REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DE MEDEA

FACULTE DES SCIENCES



Mémoire

Présenté pour obtenir le diplôme de **Master**

En chimie

Spécialité :
Chimie pharmaceutique

Par :

SADOUKI Abderrahmane
BOUCHEREF Djamal

Thème :

Prédiction de l'activité thérapeutique des analogues d'Anacétrapib en tant qu'inhibiteurs de la CETP à partir d'une modélisation QSAR

Soutenue publiquement le : 22 /06 /2020, Devant le jury composé de :

MAHDI Youcef	Grade : MCA	Université de Médéa	Président
BOUCHERIT Nabila	Grade : MCA	Université de Médéa	Examinateur
BITAM Saïd	Grade :	Etablissement	Encadreur

Année universitaire : 2019/2020

dedicaces

Au nom d'ALLAH clément et Miséricordieux et que le salut de DIEU soit sur son prophète MOHAMED.

Je dédie ce travail

A **mon père** Qui a œuvré pendant des années à notre bien être et qui a Consenti de bien Lourds efforts pour nous permettre de mener à bien nos études. Je tiens à lui témoigner Ma profonde affection.

A **ma Mère** Son soutien tout au long de mes études sa bienveillance et ses conseils ont été très précieux, je tiens à lui témoigner ma profonde affection et reconnaissance

A mes frères, A mes sœurs, A mes oncles, A mes tantes

A mes amies et mes collègues

A mon binôme

A tous ceux qui me sont chères. A tous ceux qui m'aiment.

Remerciements

Nous tenons tout d’abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d’accomplir ce Modeste travail.

En second lieu, nous tenons à remercier notre promoteur, Monsieur BITAM Said, d’avoir accepté d’encadrer ce travail de toute sa compétence scientifique, qui n’a ménagé aucuns efforts pour nous aider dans le développement de notre recherche et qui nous avons guidé durant notre travail avec une disponibilité permanente et nous avons fait bénéficier de ses connaissances en chimie informatique et théorique ; ses conseils et ses commentaires précieux nous a permis de surmonter nos difficultés et de progresser dans nos études.

الملخص

أمراض القلب والأوعية الدموية (CVD) هي واحدة من الأسباب الرئيسية لوفاة البشر. تشير التجارب الأولية إلى أن علاج تثبيط بروتين نقل الكوليسترول (CETP) يسبب الزيادة الأكثر وضوحًا في الكوليسترول. تم ذكر Anacetrapib، وهو مثبط قوي لـ CETP، في العديد من الدراسات على أنه ليس له أي آثار جانبية. لذلك سيكون من المرغوب فيه تطوير طرق حاسوبية لتسهيل فحص اشباهها. في العمل الحالي، سيتم تطوير نماذج العلاقات الكمية الهيكلية والنشاط (QSAR) للتنبؤ بالقدرة الـ ملاجية لنظائر Anacetrapib

الكلمات المفتاحية:أمراض القلب و لأوعية الدموية،الشبكات العصبية، لانحدار الخطيالمتعدد، تثبيط بروتين نقل الكوليسترول،أناسيتابيب.

Résumé

Les maladies cardiovasculaires (MCV) sont l'une des principales causes de décès chez l'homme. Les essais préliminaires indiquent que le traitement d'inhibition de la protéine de transfert d'ester de cholestérol (CETP) provoque l'augmentation la plus prononcée du cholestérol HDL rapportée jusqu'à présent. L'anacétrapib, un puissant inhibiteur de la CETP, a été bien mentionné dans plusieurs études qu'il n'a pas d'effets indésirables. Il serait donc souhaitable de développer des méthodes informatiques pour faciliter le criblage de ses analogues. Dans le présent travail, des modèles de relations quantitatives structure-activité (QSAR) vont être développés pour prédire le potentiel thérapeutique des analogues de l'Anacétrapib.

Mots-clés: QSAR; RNA ; MLR; Inhibiteurs de la CETP; Anacétrapib.

Abstract

Cardiovascular disease (CVD) is one of the leading causes of death in humans. Preliminary trials indicate that treatment for inhibition of the cholesterol ester transfer protein (CETP) causes the most pronounced increase in HDL cholesterol reported so far. Anacetrapib, a potent CETP inhibitor, has been well mentioned in several studies as having no side effects. It would therefore be desirable to develop computer methods to facilitate the screening of its analogs. In the present work, models of quantitative structure-activity relationships (QSAR) will be developed to predict the therapeutic potential of analogues of Anacetrapib.

Keywords: QSAR; RNA; MLR; CETP inhibitors; Anacetrapib..



Table de Matières

Table des matières

INTRODUCTION GENERALE : 1
I.1 LES MALADIES CARDIOVASCULAIRES 3
I.2 CONCEPTION DE MEDICAMENTS 3
I.2.1 Phases de conception actuelle d'un médicament 3
I.2.1.1 Première phase : recherche exploratoire 4
I.2.1.2 Seconde phase : Etudes précliniques 6
I.2.1.3 Troisième phase : Recherche clinique 6
I. 3 ACTIVITE INHIBITRICE D'UN MEDICAMENT 7
I.4 LA MODELISATION MOLECULAIRE : 7
I.4.1 Les méthodes QSAR : 9
I.4.1-1 Définition : 9
I.4.1-2 Les différents types de descripteurs : 9
I.4.2 Les différents modèles QSAR (RLM, RNA, …) 10
I.4.2-1 La régression linéaire multiple RLM : 10
I.4.2-2 Les réseaux de neurones artificiels : 10
I.4.2-3 Types de réseau de neurones : 11
I.4.2-3-a) Les réseaux de neurones non bouclés 12
I.3.2-3-b) Les réseaux de neurones bouclés 13
II.1 INTRODUCTION 15
II.2 PREPARATION DE LA BASE DE DONNEES (TRAINING, VALIDATION) 15
II.3 LES CALCULS : 18
II.3.1 Génération des descripteurs 18
II.3-2 Sélection des descripteurs 20
II.3-2-a) La sélection objective : 20
II.3-2-b) La sélection subjective : 20
II.4 DEVELOPPEMENT DES MODELES QSAR 22
II.4-1 Description du modèle MLR 22
II.4-2 Modèle neuronal RNA (MLP) 23
II.4-3 Les paramètres de validation 26
III.1 INTRODUCTION : 30
III.2 SELECTION DES DESCRIPTEURS 30
III.3 MODELE MLR 32
III.4 MODELE NEURONAL RAN (ANN): 35
III.5 COMPARAISONS ENTRE LE MODELE RNA ET RLM : 37

III.6 CONCLUSION 40
CONCLUSION GENERALE : 43
Annexe 47

Liste des figures

LISTE DES FIGURES :

Figure I.1 Répartition des molécules utilisées actuellement comme médicament en fonction de la nature de leurs cibles

Figure I.2 Neurone formel et neurone biologique (https://fr.slideshare.net/Oussama Werfelli/réseaux-de-neurones-artificiels consulté le 02/04/2020).

Figure I.3 Un Perceptron multicouche

Figure I.4 réseau de neurones bouclé ou récurrent

Figure II.1 Capture d'écran du site www.bindingdb.com

Figure II.2 Structure moléculaire de L'anacétrapib

Figure II.3 Capture d'écran du logiciel PaDEL.

Figure II.4 Capture d'écran de l'utilisation de la méthode pas à pas

Figure II.5 Les options choisis pour construire le modèle RLM

Figure II.6 La sélection des variables dépendantes et indépendantes.

Figure II.7 la sélection des sous-ensembles d'apprentissage et de test

Figure II.8 la sélection des paramètres du réseau neurone

Figure II.9 Le choix des fonctions d'activations

Figure III.1 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et celle prédite par le modèle MLR pour la série d'Ensemble d'apprentissage des analogues de L'anacétrapib

Figure III.2 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et prédite par le modèle MLR pour la série d'Ensemble de test des analogues de L'anacétrapib

Figure III.3 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et celle prédite par le modèle ANN pour la série d'Ensemble d'apprentissage des analogues de L'anacétrapib

Figure III.4 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et celle prédite par le modèle ANN pour la série d'Ensemble de test des analogues de L'anacétrapib

Figure III.5 Contribution des descripteurs dans le cas du modèle MLP

Liste des tableaux

LISTE DES TABLEAUX :

Tableau II.1 Analogues de l'anacétrapib utilisés et valeurs des concentrations inhibitrices expérimentales

Tableau III.1 Descripteurs utilisés pour le développement des modèles QSAR

Tableau III.2 Valeurs pIC_{50} expérimentales avec les descripteurs sélectionnés.

Tableau III.3 les différant paramètre MLR des deux Ensemble (apprentissage et test)

Tableau III.4 Valeurs du coefficient VIF des descripteurs sélectionnés

Tableau III.5 : Matrice des coefficients de corrélation des descripteurs sélectionnés

Tableau III.6 les différant paramètre MLP des deux Ensemble (apprentissage et test)

Tableau III.7 analogues de L'anacétrapib utilisés et valeurs des concentrations inhibitrices expérimentales et prédites par MLR, MLP (ANN)

Tableau III.8 Coefficients de détermination par les modèles RLM et ANN.



Liste des Abréviation



LISTE DES ABREVIATIONS

MCV: Les maladies cardiovasculaires.

CETP : la protéine de transfert d'ester de cholestérol

RLM: Régression linéaire multiple.

QSAR : Relations Quantitatives Structure -Activité (Quantitative structure-activity Relationship).

QSPR : Relations Quantitatives Propriété-Activité (Quantitative structureproper-tyrelationship).

IC_{50} :la concentration inhibitrice médiane

R : Coefficient de corrélation

R^2 : Coefficient de détermination.

R^2_{adj} : Coefficient de détermination ajusté.

F : Statistique de Fisher

RNA:Réseaux de neuronesartificiels

MLP:Le perceptron multicouche



Introduction General



INTRODUCTION GENERALE :

Environ 500 000 ± 700 000 nouvelles structures chimiques sont synthétisées chaque année .Avec le développement de de la chimie en phase solide, et en particulier la chimie combinatoire et la synthèse organique, ce nombre devrait augmenter fortement. Le développement de technologies rentables et efficaces pour tester et prédire les propriétés physiques, chimiques et biologiques des nouveaux composés, ce qui permettrait une estimation des dangers potentiels des composés organiques anciens et nouveaux et une évaluation efficace des risques. Les méthodes des relations quantitatives Structure - propriété (QSPR) et quantitatives structure-activité (QSAR) ont été Utilisé au cours des années pour développer des modèles à estimer / prédire l'activité thérapeutique en fonction des structures chimiques.

Les Maladies Cardiovasculaires (MCV) sont la première cause de mortalité dans le monde. Selon l'Organisation Mondiale de la Santé (OMS), elles représentaient, en 2015, 31% de toutes les causes de décès soit 17,7 millions de personnes. 37% des décès sont liés aux MCV dans l'Union Européenne et ce pourcentage est de 28% en France. Selon les estimations, en 2030 les Maladies Cardiovasculaires seront responsables de plus de 23,6 millions de décès dans le monde. Dans les Pays en voie de développement, elles seront responsables de plus de décès que la somme des maladies infectieuses, nutritionnelles, maternelles et périnatales.

Le principal objectif de ce travail est l'application de la modélisation moléculaire dans le développement des méthodes QSAR des propriétés chimiques pour prédire activité thérapeutique attendue dans des nouvelles molécules bioactives pour des analogues de l'anacétrpib

En plus d'une introduction et d'une conclusion générales, notre travail comporte trois chapitres, dans le premier chapitre nous présenterons une étude bibliographique consiste des généralités sur analogues de l'anacétrpib.

Dans le deuxième chapitre nous présenterons des généralités sur les méthodes QSAR et la régression linéaire.

Le troisième chapitre de ce travail, présente les résultats obtenus par l'étude QSAR de l'activité thérapeutique analogues de l'anacétrpib et leurs discussions.



CHAPITRE I

RAPPELS BIBLIOGRAPHIQUES



CHAPITRE I : Rappels bibliographiques

.1 LES MALADIES CARDIOVASCULAIRES

Les Maladies Cardiovasculaires (MCV) sont la première cause de mortalité dans le monde. Selon l'Organisation Mondiale de la Santé (OMS), elles représentaient, en 2015, 31% de toutes les causes de décès soit 17,7 millions de personnes **[1]**. 37% des décès sont liés aux MCV dans l'Union Européenne **[2]**et ce pourcentage est de 28% en France **[3].** Selon les estimations, en 2030 les Maladies Cardiovasculaires seront responsables de plus de 23,6 millions de décès dans le monde **[4].** Dans les Pays en voie de développement, elles seront responsables de plus de décès que la somme des maladies infectieuses, nutritionnelles, maternelles et périnatales **[5].** Le taux de mortalité par Maladies Cardiovasculaires est plus haut dans les niveaux socio-économiques les plus faibles. Plus de 80% des décès par MCV surviennent dans les Pays en voie de développement **[6]**. Les maladies cardiovasculaires (MCV) regroupent un certain nombre de troubles affectant le cœur et les vaisseaux sanguins comme :

- L'hypertension artérielle (élévation de la tension) ;
- Les cardiopathies coronariennes (crise cardiaque ou infarctus) ;
- Les maladies cérébrovasculaires (accident vasculaire cérébral) ;
- Les artériopathies périphériques ;
- L'insuffisance cardiaque ;
- Les cardiopathies rhumatismales ;
- Les cardiopathies congénitales ;
- Les cardiomyopathies **[7]**.

I.2 CONCEPTION DE MEDICAMENTS

Le médicament, depuis sa conception jusqu'à commercialisation, en passant par sa fabrication, il est soumis à une législation spécifique et rigoureuse ayant pour but de lui garantir la qualité, l'efficacité et la sécurité requises.

I.2.1 Phases de conception actuelle d'un médicament

Les symptômes qui accompagnent une maladie sont dus au dysfonctionnement de processus biologiques clés. Le dérèglement initial d'un seul de ces processus au niveau de la cellule par le jeu des interactions, suffit le plus souvent, à entraîner le dérèglement progressif et souvent irréversible de tout un édifice de réactions biologiques. Il s'ensuit une perte progressive, parfois définitive, d'une fonction

physiologique vitale (cardiaque, cérébrale, rénale, respiratoire, …) dont l'issue peut être fatale. C'est seulement en restaurant le fonctionnement normal des processus responsable de cet "effondrement "biologique qu'on peut espérer interrompre la progression d'une maladie. Il en résulte que seule la connaissance de l'origine d'une maladie et de ses facteurs (étiopathogénie) à un niveau moléculaire permettra de concevoir de façon rationnelle des médicaments capables d'en arrêter la progression (*drug design*) **[8].**

Le développement d'un nouveau médicament est un processus long (entre 12 et 14ans), complexe avec un coût total de l'ordre de 800 millions à 1,8 milliard de dollars **[9].**

Ce processus est scindé en trois phases : une première phase dédiée à la recherche exploratoire, une seconde phase englobe les études précliniques et une troisième phase consacrée à la recherche clinique.

I.2.1.1 Première phase : recherche exploratoire

Lors de cette phase, il s'agit de comprendre les mécanismes de la maladie afin de déterminer la cible potentielle que le médicament devra atteindre. Une fois cette cible identifiée, c'est l'étape dite du criblage ou de très nombreuses molécules (plusieurs milliers) sont testées afin de ne retenir que celles éventuellement efficaces (généralement une centaine).

- **Identification de la cible**

Ces cibles sont nombreuses. On y dénombre plusieurs familles :

- Les enzymes visées par environ 25 % des molécules actives actuelles. Ces molécules sont essentiellement des inhibiteurs de l'activité enzymatiques.
- Les pompes, transporteurs et canaux ioniques, protéines membranaires qui régissent les équilibres transmembranaires des principaux ions, visés par environ 15 % des molécules actives, des inhibiteurs des mouvements ioniques.
- Les récepteurs membranaires visés par environ40 % des molécules médicamenteuses, agonistes ou antagonistes. Leur diversité est illustrée sur la **(Figure I.1)**
- Les récepteurs nucléaires visés par environ 10 % des molécules actives, agonistes ou antagonistes.

