

Année Universitaire : 2018-2019

**Master Sciences et Techniques : CMBA
Chimie des Molécules Bio Actives**



MEMOIRE DE FIN D'ETUDES

Pour l'Obtention du Diplôme de Master Sciences et Techniques

Etude de la Relation Quantitative Structure Activité de l'acide ursolique et ses dérivés

Présenté par :

Fatima EN-NAHLI

Encadré par :

- Pr. Hamid TOUFIK (FPT Taza)
- Pr. Youssef KANDRI RODI (FST Fès)

Co-encadré par :

- Pr. Fatima LAMCHOURI (FPT Taza)

Soutenu Le 20 Juin 2019 devant le jury composé de :

- Pr. Youssef KANDRI RODI (FST Fès)
- Pr. El Hadi LAMCHARFI (FST Fès)
- Pr. Fatima LAMCHOURI (FPT Taza)
- Pr. Khadija MOUGHAMIR (FST Fès)
- Pr. Hamid TOUFIK (FPT Taza)

Stage effectué au Laboratoire Matériaux, Substances Naturelles, Environnement & Modélisation (LMSNEM)-Faculté Polydisciplinaire de Taza (FPT), Université Sidi Mohamed Ben Abdellah de Fès (USMBA).

ABSTRACT

According to the World Health Organization (WHO), cancer is one of the causes of death in the world with 18.1 million new cases and 9.6 million deaths by cancer in 2018. This requires the identification of new molecules that may become new anti-cancer drugs, but the process of discovering new therapeutic agents is costly and requires a lot of time. For this reason, the pharmaceutical industry uses theoretical modelling methods, including the study of the Quantitative Structure-Activity Relationship, for predicting the biological activity before synthesizing these agents.

The objective of this work is to study the anticancer activity of a series of 30 Terpene derivatives by using two modeling methods: The first is semi-empirical (Austin Model 1 "AM1") and the second is empirical (Density Functional Theory "DFT"). Several molecular descriptors were calculated using Gaussian 09 and ChemBioOffice software's, all descriptors were subjected to statistical studies using Partial Least Squares method (PLS) method, validated by external and internal validations.

The two models that we have chosen are highly significant. Indeed, the established 2D-QSAR model obtained by the DFT method and topological, depends on three descriptors ($R^2 = 0.8$; $Q_2 = 0.73$ and $R^2_{\text{pred}} = 0.81$), while the AM1 method and topological has giving a model with four descriptors ($R^2 = 0.77$; $Q^2 = 0.67$ and $R^2_{\text{pred}} = 0.50$). At the end of this work, we can conclude that the DFT empirical method has given better results than the AM1 method and that the activities determined by the DFT method are closer to those calculated experimentally

Keywords : Quantitative Structure-Activity Relationship (QSAR) ; Austin Model 1 (AM1) ; Density Functional Theory (DFT) ; Descriptors ; Anticancer Activity.

Avant-propos

Ce travail a été réalisé au Laboratoire Matériaux, Substances Naturelles, Environnement & Modélisation (LMSNEM) de la Faculté Polydisciplinaire de Taza (FPT), Université Sidi Mohamed Ben Abdellah de Fès, sous la direction du **Professeur Hamid TOUFIK** & la co-direction du **Professeur Fatima LAMCHOURI**, dans le cadre du **Master Chimie des Molécules Bioactives (CMBA)** de la Faculté des Sciences et Techniques de Fès.

DEDICACE

Je dédie ce travail

À ma mère

En vous, je vois la maman parfaite, toujours prête à se sacrifier pour le bonheur de ses enfants.

À mon père

Qui m'a indiqué la bonne voie en me rappelant que la volonté fait toujours des grandes personnes.

À mes chers frères M'hammed, Fouad et Younes

Pour m'avoir toujours soutenue dans mes décisions, et pour leur amour et leur tendresse.

À mes amis STITOU Mourad et EL AISSOUQ Abdellah en souvenirs des bons moments passés ensemble.

Remerciements

Avant tout développement sur cette expérience professionnelle, il apparaît opportun de commencer ce rapport par des remerciements, à ceux qui m'ont beaucoup appris au cours de ce stage.

J'exprime mes profonds remerciements et ma vive connaissance à Monsieur **Hamid TOUFIK**, Professeur à la Faculté Polydisciplinaire de Taza, pour avoir encadré et dirigé ce travail avec une grande rigueur scientifique, sa disponibilité, ses conseils et la confiance qu'elle m'a accordée m'ont permis de réaliser ce travail.

J'adresse aussi mes remerciements au Professeur **Fatima LAMCHOURI**, Directrice du Laboratoire Matériaux, Substances Naturelles, Environnement & Modélisation (LMSNEM) à la Faculté Polydisciplinaire de Taza, qui a co-encadré ce travail de recherche pour ses conseils et son importante contribution à ce travail. Qu'elle trouve ici l'expression de mon profond respect.

Mes remerciements s'adressent aussi à **Mr. Youssef KANDRI RODI**, Professeur à la Faculté des Sciences et Techniques de Fès, pour avoir accepté d'encadrer ce travail et pour son soutien.

J'exprime mes plus sincères remerciements au Professeur **Fouad OUAZZANI CHAHDI**, Coordinateur du Master Chimie des Molécules Bioactives (CMBA), et aux professeurs **EIH. LAMCHARFI** et **Kh. MOUGHAMIR** d'avoir accepté d'être membres du jury.

Un immense merci à mon ami **Mourad STITOU** pour sa patience, son calme bienfaisant et son soutien au cours de ce travail.

La gentillesse, la disponibilité et la chaleur humaine de **EL AISSOUQ Abdellah** m'ont particulièrement touchés, un grand merci mon cher ami.

Liste des abréviations

MM : Modélisation Moléculaire.

QSAR : Quantitative Structure Activity Relationships (Relation Quantitative Structure-Activité).

DFT : Density Functional Theory (Théorie de la Fonctionnelle de la Densité).

AM1 : Austin Model 1.

R² : Coefficient de détermination.

R : Coefficient de corrélation.

RMSE : Root Mean Square Error (Racine de l'erreur moyenne de prédiction).

B3LYP : Becke-Lee-Yang-Parr.

E_{HOMO} : Highest occupied molecular orbital energy (Energie de la plus haute orbitale moléculaire occupée).

E_{LUMO} : Lowest unoccupied molecular orbital energy (Energie de la plus basse orbitale moléculaire inoccupée).

μ : Moment dipolaire.

ACP : Analyse en composantes principales.

PLS : Partial Least Squares (régression des moindres carrés partiels).

RLM : Régression linéaire multiple.

VC : Validation croisée.

DA : Domaine d'applicabilité.

2D : Bidimensionnelle.

3D : Tridimensionnelle.

N : Dimension de la population (échantillon).

P : nombre de descripteurs dans le modèle.

CI50 : Concentration inhibitrice 50.

X_i : Descripteurs moléculaires.

Liste des figures

Figure 1 : Etapes de la cancérogenèse.

Figure 2 : Méthodologie globale de l'étude de la Relation Quantitative Structure-Activité.

Figure 3 : Représentation d'un réseau de neurones artificiels.

Figure 3 : Structure de l'acide ursolique.

Figure 4 : Les différentes couches de réseau de neurones.

Figure 5 : Méthode de la validation interne.

Figure 6 : Structure du squelette carboné de l'acide ursolique.

Figure 7 : Cercle de corrélation des deux méthodes : (a) Méthodes DFT et topologique (b) méthodes AM1 et topologique.

Figure 8 : Relation entre les pIC₅₀ (Obs) et pIC₅₀(Pred) obtenue par la PLS : (a) Méthode DFT et topologique (b) Méthode AM1 et topologique.

Figure 9 : Corrélation entre les activités prédites (pIC₅₀ préd) et observées (pIC₅₀ Obs) obtenue par la validation croisée : (a) Méthode DFT (b) Méthode AM1.

Figure 10 : Corrélation entre les activités prédites (pIC₅₀ préd) et observées (pIC₅₀ Obs) obtenue par la validation externe : (a) Méthode DFT (b) Méthode AM1.

Figure 11 : Diagramme du domaine d'applicabilité : (a) Méthode AM1, (b) Méthode DFT

Liste des tableaux

Tableau 1 : Différentes classes des descripteurs moléculaires.

Tableau 2 : Structure moléculaire des dérivées de l'acide ursolique avec leurs activités pIC₅₀.

Tableau 3 : Descripteurs obtenus par les logiciels d'études.

Tableau 4 : Matrice de corrélation des descripteurs moléculaires calculés par les méthodes AM1 et topologique.

Tableau 5 : Matrice de corrélation des descripteurs moléculaires calculés par les méthodes DFT et topologique.

Tableau 6 : Test de Student des descripteurs sélectionnés dans les deux modèles PLS retenus.

Tableau 7 : Valeurs prédites et observées et leurs résiduels pour les deux modèles sélectionnés.

Tableau 8 : Test de randomisation pour les deux modèles 2D-QSAR sélectionnés.

Sommaire

INTRODUCTION GENERALE	1
Chapitre I : Le cancer.....	3
Introduction.....	3
I. Définition du cancer.....	3
II. Les différents types du cancer.....	3
III. Les causes du cancer.....	4
IV. Physiopathologie du cancer.....	4
V. Les composés naturels utilisés dans le traitement et la prévention du cancer	5
V.1. La prévention	5
V.2. Le traitement	6
Conclusion	7
Chapitre 2 : Les composés terpéniques.....	8
Introduction.....	8
I. Terme et signification	8
II. Structure générale et classification.....	8
III. Fonctions des terpènes	9
III.1. Fonction métabolique primaire	9
III.2. Fonction métabolique secondaire.....	9
IV. Les activités biologiques des composés terpéniques	10
Conclusion	10
Chapitre 3 : Modélisation Moléculaire (MM) et Relations Quantitatives Structure-Activité (2D-QSAR).....	11
Introduction.....	11
I. Modélisation moléculaire.....	11
I.1. Généralités	11
I.2. Méthode de la modélisation moléculaire	11
II. Relations Quantitatives Structure Activité (QSAR).....	14
II.1. Définition et principe.....	14
II.2. Stratégie globale	14
II.3. Descripteurs moléculaires.....	15
II.4. Outils en méthodologie 2D-QSAR.....	17
II.5. Coefficients et tests statistiques standards	19
Conclusion	22
Chapitre 4 : Etude 2D-QSAR de l'activité anticancéreuse de l'acide ursolique et ses dérivées.....	24
Introduction.....	24
I. Matériel et méthodes.....	25
I.1. Présentation de la série de molécules étudiées.....	25
II-2- Calcul des descripteurs	29
II-3 Analyses statistique.....	29
II. Résultats et discussion	29
III-1 Analyse par la régression aux moindres carrés partiels (PLS).....	29
III-1-1 Sélection des descripteurs.....	30
III-2 Modèles 2D-QSAR élaborés.....	34
III-3 Pouvoirs prédictifs des modèles.....	37
III.2. Test de Randomisation	38
III.3. Validation externe	40
III-4 Domaine d'applicabilité (DA)	41
Conclusion	42
Conclusion générale et perspectives	43
bibliographiques.....	44

INTRODUCTION GENERALE

Le cancer est considéré comme la deuxième cause de mortalité dans le monde, ce qui pousse les chercheurs à développer de nouveaux traitements. La chimiothérapie est l'une des méthodes de traitement du cancer qui utilise des substances chimiques pour traiter les différents types de cancer. Néanmoins, les médicaments ne parviennent pas à détruire les cellules cancéreuses, d'où la nécessité de rechercher de nouvelles molécules anticancéreuses.

Les substances naturelles comme les alcaloïdes, les flavonoïdes et les terpènes sont des sources importantes qui méritent d'être explorées pour identifier de nouveaux agents thérapeutiques.

La découverte de nouveaux médicaments est un processus long et coûteux, c'est pour cette raison que l'industrie pharmaceutique explore de nouvelles techniques et méthodes pour optimiser leurs mises au point. La modélisation moléculaire (MM) dont les études des Relations Quantitatives Structure-Activité (QSAR) visent à répondre à ce besoin.

Dans ce travail, nous nous sommes intéressés à l'étude de l'activité anticancéreuse de molécules dérivées de l'acide ursolique en utilisant l'analyse 2D-QSAR avec deux méthodes, l'une semi empirique AM1 et l'autre empirique DFT et ce dans l'objectif de prédire les activités anticancéreuses de cette série de composés et de comparer les résultats obtenus par les deux méthodes.

Notre mémoire comporte deux parties :

La première est consacrée à une étude bibliographique où le premier chapitre présente des généralités sur le cancer, ses différents types et ses causes. Le deuxième chapitre décrit tout ce qui concerne les composés terpéniques, leurs structures générales et leurs activités biologiques. Le troisième chapitre présente des généralités sur la modélisation moléculaire, les Relations Quantitatives Structure-Activité et les méthodes d'analyse statistiques.

La deuxième partie est dédiée à la partie expérimentale où nous présentons les différentes méthodes de Relation Quantitative Structure-Activité employées, ainsi que les résultats obtenus, et nous terminons par une conclusion générale et perspectives sur les futurs travaux.

Partie

bibliographique

Chapitre 1 : Le cancer

Introduction

Le cancer est le résultat d'un dysfonctionnement des cellules ayant échappées au processus de la mort cellulaire, ayant par conséquence la formation d'une masse cellulaire qui occupe la place des autres cellules et subissent de nombreuses transformations pouvant durer plusieurs années. Ainsi, la cancérogénèse est un processus complexe causé par différentes origines nommées : facteurs de risques qui favorisent la transformation des cellules saines en cellules précancéreuses qui se développent pour atteindre un stade précoce du cancer [1].

I. Définition du cancer

Le cancer est une anomalie due à la multiplication intense et non contrôlée d'une ou plusieurs cellules anormales. Cette prolifération est dite maligne, et peut rester localisée (tumeur) ou propagée dans d'autres sites, ce qu'on appelle la métastase.

Le cancer est l'une des maladies générées par une mutation ou une modification dans la structure de l'ADN. Dans ce cas, la cellule perd sa forme spécifique et cesse de réagir aux signaux extérieurs, notamment ceux d'inhibition de la croissance, ce qui mène la cellule ainsi que ses descendants à se multiplier indéfiniment.

II. Les différents types du cancer

Le corps humain est constitué de nombreuses cellules regroupées pour former les tissus et les organes. Vu une panoplie de facteurs, chaque type de cellules peut être l'origine d'un cancer particulier. A cet effet, il existe de nombreux cancers dont tous les organes et les systèmes peuvent être touchés.

