Student à (n-p-2) degrés de liberté.

L'indice de Student (le t-test de Student) est employé aussi pur évaluer la significativité du modèle complet. Le test s'écrit : H_0 : r = 0 et H_1 : r≠0

Si le coefficient de corrélation est différent de zéro on rejette l'hypothèse H_0 (l'hypothèse nulle) et on accepte H_1 donc le modèle est significatif.

Sous H_0, la loi de Student à (n-p-1) degré de liberté $\mathbf{t_{calc}}$ s'écrit :

$$t_{calc} = \left[\frac{r}{\sqrt{\frac{1-r^2}{(n-p-1)}}} \right] \qquad (20)$$

On rejette H_0 d'après (l'hypothèse nulle) lorsque :

$$t_{calc} > t_{\left(1-\frac{\alpha}{2}\right),(n-p-1)}$$

$\mathbf{t_{(1-\alpha/2),(n-p-1)}}$ **:**La valeur de la loi de Student, à (n-p-1) degré de liberté, à une Probabilité (1-α/2) [84-90].

II.8 Validation du modèle :

Après l'obtention de l'équation du modèle, outre la stabilité et la qualité de l'ajustement du modèle, il est également important d'estimer la puissance et la validité du modèle avant de l'utiliser pour prédire l'activité biologique. La validité consiste à établir la fiabilité et la signification de la méthode pour un usage particulier. Par conséquent, la validation d'un modèle QSAR doit être effectuée [91].

Cette qualité est vérifiée par ce que l'on appelle validation. Sa robustesse, c'est-à-dire l'influence des composés de la série d'apprentissage sur le modèle, est estimée par des méthodes de validation interne. Afin d'estimer son pouvoir prédictif, des données expérimentales supplémentaires sont nécessaires afin de déterminer la capacité du modèle à

prédire ces valeurs c'est ce que l'on appelle validation externe. Enfin, il est important de savoir quel type de molécules utilisées avec quel modèle. On parle alors de domaine d'applicabilité [112].

II.8.1 Validation interne :

Afin de déterminer la stabilité prédictive d'un modèle et de tester l'influence de chaque échantillon (composé) sur le modèle final, des procédures de validation croisée (en anglais : cross-validation) sont souvent utilisées [92].

Généralement, Ces techniques de validation permettent l'évaluation de la robustesse du modèle, autrement dit la stabilité des paramètres du modèle RQSA/RQSP vis-à-vis des molécules du jeu d'entraînement. Cela dit, qu'elles ne permettent en aucun cas de démontrer le pouvoir prédictif des modèles [92,93].

La validation interne est souvent la technique la plus employée, dans les études QSAR pour déterminer la stabilité du modèle et de tester l'influence de chaque échantillon de l'ensemble d'apprentissage sur le modèle final [93]. Pour ce faire, on emploie les techniques de la validation croisée (cross validation CV) leave-one-out et leave-*n*-out.

Dans ces dernières années, d'autres méthodes sont utilisées pour faire la validation interne, tel que la hasardisation de la réponse (Y-Randomization).

Le principe de ces méthodes consiste à extraire un certain nombre de molécules du jeu d'apprentissage et à construire un nouveau modèle avec les molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent). Ce nouveau modèle est alors utilisé pour la phase de prédiction sur les molécules retirées. Ce processus est ensuite répété pour retirer et prédire les valeurs de toutes les molécules du jeu d'entraînement .Le coefficient de corrélation Q^2 (ou R^2_{cv}) entre les activités ainsi calculées et les activités observées exprime le pouvoir de prévision interne du modèle, plus la valeur du coefficient se rapproche de 1 plus le pouvoir de prévision sera meilleur. Pour que le modèle soit acceptable le pouvoir de prévision interne doit être supérieur à 0,5 [94].

II.8.1.1 La procédure leave-n-out :

Cette procédure consiste à extraire chaque fois un certain nombre *n* de molécules de l'ensemble initial d'apprentissage à *k* molécules et à construire un nouveau modèle avec les *k-n* molécules restantes en utilisant les descripteurs choisis (les mêmes descripteurs utilisés dans la construction du modèle original) [95].

Le nouveau modèle est ensuite utilisé pour la prédiction sur les *n* molécules retirées. Ce processus est ensuite réitéré *p* fois pour retirer et prédire les valeurs de toutes les molécules de l'ensemble d'entraînement [92](Figure II. 2)

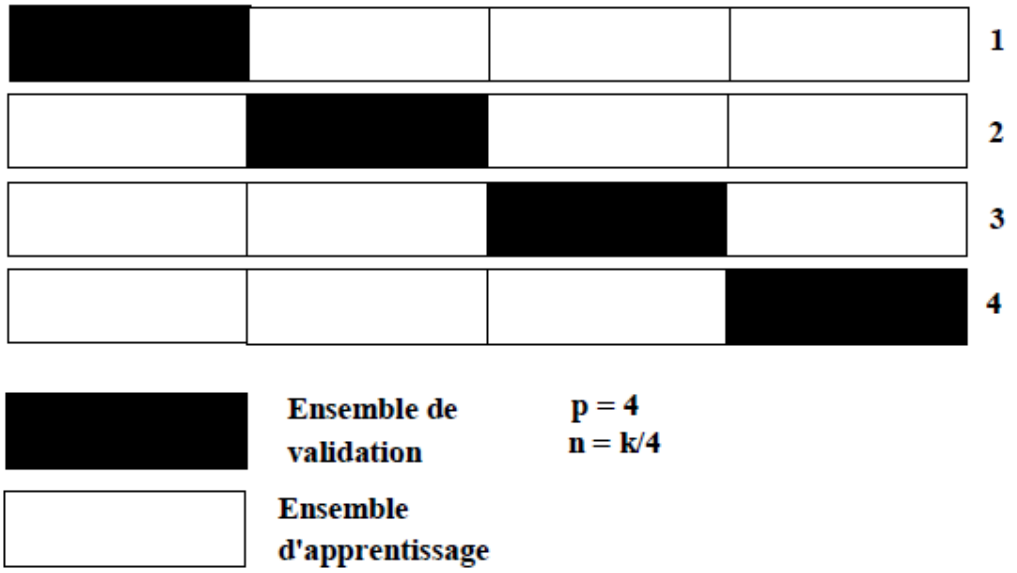


Figure II.2: principe de la validation croisée leave-*n*-out.

II.8.1.2 La procédure leave-one-out:

Pour un nombre *N* d'exemples d'apprentissage, on retire à chaque itération un exemple I de l'ensemble d'apprentissage initial. Une série d'apprentissage est réalisée pour les *N-1* molécules restantes et la molécule retirée est prédite par le modèle formé [92].

La performance des modèles de régression est estimée avec les paramètres statistiques de la validation croisée décrits par les équations ci-dessous :

L'écart type de la validation croisée $S_{CV\text{-}LOO}$:

$$S_{CV-LOO} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_{icv})^2} \qquad (21)$$

Le coefficient de détermination de la validation croisée q^2 (ou r^2_{cv}) :

$$Q^2_{cv-loo} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2 - \sum_{i=1}^{N}(y_i - y_{icv})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (22)$$

Où :

y_{icv}**:** Représente la valeur de l'activité de la molécule i calculée par la méthode de la validation croisée .

$\bar{y}$**:** La valeur moyenne de l'activité,

y_i**:** Les valeurs observées de l'activité

II.8.1.3 Test de randomisation :

Le test de randomisation permet d'affirmer que la corrélation de chance ne joue aucun rôle durant le développement du modèle. Ces tests sont exécutés quantitativement avec les modèles de QSAR et qualitativement avec les modèles issus de la classification. Les observations sont aléatoirement désorganisées dix fois. C'est-à-dire que la colonne des observations (propriétés) sera changée aléatoirement, en revanche la colonne des descripteurs reste inchangée. A la fin, on obtient dix modèles avec des caractéristiques statistiques spécifiques [96].

La supposition sous-jacente de l'essai de randomisation est la suivante :
Si les capacités prédictives du modèle ne sont pas dues aux corrélations de chance, alors la désorganisation aléatoire des observations conduira à des modèles (quantitatifs ou qualitatifs) de prévisions faibles, et vice versa.

II.8.2 Validation externe :

Cette méthode consiste à prédire la propriété/activité d'une série de molécules appelée généralement série de test qui ne sont pas dans la série de développement du modèle, cette validation est caractérisée par les paramètres R^2(test) R^2_{CV} (test). Récemment plusieurs études [93,97] ont montré l'insuffisance des paramètres R^2, R^2_{CV} pour vérifier le pouvoir prédictif des modèles QSAR. Par conséquent, d'autres paramètres doivent être vérifiés pour cet objectif. Ces paramètres sont connus sous le nom « critères de validation externe » ou souvent appelés « critères de Trophsa » (Tropsha criteria) [93].

Critères de validation Externe (série de test) :

- $R^2_{CV} > 0.5$ (critère 1)
- $R^2_{Pred} > 0.6$ (critère 2)
- $R^2 - R^2_0 / R^2 < 0.1$ et $0.85 \leq k \leq 1.15$ (critère 3)
- $R^2 - R'^2_0 / R^2 < 0.1$ et $0.85 \leq k' \leq 1.15$ (critère 4)
- $\left| R_0^2 - R'^2_0 \right| \leq 0.3$ (critère 5)

Avec :

- **R^2** : Coefficient de corrélation pour les molécules de la série de test.
- **R^2_0** : coefficient de corrélation entre les valeurs prédites et expérimentales pour la série de test.
- **R'^2_0**: coefficient de corrélation entre les valeurs expérimentales et prédites pour la série de test.
- **k** : est la constante de la droite (à l'origine) de corrélation (valeurs prédites en fonction des valeurs expérimentales)
- **k'** : est la constante de la droite (à l'origine) de corrélation (valeurs expérimentales en fonction des valeurs prédites)

II.9 Domaines d'application :

Un modèle RQSA/RQSP ne peut pas être considéré comme un modèle universel, parce qu'il est développé sur un nombre limité de composés qui ne couvrent pas tout l'espace chimique. Par conséquent, l'activité/propriété prédite d'un composé, chimiquement dissimilaire au jeu d'apprentissage, ne pourra pas être considérée fiable [98,99].

Un modèle idéal est celui qui est capable de prédire l'acticité ou la propriété de n'importe quelle molécule imaginable. Cependant cela est souvent loin d'être possible. La taille limitée du jeu d'entrainement rend l'espace chimique des modèles construits limité. Et par conséquent, lorsqu'une molécule se situe en dehors de cet espace chimique, la prédiction ne sera plus fiable [100].

Il existe plusieurs méthodes pour la détermination du domaine d'applicabilité d'un modèle RQSA/RQSP, parmi ces méthodes on trouve la méthode de leviers (en anglais : *leverage*) qui est la plus utilisée, elle est basée sur la variation des résiduels standardisés de la

variable dépendante avec la distance entre les valeurs des descripteurs et leurs moyennes appelée leviers h_i . Si un composé a un résiduel et un levier qui dépasse le seuil h*=3p/n (où **p** est le nombre de descripteurs plus 1 et n le nombre d'observations), ce composé est considéré en dehors du domaine d'applicabilité du modèle élaboré.

Le domaine d'applicabilité sera discuté à l'aide du diagramme de Williams qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers h_i **.** [92,101]

Pour chaque composé *i* dans l'espace original des variables indépendantes (xi), la valeur de h_i est calculée par la relation suivante [102]:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i = 1,2, \dots, n) \qquad (23)$$

Avec : xi est le vecteur ligne des descripteurs du composé i ; X (n*k-1) est la matrice du modèle déduit des valeurs des descripteurs de l'ensemble de d'entrainement ; L'indice T désigne la matrice transposée. La valeur critique du levier (h*) est fixée à :

$$h^* = \frac{3(K+1)}{n} \quad [103] \qquad (24)$$

Avec **n** est le nombre de composés utilisés de test ; **k** est le nombre des descripteurs du modèle.

Si **hi<h*:** la probabilité d'accord entre les valeurs mesurée et prédite du composé « i » est aussi élevée que celle des composés de la base de données. Les composés avec **hi>h*** renforcent le modèle quand ils appartiennent à l'ensemble d'entrainement, mais auront, sinon, des valeurs prédites douteuses sans pour autant être forcément aberrantes, les résidus pouvant être bas [104].

II.10 Application de QSAR :

Il existe un grand nombre d'applications des modèles QSAR en milieu industriel, dans le domaine de la recherche universitaire, en économie, en prévision météorologique… etc [48].

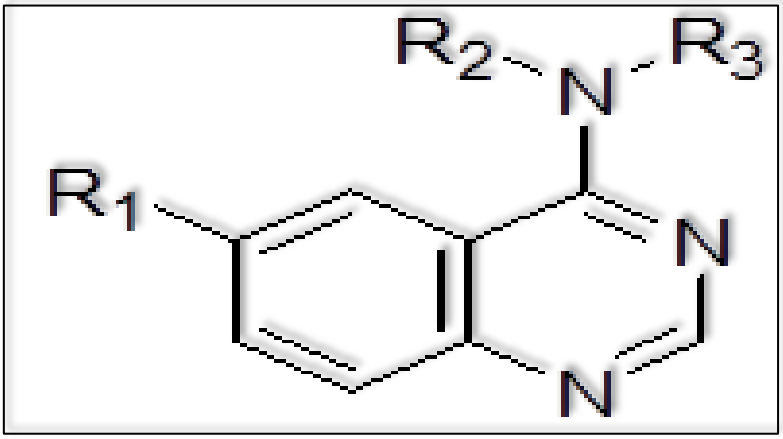
Un petit nombre d'utilisations potentielles sont énumérées ci-dessous [48] :

- L'identification rationnelle des nouvelles pistes avec une activité pharmacologique, biocide ou pesticide.
- L'optimisation de l'activité pharmacologique, biocide ou pesticide.
- La conception rationnelle de nombreux autres produits tels que les agents tensio-actifs, les parfums, les colorants et les produits chimiques fins.
- L'identification des composés dangereux aux premiers stades du développement du produit ou le criblage des inventaires des composés existants.
- La conception de la toxicité et des effets secondaires dans les nouveaux composés.
- La prédiction de la toxicité pour l'homme par une exposition délibérée, occasionnelle et professionnelle.
- La prédiction de la toxicité pour les espèces environnementales.
- La sélection de composés ayant des propriétés pharmacocinétiques optimales, que ce soit la stabilité ou la disponibilité dans les systèmes biologiques.
- La prédiction d'une variété de propriétés physico-chimiques des molécules (qu'il s'agisse de produits pharmaceutiques, de pesticides, de produits personnels, de produits de chimie fine, etc.).
- La prédiction du devenir des molécules libérées dans l'environnement.
- La rationalisation et la prédiction des effets combinés des molécules, que ce soit dans des mélanges ou des formulations.

La caractéristique clé du rôle des technologies in silico dans tous ces domaines est que les prédictions peuvent être faites à partir de la seule structure moléculaire.

Chapitre 03 :

Etude quantitative des propriétés QSAR d'une série de derivés de 6-arylquinazolin-4-amine



III.1 INTRODUCTION :

Les études quantitatives structure-activité (QSAR) constituent une méthode puissante pour la conception des composés biologiquement actifs ainsi que pour la prévision de l'activité selon les propriétés physiques et chimiques résultantes [42], [115]-[117].

Comme toute activité biologique, une compréhension de l'inhibition de DYRK1A peut être obtenue en utilisant l'approche QSAR. Ces méthodes sont basées sur la déduction de l'activité biologique de descripteurs reflétant les propriétés structurelles des molécules. Cependant, il existe un grand nombre de descripteurs et seuls quelques-uns s'avèrent pertinents, le fait que leur utilisation permette d'obtenir des modèles QSAR efficaces. Afin de déterminer les descripteurs les plus pertinentes pour la série 6-arylquinazolin-4-amine, il est nécessaire de déterminer les méthodes d'identification des descripteurs appropriées. En utilisant d'une variété de techniques statistiques pour produire des modèles QSAR.

L'objectif de ce travail est d'étudier et de prédire l'activité de l'inhibition DYRK1A pour 32 de dérives de 6-arylquinazolin-4-amine, en utilisant des descripteurs ciblés qui peuvent expliquer le mécanisme de cette activité étudiée et avec le respect de toute la méthodologie d'une étude QSAR .

III.2 Matériels et méthodes :

III.2.1 Ensembles des données:

Pour cette étude, nous avons choisi de travailler sur 32 molécules de la série 6-arylquinazolin-4-amine (tableau III.1) [118] **.** Le logiciel ChemDrew[119] a été utilisé pour dessiner les composés.