Ces quatre familles de cibles correspondent à près de 90 % des médicaments. Il faut y ajouter 5 % des molécules se liant à des cibles diverses, par exemple les protéines des

microtubules qui lient la colchicine, et les alcaloïdes de la pervenche. Enfin, les cibles d'environ 5 % des molécules actives reconnues ne sont pas déterminées, par exemple la cible du paracétamol **[10].**

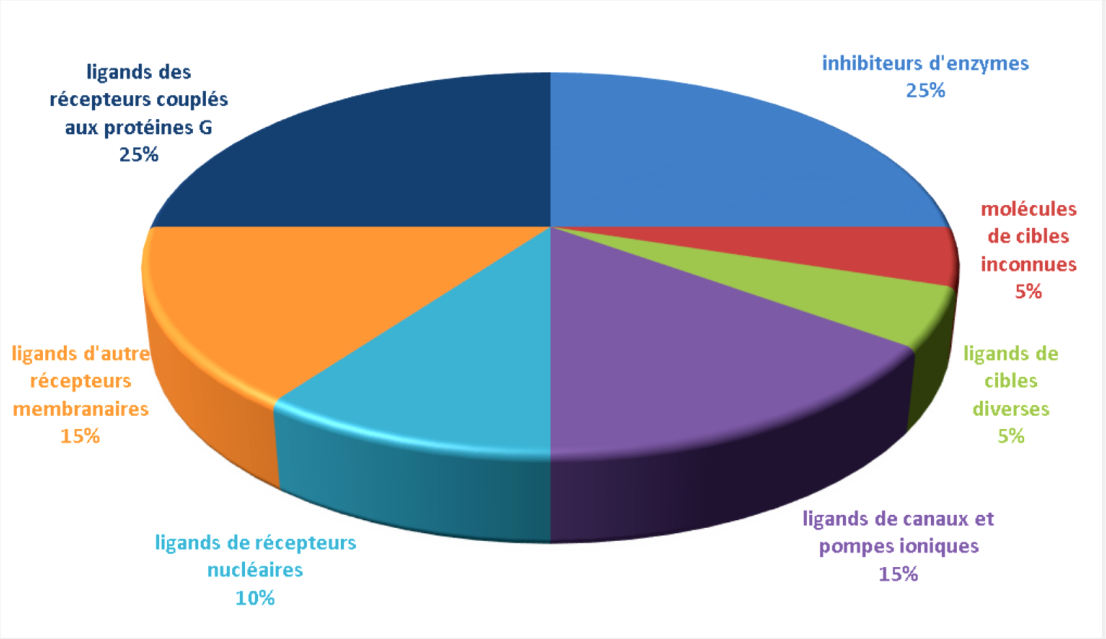


Figure I.1 Répartition des molécules utilisées actuellement comme médicament en fonction de la nature de leurs cibles **[10]**.

- **Criblage**

Une fois que la cible est identifiée, on recherche des composés capables de l'inhiber de façon spécifique, importante et durable. Ces composés sont appelés "*hits*".

La recherche de hits se fait par criblage *in vitro* (ou *screening*) de larges banques de composé appelées chimiothèques et par la réalisation de tests biologiques. Les coûts aussi bien financiers que temporels de ce criblage, où des milliers de molécules sont utilisées, s'avèrent énormes **[11].** Aussi, dans le but de minimiser les coûts de cette approche classique *in vitro*, le développement d'approches *in silico* pour le criblage virtuel des composés chimiques est une alternative prometteuse

Ces approches *in silico* font appel à des outils informatiques. Ce terme "*in silico*" fait référence au silicium, matériau principal retrouvé dans les puces informatiques de tous les ordinateurs. Ces approches permettent de simuler ou de modéliser un phénomène biologique à l'aide de l'outil informatique. Une des plus prometteuses approches *in silico* est celle désignée sous l'acronyme anglais QSAR (Quantitative

Structure Activity Relationship). Elle se base sur des relations quantitatives entre structures et propriétés des composés.

Le criblage virtuel de l'approche QSAR a lieu dès lors que des molécules actives sur la cible sont trouvées. La modélisation informatique de la structure chimique de ces molécules et de leur interaction avec la cible permet de concevoir de nouvelles molécules qui sont synthétisées et testées pour évaluer leur activité sur la cible. A l'issue de ce Cycle "modélisation-synthèse-tests", 10 à 100 molécules sont choisies parmi celles ayant une activité optimale. Ces molécules sont désignées sous le terme de "candidat médicament".

I.2.1.2 Seconde phase : Etudes précliniques

Au cours de cette seconde phase, les candidats médicament précédemment sélectionnés subissent une série de tests précliniques (sur cellules et sur animaux) obligatoires avant les tests sur l'homme. Ces tests ont pour but d'évaluer minutieusement les risques d'effets secondaires de ces candidats médicament.

Ces tests englobent des études de toxicité aigüe et chronique, des examens de la fonction reproductrice, des études de la toxicité embryofoetale et périnatale, une étude Pharmacocinétique et métabolisme du médicament, étude du pouvoir mutagène et cancérogène et une étude de la tolérance totale.

I.2.1.3 Troisième phase : Recherche clinique

Si les tests précliniques sont satisfaisants, on procède alors aux tests cliniques, c'est-à-dire qu'on passe aux tests sur l'être humain. Beaucoup de candidats médicaments sont éliminés avant ce stade atteint par seulement un médicament sur quinze. Cette recherche se déroule chez l'humain en trois phases principales (phase I, II, III) dans des conditions bien réglementées et dans un cadre juridique. Les essais de la phase I ont pour but de rechercher la dose toxique au-delà de laquelle le candidat médicament ne pourrait pas être utilisé, tandis que pour la phase II, il est question de rechercher les doses ayant une activité pharmacologique favorable. La phase III, quant à elle, est dédiée à la mesure Précise de l'efficacité du candidat médicament. Après ces trois phases, un dossier d'Autorisation de Mise sur le Marché (A.M.M) est déposé pour un seul candidat Médicament parmi ceux ayant été soumis aux tests cliniques **[12].**

I. 3 ACTIVITE INHIBITRICE D'UN MEDICAMENT

Dans la phase de découverte de nouveaux médicaments, l'activité des composés chimiques sur une cible biologique est mesurée afin de mettre en évidence une action inhibitrice. En thérapeutique, l'activité des molécules est appréciée globalement : c'est la résultante de toutes les interactions entre la molécule active et les différents récepteurs auxquels elle se lie [**16**].

L'inhibition de l'activité enzymatique est une stratégie majeure utilisée dans la conception de médicaments [**17**]. Aussi, la concentration inhibitrice médiane (IC_{50}) est la mesure in vitro la plus pertinente utilisée pour évaluer le potentiel thérapeutique (ou efficacité) d'un candidat médicament. IC_{50} est définie comme étant la concentration d'un médicament nécessaire pour 50% d'inhibition d'un processus biologique ou d'un composant d'un processus comme une enzyme, une cellule, un récepteur cellulaire ou un microorganisme. Pour les candidats médicaments qui sont des inhibiteurs.

I.4 LA MODELISATION MOLECULAIRE :

La recherche et la synthèse de nouveaux composés chimiques sont aujourd'hui souvent associées à une étude par modélisation moléculaire. La modélisation moléculaire est une technique permettant, non seulement de représenter les propriétés et les réactions chimiques mais aussi de manipuler les modèles des structures en deux ou trois dimensions.

La modélisation moléculaire implique l'utilisation des méthodes de calcul théoriques (mécanique moléculaire, dynamique moléculaire, mécanique quantique *ab-initio* ou *semi-empirique*, ...) permettant de déterminer la représentation graphique de la géométrie ou de la configuration des atomes d'une molécule et d'évaluer les propriétés physico-chimiques de la molécule étudiée. La modélisation moléculaire associée à une représentation infographique des stéréochimies permet d'interpréter des phénomènes physico-chimiques, de suggérer des nouvelles expériences et d'analyser ainsi des résultats d'une façon plus critique que les expériences classiquement utilisées, mais ces deux approches purement théoriques ou expérimentales sont complémentaires.

CHAPITRE I : Rappels bibliographiques

Le comportement des molécules peut être simulé de façon statique ou dynamique. Cette simulation inclut notamment le comportement et la flexibilité des molécules en solution, les mécanismes réactionnels et interactions mises en jeu dans la catalyse enzymatique, la reconnaissance ou la complexation, la prédiction de propriétés et de fonctionnalités à partir d'analogies de structure. La modélisation moléculaire à pour but de prévoir la structure et la réactivité des molécules ou des systèmes de molécules. Les méthodes de la modélisation moléculaire peuvent être rangées en trois catégories**[13]**:

- Les méthodes quantiques.
- La mécanique moléculaire.
- La dynamique moléculaire.

Sous le vocable de modélisation moléculaire, on trouve différentes techniques de visualisation, de manipulation, d 'analyse et de calcul des structures moléculaires. Schématiquement, on distingue les techniques de graphisme moléculaire permettant de représenter sur un écran la structure 2D ou 3D d'une molécule, de la manipuler (rotation, translation, changement de conformation, superposition, etc.) de façon interactive (tout au moins sur les stations graphiques suffisamment puissantes) et de l'analyser (calculs de paramètres géométriques tels que distances, angles, surfaces accessibles, etc.).

Ces structures peuvent être obtenues par construction à partir d'éléments (atomes, groupements, résidus nucléotidiques ou peptidiques) choisis dans une bibliothèque de programme. Dans d'autres cas, elles sont extraites d'une banque de données dans laquelle sont stockées les structures établies par radiocristallographie (Cambridge Structural Data, Protein Data Bank). Dans d'autres cas, enfin elles sont construites à partir de données expérimentales c'est-à-dire de résultats de radiocristallographie ou de RMN.

Pour passer du modèle initial au modèle final, il faut franchir plusieurs étapes d'optimisation faisant appel à différentes techniques de calcul de mécanique moléculaire et/ou de dynamique moléculaire.

Ces techniques constituent donc un second volet de la modélisation moléculaire. Elles reposent sur l'emploi de différents logiciels permettant de faire des calculs d'énergie, des optimisations, des simulations de mouvements moléculaires, à l'aide d'ordinateurs (les calculs de dynamique sont souvent effectués sur un CRAY). Les logiciels de modélisation proprement dits tels que SYBYL, MACROMODEL, etc. sont donc interfacés avec des programmes de calculs tels que MM2, AMBER, CHARMM, etc. Ces calculs peuvent être effectués sur un ordinateur local telle une station Silicone Graphique adjointe à un centre de calcul universitaire **[14].**

La stabilité de la structure tridimensionnelle d'une molécule est déterminée par les interactions intramoléculaires et les interactions avec le milieu extérieur. La recherche des conformations stables d'une molécule consiste à déterminer les minima de l'énergie globale d'interaction. Cette énergie peut être calculée par des méthodes quantiques généralement longues et onéreuses. Pour faciliter les calculs, on considère généralement que le terme variable de cette énergie dépend de la construction de la molécule et de l'arrangement de ses atomes : c'est le principe des méthodes empiriques (mécanique moléculaire, dynamique moléculaire).

Dans la plupart de ces méthodes, il n'est pas tenu compte des interactions avec le solvant, mais uniquement des interactions entre les atomes constitutifs de la molécule. La recherche d'une conformation consiste alors à faire une minimisation de l'énergie intramoléculaire. Cette énergie potentielle est fractionnée en un certain nombre de termes additifs indépendants.

I.4.1 Les méthodes QSAR :

I.4.1-1 Définition :

Les modèles quantitatifs de QSAR présentent une solution statistique du problème de la difficulté du calcul direct des propriétés physiques et biologiques à partir de la structure. L'intérêt d'un modèle de QSAR est de tirer des informations à partir de l'ensemble des descripteurs numériques caractérisant la structure moléculaire et prédire ainsi les activités biologiques de nouvelles structures **[15].**

I.4.1-2 Les différents types de descripteurs :

Il existe à ce jour plus de **3000** descripteurs répertoriés, qui sont classés en 4 types :

a) Les descripteurs **0-D** : Ils sont dérivés de la liste d'atomes de la molécule, comme le poids moléculaire.

b) Les descripteurs **1-D** : Ils sont calculés à partir de listes de sous structures de la molécule, comme les groupements fonctionnels, les cycles présents, etc[16].

c) Les descripteurs **2-D** : Ils sont obtenus à partir des représentations des molécules sous forme de graphes bidimensionnelles [17].

d) Les descripteurs **3-D** : Ils sont calculés à partir de la représentation tridimensionnelle de la molécule, on cite parmi eux, les descripteurs électroniques (moment dipolaire, l'électronégativité, …), le volume moléculaire, la surface moléculaire superficielle…etc [17-20].

e) Les descripteurs **4-D** : Obtenus par le calcul des champs d'interactions moléculaire entre deux molécules.

I.4.2 Les différents modèles QSAR (RLM, RNA, …)

I.4.2-1 La régression linéaire multiple RLM :

La régression linéaire multiple est la plus simple méthode statistique de modélisation et la plus appliquée dans les études de la relation structure-activité **[21].**

La méthode RLM repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante Y (dans notre cas l'activité biologique) et une série de n variables indépendantes Xi (ici, les descripteurs moléculaires) **[22].** L'objectif est d'arriver à une équation mathématique de la forme :

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \cdots + a_nX_n$$

Où a_i (i = 0, 1, …, n) sont les coefficients de la régression.

I.4.2-2 Les réseaux de neurones artificiels :

L'idée des réseaux neuronaux artificiels (RNA) a été inspirée de la manière par laquelle les neurones biologiques**(Figure I.2)** traitent les informations **[23].** Ce concept est utilisé pour mettre en œuvre des logiciels de simulations des processus massivement parallèles qui impliquent des éléments de traitement connectés dans une architecture de réseau.

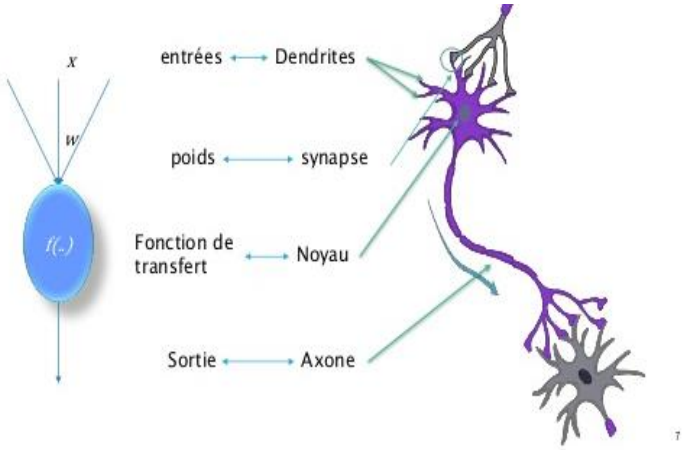


Figure I.2 Neurone formel et neurone biologique (https://fr.slideshare.net/Oussama Werfelli/réseaux-de-neurones-artificiels consulté le 02/04/2020).

- Une première étape de prétraitement des entrées consistant en la combinaison linéaire des entrées selon l'expression suivante :

$$v = w_0 + \sum_{i=1}^{n} (w_i x_i)$$

Où x_i sont les neurones d'entrée (descripteurs).$w_0$est appelé biais. Le biais peut être considéré comme la pondération de l'entrée 0 fixée à 1.$v$est appelé potentiel du neurone

- Dans la seconde étape, une fonction f, appelée fonction d'activation, est appliquée à ce potentiel v. Le choix d'une fonction d'activation serévèle dans certains cas être un élément constitutif important des réseaux de neurones. Il en existe plusieurs types. Ainsi, le neurone formel peut être défini mathématiquement comme étant une fonction algébrique, a priori non linéaire (suivant la fonction d'activation f) et bornée, avec des entrées $x_i$et paramétrée parles poidsw_i. Un neurone formel, réalise donc l'opération suivante :

$$y = f(v) = f\,[w_0 + \sum_{i=1}^{n} (w_i x_i)]$$

I.4.2-3 Types de réseau de neurones :

On distingue deux grands types d'architectures de réseaux de neurones : les réseaux de neurones non bouclés et les réseaux de neurones bouclés.