Parmi ceux les plus répandus, nous pouvons citer : le cancer du poumon, du sein, de la prostate, de la thyroïde, de l'utérus, du rein, du foie, du pancréas, des os, de la peau, de l'estomac, les leucémies... et d'autres types selon les organes touchés.

III. Les causes du cancer

Le processus de la transformation d'une cellule normale à une cellule cancéreuse est long et parfois trop complexe. Différents facteurs et mécanismes s'accommodent afin de générer des mutations dans les gènes, parmi ces éléments de risque, nous citons :

- ✓ Les radiations : certains cancers sont beaucoup plus répandus chez les personnes exposées aux radiations atomiques (les japonais lors des bombardements d'Hiroshima et de Nagasaki) ou aux rayons X (radiologues avant la mise au point des méthodes modernes de protection) ;
- ✓ Les produits chimiques : plusieurs substances sont considérées comme cancérogènes à savoir : le nickel qui induit les cancers des os et du poumon, l'oxyde de fer qui induit le cancer broncho-pulmonaire ;
- ✓ Les virus : par exemple les virus des hépatites B et C qui favorisent le cancer du foie ;
- ✓ Les bactéries non intestinales : comme *Helicobacter pylori*, qui se multiplie au niveau du mucus gastrique, en provoquant certains cancers de l'estomac [2] ;
- ✓ L'alimentation : plusieurs études scientifiques ont montré l'existence d'une forte corrélation entre la consommation de certains aliments et le développement de certains types de cancer. La fréquence d'un type de cancer diffère d'un pays à un autre, selon le mode de vie et les habitudes alimentaires. Ces variables génèrent un excès de polyamines alimentaires qui favorisent la multiplication des cellules malignes. Parmi ces aliments qui augmentent le risque du cancer, on trouve : les viandes, les graisses saturées d'origine animale, les sucres en excès, les aliments conservés, le sel...etc.

IV. Physiopathologie du cancer

Le développement des tumeurs est spécifique. Il se diffère, des cellules normales, par leurs croissances et leurs caractéristiques morphologiques. L'évolution de ces cellules malignes est provoquée par l'accumulation des gènes responsables du contrôle de la prolifération cellulaire, de la mort cellulaire et du maintien de l'intégrité génétique. Ce processus s'effectue en trois étapes principales :

- ✓ L'initiation, où se produit une lésion au niveau de l'ADN ;
- ✓ La promotion qui conduit à une prolifération intense en formant un groupe de cellules identiques ;
- ✓ La progression où ces cellules prennent un caractère cancéreux et commencent à se développer d'une façon anarchique localement ou par métastase.

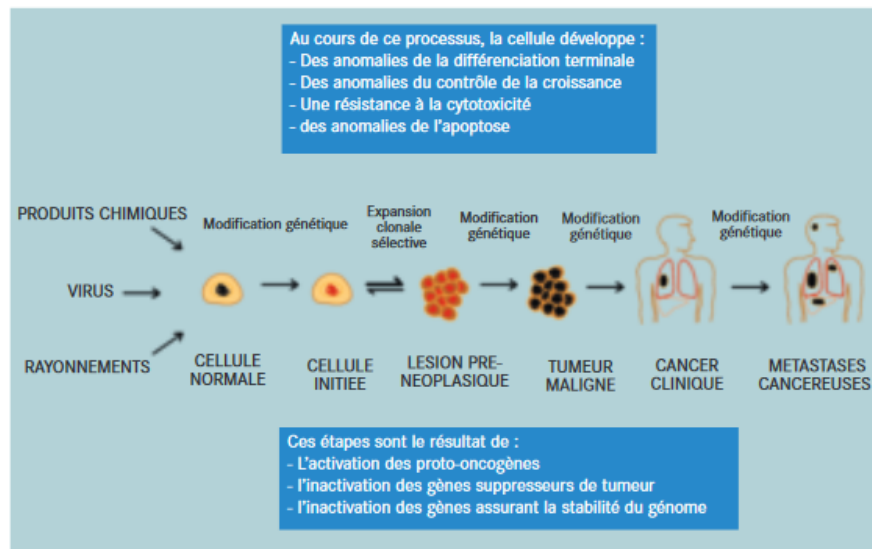


Figure 1 : Les étapes de la cancérogenèse [3]

L'accumulation des cellules cancéreuses forment des lignées cellulaires ; c'est une population homogène de cellules [4]. Les années 50 ont été marquées par l'utilisation des premières lignées de cellules cancéreuses humaines par un chercheur de l'hôpital John Hopkins de Baltimore, de la célèbre lignée HeLa à partir d'un échantillon tumoral d'Henrietta Lacks, extrait à partir d'une patiente atteinte du cancer du col de l'utérus. Depuis ce temps, nombreux sont les laboratoires qui ont pris en charge l'isolation d'un très grand nombre de lignées d'origine tissulaire. Ces dernières sont toujours cultivées, afin de servir aux chercheurs comme matériel de choix permettant d'accélérer la découverte de nouveaux agents anticancéreux ou de mécanismes de signalisation cellulaire régulant la croissance tumorale [4].

V. Les composés naturels utilisés dans le traitement et la prévention du cancer

V.1. La prévention

De nos jours, vu la pollution, la rareté des activités physiques et les régimes alimentaires non-appropriés, les populations sont de plus en plus vulnérables devant différentes maladies comme le cancer. Ce syndrome est parmi les principales causes de morbidité et de mortalité dans le monde. Il représente le deuxième facteur de décès dans le monde avec un taux de 9,6 millions de morts en 2018 [5]. D'après des études récentes environ un tiers des décès par cancer sont liés aux facteurs de risques alimentaires et

comportementaux [5], à savoir une faible consommation des produits naturels corrèle positivement avec l'apparition et le développement du cancer. De nombreuses substances naturelles ont été décrites comme étant capables d'interagir sur les processus de développement de la maladie ou de métastases. Ces substances se retrouvent naturellement au sein même de notre alimentation, notamment dans les fruits, les légumes, les huiles végétales, les huiles de poisson, les céréales, les épices, et également dans certains végétaux ou produits d'origine marine. Ces substances sont considérées comme étant chimio préventives.

Plusieurs métabolites et agents naturels doués de pouvoir chimio-préventif sont largement distribués dans la nature et principalement dans l'alimentation. Nous pouvons notamment citer les flavonoïdes (carottes), les isothiocyanates (choux, brocolis, navets...), les lycophènes (tomates), les indoles, les organo-sulfures (ail, oignons, champignons) et les polyphénols (curcuma) et le β -carotène trouvé dans les carottes et les potirons. Ils sont considérés comme des antioxydants naturels capables de protéger la membrane et l'ADN des cellules contre les dommages oxydatifs par piégeage des radicaux libres.

V.2. Le traitement

Jadis, l'Homme a développé une vie centrée sur les produits naturels en utilisant la matière végétale non seulement pour la nutrition, mais aussi pour des fins curatives. Ces substances d'origine végétale sont reconnues comme étant une base de la majorité des médicaments depuis l'antiquité. Actuellement, dans le cadre du traitement par chimiothérapie plusieurs composés naturels sont reconnus pour leurs activités biologiques importantes, notamment l'activité anticancéreuse. Parmi les médicaments anticancéreux d'origine végétale commercialisés sur le marché, on trouve à titre d'exemple : Le Taxol® (paclitaxel), un anticancéreux extrait de l'écorce d'if du pacifique (*Taxus brevifolia*) et de l'écorce d'if du Canada (*Taxus canadensis*), le Velban® et l'Oncovin® dont les molécules actives sont respectivement la (vinblastine) et (vincristine), extraites à partir de pervenche de Madagascar (*catharanthus roseus*) [6].

Il existe aussi des plantes agissant directement sur le cancer. Leurs extraits constituent des compléments aux traitements conventionnels, citons par exemples : les champignons, l'ail, le jus de grenade, le thé vert, la propolis, les produits riches en fibres, les huiles végétales et les huiles de poisson [7].

En plus de la chimiothérapie, il existe d'autres méthodes de traitement du cancer dont la chirurgie, la radiothérapie, l'hormonothérapie, l'immunothérapie.

Conclusion

Le cancer est l'une des principales causes de mortalité dans le monde entier dont le traitement repose principalement, soit sur la chirurgie et la radiothérapie ayant des traitements locaux, soit sur l'hormonothérapie et la chimiothérapie agissant sur l'ensemble de l'organisme. En revanche, ce type de traitement (la chimiothérapie) présente plusieurs effets secondaires nuisibles, ce qui fait aujourd'hui de la phytothérapie un atout considérable dans le traitement de cette maladie en utilisant souvent des terpènes considérés comme des agents anticancéreux.

Chapitre 2 : Les composés terpéniques

Introduction

Les plantes constituent un réservoir immense et non limité en molécules bioactifs douées d'activités biologiques, utilisées dans plusieurs domaines, telles que l'industrie cosmétique, agroalimentaire et pharmaceutique.

Actuellement, vu l'intérêt croissant de l'utilisation des composés d'origine naturelle, le marché des plantes aromatiques et médicinales (PAM) a connu un développement remarquable. L'augmentation de la production de certaines plantes est liée particulièrement aux propriétés biologiques de leurs huiles essentielles (HE) contenant principalement des terpènes.

Plusieurs voies de valorisation peuvent être allouées aux terpènes, à savoir la production des différents anticancéreux à usage humain. Les terpénoïdes constituent la famille des produits naturels ayant des structures chimiques diversifiées, avec plus de 55 000 molécules identifiées [8].

I. Terme et signification

Dans le règne végétal, les terpènes, à côté des polyphénols et alcaloïdes, constituent les trois classes majeures des métabolites secondaires. La classification des composés terpéniques est basée sur le nombre de répétitions de l'unité de base isoprène (à 5 atomes de carbone).

Chez les végétaux, les terpènes sont synthétisés dans toutes les parties de la plante, où ils jouent le rôle de métabolites primaires et secondaires [9].

En raison de leurs nombreuses structures, les terpènes constituent un groupe hétérogène de molécules avec différentes propriétés physiques et chimiques. Ils peuvent être volatiles, semi-volatiles ou non-volatiles, saturées et insaturées, à chaînes droites, à chaînes ramifiées, cycliques ou acycliques, chiraux ou achiraux, portant éventuellement divers groupes fonctionnels oxygénés (ex. alcools, aldéhydes, cétones, esters et éthers) ou contenant de l'azote ou du soufre, soient solubles ou insolubles dans l'eau [10].

II. Structure générale et classification

Les huiles essentielles constituent particulièrement des composés de type terpénique rencontrés soit sous forme cyclique, soit à chaîne ouverte. Les composés terpéniques sont

formés par la condensation de plusieurs unités du 2-méthylbuta-1,3-diène (molécule de base), nommé l'isoprène C_5H_8 .

La formule brute de ces hydrocarbures terpéniques est $(C_5H_x)_n$ dont le x est variable en fonction du degré d'insaturation de la molécule et qui peut prendre des valeurs (1-8) sauf dans les polyterpènes qui peuvent atteindre plus de 100 (exemple : le caoutchouc).

Ainsi, selon le nombre de carbone constituant les molécules de ce groupe, On distingue :

- ✓ Les terpènes simples ou monoterpènes, formés de deux isoprènes, $C_{10}H_{16}$;
- ✓ Les sesquiterpènes, formés de trois isoprènes, $C_{15}H_{24}$;
- ✓ Les diterpènes, formés de quatre isoprènes, $C_{20}H_{32}$;
- ✓ Les triterpènes (six isoprènes) qui, par oxydation, conduisent à de nombreuses résines ;
- ✓ Les tétraterpènes (huit isoprènes) qui conduisent aux caroténoïdes ;
- ✓ Les polyterpènes (n isoprènes) qui comprennent en particulier, le caoutchouc.

III. Fonctions des terpènes

Les terpènes sont la plus large famille des produits naturels dont certains tels que les stérols et les carotènes font partie du métabolisme primaire et responsable de la croissance et du développement cellulaire. Cependant, la majorité des terpènes végétaux sont des métabolites secondaires, qui interviennent dans les processus physiologiques comme la pollinisation et la défense contre les prédateurs [11].

III.1. Fonction métabolique primaire

Certains terpénoïdes sont impliqués dans la croissance, le développement et le métabolisme général [11]. Ils peuvent également jouer un rôle dans la reproduction d'un organisme ou d'une cellule. Ces composés ont généralement une fonction physiologique dans les organismes, c'est-à-dire une fonction intrinsèque.

III.2. Fonction métabolique secondaire

La plupart des terpénoïdes sont classés comme étant des métabolites secondaires qui n'interviennent pas dans les fonctions vitales de la plante telles que la croissance et le développement cellulaire. Néanmoins, ils sont importants car ils remplissent plusieurs rôles vu la diversité structurale et fonctionnelle des composés terpénoïdiques synthétisés. Ces métabolites secondaires sont connus, ou présumés, avoir des fonctions spécialisées associées à des interactions des plantes avec d'autres organismes vivants dans le contexte de la reproduction, la défense ou la symbiose [9].

En outre, parmi les fonctions secondaires des terpénoïdes s'inscrit leur implication dans les systèmes de défense directs et indirects contre les herbivores et les pathogènes. Ils permettent à la plante d'agir envers l'attaque des bactéries, des champignons, des insectes ou les herbivores en produisant des toxines, des antibiotiques ou des répulsifs [10].

IV. Les activités biologiques des composés terpéniques

La diversité des composés terpéniques, du point de vue structurale, explique leurs activités biologiques multiples. Ils peuvent être utilisés sous différentes formes d'huile essentielle, de résines ou de cires, ainsi comme des arômes ou saveurs, parfums, compléments alimentaires sous forme de vitamines et comme pesticides.

La grande diversité des structures et des fonctions des terpénoïdes présente un intérêt majeur dans la recherche de nouvelles applications médicales, particulièrement dans la chimio-prévention et la chimiothérapie. En outre, ils sont doués des propriétés antimicrobiennes, antifongiques, antiparasitaires, antivirales, antioxydantes, anti-allergènes, et anti-inflammatoires. Les terpénoïdes, sous forme de molécules signaux dégagés dans l'air, jouent des rôles capitaux dans la protection de la plante, contre les attaques des pathogènes et des herbivores, au sein du règne végétal. Les terpénoïdes sont aussi impliqués dans les mécanismes de division, de formation des graines et des fruits, ainsi que la pigmentation et la coloration des végétaux.

