La synthèse chimique et la détermination des propriétés soient biologiques soient chimiques des composés organiques est faite à l'aide de l'expérience, mais cette dernière est coûteuse et souvent très longue. L'un des challenges de la chimio-informatique est d'éviter ces problèmes liés au coût et à la durée des expériences. En effet, elle permet de décrire de manière simple des composés afin de pouvoir les utiliser dans des études de similarité (pour trouver de nouveaux composés potentiellement intéressants) ou de pouvoir prédire leur activité en se basant sur les informations contenues dans les composés déjà connus.

La fabrication des médicaments basée sur la corrélation activité /structure a pour but de sélectionner des molécules valables à utiliser dans le domaine pharmaceutique. La corrélation structure-activité est réalisée à l'aide de la régression linéaire multiple (RLM) et réseau neuronal artificiel (RNN) par logiciel SPSS.

Notre thèse est focalisée sur l'étude QSAR avec la modélisation des molécules par méthode DFT, 6-31G comme une base de calcul. À l'aide du logiciel gaussien view, on a validé le modèle QSAR par les tests statistiques RLM et RNN avec le logiciel SPSS. Après la validation interne et externe, nous avons créé un ensemble de molécules avec des activités très intéressantes tout en appliquant la règle Lipinski (ROF) pour la confirmation de l'utilisation de 28 molécules de dérivées d'indazole comme des médicaments à voie orale, ainsi que la création de nouvelles molécules avec des activités très importantes dans le domaine pharmaceutique..

N°d'ordre : 329/21

Simulation de molécules organiques dérivées d'indazole (logiciel gaussian - méthode QSAR) et création de nouvelles molécules thérapeutiques

LAKRIKH SIHAM

N° d'ordre : 329/2021

UNIVERSITE SULTAN MOULAY SLIMANE

Faculté des Sciences et Techniques

Béni-Mellal

Centre d'Études Doctorales : Sciences et Techniques

Formation Doctorale : Mathématiques et Physique Appliquées

THÈSE

Présentée par

LAKRIKH Siham

Pour l'obtention du grade de

DOCTEUR

Spécialité ***: Chimie / Chemoinformatique***

Option ***: Modélisation moléculaire***

Simulation de molécules organiques dérivées d'indazole (logiciel gaussian - méthode QSAR) et création de nouvelles molécules thérapeutiques

Soutenue le Jeudi 01 Avril 2021 à 10h devant la commission d'examen:

Pr.Mohammed CHIGR	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc.	Président
Pr.Ahmed LEBKIRI	Professeur, Université Ibn Tofaïl, Faculté des Sciences, Kenitra, Maroc	Rapporteur
Pr.Mustapha BOULGHALLAT	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc	Rapporteur
Pr.Mohamed BERKANI	Professeur, Université Sultan Moulay Slimane, F.P. Béni-Mellal, Maroc	Rapporteur
Pr.El Hassan EL JAOUI	Professeur Assistant, Université Sultan Moulay Slimane, E.S.E.F. Béni-Mellal, Maroc	Examinateur
Pr.Ahmed JOUAITI	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc	Directeur de Thèse
Pr. Latifa LAALLAM	Professeur, Université Sultan Moulay Slimane, F.S.T. Béni-Mellal, Maroc	Co-Directeur de Thèse

N° d'ordre : …./…./2020

UNIVERSITÉ SULTAN MOULAY SLIMANE

FACULTÉ DES SCIENCES ET TECHNIQUES

Centre d'Études Doctorales : Sciences et Techniques

Formation doctorale : Ressources Naturelles, Chimie, Environnement et Santé (RNCES)

THÈSE

Présentée par :

LAKRIKH Siham

Pour obtenir le grade de

Docteur

Discipline : Chimie

Spécialité : Chimie / Chémoinformatique

Simulation de molécules organiques dérivées d'indazole (logiciel gaussian - méthode QSAR) et création de nouvelles molécules thérapeutiques

---

Soutenue le **01 / 04 / 2021** devant les membres de jury:

Mr.	Pr. Mohammed CHIGR PES FST –Béni Mellal	Président
Mr.	Pr. Ahmed LEBKIRI PES FS- Kenitra	Rapporteur
Mr.	Pr. Mustapha BOULGHALLAT PES FST–Béni Mellal	Rapporteur
Mr.	Pr. Mohamed BERKANI PES FST–Béni Mellal	Rapporteur
Mr.	Pr. El Hassan EL JAOUI PA ESEF- Béni Mellal	Examinateur
Mr.	Pr .Ahmed JOUAITI PES FST –Béni Mellal	Directeur de thèsé
Mme.	Pr. Latifa LAALLAM PES FST- Béni Mellal	Codircteur de thèse

Année Universitaire 2020/2021

Resume

La synthèse chimique et la détermination des propriétés soient biologiques soient chimiques des composés organiques est faite à l'aide de l'expérience, mais cette dernière est coûteuse et souvent très longue.

L'un des challenges de la chimio-informatique est d'éviter ces problèmes liés au coût et à la durée des expériences. En effet, elle permet de décrire de manière simple des composés afin de pouvoir les utiliser dans des études de similarité (pour trouver de nouveaux composés potentiellement intéressants) ou de pouvoir prédire leur activité en se basant sur les informations contenues dans les composés déjà connus.

La fabrication des médicaments basée sur la corrélation activité /structure a pour but de sélectionner des molécules valables à utiliser dans le domaine pharmaceutique. La corrélation structure-activité est réalisée à l'aide de la régression linéaire multiple (RLM) et réseau neuronal artificiel (RNN) par logiciel SPSS.

Notre thèse est focalisée sur l'étude QSAR avec la modélisation des molécules par méthode DFT, 6-31G comme une base de calcul. À l'aide du logiciel gaussien view, on a validé le modèle QSAR par les tests statistiques RLM et RNN avec le logiciel SPSS. Après la validation interne et externe, nous avons créé un ensemble de molécules avec des activités très intéressantes tout en appliquant la règle Lipinski (ROF) pour la confirmation de l'utilisation de 28 molécules de dérivées d'indazole comme des médicaments à voie orale, ainsi que la création de nouvelles molécules avec des activités très importantes dans le domaine pharmaceutique.

Mots-clef : QSAR, RLM, Modèle, Descripteurs, DFT, Medicaments

Remerciement

Il n’est pas possible de se développer ou d'évoluer dans quoi que ce soit par soi-même. Cette thèse a été soulevée parles contributions essentielles de différentes personnes de différents pays. Le soutien, l’aide, mais le plus important, les sourires que j’ai trouvés, le café les pauses et les mots d’encouragement. J’espère que ces personnes pourront trouver dans ces mots ma plus sincère gratitude.

Ce travail de thèse de doctorat a été réalisé au sein du laboratoire de développement durable (L2D) à la Faculté des Sciences et Techniques - Université Sultan Moulay Slimane. Il a été suivi scrupuleusement et sous la direction du Monsieur Ahmed JOUAITI, Professeur à la Faculté des Sciences et Techniques, Université Sultan Moulay Slimane-Béni Mellal, Maroc et Madame Latifa LAALLAM, Professeure à la Faculté des Sciences et Techniques, Université Sultan Moulay Slimane-Béni Mellal, Maroc.

Je tiens à exprimer ma plus sincère gratitude à mes superviseurs, les docteurs Ahmed JOUAITI et Latifa LAALLAM, pour leur direction, leurs formations et leurs encouragements continus. Je les remercie pour leur accueil, leurs conseils, leur confiance, leur patience et pour leur disponibilité tout au long de ce travail.

Mes remerciements sont aussi destinés aux rapporteurs et aux autres membres du jury pour l’honneur qu’ils m’ont fait en évaluant mes travaux de thèse.

Merci aux Mr Ibrahim BELAYACHI, Mohammed ELKHALLOUFI et Hajar ATMANI pour la lecture de la thèse.

Un profond merci à mes parents et mes sœurs pour leur soutien permanent tout au long de ma vie, j’offre ce travail à mon frère Amine LAKRIKH qui est décédé en 2017.

Enfin, je tiens à remercier l’ensemble des personnes avec qui j’ai pu travailler ou simplement partager le quotidien durant ces quatre années passées dans le laboratoire de L2D.

J’adresse également mon profond remerciement à mes chers collègues enseignants de la faculté des Sciences et Techniques de Béni Mellal.

Liste des figures

Liste des tableaux

Liste des équations

Liste des abreviations

1D	Unidimensionnel
2D	Bidimensionnel
3D	Tridimensionnel
4D	Quadridimensionnel
ADME	**A**bsorption, **D**istribution, **M**étabolisme et **E**limination
ADN	**A**cide **D**ésoxyribo **N**ucléique
A	**A**ffinité **E**lectronique
ANOVA	**A**nalyse de **V**ariance
B3LYP	**B**ecke3-**P**arameter**L**ee-**Y**ang-**P**arr
CV	**C**oefficient de **V**ariation
CLOA	**C**ombinaison **L**inéaire d'**O**rbitales **A**tomiques
D	**D**ensité
DM	**D**ipôle **M**oment
EH	**E**nergie de la **H**ydrations
E	**E**nergie d'excitation
GGA	**G**eneralized **G**radient **A**pproximation
GTO	**G**aussian**T**ype**O**rbital
DFT	**D**ensité **F**onctionnelle **T**héorique
HOMO	**H**ighest **O**ccupied **M**olecular **O**rbital
HF	**H**artree-**F**ock
LUMO	**L**owest **U**noccupied **M**olecular **O**rbital
LDA	**L**ocal **D**ensity **A**pproximation
LOO	**L**eave **O**ne **O**ut
OECD	**O**rganisation de **C**oopération et de **D**éveloppement **E**conomique
OA	**O**rbitale **A**tomique
OMF	**O**rbitales **M**oléculaires **F**rontières
OM	**O**rbitales **M**oléculaires
OF	**O**rbitales **F**rontières
PRESS	**S**omme des **C**arrés des **E**rreurs **R**ésiduelle **P**rédite
PSA	**S**uperficie de la **S**urface **P**olaire
PI	**P**otentiel d'**I**onisation
P	**P**olarisabilite
PM3	**P**arametric **M**ethod 3
PM6	**P**arametric **M**ethod 6
RLM	**R**égression **L**inière **M**ultiple
ROF	**R**ule **O**f **F**ive (Règles de Lipinski)
RNN	**R**éseaux **N**eurones **A**rtifériels
RVDW	**R**ayon de **V**an **D**er **W**aals
T	**T**emperature
TCE	**T**ettra-**c**yano**é**thylène

SAR	**S**tructure **A**ctivité **R**elationship
SAG	**G**rille de **S**urface
SVDW	**S**urface de **V**an **D**er **W**aals
QSAR	**Q**uantitative **S**tructure **A**ctivité **R**elationship
QSPR	**Q**uantitative **S**tructure **P**roperty **R**elationship
WM	**W**ight **M**olaire
MR	**R**éfraction **M**oleculaire
MD	**M**oment **D**épolaire
MPA	**M**ulliken **P**opulation **A**nalysis
NHD	**N**ombre de **D**onneurs de liaisons **H**ydrogène
NHA	**N**ombre d'**A**ccepteurs de liaisons **H**ydrogène
VIF	Le **F**acteur d'**I**nflation de la **V**ariance
VM	**V**olume **M**oleculaire
$\mathbf{IC_{50}}$	**50**% **C**oncentration **I**nhibitrice
I	**I**onisation
λ	Longueur d'onde
Log P	**LOG**arithme du coefficient de **P**artition eau/octanol P

Sommaire

Introduction générale

Découvrir de nouveaux médicaments de la manière la plus efficace et la moins coûteuse possible constitue un enjeu majeur pour les années à venir. Il est admis que, en moyenne, pour une molécule qui arrive sur le marché en tant que médicament innovant, 10 000 molécules sont synthétisées et testées. De plus, le développement d'un médicament demande généralement entre 10 et 15 ans de recherche. Il s'agit en effet de trouver une molécule qui doit à la fois présenter des propriétés thérapeutiques particulières, et posséder le minimum d'effets secondaires indésirables [1].

L'un des challenges de la chimio-informatique est d'être capable de décrire de manière simple des composés afin de pouvoir les utiliser dans des études de similarité (pour trouver de nouveaux composés potentiellement intéressants) ou de pouvoir prédire leur activité en se basant sur les informations contenues dans les composés déjà connus.

Le prix de revient d'un médicament est essentiellement dû à ces synthèses longues, coûteuses et finalement inutiles. Pour cette raison, l'industrie pharmaceutique s'oriente vers de nouvelles méthodes de recherche, qui consistent à prédire les propriétés et les activités des molécules avant même que celles-ci ne soient synthétisées. Deux disciplines de la « chimie computationnelle » se sont développées pour répondre à ce besoin : les relations structure-activité ou QSAR (Quantitative Structure-Activity Relationship), et les relations structure-propriété ou QSPR (Quantitative Structure-Property Relationship).

D'un autre côté, ces types de relations consistent essentiellement à la recherche de similitudes entre molécules dans de grandes bases de données de molécules existantes dont les propriétés sont connues. La découverte d'une telle relation permet de prédire les propriétés physiques et chimiques ainsi que l'activité biologique des composés, et de développer de nouvelles théories ou de comprendre les phénomènes observés. Elle permet également de guider la synthèse de nouvelles molécules, sans avoir à les réaliser, ou à analyser les familles entières des composés. Les relations entre les structures des molécules et leurs propriétés ou leurs activités sont généralement établies à l'aide des méthodes de modélisation par apprentissage statistique. Les techniques usuelles reposent sur la caractérisation des molécules Cette thèse contient trois chapitres :

- Le premier chapitre de cette thèse présente des généralités sur le QSAR, ainsi que les principaux types de descripteurs et la façon de sélection.

- Le chapitre suivant introduit des notions de base sur les méthodes statistiques utilisées, tout en montrant comment ces fonctions permettent d'établir une relation entre des données structurées et l'activité biologique des molécules testées (IC_{50}), et les étapes de la validation du modèle QSAR.

- Le troisième chapitre présente les résultats de notre travail de thèses et la prédiction de propriétés et d'activités moléculaires (on a effectué une étude QSAR de deux types d'activité pour une série de 28 dérivés d'indazole à l'aide de divers types de descripteurs moléculaires), et décrit la méthodologie que nous avons élaborée dans notre thèse. On avait montré également que les modélisations obtenues se révèlent généralement de meilleure qualité. Après, on a appliqué la méthode *lipinski* pour confirmer l'utilisation de cette série des molécules testées comme des médicaments par voie orale.

- Enfin, nous terminerons par une conclusion générale et les perspectives envisagées pour ce travail, et à la fin de ce manuscrit, une annexe est consacrée à la description des méthodes de la chimie quantique utilisée pour l'optimisation des structures moléculaires.