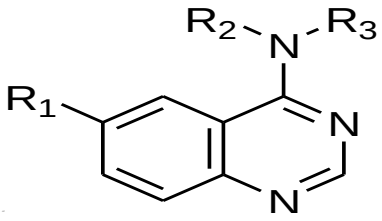
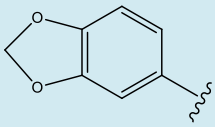
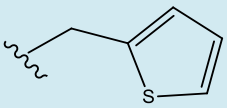
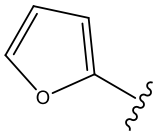
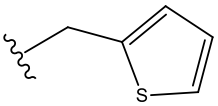
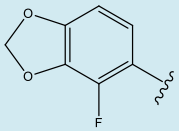
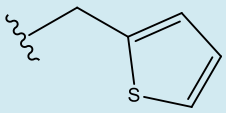
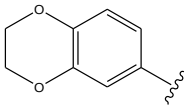
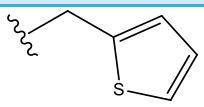
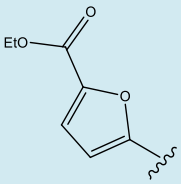
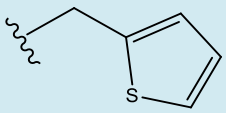
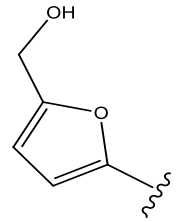
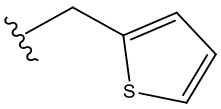
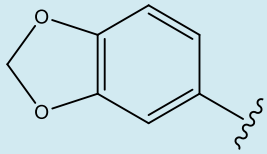
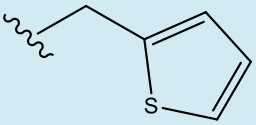
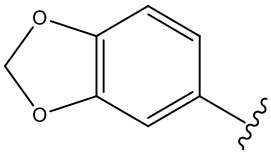
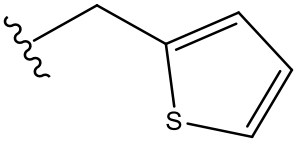
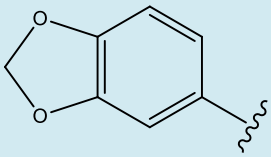
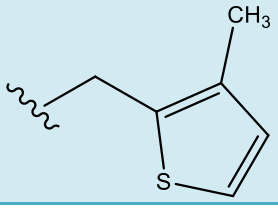
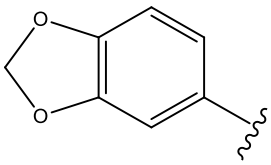
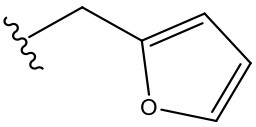
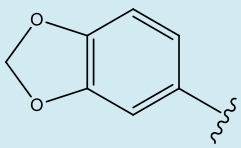
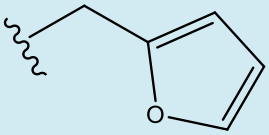
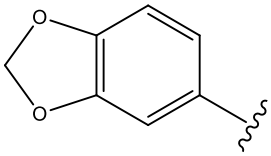
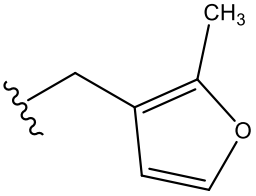
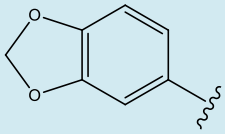
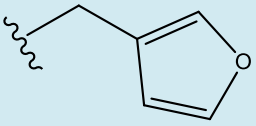
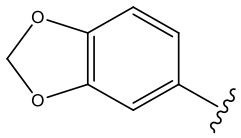
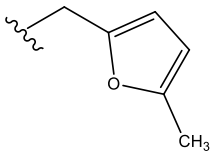
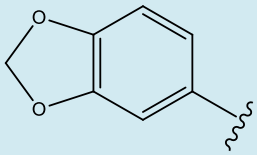
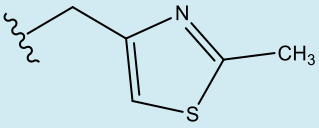
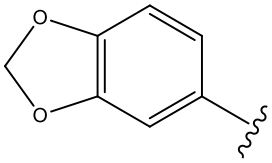
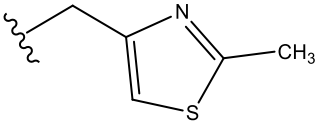
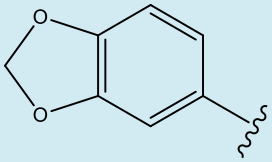
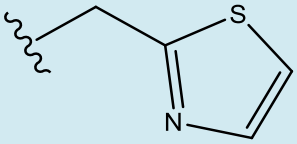
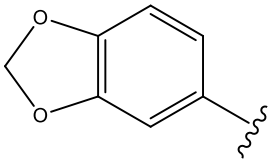
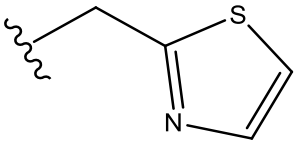
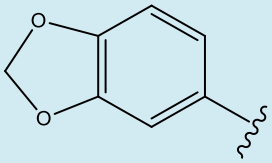
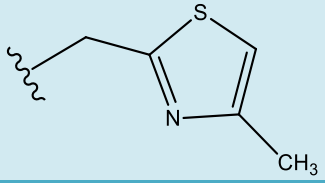
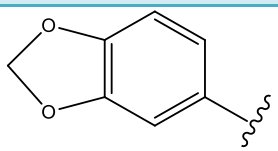
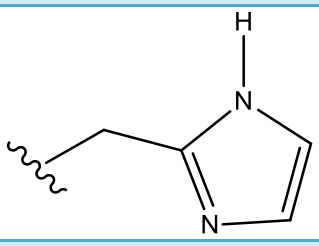
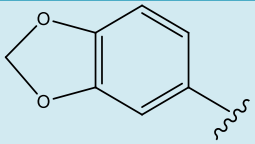
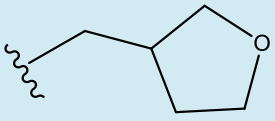
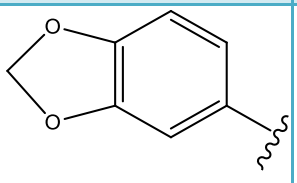
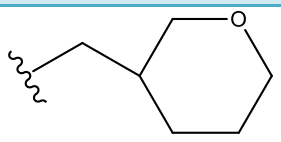


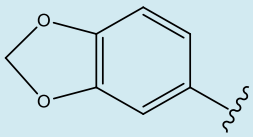
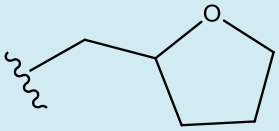
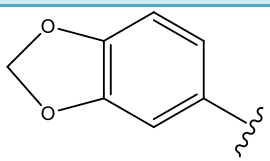
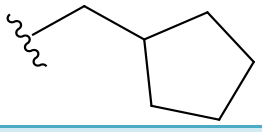
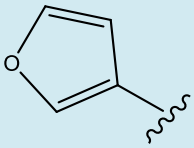
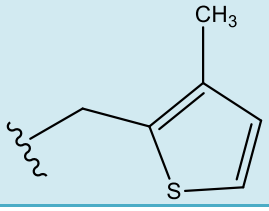
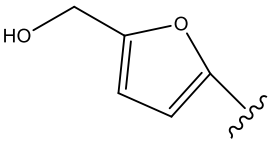
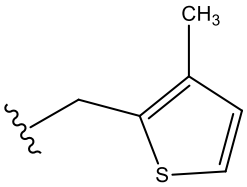
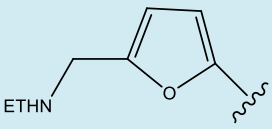
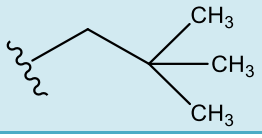
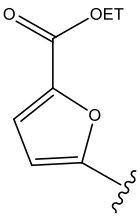
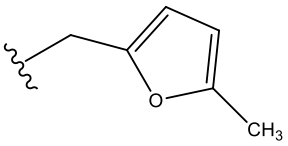
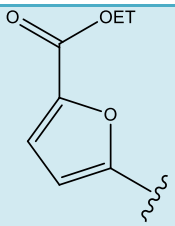
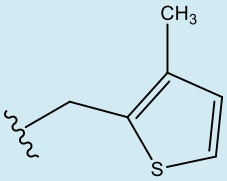
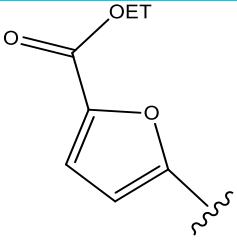
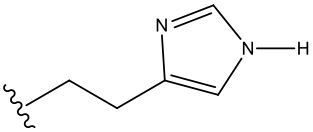
Figure III.1: Structure générale de 6-arylquinazolin-4-amine.

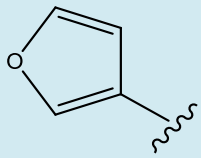
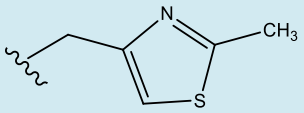
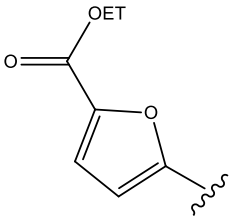
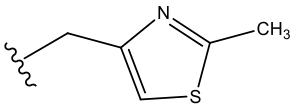
Tableau III.1: Structure chimique des dérivés de 6-arylquinazolin-4-amine

Mol	R1	R2	R3
1		H	
2		H	
3		H	
4		H	
5		H	

6		H	
7^{t}		CH_3	
8^{t}		CH_2CH_3	
9		H	
10^{t}		H	
11		CH_3	
12		H	
13		H	

14		H	
15		H	
16		CH_3	
17^t		H	
18		CH_3	
19^t		CH_3	
20^t		CH_3	
21		H	
22		H	

23		H	
24		H	
25		CH_3	
26		CH_3	
27		H	
28^t		H	
29		CH_3	
30		H	

31		H	
32		H	

Les composés de test sont marqués d'un astérisque ([t])

L'activité biologique inhibitrice est rapportée en terme IC50 : concentration micro molaire d'une drogue, nécessaire pour inhiber 50% (la moitié) de l'activité enzymatique. Pour notre cas nous avons exprimé l'activité inhibitrice par le rapport logarithmique pIC50 [Log (1/IC50)].

III.2.2 Sélection des descripteurs et méthodes de calcul :

Un descripteur moléculaire peut être considéré comme la conséquence d'un processus logique et mathématique, appliqué à l'information chimique codifiée à travers la représentation d'une molécule [120]. Le choix des descripteurs dépend des outils dont on dispose, de la nature des composés décrits et de la propriété ciblée. L'information codée d'un descripteur moléculaire dépend du type de représentation moléculaire employée et de l'algorithme défini pour son calcul.

Tout d'abord, les 32 molécules de 6-arylquinazolin-4- amine ont été pré-optimisées au moyen de la mécanique moléculaire, champ de force (MM+) en utilisant logiciel HyperChem version 8.0.6 [121].

Le module "propriétés QSAR" de l'Hyper Chem 8.0.6 a été utilisé pour calculer les paramètres suivants: la polarisabilité **(P)**, la réfractivité molaire **(R)**, le coefficient de partage octanol/eau **(log P)**, l'énergie d'hydratation **(H)**, le volume molaire **(V)**, la surface moléculaire **(S)** et le poids moléculaire **(M)** (tableau III.2) .

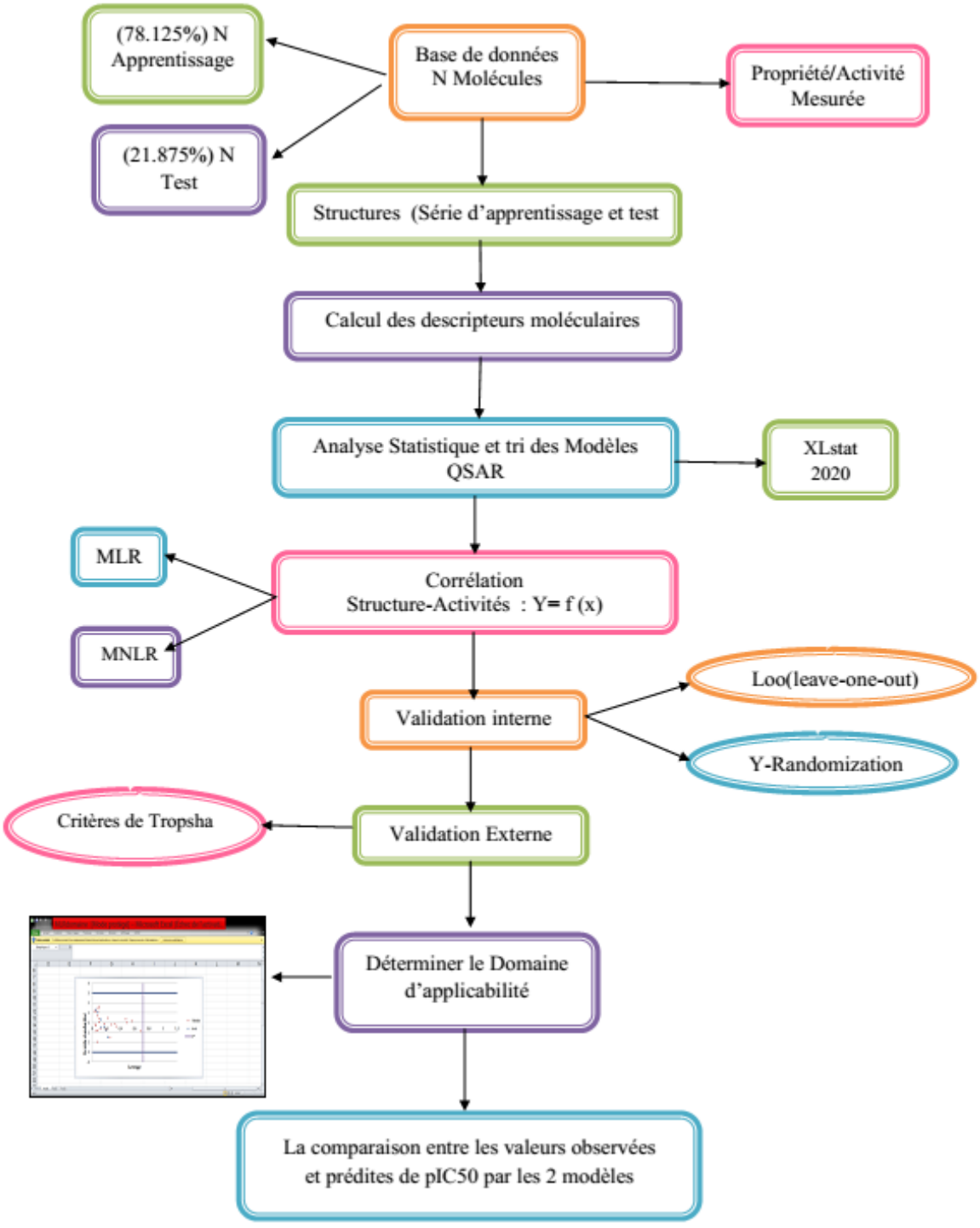
L'énergie des orbitales frontières **(HOMO, LUMO)**, les charges atomiques NBO **(q(N_{sub}); q(C_{sub}))** , l'énergie total **(ET)** et le moment dipolaire **(Md)** sont calculés par logiciel Gaussian 09 [55] avec la méthode DFT et la base 6-31G (d,).

III.2.3 Analyse statistique :

Les modèles Structure-activité biologique ont été générés en utilisant la méthode de régression multilinéaire (MLR) et la méthode Régression non linéaire multiple (MNLR) au moyen du logiciel XLstat 2020 pour Windows 7 [122].

En utilisée les valeurs pIC50 comme variable dépendante, les descripteurs physico-chimiques et quantiques comme variables indépendantes. Les modèles sont évalués par la valeur de R^2 (coefficientde détermination), le R^2ajusté , la valeur MSE (erreur quadratique moyenne) et la valeur F (statistique Fischer).