Conclusion

Les terpènes sont, vu leurs bons indices thérapeutiques et leurs tolérances sans effets secondaires, largement utilisés dans l'industrie, la parfumerie, les additifs alimentaires et la médecine traditionnelle, comme étant soit des antibiotiques, des anti-inflammatoires, des antioxydants et une source pour développer de nouveaux agents anticancéreux.

Ainsi, l'acide ursolique, au vu de ses propriétés thérapeutiques, est l'un des composés terpéniques les plus bénéfiques. En effet, il est propice dans le traitement du diabète en raison de sa capacité à améliorer les effets de l'insuline sur l'absorption du sucre. Aussi, certaines études ont montré que l'acide précité a des bienfaits pour le cœur, peut aider aussi à réduire les graisses accumulées dans le corps et prévenir l'obésité.

Egalement, L'acide ursolique, naturellement abondants dans de nombreuses plantes alimentaires, telles que les épices, les fruits, les pommes, les canneberges, la lavande, le thym et les pruneaux, présente un effet anticancéreux [12] en raison de sa tendance à inhiber la croissance des cellules cancéreuses, ainsi ce composé et ses dérivés est l'objet de notre étude.

Chapitre 3 : Modélisation Moléculaire (MM) et Relations Quantitatives Structure-Activité (2D-QSAR)

Introduction

L'expérience est un moyen direct pour obtenir des données de l'activité des composés organiques, mais elle coûte chère et nécessite beaucoup de temps. Il est donc crucial d'utiliser des méthodes théoriques avant de passer à l'expérience pour prédire les activités des composés.

Le développement important de l'informatique, ainsi que des études théoriques de la chimie quantique permettent aux chercheurs d'obtenir des paramètres physico-chimiques et quantiques des composés en un temps plus court.

Les paramètres structuraux ainsi que l'introduction des études de relation quantitative structure-activité QSAR peuvent interpréter et prédire l'activité pour de nouveaux composés organiques. Ceci est utilisé pour examiner la relation entre des descripteurs moléculaires d'un ensemble de composés et leurs activités biologiques ou propriétés physico-chimiques.

I. Modélisation moléculaire

I.1. Généralités

La recherche et la synthèse de nouveaux médicaments et de nouveaux composés chimiques est associée à une étude de modélisation moléculaire, cette méthode permet de représenter les propriétés d'une molécule et de manipuler les modèles des structures en deux ou trois dimensions.

La modélisation moléculaire est capable de prédire, confirmer l'expérience en utilisant des moyens informatiques et des méthodes statistiques. Elle est généralement utilisée dans plusieurs domaines. Cette méthode peut être classée en trois catégories : les méthodes quantiques, la mécanique moléculaire et la dynamique moléculaire.

I.2. Méthode de la modélisation moléculaire

I.2.1. Mécanique moléculaire

La mécanique moléculaire (MM) s'appuie sur le champ de force, elle représente les interactions liantes et non liantes à l'intérieur d'une molécule ou entre plusieurs molécules, en

considérant les atomes comme des sphères incompressibles portant des charges partielles fixes [18].

I.2.2. Dynamique moléculaire

La dynamique moléculaire (DM) a pour objectif d'étudier l'évolution du système en fonction du temps en s'appuyant sur les mouvements intra et intermoléculaires [13].

I.2.3. Méthode quantique

Les méthodes de la mécanique quantique (MQ) s'appuient sur la résolution de l'équation de Schrödinger, elles permettent la détermination de la configuration électronique de la molécule, sa structure tridimensionnelle, son énergie et ses propriétés physico-chimiques. Ces méthodes nécessitent des temps de calcul souvent élevés.

La mécanique quantique contient trois méthodes :

I.2.3.1. Méthode ab-initio :

Ce sont des méthodes non empiriques qui traitent toute particule (noyaux et électrons) et aucun paramètre empirique dans le calcul de l'énergie n'est utilisé [14].

Ab-initio est la méthode la plus simple de calcul de structure électronique, dans laquelle la répulsion coulombienne électron-électron n'est pas spécifiquement prise en compte [14].

I.2.3.2. Méthode semi empirique :

Elles sont utilisées pour modéliser les systèmes moléculaires de grande dimension et sont basées sur deux approximations. La première consiste à ne considérer que la couche de valence : les électrons de valence qui interviennent dans les liaisons chimiques et définissent donc les propriétés du système. La seconde annule les intégrales de répulsion électroniques multi-centres [14]. Les principales méthodes semi-empiriques sont les suivantes :

CNDO : Complete Neglect of Differential Overlap : Elle néglige complètement les intégrales des répulsions entre atomes non liés chimiquement [15].

INDO : Intermediate Neglect of Differential Overlap : les approximations introduites sont presque les mêmes que celles de CNDO, sauf par l'estimation des intégrales biélectroniques [15].

MINDO : Modified Intermediate Neglect of Differential Overlap : les intégrales biélectroniques à deux centres sont estimées selon l'approximation d'Ono [15]. Elles constituent une bonne alternative pour le calcul exact des énergies de formation, des potentiels d'ionisation et des distances interatomiques dans la plupart des molécules de dimension moyenne [15].

AM1 : La méthode Austin Model 1 (AM1) est une méthode proposée par Dewar en 1985 [15]. Elle permet d'obtenir de meilleurs résultats par rapport aux autres méthodes semi empiriques [15] et a pour objectif de corriger certains défauts liés à la méthode MNDO qui reproduit des erreurs au niveau des liaisons d'hydrogène et au niveau des énergies d'activation [15].

Aujourd'hui, cette méthode est très utilisée afin d'optimiser la géométrie moléculaire en raison de ses meilleurs résultats qui sont en accord avec ceux expérimentaux. Ceci est dû à plusieurs améliorations telles que celles apportées à l'énergie de répulsion entre les noyaux, ainsi qu'à la différentiation des exposants des orbitales s et ceux de l'orbitale p. La méthode MNDO considère que les deux exposants s et p sont pris égaux.

I.2.3.3. Théorie de la Fonctionnelle de la Densité (DFT) :

C'est une méthode empirique très répandue, et possède encore d'énormes avantages en calcul par rapport aux méthodes ab initio.

Le but des méthodes DFT est de déterminer des fonctionnelles qui permettent de donner une relation entre la densité électronique et l'énergie. La DFT a véritablement débuté avec les théorèmes fondamentaux de Hohenberg et Kohn en 1964, qui établissent une relation fonctionnelle entre l'énergie de l'état fondamental et sa densité électronique. Les deux théorèmes montrent l'existence d'une fonctionnelle de la densité qui permet de calculer l'énergie de l'état fondamental d'un système [16].

- Le premier théorème de Hohenberg et Kohn : montre que la densité électronique est la seule fonction nécessaire pour obtenir toutes les propriétés électroniques d'un système [16].

- Deuxième théorème de Hohenberg et Kohn : stipule que la fonctionnelle de la densité qui permet d'accéder à l'énergie de l'état fondamental donne la plus basse énergie si la densité est celle de l'état fondamental. Ce théorème se base sur le principe analogue à celui proposé dans l'approche de Hartree-Fock pour une fonctionnelle de la fonction d'onde, mais appliqué cette fois à une fonctionnelle de la densité électronique [16].

- Equations de Kohn-Sham et leur résolution : Ce sont Kohn et Sham, en 1965, qui ont élaboré une méthode pratique pour trouver l'énergie de l'état fondamental à partir de la densité électronique de l'état fondamental, parce que les théorèmes de Hohenberg et Kohn ne donnent pas une procédure pour calculer cette énergie. L'idée de Kohn-Sham est d'utiliser un système fictif de n électrons non interagissant, de même densité électronique que le système réel en interaction [16].

II. Relations Quantitatives Structure Activité (QSAR)

II.1. Définition et principe

La méthode 2D-QSAR est basée sur l'hypothèse que l'activité d'un composé chimique est liée à sa structure, cette méthode se rapporte à la construction de modèles prédictifs d'activités biologiques, en fonction de l'information de structure moléculaire d'une série de composés.

C'est une méthode qui relie de manière quantitative les propriétés moléculaires aussi bien électroniques que géométriques, appelées descripteurs, avec une activité biologique [17], suivant un modèle mathématique :

$$\text{Activité} = f(\text{descripteurs moléculaires})$$

II.2. Stratégie globale

Un modèle 2D-QSAR relie la structure des molécules à une activité donnée, la stratégie globale suit les étapes suivantes :

Tout d'abord, il faut construire la structure des molécules par le traçage de leurs squelettes moléculaires en utilisant des logiciels : ChemDraw, ChemSketch, MarvinSketch. Deuxièmement, la présentation de l'activité des molécules à partir des données expérimentales de la littérature. Par la suite, il faut sélectionner les descripteurs moléculaires en utilisant plusieurs logiciels, par exemple : Gaussian, Chemoffice, ChemSketch, MarvinSketch, MOE... Puis, Il faut passer à la construction des modèles à l'aide des méthodes statistiques, et finalement valider ces modèles en utilisant un ensemble de test et appliquer des validations interne et externe.

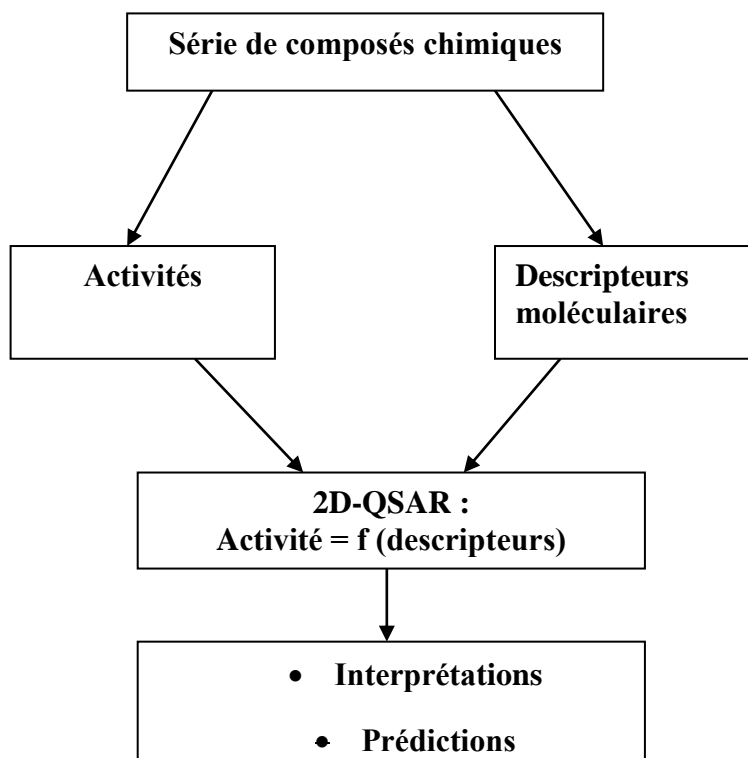


Figure 2 : Méthodologie globale de l'étude de la Relation Quantitative Structure-Activité (QSAR)

II.3. Descripteurs moléculaires

Les descripteurs moléculaires ont pour but de décrire de manière numérique la structure d'une molécule afin d'établir une relation entre les descripteurs et l'activité biologique [18]. Deux grands schémas pour la classification des descripteurs moléculaires ont été établis : l'un en fonction de leurs origines (constitutionnel, topologique, géométrique, quantique, thermodynamique...) et l'autre selon leurs dimensionnalités (1D, 2D, 3D).

Tableau1 : Différentes classes de descripteurs moléculaires

Descripteurs	Types	Exemples
<p>1D : sont calculés à partir de la formule brute de la molécule et représentent des propriétés générales du composé.</p>		<ul style="list-style-type: none"> • Nombre d'atomes • Masse molaire • Poids moléculaire (MW)
<p>2D : sont obtenus à partir de la structure plane de la molécule.</p>	<p>Descripteurs constitutionnels</p>	<ul style="list-style-type: none"> • Nombre total de liaisons • Nombre de types de liaisons • Nombre de cycles • Poids atomique moyen
	<p>Descripteurs topologiques</p>	<ul style="list-style-type: none"> • Indices de Wiener (IW) • Indice de Balaban (IB) • Connectivité Totale de Valence (TVcon) • Connectivité Totale (Tcon)
<p>3D : sont calculés à partir des positions des atomes dans l'espace, leurs calculs nécessitent de connaître la géométrie 3D de la molécule. Ces descripteurs sont coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent de la structure 3D.</p>	<p>Descripteurs électroniques/quantiques</p>	<ul style="list-style-type: none"> • Electronégativité (χ) • Moment dipolaire (μ) • Energie HOMO (E_{homo}) • Energie LUMO (E_{lumo}) • Energie moléculaire
	<p>Descripteurs géométriques</p>	<ul style="list-style-type: none"> • Nombre de liaisons rotatives (Nrot) • Volume de van der waals (VVDW) • Volume moléculaire (MV)
	<p>Descripteurs thermodynamiques</p>	<ul style="list-style-type: none"> • Constante d'Henry • Température critique • Energie de gibbs
	<p>Descripteurs physico-chimiques</p>	<ul style="list-style-type: none"> • Réfractivité molaire (MR) • Nombre de donneurs de liaisons Hydrogène (NHD) • Nombre d'accepteurs de liaisons Hydrogène (NHA)

Le modèle 2D-QSAR doit avoir le minimum de descripteurs afin d'expliquer les activités/propriétés ciblées, c'est pour cela qu'il est nécessaire d'effectuer une analyse statistique des données.

II.4. Outils en méthodologie 2D-QSAR

II.4.1. Analyse statistique

L'objectif de l'analyse statistique est d'identifier les descripteurs qui sont corrélés à l'activité, ainsi d'identifier les descripteurs qui sont corrélés entre eux pour ne garder que les principaux. L'activité étudiée est appelée variable dépendante ou variable Y , c'est la caractéristique que l'on cherche à optimiser et à expliquer. Les descripteurs, sont appelés variables indépendantes ou variables X . Le modèle statistique est une équation qui donne la valeur de l'activité en fonction de la somme des valeurs des descripteurs.

Les principaux outils statistiques employés pour obtenir un modèle QSAR sont :

- L'analyse en composantes principales (ACP) : est une méthode qui permet la réduction d'un nombre important de descripteurs en un nombre plus faible, cette méthode repose sur l'étude de la matrice de corrélation en éliminant ceux qui présentent une corrélation importante, c'est une étape très utile avant d'effectuer une régression linéaire multiple ;
- La régression des moindres carrés partiels (PLS) : est une méthode de réduction des descripteurs, qui est utilisée lorsque le nombre de variables est élevé [19] ;
- La régression linéaire multiple (RLM) : Cette méthode repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante Y (l'activité) et une série de n variables indépendantes X_i (les descripteurs moléculaires). L'objectif de cette méthode est d'obtenir une équation mathématique de la forme :

$$Y = a_0 + a_1X_1 + a_2X_2 + \dots + a_nX_n$$

Avec :

Y : l'activité.

a_i ($i = 0, 1, \dots, n$) : les coefficients de la régression.