I.4.2-3-a) Les réseaux de neurones non bouclés

Un réseau de neurones non bouclé est représenté graphiquement par un ensemble de neurones "connectés" entre eux, l'information circulant des entrées vers les sorties sans "retour en arrière"; si l'on représente le réseau comme un graphe dont les nœuds sont les neurones et les arêtes les "connexions" entre ceux-ci, le graphe d'un réseau non bouclé est *acyclique.* Le terme de "connexions" est une métaphore : dans la très grande majorité des applications, les réseaux de neurones sont des formules algébriques dont les valeurs numériques sont calculées par des programmes d'ordinateurs, non des objets physiques (circuits électroniques spécialisés) ; néanmoins, le terme de connexion, issu des origines biologiques des réseaux de neurones, est passé dans l'usage, car il est commode quoique trompeur. Il a même donné naissance au terme de *connexionnisme.***[24]**

La **Figure I.3** représente un réseau de neurones non bouclé qui a une structure particulière ; très fréquemment utilisée : il comprend des entrées, une couche de neurones "cachés" et des neurones de sortie. Les neurones de la couche cachée ne sont pas connectés entre eux. Cette structure est appelée *Perceptron multicouche.*

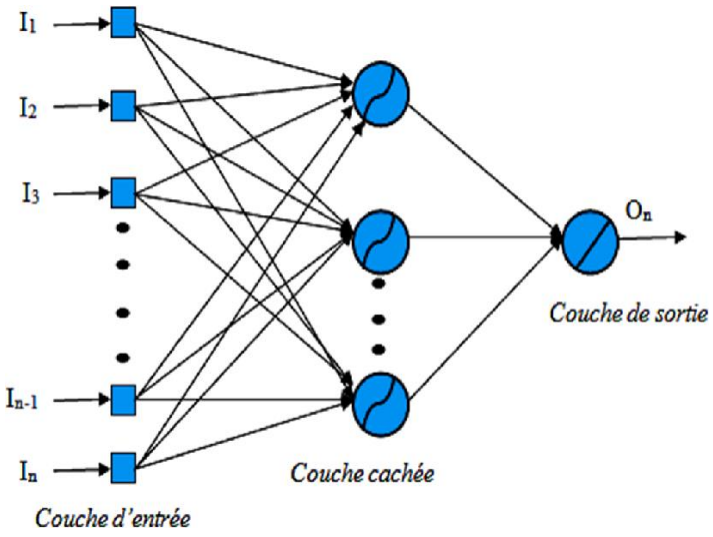


Figure I.3 Un Perceptron multicouche

I.3.2-3-b) Les réseaux de neurones bouclés

Contrairement aux réseaux de neurones non bouclés dont le graphe de connexions est acyclique, les réseaux de neurones bouclés peuvent avoir une topologie de connexions quelconque, comprenant notamment des boucles qui ramènent aux entrées la valeur d'une ou plusieurs sorties. Pour qu'un tel système soit causal, il faut évidemment qu'à toute boucle soit associé un retard : un réseau de neurones bouclé est donc un système dynamique, régi par des équations différentielles ; comme l'immense majorité des applications sont réalisées par des programmes d'ordinateurs, on se place dans le cadre des systèmes à temps discret, où les équations différentielles sont remplacées par des équations aux différences **[25].**

Les réseaux de neurones bouclés (**Figure I.4)** sont utilisés pour effectuer des tâches de modélisation de systèmes dynamiques, de commande de processus, ou de filtrage.

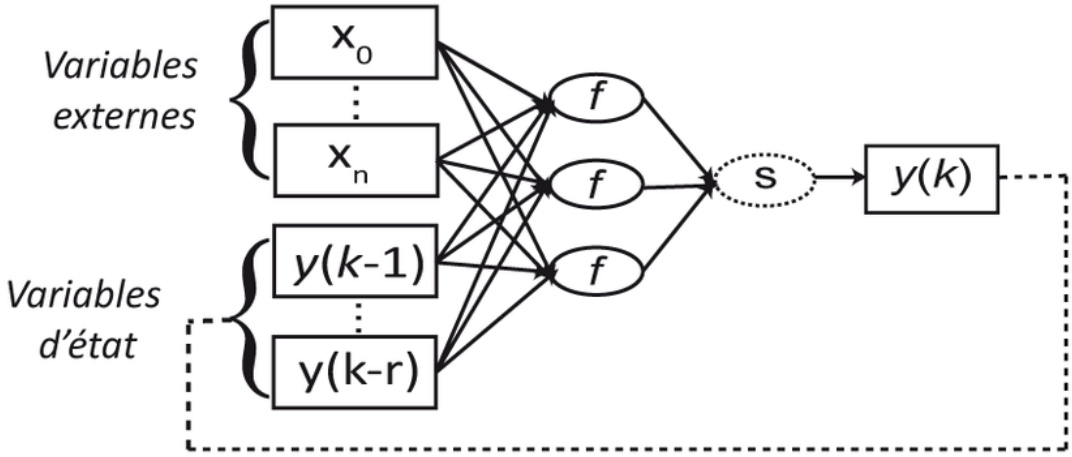


Figure I.4réseau de neurones bouclé ou récurrent [26]



CHAPITRE II
DEVELOPPEMENT DE MODELES QSAR

CHAPITRE II : DEVELOPPEMENT DE MODELES QSAR

I.1 INTRODUCTION

L'objectif majeur de l'analyse QSAR est de trouver une relation mathématique entre l'activité et les descripteurs liés à la structure de la molécule **[27, 28]**. Ces études peuvent éclairer et même élucider le mécanisme par lequel l'activité en question est liée à la structure chimique.

L'objectif de ce travail est d'établir une relation ou une équation mathématique reliant des descripteurs structurels d'une série de composés étudiés aux activités biologiques.

Nous avons étudié 61 analogues de l'anacétrapib (**Figure II.1)** par 2 méthodes :

- La régression linéaire multiple
- Les réseaux de neurones

II.2 PREPARATION DE LA BASE DE DONNEES (TRAINING, VALIDATION)

Le choix de la base de données expérimentale initiale est une étape critique pour le développement des modèles QSAR. Elle doit être composée de données expérimentales fiables.

Pour cette étude nous avons utilisé La concentration inhibitrice (CI_{50}) desanalogues de l'anacétrapib récupérés à partir de la base de données BindingDB[29]

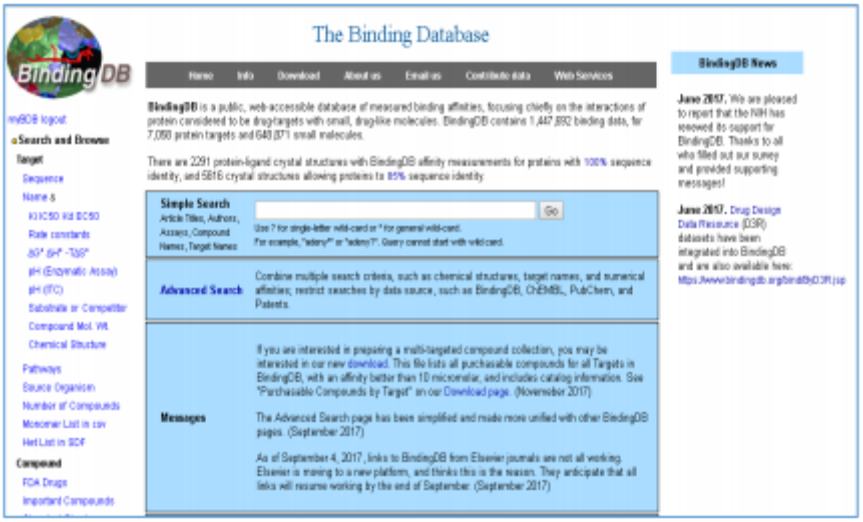


Figure II.1 Capture d'écran du site www.bindingdb.com

CHAPITRE II : DEVELOPPEMENT DE MODELES QSAR

Les 61 analogues de L'anacétrapib (**Figure II.2**) ayant une activité inhibitrice sur la protéine de transfert d'ester de cholestérol (CETP). Les valeurs CI_{50} sont converties en échelle logarithmique négative [pCI_{50} = -log CI_{50} (µmol/l)] (**Tableau II.1**).

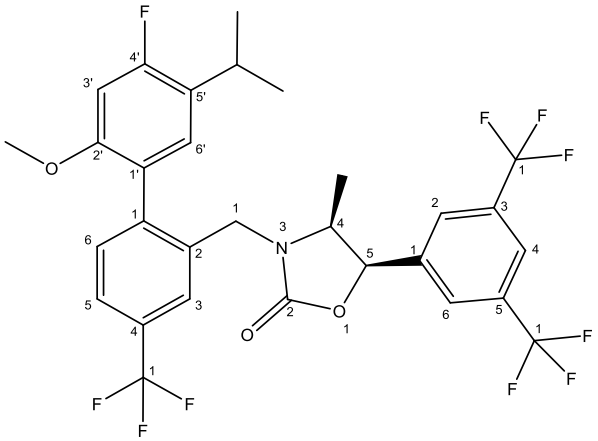


Figure II.2 Structure moléculaire de L'anacétrapib

Tableau II.1 Analogues de l'anacétrapib utilisés et valeurs des concentrations inhibitrices expérimentales

N°	PubChem CID	Structure	pIC_{50}
1	89522387	C37H27F10NO5	-2,85582
2	71526289	C37H27F10NO5	-2,07918
3	71526635	C37H27F10NO5	-2,80686
4	71526291	C37H28F9NO5	-1,93952
5	71526636	C36H25F10NO5	-2,23553
6	71526290	C36H27F9N2O5	-2,02531
7	71526382	C36H27F9N2O5	-2,09342
8	71526551	C36H27F9N2O5	-1,6721
9	89522406	C35H23F9N2O5	-2,29885
10	71526200	C35H30F9NO5	-2,38382
11	89522504	C34H29F9N2O5	-2,39094
12	71526465	C35H25F9N2O4	-3,11126
13	71526555	C34H29F9N2O4	-2,29667
14	71525739	C35H26F7NO5	-2,35965
15	71526637	C33H26F9NO5	-1,98471
16	71526288	C33H28F9NO5	-2,62118
17	71525834	C34H25F7N2O5	-3,24105
18	89522501	C32H25F9N2O5	-1,94939
19	89522372	C33H27F9N2O4	-2,66839

20	89522213	C32H25F10NO5	-3,21192
21	57390584	C30H25F10NO5S	-1,43297
22	57397759	C32H32F7NO5S	-3,41481
23	89522469	C32H27F10NO3	-3,08207
24	57394194	C34H32F7NO3	-2,40518
25	71525924	C31H25F10NO3	-1,92901
26	71526105	C31H23F10NO3	-2,08636
27	71526722	C28H18F12N2O3	-2,56984
28	57392491	C30H25F10NO3S	-1,29226
29	57394193	C33H34F7NO3	-1,3962
30	57390685	C32H31F8NO3	-1,25527
31	11721583	C30H28F7NO5S	-1,52504
32	57402947	C32H32F7NO4	-1,85612
33	11628674	C32H30F7NO3	-1,39445
34	71526721	C29H19F9N2O3	-2,35064
35	11657340	C32H32F7NO3	-1,7332
36	11527347	C29H25F7N2O5	-1,12385
37	11650234	C30H28F7NO3S	-1,20952
38	11706999	C30H25F7N2O3	-1,36736
39	89522451	C29H22F9NO3	-2,41963
40	71525462	C28H19F10NO3	-2,21932
41	71525381	C28H19ClF9NO3	-2,36791
42	71525833	C30H25F8NO3	-2,67624
43	71526723	C27H18ClF9N2O3	-2,46909
44	71525552	C27H16F11NO2	-2,01745
45	71525550	C27H16ClF10NO2	-2,273
46	71525643	C28H19ClF9NO2	-3,08743
47	89522333	C27H19F9N2O3	-2,5001
48	71525463	C28H17F9N2O2	-2,24822
49	11649992	C30H28F7NO3	-1,22272
50	11664298	C29H25F8NO3	-1,39794
51	11556473	C29H25BrF7NO3	-1,29667
52	57395961	C29H25F7INO3	-1,37107
53	11578015	C29H25ClF7NO3	-1,21748
54	57402946	C29H26F7NO3	-1,49429
55	89522516	C27H17ClF9NO2	-2,65591
56	71526724	C26H15F9N2O2S	-2,42488
57	89522295	C26H18F9NO3	-2,5358
58	89522453	C27H22F9NO2	-1,86705
59	89522402	C26H18F9NO2S	-2,08707
60	71525549	C26H20F9NO2	-1,75305
61	71525461	C25H16F9NO2S	-2,01199

II.3 LES CALCULS :

II.3.1 Génération des descripteurs

Une étape importante dans l'obtention d'un modèle QSAR est la représentation numérique des caractéristiques structurelles des molécules, appelées descripteurs moléculaires.

Les descripteurs sont des valeurs numériques qui encodent les caractéristiques physicochimiques des molécules à partir de leurs représentations structurales.

La génération des descripteurs constitue l'étape cruciale lors de l'analyse de QSAR et la qualité d'un modèle dépendra essentiellement de la nature des descripteurs générés.

Plusieurs logiciels sont disponibles pour faire les calculs des descripteurs, parmi ceux, que nous avons utilisé dans nos travaux, on trouve : ChemOffice**[30],**PaDEL**[31],** IBM SPSS 21**[32].**

Dans ce travail, nous avons utilisé le logiciel PaDEL pour générer les descripteurs, il est capable de calculer plus de 1400 descripteurs.

Afin de générer les descripteurs on a procédé comme suit:

- Sélectionnez un seul fichier structurel ou un répertoire contenant les fichiers structurels des molécules. Les formats de fichiers les plus courants (par exemple MDL mol, SMILES) sont pris en charge, mais le format de fichier recommandé est MDL mol.
- Sélectionnez un fichier pour enregistrer les descripteurs calculés. Les descripteurs seront enregistrés au format de fichier CSV (valeurs séparées par des virgules).
- Sélectionnez l'option "1D & 2D" car nous allons calculer les descripteurs 1D et 2D
- Sélectionnez l'option "Enlèvement de sel" pour supprimer les sels comme Na, Cl de la molécule avant le calcul des descripteurs.
- Cochez l'option "Détecter l'aromaticité" pour supprimer les informations d'aromaticité existantes et détecter automatiquement l'aromaticité dans la molécule avant le calcul des descripteurs.

- Cochez l'option "Normaliser les groupes nitro pour normaliser les groupes nitro à N (: O): O. Ceci est nécessaire pour que les descripteurs étendus des atomes topochimiques (ETA) soient calculés correctement.
- Cochez l'option "Journal" pour générer un fichier journal.
- Cochez l'option "Conserver l'ordre des molécules" pour conserver l'ordre des molécules dans les fichiers structurels pour le fichier descripteur. **[33]**

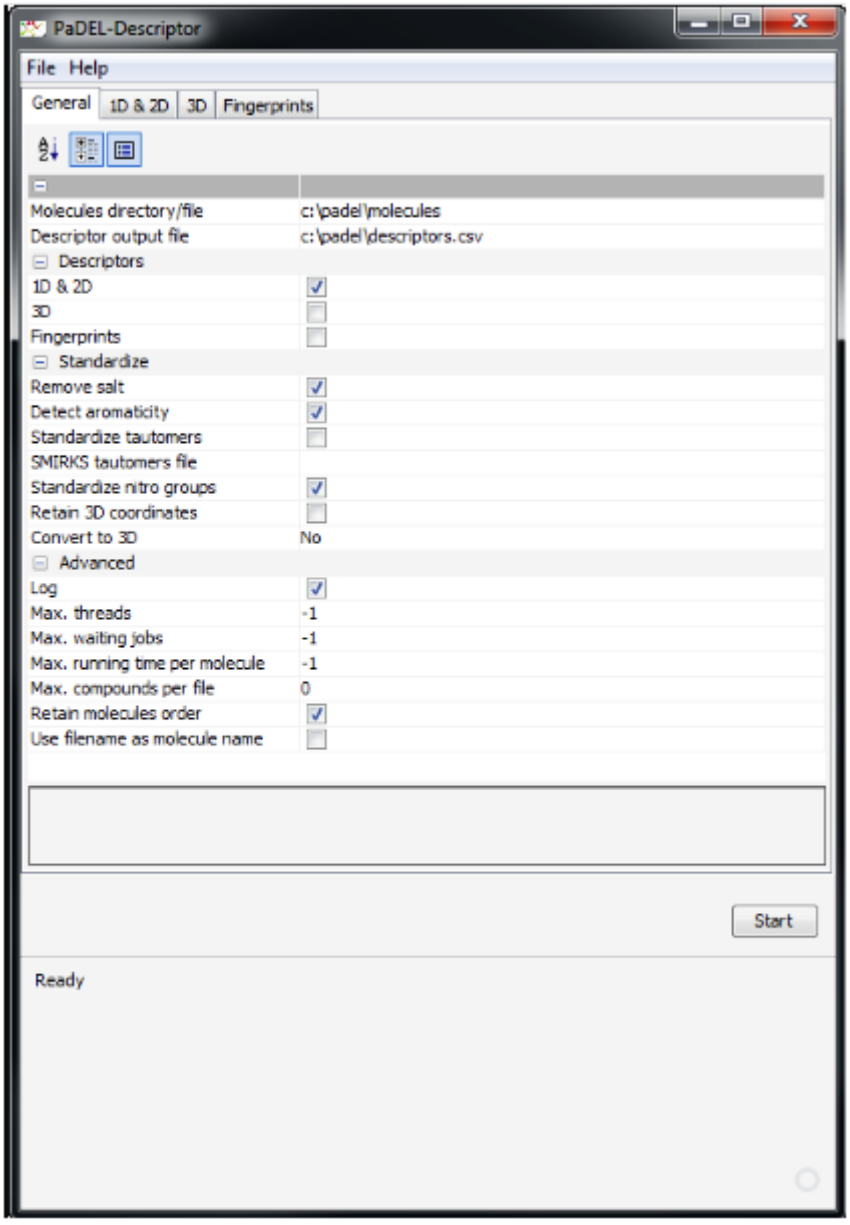


Figure II.3 Capture d'écran du logiciel PaDEL.

II.3-2 Sélection des descripteurs

Un grand nombre de descripteurs différents sont collectés après le calcule, Plus de 1400 descripteurs moléculaires sont disponibles et sont principalement calculés par le logiciel PaDEL.

Afin d'éviter la formation de modèles QSAR 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.