Chapitre I : Données bibliographiques et approches QSAR

1 Introduction

La fabrication d'un médicament dans le domaine pharmaceutique basé sur la détermination d'équation de la corrélation entre l'activité biologique des médicaments ou des molécules testées, les propriétés physiques et chimiques, par la chémoinformatiques qui est pour objectif de fournir des outils et des méthodes pour analyser et traiter des données issues des différents domaines de la chimie.

Elle est notamment utilisée en pharmacologie pour la découverte de nouvelles molécules activées et la prédiction de propriétés des structures moléculaires. Ce domaine de la science repose ou base sur le physique et la chimie quantique ainsi que la modélisation de la confirmation de la structure.

1.1 Bio-informatique

La bio-informatique est un domaine scientifique multidisciplinaire qui relie entre la biologie et l'informatique, elle est utilisée dans le but de stocker les informations biologiques (séquences nucléotidiques ou d'acides aminés), pour faciliter la compréhension des phénomènes biologiques. Toutes ces informations sont stockées dans des bases de données afin d'être accessibles et utilisables via des outils informatiques. La grande difficulté de ce domaine est la gestion d'une quantité très importante de données (multiplicité des espèces, cellules, gènes, protéines…), qu'il faut standardiser et nettoyer.

À cela, ajoutons qu'un grand soin doit être apporté à la gestion des données, afin d'éviter toute redondance des informations dans les bases de données. Une fois ces données traitées, elles doivent être rendues accessibles à travers des plateformes faciles à prendre en main par les équipes de recherche. L'analyse de ces données et l'exploitation qui en est faite doit ainsi permettre d'identifier les cibles d'intérêt thérapeutique, de regrouper les protéines au sein de familles et de comprendre les modifications cellulaires [2]. Nombreuses sont les bases de données bio-informatiques. Les plus utilisées été Uniport, pour les séquences de protéines, et la PDB pour les structures tridimensionnelles des protéines [3][4] .

1.2 Chémoinformatique

Introduit à la fin des années 1990, le terme de la chémoinformatique, est apparu afin de décrire l'utilisation en plein essor de l'informatique pour résoudre des problèmes chimiques.

La chémoinformatique mélange entre la chimie et l'informatique. Pour stocker les informations chimiques, l'analyse, la recherche d'entités chimiques et le développement d'outils prédictifs d'inhibiteurs. La chémoinformatiques nécessite de gérer une grande quantité d'informations, qu'il faut standardiser et dont la redondance doit être exclue. La chémoinformatiques est aussi vue comme l'outil d'étude et d'exploration de l'espace chimique. Espace chimique dans lequel les molécules sont représentées par des descripteurs moléculaires [5].

Les définitions les plus populaires et les plus anciennes de la chémoinformatiques sont les suivantes en englais :

The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug leads identification and organization [6]**.**

Chem(o)informatics is a generic term that encompasses the design, creation ,organization,management,retrieval,analysis,dissemination,visualization and use of chemical information [7].

En chimie thérapeutique, l'espace chimique décrit l'ensemble des petites molécules organiques, contenant au maximum 30 atomes lourds, qu'il est théoriquement possible de synthétiser [8][9].Il contiendrait 1060 entités, contre 1023 étoiles dans notre univers observable.[10] Il est important de noter qu'à l'heure actuelle, plus de 99,9% des molécules de l'espace chimique n'ont jamais été synthétisés.[11], en septembre 2015, la base de données CAS référençait plus 100 millions de molécules enregistrées. De ce vaste espace, des librairies de molécules peuvent être générées selon les besoins afin, par exemple, de regrouper toutes les molécules d'un même chémotype, ou bien d'identifier tous les isomères

d'une même formule brute. De la volonté de créer des bases de données de molécules, a vite découlé le besoin de les comparer. Bien qu'aujourd'hui très contestée, c'est sur cette hypothèse que repose une grande partie des décisions prisent en chimie médicinale [12].

Afin d'accéder aux propriétés des molécules, il est d'usage de passer par des bases de données. Ces dernières se cantonnent principalement à référencer des molécules qui ont été synthétisées et testées. Elles fournissent ainsi les informations d'activité nécessaires à la création de modèles statistiques. Elles peuvent aussi renseigner sur les fournisseurs à contacter pour se procurer lesdites molécules. C'est notamment le cas de la ChEMBL et de PubChem [13] [14].

2 Relation quantitative structure activité (QSAR)

Parmi des techniques les plus récentes dans la modélisation moléculaire en domaine de la pharmaceutique le QSPR (en anglais QSPR : Quantitative Structure Property Relationship) et la QSAR (en anglais QSAR : Quantitative Structure-Activity Relationship) ; elles utilisent pour la plupart sur « la recherche d'une relation entre les descripteurs ou propriétés moléculaires, et la propriété biologique ou thermodynamique que l'on souhaite prédire ». Ces méthodes permettent de justifier les données expérimentales disponibles et de prédire les propriétés/activités pour de nouveaux composés ou des composés pour lesquels les données expérimentales ne sont pas disponibles.

Dans ce chapitre, nous avons parlé en général sur l'étude QSAR (on va utiliser dans notre thèse cette abréviation pour indiquer sur la relation quantitative structure-activité), la validation de modèle QSAR (les différentes étapes de développement) et les méthodes statistiques utilisées (les tests statistiques ou les méthodes du chemiometrie utilisable dans ce travail).

2.1 Définition

Les méthodes QSARs basées sur l'hypothèse que l'activité biologique ou la propriété d'un composé chimique est liée à sa structure et les paramètres de la molécule, plus précisément cette approche affirme que l'activité (ou la propriété) et la structure d'un composé chimique sont liées d'un certain algorithme ou équation mathématique, cela est basé sur le postulat de base « les composés chimiques similaires ont des activités similaires ». De

plus, lorsque les paramètres moléculaires sont exprimés par des chiffres, on peut proposer une équation mathématique, ou relation quantitative structure-activité/propriété, entre les deux l'activité en fonction des propriétés physique et chimique.

2.2 Principe

Le principe de méthode QSAR est d'établir une relation mathématique reliant de manière quantitative des propriétés moléculaires, appelées descripteurs, avec une observable macroscopique (activité biologique, toxicité, etc.), pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données. Selon l'équation (1). L'objectif d'un modèle est alors de capter la relation trouvée entre les descripteurs moléculaires et l'activité, afin de créer des règles génériques permettant d'expliquer l'activité étudiée. Le but est ensuite d'appliquer ces règles afin de prédire l'activité de molécules inconnues à partir de leurs descripteurs moléculaires.

Ceci peut être traduit par l'équation suivante : **(1)**

Activité biologique = f (propriétés physico-chimiques)

Remarque : L'expression mathématique de l'étude QSAR obtenue peut alors être utilisée comme moyen prédictif de l'activité/propriété étudiée pour de nouvelles molécules ou des molécules pour lesquelles les données expérimentales ne sont pas disponibles.

Équation 1 : Relation de QSAR

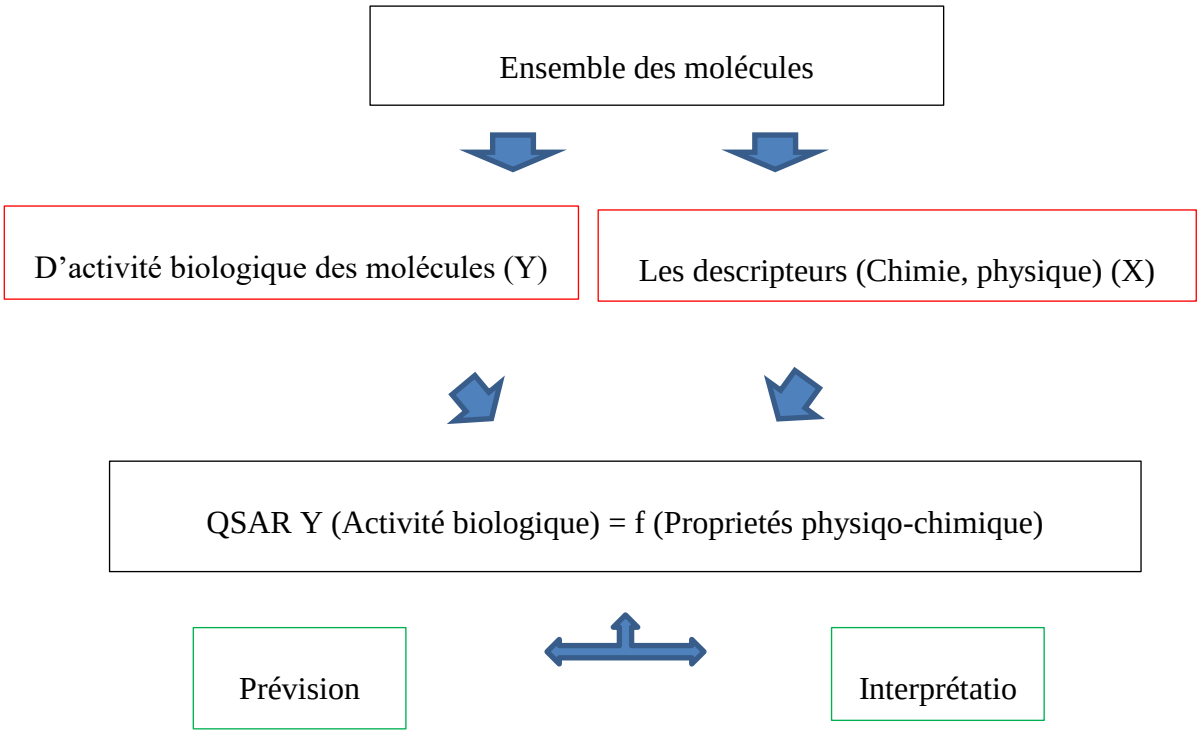


Figure 1 : Étude relation quantitative structure activité (QSAR)

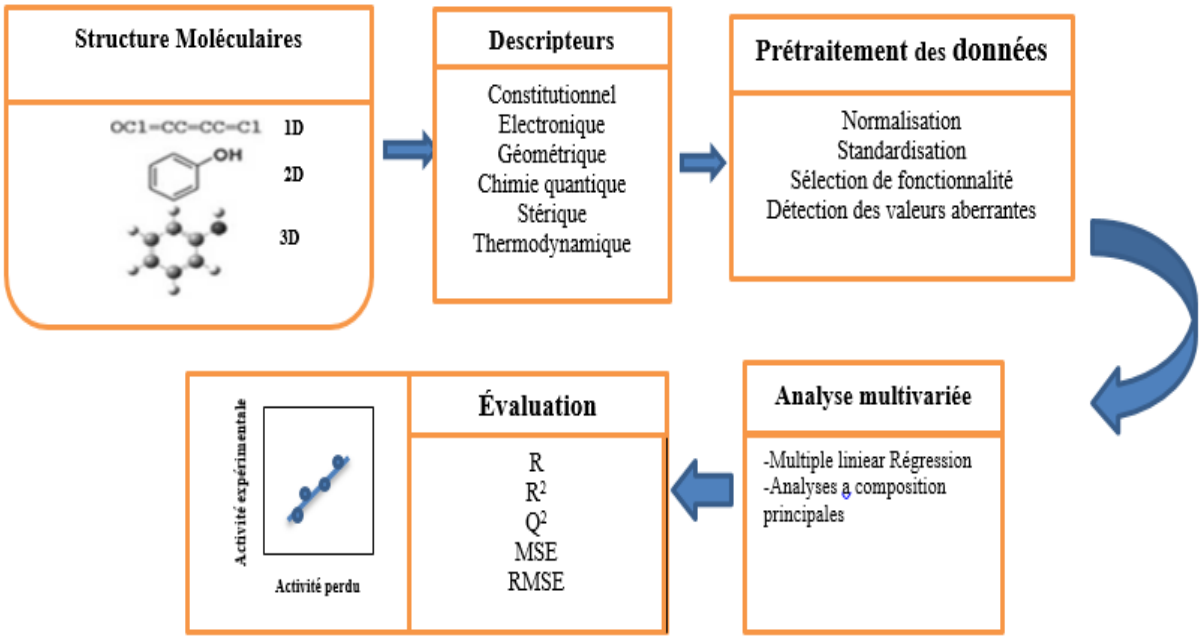


Figure 2 : Étapes d'étude QSAR et leurs étapes de la validation de modèle QSAR [15].

Les descripteurs les plus utilisés lors des études QSAR peuvent être des descripteurs topologiques (2D-QSAR), des descripteurs géométriques ou basés sur des grilles (3D-QSAR). De nouvelles approches plus récentes, QSAR-4D, QSAR-5D et QSAR-6D ont été développées pour améliorer les approches QSAR-3D, en prenant en compte différentes conformations des ligands (4D[16]), et l'adaptation structurale de la cible au ligand (5D[17]) et les effets de solvants (6D[18]).

2.3 Objectifs de QSAR

La plupart des méthodes QSAR se concentrent sur les objectifs suivants:

• corréler quantitativement entre les tendances des modifications de la structure chimique et les modifications respectives de l'activité biologique afin de déterminer quelles propriétés chimiques les plus probables de leurs activités biologiques;

• optimiser les pistes existantes pour améliorer leurs activités biologiques;

• prédire les activités biologiques de composés non testés et parfois encore indisponibles l'association des variations de l'activité aux paramètres structuraux permet d'obtenir un système d'équations qui donne, pour une série chimique donnée et pour une activité définie, une équation de corrélation. L'intérêt essentiel de cette équation est qu'elle doit permettre de déterminer les valeurs des paramètres qui correspondent à une activité maximale et ainsi de prévoir l'activité des molécules qui n'ont pas encore été synthétisées.

2.4 Élaboration de modèles QSAR

Ce travail présente la modélisation et la simulation des molécules; pour ce bute nous travaillons pour déterminer certaine propriété chimique physique, électrique et thermodynamique et la réactivité chimique des molécules.

L'évaluation d'un modèle débute par la recherche du maximum possible des données expérimentales fiables. Ensuite, le développement d'une série de descripteurs qui caractérisent les structures moléculaires des composés de la base de données en vue de les relier à l'activité/propriété expérimentale étudiée. Une fois développé, le modèle doit être validé en termes de corrélation (sur le jeu de données d'entraînement). Construire le modèle

QSAR (activité–structure). À fin de valider le modèle QSAR, confirme la validation du modèle par la validation externe (test de validation), et après crée de nouvelles molécules avec d'activité biologique importante et la fin de premières parties de nos résultats. On applique en pareille le ROF Règle Lipinski sur les 28 molécules testées pour vérifier la confirmative des molécules dans le domaine pharmaceutique et dans la production des médicaments pour l'utilisation orale.