X_i ($i = 1, 2, \dots, n$) : les descripteurs.

- Le réseau de neurones artificiels (RNA) : est un réseau neuronal qui s'inspire du fonctionnement des neurones biologiques et prend corps dans un ordinateur sous forme d'un algorithme. Le réseau neuronal peut se modifier lui-même en fonction des

résultats de ses actions, ce qui permet l'apprentissage et la résolution des problèmes sans algorithme, donc sans programmation classique [20] ;

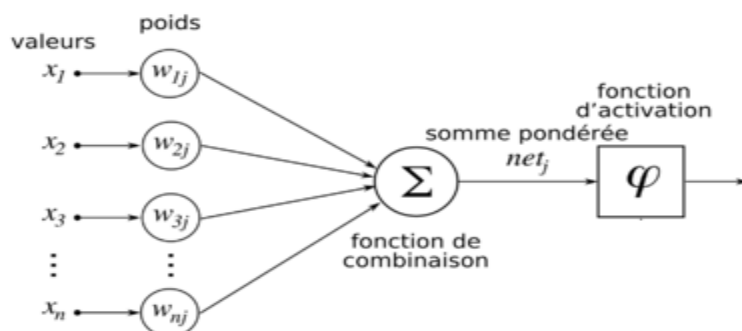


Figure 3 : Représentation d'un réseau de neurones artificiels.

X_k (où $k = 1, 2, \dots, n$) : descripteurs

W_k (où $k = 0, 1, 2, \dots, n$) : les poids synaptiques

Le réseau de neurone réalise une ou plusieurs fonctions algébriques de ses entrées, par composition des fonctions réalisées par chacun de ses neurones, il s'agit donc d'un ensemble de neurones connectés entre eux, l'information circulant des entrées vers les sorties sans retour en arrière possible. On parle souvent de perceptron multicouche à cause de la présence de neurones cachés.

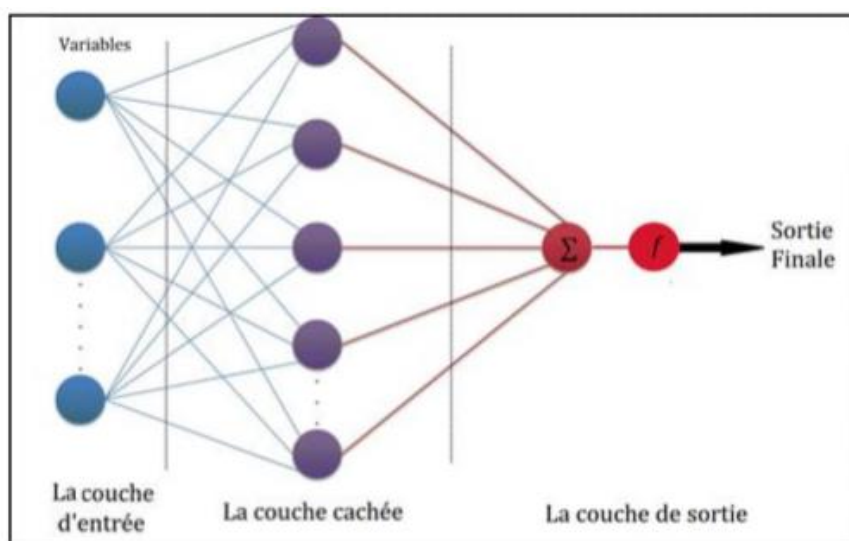


Figure 4 : Les différentes couches de réseau de neurones.

- Régression Non Linéaire Multiple (RNLM) : est une méthode non linéaire (exponentielle, logarithmique, polynomiale, ...) qui permet de déterminer le modèle mathématique d'une activité y en fonction des descripteurs moléculaires. Les descripteurs proposés par le modèle linéaire qui seront élevés à la puissance 2 selon l'équation suivante [21] :

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

Avec : **Y** est la variable dépendante (à expliquer ou à prédire) ;

X_i sont les variables indépendantes (explicatives) ;

i est le nombre de variables explicatives ;

a₀ est la constante de l'équation du modèle ;

a_i et **b_i** sont les coefficients de descripteurs dans l'équation du modèle.

II.5. Coefficients et tests statistiques standards

La qualité du modèle peut être interprétée par la détermination de plusieurs coefficients :

II.5.1. Coefficient de corrélation R

Ce coefficient détermine la variance de l'activité cible qui est expliquée par le modèle 2D-QSAR, c'est à dire par la régression de l'activité cible en fonction de l'activité initiale [19].

- Une bonne corrélation entre l'activité cible et l'activité initiale est observée lorsque R est plus proche de 1.

- Une corrélation non linéaire entre l'activité cible et l'activité initiale est obtenue quand R est proche de 0.

$$r = \sqrt{1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}}$$

Avec : R est le coefficient de corrélation ; y_i et \hat{y}_i sont respectivement les valeurs observées et calculées de la variable dépendante ; \bar{y} est la valeur moyenne des valeurs observées.

II.5.2. Coefficient de détermination R²

Ce coefficient de détermination R² donne le taux d'explication ou pourcentage de la variation de Y expliquée par la variation de X ; par exemple si R² = 0,80 ce chiffre signifie que 80% de la variable Y est attribuable à la variation de la variable X [19].

Plus la valeur de R² est proche de 1, plus les valeurs des activités calculées et observées sont corrélées.

Plus la valeur de R² est proche de 0, plus les valeurs des activités calculées et observées ne sont pas corrélées.

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

Où :

y : La valeur expérimentale de l'activité.

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

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

II.5.3. Racine de l'erreur moyenne de prédiction RMSE

La précision du modèle est également évaluée par un coefficient dit racine de l'erreur moyenne de prédiction (Erreur Root Mean Square : RMSE).

Plus la valeur de RMSE est proche de 0, plus la qualité du modèle est meilleure [19].

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (p_{\text{exp}} - p_{\text{pred}})^2}$$

II.5.4. Test de Fisher

L'indice de Fisher F est couramment employé pour mesurer le niveau de signifiante statistique du modèle, c'est-à-dire la qualité du choix des paramètres. Après le calcul de F (observé) on le compare avec le F théorique obtenu à partir des tables statistiques usuelles (la table de Fisher).

Si F observé est plus grand que le F théorique : refus de l'hypothèse nulle H_0 (les variances des échantillons sont homogènes) et cela signifie que les variances des échantillons sont trop différentes pour être considérées comme homogènes [22].

Si F observé est plus petit que le F théorique : acceptation de l'hypothèse nulle H_1 (les variances des échantillons ne sont pas homogènes) et cela signifie que les deux variances ont des valeurs suffisamment proches pour qu'on accepte l'idée qu'elles soient homogènes.

II.5.5. Test de student

L'indice de Student (le t-test de Student) est employé afin de tester les descripteurs comme significatifs ou non. Il est employé aussi pour évaluer la significativité du modèle complet. Le test s'écrit : $H_0 : r = 0$ et $H_1 : r \neq 0$

Si le coefficient de corrélation est différent de zéro, on rejette l'hypothèse H_0 (l'hypothèse nulle) et on accepte H_1 donc le modèle est significatif [22].

II.5.6. La validation interne

- La validation croisée

La validation croisée « cross-validation » est l'une des méthodes les plus employées pour déterminer la stabilité du modèle prédictif et de tester l'influence de chaque échantillon sur le modèle final.

Ce processus consiste à extraire un certain nombre n de molécules du jeu initial à k molécules et à construire un nouveau modèle avec les $n-k$ molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent). Ce nouveau modèle est alors utilisé pour la phase de prédiction sur les n molécules retirées. Ce processus est ensuite réitéré pour retirer et prédire les valeurs de toutes les molécules du jeu d'entraînement [23].

Plus la valeur du coefficient de corrélation q^2 entre les activités ainsi calculées et les activités observées se rapproche de 1, plus le pouvoir de prédiction sera meilleur. Pour que le modèle soit acceptable, le pouvoir de prédiction interne doit être supérieur à 0,5 [24].

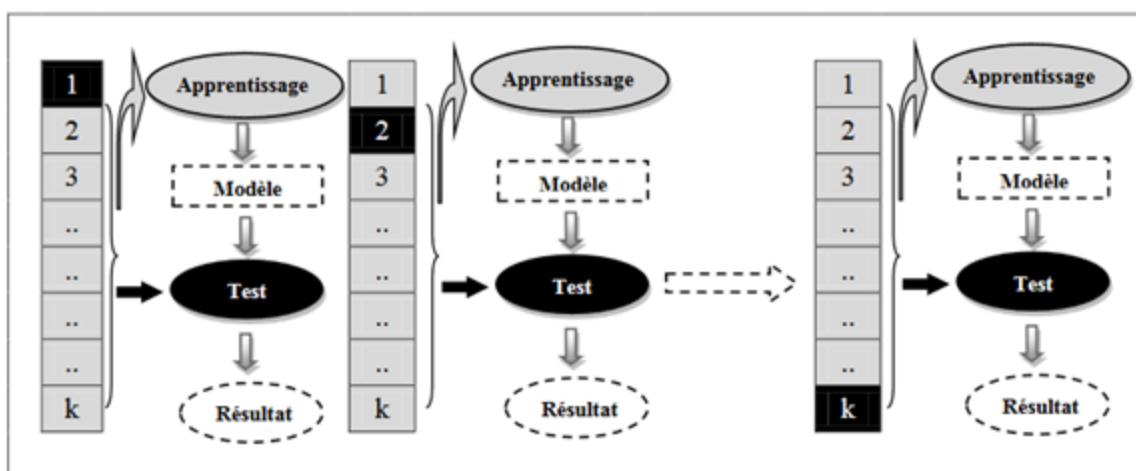


Figure 5 : méthode de la validation interne

- Test de randomisation

La robustesse du modèle peut être analysée par un test de randomisation Y . En cas de randomisation de l'activité (Y), les entrées de la colonne Y sont randomisées aléatoirement et de nouveaux modèles QSAR sont développés sans perturber la disposition de la matrice de descripteurs d'origine. Pour qu'un modèle QSAR soit acceptable, les coefficients de détermination R^2 et de la validation interne Q^2 des modèles randomisés doivent être inférieurs aux coefficients de détermination du modèle non-randomisé [23]. Ce paramètre garantit que les modèles ainsi développés ne sont pas obtenus par hasard.

II.5.7. La validation externe

Afin de tester de manière fiable le pouvoir prédictif, l'utilisation d'un jeu de validation externe est nécessaire. Pour que le jeu de données initial soit suffisamment large, ce dernier peut être aisément divisé en deux : un jeu d'entraînement sur lequel le modèle est développé et un jeu de validation utilisé pour déterminer son pouvoir prédictif [25].

II.5.8. Domaine d'applicabilité (DA)

Le domaine d'applicabilité des modèles 2D-QSAR joue un rôle critique dans l'estimation de l'incertitude de prédiction d'une molécule en fonction de sa similitude avec les composés utilisés dans l'ensemble d'apprentissage. Ainsi, la prédiction d'une activité à l'aide du modèle 2D-QSAR n'est valable que si le composé à prédire se situe dans le DA, car il est impossible de prédire l'activité de tous les composés chimiques en utilisant un modèle 2D-QSAR unique. Bien qu'il existe plusieurs façons de présenter un DA, le graphe de William est considéré comme l'une des méthodes les plus courantes basées sur la distance de levier h [25], [26]. Cette dernière est calculée à tous les composés i situés dans l'ensemble d'apprentissage à partir de la formule suivante :

$$h_i = x_i(X^T * X)^{-1} * x_i^T \quad (i = 1. \dots \dots \dots n)$$

La valeur du levier critique (h^*) est utilisée pour prouver le degré d'extrapolation de la prédiction, généralement déterminée par $h^* = 3(N + 1) / M$. Lorsque $h_i > h^*$, la prédiction du composé i est en dehors du domaine d'applicabilité.

Avec h représente l'effet de levier, N le nombre de descripteurs utilisés dans l'ensemble d'apprentissage et M le nombre de composés dans l'ensemble d'apprentissage.

Conclusion

Actuellement les méthodes QSAR sont employées pour prédire l'activité d'une molécule avant de faire sa synthèse. C'est pour cette raison que les relations quantitatives structures-activités sont devenues de plus en plus utilisées en chimie organique, médicinale et en particulier dans le cadre de la conception de nouvelles entités chimiques utilisées dans les médicaments.

Partie

expérimentale

Chapitre 4 : Etude 2D-QSAR de l'activité anticancéreuse de l'acide ursolique et ses dérivées.

Introduction

Les plantes sont l'une des sources de composés naturels à partir desquelles des principes actifs sont extraits (terpènes, flavonoïdes, alcaloïdes...). Dans ce travail, nous nous sommes intéressés aux terpènes vus leurs intérêts médicaux.

Les méthodes d'étude de la Relation Quantitative Structure Activité (QSAR) sont des outils très employés dans divers domaines de recherche, notamment dans le processus de découverte de nouveaux médicaments.

L'objectif de la présente étude est d'une part établir des modèles 2D-QSAR à l'aide de descripteurs moléculaires qui permettent de trouver une relation entre les descripteurs et les activités biologiques et prédire l'activité antiproliférative d'une série de composés dérivés de l'acide ursolique. Et d'autre part, comparer les résultats obtenus par les méthodes semi-empirique AM1 et empirique DFT.

L'acide ursolique est un triterpénoïde pentacyclique présent dans les plantes médicinales, les légumes et les fruits, et qui a été isolé pour la première fois par le pharmacien Johann TROMMSDORFF [27].