Cela signifie que nous allons réduire le nombre de descripteurs au plus petit nombre possible

La procédure de sélection des variables peut être divisée en deux étapes **[34].**

* Sélection objective
* Sélection subjective

II.3-2-a) 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. La sélection objective se fait manuellement.

II.3-2-b) La sélection subjective :

La sélection subjective des variables de l'ensemble des descripteurs déjà réduit est alors employée dans la recherche de sous-ensembles de descripteurs optimaux en utilisant l'activité visée. A cet effet, diverses méthodes d'optimisations sont utilisées pour parcourir, d'une manière efficace, la dimension du modèle (le nombre optimal de descripteurs par modèle) sans pour autant examiner toutes les combinaisons possibles.

Pour cette étape en à utiliser le logiciel IBM SPSS Statistics 21.

Nous avons utilisé la régression linéaire multiple pour réduire le nombre de descripteurs.

- **La sélection Ascendante pas à pas (ForwardStepwise)**

Les méthodes de type pas à pas consistent à considérer d'abord un modèle faisant intervenir un certain nombre de variables explicatives. Puis nous procédons par élimination ou ajout successif de variables.

La Sélection par la méthode pas à pas est une étape importante dans l'obtention d'un modèle QSAR, cette méthode vient sous l'utilisation de la régression linéaire multiple à l'intérieure du logiciel IBM SPSS Statistics 21(**Figure II.4).**

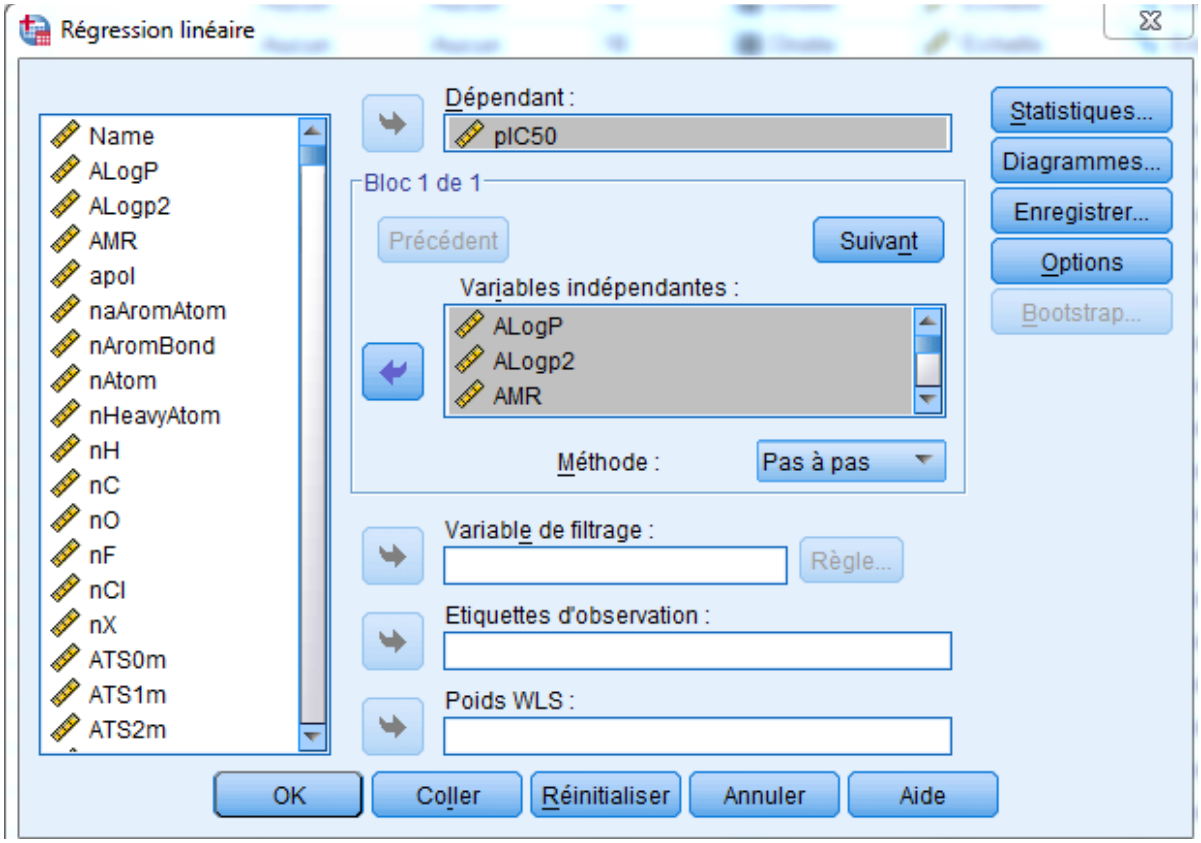


Figure II.4 Capture d'écran de l'utilisation de la méthode pas à pas

- **Elimination par bloc :**

Procédure de sélection de variables dans laquelle toutes les variables d'un bloc sont supprimées en une seule étape

- **Elimination descendante :**

C'est une procédure de sélection de variables au cours de laquelle toutes les variables sont introduites dans l'équation, puis éliminées une à une. La variable ayant la plus petite corrélation partielle avec la variable dépendante est la variable dont la suppression est étudiée en premier. Si elle répond aux critères d'élimination, elle est supprimée. Une fois la première variable éliminée, l'élimination de la variable suivante restant dans l'équation et ayant le plus petit coefficient de corrélation

partielle est étudié. La procédure prend fin quand plus aucune variable de l'équation ne satisfait aux critères de suppression.

- **Sélection ascendante :**

Procédure de sélection étape par étape de variables, dans laquelle les variables sont introduites séquentiellement dans le modèle. La première variable considérée est celle qui a la plus forte corrélation positive ou négative avec la variable dépendante. Cette variable n'est introduite dans l'équation que si elle satisfait le critère d'introduction. Si la première variable est introduite dans l'équation, la variable indépendante externe à l'équation et qui présente la plus forte corrélation partielle est considérée ensuite. La procédure s'interrompt lorsqu'il ne reste plus de variables satisfaisant au critère d'introduction.

II.4 DEVELOPPEMENT DES MODELES QSAR

La base de données contenant 61 composés a été divisée en deux ensembles: un ensemble d'apprentissage englobant 80% et un ensemble de avec 20% du nombre total de composés. L'ensemble d'apprentissage avec 49 dérivés est dédié au développement du modèle QSAR. Par contre, l'ensemble de test, qui comprend 12 dérivés de L'anacétrapib qui n'avaient pas été utilisés pour le développement des modèles QSAR, a été laissé pour la validation externe.

Pour le développement des modèles QSAR, deux méthodes d'analyse de données ont été utilisées: la régression linéaire multiple (RLM ou MLR en anglais), les réseaux de neurones artificiels (RNA ou ANN en anglais). L'approche MLR a été réalisée à l'aide du logiciel IBM SPSS Statistics 21. Les approches ANN et, quant à elle, été réalisée à l'aide du logiciel STATISTICA.

II.4-1 Description du modèle MLR

Pour l'obtention du model MLR en a utilisé le du logiciel IBM SPSS Statistics 21.

Et en suivant les étapes ci-dessous :

1. d'abord, nous ouvrons le fichier Excel qui contient la base de données dans le programme IBM SPSS (nombre de descripteurs réduite + les analogues de L'anacétrapib)

2. Pour réaliser l'analyse, cliquez sur **Analyse**, **Régression**, puis **Linéaire**.

3. En cliquant sur « «ajouter», pour insérer la variable dépendante et les variable(s) indépendante(s) dans les boîtes appropriées.

4. En a choisi la méthode **Ascendante pas à pas**.

Pour la régression multiple, nous avons examiné les options illustrées sur la (**Figure II.5**) ci-dessous:

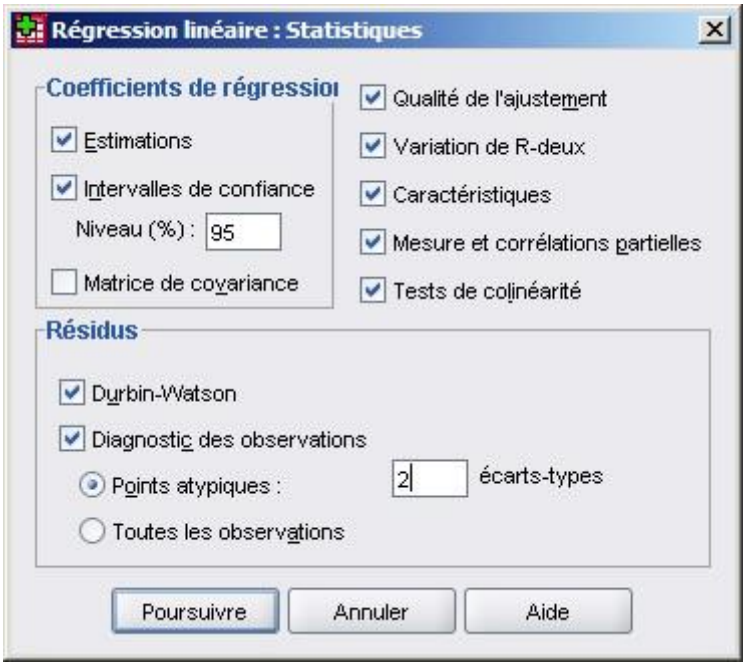


Figure II.5 Les options choisis pour construire le modèle RLM

II.4-2 Modèle neuronal RNA (MLP)

L'utilisation des réseaux de neurones artificiels (RNA) est nécessaire et est devenue une technique de modélisation importante. Elle est largement utilisée pour les études de QSAR non linéaires et est un outil puissant pour construire des modèles prédictifs. Dans le cadre de ce travail, les réseaux de neurones artificiels (RAN) ont été utilisés pour construire un modèle non linéaire sur la base des mêmes descripteurs utilisés dans le cas du modèle MLR. Un réseau typique à trois couches de type perceptron multicouche (MLP) avec une couche d'entrée, une couche cachée et une couche de sortie a été adoptée dans cette étude. Des travaux théoriques ont montré qu'une seule couche cachée suffit pour que l'RNA se rapproche d'une fonction non linéaire complexe et que de nombreux résultats expérimentaux semblent confirmer qu'une couche cachée peut suffire à la plupart des problèmes de prédiction **[35].**

Il n'y a pas de principes théoriques rigoureux pour choisir la topologie du réseau appropriée. Ainsi, l'utilisation d'une régression neuronale nécessite l'optimisation de

l'architecture du réseau neuronal. Dans cette étude, la fonction sigmoïde a été utilisée comme fonction de transfert de la couche cachée et la fonction identité comme fonction de transfert de la couche de sortie. Le réseau a été formé à l'aide de l'algorithme des méthodes quasi-Newton BFGS. Un neurone de sortie a été utilisé pour représenter la pCI_{50} observée. La couche d'entrée est formée des 5 descripteurs (ou variables) sélectionnés auparavant. L'ensemble de données (61 composés) des dérivés de l'anacétrapib a été divisé en deux lots: un ensemble d'apprentissage et un ensemble de test (ou de validation) composés respectivement de 49 et 12 composés. Pour l'obtention du model RNA en a utilisé le du logiciel STATISTICA (STATISTICA 10.0, StatSoft, Inc., Tulsa, OK, USA).

On a procéder comme suit :

1. On ouvre le fichier Excel qui contient la base de données (nombre de descripteurs réduite + les analogues de l'anacétrapib) dans le programme STATISTICA

2. Dans le menu « **Data mining** » en sélectionne : « **Neural Network** », « **Régression** », puis « **variable** ».

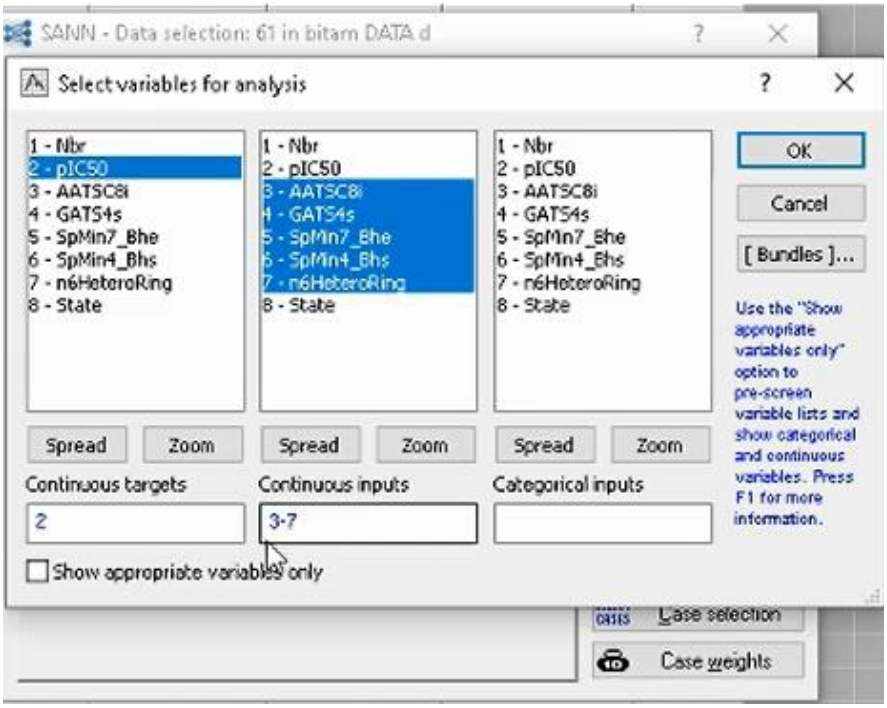


Figure II.6 La sélection des variables dépendantes et indépendantes.

Dans le volet « **continuous Target »** en a choisi **pIC$_{50}$ (Figure II.6),** et dans le volet « **continuous inputs «** en a choisi les 5 descripteurs réduite

3.Dans l'option « **sampling(CNN and ANS) »** on sélectionne les deux échantillons d'apprentissage et de test **(Figure II.7)** :