Un modèle QSAR relie, d'une manière qualitative ou quantitative, les propriétés (physique et chimique) et l'activité donnée. La stratégie de développement de tels modèles, en respectant les cinq règles qui sont mises en place par l'OCDE [19] (Organisation de Coopération et de Développement Economique) pour la validation des modèles QSAR voir plus loin : les principes OCDE de validité des modèles QSAR [20][21], l'évaluation de chacun des cinq principes est une condition importante afin de proposer des modèles applicables dans le plan expérimental, ce qui était le but de cette thèse.

3 Principes de l'OCDE

Lors du congrès QSAR de Setubal (Portugal) en mars 2002, les lignes directrices pour déterminer la validité de modèle QSAR, en particulier à des fins réglementaires [22] ont été définies. Suite à ce congrès, les membres de l'OCDE ont convenu de 5 principes fondamentaux à suivre pour établir la validité scientifique d'un modèle QSAR.Il est intéressant de noter que des observations similaires ont été proposées par Unger et Hansch en 1973[23]. Ces principes sont un aperçu des points impératifs auxquels doit répondre le modèle pour être considéré comme cohérent, fiable et reproductible [24]. Les cinq principes adoptés par l'OCDE sont les suivants :

i) **Une activité définie** pour s'assurer que les données modélisées soient homogènes (même activité, même unité et dans la mesure du possible même protocole expérimental).

ii) **Un algorithme non ambigu**, les méthodes statistiques utilisées pour construire un modèle QSAR doivent être explicitement détaillées dans la mesure du possible, afin d'assurer la reproductibilité des prédictions.

iii) **Un domaine d'applicabilité défini**, les modèles sont construits sur des sous-ensembles spécifiques de l'espace chimique. Des prédictions peu fiables peuvent être obtenues pour des molécules qui n'appartiennent pas au sous-espace chimique couvert par le modèle.

iv) **Des mesures appropriées de la qualité**, de la robustesse et de la prédictivité du modèle et les performances des modèles doivent être évaluées à l'aide de métriques détaillées suite à une validation interne puis une validation externe. La validation interne consiste à estimer la qualité statistique du modèle sur le jeu d'apprentissage (jeu de données utilisées pour créer le modèle). La validation externe consiste à estimer la capacité du modèle à prédire de nouvelles molécules (pouvoir prédictif) à l'aide d'un jeu de test.

v) **Une interprétabilité du modèle (si possible)**, le modèle doit être interprétable chimiquement, c'est-à-dire qu'il doit permettre d'expliquer l'importance de chaque descripteur moléculaire sur la propriété modélisée, afin de définir des règles génériques. Le respect de ce principe n'est pas toujours évident à mettre en œuvre, car certaines méthodes d'apprentissages (algorithmes) ainsi que des descripteurs peu explicatifs, comme les empreintes moléculaires, ne permettent pas une interprétation facile. Toutefois, si ce dernier principe n'est pas respecté, un modèle statistique disposant de bonnes performances peut tout de même être utilisé s'il respecte les principes précédemment énoncés.

La mise en place d'un modèle de prédiction nécessite avant tout une étape de collecte de données. Le choix de la base de données expérimentale initiale est une étape critique pour le développement de modèle QSAR. Quelle que soit son origine, il arrive qu'un échantillon ne soit pas pur, mais corresponde à un mélange racémique. Le résultat du test d'un tel échantillon pose problème : il est impossible de savoir quelle est la contribution de chaque énantiomère dans l'activité observée. Les structures dont la propriété étudiée est mesurée sur un mélange racémique ne peuvent pas être utilisées dans l'étude de QSAR [21]**.**

Pour être de qualité, une base de données doit être composée de données expérimentales fiables, puisque les barres d'erreurs sur celles-ci se propageront dans le modèle final. Il est donc important de choisir des données présentant de faibles incertitudes afin de limiter les barres d'erreur expérimentales.

D'autre part, l'homogénéité des données est fondamentale. Si l'on veut comparer l'activité/propriété d'une série de molécules, il faut s'assurer, si cela est possible, qu'elle est le résultat de leur interaction avec une seule et même cible et plus précisément avec le même site actif, et l'activité doit être mesurée par un seul et même test, avec des conditions expérimentales identiques pour chaque molécule.

En fin, la diversité des structures est un facteur important dans la qualité des modèles construits, elle définit l'espace chimique que l'analyse va couvrir.

4 Activités ciblées dans ce travail

Les activités biologiques ou les propriétés physicochimiques des molécules peuvent être exprimées de manière quantitative par des chiffres (valeurs numériques).

Dans ce travail, nous avons présenté les études suivantes qui sont réalisées dans notre parcours de thèse :

- Modèle de l'activité anticancéreuse (cancer de l'ovaire, cancer du poumon).

- Création des nouvelles molécules thérapeutiques.

5 Paramètres biologiques

Les données biologiques sont habituellement exprimées sur une échelle logarithmique en raison de la relation linéaire entre la réponse et le logarithme de dose dans la région centrale de la courbe de log dose-réponse. Les logarithmes inverses de l'activité (Log 1/C) sont également utilisés pour obtenir des valeurs mathématiques plus élevées lorsque les structures sont biologiquement très efficaces. Des exemples de données biochimiques ou biologiques, utilisés dans l'analyse de QSAR, sont décrits dans le tableau 1 [25].

Tableau 1 : Paramètres biologiques

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

6 Descripteurs moléculaires

6.1 Définition

Un descripteur est une valeur numérique ou textuelle résultant d'une opération réalisée à partir d'une certaine représentation de la molécule à décrire. Les descripteurs peuvent être regroupés suivant la manière dont ils sont encodés (représentation textuelle, numérique ou vectorielle), suivant le type d'information qu'ils portent (descripteur physico-chimique, topologique, pharmacophorique, etc.), ou suivant la dimensionnalité de la représentation de la molécule à partir de laquelle ils ont été calculés (1D, 2D, 3D)[26].

La structure moléculaire d'un composé contient implicitement toutes ses informations chimiques. Théoriquement, il est possible de définir des données numériques (descripteurs) capables d'extraire une partie des informations chimiques [27].

Depuis plusieurs décennies, des recherches se sont concentrées sur la façon de capturer et de convertir l'information codée dans la structure moléculaire en un ou plusieurs descripteurs. L'intérêt de la communauté scientifique pour les descripteurs moléculaires est attesté par le grand nombre de descripteurs proposés et calculables à l'aide d'outils logiciels

dédiés. Le nombre de descripteurs augmente continuellement avec la complexité croissante des systèmes chimiques étudiés. De ce fait, nous ne présenterons que les descripteurs moléculaires liés aux petites molécules organiques.

6.2 Types de descripteurs

L'importance du nombre des descripteurs (plus de 6000 descripteurs répertoriés [28]) pouvant décrire une molécule rend toute classification ou présentation de ces descripteurs non exhaustive.

Dans ce qui suit, nous allons présenter que les descripteurs moléculaires les plus utilisés sont ceux qui ont été utilisés dans l'ensemble de nos travaux, en commençant par les descripteurs les plus simples, qui nécessitent peu de connaissances sur la structure moléculaire, nous verrons ensuite comment les progrès de la modélisation moléculaire ont permis d'accéder à la structure 3D de la molécule, et de calculer des descripteurs à partir de cette structure.

Historiquement, deux grands schémas pour la classification des descripteurs moléculaires ont été établis : l'un en fonction de leur origine (constitutionnel, topologique, géométrique, quantique, thermodynamique...), et un autre sur leur dimensionnalité (1D, 2D, 3D ou 4D) [29].

6.2.1 Descripteurs 1D

Sont accessibles à partir de la formule brute de la molécule et décrivent des propriétés globales du composé comme le nombre d'atomes et la masse moléculaire, etc. Ces descripteurs sont couramment utilisés du fait de leur extrême simplicité. Cependant, ils peuvent poser problème pour une bonne interprétation des mécanismes d'interaction du fait qu'ils ne permettent pas de tenir en compte des effets stériques et d'isomérie.

Dans nos travaux nous avons utilisé :

- Le poids moléculaire : est appelé aussi le poids de formule, mesuré en daltons (Da). C'est la somme des poids atomiques des différents atomes constituant la molécule. Il est utilisé dans l'étude de transport dont la diffusion et le mode de fonctionnement. Les composés avec des

poids plus élevés sont moins susceptibles d'être absorbés et donc ne peuvent pas atteindre le site d'action. Ainsi, essayer de garder des poids moléculaires aussi bas que possible devrait être l'objectif pour établir un médicament [29] [30]. Pour les médicaments délivrés par voie orale le poids moléculaire doit être inférieur ou égal à 500 daltons (optimum autour de 300 daltons) [31].

- Le pourcentage massique : Est défini par la formule suivante : **(2)**

$$\%\ \boldsymbol{massique} = \frac{la\ masse\ de\ l'élément\ dans\ une\ mole\ du\ composé}{la\ masse\ d'une\ mole\ du\ composé} * 100$$

Équation 2 : Pourcentage massique

Les descripteurs 1D sont faciles à calculer, leurs valeurs sont précises, essentielles et interviennent régulièrement dans les modèles QSAR, mais ils ne permettent pas de distinguer les isomères de constitution et ne permettent pas d'élaborer des modèles plus complexes, c'est-à-dire, si on développe des modèles avec ce type de descripteurs seulement, on aura des problèmes au niveau de l'interprétation des mécanismes d'interaction mis en jeu pour l'activité ou la propriété étudiée [32] . Or, pour la grande majorité des propriétés, la position d'un substituant modifie la valeur de celle-ci, les descripteurs 1D sont, dans de tels cas, défaillants. Il faut alors recourir à d'autres classes de descripteurs.

6.2.2 Descripteurs 2D

Ils sont eux généralement basés sur la topologie de la molécule. Parmi eux, on trouve tout un ensemble de descripteurs capturant diverses informations calculées à partir du graphe de connectivité de la molécule. Les plus connus sont les indices de Kier & Hall [33][34], Randic [35] ou Weiner [36]. De par leur nature, ils permettent de distinguer les molécules plus finement et notamment des molécules cycliques, linéaires, ou par fois même chirales [37] .Également rapides à calculer, ils sont aussi beaucoup utilisés pour décrire et analyser des chimiothèques, ainsi que pour des études de diversité ou de QSAR.

Les descripteurs 2D sont obtenus à partir de la structure plane de la molécule. Dans cette catégorie on trouve principalement les descripteurs topologiques et constitutionnels. Les indices 2D constitutionnels : qui caractérise les différents composants de

la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles, etc.

Les indices topologiques : décrivent les connectivités atomiques dans la molécule. Ce sont des descripteurs plus "sophistiqués" qui n'ont pas forcément un sens chimique évident, mais ils contiennent en leur sein des informations sur la taille globale du système, sa forme globale et ses ramifications [27]. Le principe est de trouver une valeur différente pour chaque squelette moléculaire.

Ces descripteurs sont faciles à calculer, leurs valeurs sont généralement précises, ils interviennent souvent dans les modèles. Ils sont issus de la théorie des graphes développée par Euler en 1736 [38] ; cette théorie est appliquée à la table de connectivité, qui est une représentation compacte de la connectivité interatomique au sein de la molécule.

Un graphe est un ensemble de points, certains reliés par des lignes ; il permet de représenter la topologie de la molécule sans se soucier de la géométrie spatiale exacte de cette dernière [39] .Ces descripteurs 2D permettent de prédire les propriétés physiques, mais sont insuffisants pour expliquer certaines propriétés et activités biologiques comme la toxicité.

-**La superficie de la surface polaire** [40], notée (PSA), en ($Å^2$), est un paramètre très utile pour la prédiction des propriétés du transport des médicaments. Elle est définie comme la somme des surfaces des atomes polaires (habituellement, l'oxygène, l'azote, le soufre, le chlore et l'hydrogène ci-joints) dans une molécule.

- **Surface moléculaire** : c'est une enveloppe entourant les atomes à la périphie, qui explique la surface de contact ligand-Récepteur appelé l'affinité de liaison L-R [41].

- **La surface grille (SAG)** : est calculée par l'hyperChem, c'est une méthode de grille ou une méthode plus rapide plus approximative quelle que soit la zone accessible au solvant ou la Surface de Van der Waals [42].Dans cette théorie, chaque atome de la molécule est représenté par une sphère.la surface extérieure de toutes les sphères atomiques définit la **surface de Van der Waals** .

Les descripteurs 2D sont employés pour obtenir des modèles QSAR plus simples, mais leur défaillance, comme pour les descripteurs 1D, qu'il ne permet pas la bonne interprétation des mécanismes d'interaction mis en jeu pour l'activité/propriété étudiée.

6.2.3 Descripteurs 3D

Ce type de descripteurs nécessitent une conformation 3D de la molécule ; ils sont évalués à partir des positions relatives de leurs atomes dans l'espace et décrivent des caractéristiques plus complexes ; leurs calculs nécessitent donc de connaître, le plus souvent par « modélisation moléculaire empirique » ou « ab-initio », la géométrie 3D de la molécule. La plupart de ces descripteurs s'avèrent relativement 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. On distingue plusieurs familles de descripteurs 3D :

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

(3)

$$\mathbf{MV}=\frac{\boldsymbol{MW}}{\boldsymbol{d}}$$

Équation 3 : Volume moléculaire

Avec : **MW** est le poids moléculaire et **d** la densité.

-**Le nombre de liaisons rotatives** *:* la liaison rotative est définie comme une liaison d'un composé non cyclique, associée à un atome non lourd (qui n'est pas l'hydrogène). Les liaisons CN (amide) ne sont pas considérées en raison de leur barrière d'énergie de rotation élevée. Le nombre de liaisons rotatives, notées NROT, est utilisé pour identifier la flexibilité de la molécule, il a été montré pour être un descripteur de très bonne biodisponibilité orale de médicaments, et pour qu'une structure chimique puisse présenter de bons effets inhibiteurs et être similaire aux médicaments, selon la règle de Lipinski, il faut que le nombre de liaisons rotatives soit inférieur ou égal à 5 [31].

- La surface de Van Der Waals : notée SVDW, est décrite comme résultant de l'ensemble des surfaces atomiques définies par le rayon de Van Der Waals de chaque atome composant la molécule (Figure 3). Plus cette surface est grande et plus importantes sont les possibilités d'interactions.