Plusieurs travaux ont été effectués sur l'acide ursolique et ont mis en évidence ses activités biologiques, telles que :

- Activité antibactérienne (Nascimento et al., 2014) ;
- Activité antifongique et antivirale (Liu, 1995) ;
- Activité hépatoprotectrice (Liu, 1995) ;
- Activité anti-inflammatoire (Hirota et al., 1990) ;
- Activité anti-tumorale (Lee et al., 1994) ;
- Activité antiproliférative (Andersson et al., 2003) ;

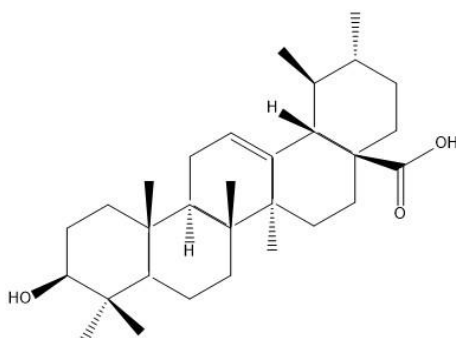


Figure 6 : Structure du squelette carboné de l'acide ursolique

I. Matériel et méthodes

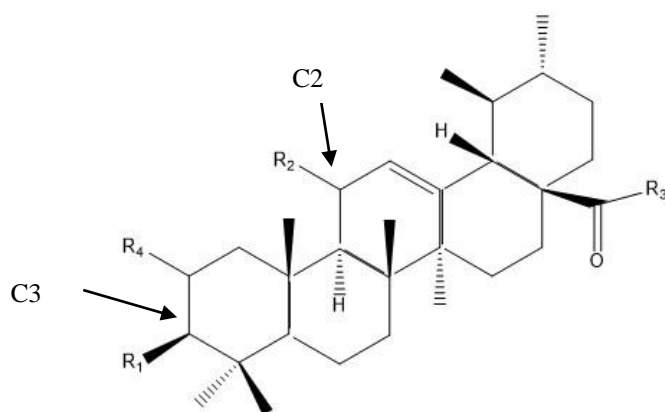
I.1. Présentation de la série de molécules étudiées

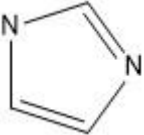
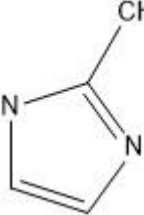
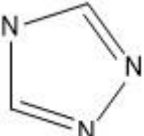
L'étude expérimentale relative à la série étudiée a été menée par Leal et ses collaborateurs en 2012. Nous présentons ici, de présenter le protocole expérimental, les résultats obtenus et les conclusions tirées [28].

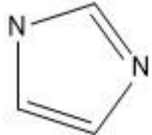
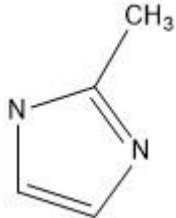
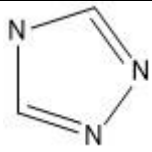
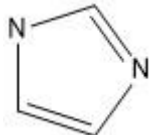
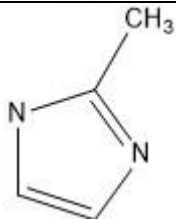
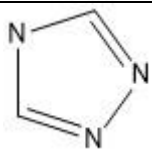
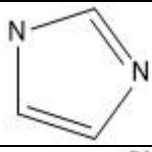
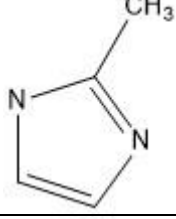
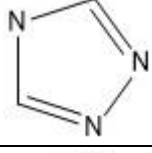
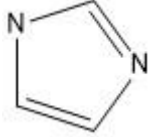
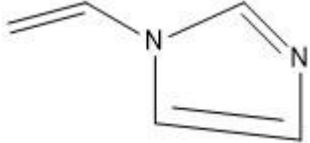
Dans leur travail, trente (30) molécules d'acide ursolique et dérivés sont synthétisés par réaction de l'acide ursolique avec les réactifs hétérocycliques appropriés, afin d'améliorer leurs activités antiprolifératives. La technique colorimétrique du bromure de 3-[4,5-diméthylthiazol-2-yl] - 2,5-diphényltétrazolium (MTT) a été utilisée pour tester leurs activités antiprolifératives vis-à-vis des lignées cancéreuses (AsPC-1, MIA PaCa-2, PANC-1, Hep G2, PC-3, MCF7 et A549). Les résultats de l'étude biologique des molécules de l'acide ursolique et ses dérivés sont rapportées dans le tableau 2. Ils ont montré que la synthèse des nouveaux dérivés de l'acide ursolique par l'introduction de noyaux hétérocycliques d'imidazole, de méthylimidazole ou de triazole sur le squelette de l'acide ursolique a permis d'obtenir des composés ayant des activités antiprolifératives améliorées par rapport à l'acide ursolique.

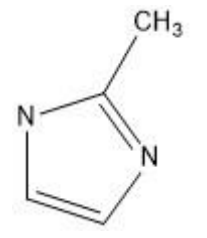
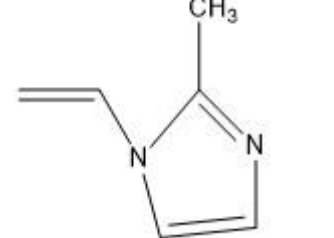
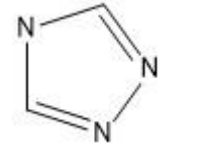
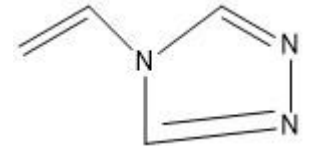
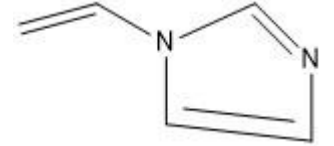
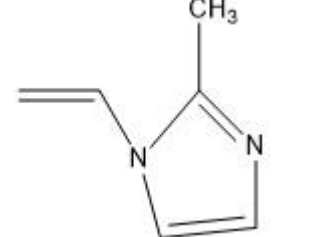
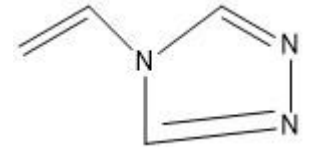
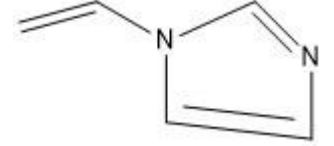
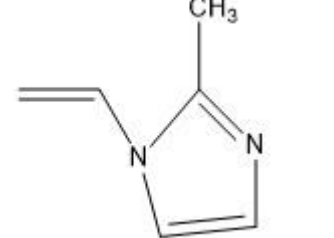
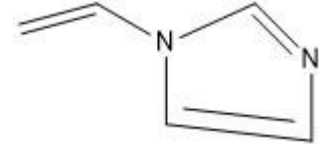
Le tableau 2 rassemble les structures chimiques des 30 composés étudiés et valeurs des activités antiprolifératives à l'échelle logarithmique ($pCI_{50} = \log 1 / CI_{50}$). Avant d'entamer l'étude 2D-QSAR, les molécules ont été réparties au hasard en deux sous-ensembles : un ensemble d'apprentissage qui contient 25 molécules pour construire le modèle 2D-QSAR et un ensemble de test de 5 molécules pour tester le pouvoir prédictif du modèle retenu.

Tableau 2 : Structure moléculaire des dérivées de l'acide ursolique avec leurs activités pIC₅₀



N° de la molécule	R1	R2	R3	R4	pIC ₅₀
1	OH	H	OH	H	4,899
2	AcO	H	OH	H	4,148
3	AcO	$\text{C}=\text{O}$	OH	H	4,474
4	$\text{C}=\text{O}$	H	OH	H	4,610
5	$\text{C}=\text{O}$	$\text{C}=\text{O}$	OH	H	4,260
6	$\text{C}=\text{O}$	H	OH	$\text{CH}_2=\text{CH}-\text{OH}$	4,872
7	OH	H	OMe	H	4,752
8	$\text{C}=\text{O}$	$\text{C}=\text{O}$	OMe	$\text{CH}_2=\text{CH}-\text{OH}$	5,045
9	$\text{C}=\text{O}$	H	OMe	$\text{CH}_2=\text{CH}-\text{OH}$	5,142
10	OAc	H		H	4,818
11	OAc	H		H	4,694
12	OAc	H		H	4,632

13	OAc	=O		H	4,856
14	OAc	=O		H	4,815
15	OAc	=O		H	4,991
16	=O	H		H	4,818
17	=O	H		H	4,707
18	=O	H		H	4,716
19	=O	=O		H	4,931
20	=O	=O		H	5,292
21	=O	=O		H	5,004
22	=O	H			5,292

23	=O	H			4,946
24	=O	H			5,408
25	=O	=O	OMe		5,638
26	=O	=O	OMe		5,677
27	=O	=O	OMe		5,091
28	=O	H	OMe		5,236
29	O	H	OMe		5,677
30	=O	H	OMe		5,721

* CI_{50} (μM) = Concentration inhibitrice nécessaire pour réduire de 50% la densité optique des cellules traitées par rapport aux cellules non traitées [29].

II-2- Calcul des descripteurs

Les structures chimiques des composés ont été dessinées à l'aide du logiciel ChemDraw professional 15 [30]. Les optimisations initiales de la géométrie ont été effectuées avec la méthode de la mécanique moléculaire (MM) en utilisant le champ de forces MM2 implémenté dans le logiciel ChemBioOffice [30]. Les descripteurs topologiques calculés (Voir tableau 3) ont été obtenus en utilisant un ordinateur (Core I5, 6Go de RAM, 1000 Go de disque D et 850 Go de disque C).

Le logiciel Gaussian 09 [31] a été utilisé pour calculer les descripteurs électroniques 3D avec les deux méthodes (AM1 et DFT). Les structures optimisées obtenues par le champ de forces MM2 ont été optimisées par la méthode semi empirique AM1 afin de calculer les descripteurs de réactivité, notamment E_{LUMO} , E_{HOMO} , χ , μ , ω , ... (Voir Tableau 3). Par la suite, la méthode semi empirique ZINDO a été utilisée pour calculer l'énergie de chaque structure optimisée par la méthode AM1, à savoir E (ZINDO), f , ΔE , λ_{max} . De même, la théorie de la fonctionnelle de la densité DFT a été utilisée pour calculer les mêmes descripteurs obtenus par la méthode AM1 en utilisant la fonction hybride à trois paramètres (B3LYP) avec l'ensemble de bases 6-31 G (d), tandis que la méthode de la théorie de la fonctionnelle de la densité dépendante du temps TD-DFT a été utilisée pour calculer les descripteurs énergétiques (Tableau 3). Les différents calculs ont été effectués en utilisant deux ordinateurs : (ordinateur 1 : Core 2 Duo, 1,96 Go de RAM, 195 Go de disque D et 74,2 Go de disque C. ordinateur 2 : Pentium 4, 248 Mo de RAM, 19,2 Go de disque D et 19,0 Go de disque C).

II-3 Analyses statistique

Des modèles 2D-QSAR ont été établis en utilisant la régression des moindres carrés partiels PLS implémentée dans les logiciels de travail XLSTAT [32] et Minitab version 17 [33]. L'activité pCI50 est prise comme variable dépendante, tandis que les variables indépendantes sont les descripteurs moléculaires calculés.

II. Résultats et discussion

III-1 Analyse par la régression aux moindres carrés partiels (PLS)

Nous avons calculé un ensemble de 38 descripteurs basés sur les méthodes quantiques (AM1 et DFT) et topologiques. La méthode des moindres carrés partiels PLS a été utilisée pour choisir à partir des 38 descripteurs initiaux un ensemble de 14 descripteurs. Vu le nombre élevé de descripteurs par rapport au nombre de molécules.

Tableau 3 : Descripteurs obtenus par les logiciels d'étude

ChemBioOffice	Gaussian 09
H accepteur	E_{HOMO} (ev)
H donneur	E_{LUMO} (ev)
Ovalité	Egap
La réfractivité molaire	μ : Potentiel chimique
Coefficient de partition	DM : Moment dipolaire
Log P	χ : Electronegativité
Log S	η : Dureté chimique
Indice de Balaban (IB)	ω : Indice d'électrophilie globale
Indice de Wiener (IW)	ΔE (ev) : Energie d'activation
La surface polaire	λ_{max} : Longueur d'onde du maximum d'absorption
Nombre de liaisons rotatables	f : Force d'oscillation
La Connectivité totale	E (ZINDO) : Energie calculée par la méthode ZINDO
La connectivité totale de valence	E (TD-DFT) : Energie calculée par la méthode TD-DFT
Charge partielle de Mullikan (C1, C2, C3)	
Masse molaire	
Le volume critique	
Le volume accessible au solvant	
La surface moléculaire	
La surface accessible au solvant	
La Constante d'Henry	
Moment d'inertie principal	
L'énergie libre de Gibbs	

III-1-1 Sélection des descripteurs

En raison du problème de colinéarité inhérent au développement du modèle 2D-QSAR, l'analyse des corrélations est indispensable pour sélectionner les descripteurs les plus corrélés avec l'activité cytotoxique.

L'analyse de la corrélation a été réalisée pour les 14 descripteurs sélectionnés à partir d'un nombre initial de 38 descripteurs moléculaires calculés. Les deux axes principaux F1 et F2 obtenus par la méthode DFT représentent respectivement 71,12% et 15,47% de la variance totale et les deux représentent 86,59% de l'information totale, tandis que ceux retenus par la méthode AM1 représentent seulement 74,07% (Figure 6.b). Les matrices de corrélation pour les deux méthodes étudiées sont indiquées dans les tableaux 4 et 5. Les coefficients de corrélation dans la matrice obtenue fournissent les informations sur la haute ou la basse inter-corrélation entre les descripteurs. De plus, afin de réduire la redondance existante dans cette matrice de données, les descripteurs hautement corrélés ($R \geq 0,75$) ont été exclus [34].

En se basant sur les résultats des matrices et les cercles de corrélation, nous avons retenu quatre (4) descripteurs déterminés par les méthodes AM1 et topologique : Egap, C3, H

accepteur et TV connectivité, et trois (3) descripteurs calculés par les méthodes DFT et topologique : Egap, C3 et TV connectivité.

Tableau 4 : Matrice de corrélation des descripteurs moléculaires calculés par les méthodes AM1 et topologique.