III.2.4 Méthodologie générale d'une étude QSAR :



FigureIII.2. Schéma de La méthodologie générale d'une étude QSAR

les logiciels Gaussian 09W , GaussView 5.0.8 [54] et Hyper Chem (8.06) ont été utilisé pour obtenir les descripteurs moléculaires des dérivés de 6-arylquinazolin-4- amine. (Tableau III.2)

TableauIII.2 : Descripteurs moléculaires utilisées dans l'étude QSAR

Mol	ET (a.u.)	Md	Homo	Lumo	q(Nsub)	q(Csub)
1	-1484.031	4.818	- 0.2418	0.0093	-0.621	-0.052
2	-1293.282	3.615	- 0.2007	- 0.0525	-0.650	-0.115
3	-1583.256	4.523	- 0.2039	-0.0460	-0.616	-0.094
4	-1523.350	5.322	- 0.2399	0.0089	-0.631	-0.047
5	-1560.482	3.780	- 0.2321	0.0035	-0.615	-0.114
6	-1368.505	6.045	- 0.2407	0.0132	-0.608	-0.097
7	-1523.338	4.145	- 0.2428	0.0087	-0.452	-0.050
8	-1562.649	3.131	- 0.2444	0.0075	-0.463	-0.047
9	-1523.349	4.977	- 0.2413	0.0083	-0.633	-0.051
10	-1161.053	4.112	- 0.2437	0.0110	-0.637	-0.048
11	-1200.358	3.779	- 0.2431	0.0118	-0.456	-0.047
12	-1200.371	4.226	- 0.2397	0.0111	-0.627	-0.053
13	-1163.418	2.804	- 0.2228	0.0417	-0.720	-0.083
14	-1200.375	4.241	- 0.2395	0.0096	-0.622	-0.050
15	-1539.402	5.328	- 0.2390	-0.0089	-0.617	-0.054
16	-1578.704	3.575	- 0.2015	- 0.0526	-0.429	-0.088
17	-1500.079	3.679	- 0.2391	- 0.0109	-0.629	-0.047
18	-1539.382	3.720	- 0.2393	- 0.0086	-0.449	-0.046
19	-1578.705	3.572	- 0.2442	0.00021	-0.453	-0.046
20	-1196.555	2.230	- 0.2500	- 0.0067	-0.478	-0.041
21	-1163.474	3.553	- 0.2436	0.0062	-0.620	-0.050
22	-1202.795	3.872	- 0.2418	0.0073	-0.618	-0.052
23	-1163.481	4.926	- 0.2393	0.0087	-0.618	-0.050
24	-1127.584	4.741	- 0.2404	0.0094	-0.618	-0.052
25	-1371.899	2.771	- 0.2450	0.0082	-0.448	-0.055
26	-1486.426	3.148	- 0.2425	0.0102	-0.446	-0.097
27	-1072.023	5.517	- 0.2374	0.0129	-0.513	-0.106
28	-1276.826	5.454	- 0.2483	0.0067	-0.636	-0.107
29	-1639.104	4.932	- 0.2487	0.0038	-0.445	-0.107
30	-1273.015	9.660	- 0.2389	0.0023	-0.652	-0.112
31	-1348.647	3.702	- 0.2400	- 0.0087	-0.619	-0.056
32	-1615.853	2.974	- 0.2488	- 0.0106	-0.628	-0.108

Tableau III.2: La suit

Mol	S (Grid)	V	H (Kcal /mol)	Log p	R	P	M (a m u)
1	517.93	899.56	-9.76	-1.59	111.96	39.56	361.42
2	471.02	800.05	-7.28	-1.53	97.22	34.39	307.37
3	507.56	892.27	-9.10	-2.19	112.09	39.47	379.41

4	533.71	933.04	-8.51	-1.96	117.15	41.40	375.44
5	560.29	969.64	-6.62	-1.68	114.02	40.61	379.43
6	509.97	842.83	-15.04	-0.86	98.27	35.02	323.37
7	582.25	971.41	-9.34	-1.23	117.26	41.40	375.44
8	595.88	1020.79	-8.23	-0.89	122.00	43.23	389 .47
9	519.56	924.99	-8.43	-1.44	116.24	41.40	375.44
10	505.75	877.38	-10.02	-1.93	105.52	37.20	345.36
11	530.65	926.31	-7.59	-1.57	110.81	39.03	359.38
12	515.06	907.82	-9.82	-2.00	110.03	39.03	359.38
13	500.09	888.61	-11.16	-2.36	104.51	38.03	349.39
14	518.60	910.15	-9.77	-1.91	111.31	39.03	359.38
15	535.87	942.83	-8.91	-1.95	113.27	40.69	376.43
16	559.86	990.31	-7.24	-1.59	118.56	42.52	390.46
17	512.62	890.65	-10.46	-1.75	107.64	38.85	362.41
18	532.62	937.33	-7.48	-1.39	112.93	40.69	376.43
19	567.48	1000.74	-6.68	-1.36	118.72	42.52	390.46
20	528.20	925.12	-8.62	-1.91	108.73	39.04	359.39
21	530.75	918.20	-9.86	-1.31	104.51	37.58	349.39
22	547.27	959.14	-9.18	-0.91	109.11	39.42	363.42
23	533.39	922.36	-8.22	-0.98	104.17	37.58	349.39
24	538.71	945.15	-7.75	0.29	107.06	38.78	347.42
25	494.44	865.54	-5.70	-1.10	105.53	38.06	335.42
26	547.93	951.06	-8.87	-1.77	114.13	40.53	365.45
27	602.38	1036.49	-4.87	0.21	107.83	39.40	338.45
28	593.96	1014.52	-6.58	-1.99	113.36	40.09	377.40
29	616.02	1065.06	-4.07	-1.16	123.59	44.28	407.49
30	612.09	1031.23	-12.19	-2.71	111.92	40.09	377.40
31	469.79	822.64	-8.16	-1.98	97.25	35.51	322.38
32	584.66	1017.51	-7.18	-2.04	115. 32	41.74	394.45

III.3 Résultats et discussions :

III.3.1 Analyse des composants principaux :

Dans cette partie, ACP a été utilisé pour déterminer les descripteurs directement liés aux activités des inhibiteurs de la 6-arylquinazoline-4-amine de DYRK1A.

Tableau III.3: La matrice de corrélation (Pearson (n)) entre les différents descripteurs obtenus

Var	V	S	H	LogP	R	P	M	LUMO	HOMO	Md	ET	Csub	Nsub	pIC50
V	**1**													
S	**0.961**	**1**												
H	0.421	0.323	**1**											
LogP	0.161	0.196	0.319	**1**										
R	0.775	0.696	0.377	-0.068	**1**									
P	0.826	0.730	0.428	-0.009	**0.982**	**1**								
M	0.772	0.680	0.279	-0.222	**0.938**	**0.941**	**1**							
LUMO	0.135	0.151	-0.205	0.159	-0.006	0.012	-0.064	**1**						
HOMO	-0.306	-0.293	-0.025	-0.152	-0.179	-0.204	-0.162	-0.688	**1**					
Md	0.260	0.386	-0.302	-0.060	0.032	0.014	0.064	0.092	-0.050	**1**				
ET	-0.257	-0.252	-0.199	0.211	-0.635	-0.624	-0.647	0.380	-0.101	0.094	**1**			
Csub	-0.269	-0.366	-0.137	0.117	0.002	-0.009	-0.053	0.257	-0.371	-0.310	0.137	**1**		
Nsub	0.374	0.328	0.472	0.321	0.469	0.481	0.322	-0.124	-0.150	-0.316	-0.335	0.131	**1**	
pIC50	0.065	-0.014	-0.076	0.164	0.334	0.311	0.265	0.079	0.098	-0.143	-0.022	0.655	0.193	**1**

Les 13 descripteurs (variables) codant pour les 32 molécules ont été soumis à une analyse en composantes principales (ACP). Les deux premiers axes F1 et F2 contribuant respectivement à 36,643% et 51,315% à la variance totale, l'information totale est estimée à un pourcentage de 87,958%.

L'ACP a été réalisée pour identifier le lien entre les différentes variables ; Il est également utile pour comprendre la distribution des composés [123] . Les corrélations entre les treize descripteurs sont présentées dans le tableau III.3 sous forme de matrice de corrélation, ces descripteurs sont représentés dans un cercle de corrélation. La matrice obtenue fournit des informations sur la corrélation négative ou positive entre les variables.

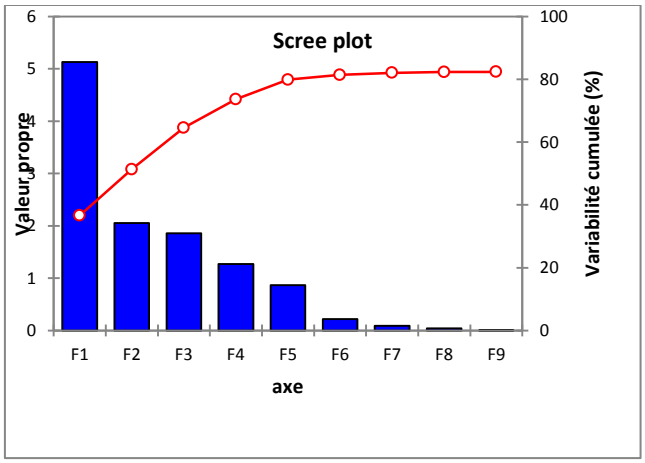


Figure III.3: Les principales composantes et leurs variances

Le cercle de corrélation a été fait pour détecter la connexion entre les différents descripteurs. Chaque variable (descripteur) était représentée par un vecteur. La direction et la longueur de chaque vecteur fournissent des informations sur l'impact des descripteurs sur la hache individuelle (comme le montre la figure III.3) L'analyse en composantes principales révélée par le cercle de corrélation (figure III.4) montre que l'axe F1 (36,64% de la variance) est localisé par les autres paramètres physico-chimique, tandis que l'axe F2 (51,31% de la variance) est principalement dû à la carbone substituant q(C_{sub}) .

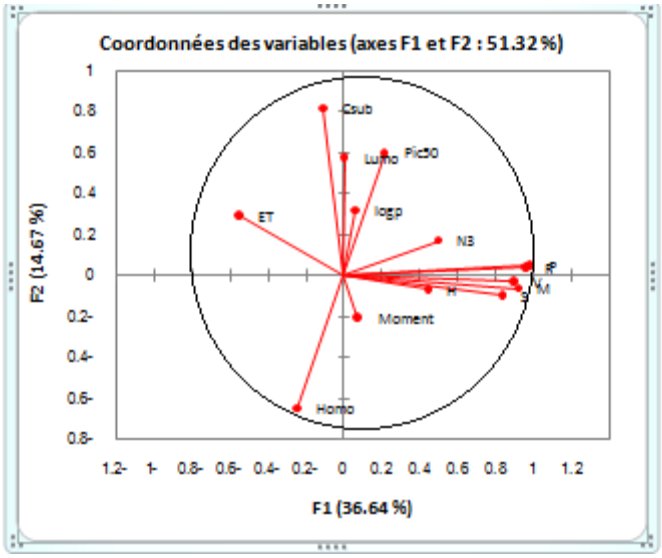


Figure III.4: Cercle de corrélation pour les axes F1, F2

La matrice obtenue fournit des informations sur l'interrelation élevée ou faible entre les descripteurs et identifie leur colinéarité potentielle. En général, une bonne colinéarité (R>0,5) a été observée entre la plupart des variables. Une interdépendance élevée a été observée (R ≈ 1,00). De plus, pour diminuer la redondance existante dans notre matrice de données, les

descripteurs qui sont fortement corrélés (R≥ 0,9) ont été exclus. Afin de réduire le nombre de paramètres non significatifs et de former les résultats obtenus par la matrice et le cercle des corrélations, les corrélations suivantes sont observées :

- S et V : est corrélé (r (S, V) = 0.961)

On élimine le descripteur V et on laisse S

- (R; P), (R; M), (P; M) (sont parfaitement corrélés (r ≈1). Par conséquent, R et P sont élimines .

III.3.2 Les modèles QSAR :

III.3.2.1 La méthode de régression multilinéaire (MLR) :

Le développement d'un modèle QSAR exige un ensemble de diverses données, pour cet effet, un nombre considérable de descripteurs doivent être considérés dans l'étude QSAR. Les descripteurs sont des valeurs numériques qui codent les différentes caractéristiques structurelles des molécules. La sélection d'un ensemble de descripteurs appropriés à partir d'un grand nombre de descripteurs utilisés, nécessite un procédé qui est capable de discriminer entre les paramètres. La matrice de corrélation de Pearson a été effectuée sur tous les descripteurs en utilisant le logiciel XLstat version 2020 .

Dans ce travail, nous avons un ensemble de 32 molécules réparties au hasard en deux sous-groupes: un sous-ensemble de 7 molécules sélectionnées au hasard. Les 25 autres molécules constituent le TSET utilisé pour construire le modèle QSAR.

L'analyse de régression linéaire multiple (MLR) est utilisée pour modéliser la relation structure-activité. Cette technique mathématique minimise la différence entre les valeurs réelles et prévues.

De nombreuses tentatives ont été faites pour développer une relation de pIC50 , mais la meilleure relation obtenue avec cette méthode ne correspond qu'à la combinaison linéaire de plusieurs descripteurs sélectionnés:M / LUMO / HOMO / q(Csub) / S, il est déterminé à partir de l'analyse des paramètres statistiques obtenus. (tableau III.4).

Tableau III.4: Modèles sélectionnés et paramètres statistiques des corrélations entre les propriétés moléculaires et l'activité biologique.

Model	Variables	R^2	R^2 ajusté	F
1	V/S/H/logP/q(Nsub)	0.671	0.606	12.740
2	R /M /LUMO / HOMO /q(Csub)	0.810	0.620	8.255
3	R / P/ LUMO / HOMO / q(Csub)	0.640	0.568	13.850
4	M / LUMO / HOMO /q(Csub) / S	0,867	0,831	24,682

La corrélation entre l'activité biologiques et les descripteurs exprimés par les relations suivantes(Equ 1):

$$pIC50=14.889+5.431^{3}.\ M+15.378\ .\ LUMO + 59.127.\ HOMO +32.950\ .\ C_{sub} + 1.131^{-2}.S$$

$N= 25$; $R^2= 0.867$; R^2 ajusté$= 0.831$;$F= 24.682$; $P< 0.0001$; $MSE= 0.103$, $N_{test}= 7$, $R^2_{test}= 0.881$

Où : **N** est le nombre de composés ; **R^2**: coefficient de détermination ;

R^2ajusté: est coefficient de détermination ajusté ; **F :** est la statistique de Fischer;

MSE: erreur quadratique moyenne ; **N_{test} :** le nombre de composés de la validation externe ;

R^2_{test}: La coefficient de détermination issu de la validation externe

Une amélioration significative de la qualité du modèle QSAR est obtenue avec la combinaison des cinq paramètres à savoir, la masse moléculaire (M), la surface molaire (S), Les orbitales frontières HOMO et LUMO et la charge l'atome q(C_{sub}).

Le modèle QSAR doit considérer un **R^2 > 0.6** pour qu'il soit valide [124]. Par exemple comme notre cas, la valeur R^2 = 0,867; nous permis d'indiquer fermement la corrélation entre les différents descripteurs utilisés (variables indépendantes) et les activités biologiques et une

erreur quadratique moyenne plus faible (MSE) indique que le modèle proposé est prédictif et fiable.

La valeur F est jugée la signification statistique au niveau de 95%, pour toutes les valeurs de F calculées est supérieures par rapport aux valeurs lus dans la table du Fischer (voire l'annexe).

- **Tests de colinéarité :**

Le problème de colinéarité entre les descripteurs inclus dans le modèle final de QSAR est testé par l'examen de la matrice de corrélation, en calculant le coefficient de corrélation pour toutes les combinaisons paires possibles des cinq descripteurs. Les valeurs élevées du coefficient de corrélation $R \geq 0,9$ correspondent aux fortes corrélations entre les descripteurs du modèle (Equ 1):Les résultats obtenus sont récapitulés dans le tableau III.5.

Tableau III.5: Matrice de corrélation

Variables	M	LUMO	HOMO	C_{sub}	S	pIC50
M	**1,000**					
LUMO	-0,090	**1,000**				
HOMO	-0,117	-0,722	**1,000**			
C_{sub}	-0,108	0,305	-0,363	**1,000**		
S	0,648	0,111	-0,294	-0,430	**1,000**	
pIC50	0,242	0,070	0,132	0,628	-0,082	**1,000**

L'examen de la matrice de corrélation (Tableau III.5) confirme l'absence du problème de colinéarité entres les descripteurs du modèle, expliquée par les faibles valeurs des coefficients de corrélation ($R < 0,9$).