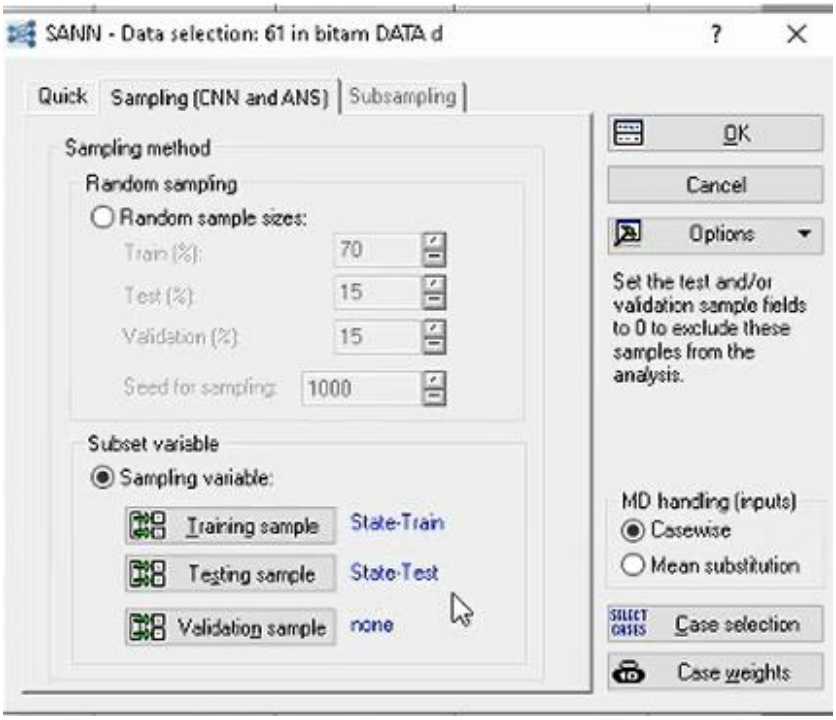


Figure II.7 la sélection des sous-ensembles d'apprentissage et de test

4. Les différents paramètres du réseau RNA sont choisis comme suit **(Figure II.8) :**

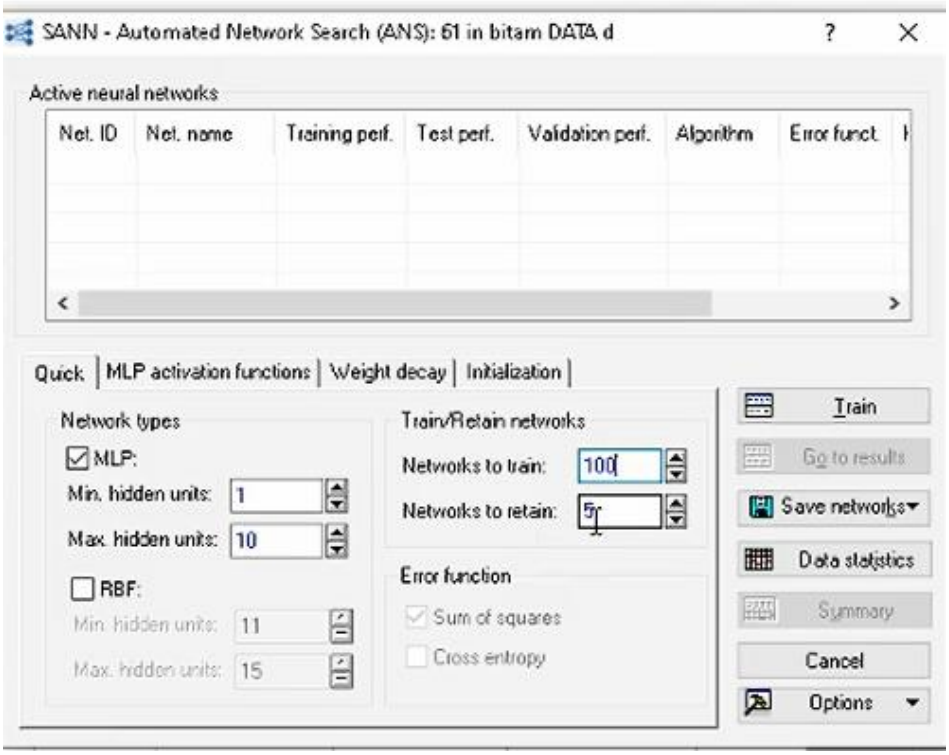


Figure II.8la sélection des paramètres du réseau neurone

5. Dans l'option MLP, on va choisir les fonctions d'activations**(Figure II.9)** :

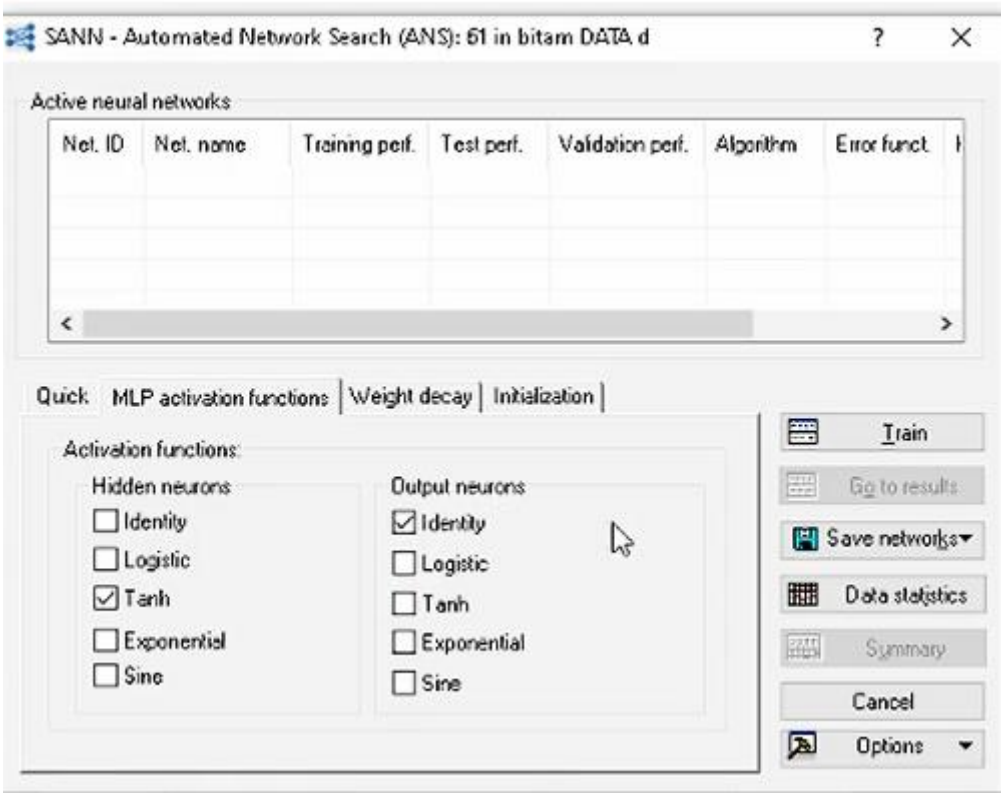


Figure II.9 Le choix des fonctions d'activations

Finalement en choisis le meilleur RNA parmi les cinq réseaux retenus.

II.4-3 Les paramètres de validation

Afin de déterminer la qualité d'un modèle, différents indicateurs statistiques sont employés

- **Coefficient de détermination** (R^2) :

C'est le coefficient de corrélation de Bravais Pearson entre Y_{exp} et Y_{pred}, c'est à dire entre valeurs observées et prédites par le modèle de régression, il est noté R, sa valeur variée entre 0 et 1

$$R^2 = 1 - \frac{\sum(Y_{exp} - Y_{pred})^2}{\sum(Y_{exp} - \bar{Y}_{exp})^2}$$

Plus la valeur de R^2 sera proche de 1 (cas idéal) et plus les valeurs prédites et observées sont corrélées.

- **Coefficient de détermination ajusté($R_{adj}{}^2$) :**

Le coefficient de détermination ajusté tient compte du nombre de variables. En effet, le principal défaut du $\boldsymbol{R^2}$ est de croître avec le nombre de variables explicatives **[36]**. Or, on sait qu'un excès de variables produit des modèles peu robustes. C'est pourquoi on s'intéresse davantage à cet indicateur qu'au $\boldsymbol{R^2}$. Voici l'expressions du $\boldsymbol{R_{adj}{}^2}$, sachant que certains auteurs lui donnent une définition légèrement différente :

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

- **Test Fisher (F):**

Le test Fisher permet de justifier la liaison globale entre Y_n et X_n, une version dérivée de cet indicateur peut juger du degré de pertinence des variables du modèle.

On peut définir quelques paramètres utilisés dans la régression :

Somme des carrés totaux : $TSS = \sum(Y_{exp} - \bar{Y}_{exp})^2$

Somme des carrés expliquée : $ESS = \sum(Y_{pred} - \bar{Y}_{exp})^2$

Somme des carrée résiduels : $RSS = \sum(Y_{exp} - Y_{pred})^2$

Ainsi, $TSS = ESS + RSS$

Fpeut s'écrire comme suit :

$$F = \frac{ESS}{p}\frac{n - p - 1}{RSS}$$

Ou n représente le nombre des molécules, p représente le nombre des descripteurs.

- **Ecart type (s):**

L'écart type (s) est un autre paramètre habituellement rapporté; il indique dans quelle mesure la fonction de régression prédite les données observées **[37]**, ce paramètre est donnée par:

$$s = \sqrt{\frac{RSS}{n-p-1}}$$

- **Coefficient de prédiction (Q^2):**

Si R² est une mesure de la justesse de la corrélation, Q^2 est une mesure de la justesse de la prédiction. Il mesure la capacité prédictive d'un modèle**[37]** et donné par :

$$Q^2 = 1 - \frac{PRESS}{SSY}$$

$$PRESS = \sum (Y_{exp} - Y_{pred})^2$$

$$SSY =: TSS = \sum (Y_{exp} - \bar{Y})^2$$

- **Les facteurs d'inflation de la variance (VIF) :**

Au sens strict, on parle de multicolinéarité parfaite lorsqu'une des variables explicatives d'un modèle est une combinaison linéaire d'une ou plusieurs autres variables explicatives introduites dans le même modèle **[38]**. L'absence de multicolinéarité parfaite est une des conditions requises pour pouvoir estimer un modèle linéaire.

L'approche la plus classique consiste à examiner les *facteurs d'inflation de la variance* (*FIV*) ou *variance inflation factor* (*VIF*) en anglais :

$$VIF = \frac{1}{1 - R_i^2}$$

La valeur du paramètre VIF, calculée doit se situer dans la gamme de valeur de 1 à 5 **[39]**



CHAPITRE III
RESULTATS ET DISCUSSION



CHAPITRE III :RÉSULTAT ET DISCUSSION

III.1 INTRODUCTION :

Dans ce chapitre nous avons discuté les études linéaire et non linéaire de la relation quantitative structure-activité inhibitrice de la protéine de transfert d'ester de cholestérol (CETP) provoque l'augmentation la plus prononcée du cholestérol HDL rapportée jusqu'à présent. L'anacétrapib, un puissant inhibiteur de la CETP. Nous avons fait appel aux réseaux de neurones en tant que méthode d'apprentissage et de sélection de variables respectivement.

III.2 SELECTION DES DESCRIPTEURS

Il est nécessaire et important de réduire le nombre de descripteurs jusqu'à ce que le rapport "nombre de dérivés des il est nécessaire et important de réduire le nombre de descripteurs jusqu'à ce que le rapport "nombre de dérivés des anacétrapib /descripteurs" soit >5 et qu'ils ne soient pas corrélés entre eux **[40].**

Le nombre de descripteurs pertinents obtenus après la sélection effectuée est de 05:**AATSC8i; GATS4s; SpMin7_Bhe; SpMin4_Bhs; n6HeteroRing**. Ces descripteurs ont été utilisés pour le développement des modèles prédictifs. La signification de ces descripteurs est donnée sur le **Tableau III.1** ci-dessous.

Tableau III.1 Descripteurs utilisés pour le développement des modèles QSAR [31]

Type	Descripteur	La description	Classe de descripteur
2D	AATSC8i	Auto corrélation Broto-Moreau centrée moyenne - décalage 8 / pondéré par le premier potentiel d'ionisation	Descripteur de corrélation automatique
2D	GATS4s	Auto corrélation Geary - décalage 4 / pondéré par l'état I	Descripteur de corrélation automatique
2D	SpMin7_Bhe	Plus petite valeur propre absolue de la matrice modifiée de Burden - n 7 / pondérée par les électronégativités de Sanderson relatives	Descripteur des valeurs propres modifiées de la charge
2D	SpMin4_Bhs	Plus petite valeur propre absolue de la matrice modifiée de Burden - n 4 / pondérée par l'état I relatif	Descripteur des valeurs propres modifiées de la charge
2D	n6HeteroRing	Nombre de cycles à 6 chaînons contenant des hétéroatomes (N, O, P, S ou halogènes)	Descripteur du nombre de sonneries

Tableau III.2 Valeurs pIC_{50} expérimentales avec les descripteurs sélectionnés

N°	$pIC_{50\ épx}$	AATSC 8i	GATS4s	SpMin7_Bhe	SpMin4_Bhs	n6HeteroRing	Etat
1	-2.85582	-0.00141	0.85799	1.22396	1.76923	1	Train
2	-2.07918	-0.01347	0.85746	1.22441	1.66043	0	Train
3	-2.80686	0.11114	0.9333	1.26128	1.66607	0	Train
4	-1.93952	0.11591	0.86622	1.23328	1.66676	0	Train
5	-2.23553	0.08193	0.8925	1.22333	1.66107	0	Train
6	-2.02531	0.21189	0.86002	1.22309	1.65111	1	Train
7	-2.09342	0.31473	0.87995	1.23308	1.65186	1	Train
8	-1.6721	0.33156	0.87526	1.22594	1.65103	1	Train
9	-2.29885	0.24791	0.92298	1.22109	1.58859	1	Train
10	-2.38382	-0.01862	0.85401	1.22967	1.66038	0	Train
11	-2.39094	-0.06744	0.80944	1.17966	1.63479	1	Train
12	-3.11126	0.0266	0.87636	1.24401	1.65672	1	Train
13	-2.29667	-0.06056	0.83275	1.19503	1.63688	1	Train
14	-2.35965	-0.04322	0.86879	1.21171	1.67355	0	Train
15	-1.98471	-0.06311	0.90428	1.16586	1.64204	0	Train
16	-2.62118	-0.05321	0.86093	1.37964	1.60769	0	Train
17	-3.24105	-0.12262	0.85313	1.19229	1.65792	1	Train
18	-1.94939	-0.03893	0.8757	1.17196	1.62158	0	Train
19	-2.66839	-0.07223	0.84023	1.36733	1.62282	0	Train
20	-3.21192	-0.14117	0.88083	1.16586	1.67509	1	Train
21	-1.43297	-0.1594	0.80425	1.21208	1.62912	0	Train
22	-3.41481	0.28486	1.03856	1.23214	1.71603	0	Train
23	-3.08207	-0.22399	0.84653	1.16586	1.69727	1	Train
24	-2.40518	0.28116	0.89595	1.3447	1.61217	0	Train
25	-1.92901	-0.1747	0.88154	1.16586	1.62703	0	Train
26	-2.08636	-0.24865	0.90835	1.16586	1.5387	0	Train
27	-1.3962	-0.00676	0.92731	1.16586	1.59097	1	Train
28	-2.56984	-0.15223	0.78829	1.16586	1.5852	0	Train
29	-1.29226	0.10283	0.81016	1.21683	1.66195	0	Train
30	-1.25527	0.07458	0.86631	1.21198	1.61097	0	Train
31	-1.52504	0.16366	0.90063	1.21252	1.65848	0	Train
32	-1.85612	0.10896	0.85057	1.21225	1.61227	0	Train
33	-1.39445	0.16376	0.8957	1.16586	1.58228	0	Train
34	-2.35064	-0.06485	0.92226	1.16586	1.63255	0	Train
35	-1.7332	0.12642	0.86567	1.21208	1.62843	0	Train
36	-1.12385	0.12814	0.91229	1.16586	1.59036	0	Train
37	-1.20952	0.19708	0.90606	1.16586	1.59951	0	Train
38	-1.36736	0.1966	0.91625	1.16586	1.60279	0	Train
39	-2.41963	-0.05961	0.88451	1.16586	1.62743	0	Train
40	-2.21932	-0.3418	0.90398	1.12888	1.61381	0	Train
41	-2.36791	-0.13678	0.91198	1.12894	1.61947	0	Train

42	-2.67624	-0.26012	0.87114	1.16586	1.6311	0	Train
43	-2.46909	0.098	0.93023	1.12864	1.59586	1	Train
44	-2.01745	-0.22213	0.93459	1.12887	1.6043	0	Train
45	-2.273	0.04194	0.93928	1.12891	1.61232	0	Train
46	-3.08743	-0.43645	0.88092	1.12905	1.62167	0	Train
47	-2.5001	0.06655	0.92879	1.12862	1.58396	1	Train
48	-2.24822	-0.15864	0.91908	1.16586	1.6328	0	Train
49	-1.22272	0.17583	0.88864	1.16586	1.59832	0	Train
50	-1.39794	0.09727	0.89201	1.16586	1.5877	0	Test
51	-1.29667	0.22575	0.90942	1.16586	1.59126	0	Test
52	-1.37107	0.25276	0.90918	1.16586	1.59133	0	Test
53	-1.21748	0.2016	0.91099	1.16586	1.59081	0	Test
54	-1.49429	0.18789	0.91047	1.16586	1.59423	0	Test
55	-2.65591	-0.31096	0.9075	1.12888	1.61813	0	Test
56	-2.42488	-0.69686	0.92664	1.16586	1.48204	0	Test
57	-2.5358	-0.40738	0.89961	1.16586	1.51369	0	Test
58	-1.86705	-0.24756	0.85697	1.12904	1.59324	0	Test
59	-2.08707	-0.6473	0.8822	1.16586	1.52087	0	Test
60	-1.75305	-0.35718	0.89352	1.16586	1.55782	0	Test
61	-2.01199	-0.44643	0.91264	1.16586	1.46821	0	Test

III.3 MODELE MLR

Le modèle MLR obtenu pour la prédiction de la concentration inhibitrice des dérivésde l'anacétrapib, en utilisant 61 composés, est donné par la relation linéaire ci-dessous. Les valeurs de **pCI_{50}**prédites par le modèle MLR pour l'ensemble des données (61 composés) sont présentées dans le **Tableau III.2**

La meilleure corrélation est exprimée par la relation mathématique suivante:

pCI_{50} = 20,392 + 2,133 **AATSC8i** -7,780 **GATS4s** - 4,530 **SpMin7_Bhe** - 6,183 **SpMin4_Bhs** - 0,**650 n6HeteroRing** **eq(III.1)**

Tableau III.3 les différant paramètre MLR des deux Ensemble (apprentissage et test)

Les paramètres de validation (RLM)	
N_{appr} = 49	**N_{test}**=12
S (écart type) = 0,946	**S** = 0,798
R = 0,900	**R** = 0,920
R-deux = 0,810	**R-deux** = 0,847
R-deux ajusté =0 ,788	**R-deux ajusté**= 0,759
variation de F= 36.695	**Variation de F** = 9.672

Les signes des coefficients de régression suggèrent la direction de l'influence des descripteurs dans un modèle donné. D'après les coefficients de régression de l'équation (**eq. III.1**), il convient de noter que le coefficient de régression du descripteur AATSC8i est de signes positif. Ce descripteur est eu une influence positive et les valeurs les plus élevées vont contribuer à améliorer le pouvoir inhibiteur des analogues de l'anacétrapib.

D'après les résultats statistiques, les Cinq descripteurs pertinents (variables) de l'équation. Pourraient expliquer 78.8 % de la variance (coefficient de variation ajusté) de la Concentration inhibitrice.