Variables	E _{HOMO}	E _{LUMO}	Egap	μ	X	η	ω	f	λ _{max}	E (AM1)	H accep	ovality	C2	C3	PIC50
E _{HOMO}	1.00														
E _{LUMO}	0.21	1.00													
Egap	-0.40	0.81	1.00												
μ	0.66	0.88	0.42	1.00											
X	-0.66	-0.88	-0.42	-1.00	1.00										
η	-0.40	0.81	1.00	0.42	-0.42	1.00									
ω	-0.91	-0.61	-0.02	-0.92	0.92	-0.02	1.00								
f	0.16	-0.05	-0.15	0.04	-0.04	-0.15	-0.13	1.00							
λ _{max}	-0.25	-0.71	-0.51	-0.67	0.67	-0.51	0.52	-0.29	1.00						
E (AM1)	0.20	-0.69	-0.77	-0.43	0.43	-0.77	0.13	0.34	0.25	1.00					
H accep	-0.13	-0.84	-0.71	-0.71	0.71	-0.71	0.47	0.18	0.63	0.76	1.00				
ovality	0.41	-0.59	-0.80	-0.25	0.25	-0.80	-0.08	0.33	0.39	0.60	0.67	1.00			
C2	-0.28	-0.06	0.11	-0.18	0.18	0.11	0.26	-0.11	0.17	-0.20	-0.08	-0.22	1.00		
C3	-0.40	-0.19	0.07	-0.34	0.34	0.07	0.42	-0.37	0.43	-0.27	0.04	-0.19	0.89	1.00	
PIC50	0.41	-0.46	-0.67	-0.15	0.15	-0.67	-0.13	0.27	0.40	0.58	0.64	0.63	-0.28	-0.24	1.00

Tableau 5 : Matrice de corrélation des descripteurs moléculaires calculés par les méthodes DFT et topologique.

Variables	E _{HOMO}	E _{LUMO}	Egap	μ	X	η	ω	f	λ _{max}	E (ev)	H accep	ovality	C3	C2	PIC50
E _{HOMO}	1.00														
E _{LUMO}	0.21	1.00													
Egap	-0.10	0.95	1.00												
μ	0.48	0.96	0.83	1.00											
X	-0.48	-0.96	-0.83	-1.00	1.00										
η	-0.10	0.95	1.00	0.83	-0.83	1.00									
ω	-0.28	-0.99	-0.91	-0.97	0.97	-0.91	1.00								
f	0.65	0.16	-0.05	0.33	-0.33	-0.05	-0.13	1.00							
λ _{max}	0.11	-0.84	-0.89	-0.73	0.73	-0.89	0.81	0.00	1.00						
E (ev)	0.26	0.70	0.63	0.71	-0.71	0.63	-0.71	-0.02	-0.56	1.00					
H accep	-0.23	-0.85	-0.80	-0.83	0.83	-0.80	0.90	-0.02	0.72	-0.75	1.00				
ovality	0.11	-0.60	-0.64	-0.51	0.51	-0.64	0.60	0.30	0.56	-0.82	0.66	1.00			
C3	0.08	-0.17	-0.20	-0.13	0.13	-0.20	0.14	-0.13	0.33	0.11	0.03	-0.21	1.00		
C2	0.10	-0.06	-0.09	-0.03	0.03	-0.09	0.04	-0.07	0.14	0.23	-0.09	-0.22	0.89	1.00	
PIC50	0.30	-0.55	-0.66	-0.41	0.41	-0.66	0.58	0.40	0.57	-0.40	0.66	0.64	-0.23	-0.28	1.00

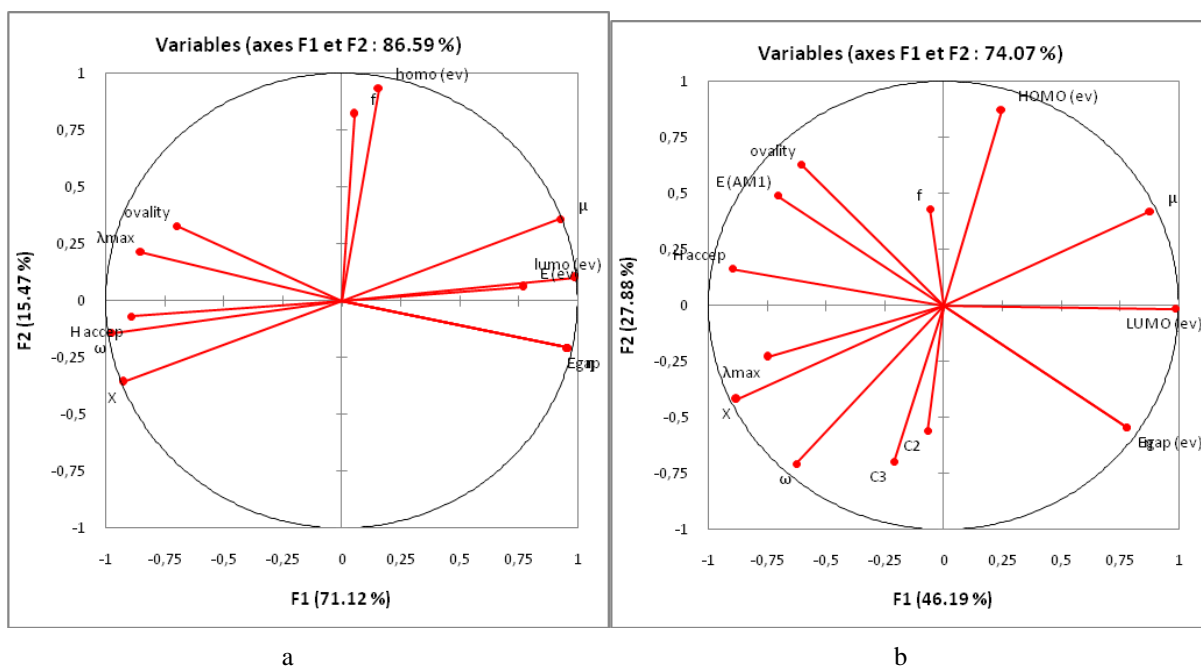


Figure 7 : Cercle de corrélation des deux méthodes : (a) Méthodes DFT et topologique
(b) Méthodes AM1 et topologique.

Dans ce paragraphe, nous présentons la signification des descripteurs représentés dans la matrice de corrélation :

E_{HOMO} : est l'énergie de l'orbitale moléculaire la plus haute occupée, mesurée en eV ; c'est le niveau d'énergie le plus élevé dans la molécule qui contient des électrons. Elle est directement liée au potentiel d'ionisation. Lorsqu'une molécule agit comme une base de Lewis (un doublet d'électrons donneur) dans la formation d'une liaison, les électrons sont alimentés à partir de cette orbitale. Elle mesure la nucléophilie d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par des électrophiles ;

E_{LUMO} : est l'énergie de l'orbitale moléculaire la plus basse inoccupée, mesurée en eV ; c'est le niveau d'énergie le plus bas dans la molécule qui ne contient pas d'électrons. Elle est directement liée à l'affinité d'électron. Lorsqu'une molécule agit comme un acide de Lewis (un doublet d'électrons accepteur) dans la formation de liaisons, des doublets d'électrons entrants sont reçus dans cette orbitale. Elle mesure l'électrophilicité d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par les nucléophiles ;

E_{gap} : est la différence entre E_{HOMO} et E_{LUMO}. Elle traduit l'énergie comprise entre l'orbitale moléculaire la plus haute occupée et la plus basse vacante, c'est un indice de stabilité. Cette différence d'énergie sert de mesure de l'excitabilité d'une molécule. Ainsi, plus l'intervalle d'énergie sera faible et plus la molécule pourra interagir avec l'environnement.

Un grand écart HOMO-LUMO implique une grande stabilité pour la molécule dans le sens de sa faible réactivité dans les réactions chimiques, et de même, un faible écart implique une grande réactivité de la molécule. L'écart HOMO-LUMO a également été utilisé comme une approximation de la plus faible énergie d'excitation de la molécule ;

χ : L'électronégativité mesure la capacité d'un atome à attirer vers lui les électrons, elle est déterminée par la relation suivante

$\chi = - (E_{LUMO} + E_{HOMO})/2$. Elle est définie comme l'opposé du potentiel chimique qui mesure la tendance du nuage électronique à s'échapper de la molécule, c'est un paramètre global du système moléculaire égal à la pente de l'énergie en fonction du nombre d'électrons N à potentiel externe $v(r)$ constant telle que définie par Parr et Mulliken [35] ;

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

λ_{max} : est la longueur d'onde du maximum d'absorption, Les transitions électroniques correspondent au passage des électrons des orbitales moléculaires liantes et non liantes remplies, vers des orbitales moléculaires anti-liantes non remplies. L'absorption provient de ce passage (transition) entre les deux orbitales moléculaires. La longueur d'onde d'absorption et l'énergie d'activation, dépendent de la nature des orbitales mises en jeu ;

f : la force d'oscillation qui correspond à la probabilité pour que la transition électronique soit permise ;

η : la dureté qui peut être obtenue par la relation suivante : $\eta = (E_{LUMO} - E_{HOMO})/2$ [36] ;

ω : l'indice d'électrophilicité utilisé pour caractériser la capacité d'une molécule à engendrer un transfert d'électrons, il est calculé selon la formule suivante : $\omega = \chi^2/2\eta$ [36] ;

E_T : Energie totale, qui est calculée à partir de la conformation optimisée de la géométrie la plus stable, pour une molécule isolée à l'état fondamental. L'énergie totale calculée, mesurée en eV, peut être utilisée comme descripteur moléculaire quantique. Cette énergie approximative a été calculée pour une conformation optimisée de la géométrie la plus stable dont la structure d'énergie est minimale.

H accepteur : est le nombre d'accepteurs de liaisons hydrogène, il s'agit du nombre d'atomes possédant des doublets non liants (azote, oxygène, ou fluor) et capables de se lier par liaison hydrogène à d'autre molécules.

Selon la « règle des cinq de Lipinski », lors de l'identification de la similarité médicamenteuse, les médicaments délivrés par voie orale doivent avoir un nombre d'accepteurs de liaisons hydrogène (NHA) inférieur ou égal à 10 (optimum d'environ 5) et un nombre de donneurs de liaisons hydrogène (NHD) inférieur ou égal à 5 (optimum de 2).

TVconn : définit la connectivité des hétéroatomes constituant la molécule ;

C₂ : la charge partielle de Mulliken du carbone 2 ;

C₃ : la charge partielle de Mulliken du carbone 3.

III-2 Modèles 2D-QSAR élaborés

La méthode des moindres carrés partiels a été utilisée dans cette étude pour relier les descripteurs retenus par les deux méthodes avec l'activité cytotoxique de l'acide ursolique et ses dérivés selon les équations suivantes :

- Le modèle retenu en utilisant les méthodes AM1 et topologique est le suivant :

$$\text{pCI}_{50} = 9,35197 - 0,53044*\text{Egap} + 0,13546*\text{H accep} + 1159234*\text{TVConn} - 0,08807*\text{C3}$$

(Eq : 1)

$$\text{N}=25 \quad \text{R}^2 = 0,77 \quad \text{R}^2 \text{ ajusté} = 0,72 \quad \text{RMSE} = 0,19 \quad \text{F}=15,86 \quad \alpha=1\%$$

- Le modèle retenu en utilisant les méthodes DFT et topologique est le suivant :

$$\text{pCI}_{50} = 8,43702 - 0,72727*\text{Egap} + 1143592*\text{TVConn} - 0,17283*\text{C3} \quad \text{(Eq : 2)}$$

$$\text{N} = 25 \quad \text{R}^2 = 0,83 \quad \text{R}^2 \text{ ajusté} = 0,81 \quad \text{RMSE} = 0,16 \quad \text{F}=33,73 \quad \alpha=1\%$$

Avec N est le nombre de molécules dans l'ensemble d'apprentissage ;

R² : Coefficient de détermination ;

RMSE : Racine de l'erreur moyenne ;

R² ajusté : Coefficient de détermination ajusté ;

F : Test de Fisher.

La signification de chaque descripteur a été examinée par le test de Student, qui nous a montré que les descripteurs retenus sont hautement significatifs au seuil $\alpha = 1\%$ (Tableau 6).

La corrélation entre les valeurs prédites et observées pour les deux modèles 2D-QSAR retenus est illustrée sur la figure 7. Les valeurs résiduelles entre les activités prédites et observées sont présentées dans le tableau 7.

La multi-colinéarité entre les descripteurs retenus dans chaque modèle 2D-QSAR ci-dessus a été vérifiée en calculant les facteurs de variation d'inflation VIF. Le VIF [37] est défini comme $1 / (1 - R^2)$, où R est le coefficient d'inter-corrélation entre un descripteur et tous les autres descripteurs du modèle proposé. Une valeur VIF supérieure à 5 indique que le modèle est instable et une valeur comprise entre 1 et 4 indique que le modèle est acceptable.

Ces valeurs sont rassemblées dans le tableau 4 et leurs analyses montrent que les descripteurs utilisés dans les modèles proposés ont une corrélation très faible.

Le niveau théorique de calcul est une tâche importante pour la précision du calcul des descripteurs. Les résultats présentés ci-dessus démontrent l'effet du niveau théorique utilisé pour les calculs. Les méthodes semi-empiriques telle que AM1 utilisent les paramètres empiriques ou expérimentaux pour traiter l'équation de Schrödinger et omettent certains calculs d'intégrale moléculaire. Elles sont donc beaucoup plus rapides que la méthode DFT-B3LYP. Par conséquent, ils sont plus largement utilisés dans les calculs de propriétés moléculaires, mais la précision de leurs résultats est inférieure aux méthodes DFT [38].

Les modèles 2D-QSAR ci-dessus, obtenus à l'aide des descripteurs issus des calculs DFT (3 descripteurs), sont bien meilleurs que ceux obtenus à l'aide des méthodes semi-empiriques AM1 (4 descripteurs) en termes de critères d'ajustement. La figure 8 illustre la corrélation entre les activités observées et prédites pour les deux niveaux de calculs DFT et AM1.

Afin de déterminer l'importance de chaque descripteur dans les modèles 2D-QSAR obtenus, l'analyse des coefficients normalisés en fonction des descripteurs a été appliquée afin de rechercher les descripteurs les plus importants et qui sont le mieux corrélés à l'activité biologique des molécules.