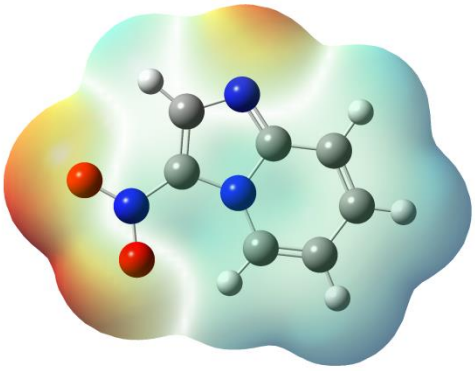


Figure 3 : Surface de Van Der Waals

-Le volume de Van Der Waals : Noté VVDW, est le volume occupé par l'enveloppe de Van Der Waals, ces valeurs numériques dépendent de la méthode de calcul et des rayons de Van Der Waals (RVDW) atomique. Ces derniers déterminent la position la plus favorable d'un atome par rapport à un autre, la distance adéquate où les potentiels répulsifs et attractifs des atomes s'équilibrent. Ils sont particulièrement utilisés pour modéliser comment les molécules organiques "s'approchent" les unes des autres.

6.2.4 Descripteurs physico-chimiques

Les descripteurs physicochimiques (ou indices physicochimiques) certains d'entre eux reflètent la composition moléculaire du composé (le nombre et le type d'atomes et de liaisons présents dans la molécule, le nombre de cycles, les propriétés donneur/accepteur de liaison H, cation, anion, etc.…) [43]. D'autres représentent le caractère hydrophile ou lipophile de la molécule généralement évalué à partir du coefficient de partage Octanol/eau représenté par le *log P* [44] **.** Parmi ceux que nous avons utilisés dans nos travaux, on trouve :

-Le coefficient de partage Octanol/Eau : Noté (*Log P*), qui mesure la solubilité différentielle d'un soluté dans ces deux solvants non miscibles [45]. C'est une mesure importante pour l'identification de la similarité médicamenteuse, selon la règle de Lipinski,

les médicaments délivrés par voie orale doivent avoir des valeurs de Log P supérieures ou égales à -2 et inférieures ou égales à 5[31]. Il est défini par la formule suivante :

(4)

$$\boldsymbol{Log\ P = log([Solut\acute{e}]_{oct})/([Solut\acute{e}]_{eau})}$$

Équation 4 : Coefficient de partage

$[Soluté]_{oct}$ et $[Soluté]_{eau}$ sont les concentrations du soluté dans l'Octanol et l'eau.

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

Le coefficient de partage est largement utilisé dans des études de relations structure activité quantitative (QSARs) dans les sciences pharmaceutiques, biochimiques, toxicologiques et dans les sciences de l'environnement. La lipophilie intéresse donc tout autant la communauté qui étudie les problèmes de santé humaine que celle qui est impliquée dans les problèmes de l'environnement.

-**Énergie d'hydratation (HE) :** l'hydratation est la formation d'une solution implique l'interaction du soluté avec des molécules de solvant, différents liquides peuvent être utilisés comme solvants, mais l'eau est le solvant le plus couramment utilisé.

La liaison hydrogène joue un rôle primordial dans la solubilité des molécules médicamenteuses et leurs interactions avec les récepteurs biologiques [46].Dans les milieux, les molécules polaires ne s'entourent pas des molécules d'eau, ce qui fait apparaitre des liaisons hydrogène entre eux; évidement, les sites donneurs de porton interagissent avec l'atome d'oxygène de l'eau et les sites accepteurs de protons avec l'atome d'hydrogène.

L'oxygène (O) est appelé l'accepteur (accepteur de proton H+) et l'azote (N)ou le donneur (donneur de proton H^+) présenté dans la figure 4.

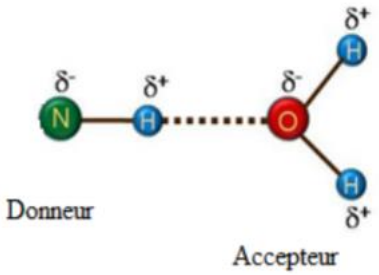


Figure 4 : Liaison hydrogène

L'énergie d'hydratation est un facteur déterminant de la stabilité des différentes conformations moléculaires dans les solutions aqueuses[47].la calcul d'énergie d'hydratation est basé sur la surface exposée qui dépend du type d'atome des groupements moléculaires qui peuvent être donneurs des liaisons hydrogène tels que :O-H,N-H,P-H……Ou bien des groupements accepteurs qui portent des doublets libres tels que :O,N,S,P.

-**La réfractivité moléculaire :** Notée (MR), en m^3/mol, est le volume de la substance absorbée par mole de cette substance. Elle est définie par Lorentz-Lorenz [48] par la formule suivante :

(5)

$$\mathbf{MR} = \frac{n^2-1}{n^2+2}\frac{MW}{d} = \frac{n^2-1}{n^2+2}MV$$

Équation 5 : Réfractivité moléculaire

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

La réfractivité moléculaire est également proportionnelle à la polarisabilité P, par la relation suivante [49] :

$$MR = 4/3\pi\ NA\ P$$

Où : NA est le nombre d'Avogadro qui est, le nombre de molécules dans une mole de substance, NA = 6,022 10^{23}.

-L'indice de réfraction : noté n, est défini par la formule de Lorentz suivante [48] :

(6)

$$n = \sqrt{\frac{2MR + MW}{MV - MR}}$$

Équation 6 : Indice de réfraction

-La polarisabilité : notée (P), en (m^3), est l'aptitude à la déformation du nuage électronique de la molécule sous l'influence d'un champ électrique uniforme. C'est l'un des paramètres qui traduisent les propriétés moléculaires liées à l'hydrophobie et par conséquent aux activités biologiques [50][51]. Elle est calculée à partir de la réfractivité molaire ou du volume molaire comme suit :

(7)

$$P = 0.3964308 \times MR = 0.3964308 \times \frac{n^2 - 1}{n^2 + 2} MV$$

Équation 7 : Polarisabilité

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

(8)

$$d = \frac{MW}{MV}$$

Équation 8 : Densité

L'augmentation de la pression augmente la densité, alors que l'augmentation de la température diminue généralement la densité, mais il y a des exceptions (par exemple, l'eau).

- Le nombre de donneurs de liaisons hydrogène : noté (NHD), calcule le nombre de donneurs de liaison hydrogène dans la molécule. Il s'agit du nombre d'atomes possédant une case quantique vide et contenant un hydrogène acide, c'est-à-dire un atome d'hydrogène lié à un hétéroatome (comme dans les amines, alcools, thiols).

- **Le nombre d'accepteurs de liaisons hydrogène** : noté (NHA), calcule le nombre d'accepteurs de liaison hydrogène dans la molécule. Il s'agit du nombre d'atomes possédant des doublets non liants (azote, oxygène ou fluor) et capables de se lier par liaisons hydrogène à d'autres molécules.

6.2.5 Descripteurs quantiques/électroniques

Ces descripteurs caractérisent la distribution de charge des molécules (polarité des molécules) mais aussi les paramètres de la chimie quantique qui, pour être calculés de manière fiable, nécessitent des calculs plus sophistiqués.

Les approches de la chimie quantique nous donnent accès à des informations supplémentaires telles que des données structurales, énergétiques, électroniques et spectroscopiques des systèmes étudiés [52]**.**

Les structures étudiées dans ce travail ont été optimisées en utilisant la base 6-31G de la fonctionnelle B3LYP qui est une sorte de la méthode de théorie de la fonctionnelle de la densité DFT, détaillée dans l'annexe. Le calcul des descripteurs commence par le dessin des molécules dans le logiciel GaussView 05 [53] puis l'ouverture de ces structures dans le programme Gaussian 09 et ensuite l'exécution de l'optimisation (les calculs). À la fin de ces calculs, des propriétés électroniques seront obtenues.

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

-L'énergie totale

Pour une molécule isolée à l'état fondamental, l'énergie totale calculée, notée et, mesurée en eV, peut être utilisée comme descripteur moléculaire quantique. Cette énergie approximative a été calculée pour une conformation optimisée de la géométrie la plus stable dont la structure

d'énergie est minimale. Les expressions de l'énergie totale de l'état fondamental d'un système sont décrites en détail dans l'annexe.

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

- **Les énergies des orbitales frontières** jouent un rôle majeur dans de nombreuses réactions chimique et dans les mécanismes réactionnels. Les énergies de ces orbitales sont des paramètres très populaires dans la chimie quantique et dans les études QSAR [54] [55]:

- $\mathbf{E_{HOMO}}$: notée E_{HOMO}, mesurée en eV, est l'énergie de l'orbitale moléculaire la plus haute occupée, se réfère à l'aptitude électrodonneur de la molécule. Plus l'énergie de cette OM est élevée, plus la molécule cédera facilement des électrons, il est directement lié au potentiel d'ionisation. Lorsqu'une molécule agit comme une base de Lewis (un doublet d'électrons donneur) dans la formation d'une liaison, les électrons sont alimentés à partir de cette orbite. Il mesure la nucléophilie d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par des électrophiles [56][57]**.**

-$\mathbf{E_{LUMO}}$: notée E_{LUMO}, mesurée en eV, est l'énergie de l'orbitale moléculaire la plus basse inoccupée, se réfère à l'aptitude électro-accepteur de la molécule. Plus l'énergie de cette OM est faible, plus la molécule acceptera facilement des électrons, il est directement lié à l'affinité d'électron. Lorsqu'une molécule agit comme un acide de Lewis (un doublet d'électrons accepteur) dans la formation de liaisons, des doublets d'électrons entrants sont reçus dans cette orbite. Il mesure l'électrophilicité d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par les nucléophiles [57]**.**

- Gap énergétique :

Dans le cas des molécules organiques conjuguées, les niveaux HOMO et LUMO se rapprochent, entre ces deux bandes se trouve un interval d'énergie dans lequel un porteur de charge ne peut pas se retrouver, il s'agit d'une bande interdite. L'interval d'énergie entre les deux bandes est appelé gap. Plus l'énergie Egap d'une molécule est plus faible plus la conductivité est importante. L'énergie Egap est calculée par la formule suivante :

(9)

$$\mathbf{Egap = E_{LUMO} - E_{HOMO}}$$

Équation 9 : Gap énergétique

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 [58][59]**.** L'écart HOMO-LUMO a également été utilisé comme une approximation de la plus faible énergie d'excitation de la molécule [60] **.**

-Potentiel d'ionisation I :

Le potentiel d'ionisation ou l'énergie d'ionisation d'un atome ou d'une molécule est l'énergie qu'il faut fournir à un atome neutre pour arracher un électron et former un ion positif, il est calculé par la formule [57]:

(10)

$$\mathbf{I = - E_{HOMO}}$$

Équation 10 : Potentiel d'ionisation I

C'est une grandeur qui est toujours positive, ce qui signifie qu'il faut fournir de l'énergie à un atome pour lui arracher un électron.

-Affinité électronique (A) :

L'affinité électronique se réfère à l'aptitude d'un atome neutre ou molécule à capter un électron supplémentaire, elle est calculée par la formule [57][61]:

(11)

$$\mathbf{A = - E_{LUMO}}$$

Équation 11 : Affinité électronique

-Potentiel chimique électronique PCE :

Le potentiel chimique sert à déterminer le sens du transfert d'électrons lors d'une condensation entre deux molécules, en effet : soient deux molécules A et B, si le potentiel chimique de la molécule A est supérieur à celui de la molécule B, alors ceci implique que le transfert d'électrons aura lieu de la molécule A vers la molécule B et vice versa ; il est calculé par la formule [62][61]:

(12)

$$\mathbf{PCE = (E_{HOMO} + E_{LUMO}) / 2}$$

Équation 12 : Potentiel chimique électronique

- Électronégativité : notée χ, mesurée en eV, est 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***[63][64]** :

(13)

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

Équation 13 : Électronégativité

Avec : $\mathbf{E=E[N, v(r)]}$

Et aussi

$$\chi = (I+A) / 2.$$

Avec :

I : potentiel d'ionisation de la molécule

A : affinité électronique de la molécule.

- La dureté et la mollesse : notée η, et son inverse la mollesse, notée S, peuvent être obtenues à partie de la première dérivée du potentiel chimique [65][66]: **(14)**

$$\eta = -\left(\frac{\partial E}{\partial N}\right)_{v(r)} = \left(\frac{\partial^2 E}{\partial N^2}\right)_{v(r)} = \frac{1}{S} = \frac{(E_{LUMO}+E_{HOMO})}{2}$$

Équation 14 : Dureté et la mollesse

$$\mu = \left(\frac{\partial E}{\partial N}\right)_{v(r)} = -\frac{I+A}{2} = -\chi$$

Où : I est le potentiel d'ionisation et A est l'affinité électronique

La dureté donne une idée sur la durée relative d'une molécule à conserver les électrons dans son environnement, en effet : soient deux molécules A et B, si la dureté de la molécule A est inférieure à celle de la molécule B, alors ceci signifie que la molécule A conservé peu les électrons dans son environnement par rapport à la molécule B, elle est calculée par la formule [67] [59]:

η = ELUMO – EHOMO

-L'indice d'électrophilicité : notée ω, utilisée pour caractériser la capacité d'une molécule à engendrer un transfert d'électron, elle est calculée selon la formule suivante [69] **:**

(15)

$$\boldsymbol{W = \frac{\chi^2}{2\eta}}$$

Équation 15 : Indice d'électrophilicité

- La charge négative totale : notée TNC, est la somme de toutes les charges négatives des atomes dans une molécule, elle est calculée selon la formule suivante :

(16)

$$\boldsymbol{- TNC = \sum_i \overline{q}_i}$$

Équation 16 : Charge négative totale

Où : $\overline{q}_i$sont les charges négatives atomiques nettes.

-Charges de Mulliken :

L'analyse de population de Mulliken (MPA) est le schéma de population plus utilisé et la plus triviale [69]. Il permet de partitionner la densité électronique de façon arbitraire sur les atomes, en considérant que toute la densité électronique Pμμ d'une orbitale $Ø_μ$ est assignée à l'atome sur laquelle elle est localisée. Le reste de la densité, associée à la population de recouvrement Pμϑ, est également partagée entre les deux atomes sur lesquelles les orbitaux Qμ et Qϑ sont situés. La charge nette de l'atome A est calculé par l'équation suivante :

(17)

$$\mathbf{q_A = Z_A - \sum_{\mu=1;\mu sur A}^{k} P_{\mu\mu} - \sum_{\mu=1;\mu\ sur\ A}^{K} \sum_{\vartheta=1;\vartheta\neq\mu}^{K} P_{\mu\vartheta} S_{\mu\vartheta}}$$

Équation 17 : Charges de Mulliken

Avec Z_A : Charge nucléaire Z_A

λ max : La longueur d'onde du maximum d'absorption et l'énergie d'activation : C'est la longueur qui correspond au maximum d'absorption des radiations. La longueur d'onde du maximum d'absorption des molécules organiques varie en fonction de la longueur de la chaine de conjugaison de la molécule [70].

F : La force d'oscillation, notée, F est la probabilité pour que la transition électronique soit permise. Soit E le champ électrique généré par la lumière. En l'absence de lumière, l'électron se trouve à une distance moyenne r du centre de gravité électrique de la molécule.