D'après la figure III.5, les descripteurs M ; HOMO ; LUMO ; C_{sub} et S positives montrent que toute augmentation dans les valeurs de ces paramètres entraine une augmentation de l'activité biologique. Nous notons également que C_{sub} a un impact significatif sur l'activité biologique alors que la masse n'affecte pas significativement l'activité biologique

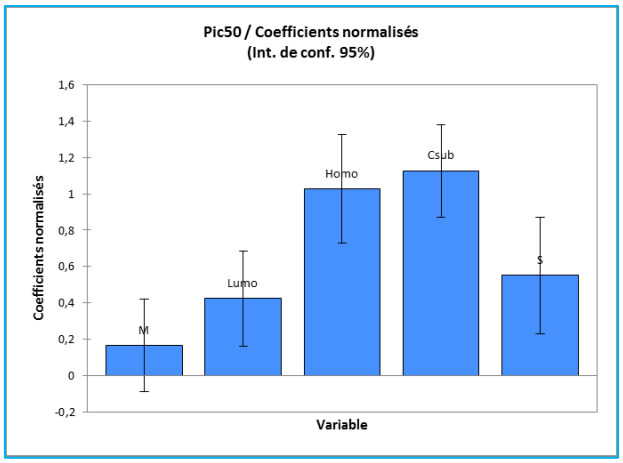


Figure III.5: Diagramme de la contribution en pourcentage de chaque descripteur dans le modèle pIC50 développé expliquant la variation de l'activité.

Dans le modèle obtenus dans l'équation (Equ 1): le coefficient de M (masse moléculaire) dans l'équation est positif, ce qui indique que l'activité biologique (pIC50) des composés est positivement corrélée avec le poids moléculaire (M).

L'HOMO est l'orbitale qui agit principalement comme un donneur d'électrons et le LUMO est l'orbitale qui agit en grande partie comme accepteur d'électrons . Alors que l'énergie de l'HOMO est directement liée au potentiel d'ionisation d'une molécule et il est caractérise comme un composé électrophile, l'énergie LUMO est directement liée à l'affinité d'une molécule et il caractérise la sensibilité de la molécule d'être attaqué par les nucléophiles, d'après l'équation de corrélation on remarque que l'activité biologique dépend de potentiel d'ionisation et l'affinité électronique

Nous avons observé que La charge l'atome C_{sub} diminue lorsque R1 est un cycle phényle substitué par un groupe hydrophile et attracteur d'électrons.d'après l'équation, on remarque que l'activité biologique dépend des effets électroniques .

Aussi, l'augmentation de la surface moléculaire en accord avec l'augmentation de surface de contact ligand-récepteur, donc on observe le descripteur positif qui montre que toute augmentation de S entraine une augmentation de l'activité inhibitrice.

Les valeurs élevées des descripteurs mentionnés ci-dessus résument la corrélation des cinq descripteurs avec l'activité biologique. Cette corrélation est vérifiée en examinant le tableau d'ANOVA et le test t.

- **Analyse de la variance ANOVA :**

A partir du tableau d'ANOVA, la présence ou l'absence de corrélation entre l'ensemble de descripteurs et l'activité biologique est vérifiée en examinant la statistique de Fisher F_{obs}. Pour la réaliser nous avons fait appel aux deux hypothèses : hypothèse nulle et hypothèse alternative.

L'hypothèse nulle (H0) : « aucun descripteur n'est lié à l'activité biologique : $\beta j=0$ avec (j=0 ; 1 ; 2 ; 3 ; 4); *Bj:* coefficient correspond au descripteur *j* ». Cette hypothèse est acceptée si la valeur de la statistique de Fischer observée est inférieure à la valeur $F_{(\nu_1;\nu_2)}$ (voire l'annexe A)

L'hypothèse alternative (H1) : « il existe au moins un descripteur corrélé avec l'activité biologique ». Cette hypothèse est acceptée lorsque la valeur de la statistique de Fischer observée est supérieure de la valeur $F_{(\nu_1;\nu_2)}$ à un niveau de confiance (1- α = 0.95 donc α= 0.05)

Où: ν_1**:** est le degré de liberté du numérateur

ν_2**:** est le degré de liberté du dénominateur .

Tableau III.6: Analyse de la variance ANOVA:

Source	DDL	Somme des carrés	Moyenne des carrés	F_{obs}	Pr > F
Modèle	5	12,756	2,551	24,682	< 0.0001
Erreur	19	1,964	0,103		
Total corrigé	24	14,720			

Où : **DDL:** est le degré de liberté , ***Pr* :** est la probabilité et $\mathbf{F_{obs}}$: la statistique de Fischer observée

- ν_1= p = 5
- ν_2 = N –p – 1= 25- 5 -1= 19

Où : **N:** le nombre d'observations; **p:** est le nombre de variables indépendantes (les descripteurs).

D'après le tableau d'ANOVA, la statistique de Fischer observée (F_{obs}= 24,682) est supérieure à ($F_{(5;19)}$= 2.74), ce qui nous permet d'accepter l'hypothèse alternative et de confirmer qu'il existe au moins un coefficient différent de zéro c'est-à-dire un descripteur corrélé avec l'activité inhibitrice expliquée par pIC50.

- **Test de Student (t):**

Nous avons ensuite examiné le tableau des Paramètres du modèle pour vérifier la signification de chaque descripteur et sa contribution dans l'explication de l'activité biologique. L'utilisation des valeurs de la statistique T de Student, affichées dans le Tableau ci-dessous pour chaque descripteur, nous permet de vérifier la présence ou l'absence de corrélation entre chaque descripteur et l'activité biologique en se basant sur les deux hypothèses : hypothèse nulle et hypothèse alternative.

L'hypothèse nulle (H0) :« le descripteur n'est pas lié à l'activité biologique : $\beta j = 0$ avec (j=0 ; 1 ; 2 ; 3 ; 4), βj : coefficient correspond au descripteur *j* » Cette hypothèse est acceptée si la valeur de la statistique de Student t observée est inférieure à la valeur T ($1 - \frac{\alpha}{2}$, DDL) à un niveau de confiance α =0.05.

L'hypothèse alternative (H1) :« le descripteur est corrélé avec l'activité biologique : $\beta j \neq 0$ ». Cette hypothèse est acceptée lorsque la valeur de la statistique de Student t observée T_{obs} est supérieure de la valeur T ($1 - \frac{\alpha}{2}$, DDL)

Où : T ($1 - \frac{\alpha}{2}$, DDL) la valeur de la table de Student avec le degré de liberté DDL (N-p-1 = 25-5-1=19) et ($1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975$) est la probabilité

D'après le table de Student (voire annex B), la valeur de $T_{(0.975,19)}$= 2.093

Tableau III.7: Tableau des coefficients

Source	Valeur	Erreur standard	T_{obs}	Pr > |t|
Constant	14,889	1,836	8,108	< 0.0001
S	0,011	0,003	3,597	0,002
M	0,005	0,004	1,353	0,192
Lumo	15,378	4,565	3,369	0,003
Homo	59,128	8,244	7,172	< 0.0001
Csub	32,951	3,556	9,267	< 0.0001

Où : $\mathbf{T_{obs}}$ **:** est la valeur de la statistique de Student t observée

Pr: est la probabilité.

D'après ce tableau, les valeurs de la statistique observées T_{obs} sont plus élevées par rapport à celle du Tableau de la distribution $T_{(0.975,19)}$. Cela nous permet de rejeter l'hypothèse nulle, c'est-à-dire que les coefficients inclus dans le modèle diffèrent considérablement de zéro. Ce jugement est consolidé par les faibles valeurs de probabilité (Pr < α) pour les cinq paramètres (constante + 4 descripteurs) de l'équation (Equ III.1).

Sauf , le test-t pour le paramètre M est égale à 1.353 est inférieur à la valeur de la table de Student. La valeur de *Pr* justifie aussi la non contribution de ce descripteur :Pr = 0.192> α; (α=0.05).

- **Test Multicolinéarité :**

Nous avons ainsi examiné la multicolinéarité par l'approche de la valeur d'inflation de la variance (VIF: Variance Inflation Factor).

Le VIF a été défini comme $\frac{1}{(1-r\,2)}$ [125]; où **r** était le coefficient de corrélation d'une variable indépendante par rapport à tous les autres descripteurs du modèle. Les variables avec un VIF supérieur à 5 sont instables et doivent être éliminées, les modèles avec des valeurs VIF comprises entre 1 et 4 signifient que les modèles peuvent être acceptés. On peut aussi confirmer la présence ou l'absence de la forte multicolinéarité par une autre grandeur, similaire au VIF, appelée facteur de tolérance TF .

Les valeurs de TF varient dans l'intervalle 0 < TF < 1. Les valeurs de TF > 0,5 et TF < 0,5 correspondent à la présence et à l'absence de fortes multicolinéarité entre les descripteurs respectivement.

Les valeurs correspondantes au VIF(x_k) issues des cinq descripteurs de notre modèle sont réunies dans le tableau III.8.

Tableau III.8: Valeurs des critères VIF et TF pour les descripteurs significatifs

Statistique	M	Lumo	Homo	Csub	S
Tolérance	0,473	0,446	0,342	0,475	0,299
VIF	2,113	2,242	2,923	2,104	3,343

D'après ce Tableau, les cinq descripteurs ne présentent aucun problème de multicolinéarité avec une valeur maximale de VIF(s) = 3.343< 5 et les valeurs de TF toutes inférieures à 0,5.

III.3.2.2 Régression non linéaire multiple (MNLR) :

La méthode statistique de régression non linéaire a été utilisée pour quantifier l'activité attendue (pIC50), Il a également été utilisé pour améliorer la relation entre la structure et l'activité pour une évaluation quantitative de l'effet de l'alternative. Nous avons appliqué à la matrice de données clairement formée des descripteurs suggérés par le MLR correspondant pour 25 molécules. Les coefficients R^2 et R^2 ajusté sont utilisés pour déterminer les meilleures offres de régression.

La matrice de corrélation de Pearson a été effectuée sur tous les descripteurs en utilisant le logiciel XLstat 2020. Les résultats obtenus par l'analyse MNLR Il a pris en compte les 5 descripteurs choisis, L'équation résultante est (Equ 2):

$$pIC50= 61.364+5.363^{-2}. M+21.511. Lumo+711.883. Homo+20.001.Csub+8.940^{-2}.S - 7.114^{5}. M^{2} -378.445.Lumo^{2} +1396.273. Homo^{2} – 60.384. Csub^{2} -7.205^{-5}. S^{2}$$

$N = 25$; $R^2 = 0.902$; MSE = 0.103 ; R^2 ajusté = 0.876 ;$N_{test}= 7$; $R^2_{test}= 0.850$

Où : **N :** est le nombre de composés .

R^2: coefficient de détermination .

R^2 ajusté: coefficient de détermination ajusté .

N_{test} : le nombre de composé de la validation externe .

R^2_{test}: La coefficient de détermination issu de la validation externe .

la qualité du modèle QSAR est obtenue avec la combinaison des cinq paramètres à savoir, la masse moléculaire (M), la surface molaire (S) , Les orbitales frontières HOMO et LUMO et la charge l'atome q(Csub).

Considérant que, les valeurs des activités prévues pIC50 MNLR sont calculées à partir de l'équation (Equ 2) et les valeurs observées sont présentées dans le tableau III.14. Montre les corrélations entre les activités prévues . La corrélation entre les activités MNLR calculées et expérimentales est très significative , comme indiqué par les valeurs R^2 (comme notre cas,

la valeur R^2 = 0,902 ; cela indique la force de la relation entre les différentes descripteurs utilisées et les activités biologiques). La valeur de R^2 indique que le modèle proposé est prédictif et fiable.

III.3.3 Validation des modèles :

La validation des modèles QSAR obtenu est nécessaire pour estimer sa fiabilité. Dans cette étude nous avons utilisé trois méthodes de validation : la validation interne (validation croisée et test de randomisation) et la validation externe. La validation croisée est effectuée par la procédure leave-one-out, on retire successivement une molécule de l'ensemble d'apprentissage. Cette procédure est réitérée n fois (n est le nombre des molécules qui constituent l'ensemble d'apprentissage) afin de prédire les propriétés de toutes les molécules.

III.3.3.1 Validation interne:

Pour la validation de modèle on utilisant la méthode de validation croisée, est une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage. Le leave-one-out (LOO) a été utilisé pour cette proposition dans laquelle un composé est retiré du jeu de données et reconstruit le modèle. Ceci est répété de telle sorte que chaque observation dans l'échantillon est utilisée une fois comme données de validation . Il s'agit d'utiliser une seule observation de l'échantillon d'origine comme donnée de validation et les observations restantes comme données d'apprentissage. Un certain nombre d'ensembles de données modifiés sont créés en supprimant dans chaque cas une ou un petit groupe de molécules [93]. Cette dernière validé notre model par calcule de ses paramètres statistiques :Press, SST, Press/SST, S, R^2cv et ; R^2aju

Tableau III.9: Les valeurs de validation croisée

Modèle	PRESS	SST	PRESS/SST	R^2_{cv}	R^2_{aju}
MLR	1.964	14.720	0.133	0.867	0.831
NMLR	0.156	1.600	0.098	0.902	0.876

Où : **PRESS**: est Prédictive somme des carrés résiduelle ; **SST**: est la somme des carrés totaux et $\mathbf{R^2_{cv}}$: le coefficient de corrélation prédictive

D'après ce tableau, Le PRESS est un paramètre important. Sa valeur inférieure à indique que les modèles proposées à un bon pouvoir prédictif et ce sont mieux que le hasard.

Tell que dans le (Tableau III.9) les valeurs de PRESS sont :dans le modèle MLR est égale à 1.964 et MNLR est égale à 0.156, les valeurs de SST sont: (MLR)= 14.720 et (MNLR) est égale à 1.600.

Dans notre cas, (Tableau III.9) pour les deux modèles proposées PRESS < SST indiquant qu'il est un bon pouvoir prédictif et sont mieux que le hasard.

Pour un modèle bon (PRESS / SST) devrait être inférieur à 0.4 et sa valeur dans notre modèles sont 0.133 et 0.098 [126].

Autre paramètres de validation croisée important est le coefficient de corrélation prédictive, la valeur R^2_{cv} la plus élevée qui a donné l'excellent modèle, et dans notre étude (tableau III.9) (R^2_{cv}= 0.867), pour (R^2_{aju} est 0.831) aussi la valeur plus élevée exprime un meilleur modèle.

D'un autre côté, Les paramètres de performance obtenus par La méthode de régression non-linéaire indique que ce modèle est plus fiable.Le modèle obtenu a été validé par la technique de validation croisée. La valeur de (R^2_{cv} = 0.906) est supérieure à 0,6 et (R^2_{aju}=0.876) ; cela indique le meilleur modèle MNLR prédictif.

III.3.3.2 Validation externe :

La bonne mesure de prévisibilité du modèle de QSAR obtenu par la validation interne (validation croisée) ne suffit pas. En effet, nous devons en plus généraliser ces prévisions en les appliquant sur un échantillon externe.

Selon Tropsha et ses collaborateurs [93] , un modèle de QSAR ne possède une puissance prédictive acceptable que si les critères de Tropsha sont vérifiés (Comme nous l'avons mentionné dans le deuxième chapitre)

Tableau **III.10**: Critères de Tropsha

Modèles / **Critères**	**MLR**	**MNLR**
$R^2_{CV} > 0.5$	0.867	0.902
$R^2_{pred} > 0.6$	0.881	0.850

$\frac{(R^2 - R_0^2)}{R^2} < 0.1$	-0.127	-0.171
$\frac{(R^2 - R'^2_0)}{R^2} < 0.1$	-0.130	-0.174
$\lvert R_0^2 - R'^2_0 \rvert < 0.3$	0.003	0.003
k ($0.85 \leq k \leq 1.15$)	0.991	0.993
k'($0.85 \leq k' \leq 1.15$)	1.006	1.004

Avec :

R^2: (des équations1 et 2), est le coefficient de détermination entre les valeurs observées et celles prédits par le modèle (seulement pour l'ensemble d'apprentissage : 25 molécules).

R_0^2 : coefficient de détermination issu de la régression des activités observées sur les activités prédites pour tout l'ensemble (32 molécules).

R'^2_0: coefficient de détermination issu de la régression des activités prédites sur les activités observées pour tout l'ensemble (32 molécules).

k et k ': sont les pentes des lignes de régression qui passe par l'origine.

D'après ce tableau, tous les critères de Tropsha sont vérifiés soit le modèle MLR ou le modèle MNLR .