Tableau 6 : Test de Student des descripteurs sélectionnés dans les deux modèles
PLS retenus

Descripteurs	Coefficients	Erreur standard	t	p-value	VIF
Méthodes AM1 et topologique					
Egap	-0,796	0,1924	-4,1373	0,0006	3,05
H accep	0,471	0,1512	3,1145	0,0057	1,88
TVConn	0,5576	0,1835	3,0394	0,0067	2,77
C3	-0,2903	0,113	-2,5692	0,008	1,05
Méthode DFT et topologique					
Egap	-0,7273	0,0872	-8,3366	< 0,0001	2,56
TVConn	1143591,638	296442,7153	3,8577	0,001	2,46
C3	-0,1728	0,0284	-6,0851	< 0,0001	1,06

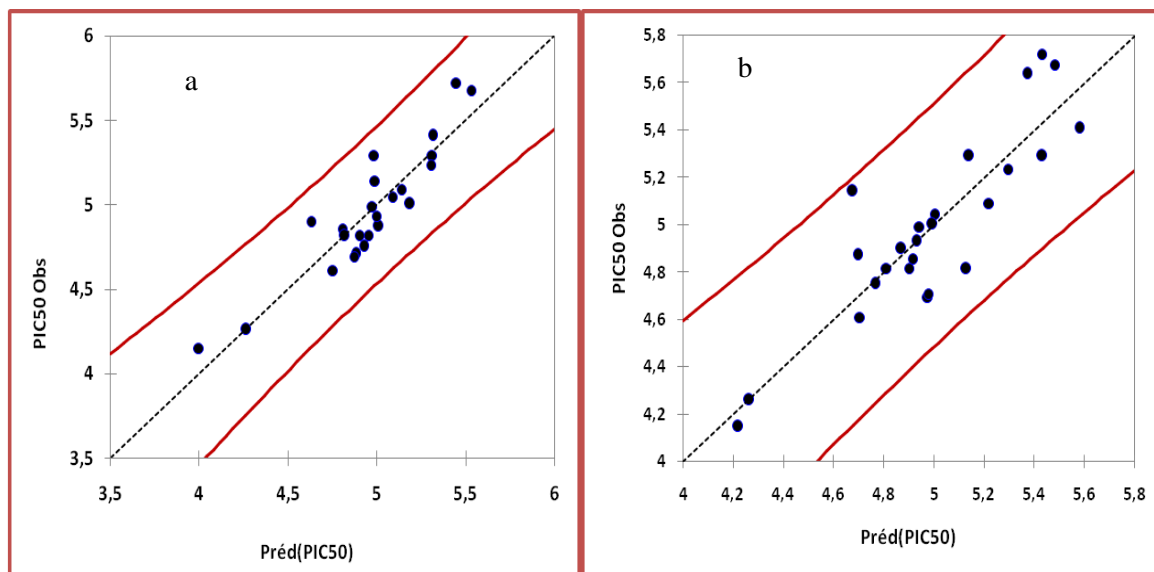


Figure 9 : Relation entre les pCI50 (Obs) et pCI50(Pred) obtenues par PLS :
(a) Méthode DFT et topologique (b) Méthode AM1 et topologique.

Tableau 7 : Valeurs prédites et observées et leurs résiduels pour les deux modèles sélectionnés

Molécules	PLS (Méthodes DFT et topologique)			AM1 (Méthodes AM1 et topologique)	
	PIC50	Préd(pCI50)	Résidu	Préd(pCI50)	Résidu
M2	4,1487	3,9986	0,1502	4,8703	0,0293
M4	4,6108	4,7533	-0,1425	4,2211	-0,0724
M6	4,8729	5,0105	-0,1376	4,7051	-0,0943
M7	4,752	4,9281	-0,1761	4,2645	-0,0041
M8	5,0458	5,0892	-0,0435	4,6974	0,1755
M10	4,8182	4,901	-0,0828	4,7662	-0,0142
M13	4,857	4,8064	0,0505	5,003	0,0428
M14	4,8153	4,8192	-0,0039	4,6769	0,4658
M15	4,9914	4,9759	0,0155	4,8123	0,0059
M16	4,8182	4,9515	-0,1333	4,9765	-0,2819
M17	4,7077	4,883	-0,1752	4,9158	-0,0588
M19	4,9318	4,9987	-0,0669	5,1291	-0,3138
M21	5,0044	5,1862	-0,1818	4,9425	0,0489
M22	5,2924	5,3132	-0,0208	4,9004	-0,0823
M24	5,4089	5,3136	0,0954	4,9769	-0,2691
M27	5,0915	5,1393	-0,0478	4,9293	0,0025
M28	5,2366	5,3063	-0,0697	5,1398	0,1526
M29	5,6778	5,5308	0,147	4,9946	0,0097
M30	5,7212	5,4476	0,2737	5,4313	-0,1389
M11	4,6946	4,8705	-0,1759	5,5804	-0,1715
M20	5,2924	4,9841	0,3083	5,3718	0,2665
M1	4,8996	4,6298	0,2698	5,2197	-0,1282
M9	5,1427	4,9905	0,1522	5,2956	-0,0591
M5	4,2604	4,2651	-0,0047	5,4804	0,1974

III-3 Pouvoirs prédictifs des modèles

III-3-1 Validation interne (Leave-One-Out)

La capacité prédictive interne des deux modèles retenus a été testée par la validation croisée (LOO). Le meilleur coefficient de détermination de la validation interne a été obtenu par les méthodes DFT et topologique ($Q^2=0,73$), comparativement à celui des méthodes AM1 et topologique ($Q^2=0,67$). D'après cette étude, nous pouvons conclure que le modèle obtenu par la méthode DFT présente une meilleure stabilité interne que celui du modèle obtenu par la

méthode AM1. La Figure 9 illustre la relation entre les valeurs prédites en fonction de celles observées, obtenue par la validation croisée (LOO).

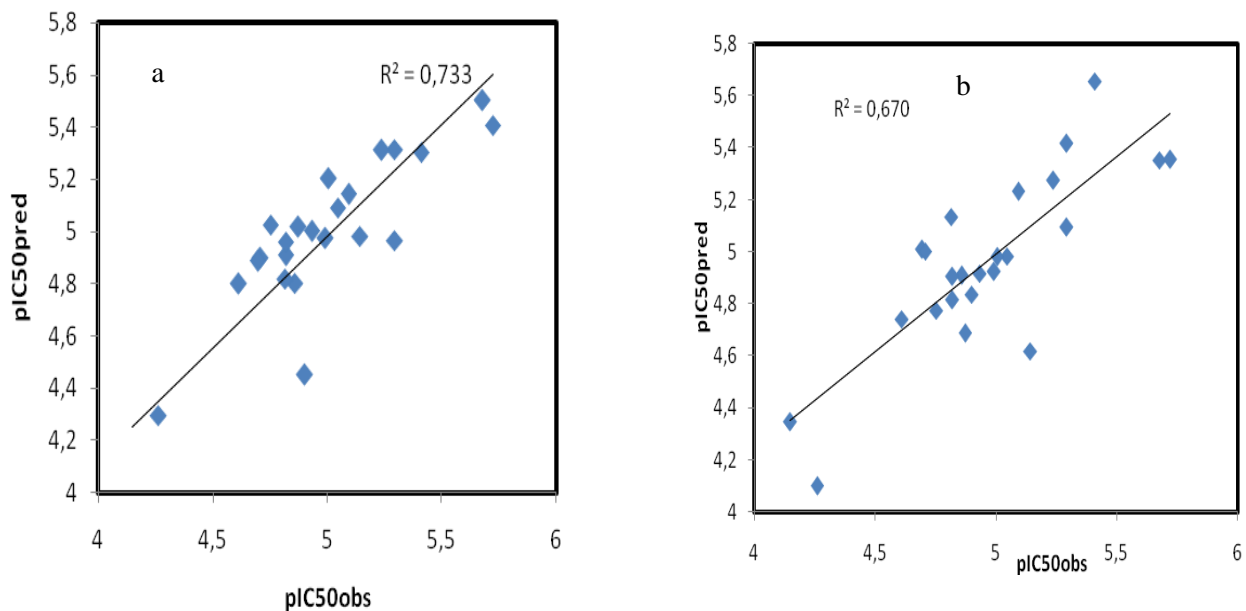


Figure 9 : Corrélation entre les activités prédites (pCI50 préd) et observées (pCI50 Obs) obtenue par la validation croisée : (a) Méthode DFT (b) Méthode AM1.

III.2. Test de Randomisation

Le test de randomisation a été utilisé pour éviter les corrélations fortuites (« chance corrélation ») entre les activités et les descripteurs. A partir du tableau 8, nous pouvons constater que les modèles initiaux présentent une bonne performance par rapport aux modèles randomisés, ce qui nous a permis de conclure que la probabilité de trouver une corrélation fortuite dans les modèles obtenus est faible.

Tableau 8 : Test de randomisation pour les deux modèles 2D-QSAR sélectionnés.

Itération	Test de randomisation des deux modèles 2D-QSAR			
	Méthode DFT et topologique		Méthode AM1 et topologique	
	R ²	Q ²	R ²	Q ²
Modèle initial	0,83	0,73	0,77	0,67
Random 1	0,299	-0,073	0,085	-0,319
Random 2	0,292	-0,097	0,144	-0,594
Random 3	0,19	-0,163	0,168	-0,204
Random 4	0,068	-0,141	0,088	-1,025
Random 5	0,095	-0,297	0,196	-1,887
Random 6	0,063	-0,295	0,042	-1,329
Random 7	0,019	-0,178	0,161	-2,544
Random 8	0,14	-0,699	0,147	-0,495
Random 9	0,348	-0,013	0,119	-0,238
Random 10	0,174	-0,785	0,065	-0,46
Random 11	0,178	-0,211	0,116	-1,68
Random 12	0,161	-0,474	0,286	-0,185
Random 13	0,041	-0,237	0,062	-1,757
Random 14	0,163	-0,285	0,031	-0,657
Random 15	0,217	-0,156	0,121	-1,281
Random 16	0,134	-0,992	0,079	-0,923
Random 17	0,086	-0,511	0,09	-0,407
Random 18	0,21	-0,017	0,054	-0,904
Random 19	0,405	0,203	0,059	-0,454
Random 20	0,071	-0,597	0,131	-0,45
Random 21	0,082	-0,637	0,352	-0,576
Random 22	0,142	-1,032	0,177	-2,326
Random 23	0,077	-0,738	0,107	-0,788
Random 24	0,29	-0,753	0,382	-0,106
Random 25	0,075	-0,998	0,282	-0,062
Random 26	0,393	0,071	0,123	-0,346
Random 27	0,043	-0,289	0,135	-0,181
Random 28	0,137	-0,416	0,218	-0,259
Random 29	0,064	-0,348	0,149	-1,172
Random 30	0,199	-0,088	0,075	-0,279

III.3. Validation externe

La validation externe a été utilisée dans cette étude afin de tester le pouvoir de prédiction externe des modèles 2D-QSAR retenus. Le modèle 2D-QSAR obtenu par la méthode DFT présente un meilleur pouvoir prédictif externe ($R^2_{\text{pred}}=0,81$) que celui obtenu par la méthode AM1 ($R^2_{\text{pred}}=0,51$). Par conséquent, le modèle 2D-QSAR obtenu par la méthode DFT peut être considéré comme fiable pour prédire l'activité de nouveaux composés appartenant à la même série de molécules étudiées. La corrélation entre les activités prédites et observées, obtenue par la validation externe est illustrée dans la figure 10.

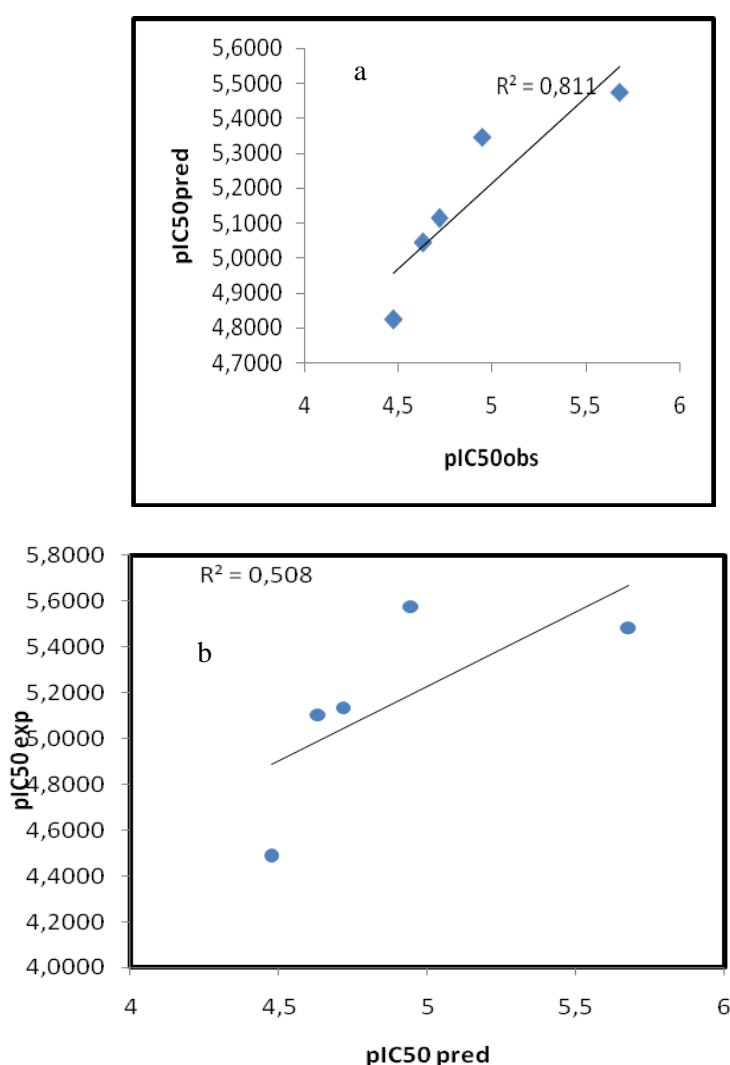
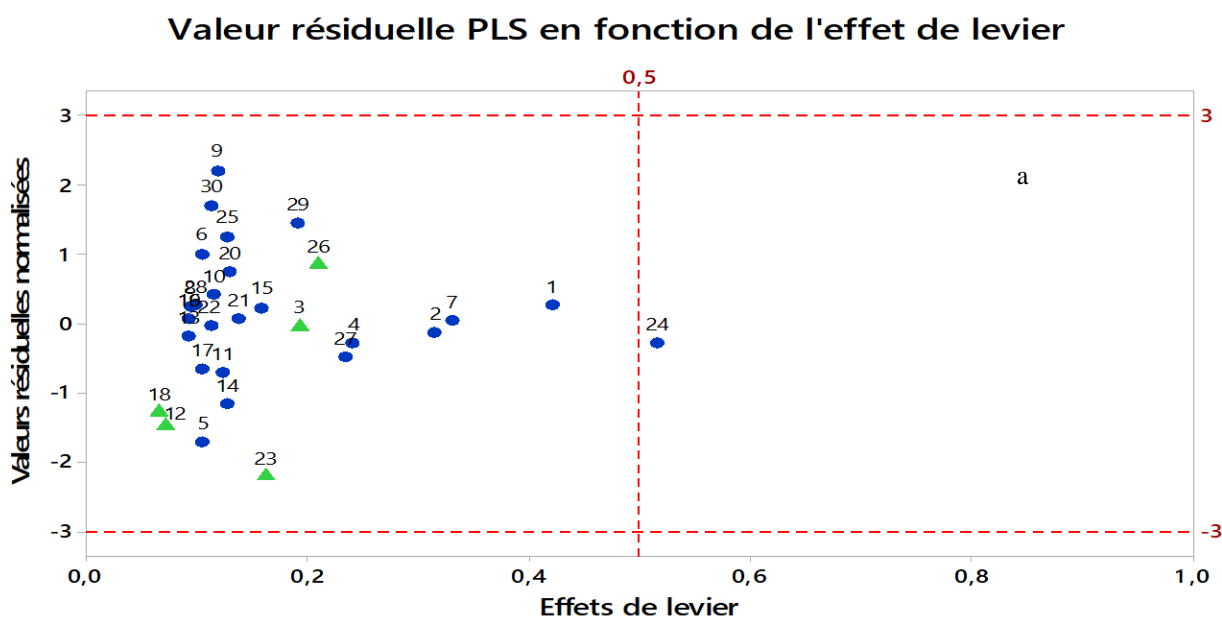


Figure 10 : Corrélation entre les activités prédites (pCI50 préd) et observées (pCI50 Obs) obtenue par la validation externe : (a) Méthode DFT (b) Méthode AM1.