Quand on applique le champ E, une force eE va déplacer l'électron de δr. La nouvelle distribution électronique est associée à un moment dipolaire induit appeler dipôle de transition :

$$\mu = -e\,\delta r$$

La force d'oscillateur f est définie par la formule suivante [71]:

(18)

$$f = \frac{4\pi\upsilon m}{3\hbar e^2}|\mu|^2$$

Équation 18 : Force d'oscillateur

Avec :

μ: moment dipolaire induit ou dipôle de transition

υ : est la fréquence de résonnance

m : la masse de l'électron

ħ = h/2π ; avec h : Constante de Planck

e : la charge élémentaire

-L'énergie d'excitation E : c'est l'énergie nécessaire pour faire passer un électron d'un niveau d'énergie stable à un niveau élevé, cette énergie qui est inversement

proportionnelle à la longueur d'onde du maximum d'absorption λmax, est calculée en utilisant la formule [71]:

(19)

$$E = h\,c / \lambda_{max}$$

Équation 19 : Énergie d'excitation E

Où h : la constante de Planck

c : la célérité de lumière

λmax : la longueur d'onde du maximum d'absorption

-L'indice de nucléophilie globale N : le caractère nucléophile d'une molécule peut être lié à son aptitude à négliger sa densité électronique. Aux valeurs élevées de nucléophilie correspondent des valeurs faibles de potentiel d'ionisation et inversement. Domingo est calculé cet indice en utilisant les énergies HOMO obtenues par la méthode de *Kohn – Sham,* comme suit[61][62]:

(20)

$$N = E_{HOMO} - E_{HOMO}\,(TCE)$$

Équation 20 : Indice de nucléophilie globale N

TCE : Tétra-cyanoéthylène

6.2.6 Descripteurs thermodynamiques

Ce sont des descripteurs peu utilisés dans les études QSAR. Ils peuvent être exprimés par la fonction de partition Q de la molécule utilisée en thermodynamique statistique ainsi que de ses dérivées [72][73]**.** Cette fonction décrit la façon avec laquelle l'énergie d'un système de molécules est répartie parmi les individus moléculaires. Sa valeur dépend du poids moléculaire, de la température, du volume moléculaire, des distances inter nucléaires, des mouvements moléculaires et des forces intermoléculaires. La fonction de partition est le point le plus commode entre les propriétés microscopiques des molécules indépendantes (niveaux

d'énergie, moments d'inertie) avec les propriétés macroscopiques (point de fusion, point d'ébullition, entropie). L'expression de cette fonction s'écrit : **(21)**

$$Q=Q_{éle}*Q_{trans}*Q_{rot}*Q_{vibr}$$

Équation 21 : Fonction de partition Q de la molécule

Avec :

$Q_{éle} = \sum_i g_i \exp(-\frac{\varepsilon_i}{KT})$Fonction de partition électronique ;

$Q_{vibr}=\prod_i(1 - \exp\left(-\frac{h\vartheta_i}{KT}\right))$ Fonction de partition de vibration ;

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

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

T est la température en °K, k est la constante de Boltzmann k=1.38 10^{-23} J/K, gi représente la dégénérescence du niveau d'énergie, h est la constante de Planck h = 6.62 10^{-34} J.s, les fréquences de vibration de la molécule, est le degré de symétrie, A, B et C sont les trois moments d'inertie par rapport aux axes x, y et z, m la masse de la particule, V le volume de la molécule.

Les descripteurs thermodynamiques que nous avons utilisés dans nos travaux sont :

-**Le point d'ébullition** : en K, est la température à laquelle les phases liquide et gazeuse d'une substance pure sont en équilibre à une pression donnée, c'est la température à laquelle la substance change d'état du liquide au gaz à une pression donnée. Le point d'ébullition normal est le point d'ébullition à la pression atmosphérique normale (1,013 10^{05} Pa).

En termes d'interactions intermoléculaires, le point d'ébullition représente la température à laquelle les molécules possèdent l'énergie thermique suffisante pour surmonter les attractions intermoléculaires liant les molécules dans le liquide (par exemple des liaisons hydrogène, les attractions dipôle-dipôle...).

Le point d'ébullition d'un composé pur augmente avec la taille, la ramification de la molécule, et avec la présence des liaisons hydrogène et des interactions dipôle-dipôle.

-La constante de Henry : notée KH, est issue de la loi de Henry qui établit une relation entre la pression partielle pi d'un corps pur gazeux et sa concentration c dans un solvant KH=Pi/c. La constante de Henry traduit la volatilité de la molécule. La constante de Henry dépend du soluté, du solvant, et de la température. Une molécule est considérée comme volatile si sa constante est supérieure à 1.10^{-5} $Pa.m^3.mol^{-1}$.

- La température critique : notée Tc, est la température au-dessus de laquelle les phases liquide et gazeuse d'une substance n'existent pas, autrement dit, la température au-dessus de laquelle un gaz ne peut être liquéfié par une augmentation de la pression. Lorsqu'on rapproche de la température critique, les propriétés des phases gazeuse et liquide deviennent les mêmes et se transforment en une seule phase fluide [74].

-La pression critique : notée Pc, est la pression de vapeur à la température critique et au volume critique, c'est la pression minimale pour liquéfier un gaz à la température critique [75].

-La chaleur de formation (appelé aussi l'enthalpie) : notée H, mesurée en KJ/mol, est l'énergie résultant de la formation d'une mole d'une substance à partir de ses éléments constitutifs à l'état standard (T à 298.15 °K et P à 1 atm).

-L'énergie libre de Gibbs : mesurée en KJ/mole est définie par la formule suivante :

(22)

$$G(p, T) = U + p V - TS \text{ ou encore : } G(p, T) = H - TS$$

Équation 22 : L'énergie libre de Gibbs

Avec : U l'énergie interne (en J) ; p est la pression (en Pa) ; V est le volume (en m^3) ; T est la température (en °K) ; S est l'entropie (en J/°K) ; H est l'enthalpie (en J).

-Le point de fusion : (Melting Point) [76]**,** est la température à laquelle la substance passe de l'état solide à liquide à la pression atmosphérique normale. La taille de la molécule et de sa symétrie augmente habituellement le point de fusion; cependant, contrairement au point

d'ébullition, le point de fusion est relativement insensible à la pression. Les points de fusion sont souvent utilisés pour caractériser la pureté des composés organiques. Le point de fusion d'une substance pure est toujours supérieur au point de fusion de cette substance quand une petite quantité d'impureté est présente. Il est utilisé pour prédire la solubilité des composés. La formule pour calculer le point de fusion est :**(23)**

$$\mathbf{T}=\frac{\Delta \boldsymbol{H}}{\Delta \boldsymbol{S}}$$

Équation 23 : Point de fusion

Où : T est la température au point de fusion en °K ;

$\Delta \boldsymbol{S}$: est la variation d'entropie en J/°K ;

$\Delta \boldsymbol{H}$: est la variation d'enthalpie en J.

-Correction au point –zéro : l'énergie de vibration d'une molécule à 0K, phénomène quantique, n'est pas nulle. Par conséquent, une fois le point stationnaire localisé (minimum ou état de transition), son énergie est inférieure à l'énergie réelle de la molécule. L'énergie doit être corrigée en ajoutant à la valeur obtenue l'énergie du niveau vibrationnel le plus bas ou ZPVE (« Zéro Point Vibrational Energy ») définie comme : **(24)**

$$\boldsymbol{ZPVE} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{\mathbf{1}}{\mathbf{2}} \boldsymbol{h\nu_i}$$

Équation 24 : Correction au point –zéro

Ou h est la constante de Plank et $\boldsymbol{\nu_i}$ les n fréquences des modes normaux de vibration .La valeur de ZPVE est obtenue par un calcul des fréquences dans l'approximation harmonique.

-Variation d'énergie interne à T : la variation d'énergie interne ΔU(0) de la réaction à 0 K est obtenue en ajoutant la variation d'énergie de vibration au point zéro à l'énergie de la réaction Eél : **(25)**

$$\mathbf{\Delta U(0)=E_{él}+\Delta ZPVE}$$

Équation 25 : Variation d'énergie interne à T

Pour déterminer la variation d'énergie interne à la température T à partir de l'énergie de réaction Eél à 0K, il faut y ajouter un terme ΔE_{th} que nous nommerons variation d'énergie thermique à T. **(26)**

$$\Delta U(T) = Eél + \Delta E_{th}$$

Équation 26 : Variation d'énergie thermique à T

Avec $\Delta E_{th}(T) = E_{th}(T)(prod) - E_{th}(T)(réact)$

Tableau 2 : Liste des descripteurs utilises dans notre travail

Descripteurs	Abréviation	Type
- Le poids moléculaire	**MW**	**Constitutionnel**
- Le coefficient de partage Octanol/Eau	***Log P***	**Physico-chimique**
- La réfractivité molaire	**MR**	
- La densité	**D**	
- Le nombre de donneurs de liaisons H	**NHA**	
- Le nombre d'accepteurs de liaisons H	**NHD**	
- Le volume moléculaire	**MV**	**Géométrique**
- L'énergie totale	**Et**	**Quantique**
- L'énergie HOMO	$\mathbf{E_{HOMO}}$	
- L'énergie LUMO	$\mathbf{E_{LUMO}}$	
- Le Gap énergétique	$\mathbf{E_{Gap}}$	
- Le moment dipolaire	**μ**	
- La dureté	**η**	
- La mollesse	**S**	
- L'électronégativité	**χ**	
- L'indice d'électrophilicité	**ω**	

6.2.7 Descripteurs 4-D

Ils correspondent à la mesure des propriétés 3D (potentiel électrostatique, d'hydrophobicité, de liaison hydrogène...) d'une molécule en tout point de l'espace. Ils permettent d'avoir l'information sur la structure de la cible (protéine). On pourra ainsi distinguer les descripteurs 4D qui nécessitent un alignement de la molécule guidé par l'étude des complexes ligand-cible (ou, au moins, par des contraintes visant d'optimiser le recouvrement spatial des champs électriques et stériques des ligands, faute d'information sur le vrai mode de fixation dans la cible) avant d'être calculés. Ces descripteurs sont obtenus par le calcul des champs d'interactions moléculaires (CoMFA, CoMSIA) entre une molécule et une sonde représentée par une autre molécule (eau, amide, ...) [77][32][78][79].

6.3 Sélection des descripteurs

Un grand nombre de descripteurs différents sont collectés pour la modélisation d'une grandeur donnée (activité ou propriété), car les facteurs déterminants du processus étudié ne sont *a priori* pas connus. Cependant, les descripteurs envisagés n'ont pas tous une influence significative sur la grandeur modélisée, et les variables ne sont pas toujours indépendantes. De plus, le nombre de descripteurs, c'est-à-dire la dimension de la base de données d'entrée, détermine la dimension du vecteur des paramètres à ajuster. Si cette dimension est trop importante par rapport au nombre des observations (molécules) de la base d'apprentissage, le modèle risque d'être sur-ajusté à ces exemples, incapable de prédire la grandeur modélisée sur de nouvelles molécules et peut contenir des informations redondantes. Les descripteurs moléculaires employés doivent être porteurs de sens et interprétables d'un point de vue chimique. Et par conséquent, lorsque les descripteurs sélectionnés sont pertinents, ils offrent des idées sur les mécanismes, et les modèles QSAR seront simples, transparents et compréhensibles [80].

Il ne s'agit d'utiliser alors que le minimum de descripteurs pour expliquer la propriété ciblée, mais il ne faut pas avoir de perdre d'information. Comme Einstein a dit : "Tout devrait être fait aussi simple que possible, mais pas plus simple.". Dans une modélisation QSAR, ce principe d'Einstein signifie que le modèle doit avoir le moins de paramètres possible tout en traduisant au mieux l'information contenue dans la propriété [81].

Avant d'entamer le développement effectif des équations de régression QSAR, il est hautement recommandé d'examiner la qualité statistique des données de départ, à la fois les données à corréler (variable dépendante) et les descripteurs utilisés dans la corrélation (variables indépendantes).

On distingue habituellement dans un tel prétraitement des données les analyses univariées des analyses bivariées [82][83]**.** Dans l'analyse univariée, il est recommandé de vérifier la conformité des données à la distribution normale. Une précaution particulière doit être prise lors de la procédure de régression subséquente si les valeurs de la propriété étudiée, ou d'un descripteur, ne suivent pas la loi de Laplace-Gauss [84]**.**

Pour un ensemble de descripteurs différents, il est nécessaire d'effectuer une analyse des données bivariée, c'est-à-dire de calculer le coefficient de corrélation entre chacune des paires de l'ensemble des descripteurs. Si ce coefficient est statistiquement significatif (R>0,95), ces deux descripteurs sont considérés comme fortement corrélés et ne peuvent être utilisés simultanément lors de l'analyse QSAR [92] et en pratique, ils seront alors enlevés dans le procédé de sélection. Ce type d'analyse est appelé l'analyse objective qui permet de réduire le nombre de descripteurs sans faire participer la variable dépendante (l'activité ou la propriété).

Les méthodes statistiques, qu'on discutera dans la partie suivante de ce chapitre, sont aussi utilisées pour l'élimination des descripteurs qui n'interviennent pas dans les modèles proposés, c'est-à-dire qui n'influencent pas l'activité ou la propriété étudiée dans ce qu'on appelle l'analyse subjective (spécialement l'analyse en composantes principales et la régression linéaire descendante).

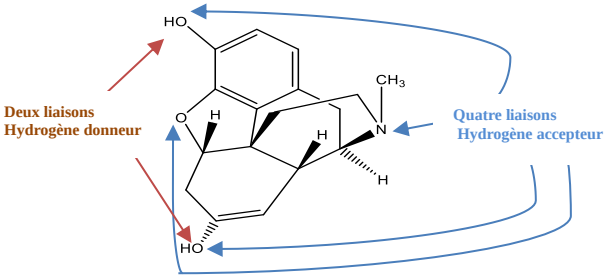
Finalement, pour que les relations QSAR ne soient pas statistiquement non significatives ou en cas d'erreur ponctuelles, il faut que le rapport composés/descripteurs doive être supérieur à 5 [86] [87]**.**

7 Règles de Lipinski

En 1997 Lipinski est proposé des règles simples permettant d'identifier rapidement et à grande échelle des molécules à caractère « drug-like », plus susceptibles de présenter les

caractéristiques de biodisponibilité nécessaires au développement d'un candidat médicament. [31] Ces règles, communément appelées « règles de Lipinski » ou « la règle de 5 », comportent quatre critères physico-chimiques qui décrivent la molécule : poids moléculaire ≤ 500 Da, *LogP* ≤ 5, nombre d'accepteurs de liaisons hydrogène ≤ 10 et nombre de donneurs de liaisons hydrogène ≤ 5. Particulièrement, la mesure du *LogP* caractérise la polarité du composé (estimée par le coefficient de partition octanol/eau), permettant ainsi d'estimer la distribution du composé dans l'organisme. Les molécules hydrophobes (hautes valeurs du *LogP*) sont principalement distribuées dans les régions hydrophobes, comme la bicouche lipidique des cellules. Inversement, les molécules hydrophiles sont retrouvées principalement dans des régions aqueuses, comme le sérum sanguin. La 5e règle de Lipinski stipule que les adjuvants et assimilés font exception aux quatre autres règles.