Afin , d'après les résultats on peut voir clairement que le MNLR est statiquement meilleur que le modèle MLR en termes de coefficient de détermination, mais le modèle MLR a la capacité de prédiction la plus élevée pour l'ensemble de test (R^2_{pred}= 0,881).

Cependant, les résultats obtenus par le MLR et le MNLR doivent être considérés comme satisfaisants pour prédire l'activité de 6-arylquinazolin-4-amine à l'aide des descripteurs proposés.

III.3.3.3 Y-Randomisation:

Évaluation de la randomisation Y ou le test de randomisation Y est un test de confirmation pour montrer que le modèle QSAR développé est fiable, solide et robuste et n'a pas été obtenu par hasard. Ce test a été effectué sur les données de l'ensemble d'entraînement [92]. Des modèles de régression multilinéaire (MLR) et (MNLR) ont été générés en

mélangeant aléatoirement la variable dépendante (données d'activité) tout en gardant les variables indépendantes (descripteurs) inchangées. Il est prévu que le modèle QSAR développé devrait avoir des valeurs R^2 et Q^2 significativement basses pour les nombres d'essais afin de vérifier que le QSAR développé le modèle est robuste. Coefficient de randomisation Y (cR_p^2) est un autre paramètre important qui devrait être supérieur à 0,5 pour réussir ce test.(Equ III. 3)

$$cR_p^2 = R \times [R^2 - (R_r)^2]^2$$

Avec **:** $\mathbf{cR_p^2}$ **:**est le coefficient de randomisation Y,

R: est la corrélation coefficient de randomisation Y.

R_r: est le «R» moyen de modèles aléatoires.

À l'étape suivante, tous les calculs ont été répétés avec des activités randomisées des composés de l'ensemble d'apprentissage afin d'évaluer la robustesse du modèle (randomisation y tester). Dans le cas présent, 50 essais aléatoires ont été exécutés pour le modèle MLR et MNLR [127]. Aucun des essais aléatoires n'a pu correspondre au modèle d'origine (tableau III.11).

Tableau III.11: les 50 premières itérations de Y-randomisation

	MLR			**MNLR**		
Model	R	R^2	Q^2	R	R^2	Q^2
Random 1	0,348	0,121	-0,458	0,743	0,552	-0,949
Random 2	0,573	0,329	-0,232	0,700	0,490	-1,839
Random 3	0,530	0,281	-0,199	0,689	0,475	-0,222
Random 4	0,467	0,218	-0,396	0,668	0,447	-2,877
Random 5	0,330	0,109	-0,925	0,734	0,539	-3,528
Random 6	0,523	0,273	-0,138	0,659	0,435	-1,300
Random 7	0,457	0,209	-0,418	0,579	0,336	-1,844
Random 8	0,492	0,242	-0,296	0,615	0,378	-1,261
Random 9	0,537	0,288	-0,324	0,598	0,358	-1,048
Random10	0,395	0,156	-0,579	0,778	0,605	-2,503
Random11	0,567	0,321	-0,114	0,680	0,463	-7,662
Random12	0,450	0,203	-0,390	0,546	0,298	-9,569
Random13	0,561	0,315	-0,100	0,466	0,217	-2,315
Random14	0,443	0,196	-0,597	0,792	0,627	-0,953
Random15	0,405	0,164	-0,276	0,518	0,268	-4,284
Random16	0,439	0,192	-0,601	0,460	0,212	-1,551

Random17	0,617	0,381	-0,068	0,711	0,505	-4,817
Random18	0,522	0,272	-0,261	0,605	0,366	-2,851
Random19	0,599	0,359	-0,121	0,691	0,478	-2,755
Random20	0,402	0,162	-0,701	0,732	0,535	-0,536
Random21	0,476	0,226	-0,579	0,561	0,315	-2,005
Random22	0,476	0,226	-0,522	0,827	0,684	0,308
Random23	0,381	0,145	-0,351	0,645	0,416	-3,065
Random24	0,559	0,312	-0,919	0,702	0,492	-1,080
Random25	0,441	0,195	-0,469	0,421	0,177	-4,870
Random26	0,443	0,196	-0,493	0,583	0,340	-4,007
Random27	0,345	0,119	-0,509	0,513	0,264	-6,684
Random28	0,375	0,140	-0,232	0,456	0,208	-3,273
Random29	0,682	0,465	-0,154	0,763	0,582	-0,547
Random30	0,342	0,117	-0,602	0,728	0,529	-2,022
Random31	0,354	0,125	-0,747	0,711	0,506	-1,697
Random32	0,417	0,174	-0,248	0,676	0,457	-2,675
Random33	0,489	0,239	-0,434	0,849	0,722	-0,258
Random34	0,506	0,256	-0,234	0,383	0,147	-5,593
Random35	0,555	0,309	-0,549	0,577	0,333	-3,081
Random36	0,509	0,259	-0,348	0,702	0,493	-2,887
Random37	0,283	0,079	-0,309	0,758	0,575	-1,033
Random38	0,612	0,375	-0,386	0,555	0,308	-1,911
Random39	0,354	0,126	-1,023	0,641	0,411	-1,765
Random40	0,176	0,031	-0,789	0,718	0,515	-3,764
Random41	0,452	0,204	-0,462	0,703	0,495	-1,296
Random42	0,264	0,069	-0,857	0,717	0,514	-2,566
Random43	0,115	0,013	-0,882	0,661	0,437	-2,230
Random44	0,621	0,385	-0,301	0,785	0,616	-2,120
Random45	0,477	0,228	-0,574	0,794	0,629	-0,604
Random46	0,575	0,331	-0,096	0,624	0,389	-0,764
Random47	0,598	0,358	-0,099	0,691	0,477	-1,073
Random48	0,454	0,206	-0,814	0,563	0,317	-8,006
Random49	0,550	0,303	-0,508	0,685	0,469	-1,063
Random50	0,462	0,214	-0,407	0,682	0,465	-3,467

Tableau III.12: Random Models Parameters

Modèle	**Average R :**	**Average R ² :**	**Average Q ² :**	**cR_p^2:**
MLR	0.460	0.224	-0.442	0.757
NMLR	0.653	0.437	-2.595	0.656

Les valeurs cR_p^2est égale à 0,759 et 0.792 pour le modèle MLR et MNLR respectivement (plus de 0,5) qui a été rapportée dans le tableau III.11 soutenant l'affirmation selon laquelle le modèle généré est puissant et non déduit par hasard . et tous les nouveaux modèles QSAR ayant des valeurs R^2_{CV} (Q^2) et R^2 significativement basses pour les 50 essais, qui confirment que les modèles QSAR développés sont robustes. D'après les résultats de la validation interne on peut conclure que ces modèles sont stables et a un pouvoir explicatif très acceptables.

III.3.4 Domaine d'applicabilité du modèle QSAR :

Pour estimer la fiabilité d'un modèle QSAR et sa capacité à prédire de nouveaux composés, le domaine d'applicabilité doit être essentiellement défini [128]. Les composés prévus qui entrent dans ce domaine peuvent être considérés comme fiables. Les domaines d'applicabilité étaient discuté avec les graphiques de Williams dans les figures III.6 et III.7 des modèles MLR et MNLR respectivement :

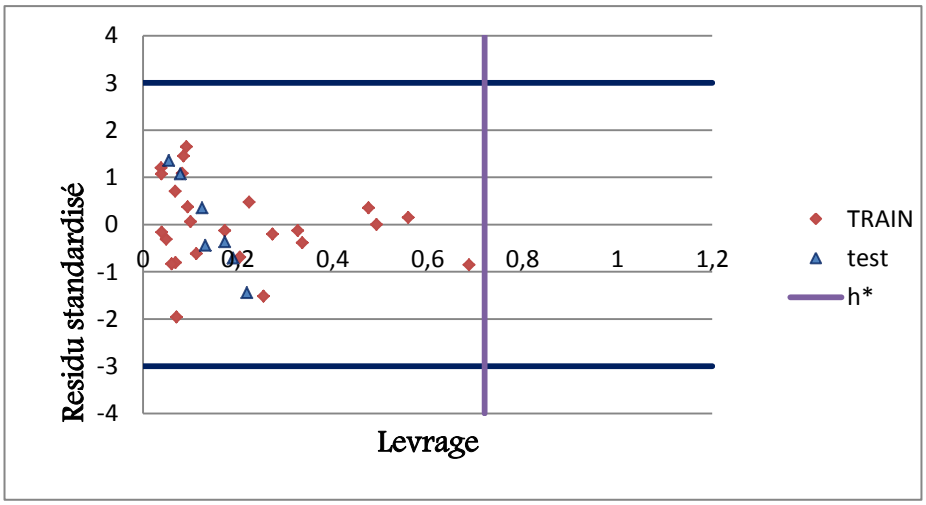


Figure III.6: La courbe de Williams pour le modèle MLR

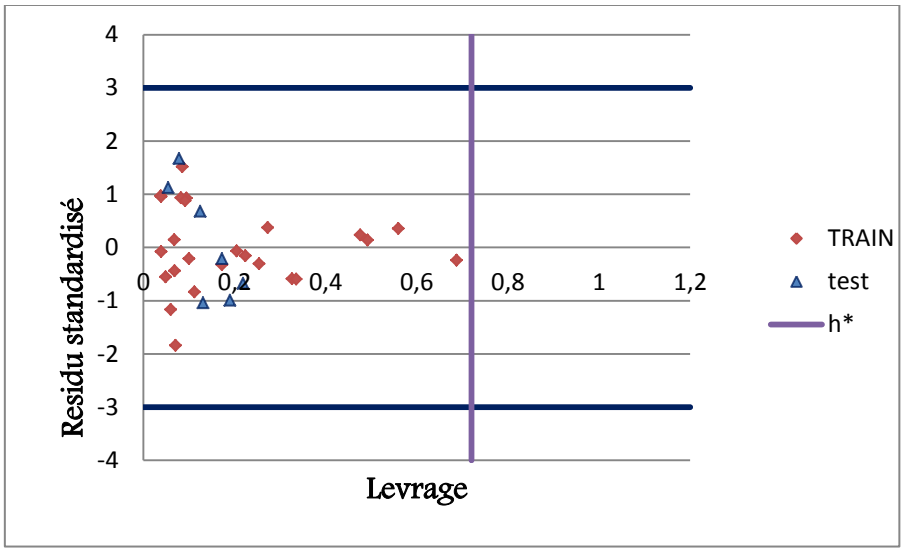


Figure III.7: la courbe de Williams pour le modèle MNLR

Le domaine a été discuté sur la base du graphique Williams sur les figures (III.6 et III.7), où les résidus et l'effet de levier normalisés les valeurs (hi) sont tracées. Sur la base du calcul de la tirer parti de salut pour chaque composé, le modèle QSAR est utilisé pour prédire l'activité du composé (basant sur l'équation mathématique précédente Equ 23 *(mentionnée au chapitre deux).*

Le calcul des valeurs de levier *hi* et des valeurs résiduelles σ_i (tableau III.13), pour tous les composés de la base de données, nous a permis de définir le graphe de Williams. Ce dernier est une représentation graphique des valeurs résiduelles normalisées en fonction des valeurs de levier de chaque molécule (Figures III.6 et III.7) à partir du quel on peut déterminer les molécules correspondantes aux points aberrants .

D'après les figures III.6 et III.7 des modèles MLR et MNLR, la plupart des composés de l'ensemble de données se trouvent dans cette zone, respectivement Ce graphique montre que les valeurs de levier (hi) de tout composé dans les ensembles d'apprentissage et de test sont inférieures à la valeur critique (h * = 0,72). De plus, les résidus normalisés de tous les composés dans les ensembles d'apprentissage et de test sont inférieurs à trois unités valeurs résiduelles (± 3σ) (valeurs bornées entre -3 et 3) , ce qui prouve l'absence d'observations aberrantes (outliers).

Les résultats de validation interne, externe et le domaine d'applicabilité montrent que le modèle QSAR élaboré dans ce travail peut être utilisé pour des objectifs prédictifs pour les composés . Par conséquent, l'activité prévue par les modèles MLR et MNLR développés est

fiable. En outre, toute valeur des données pIC50 attendues doit être considérée comme fiable pour les véhicules qui entrent dans cette déclaration de modèle.

Les résultats obtenus lors de l'analyse des points aberrants et points à grand valeur de levier pour notre modèle sont affichés dans le tableau III.13.

Tableau III.13 : Valeurs résiduelles normalisées et valeurs de levier MLR et MNLR :

Modèle	**MLR**		**MNLR**	
N	**δi**	***Hi***	**δi**	***Hi***
1	1,203	0,038	0,971	0,038
2	0,149	0,559	0,354	0,559
3	-0,126	0,326	-0,587	0,326
4	-1,956	0,070	-1,844	0,070
5	-0,128	0,172	-0,326	0,172
6	0,473	0,224	-0,156	0,224
7 [t]	-0,365	0,172	-0,208	0,172
8 [t]	-1,444	0,219	-0,673	0,219
9	1,452	0,085	1,518	0,085
10 [t]	1,359	0,054	1,124	0,054
11	-0,311	0,049	-0,556	0,049
12	1,070	0,039	0,954	0,039
13	-0,851	0,687	-0,243	0,687
14	-0,161	0,039	-0,077	0,039
15	-0,804	0,068	-0,44	0,068
16	0,356	0,475	0,233	0,475
17 [t]	1,075	0,078	1,675	0,078
18	0,375	0,094	0,924	0,094
19 [t]	0,351	0,124	0,677	0,124
20 [t]	-0,710	0,189	-0,99	0,189
21	0,703	0,068	0,144	0,068
22	-0,833	0,060	-1,165	0,060
23	1,085	0,082	0,934	0,082
24	0,063	0,099	-0,209	0,099
25	-0,615	0,112	-0,837	0,112
26	1,647	0,091	0,875	0,091
27	0,001	0,492	0,137	0,492
28 [t]	-0,437	0,131	-1,044	0,131
29	-0,201	0,273	0,37	0,273
30	-0,689	0,204	-0,069	0,204
31	-1,517	0,254	-0,306	0,254
32	-0,385	0,335	-0,598	0,335

Avec t désigne les molécules de test

TableauIII.14 :Les valeurs expérimentales, prédites et résiduelles de pIC50 par MLR et MNLR.

Molécule	Pic50	Préd(Pic50)MLR	Résidu	Préd(Pic50)MNLR	Résidu
1	7,210	6,842	0,368	6,898	0,312
2	5,460	5,422	0,038	5,346	0,114
3	6,790	6,830	-0,040	6,978	-0,188
4	6,740	7,372	-0,632	7,332	-0,592
5	5,840	5,865	-0,025	5,945	-0,105
6	5,330	5,193	0,137	5,380	-0,050
7^{t}	7,510	7,645	-0,135	7,577	-0,067
8^{t}	7,420	7,861	-0,441	7,636	-0,216
9	7,460	6,988	0,472	6,973	0,487
10^{t}	7,080	6,666	0,414	6,719	0,361
11	7,010	7,101	-0,091	7,189	-0,179
12	7,240	6,920	0,320	6,934	0,306
13	6,900	7,179	-0,279	6,978	-0,078
14	7,030	7,049	-0,019	7,055	-0,025
15	6,690	6,949	-0,259	6,831	-0,141
16	7,850	7,723	0,127	7,775	0,075
17^{t}	7,150	6,801	0,349	6,612	0,538
18	7,290	7,162	0,128	6,993	0,297
19^{t}	7,590	7,480	0,110	7,373	0,217
20^{t}	6,040	6,579	-0,539	6,786	-0,746
21	7,030	6,838	0,192	6,984	0,046
22	6,870	7,155	-0,285	7,244	-0,374
23	7,520	7,161	0,359	7,220	0,300
24	7,120	7,089	0,031	7,187	-0,067
25	5,940	6,132	-0,192	6,209	-0,269
26	6,250	5,696	0,554	5,969	0,281
27	6,230	6,212	0,018	6,186	0,044
28^{t}	5,440	5,553	-0,113	5,775	-0,335
29	5,820	5,899	-0,079	5,701	0,119
30	5,850	6,083	-0,233	5,872	-0,022
31	5,310	5,785	-0,475	5,408	-0,098
32	5,080	5,212	-0,132	5,272	-0,192

Les molécules de test sont marqués d'un astérisque (^{t})

La figure III.8 représente la représentation linéaire des valeurs prédites en fonction des valeurs expérimentales pour les deux sous ensembles d'apprentissage et de validation (TSET et PSET). Nous observons que les valeurs prédites projetées en fonction du score forment approximativement une droite linéaire dans un intervalle de confiance de 95%. Ce qui montre que les valeurs sont distribuées normalement soit dans le modèle MLR ou MNLR .ll indique que les modèles peut être appliqué avec succès pour prédire activités des inhibiteurs de la 6-arylquinazoline-4-amine de DYRK1A.