Pour évaluer la multi-colinéarité entre les cinq descripteurs pertinents, leurs facteurs de variation d'inflation (VIF) ont été calculés. Si VIF est égal à 1, aucune intercorrélation n'existe pour chaque variable; si VIF est situé dans l'intervalle de 1 à 5, la sélection des descripteurs est acceptable et si VIF est supérieur à 10, le modèle associé aux descripteurs sélectionnés est instable et une nouvelle sélection est nécessaire **[41].** D'après le **Tableau III.4**, tous les descripteurs ont des valeurs VIF inférieures à 2, ce qui indique que la sélection effectuée est bonne et que les descripteurs obtenus ne sont pas corrélés entre eux **[39]**.

Tableau III.4 Valeurs du coefficient VIF des descripteurs sélectionnés

Descripteur	AATSC8i	GATS4s	SpMin7_Bhe	SpMin4_Bhs	n6HeteroRing
VIF	1,438	1 ,203	1,288	1,418	1,142

Afin d'étudier la corrélation entre les 5 descripteurs sélectionnés, la matrice de corrélation a été établie. La valeur du coefficient de corrélation de chaque paire de descripteurs sélectionnés a été examinée (**Tableau III.5**).La plus grande valeur du coefficient de corrélation appartenant à la paire de descripteurs SpMin4_Bs /AATSC8i étant égale à 0,430 (donc < 0.7), cela signifie que tous les descripteurs sélectionnés sont indépendants les uns des autres.

Tableau III.5 : Matrice des coefficients de corrélation des descripteurs sélectionnés

	pIC_{50}	AATSC8i	GATS4s	SpMin7_Bhe	SpMin4_Bs	n6HeteroRing
pIC_{50}	**1.000**	0.320	-0.155	-0.173	-0.344	-0.411
AATSC8i		**1.000**	0.122	0.335	0.430	0.187
GATS4s			**1.000**	-0.249	-0.182	-0.144
SpMin7_Bhe				**1.000**	0.310	0.028
SpMin4_Bhs					**1.000**	0.308
n6HeteroRing						**1.000**

Sur les **Figures III.1 et III.2** sont représentées les concentrations inhibitrices prédites en fonction des concentrations inhibitrices expérimentales pour l'ensemble d'apprentissage et de test.

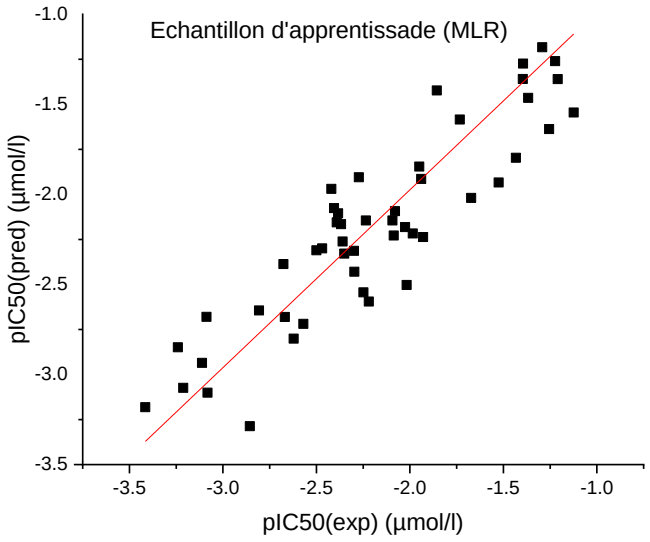


Figure III.1 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et celle prédite par le modèle MLR pour la série d'Ensemble d'apprentissage des analogues de L'anacétrapib

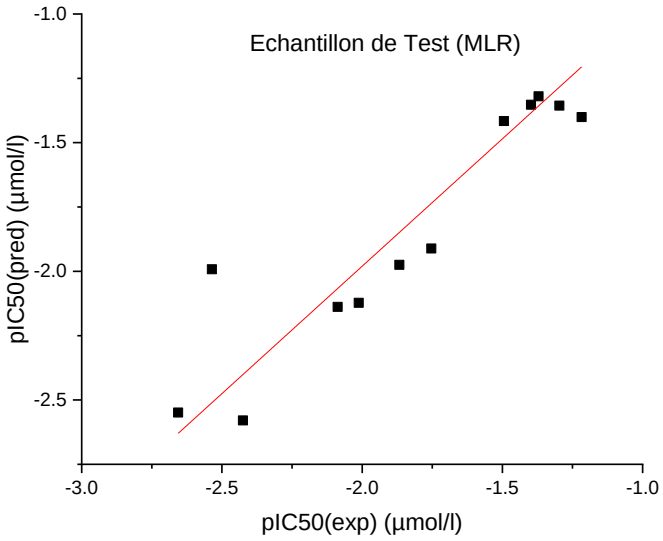


Figure III.2 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et prédite par le modèle MLR pour la série d'Ensemble de test des analogues de L'anacétrapib

III.4 MODELE NEURONAL RAN (ANN):

Les réseaux de neurones artificiels sont très peu utilisés en anthropologie biologique, alors que leur application est en plein essor dans de nombreuses disciplines. Ils constituent, entre autre, une alternative intéressante aux statistiques traditionnelles pour le traitement des données.**[42]**

L'objectif général d'un RNA est de trouver la configuration des poids de connexion entre neurones pour qu'il associe à chaque configuration d'entrée, une réponse adéquate. L'utilisation d'un RNA se fait en deux temps. Tout d'abord une phase d'apprentissage qui est chargée d'établir des valeurs pour chacune des connexions du réseau, puis une phase d'utilisation proprement dite, où l'on présente au réseau une entrée et où il nous indique en retour « sa » sortie calculée. **[42]**

Les concentrations inhibitrices prédites par le modèle ANN pour l'ensemble des données (61 composés) ont été obtenus; elles sont rassemblées dans le **Tableau III.6.**

Sur les **Figures III.3 et III.4** sont représentées les concentrations inhibitrices prédites en fonction des concentrations inhibitrices expérimentales pour l'ensemble d'apprentissage et de test.

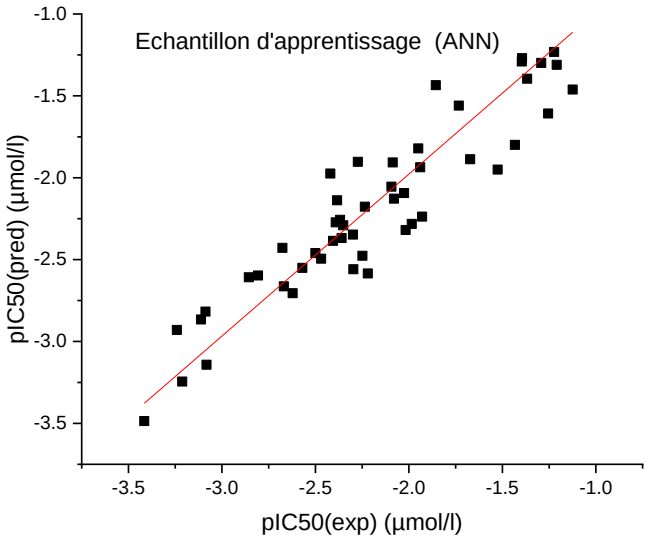


Figure III.3 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et celle prédite par le modèle ANN pour la série d'Ensemble d'apprentissage des analogues de L'anacétrapib

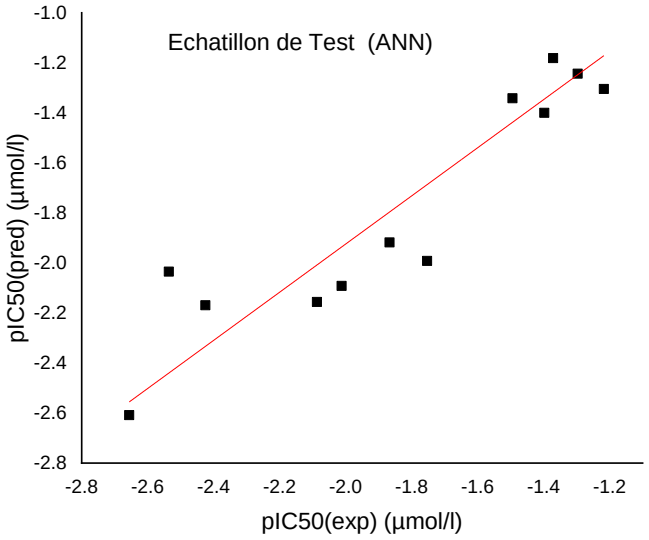


Figure III.4 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et celle prédite par le modèle ANN pour la série d'Ensemble de test des analogues de L'anacétrapib

D'après les **Figures III.3 et III.4**, en remarque une bonne corrélation entre les valeurs prédites par le modèle MLP et les valeurs expérimentale a été constatée. Le modèle non linéaire des réseaux de neurones (modèle MLP) a donné de bons résultats avec des coefficients de corrélation R^2 élevés.

Tableau III.6 les différant paramètre MLP des deux Ensemble (apprentissage et test)

Les paramètres de validation (ANN)	
N $_{appr}$= 49	**N**$_{test}$=12
S (écart type) = 0,230	; **S** = 0,103
R = 0,930	**R**= 0,915
R-deux = 0,864	**R-deux** = 0,849
R-deux ajusté =0 ,848	**R-deux ajusté**= 0,763
Q^2 = 0.848	**Q**^{2} = 0.700
variation de F= 54.617	**Variation de F** = 6.140

Selon la recommandation de Tropsha et Golbraikh**[43]**, si la différence entre **R**^{2} et **Q**^{2}ne dépasse pas 0.3, il n'y'a pas de sur-apprentissage dans le modèle. Dans le présent travail, cette différence est inferieur à 0.3. En conséquence, il y'a absence du sur-apprentissage dans le modèle MLP

III.5 COMPARAISONS ENTRE LE MODELE RNA ET RLM :

Tableau III.7 analogues de L'anacétrapib utilisés et valeurs des concentrations inhibitrices expérimentales et prédites par MLR, MLP (ANN)

N°	pIC$_{50}$ expérimentale	pIC$_{50}$ - prédite par ANN	pIC$_{50}$ - prédite par RLM
Ensemble d'apprentissage			
1	-2.856	-2.608	-3.287
2	-2.079	-2.128	-2.093
3	-2.807	-2.598	-2.645
4	-1.940	-1.936	-1.915
5	-2.236	-2.177	-2.145
6	-2.025	-2.094	-2.182
7	-2.093	-2.055	-2.145
8	-1.672	-1.887	-2.020
9	-2.299	-2.347	-2.313
10	-2.384	-2.138	-2.105
11	-2.391	-2.273	-2.156
12	-3.111	-2.866	-2.935
13	-2.297	-2.559	-2.429
14	-2.360	-2.368	-2.262

15	-1.985	-2.282	-2.217
16	-2.621	-2.706	-2.801
17	-3.241	-2.930	-2.849
18	-1.949	-1.821	-1.846
19	-2.668	-2.663	-2.680
20	-3.212	-3.245	-3.074
21	-1.433	-1.799	-1.797
22	-3.415	-3.486	-3.181
23	-3.082	-3.142	-3.101
24	-2.405	-2.385	-2.077
25	-1.929	-2.237	-2.237
26	-2.086	-1.907	-2.228
27	-2.570	-2.551	-1.359
28	-1.292	-1.299	-2.719
29	-1.396	-1.290	-1.183
30	-1.255	-1.608	-1.638
31	-1.525	-1.950	-1.934
32	-1.856	-1.435	-1.423
33	-1.394	-1.270	-1.274
34	-2.351	-2.291	-2.329
35	-1.733	-1.560	-1.584
36	-1.124	-1.462	-1.546
37	-1.210	-1.311	-1.360
38	-1.367	-1.396	-1.464
39	-2.420	-1.975	-1.969
40	-2.219	-2.585	-2.595
41	-2.368	-2.257	-2.165
42	-2.676	-2.428	-2.387
43	-2.469	-2.494	-2.300
44	-2.017	-2.320	-2.503
45	-2.273	-1.903	-1.905
46	-3.087	-2.818	-2.680
47	-2.500	-2.460	-2.310
48	-2.248	-2.476	-2.544
49	-1.223	-1.232	-1.260
Ensemble de test			
50	-1.398	-1.400	-1.353
51	-1.297	-1.243	-1.356
52	-1.371	-1.182	-1.320

53	-1.217	-1.305	-1.401
54	-1.494	-1.342	-1.416
55	-2.656	-2.608	-2.549
56	-2.425	-2.169	-2.579
57	-2.536	-2.035	-1.992
58	-1.867	-1.918	-1.975
59	-2.087	-2.156	-2.138
60	-1.753	-1.992	-1.912
61	-2.012	-2.092	-2.122

Au cours de ce travail nous avons appliqué la méthode des réseaux de neurones (RNA) Et la régression linéaire multiple (RML) dans le développement d'un modèle décrivant la relation quantitative structure-activité inhibitrice de l'anacétrapib. Pour cela nous avons fait appel à la méthode des réseaux de neurones artificiels (RNA) à multicouches en tant que méthode d'apprentissage. Cinq descripteurs moléculaires ont été utilisés dans l'apprentissage des RNA. qui sont : AATSC8i; GATS4s; SpMin7_Bhe; SpMin4_Bhs; n6HeteroRing La comparaison des bons résultats obtenus par l'apprentissage des RNA avec ceux obtenus par la RLM montre que les RNA est une méthode adaptée à la modélisation de l'activité inhibitrice de la protéine de transfert d'ester de cholestérol (CETP). Ils permettent non seulement d'obtenir une prédiction mais également de donner de meilleurs résultats que l'analyse par la régression linéaire multiple (RLM). Dans la RLM. L'activité inhibitrice est une combinaison linéaire des descripteurs. Par contre, dans les RNA leur pouvoir prédictif provient de leur capacité à produire des combinaisons non linéaires. L'étude de la contribution des cinq descripteurs sélectionnés a montré que le descripteur moléculaire AATSC8i, exerce l'effet le plus important sur le pouvoir prédictif de l'activité inhibitrice de l'anacetrapib.

Tableau III.8 Coefficients de détermination par les modèles RLM et ANN.

	RLM		**RNA**	
	R^2	**R^2_{aju}**	**R^2**	**R^2_{aju}**
Ensemble d'apprentissage	0.810	0.788	0.864	0848
Ensemble de test	0.847	0.759	0.849	0.763

En se basant sur les résultats statistiques donnés par les deux modèles RNA et RLM, nous observons que le modèle RNA a une fiabilité et une capacité prédictive

nettement supérieures à celles données par le modèle RLM. Les erreurs types de calcul sont plus faibles, les coefficients de détermination des ensembles d'apprentissage et de prédiction sont plus élevés avec le modèle RNA qu'avec le modèle RLM. Par conséquent, nous considérons que l'approche non linéaire est la plus convenable pour analyser la relation quantitative structure-activité inhibitrice de de la protéine de transfert d'ester de cholestérol (CETP) par les analogues de L'anacétrapib.

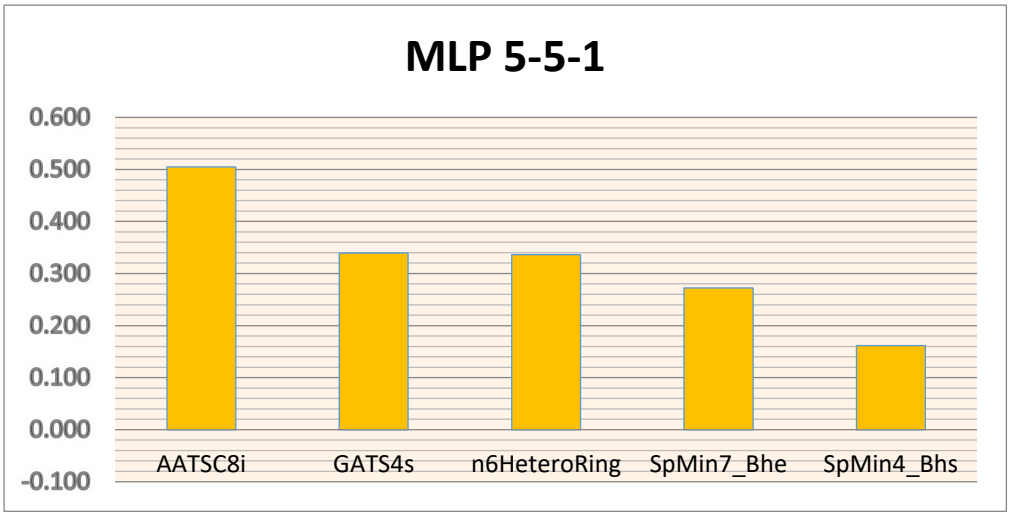


Figure III.5 Contribution des descripteurs dans le cas du modèle MLP

III.6 CONCLUSION

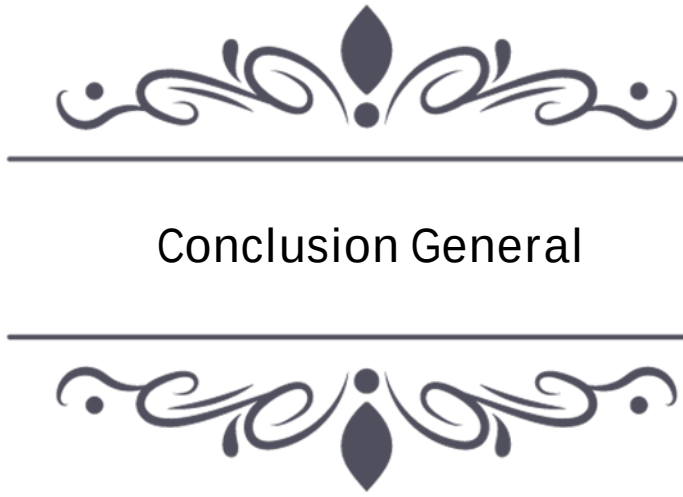
Un réseau linéaire et neuronal multiple a été utilisé pour construire. Le réseau de neurones artificiels avait un pouvoir prédictif significativement meilleur que les autres modèles, avec un pouvoir prédictif supérieur. Nous avons établi des relations satisfaisantes entre de nombreuses descriptions et l'activité des récepteurs d'antigènes (CETP), avec vérification mutuelle.