III-4 Domaine d'applicabilité (DA)

La méthode du levier a été utilisée pour déterminer le domaine d'applicabilité des modèles 2D-QSAR élaborés. Cette méthode est basée sur la variation des résiduelles normalisées de la variable dépendante avec la distance entre les valeurs des descripteurs et leurs moyennes appelées leviers h_i . Si un composé a un résiduel et un levier qui dépassent le seuil $h^* = 3(p + 1)/n$ (où p est le nombre de descripteurs et n le nombre de molécules), ce composé est considéré en dehors du domaine d'applicabilité.

Le domaine d'applicabilité du modèle 2D-QSAR élaboré par la méthode AM1 est illustré dans la figure 11. L'analyse de cette figure montre que le composé 24 présente un effet de levier supérieur au seuil h^* ($h^* = 0,5$) et par conséquent ce composé peut être considéré en dehors du domaine d'applicabilité. Cependant, le DA du modèle 2D-QSAR obtenu par la méthode DFT révèle que toutes les composées sont à l'intérieur du domaine d'applicabilité et sont tous bornés à l'intérieur de l'intervalle $\pm 3\sigma$, à l'exception de la molécule M5 qui sort du domaine d'applicabilité.



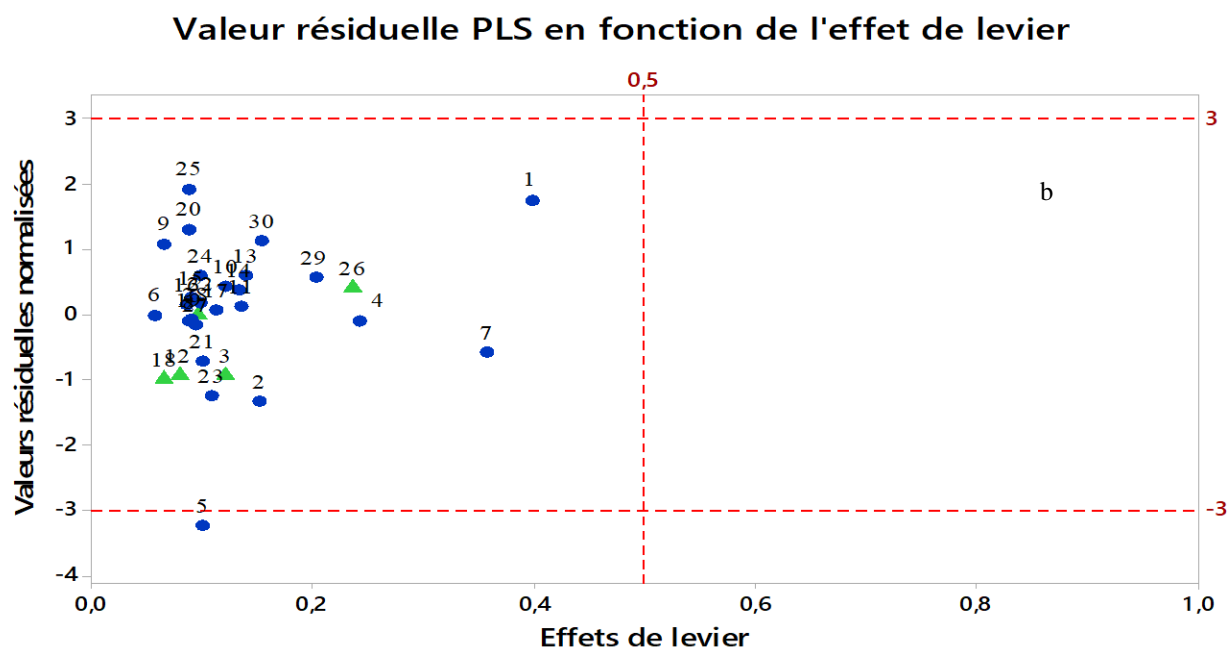


Figure 11 : Diagramme du domaine d'applicabilité : (a) Méthode AM1, (b) Méthode DFT.

Conclusion

Dans ce travail, nous avons étudié l'activité anticancéreuse d'une série de 30 molécules d'acide ursolique et ses dérivées. Nous avons utilisés la méthode semi-empirique AM1 et la méthode empirique DFT. Ces méthodes nous ont permis de construire des modèles prédictifs reliant l'activité biologique aux descripteurs moléculaires, en utilisant des outils statistiques. La stabilité et le pouvoir prédictif de ces modèles ont été confirmés par la validation interne et la validation externe. Les résultats de ce travail montrent que la méthode empirique DFT à donner de bons résultats par comparaison avec la méthode AM1.

Conclusion générale et perspectives

A l'heure actuelle, le cancer est considéré parmi les principales causes de mortalité dans le monde, ce qui nécessite des recherches pour la découverte de nouveaux médicaments en utilisant des études QSAR. Ces méthodes s'appuient sur le principe que les propriétés physico-chimiques et les activités biologiques des molécules dépendent fortement de leurs structures chimiques et de certains descripteurs.

Notre travail a porté sur une étude QSAR de 30 molécules d'acide ursolique et ses dérivés, en utilisant deux méthodes, la première est semi empirique AM1 et la deuxième est empirique DFT.

La méthode semi empirique AM1 nous a permis d'obtenir un modèle avec quatre descripteurs (TVconn, C3, Egap, Haccep) qui ont une influence significative sur l'activité biologique avec un coefficient de détermination égal à 0,77 et une racine de l'erreur moyenne égale à 0,19. Tandis que, la méthode empirique DFT nous a permis d'obtenir un modèle avec seulement trois descripteurs bien corrélés avec l'activité biologique (TVconn, C3, Egap) avec un coefficient de détermination égal à 0,83 et une racine de l'erreur moyenne égale à 0,16.

Dans les perspectives des travaux, nous avons l'intention de prédire *in silico* l'activité de nouveaux dérivés de l'acide ursolique par les modèles 2D-QSAR obtenus et de compléter cette étude 2D-QSAR par l'analyse 3D-QSAR et le docking moléculaire.

Référence bibliographiques :

- [1] J. Pincemail, M. Meurisse, R. Limet, and J. Defraigne, 1999, “Espèces oxygénées activées, antioxydants et cancer,” vol. 4, p. 4.
- [2] R. G. Goodwin *et al.*, 1993, “Molecular and biological characterization of a ligand for CD27 defines a new family of cytokines with homology to tumor necrosis factor,” *Cell*, vol. 73, no. 3, pp. 447–456.
- [3] Mécanisme de développement tumoral. Site d’internet « [https](https://www.google.com/search?q=Mécanismes+du+développement+tumoral&rlz) : [//www.google.com/search?q=Mécanismes+du+développement+tumoral&rlz](https://www.google.com/search?q=Mécanismes+du+développement+tumoral&rlz) »
- [4] John R. (2002). HeLa cells 50 years on : the good, the bad and the ugly. Memoire de Master, Institute of Urology, University College London, 67 Riding House Street, London W1W 7EY, UK. April.
- [5] F.Bray, J.Ferlay, 2018 “Global cancer statistics 2018 : GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”, a cancer journal for clinicians.
- [6] D. Dufour, 2006, “*Évaluation de l’activité biologique du Ledum groenlandicum Retzius*” Université du Québec à Chicoutimi.
- [7] J. Seignalet, 2008 “*L’alimentation, ou, La troisième médecine*. Paris : F.-X. de Guibert, 2011.
- [8] D. W. Christianson, “Unearthing the roots of the terpenome,” *Curr. Opin. Chem. Biol.*, vol. 12, no. 2, pp. 141–150, Apr..
- [9] N. Dudareva, 2004 “Biochemistry of Plant Volatiles,” *PLANT Physiol.*, vol. 135, no. 4, pp. 1893–1902.
- [10] J. Bohlmann and C. I. Keeling, 2008, “Terpenoid biomaterials,” *Plant J.*, vol. 54, no. 4, pp. 656–669.
- [11] T. Benabdelkader, “Biodiversité, bioactivité et biosynthèse des composés terpéniques volatils des lavandes ailées, *Lavandula stoechas* sensu lato, un complexe d’espèces méditerranéennes d’intérêt pharmacologique,” p. 283.
- [12] K. Khadidja, “Contribution à l’étude phytochimique et activité antioxydante des extraits de *Myrtus communis* L. (Rayhane) de la région de Tlemcen (Honaine),” p. 118.
- [13] U. Chaube, “3D-QSAR, molecular dynamics simulations, and molecular docking studies on pyridoaminotropans and tetrahydroquinazoline as mTOR inhibitors,” *Mol Divers*, p. 19.
- [14] kh.ABDELMADJID, “Etude par DFT et TD-DFT de la structure et des propriétés optiques de quelques complexes de métaux de transition”.
- [15] Y. Yousfi, “Etude QSAR de l’activité anti-oxydante d’une série de composés phénoliques,” p. 94.

- [16] C. Frayret, "Application de la Théorie de la Fonctionnelle de la Densité à la modélisation de la diffusion de l'ion oxygène dans des électrolytes solides modèles et des conducteurs mixtes," p. 273.
- [17] *A primer on QSAR/QSPR modeling*. New York, NY : Springer Berlin Heidelberg, 2015.
- [18] M. Dehmer, K. Varmuza, and D. Bonchev, Eds., 2012 "*Statistical Modelling of Molecular Descriptors in QSAR/QSPR : DEHMER : MOL. DESCRIPTOR O-BK*". Weinheim, Germany : Wiley-VCH Verlag GmbH & Co. KGaA,".
- [19] K. Roy, S. Kar, and R. N. Das, 2015 "Statistical Methods in QSAR/QSPR," in *A Primer on QSAR/QSPR Modeling*, Cham : Springer International Publishing , pp. 37–59.
- [20] M. W. Gardner and S. R. Dorling, 1998 "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, no. 14–15, pp. 2627–2636.
- [21] P. Potočnik, 2012 "Neural Networks : MATLAB examples," p. 91.
- [22] S. Weisberg, 2005 "*Applied linear regression*", 3rd ed. Hoboken, N.J : Wiley-Interscience.
- [23] P. Pratim Roy, S. Paul, I. Mitra, and K. Roy, "On Two Novel Parameters for Validation of Predictive QSAR Models," *Molecules*, vol. 14, no. 5, pp. 1660–1701, Apr. 2009.
- [24] P. Gramatica, 2007 "Principles of QSAR models validation : internal and external," *QSAR Comb. Sci.*, vol. 26, no. 5, pp. 694–701.
- [25] P. Pratim Roy, S. Paul, I. Mitra, and K. Roy, 2009 "On Two Novel Parameters for Validation of Predictive QSAR Models," *Molecules*, vol. 14, no. 5, pp. 1660–1701.
- [26] M. Stitou, H. Toufik, T. Akabli, M. Bouachrine, B. Mounir, F. Lamchouri, 2018 "2D-QSAR method of lupane-type saponins the treatment of cancer cell line. Rhazes : Green and Applied Chemistry" 2605-6895.
- [27] G. JOST "synthèse de dérivés glycosidiques de type ursane et évaluation de leur cytotoxicité"
- [28] A. S. Leal, R. Wang, J. A. R. Salvador, and Y. Jing, 2012, "Synthesis of novel ursolic acid heterocyclic derivatives with improved abilities of antiproliferation and induction of p53, p21waf1 and NOXA in pancreatic cancer cells," *Bioorg. Med. Chem.*, vol. 20, no. 19, pp. 5774–5786.
- [29] Cao R., Fan W., Guo L., Mac Q., Zhang G., Li J., Chen X., Ren Z. and Qiu L. (2012). Synthesis and structure-activity relationships of harmine derivatives as potential antitumor agents. *J. Med. Chem* ; 60 : 135-143.
- [30] ChemBioOffice, (2015) PerkinElmer Informatics. [http : //www.cambridgesoft.com](http://www.cambridgesoft.com).
- [31] Frisch, M.J., et al. (2009) Gaussian 09, Revision B. 01. Gaussian, Inc., Wallingford.

- [32] XLSTAT 2014 software (2018) (XLSTAT Company), <http://www.xlstat.com>.
- [33] Minitab 17.1.0 Statistical Software (2017). State College, PA : Minitab, Inc. (www.minitab.com).
- [34] T. Akabli, H. Toufik, A. Yasri, H. Bih, and F. Lamchouri, 2018, "Combining ligand-based and structure-based drug design approaches to study the structure-activity relationships of a β -carboline derivative series," *Struct. Chem.*
- [35] R. Hmamouchi and M. Bouachrine, 2016, "Tentative Pratique du Relation Quantitatives Structure-Activité/Propriété," p. 13.
- [36] S. Chtita, 2017 "Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments," p. 187.
- [37] R. M. O'brien, 2007 "A Caution Regarding Rules of Thumb for Variance Inflation Factors," *Qual. Quant.*, vol. 41, no. 5, pp. 673–690.
- [38] O. Deeb and B. W. Clare, 2008 "Comparison of AM1 and B3LYP-DFT for Inhibition of MAO-A by Phenylisopropylamines : A QSAR Study," *Chem. Biol. Drug Des.*, vol. 71, no. 4, pp. 352–362.