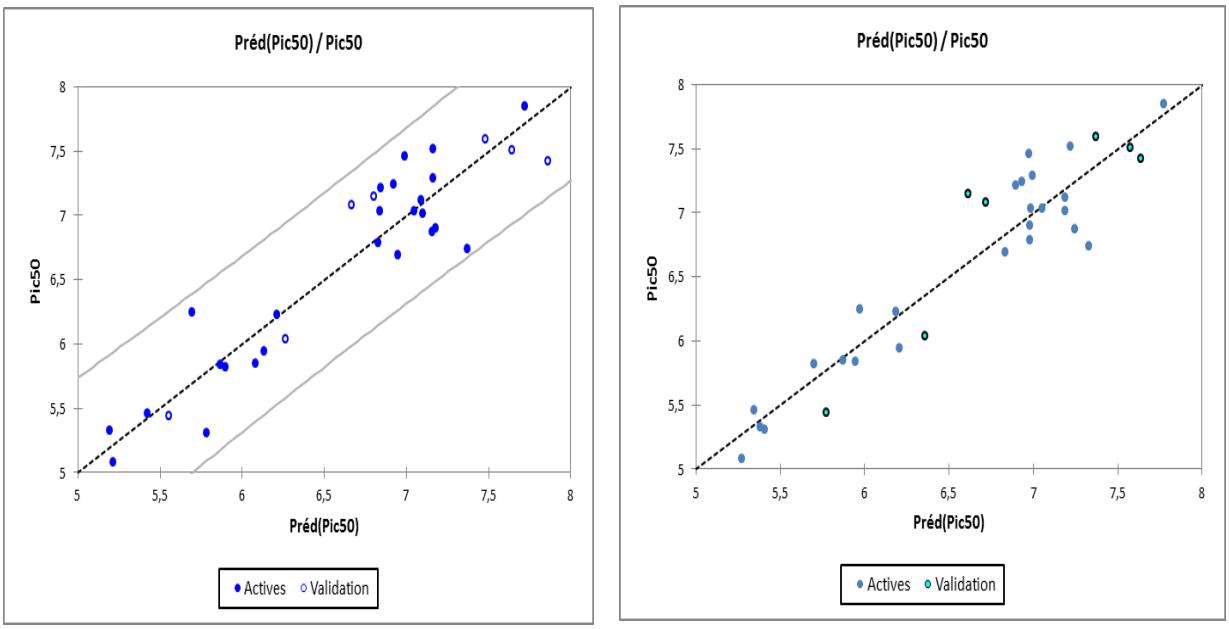


Figure III.8: les courbes des valeurs prédictives en fonction des valeurs expérimentales de pIC50 pour MLR et MNLR

Pour étudier la présence d'une erreur systématique dans le développement des modèles QSAR, les valeurs résiduelles des valeurs prédites du log d'activité biologique (1 / IC50) ont été tracées par rapport aux valeurs expérimentales

La propagation des résidus des deux côtés de zéro indique qu'il n'y a pas d'erreur systémique, comme le suggèrent Jalali-Heravi et Kyani [129].

Les figures III.9 et III.10 montrentles courbe de régression linéaire et non linéaire (MLR et MNLR) des valeurs résidues par rapport aux valeurs expérimentales de l'activité biologique de 6-arylquinazolin-4-amine .Les courbes présentent l'ensamble d'entrainement (training set) et l'ensemble d'essai (validation) pour les deux modèles. Les résultats indiquent ces deux modèles peuvent être appliqués avec succès pour prédire les activités biologiques de l'ensemble 6-arylquinazoline-4-amine utilisée dans le développement des modèles QSAR.

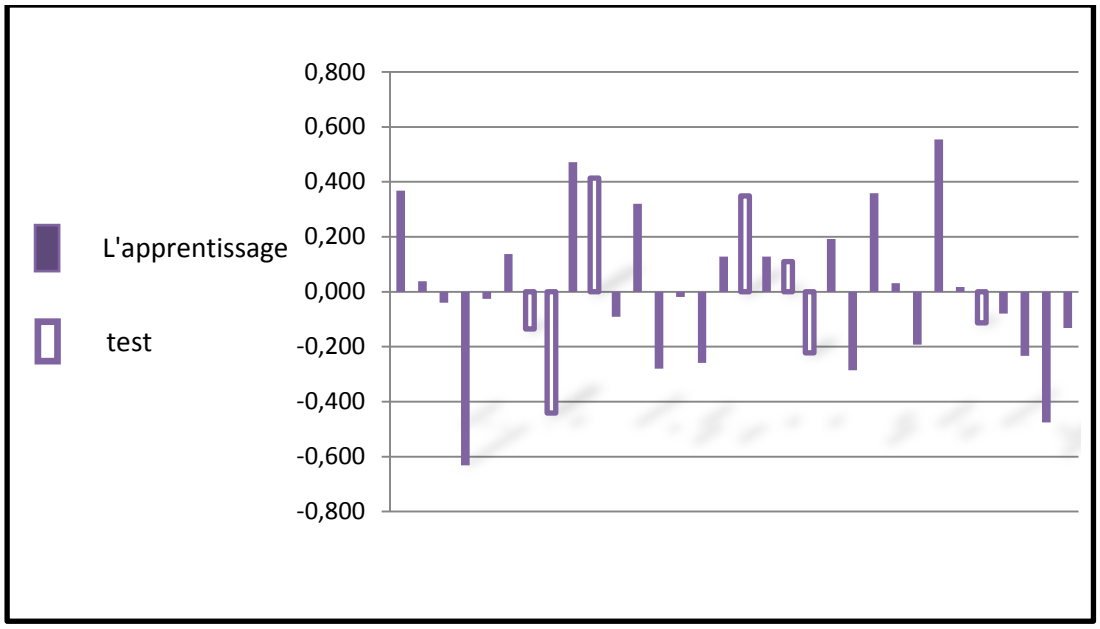


Figure III.9: La courbe des valeurs résiduelles par rapport à l'expérimentale par MLR

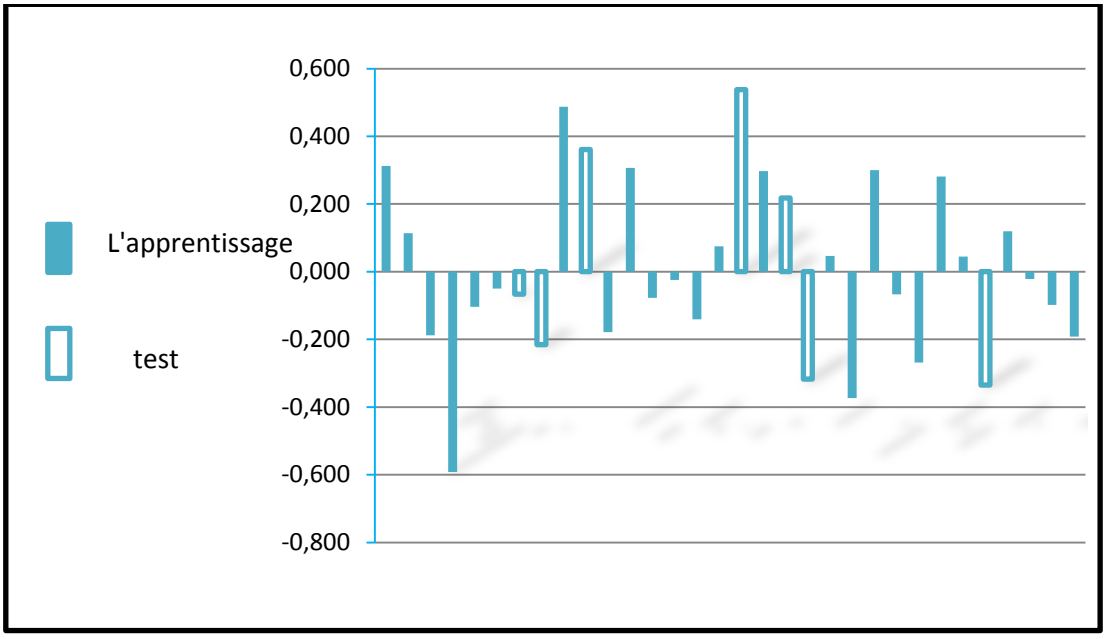


Figure III.10 : La courbe des valeurs résiduelles par rapport à l' expérimentale par MNLR

Conclusion Générale

Conclusion générale :

La recherche d'une relation entre la structure chimique et l'activité biologique ou d'autres propriétés est d'un grand intérêt pour l'industrie pharmaceutique. Connues sous le nom de relation structure-activité quantitative (QSAR), elles sont utilisées pour prédire les activités de molécules en se basant uniquement sur leurs structures chimiques. Afin d'obtenir de bonnes relations prédictives, il est nécessaire de découvrir et d'utiliser le jeu particulier de descripteurs moléculaires ayant de bonnes corrélations avec l'activité biologique cible.

Comme nous l'avons mentionné précédemment, l'objectif de notre travail est de modéliser les activités inhibitrices de DYRK1A pour former des modèles de QSAR robustes, stables, et précis capables de prédire efficacement ces activités.

QSAR se trouve à l'intersection de la chimie, les statistiques et la biologie, pour cette raison les deux premiers chapitres de cette thèse ont été consacrés à la présentation de la maladie d'Alzheimer et Influence de la protéine kinase DYRK1A sur la maladie , la méthodologie QSAR/QSPR ainsi tous les outils nécessaires pour la mise en place d'un modèle QSAR.

Une étude de la relation quantitative structure-activité (QSAR) est appliquée à un ensemble de 32 molécules dérivés des 6-arylquinazolin-4-amine , afin de prédire l'activité biologique inhibitrice de DYRK1A exprimée des composés à tester et de trouver une corrélation entre les différents paramètres moléculaires (descripteurs) de ces composés et son activité biologique, en utilisant l'analyse en composantes principales (ACP) pour la sélection de variables (descripteurs), la régression linéaire multiple (MLR), la régression multiple non linéaire (MNLR) et les modèles résultants ont été comparés. Les modèles est obtenu en utilisant la totalité des descripteurs issus du serveur logiciel XLstat version 2020.

Les résultats statistiques du MLR, du MNLR indiquent que les coefficients de détermination R^2 étaient respectivement de 0.867 ; 0.902 et coefficient de détermination R^2ajusté 0.831 ; 0.876 respectivement. Les résultats montrent que les méthodes de modélisation permettent de bien prédire l'activité étudiée et peuvent être utiles pour prédire l'activité biologique de nouveaux composés. la validation des modèles a été utilisée pour déterminer la qualité statistique et le pouvoir prédictif du QSAR des deux modèles MLR et

MNLR. les modèles MLR et MNLR bien générés présentent les coefficients de validation croisée Q^2 (R^2_{CV}) de 0,867 et 0,902 respectivement.

En outre, la capacité prédictive de ces modèles a été évaluée par la validation externe en utilisant un ensemble de test de sept composés avec des coefficients de détermination prédits R^2_{pred} de 0,881 et 0,850, respectivement et critères de Tropsha et aussi coefficient de randomisation Y (cR^2_p) 0.757 et 0.656 respectivement.

Le domaine d'applicabilité des modèles proposés a été étudié à l'aide du tracé de William pour détecter le sous-espace des structures chimiques (pour détecter les valeurs aberrantes et les composés extérieurs) qui peuvent être prédits de manière fiable par les modèles. Les méthodes proposées réduiront le temps et le coût de synthèse et de détermination de l'activité .

Enfin, nous avons pu soutirer des informations utiles à partir des équations des modèles. Ainsi, nous avons pu interpréter la contribution des descripteurs sur la variation de l'activité biologiques .

A travers les différents résultats obtenus au cours de ce travail, nous pouvons dire que le but que nous nous sommes fixés au départ a été largement atteint. Ainsi, les ensembles de molécules, les traitements statistiques et les techniques informatiques utilisés, lors du développement et l'analyse des modèles de QSAR, ont donné de bons résultats, ce qui nous permet d'entrevoir des perspectives assez prometteuses dans ce domaine par l'amélioration des traitements statistiques et par l'utilisation d'autres méthodes de sélection.

Références bibliographiques:

Références bibliographiques:

[1] Prince, M.; Wimo, A.; Guerchet, M.; Ali, G. C.; Wu, Y. T.; &Prina, M. Alzheimer's Disease International: World Alzheimer Report 2015: The Global Impact of Dementia: an Analysis of Prevalence, Incidence, Cost and Trends. 2015. Alzheimer'sDisease International: London. (2019).

[2] Herault, Y et Meijer, L .DYRK1A, la star des protéines !. la lettre de la fondation Jérôme Lejeune, 103,(2017), 6 .

[3] ACAR J. & COURVALIN, P. La fin de l'age d'or des antibiotiques: Il est grand temps d'agir: certaines bactéries échappent à tous les traitements: Antibiotiques la résistance des bactéries. Recherche (Paris, 1970), (314), (1998), 50-52.

[4] Götte, M., Rausch, J. W., Marchand, B., Sarafianos, S., & Le Grice, S. F. Reverse transcriptase in motion: conformationaldynamics of enzyme–substrate interactions. Biochimica et BiophysicaActa (BBA)-Proteins and Proteomics, 1804(5), (2010) ,1202-1212.

[5] Hu, W. S., & Hughes, S. H. HIV-1 reverse transcription. Cold Spring HarbPerspect Med 2: a006882. (2012).

[6] Bakchine,S.; Habert,M.O. Classification des démences : aspects nosologiques .Médecine Nucléaire,31 (6),(2007), 278-293.

[7] Marfai,L. Les difficultesnrencontrees lors du devloppement de nouvelles molécules thérapeutiques dans l'indication de la maladie d'Alzheimer . le diplôme d'etat de docteur en pharmacie . Université de Lorraine . France .(2013).

[8] Cummings, J.L.Alzheimer's Disease . The New England journal of medicine, 351(18),(2004),56-67.

[9] Christelle,C. Les Capacités Musicales Dans La maladie D'Alzheimer. Mémoire présenté en vue de l'obtention du certificat de capacité d'orthophoniste.Université Nice Sophia Antipolis.Nice.(**2012)**

[10] https://www.francealzheimer.org

[11] Feldman, H.H.; Woodward, M.D.The staging and assessment of moderate to severe Alzheimer disease. Neurology, 65(6), (2005) , 10-17.

[12] Mécanismes moléculaires dans les démences neurodégénératives. la maladie d'Alzheimer : aspects lésionnels et moléculaire diagnostiques et thérapeutique. [en ligne] http://www.mmdn.univ-montp2.fr/ . Consulte le [08/04/2015].

[13] Société d'Alzheimer Canada. La maladie d'Alzheimer et les facteurs de risque. [En ligne]. http://www.alzheimer.ca/fr. Consulte le : [08/04/2015].

[14] Schenk F.; Leuba, G.; Büla, C. Du vieillissement cérébral à la maladie d'Alzheimer (Autour de la notion de plasticité).De Boeck Université. (2004).

[15] Ubersax, J.A.; Ferrell Jr, J.E. Mechanisms of specificity in protein phosphorylation . Nat Rev Mol Cell Biol,8(7),(2007), 530-541

[16] Manning, G.;Whyte, D. B.;Martinez, R.;Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. Science, 298(5600), (2002), 1912-1934.

[17] Hanks, S.K.; Quinn, A.M.; Hunter, T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science ,241(4861), (1988),42–52

[18] Hubbard, S. R.; Miller,W.T. Receptor tyrosine kinases: mechanisms of activation and signaling.CurrOpin Cell Biol, 19(2),(2007) 117-123.

[19] Pawson, T.;Kofler ,M. Kinome signaling through regulated protein-protein interactions in normal and cancer cells.CurrOpin Cell Biol, **21**(2), (2009), 147-153.

[20] Lindberg, R. A.; Quinn,A.M .; Hunter ,T. Dual-specificity protein kinases: will any hydroxyl do?. Trends BiochemSci ,**17**(3),(1992),114-119.

[21] Becker, W.;Weber,Y.; Wetzel, K.; Eirmbter, K.; Tejedor, F.J.; Joost, H.G. Sequence characteristics, subcellular localization, and substrate specificity of DYRK-related kinases, a novel family of dual specificity protein kinases. J BiolChem , **273**(40), (1998), 25893-25902.