D'après ces règles, déterminées à partir de 2245 molécules extraites du World Drug Index (WDI) et ayant passé avec succès la phase II des tests cliniques, les composés ne validant pas au moins deux des critères suivants auraient de très fortes chances d'avoir des problèmes d'absorption ou de perméabilité intestinale. [31]. Une mauvaise biodisponibilité orale serait également détectée dès lors qu'une de ces règles est violée. Néanmoins, il n'est pas garanti que les composés adhérant à cette règle possèdent une excellente absorption, perméabilité ou biodisponibilité. La morphine, par exemple, satisfait toutes les règles de Lipinski mais présente une biodisponibilité orale modérée (Figure 5). De plus, un médicament qui enfreint une ou plusieurs règles peut tout de même avoir une biodisponibilité satisfaisante.



Morphine ($C_{17}H_{19}O_3N$)

Poids moleculaire =285 DA
Log(P)=1,27

Figure 5 : Structure moléculaire de la morphine avec ses sites d'interaction.

8 Applications des méthodes QSAR

Les applications des méthodes QSAR sont très nombreuses, elles touchent tous les domaines ou la structure chimique intervient, entre autres on peut citer :

Propriétés physico - chimiques :

- Point d'ébullition, point de fusion, densité, indice de réfraction, température critique, viscosité, solubilité, pression de vapeur, tension superficielle, coefficients de partition : eau/octanol, air/eau, huile/air, lait/plasma.

- Activités biologiques :

Anti VIH, Anti malaria, Anti Diabète, Anti Cancer, Anti-oxydant, Anti inflammatoire.

- Autres propriétés/activités :

- Prédiction de la toxicité aquatique des composés chimiques vis-à-vis des espèces environnementales ;

- Toxicité des nanoparticules ;

- Toxicité des pesticides et des colorants ;

- Propriétés inhibitrices de corrosion ;

- Concentration micellaire critique ;

- Prédiction de plusieurs propriétés dangereuses telle que l'explosibilité et l'inflammabilité de certaines familles de molécules chimiques ;

- Conception des médicaments et de nombreux autres produits tels que les agents tensio-actifs, parfums, les colorants et les produits de la chimie fine.

Chapitre II : Méthodes et analyse statistique

1 Intoduction

Elaboration des modèles QSAR n'est pas une chose facile. La première difficulté réside dans la différence d'échelle existant entre les données à corréler. La structure étant à une échelle moléculaire alors que les activités /propriétés à prédire sont à une échelle macroscopique. Un des problèmes importants réside également dans le traitement des données. En fait, de nombreux outils existent et il s'agit juste de trouver le moyen le plus adapté pour obtenir un modèle fiable à partir des données disponibles.

L'analyse de la régression est l'outil le plus utilisé en modélisation QSAR. L'idée est de décrire et d'évaluer la relation entre une variable (dites variable à expliquer ou variable dépendante) souvent notée **Y**, et une (ou plusieurs) variable(s), dite(s) variable(s) explicative(s) ou indépendante (s) notée (s) **X.**

Le présent chapitre revient sur les aspects théoriques des méthodes statistiques et chimiométriques qui ont été utilisées pour la modélisation et pour les études QSAR de l'ensemble des molécules testées. Nous présenterons tout d'abord les méthodes d'analyse exploratoire de données multivariées qui ont été mises en œuvre pour décrire les données. Ensuite, nous reviendrons sur les méthodes prédictives qui ont été développées pour modéliser les propriétés d'intérêt. Les principaux fondements de la régression linière multiple RLM et réseaux de neurones artificiels (RNN), les analyses à composantes principales (ACP).

Il existe une grande diversité d'algorithmes pouvant être appliqués dans le cadre des approches QSAR. Le défi consiste à choisir la méthode d'apprentissage qui convienne le mieux à l'exploration de la propriété en court d'investigation. Pour cela, des méthodes supervisées et non supervisées peuvent être utilisées et seront décrites par la suite (figure 6)

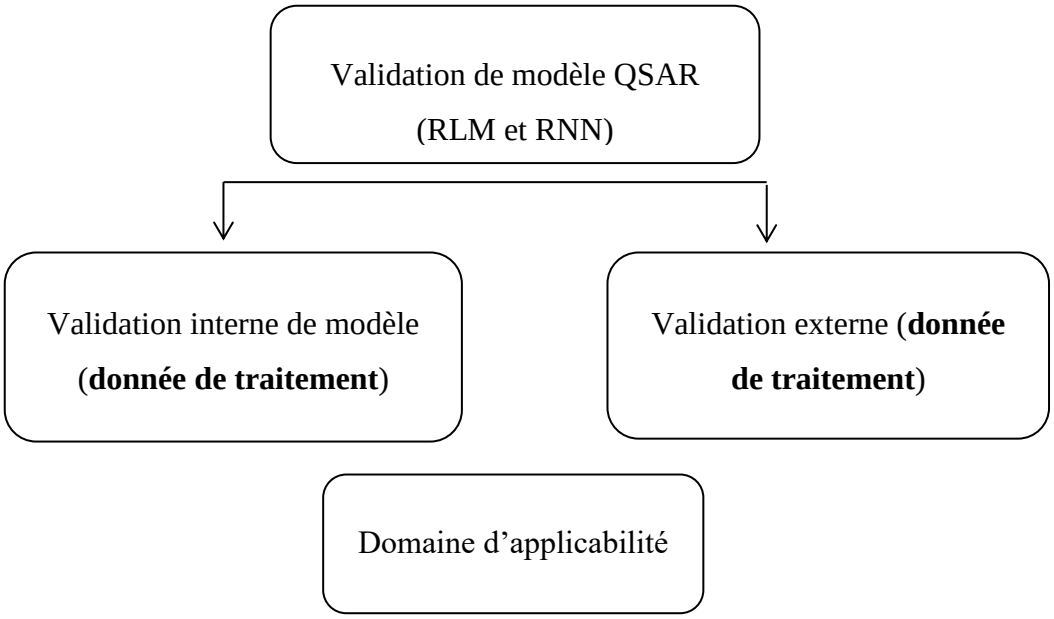


Figure 6 : Techniques statistiques permettant de créer des modèles QSAR

2 Définition des méthodes statistiques

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

2.1 Types des méthodes statistiques

Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées « variables ». Dans notre cas, les objets (ou individus) sont les molécules et les variables sont les descripteurs moléculaires précédemment décrits dans le 1er chapitre.

Après le recueil des descripteurs, la démarche statistique consiste à traiter et interpréter les informations recueillies sur ces molécules. Cette démarche comporte deux grandes classes : la statistique descriptive et la statistique décisionnelle ou prédictive.

2.1.1 Statistique descriptive

La statistique descriptive ou l'analyse des données a pour but d'extraire le maximum d'information contenue dans les données d'une façon efficace, simple et compréhensible. Elle permet de résumer les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour des études plus sophistiquées. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs statistiques. Elle est utilisée aussi pour diviser et classer les données dans des classes homogènes.

Dans l'ensemble de nos travaux, on a principalement utilisé l'analyse en composantes principales (ACP) comme technique pour l'analyse des données, et la méthode du partitionnement en k-moyennes (ou k-means en anglais).

a. Analyse en Composantes Principales (ACP)

L'ACP est une technique de réduction de dimensionnalité qui est largement utilisée pour l'analyse de données. Elle décompose un jeu de données multivariées (plusieurs descripteurs) à l'aide d'un ensemble de composantes orthogonales successives qui expliquent la variance maximale observée dans le jeu de données. Ces composantes orthogonales vont être appelées composantes principales et correspondent à des combinaisons linéaires des descripteurs moléculaires. Chaque composante principale va exprimer une part de la variance expliquée présente dans le jeu de données.

Ainsi, l'ACP définit un espace de plus faible dimensionnalité que le jeu de données initial ce qui permet au modélisateur d'analyser de façon plus aisée les données sur lesquelles il travaille. L'ACP permet alors de visualiser les molécules et les descripteurs sur l'ensemble des plans bidimensionnels définis par les combinaisons de deux composantes (Figure 7), ce qui est plus facilement interprétable.

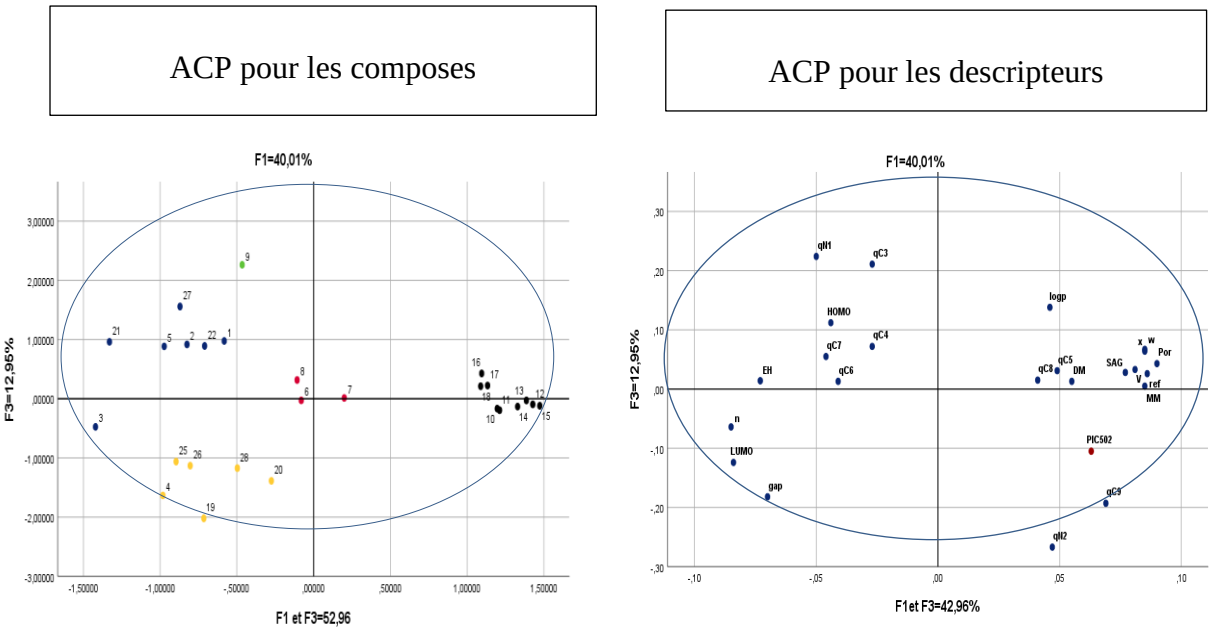


Figure 7 : Représentation schématique d'une ACP sur F3 et F1.

b. Classification des données (données de traitement et données de validation)

- Classification par méthode K-means

K-means regroupement est une méthode bien développée et K-means mise en œuvre dans de nombreux logiciels statistiques dans notre thèse par SPSS, ce qui permet un accès facile, et, par conséquent, elle est choisie comme méthode appropriée pour le sous-ensemble division. Enfin, un tiers des composés de chacun des sous-groupes créés à partir du regroupement de k-means de chaque sous-ensemble sont sélectionnés pour former un ensemble de tests, ainsi que les composés restants composer l'ensemble de formation correspondant (groupe des molécules pour le test et l'autre pour la validation) [88].

L'algorithme de mise en grappes de K-means est décrit en détail par Hartigan (1975) [89]. Une méthode efficace de la version de l'algorithme est présentée ici.

L'objectif de l'algorithme des K moyens est de diviser M points dans N dimensions en K groupes de sorte que la somme des carrés à l'intérieur du groupe soit réduite au minimum. Il n'est pas pratique d'exiger que la solution ait une somme minimale de carrés

contre toutes les partitions, sauf lorsque M et N sont petits et K = 2. Nous recherchons plutôt des optima "locaux", des solutions tel qu'aucun déplacement d'un point ou bien d'un passage d'une grappe à une autre réduira la somme des carrés à l'intérieur de la grappe [89].

2.1.2 Statistique décisionnelle ou prédictive

Dans ce type des statistiques, les probabilités jouent un rôle fondamental. Cette statistique a pour but de prendre des décisions et de faire des prévisions au vu des observations. En général, il faut pour cela proposer des modèles probabilistes du phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Dans notre cas, il faut rechercher une relation approximative entre une activité ou propriété et plusieurs variables quantitatives (descripteurs moléculaires), la forme de cette relation peut être linéaire ou non linéaire.

Dans l'ensemble de nos travaux, on a utilisé la régression linéaire multiple RLM, et modèle de réseau neuronal artificiel (RNN) pour la construction et la validation du modèle QSAR.

Dans ce travail on a appliqué la régression linéaire multiple RLM et RNN. Car la régression linéaire multiple RLM est l'une des méthodes de modélisation les plus populaires grâce à sa simplicité d'utilisation et sa facilité d'interprétation. L'avantage important de la régression linéaire multiple est qu'elle est très transparente, puisque l'algorithme est disponible, et que les prédictions peuvent être réalisées facilement. Dans la plupart de nos travaux, cette méthode a été utilisée aussi pour la sélection des descripteurs moléculaires utilisés dans les autres méthodes statistiques [90]**.**

a. Régression linéaire multiple (RLM)

La régression multiple cherche à approximer une relation trop complexe en général, par une fonction mathématique simple. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante (à expliquer) Y (ici, l'activité) et une série de k variables indépendantes (explicatives) Xi (ici, les descripteurs). L'objectif est d'obtenir une équation de la forme suivante :

(27)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Équation 27 : 1ere formule de la régression linéaire multiple

Cette équation est linéaire par rapport aux paramètres (coefficients de régression), la méthode RLM se base sur l'hypothèse que la propriété Y dépend linéairement des différentes variables (les descripteurs) x_1, x_2, x_3 ……x_n selon la relation :

(28)

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

Équation 28 : 2emme formule de la régression linéaire multiple

La taille de ces coefficients indique le degré d'influence des descripteurs moléculaires correspondants au couple l'activité/propriété cible. Un coefficient positif indique que le descripteur moléculaire correspondant contribue positivement à la cible, tandis qu'un coefficient négatif suggère la contribution négative.