Les résultats ont montré que le modèle proposé dans ce travail peut prédire l'activité avec précision et que les descriptions sélectionnées sont pertinentes. Les résultats montrent que le modèle proposé dans ce travail peut prédire l'activité avec précision et que les descriptions sélectionnées sont appropriées.

CHAPITRE III :RÉSULTAT ET DISCUSSION

La précision et la prévisibilité des modèles proposés sont illustrées en comparant les principaux termes statistiques R ou R^2 pour différents modèles (**Tableau III.8)**.

Les descripteurs sont suffisamment riches en informations chimiques, électroniques et topographiques pour coder les caractéristiques structurelles qui peuvent être utilisées avec d'autres descripteurs pour développer des modèles structurels prédictifs quantitatifs.



Conclusion General

CONCLUSION GENERALE :

Les méthodes QSAR constituent un outil puissant d'information sur les caractéristiques probables d'une substance à condition d'être maniés et interprétés par un utilisateur averti. Ils peuvent permettre d'éviter des tests expérimentaux ou de ne les effectuer que lorsque l'interprétation des résultats laisse planer une suspicion. Ils peuvent permettre aussi de compléter l'information expérimentale.

Les méthodes QSAR sont faciles à mettre en œuvre, ils fournissent un résultat immédiat et leurs coûts sont sans commune mesure avec ceux engendrés par la réalisation de tests expérimentaux. Ils peuvent être réalisés par des laboratoires équipés des logiciels adaptés ou par le demandeur lui-même s'il dispose des outils informatiques et des personnes compétentes pour valider les résultats car on ne saurait trop insister sur les risques de conclusions hâtives.

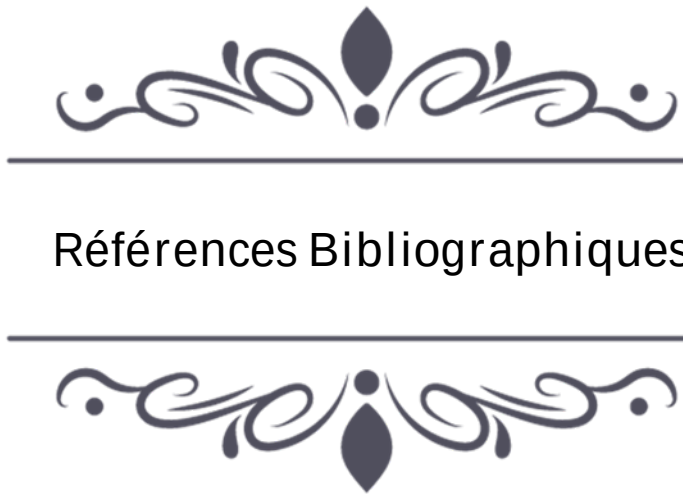
Dans ce travail nous avons réalisé une étude quantitative des relations structure-activité (QSAR) sur ensemble des analogues d'anacétrapib, afin de déterminer la relation entre l'activité biologique et leurs structures chimiques en utilisant la régression linéaire multiple (RLM) et les réseaux neurones (ANN).

Avec un pouvoir prédictif supérieur, nous avons établi des relations satisfaisantes entre de nombreuses descriptions et l'activité des récepteurs d'antigènes (CETP), avec vérification mutuelle.

Les descripteurs sont suffisamment riches en informations chimiques, électroniques et topographiques pour coder les caractéristiques structurelles qui peuvent être utilisées avec d'autres descripteurs pour développer des modèles structurels prédictifs quantitatifs.

La capacité prévisionnelle et la robustesse de modèle été examiné par les paramètresR^2 et R^2 ajusté. Les résultats ont indiqué que le modèle basés sur MLR a pu prévoir plus 0.86.La comparaison du résultat de R^2 et R^2 ajusté confirme la supériorité de modèle.

Les résultats ont montré que le modèle proposé dans ce travail peut prédire l'activité avec précision et que les descriptions sélectionnées sont pertinentes. Les résultats montrent que le modèle proposé dans ce travail peut prédire l'activité avec précision et que les descriptions sélectionnées sont appropriées.



Références Bibliographiques

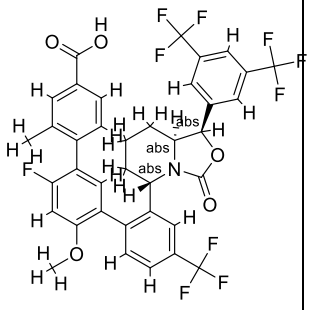
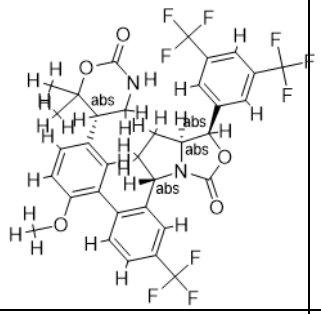
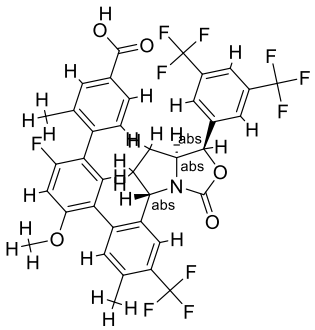
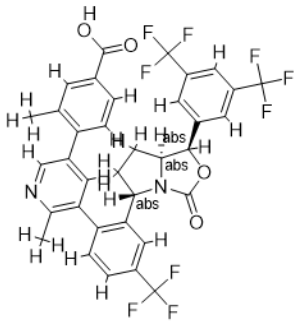
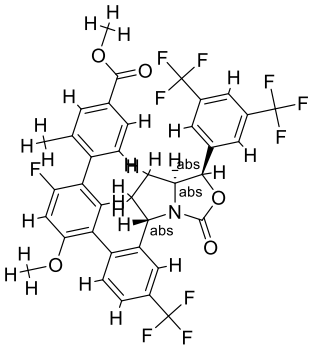
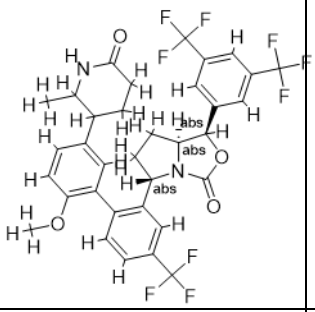
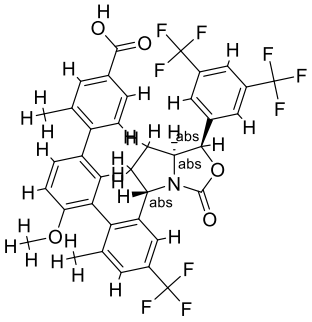
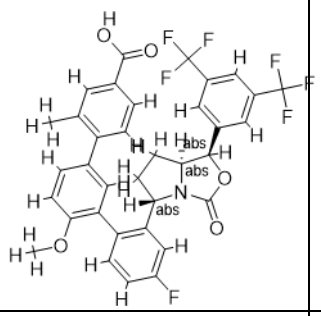
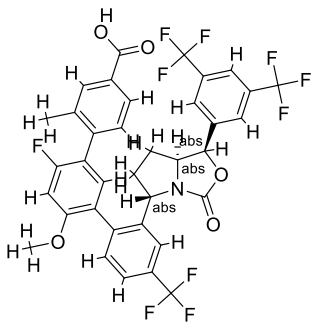
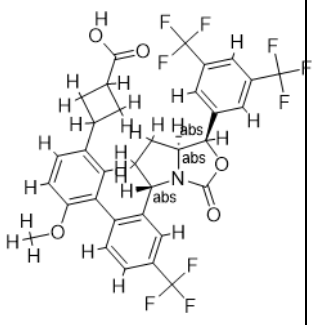
Références bibliographiques

1. https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). *28 juin 2017*. Access Date : 20 Avril 2020.
2. http://www.ehnheart.org/cvd-statistics.html, Access Date : 20 Avril 2020.
3. https://www.who.int/countries/fra/fr/, *2016.* Access Date : 20 Avril 2020.
4. http://www.who.int/cardiovascular_diseases/about_cvd/en/, Access Date : 20 Avril 2020.
5. Mathers, C.D. and D. Loncar, *Projections of global mortality and burden of disease from 2002 to 2030.*PLoS medicine, 2006. **3**(11): p. e442-e442.
6. Vayer, P., et al., *Apports de la chémo-informatique dans la recherche et l'optimisation des molécules d'intérêt thérapeutique.* médecine/sciences, 2009. **25**(10): p. 871-877.
7. https://www.who.int/cardiovascular_diseases/about_cvd/fr/#, Access Date : 10 Mai 2020.
8. Brunko, P., *Recherche et innovation pharmaceutique à l'approche du troisième millénaire*. 1994: Libbey Eurotext.
9. Morgan, S., et al., *The cost of drug development: a systematic review.*Health policy, 2011. **100**(1): p. 4-17.
10. Landry, Y. and J.P. Gies, *Pharmacologie - 3e édition: Des cibles à la thérapeutique*. 2014: Dunod.
11. Bajorath, J., *Integration of virtual and high-throughput screening.*Nature Reviews Drug Discovery, 2002. **1**(11): p. 882-894.
12. Mouly, S., et al., *Évaluation de l'efficacité d'un médicament: de la découverte à la mise sur le marché.* Journal français d'ophtalmologie, 2008. **31**(1): p. 75-79.
13. Rivail, J.L., *Éléments de chimie quantique à l'usage des chimistes (2e édition)*. 2012: EDP Sciences.
14. Cances, E., C.L. Bris, and Y. Maday, *Méthodes mathématiques en chimie quantique. Une introduction*. 2006: Springer Berlin Heidelberg.
15. Chtita, S., *Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments*. 2017.
16. Goulon-Sigwalt-Abram, A., *Une nouvelle méthode d'apprentissage de données structurées: applications à l'aide à la découverte de médicaments*. 2008, Université Pierre et Marie Curie-Paris VI.
17. Todeschini, R. and V. Consonni, *Handbook of molecular descriptors*. Vol. 11. 2008: John Wiley & Sons.
18. Balaban, A.T., *Highly discriminating distance-based topological index.* Chemical Physics Letters, 1982. **89**(5): p. 399-404.
19. Randic, M., *Characterization of molecular branching.* Journal of the American Chemical Society, 1975. **97**(23): p. 6609-6615.
20. Wiener, H., *Structural determination of paraffin boiling points.* Journal of the American Chemical Society, 1947. **69**(1): p. 17-20.
21. Rakotomalala, R., *Pratique de la regression lineaire multiple.*
22. Roy, K. and P.P. Roy, *Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques.* European journal of medicinal chemistry, 2009. **44**(7): p. 2913-2922.
23. McCulloch, W.S. and W. Pitts, *A logical calculus of the ideas immanent in nervous activity.* The bulletin of mathematical biophysics, 1943. **5**(4): p. 115-133.
24. Bishop, C.M., et al., *Neural Networks for Pattern Recognition*. 1995: Clarendon Press.
25. Dreyfus, G. and Y. Idan, *The canonical form of nonlinear discrete-time models.*Neural Computation, 1998. **10**(1): p. 133-164.

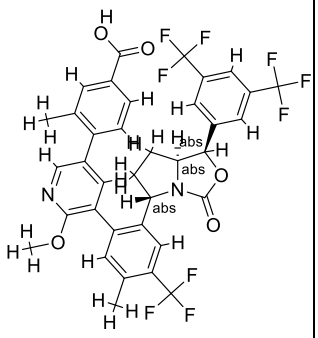
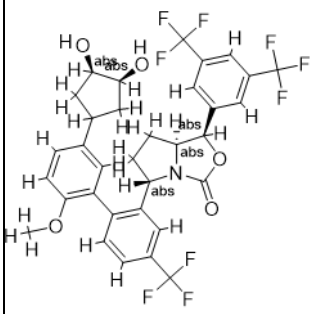
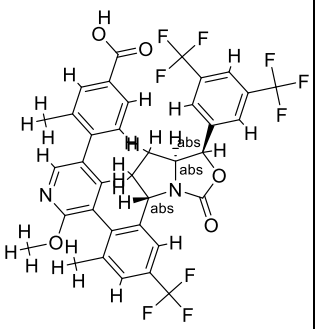
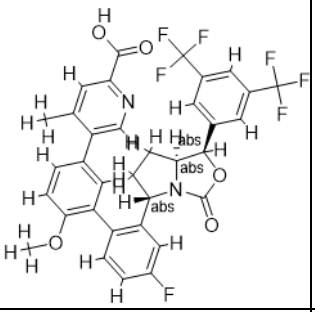
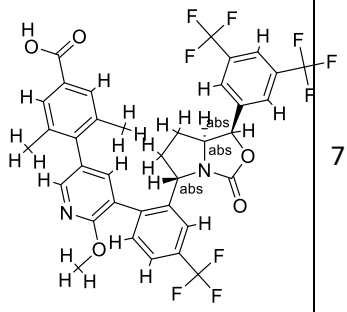
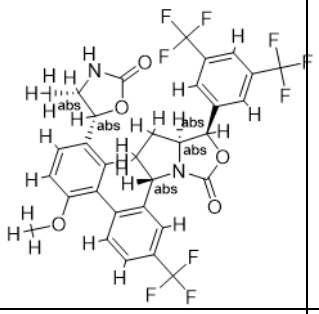
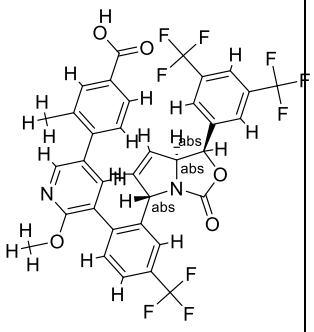
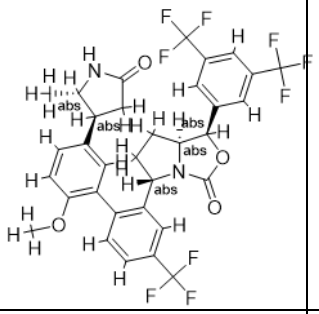
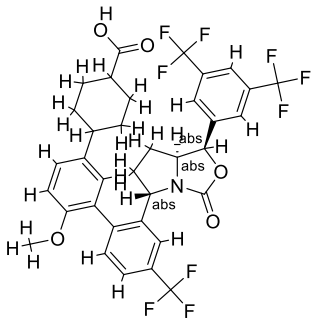
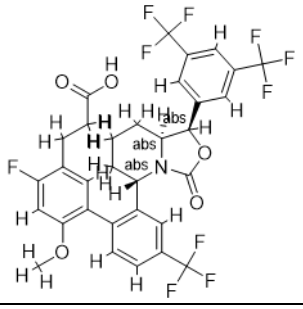
26. a Siou, L.K., *Modélisation des crues de bassins karstiques par réseaux de neurones. Cas du bassin du Lez (France)*. 2011.
27. Katritzky, A.R., M. Karelson, and V.S. Lobanov, *QSPR as a means of predicting and understanding chemical and physical properties in terms of structure.* Pure and Applied Chemistry, 1997. **69**(2): p. 245-248.
28. Katritzky, A.R., V.S. Lobanov, and M. Karelson, *QSPR: the correlation and quantitative prediction of chemical and physical properties from structure.* Chemical Society Reviews, 1995. **24**(4): p. 279-287.
29. Liu, T., et al., *BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities.* Nucleic acids research, 2007. **35**(suppl_1): p. D198-D201.
30. ChemOffice, *PerkinElmer Informatics.* 2010.
31. Yap, C.W., *PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints.* Journal of computational chemistry, 2011. **32**(7): p. 1466-1474.
32. *IBM SPSS Statistics.* 21.0. **213045**.
33. http://www.yapcwsoft.com/dd/padeldescriptor/. Access Date : 01 Avril 2020.
34. Kalivas, J.H., *Adaption of simulated annealing to chemical optimization problems*. 1995: Elsevier.
35. Othman, F. and M. Naseri, *Reservoir inflow forecasting using artificial neural network.*Int. J. Phys. Sci, 2011. **6**(3): p. 434-440.
36. Rakotomalala, R., *Pratique de la regression lineaire multiple.* Diagnostic et selection de variables, 2011.
37. Leach, A.R. and V.J. Gillet, *An Introduction to Chemoinformatics*. 2007: Springer Netherlands.
38. Foucart, T., *Colinéarité et régression linéaire.* Mathématiques et sciences humaines. Mathematics and social sciences, 2006(173).
39. Confais, J. and M. Le Guen, *Premiers pas en régression linéaire avec SAS®.* 2007.
40. Tropsha, A., P. Gramatica, and V.K. Gombar, *The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models.* QSAR & Combinatorial Science, 2003. **22**(1): p. 69-77.
41. Jaiswal, M., et al., *Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides.*Bioorganic & medicinal chemistry letters, 2004. **14**(12): p. 3283-3290.
42. Schmitt, A., et al., *Les réseaux de neurones artificiels.* Bulletins et mémoires de la Société d'Anthropologie de Paris, 2001. **13**(1): p. 2.
43. Golbraikh, A. and A. Tropsha, *Beware of q2!* Journal of molecular graphics and modelling, 2002. **20**(4): p. 269-276.