[22] Souchet, B. Implication de la protéine DYRK1A dans la pathologie Alzeimer et développement de stratégies thérapeutiques. Thèse de doctorat. Université Paris-Saclay. France.(2018).

[23] Tabouy,L. Mise au point d'un dosage de l'activité kinase de la protéine DYRK1A et régulation épigénétique de l'expression du gène codant le facteur de transcription ISL1. Thèse de doctorat. Université Paris Diderot. France .(2012).

[24] Park, J., Oh,Y.; Yoo, L.; Jung,M.S.; Song, W.J.; Lee, S.H.; Seo, H.;Chung, K.C. Dyrk1A phosphorylates p53 and inhibits proliferation of embryonic neuronal cells. J BiolChem, 285,(2010),31895-906.

[25] Wegiel, J., Gong, C.X.; Hwang, Y.W. The role of DYRK1A in neurodegenerative diseases. FEBS J ,278(2), (2011), 236-245

[26] Guedj, F.; Pereira, P.L.; Najas, S.; Barallobre, M.J.; Chabert, C.; Souchet, B.; Sebrie, C.; Verney, C.; Herault, Y.; Arbones, M.; Delabar, J.M. DYRK1A: a master regulatory protein controlling brain growth. Neurobiol Dis, 46(1), (2012), 190-203.

[27] Branca, C.; Shaw, D. M.; Belfiore, R.; Gokhale, V.; Shaw, A. Y.; Foley, C.; ... &Caccamo, A. Dyrk1 inhibition improves Alzheimer's disease-like pathology. Aging cell, 16(5),(2017), 1146-1154.

[28] http://www.maisons-de-retraite.fr/La-sante-des-seniors/Maladie-d-Alzheimer

[29] De Ferrari, G. V.; Canales, M. A.; Shin, I.; Weiner, L. M.; Silman, I.; Inestrosa, N. C.. A structural motif of acetylcholinesterase that promotes amyloid β-peptide fibril formation. Biochemistry, 40(35),(2001), 10447-10457.

[30] Schneider, L. S.; Tariot, P. N. Emerging drugs for Alzheimer's disease: mechanisms of action and prospects for cognitive enhancing medications. Medical Clinics of North America, 78(4),(1994), 911-934.

[31] Fühner,H.,Neubauer,E. Hämolyse durch Substanzen homologer reihen. Archivfür experimentelle Pathologie und Pharmakologie, 56(5-6), (1907), 333-345.

[32] Hansen O.R. Hammett Series with Biological Activity . ActaChemicaScandinavica, 16(7), (1962), 1593–1600.

[33] Hansch C.; & Fujita, T. p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. Journal of the American Chemical Society, 86(8), (1964), 1616-1626.

[34] Free S. M.; & Wilson J. W. A mathematical contribution to structure-activity studies. Journal of Medicinal Chemistry, 7(4), (1964), 395-399.

[35] Hansen, O.C. Quantitative structure-activity relationships (QSAR) and pesticides: Ministry of the Environment . Environmental Protection Agency . The Danish . (2004).

[36] Brown AC.; Fraser TR. V.—On the connection between chemical constitution and physiological action. Part. I.—On the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. Earth and Environmental Science Transactions of The Royal Society of Edinburgh, 25(1), (1868), 151-203.

[37] Srinivas Reddy A,;Kumar S,; Garg R. Hybrid-genetic algorithm based Descriptor Optimization and QSAR Models for Predicting the Biological Activity of Tipranavir Analogs for HIV Protease Inhibition. J Mol Graph Model , 28(8),(2010), 852–862.

[38] Fourches D,; Muratov E,; Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. Journal of chemical information and modeling ,50 (7), (2010), 1189–1204.

[39] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Molecular informatics, 29(6-7), (2010), 476–488.

[40] Chatterjee S,; Hadi ,A. S,; Price B. « Regression analysis by example » .Wiley-Interscience : New York. (2000).

[41] Phuong, H. T. N . « Synthèse et étude des relations structure/activité quantitatives (QSAR/2D) d'analogues Benzo [c] phénanthridiniques » .Thèse de doctorat .Université d'Angers. France . (2007).

[42] Karelson , M. Molecular descriptors in QSAR/QSPR . New York: Wiley-Interscience.(2000).

[43] Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug discovery today,21(8),(2016),1291-1302.

[44] Wiener H . Structural determination of paraffin boiling points. Journal of the American chemical society, 69(1), (1947), 17-20.

[45] Randic M., On characterization of molecular branching. Journal of the American Chemical Society , 97(23), (1975), 6609 -6615.

[46] Kier L.B.; Hall L.H. Derivation and significance of valence molecular connectivity. Journal of Pharmaceutical Sciences,70(6), (1981), 583-589.

[47] Balaban A.T. Highly discriminating distance-based topological index. Chemical Physics Letters, 89(5), (1982), 399 -404.

[48] Puzyn T.; Leszczynski J.; & Cronin M. T. (Eds.). Recent advances in QSAR studies: methods and applications , Springer Science & Business Media. London New York, (2010).

[49] Bosque R .; Sales J.; Bosch E.; Rosès M.; Garcia-Alvarez-Coque M.C.; and Torres-Lapasio J.R . A QSPR study of the p-solute polarity parameter to estimate retention HPLC. Journal of Chemical Information and Modeling, 43(4) , (2003), 1240–1247.

[50] Fukui K. Theory of Orientation and Stereoselection‖. Reactivity and Structure Concepts in Organic Chemistry, 2, (1975), 34–39.

[51] Franke, R. (1984). Theoretical drug design methods. Pharmacochemistry library, 7.(1984), 115– 123.

[52] Lewis D.F.V.; Ioannides C.; and Parke D.V. ―Interaction of a series of nitriles with the alcoholinducible isoform of P450: Computer analysis of structure—activity relationships . Xenobiotica, 24(5), (1994), 401–408.

[53] Zhou Z.; Parr R.G. Activation hardness: new index for describing the orientation of electrophilic aromatic substitution. Journal of the American Chemical Society, 112(15), (1990), 5720–5724.

[54] Dennington, R., Keith, T., &Millam, J. GaussView, version 5. (2009).

[55] Frisch, A. (2009). gaussian 09W Reference. Wallingford, USA, 25p.

[56] Atkins, P.W.; and de Paula, J. Atkins' Physical Chemistry.Oxford University Press. Oxford. (2002).

[57] Viswanadhan V.N.; Ghose A.K .; Revankar G.R .; and Robins R.K. ―Atomic physicochemical parameters for 3D structure directed quantitative structure-activity relationships: Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics‖. Journal of Chemical Information and Modeling, 29, 1989, 163–172.

[58] Taylor P.J.Hydrophobic Properties of Drugs, In Quantitative Drug Design . Pergamon Press, Oxford (UK), 4, (1990), 241–294.

[59] Lipinski C.A.; Lombardo F.; Dominy B.W.; and Feeney P.J.;―Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings . Advanced Drug Delivery Reviews, 23(1–3), (1997), 3–25.

[60] Lorentz H.A. Ueber die Beziehung zwischen der Fortpflanzungsgeschwindigkeit des Lichtes und der Körperdichte. Annalen der Physik, 245(4), (1880), 641-665.

[61] Hansen C.; Telzer B. R., and Zhang L.T. Comparative QSAR in toxicology: examples from teratology and cancer chemotherapy of aniline mustards. Critical reviews in toxicology, 25(1), (1995), 67-89.

[62] Cammarata A. An apparent correlation between the in vitro activity of chloramphenicol analogs and electronic polarizability. Journal of Medicinal Chemistry, 10(4), (1967), 525-527.

[63] Leo A.; Hansch C.; and Church C.; Comparison of parameters currently used in the study of structure-activity relationships, Journal of Medicinal Chemistry, 12(5), (1969), 766–771.

[64] Hansch C.; and Coats E . α-Chymotrypsin: A Case Study of Substituent Constants and Regression Analysis in Enzymic Structure—Activity Relationships. Journal of Pharmaceutical Sciences, 59(6), (1970), 731-743.

[65] Autin , L. Analyse des systèmes tenase et prothrombinase par bioinformatique structurale: prédiction de complexes macromoléculaires et proposition d'agents anti-coagulants, Thèse de Doctorat .Université Paris 5. France .(2005).

[66] McQuarrie ,D.A. Statistical Thermodynamics. harper row publishers. New York.(1973).

[67] Akhiezer A.I .; and Peltminskii S.V. Methods of Statistical Physics . Pergamon Press. Oxford. (1981).

[68] Atkins P.W. Physical Chemistry , 2nd edition, W.H. Freeman and Company. San Francisco. American. (1982).

[69] Kalivas , J. H.(Ed). Adaption of simulated annealing to chemical optimization problems. Elsevier Science B.V . Pays Bas . (1995).

[70] Cornillon, P. A. ;&Matzner-Lober, E. Régression: théorie et applications . Springer. (2007).

[71] Borcard, D. Régression multiple. Bio-2042 .Département de sciences biologiques. Université de Montréal.Montréal (in French). (2009).

[72] Domingo, L. R., Chamorro, E., & Pérez, P. (2008). Understanding the reactivity of captodative ethylenes in polar cycloaddition reactions. A theoretical study. The Journal of organic chemistry, 73(12), 4615-4624.

[73] Becke A.D. Density-functional thermochemistry. III. The role of exact exchange J. Chem. Phys, 98,(1993), 5648.

[74] Lee, C.;Yang, W.;& Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Physical review B, 37(2), (1988),785.

[75] Hohenberg, P.;& Kohn, W. Phys Rev 136: B864. Kohn W, Sham LJ (1965) Phys Rev, 140, (1964), A1133.

[76] Kohn W.; Becke A. D.;& Parr R. G. Density functional theory of electronic structure. The Journal of Physical Chemistry, 100(31), (1996), 12974-12980.

[77] Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to linear regression analysis. John Wiley & Sons ,(Vol. 821),(2012),672 pages.

[78] Abdi H.; & Williams L.J. Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), (2010), 433-459.

[79] Husson, F.; Lê, S.;&Pagès, J. Exploratory multivariate analysis by example using R. CRC press.(2017).

[80] Jollife, I.T. 2002. Principal Component Analysis. 2nd ed. New York: Springer- Verlag. https://goo.gl/SB86SR.

[81] An Introduction to Partial Least Squares Regression.http://www.ats.ucla.edu/ stat/sas/library/pls.pdf

[82] Wold S.; Sjöström M.; & Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 58(2), (2001),109-130.

[83] Esposito E. X.; Hopfinger A. J.; & Madura J. D. Methods for applying the quantitative structure-activity relationship paradigm. In Chemoinformatics, HumanaPress, 275, (2004), 131-213.

[84] J. Jacques, Modélisation Statistique‖. Availableat: http://eric.univ-lyon2.fr/~jjacques/Download/Cours/Mod-Cours.pdf

[85] Besse, P. Pratique de la modélisation statistique. Publications du laboratoire de statistiques et probabilités. Université Paul Sabatier, Toulouse. Disponiblea partir de l'URL http://www-sv. cict. fr/lsp/Besse. France. (2003).

[86] P. Besse, ―Apprentissage Statistique Data mining‖, Publications du laboratoire de statistique et Probabilités, University Toulouse: INSA.France .(2009).

[87] Mclachlan G.J. Discriminant analysis and Statistical Pattern Recognition. Wiley. University of Queensland .New-York. (1992).

[88] Nakache, J. P., &Confais, J. Statistique explicative appliquée: analyse discriminante, modèle logistique, segmentation par arbre. Paris. Editions Technip. (2003).

[89] Saporta, G. Probabilités, analyse des données et statistique. France. Editions Technip. (2006).

[90] Laffly D. Régression multiple : principes et exemples d'application‖, Université de Pau et des Pays de l'Adour. France. (2006).

[91] Shao J. Linear model selection by cross-validation. Journal of the American statistical Association, 88(422), (1993), 486-494.

[92] Tropsha A.; Gramatica P.; and Gombar V.K . the importance of Being Earnest: Validation is the Absolute Essential for Successful Application and interpretation of QSPR Models . QSAR and Combinatorial Sciences, 22(1), (2003), 69–77.

[93] Golbraikh A.; and TropshaA.Beware of q2!. Journal of Molecular Graphics and Modelling , 20(4), (2002), 269–276.

[94] Refaeilzadeh P., Tang L., & Liu H. Cross-Validation. Encyclopedia of database systems, 5, (2009), 532-538.

[95] Hawkins D.M. The problem of overfitting . Journal of chemical information and computer sciences, 44(1), (2004), 1-12.

[96] Zhang L.; Zhu H.; Oprea T. I.; Golbraikh A.; &Tropsha A. QSAR modeling of the blood–brain barrier permeability for diverse organic compounds. Pharmaceutical research, 25(8), (2008), 1902- 1914.

[97] artin T. M.; Harten P.; Young D. M.; Muratov E. N.; Golbraikh A.; Zhu H.; &Tropsha A. Does rational selection of training and test sets improve the outcome of QSAR modeling?. Journal of chemical information and modeling, 52(10), (2012). 2570-2578.

[98] Tetko I.V.; Sushko I.; Pandey A.K.; Zhu H.; Tropsha A.; Papa E.; Oberg T.; Todeschini R.; Fourches D.; and Varnek A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena-pyriformis: focusing on applicability domain and overfitting by variable selection‖, Journal of Chemical Information and Modeling, 48(9), (2008), 1733–1746.

[99] Jaworska, J., Nikolova-Jeliazkova, N., &Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: a review. Alternatives to laboratory animals, 33(5), (2005), 445-459.

[100] OECD, O. (2004). Principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models.

[101] Eriksson L.; Jaworska J.; Worth A. P.; Cronin M. T.; McDowell R. M.; &Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. Environmental health perspectives, 111(10), (2003), 1361-1375.

[102] Gramatica, P. Principles of QSAR models validation: internal and external. QSAR & combinatorial science, 26(5), (2007),694-701.

[103] Netzeva T.I .; Worth A.P.; Aldenberg T.;Benigni R.; Cronin M.T.D.; Gramatica P.; Jaworska J.S., Kahn S.; Klopman G.; Marchant C.A.; Myatt G.; Nikolova-Jeliazkova N.; Patlewicz G.Y.; Perkins R.; Roberts D.W.; Schultz T.W.; Stanton D.T.; Van De Sandt J.J.M.; Tong W.; Veith G.; and Yang C.; Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships :The report and recommendations of ecvam workshop 52. Alternatives to Laboratory Animals, 33(2), (2005), 155–173.

[104] Dearden J.C. the History & Development of Quantitative Structure-Activity Relationships (QSARs), International Journal of Quantitative Structure-Property Relationships, 1(1), (2016), 1-44.

[105] TOUHAMI , M . Modélisation de l'activité biologique de composés hétérocycles .Le diplôme de Doctorat LMD en « CHIMIE COMPUTATIONNELLE ». Université de TAHAR MOULAY – SAIDA- . Algérie. (2019).

[106] GUENDOUZI, A. Élaboration des modèles QSPR prédictifs des propriétés physico-chimiques à l'aide des descripteurs moléculaires . Thèse de doctorat. Université de ABOU BEKR BELKAÏD DE TLEMCEN. Algérie. (2015).

[107] Fayet, G. Développement de modèles QSPR pour la prédiction des propriétés d'explosibilité des composés nitroaromatiques. Thèse de doctorat. Université de PIERRE ET MARIE CURIE. France . (2010).

[108] Chtita, S. Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments . Thèse de doctorat. Université de MOULAY ISMAIL .Maroc. (2017).

[109] Khairedine, K. études de QSAR sur des activités biologiques utilisant des produits d'origines naturels .These de Doctorat. Université 8 mai 1945 de Guelma . Algérie . (2009).

[110] SEBAA, Z. Etude computationnelle de la relation structure- activité dans des séries de composés hétérocycliques à intérêt thérapeutique, Thèse de Doctorat LMD . Université Mohamed Boudiaf d'Oran .Algérie . 2018-2019.

[111] Benazzouz, H. ; Khebiza, A. Relation Structure Activité: Etude Qualitative et Quantitative et Développement de Recherche sur les Coumarines. Thèse de Doctorat . Université ABOU BEKR BELKAÎD DE TLEMCEN. Algérie. (2018).

[112] BELLIFA, K . Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires« Modélisation QSAR » . Thèse de Doctorat. Université ABOU BEKR BELKAÏD DE TLEMCEN. Algérie. (2015).

[113] SAIHI ,Y . Etude de la relation quantitative structure-activité inhibitrice des enzymes hydrolytiques : cas des alpha-glucosidases . Thèse de doctorat. Université BADJI MOKHTAR d'ANNABA. Algérie. (2015).