On distingue divers types de RLM, les plus utilisés sont :

- **La RLM progressive ascendante**, qui consiste à incorporer les variables au modèle une à une, en sélectionnant, à chaque étape, la variable dont la corrélation partielle avec la grandeur modélisée est la plus élevée. À l'inverse, lors de RLM progressive descendante, on débute la modélisation avec l'ensemble des descripteurs, en les éliminant un par un jusqu'à obtenir le meilleur jeu de composantes, c'est-à-dire l'obtention d'un modèle valide ayant la bonne corrélation.

- **La RLM pas à pas (Stepwise),** est une combinaison des deux méthodes évoquées précédemment. Les variables sont incorporées une à une dans le modèle, par sélection progressive. Cependant, à chaque étape, on vérifie que les corrélations partielles des variables précédemment introduites sont encore significatives.

b. Modèle de réseau neuronal artificiel (RNN)

Un RNN se compose d'un grand nombre de (nœuds ou unités), qui simulent des éléments de traitement fonctions des neurones biologiques. La figure 8 illustre un modèle simple d'oneneuron au sein d4un réseau artificiel, où une entrée le vecteur passe par le neurone pour fournir une sortie valeur. Un système à plusieurs couches se compose de données d'entrée, de données cachées et les couches de sortie. La théorie et les bases mathématiques des RNN ont été largement décrites dans Haykin en 1999 et Bishop en 1995.

Dans cette étude, une rétropropagation par feed-forward à plusieurs couches d'un réseau à trois couches a été utilisée : en entrée, cachée et en sortie couche. Ce type de réseau offre généralement de meilleures performances par rapport aux autres types[91]. Tan-fonctions de transfert sigmoïde (non linéaire) et transfert linéaire ont été sélectionnés pour les couches cachées et de sortie, respectivement. Utilisation d'une fonction tan-sigmoïde dans la couche cachée permet de ne rapprocher que les relations non linéaires présentes entre les couches d'entrée et de sortie [92]. Le nombre de neurones dans la couche cachée peut être défini à l'aide d'une formule recommandée par Fletcher et Goss en 1993 ou à l'aide d'un essai-erreur comme le suggère Chang en 2004. Le nombre de neurones dans la couche cachée est d'une grande importance, car trop de neurones peuvent causer des problèmes de surajustement[93].

Pour améliorer la généralisation des réseaux, Demuth et Beale en 2004 ont recommandé d'utiliser un réseau qui soit juste assez grand de fournir une adéquation entre le prédicteur et la réponse des variables. Pour éviter les problèmes de surajustement et pour fournir un des moyens efficaces pour mettre fin à la phase de formation qui prend beaucoup de temps, l'approche dite de "l'arrêt anticipé" a été utilisée [94].

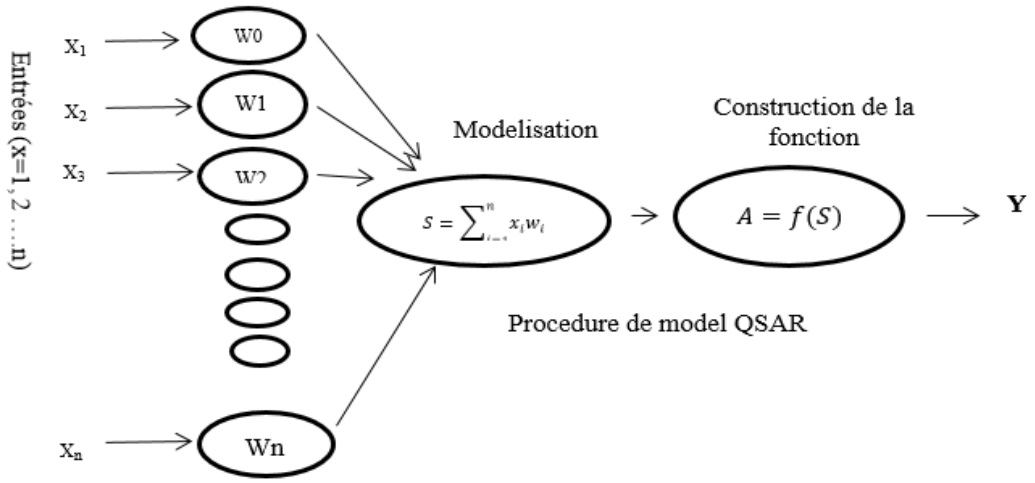


Figure 8 : Base de modèle de réseau neuronal artificiel (RNN)

La couche d'entrée du réseau utilisée dans cette étude concerne la réflexion, tandis que la couche de sortie du réseau se rapporte au activité et les descripteurs. Les techniques de prétraitement des données (centrage ou normalisation à une moyenne de zéro et à un écart-type d'un) et l'analyse en composantes principales (ACP) ont été appliquées aux intrants pour normaliser les données de réflectance et de réduire la dimension les données. Le nombre de neurones utilisé pour la formation des réseaux a été systématiquement varié entre 5 et 13 pour permettre la sélection ultérieure de la taille du réseau la plus appropriée sur la base des résultats de l'ensemble des données de test[95]. Pour la modélisation de la RNN, le logiciel SPSS et la boîte à outils ,des réseaux neuronaux ont été utilisés [94].

À chaque échelle de l'étude, chaque ensemble de données a été divisé en trois sous-ensembles, un pour la formation (la moitié des données d'entrée), un pour validation (un quart) et un autre pour les tests (un quart de l'entrée donnée). L'algorithme de Levenberg-Marquardt [102], qui permet une optimisation rapide, a été utilisé pour les réseaux formation. La performance d'un réseau formé a été évaluée pour comparer l'erreur quadratique moyenne (EQM) et la moyenne des racines d'erreur quadratique moyenne (MEQM) calculée à partir de la formation, de la validation, ainsi que tester des sous-ensembles de données. Seul un sous-ensemble de données de formation est utilisé pour la mise à jour des poids et des biais du réseau. Pendant la formation, l'erreur en ce qui concerne le sous-ensemble de données de validation est surveillé. Lorsque l'erreur de validation augmente pour un nombre d'itérations déterminé, la formation est arrêtée. Erreur concernant le test du sous-ensemble de données

n'est pas contrôlé pendant la formation, mais est quantifiée pour évaluer la performance finale d'un modèle de RNN formé.

3 Techniques de validation

La validation est indispensable et présente différents enjeux pour justifier la qualité d'un modèle QSAR. Dans le cadre de cette partie, nous verrons les phénomènes responsables du non validité d'un modèle QSAR, ainsi que les métriques de performances et les méthodes de validation utilisées pour appréhender et prévenir chacun de ces phénomènes. [96].

3.1 Validation de modèle QSAR

La validation est une étape importante qui permet de vérifier un modèle est-il statistiquement valide et performant, c'est-à-dire est ce qu'il est capable de prédire avec fiabilité l'activité biologique étudiée pour un ensemble de molécules ? L'élaboration d'un modèle prédictif peut être perturbée par différents phénomènes comme le sous-apprentissage, le sur-apprentissage ou encore la corrélation aléatoire. Comme nous l'avons énoncé précédemment, ces phénomènes peuvent être induits par différents facteurs, par exemple la taille du jeu de données, le nombre de descripteurs moléculaires, ou encore l'utilisation d'une méthode d'apprentissage particulière. Afin de pouvoir identifier les conditions les plus propices à l'obtention d'un modèle performant, il est intéressant de définir l'impact de chacun de ces facteurs sur les phénomènes perturbant la création des modèles de prédiction.

Les différentes approches utilisées à cet effet sont décrites ci-dessous :

3.1.1 Coefficients et tests statistiques standards

Afin de déterminer la qualité d'un modèle, différents paramètres statistiques sont employés, tels que les erreurs quadratiques moyennes (*EQM*), les coefficients de corrélation qui sont régulièrement utilisés dans les études QSAR, sont décrites en détail dans cette partie.

- Coefficient de corrélation r

Le coefficient de corrélation, r est une mesure du degré de linéarité de la relation. Il signifie la qualité de l'ajustement du modèle et quantifie la variance dans les

données [97]**.** Dans une situation idéale, le coefficient de corrélation doit être égal à ou proche de 1, mais en réalité, en raison de la complexité des données biologiques, toute valeur supérieure à 0,9 est appréciable. Les coefficients de corrélation pour les variables d'un jeu de données sont compilés dans une matrice de corrélation, qui montre la relation entre un descripteur et un autre. La matrice de corrélation garantit que les variables significatives sont orthogonales les unes aux autres. L'ajout de chaque nouvelle variable au modèle augmente toujours le r, sauf si la nouvelle variable est une constante d'une combinaison linéaire d'autres variables, qui ne produirait aucun effet.

L'augmentation de r provoquée par l'ajout d'une nouvelle variable signifie un ajustement excessif des données.

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

(29)

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

Équation 29 : Coefficient de corrélation r

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

Le coefficient de détermination multiple r^2 : également appelé coefficient de corrélation de Pearson, r^2 est le coefficient de corrélation au carré qui renseigne sur la qualité avec laquelle le modèle reproduit les données expérimentales [97]**.** Il s'agit d'une mesure quantitative de la précision de l'ajustement des valeurs ajustées à celles observées. Plus il se rapproche de l'unité, plus les valeurs ajustées des valeurs ajustées sont similaires à celles des valeurs expérimentales, ce qui suggère que le modèle adapte infailliblement les données. Cependant, un r^2 proche de 1 ne signifie pas que le modèle est parfait; l'ajout de tout nouveau descripteur au modèle induit une augmentation constante de r^2, même si le descripteur ajouté ne contribue pas au modèle. Ainsi, d'autres mesures sont nécessaires pour déterminer la capacité prédictive du modèle.

Le jugement sur la valeur de r ou r² est très subjectif. Bien que ce coefficient soit très facile à comprendre, il faut se garder d'y attacher trop d'importance, car il est loin de fournir un critère suffisant pour juger la qualité d'une régression. Il n'est pas recommandé d'utiliser r² pour comparer des modèles avec un nombre différent de descripteurs, le coefficient r² nous dira toujours de choisir le modèle avec le plus grand nombre de descripteurs, car son r² sera plus important (on projette sur un espace plus grand), même si les variables sont sans effets sur la réponse (l'activité ou la propriété étudiée).

La valeur de r^2 dépend de la taille de l'échantillon et le nombre de variables prédictives dans l'équation. Il garde la même valeur ou augmente lors d'une nouvelle variable de prédiction est ajoutée à l'équation de régression, même si la variable ajoutée ne contribue pas à la réduction de la variance inexpliquée. Par conséquent, un autre paramètre statistique peut être utilisé, appelé r^2 ajusté (r^2adj). Bien entendu, un autre indicateur est **l'erreur quadratique moyenne** (*EQM*), à laquelle est parfois préférée la déviation standard ***s*.**

Le coefficient de détermination ajusté r^2adj :

Ce coefficient est utilisé en régression multiple parce qu'il tient compte du degré de liberté :

(30)

$$r^2adj = \sqrt{\frac{r^2(n-1)-k}{n-k-1}}$$

Équation 30 : Coefficient de détermination ajusté r^2adj

Avec : n est le nombre des observations (les molécules) ; k est le nombre de variables indépendantes (les descripteurs ou paramètres) ; r^2 est le coefficient de détermination du modèle.

- **L'erreur quadratique moyenne « EQM » et l'erreur type résiduel « s »**

(31)

$$\boldsymbol{EQM = \frac{\sum\left|(\hat{y} - y_i)^2\right|}{n}}$$

Équation 31 : Erreur quadratique moyenne

Ou encore, l'erreur type résiduelle « $\boldsymbol{s}$ » :

(32)

$$\mathbf{s = \sqrt{\frac{\sum(\hat{y}_\iota - y_i)^2}{n - k - 1}}}$$

Équation 32 : Erreur type résiduelle

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

Ces paramètres mesurent la variation de l'activité cible non expliquée par le modèle QSAR. En particulier, plus la déviation standard est petite plus la corrélation est meilleure. Sa valeur est toujours fonction de l'unité de mesure de l'activité cible et tient également compte des erreurs expérimentales ce qui explique qu'une valeur trop petite n'a aucune signification.

- **Le facteur d'inflation de la variance VIF**

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

(33)

$$\boldsymbol{VIF = \frac{1}{1 - r_i^2}}$$

Équation 33 : Facteur d'inflation de la variance VIF

Avec : r_i^2 est le coefficient de détermination de la régression de la variable sur les autres variables. Plus x_i est linéairement proche des autres variables, plus r_i^2 est proche de 1 et le VIF est grand. L'avantage du VIF par rapport à la matrice de corrélation est qu'il prend en compte des corrélations multiples.

- Le test de Fisher F

L'indice de Fisher F-test est employé afin de mesurer le niveau de signifiance statistique du modèle à « x% » (le niveau usuel est 95%), c'est-à-dire la qualité du choix du jeu de paramètres. La conclusion obtenue ne doit pas nous faire penser que la corrélation a « x % » de la chance d'être vraie, mais seulement que la corrélation est vérifiée pour « x% » des composés pris pour référence et qu'une abstraction est faite pour les autres.

Hypothèses :

H_0 : les variances des échantillons sont homogènes

H_1 : les variances des échantillons ne sont pas homogènes la valeur à calculer est :

On calcule le F (observé) à partir de la formule :

(34)

$$\boldsymbol{F(observé)} = \frac{\sum(\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \bar{\boldsymbol{y}})^{\boldsymbol{2}}}{\sum(\boldsymbol{y}_{\boldsymbol{i}} - \hat{\boldsymbol{y}}_{\boldsymbol{i}})^{\boldsymbol{2}}} \frac{\boldsymbol{n-k-1}}{\boldsymbol{k}}$$

Équation 34 : (F) observé

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

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

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

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

- Le test de Student

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

(35)

$$|t_i| = \left|\frac{a_i}{s(a_i)}\right| > t_{1-\frac{a}{2}}^{n-k-2}$$

Équation 35 : Coefficient $|t_i|$de test de Student

Avec : t_i est le *t-test* pour le descripteur « i » ; a_i est le coefficient associé au descripteur « i » dans le modèle ; $s(a_i)$ est l'erreur type associée au descripteur « i » ; a est l'intervalle de confiance ; n est le nombre des observations (les molécules) ; k est le nombre de variables indépendantes (les descripteurs).