ANNEXES

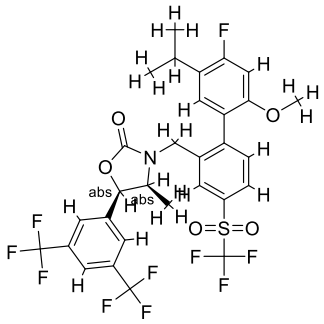
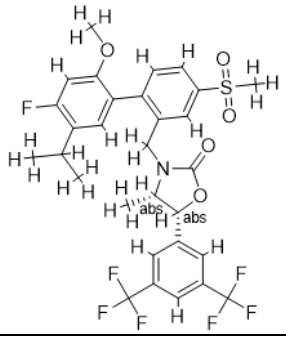
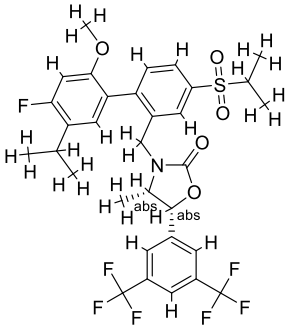
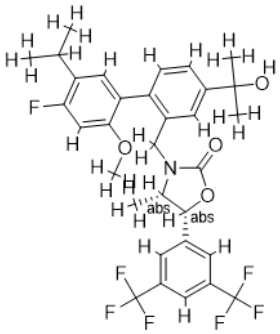
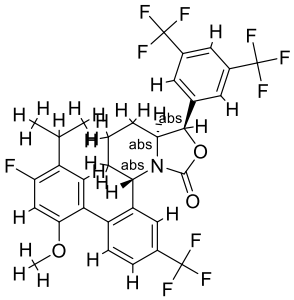
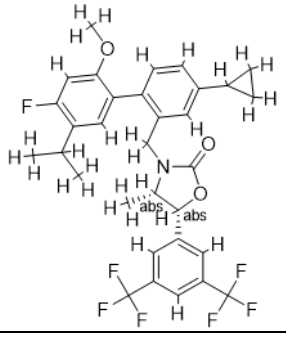
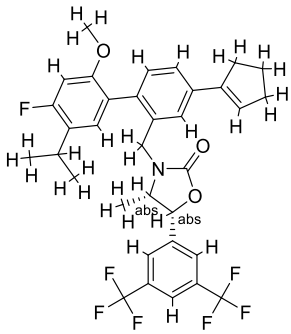
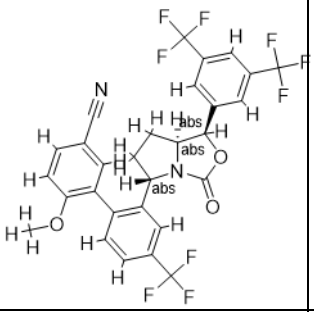
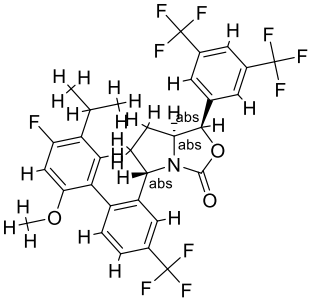
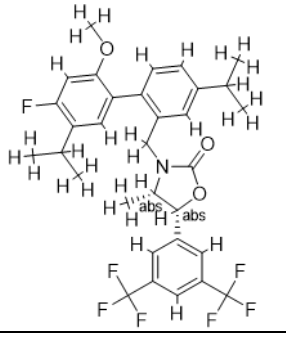
ANNEX

N°=	Structure	CID	pIC50	N°=	Structure	CID	pIC50
1		89522387	-2.391	11		89522504	-2.391
2		71526289	-3.111	12		71526465	-3.111
3		71526635	-2.297	13		71526555	-2.297
4		71526291	-2.360	14		71525739	-2.360
5		71526636	-1.985	15		71526637	-1.985

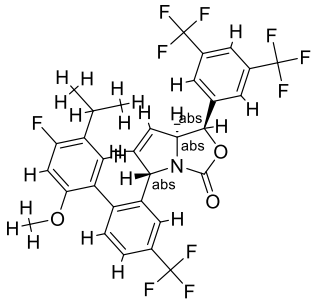
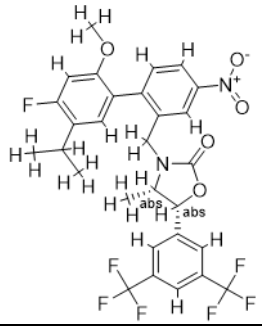
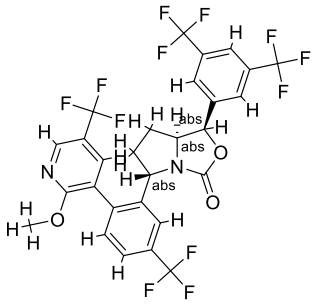
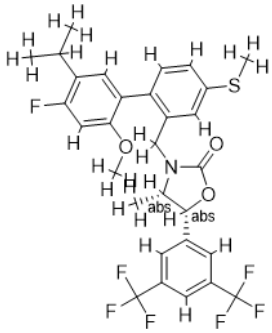
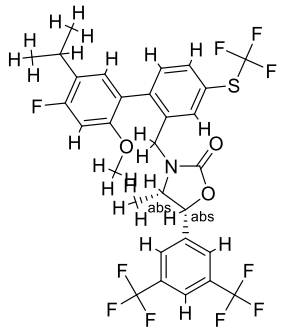
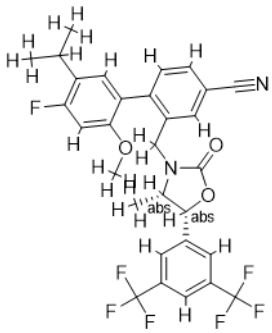
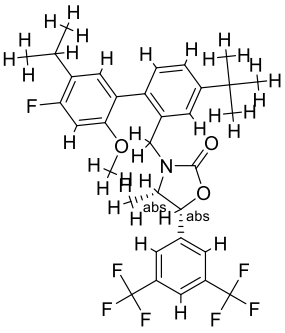
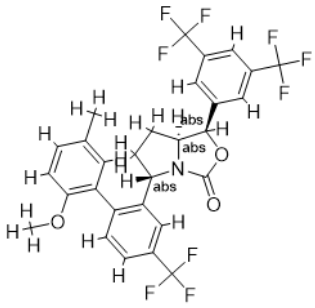
ANNEX

N°=	Structure	CID	pIC50	N°=	Structure	CID	pIC50
6		71526290	-2.621	16		71526288	-2.621
7		71526382	-3.241	17		71525834	-3.241
8		71526551	-1.949	18		89522501	-1.949
9		89522406	-2.668	19		89522372	-2.668
10		71526200	-3.212	20		89522213	-3.212

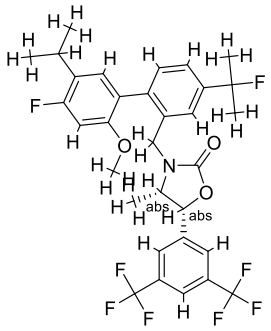
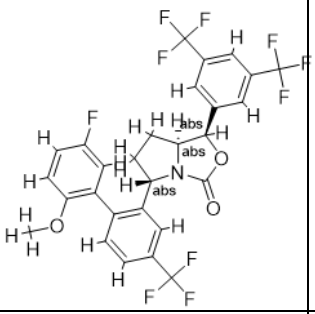
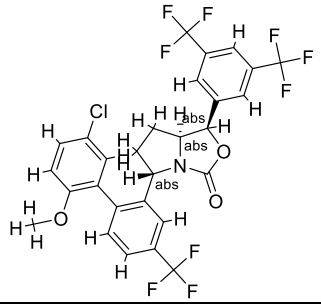
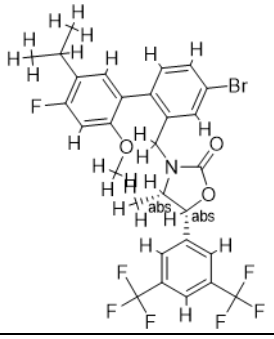
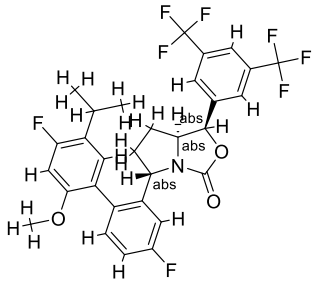
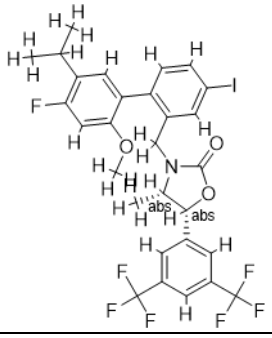
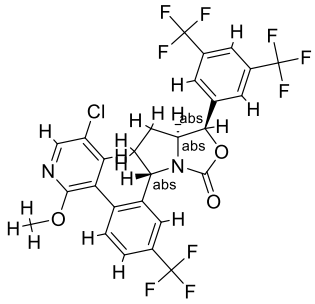
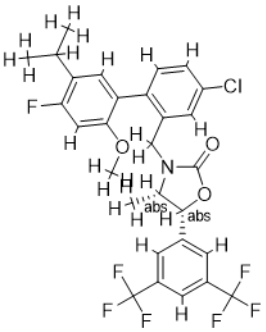
ANNEX

N°=	Structure	CID	pIC50	N°=	Structure	CID	pIC50
21		57390584	-1.433	31		11721583	-1.525
22		57397759	-3.415	32		57402947	-1.856
23		89522469	-3.082	33		11628674	-1.394
24		57394194	-2.405	34		71526721	-2.351
25		71525924	-1.929	35		11657340	-1.733

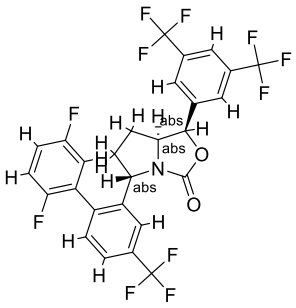
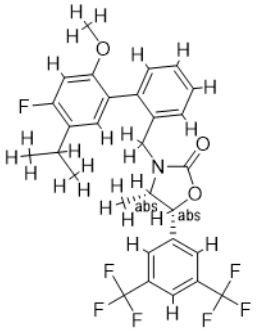
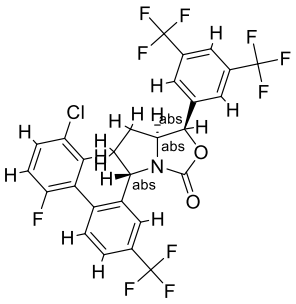
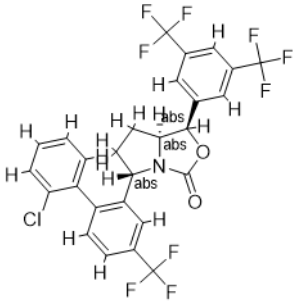
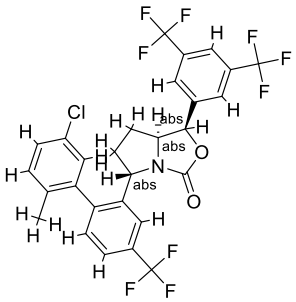
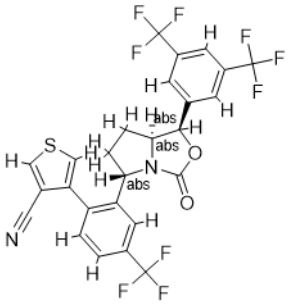
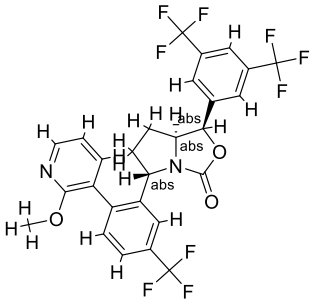
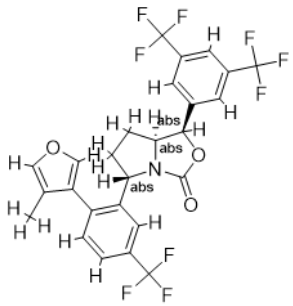
ANNEX

N°=	Structure	CID	pIC50	N°=	Structure	CID	pIC50
26		71526105	-2.086	36		11527347	-1.124
27		71526722	-2.570	37		11650234	-1.210
28		57392491	-1.292	38		11706999	-1.367
29		57394193	-1.396	39		89522451	-2.420

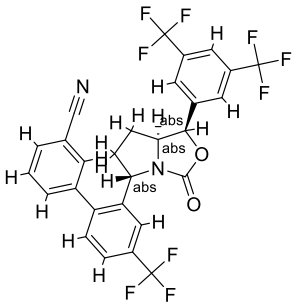
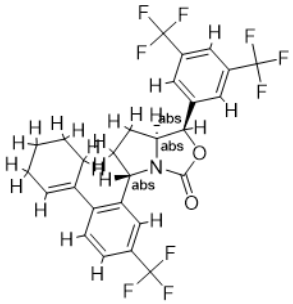
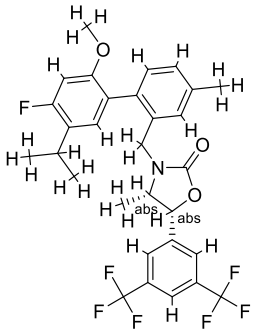
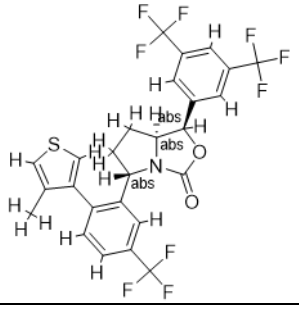
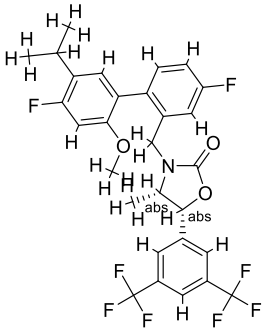
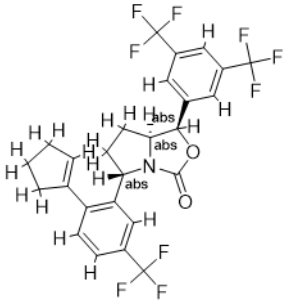
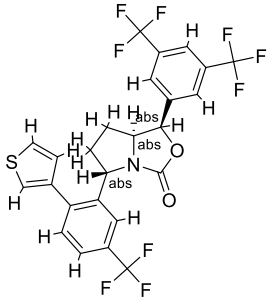
ANNEX

N°=	Structure	CID	pIC50	N°=	Structure	CID	pIC50
30		57390685	-1.255	40		71525462	-2.219
41		71525381	-2.368	51		11556473	-1.297
42		71525833	-2.676	52		57395961	-1.371
43		71526723	-1.367	53		11578015	-1.217

ANNEX

N°=	Structure	CID	pIC50	N°=	Structure	CID	pIC50
44		71525552	-2.017	54		57402946	-1.494
45		71525550	-2.273	55		89522516	-2.656
46		71525643	-3.087	56		71526724	-2.425
47		89522333	-2.500	57		89522295	-1.371

ANNEX

N°=	Structure	CID	pIC50	N°=	Structure	CID	pIC50
48		71525463	-2.248	58		89522453	-1.867
N°=	Structure	CID	pIC50	N°=	Structure	CID	pIC50
49		11649992	-1.223	59		89522402	-2.087
50		11664298	-1.398	60		71525549	-1.753
61		71525461	-2.012				