[114] YOUSFI, Y. ETUDE QSAR DE l'ACTIVITE ANTI-OXYDANTE D'UNE SERIE DE COMPOSES PHENOLIQUES, thèse de Master. Université ABOU-BEKR BELKAID - TLEMCEN. Algérie. (2017).

[115] Fernández, M .; Caballero, J.;Helguera, A. M .; Castro, E. A.; González, M. P. Quantitative structure–activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds. Bioorganic & medicinal chemistry, 13(9), (2005), 3269-3277.

[116] Saíz-Urra, L.; González, M. P.; Fall, Y.; Gómez, G. Quantitative structure–activity relationship studies of HIV-1 integrase inhibition. 1. GETAWAY descriptors. European journal of medicinal chemistry, 42(1), (2007), 64-70.

[117] Tropsha, A.; Weifan, Z. Identification of the Descriptor Pharmacophores Using Variable Selection QSAR Applications to Database Mining. Current pharmaceutical design, 7(7), (2001), 599-612.

[118] Leal, F. D.; da Silva Lima, C. H.; De Alencastro, R. B.; Castro, H. C., Rodrigues, C. R.; Albuquerque, M. G. Hologram QSAR models of a series of 6-arylquinazolin-4-amine inhibitors of a new Alzheimer's disease target: dual specificity tyrosine-phosphorylation-regulated kinase-1A enzyme. International Journal of Molecular Sciences, 16(3), (2015), 5235-5253.

[119] ACD/Labs Extension for ChemDraw Version 7.0 for Microsoft Windows User's Guide

[120] González-Díaz, H.; Olazábal, E.; Santana, L.; Uriarte, E.; González-Díaz, Y.;Castanedo, N.. QSAR study of anticoccidial activity for diverse chemical compounds: prediction and experimental assay of trans-2-(2-nitrovinyl) furan . Bioorganic & Medicinal Chemistry, 15(2), (2007), 962-968.

[121] Hyper chem 8.0.6. Molecular Visualization and Simulation Program Package, Hypercube Inc. Gainesville, 2007

[122] XLSTAT 2020 Add-in software (XLSTAT Company). www.xlstat.com..

[123] Ousaa, A.; Elidrissi, B.;Ghamali, M.; Chtita, S.;Aouidate, A.;Bouachrine, M.;Lakhlifi, T.Quantitative structure-toxicity relationship studies of aromatic aldehydes to Tetrahymenapyriformis based on electronic and topological descriptors. Journal of Materials and Environmental Science, 9(1),(2018), 256-266.

[124] Ajeet .; Bijandre, K. Quantitative structure activity relationship (QSAR) modeling of 2-x-5,8-Dimethoxy-1,4- naphthoquinone against L1210 cells. International Journal of Pharmacy and Pharmaceutical Sciences, 4(4), (2012), 445-448.

[125] Chtita, S.; Ghamali, M.; Larif, M.; Hmamouchi, R.; Bouachrine, M.; Lakhlifi, T.. Quantitative Structure–Activity Relationship Studies of Anticancer Activity for Isatin (1H-Indole-2, 3-Dione) Derivatives Based on Density Functional Theory. International Journal of Quantitative Structure-Property Relationships (IJQSPR), 2(2), (2017), 90-115

[126] Allinger, N. L. QCPE 1982, 3, 32.(b) Beckhaus, H. D. Chem. Ber, 116(115), (1983), 22-3263.

[127] Y-randomisation The standalone QSAR-tools ("Programs") available online at http://dtclab.webs.com/softwaretool sandhttp://teqip.jdvu.ac.in/QSAR_Tools ("Websites") was employed in the y-randomization.

[128] Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P.. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. Environmental health perspectives, 111(10),(2003), 1361-1375.

[129] Jalali-Heravi, M.; Kyani, A. Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: a PCA-MLR-ANN approach. Journal of chemical information and computer sciences, 44(4), (2004), 1328-1335.

ANNEXE A

Table 6: Loi du *F* de Fisher (suite)

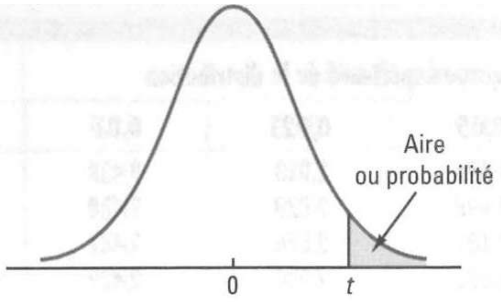
$$P(F_{\nu_1,\nu_2} < f_{\nu_1,\nu_2,\alpha}) = \alpha$$

$\alpha = 0{,}95$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	•
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	253	254	254	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,62	8,58	8,55	8,54	8,53	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,70	5,66	5,65	5,64	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,44	4,41	4,39	4,37	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,75	3,71	3,69	3,68	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,38	3,32	3,27	3,25	3,24	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,02	2,97	2,95	2,94	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,80	2,76	2,73	2,72	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,64	2,59	2,56	2,55	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,57	2,51	2,46	2,43	2,42	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,47	2,40	2,35	2,32	2,31	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,38	2,31	2,26	2,23	2,22	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,31	2,24	2,19	2,16	2,14	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,18	2,12	2,10	2,08	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,19	2,12	2,07	2,04	2,02	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,15	2,08	2,02	1,99	1,97	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,11	2,04	1,98	1,95	1,93	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,07	2,00	1,94	1,91	1,89	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,97	1,91	1,88	1,86	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,15	2,07	1,98	1,91	1,85	1,82	1,80	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,11	2,03	1,94	1,86	1,80	1,77	1,75	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,07	1,99	1,90	1,82	1,76	1,73	1,71	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,04	1,96	1,87	1,79	1,73	1,69	1,67	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,84	1,76	1,70	1,66	1,64	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,74	1,66	1,59	1,55	1,53	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,69	1,60	1,52	1,48	1,46	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,56	1,48	1,44	1,41	1,39
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,79	1,70	1,60	1,51	1,43	1,38	1,35	1,32
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,57	1,48	1,39	1,34	1,31	1,28
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,72	1,62	1,52	1,41	1,32	1,26	1,22	1,19
500	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,69	1,59	1,48	1,38	1,28	1,21	1,16	1,11
•	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,46	1,35	1,24	1,17	1,11	1,00

ANNEXE B

Table 4: Loi du *t* de Student



Les chiffres de la table correspondent aux valeurs *t* pour différentes aires ou probabilités situées dans la queue supérieure de la distribution de Student. Par exemple, avec 10 degrés de liberté et une aire de 0,05 dans la queue supérieure de la distribution, $t_{0,05}$ = 1,812. (pour test unilatéral !)

Degrés de liberté	Aire dans la queue supérieure de la distribution					
	0,20	0,10	0,05	0,025	0,01	0,005
1	1,376	3,078	6,314	12,706	31,821	63,656
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250
10	0,879	1,372	1,812	2,228	2,764	3,169
11	0,876	1,363	1,796	2,201	2,718	3,106
12	0,873	1,356	1,782	2,179	2,681	3,055
13	0,870	1,350	1,771	2,160	2,650	3,012
14	0,868	1,345	1,761	2,145	2,624	2,977
15	0,866	1,341	1,753	2,131	2,602	2,947
16	0,865	1,337	1,746	2,120	2,583	2,921
17	0,863	1,333	1,740	2,110	2,567	2,898
18	0,862	1,330	1,734	2,101	2,552	2,878
19	0,861	1,328	1,729	2,093	2,539	2,861
20	0,860	1,325	1,725	2,086	2,528	2,845
21	0,859	1,323	1,721	2,080	2,518	2,831
22	0,858	1,321	1,717	2,074	2,508	2,819
23	0,858	1,319	1,714	2,069	2,500	2,807
24	0,857	1,318	1,711	2,064	2,492	2,797
25	0,856	1,316	1,708	2,060	2,485	2,787
26	0,856	1,315	1,706	2,056	2,479	2,779
27	0,855	1,314	1,703	2,052	2,473	2,771
28	0,855	1,313	1,701	2,048	2,467	2,763
29	0,854	1,311	1,699	2,045	2,462	2,756
30	0,854	1,310	1,697	2,042	2,457	2,750
31	0,853	1,309	1,696	2,040	2,453	2,744
32	0,853	1,309	1,694	2,037	2,449	2,738
33	0,853	1,308	1,692	2,035	2,445	2,733
34	0,852	1,307	1,691	2,032	2,441	2,728

Table 4: Loi du *t* de Student (suite)

Degrés de liberté	Aire dans la queue supérieure de la distribution					
	0,20	0,10	0,05	0,025	0,01	0,005
35	0,852	1,306	1,690	2,030	2,438	2,724
36	0,852	1,306	1,688	2,028	2,434	2,719
37	0,851	1,305	1,687	2,026	2,431	2,715
38	0,851	1,304	1,686	2,024	2,429	2,712
39	0,851	1,304	1,685	2,023	2,426	2,708
40	0,851	1,303	1,684	2,021	2,423	2,704
41	0,850	1,303	1,683	2,020	2,421	2,701
42	0,850	1,302	1,682	2,018	2,418	2,698
43	0,850	1,302	1,681	2,017	2,416	2,695
44	0,850	1,301	1,680	2,015	2,414	2,692
45	0,850	1,301	1,679	2,014	2,412	2,690
46	0,850	1,300	1,679	2,013	2,410	2,687
47	0,849	1,300	1,678	2,012	2,408	2,685
48	0,849	1,299	1,677	2,011	2,407	2,682
49	0,849	1,299	1,677	2,010	2,405	2,680
50	0,849	1,299	1,676	2,009	2,403	2,678
51	0,849	1,298	1,675	2,008	2,402	2,676
52	0,849	1,298	1,675	2,007	2,400	2,674
53	0,848	1,298	1,674	2,006	2,399	2,672
54	0,848	1,297	1,674	2,005	2,397	2,670
55	0,848	1,297	1,673	2,004	2,396	2,668
56	0,848	1,297	1,673	2,003	2,395	2,667
57	0,848	1,297	1,672	2,002	2,394	2,665
58	0,848	1,296	1,672	2,002	2,392	2,663
59	0,848	1,296	1,671	2,001	2,391	2,662
60	0,848	1,296	1,671	2,000	2,390	2,660
61	0,848	1,296	1,670	2,000	2,389	2,659
62	0,847	1,295	1,670	1,999	2,388	2,657
63	0,847	1,295	1,669	1,998	2,387	2,656
64	0,847	1,295	1,669	1,998	2,386	2,655
65	0,847	1,295	1,669	1,997	2,385	2,654
66	0,847	1,295	1,668	1,997	2,384	2,652
67	0,847	1,294	1,668	1,996	2,383	2,651
68	0,847	1,294	1,668	1,995	2,382	2,650
69	0,847	1,294	1,667	1,995	2,382	2,649
70	0,847	1,294	1,667	1,994	2,381	2,648
71	0,847	1,294	1,667	1,994	2,380	2,647
72	0,847	1,293	1,666	1,993	2,379	2,646
73	0,847	1,293	1,666	1,993	2,379	2,645
74	0,847	1,293	1,666	1,993	2,378	2,644
75	0,846	1,293	1,665	1,992	2,377	2,643
76	0,846	1,293	1,665	1,992	2,376	2,642
77	0,846	1,293	1,665	1,991	2,376	2,641
78	0,846	1,292	1,665	1,991	2,375	2,640
79	0,846	1,292	1,664	1,990	2,374	2,639

Table 4: Loi du *t* de Student (suite)

Degrés de liberté	Aire dans la queue supérieure de la distribution					
	0,20	0,10	0,05	0,025	0,01	0,005
80	0,846	1,292	1,664	1,990	2,374	2,639
81	0,846	1,292	1,664	1,990	2,373	2,638
82	0,846	1,292	1,664	1,989	2,373	2,637
83	0,846	1,292	1,663	1,989	2,372	2,636
84	0,846	1,292	1,663	1,989	2,372	2,636
85	0,846	1,292	1,663	1,988	2,371	2,635
86	0,846	1,291	1,663	1,988	2,370	2,634
87	0,846	1,291	1,663	1,988	2,370	2,634
88	0,846	1,291	1,662	1,987	2,369	2,633
89	0,846	1,291	1,662	1,987	2,369	2,632
90	0,846	1,291	1,662	1,987	2,368	2,632
91	0,846	1,291	1,662	1,986	2,368	2,631
92	0,846	1,291	1,662	1,986	2,368	2,630
93	0,846	1,291	1,661	1,986	2,367	2,630
94	0,845	1,291	1,661	1,986	2,367	2,629
95	0,845	1,291	1,661	1,985	2,366	2,629
96	0,845	1,290	1,661	1,985	2,366	2,628
97	0,845	1,290	1,661	1,985	2,365	2,627
98	0,845	1,290	1,661	1,984	2,365	2,627
99	0,845	1,290	1,660	1,984	2,364	2,626
100	0,845	1,290	1,660	1,984	2,364	2,626
∞	0,842	1,282	1,645	1,960	2,326	2,576

Résumé

La kinase-1A régulée par la tyrosine à phosphorylation à double spécificité DYRK1A est une enzyme directement impliquée dans la maladie d'Alzheimer.

Notre travail consiste à étudier l'inhibition de protéine kinase DYRK1A en utilisant des modèles statistiques QSAR comme solution au problème de difficulté de calcul direct des propriétés moléculaires et biologiques de la structure. Une étude QSAR a été effectuée sur 32 molécules de dérivés de 6-arylquinazolin-4-amine . La régression linéaire multiple (MLR) et la régression non linéaire multiple (MNLR) a été utilisée pour quantifier les relations entre les descripteurs moléculaires et la propriété inhibitrice de DYRK1A des dérivés 6-arylquinazolin-4-amine . La prédiction des modèles obtenus a été confirmé par la méthode de validation interne, externe et la technique de randomisation Y leur domaine chimique structurel. Une forte corrélation a été observée entre les valeurs expérimentales et prédites des activités biologiques, se qui indique la validité et la qualité des modèles QSAR obtenus.

Mots clés: DYRK1A, Alzheimer, 6-arylquinazolin-4-amine, QSAR, MLR, MNLR.

Abstract

Dual Specificity Tyrosine-Phosphorylation-Regulated Kinase-1A Enzyme DYRK1A is an enzyme that give rise to Alzheimer's disease .

Our work is to study the inhibition of protein kinase DYRK1A inhibitor using QSAR statistical models as a solution to the problem of direct computational difficulty of the molecular and biological properties of the structure. A QSAR study was conducted on 32 molecules derived from 6-arylquinazolin-4-amine. the multiple linear regression MLR and the multiple non-linear regression MNLR was used to quantify the relationships between molecular descriptors and DYRK1A inhibitory property of 6-arylquinazolin-4-amine derivatives. The prediction of the resulting models was confirmed by the internal, external validation method and the randomization technique Y their structural chemical domain . A strong correlation was observed between the experimental and predicted values of biological activities, indicating the validity and quality of the QSAR models obtained.

Keywords: DYRK1A, Alzheimer, 6-arylquinazolin-4- amine, QSAR, MLR, MNLR

الملخص

البروتين كيناز الثنائي المنظم للتيروزين DYRK1A هو انزيم متورط بشكل مباشر في مرض الزهايمر. يتمثل عملنا في دراسة تثبيط بروتين كيناز DYRK1A و ذلك باستخدام نماذج QSAR الاحصائية كحل لمشكلة صعوبة الحساب المباشر للخصائص الفيزيائية و البيولوجية للهيكل .أجريت دراسة QSAR على اثنين وثلاثين جزيء من مشتقات6-arylquinazolin-4- amine . تم استخدام الانحدار الخطي المتعدد (MLR) والانحدار غير الخطي المتعدد MNLR لتحديد العلاقات بين الواصفات الجزيئية والخصائص المثبطة لـ DYRK1A من مشتقات 6-arylquinazolin-4- amine. تم تأكيد التنبؤ بالنماذج التي تم الحصول عليها من خلال طريقة التحقق الداخلية وتقنية التوزيع العشوائيY والتحقق الخارجي تم التحقق من مجالها الكيميائي الهيكلي . ولوحظ وجود ارتباط قوي بين القيم التجريبية والقيم المتوقعة للأنشطة البيولوجية ، مما يشير إلى صحة ونوعية نماذج QSAR التي تم الحصول عليها.

الكلمات الاساسية : DYRK1A, الزهايمر6-arylquinazolin-4- amine , QSAR,MNLR , MLR ,