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

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

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

Sous H_0, la loi de Student à (n-p-1) degré de liberté s'écrit :

(36)

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

Équation 36 : Coefficient t_{calc} de test de Student

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

$$t_{calc} > t_{\left(1-\frac{a}{i}\right),(n-k-1)}$$

(1-a/2)[106][107] .$t_{\left(1-\frac{a}{i}\right),(n-p-1)}$: la valeur de la loi de Student, à (n-k-1) degré de liberté, à une Probabilité

3.1.2 Pouvoir de prévision interne

Afin de déterminer la stabilité prédictive d'un modèle et de tester l'influence de chaque échantillon (composé) sur le modèle final, des procédures de validation croisée (en anglais : cross-validation) sont souvent utilisées [96]**.**

Généralement, ces techniques de validation permettent l'évaluation de la robustesse du modèle, autrement dit la stabilité des paramètres du modèle QSAR vis-à-vis des molécules du jeu d'entraînement. Cela dit, qu'elles ne permettent en aucun cas de démontrer le pouvoir prédictif des modèles [100][99]**.**

Le principe de ces méthodes consiste à extraire un certain nombre de molécules du jeu d'apprentissage et à construire un nouveau modèle avec les molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent). Ce nouveau modèle est alors utilisé pour la phase de prédiction sur les molécules retirées. Ce processus est ensuite répété pour retirer et prédire les valeurs de toutes les molécules du jeu d'entraînement. Le coefficient de corrélation r^2_{cv} entre les activités ainsi calculées et les

activités observées exprime le pouvoir de prévision interne du modèle, plus la valeur du coefficient se rapproche de 1 plus le pouvoir de prévision sera meilleur. Pour que le modèle soit acceptable, le pouvoir de prévision interne doit être supérieur à 0,5 [101]**.**

- Validation croisée "leave one out cross-validation"

Cette méthode est un cas particulier de la validation croisée « k-paquets » où k=n. c'est- à-dire que l'on apprend sur (n-1) observations pour construire le modèle QSAR puis on le valide sur la $n^{ième}$ observation et l'on répète cette opération n fois pour qu'en fin de compte chaque observation ait été utilisée exactement une fois comme ensemble de validation.

La validation croisée est l'une des méthodes les plus utilisées pour la validation interne d'un modèle statistique [101]**.** Dans la validation croisée, la capacité prédictive d'un modèle est estimée à l'aide d'un ensemble réduit de données structurelles. Généralement, un élément de l'ensemble est extrait à chaque fois et un nouveau modèle est dérivé sur la base de données réduite, qui est ensuite utilisée pour prédire l'activité de la molécule exclue. La procédure est répétée n nombre de fois jusqu'à ce que tous les composés aient été exclus et prévus une fois. C'est la méthode dite du «laisser-un-sortir» (LOO)[102]**.** De manière analogue, laisser de côté plus d'une molécule du jeu de données à la fois est désigné par méthode **CV** sans ou avec ou sans sortie [103]. Le résultat de la procédure LOO est un coefficient de corrélation croisé r^2_{cv} (ou q^2) qui est un critère de robustesse et de capacité prédictive du modèle:

(37)

$$\mathbf{r^2cv = (PRESS_0 - PRESS) / (PRESS_0)}$$

Équation 37 : Coefficient de corrélation croisée r^2cv

Où **$\underline{PRESS_0}$** est la moyenne de l'activité biologique observée et PRESS est la somme des carrés des différences entre les valeurs d'activité prédites et observées. De nombreux chercheurs considèrent que le q^2 élevé est la preuve ultime de la puissance prédictive élevée du modèle QSAR, ce qui est incorrect. Il a été établi que, dans les cas où des ensembles d'essais avec des valeurs connues d'activités biologiques étaient disponibles pour la

prédiction, il n'existait aucune corrélation entre le q^2 et le r^2. Par conséquent, q^2 doit être considéré comme une mesure de la cohérence interne du modèle dérivé plutôt que comme un véritable indicateur de la prévisibilité. Il convient de noter que, puisqu'il est plus facile d'ajuster les données expérimentales que de les prévoir à partir du modèle QSAR, la valeur r^2 du modèle est toujours supérieure à q^2. La validation croisée n'est pas infaillible. Dans des ensembles de données hautement redondants avec moins de degrés de liberté, cela peut donner un résultat trop optimiste. Cela peut également indiquer incorrectement un manque de corrélation si tous les composés du jeu de données sont uniques. Par conséquent, nous pouvons conclure que, malgré sa large acceptation, une valeur élevée de q^2 seule est un critère insuffisant pour qu'un modèle QSAR soit hautement prédictif.

3.2 Pouvoir de prévision externe

Un modèle avec des valeurs élevées des indices internes q^2 (ou r^2_{cv}) n'est pas encore dit valide, par conséquent, la validation interne est nécessaire, mais insuffisante.

La puissance prédictive réelle d'un modèle QSAR est de tester leur capacité à prédire parfaitement l'activité/propriété des composés à partir d'un ensemble de test externe (composés non utilisés pour le développement du modèle). Le but d'un bon modèle QSAR est non seulement de prédire l'activité des composés d'ensemble d'apprentissage, mais aussi de prévoir les activités des molécules de test [104]**.** Le modèle QSAR est bâti sur l'ensemble d'apprentissage et validé sur l'ensemble de test. La capacité prédictive du modèle est basée sur le coefficient de corrélation r^2_{test} entre les activités observées et les activités prédites pour l'ensemble de test, la valeur plus élevée de r^2_{test} (> 0,5) indique la bonne productivité du modèle.

3.3 Domaine d'applicabilité

L'activité de l'univers entier des produits chimiques ne peut pas être prédite même par un modèle QSAR robuste et validé. La prédiction d'une réponse modélisée à l'aide du QSAR est valide que si le composé prévu se trouve dans le domaine d'applicabilité du modèle. L'applicabilité est une région théorique de l'espace chimique, défini par les descripteurs du modèle et la réponse modélisée et, donc, par la nature des molécules de l'ensemble de formation [105] [106].

Un modèle idéal est celui qui est capable de prédire l'activité ou la propriété de n'importe quelle molécule imaginable. Cependant cela est souvent loin d'être possible. La taille limitée du jeu d'entrainement rend l'espace chimique des modèles construits limité. Et par conséquent, lorsqu'une molécule se situe en dehors de cet espace chimique, la prédiction ne sera plus fiable [107]**.**

Pour éviter cette extrapolation hasardeuse et prévenir ce type des problèmes, un domaine d'applicabilité (DA), qui permet de définir la zone dans laquelle un composé pourra être prédit avec confiance, doit être déterminé. Le DA correspond donc à la région de l'espace chimique incluant les composés du jeu d'apprentissage et les composés similaires, proches dans ce même espace [19]**.** Cette stratégie permet d'éliminer du jeu de test les molécules se situant en dehors de l'espace chimique du jeu d'entrainement. Cette partie de l'analyse est d'ailleurs explicitement demandée dans les démarches de validation mises en place au niveau de l'OCDE [96] [107].

Il est possible de vérifier si un nouveau produit chimique se trouve dans domaine d'applicabilité en utilisant l'approche par effet de levier. A le composé sera considéré comme n'étant pas applicable lorsque la valeur de l'effet de levier est supérieure à la valeur critique valeur de 3p/n, où p est le nombre de variables du modèle plus 1 et n est le nombre d'objets utilisés pour développer ou valider le modèle.

Récemment le domaine d'applicabilité sera discuté à l'aide du diagramme de Williams qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers [96][108].

Pour chaque composé *i* dans l'espace original des variables indépendantes (xi), la valeur d'est calculée par la relation suivante [109]:

(39)

$$\mathbf{Hi} = \boldsymbol{x_i^T}(x^T x)^{-1} x_i \ \ \mathbf{(i=1, 2 \ldots n)}$$

Équation 38 : Fonction des valeurs des leviers

Avec : xi est vecteur ligne des descripteurs du composé i ; X (n*k-1) est la matrice du

Modèle déduit des valeurs des descripteurs de l'ensemble d'entrainement ; l'indice T désigne la matrice transposée. La valeur critique du levier (h*) est fixée à : h*=3*k/n [105]**.**

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

Si hi<h*, la probabilité d'accord entre les valeurs mesurée et prédite du composé « i » est aussi élevée que celle des composés de la base de données. Les composés avec hi>h* renforcent le modèle quand ils appartiennent à l'ensemble d'entrainement, mais auront, sinon, des valeurs prédites douteuses sans pour autant être forcément aberrantes, les résidus pouvant être bas [110]**.**

4 Logiciels utilisés dans nos études QSAR

Dans notre travail on a utilisé plusieurs logiciels libres ou commerciaux disponibles dans les études QSAR. Ceux-ci comprennent des logiciels spécialisés pour dessiner les structures chimiques, générant des structures 3D, le calcul des descripteurs moléculaires et le développement de modèles QSAR. Les logiciels utilisés dans nos travaux sont :

Le dessin des molécules a été fait par Chemdrow, ChemSketch [111];

Les structures 3D des molécules ont été générées par GaussView [53];

Les descripteurs ont été calculés par hyperChem et Gaussian 09 [112] [42];

L'analyse descriptive et la validation des modèles QSAR sont été faites par SPSS [113].

References

[1] M.Touhami, “Modélisation de l’activité biologique de composés hétérocycles,” 2019.

[2] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, “The protein kinase complement of the human genome,” *Science*, vol. 298, no. 5600, pp. 1912–1934, 2002, doi: 10.1126/science.1075762.

[3] J. B. O. Mitchell B.O., “Machine learning methods in chemoinformatics,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 4, no. 5, pp. 468–481, 2014, doi: 10.1002/wcms.1183.

[4] H. M. Berman *et al.*, “The Protein Data Bank,” *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.

[5] I. Baskin and A. Varnek, “Fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening,” *ChemInform*, vol. 40, no. 20, p. i, 2009.

[6] F. K. Brown, “Chemoinformatics: what is it and how does it impact drug discovery,” *Annual reports in medicinal chemistry*, vol. 33, pp. 375–384, 1998.

[7] W. A. Warr, “Balancing the needs of the recruiters and the aims of the educators.,” in *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*, 1999, vol. 218, pp. U500–U500.

[8] J. L. Reymond, L. C. Blum, and R. Van Deursen, “Exploring the chemical space of known and unknown organic small molecules at www.gdb.unibe.ch,” *Chimia*, vol. 65, no. 11, pp. 863–867, 2011, doi: 10.2533/chimia.2011.863.

[9] J. Reymond, L. Ruddigkeit, L. Blum, and R. van Deursen, “The enumeration of chemical space,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 2, no. 5, pp. 717–733, 2012.

[10] P. Kirkpatrick and C. Ellis, “Chemical space.” Nature Publishing Group, 2004.

[11] J.-L. Reymond and M. Awale, “Exploring chemical space for drug discovery using the chemical universe database,” *ACS chemical neuroscience*, vol. 3, no. 9, pp. 649–657, 2012.

[12] Y. C. Martin, J. L. Kofron, and L. M. Traphagen, “Do structurally similar molecules have similar biological activity?,” *Journal of medicinal chemistry*, vol. 45, no. 19, pp. 4350–4358, 2002.

[13] A. P. Bento *et al.*, “The ChEMBL bioactivity database: an update,” *Nucleic acids research*, vol. 42, no. D1, pp. D1083–D1090, 2014.

[14] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, “PubChem: integrated platform of small molecules and biological activities,” in *Annual reports in computational chemistry*, vol. 4, Elsevier, 2008, pp. 217–241.

[15] M. G. Damale, S. N. Harke, F. A. K. Khan, D. B. Shinde, and J. N. Sangshetti, “Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review,” pp. 35–55, 2014.

[16] C. H. Andrade, K. F. M. Pasqualoto, E. I. Ferreira, and A. J. Hopfinger, “4D-QSAR: Perspectives in Drug Design,” pp. 3281–3294, 2010, doi: 10.3390/molecules15053281.

[17] A. Vedani and M. Dobler, “5D-QSAR: the key for simulating induced fit?,” *Journal of medicinal chemistry*, vol. 45, no. 11, pp. 2139–2149, 2002.

[18] A. Vedani, A.-V. Descloux, M. Spreafico, and B. Ernst, “Predicting the toxic potential of drugs and chemicals in silico: A model for the peroxisome proliferator-activated receptor γ (PPAR γ),” *Toxicology letters*, vol. 173, no. 1, pp. 17–23, 2007.

[19] R. D. Cramer, D. E. Patterson, and J. D. Bunce, “Comparative molecular field analysis

(CoMFA). 1. Effect of shape on binding of steroids to carrier proteins," *Journal of the American Chemical Society*, vol. 110, no. 18, pp. 5959–5967, 1988.
[20] C. Hansch and E. J. Lien, "Structure-activity relations in antifungal agents. A survey," *Journal of medicinal chemistry*, vol. 14, no. 8, pp. 653–670, 1971.
[21] S. Y. Tham and S. Agatonovic-Kustrin, "Application of the artificial neural network in quantitative structure–gradient elution retention relationship of phenylthiocarbamyl amino acids derivatives," *Journal of pharmaceutical and biomedical analysis*, vol. 28, no. 3–4, pp. 581–590, 2002.
[22] J. S. Jaworska, M. Comber, C. Auer, and C. J. Van Leeuwen, "Summary of a workshop on regulatory acceptance of (Q) SARs for human health and environmental endpoints.," *Environmental health perspectives*, vol. 111, no. 10, pp. 1358–1360, 2003.
[23] S. H. Unger and C. Hansch, "Model building in structure-activity relations. Reexamination of adrenergic blocking activity of. beta.-halo-. beta.-arylalkylamines," *Journal of medicinal chemistry*, vol. 16, no. 7, pp. 745–749, 1973.
[24] "OECD PRINCIPLES FOR THE VALIDATION, FOR REGULATORY PURPOSES, OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP MODELS These principles were agreed by OECD member countries at the 37," *Biotechnology*, no. November, pp. 3–4, 2004.
[25] C. Selassie and R. P. Verma, "History of quantitative structure–activity relationships," *Burger's Medicinal Chemistry and Drug Discovery*, pp. 1–96, 2003.
[26] R. Todeschini and V. Consonni, "@ WILEY-VCH."
[27] M. Karelson, *Molecular descriptors in QSAR/QSPR*, vol. 230. Wiley-Interscience New York, 2000.
[28] R. Todeschini, "Molecular Descriptors for Chemoinformatics. Todeschini R, Consonni V., editors." Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2009.
[29] A. Z. Dudek, "Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review," pp. 213–228, 2006.
[30] U. Constantine and S. Exactes, "" Drug Design " et synthèse de nouveaux calix [8] arènes sulfoniques flexibles à activités anticorrosive et anticoagulante," 2014.
[31] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced drug delivery reviews*, vol. 23, no. 1–3, pp. 3–25, 1997.
[32] A. Goulon-Sigwalt-Abram, "Une nouvelle méthode d'apprentissage de données structurées: applications à l'aide à la découverte de médicaments." Université Pierre et Marie Curie-Paris VI, 2008.
[33] L. H. Hall, L. B. Kier, and B. B. Brown, "Molecular similarity based on novel atom-type electrotopological state indices," *Journal of chemical information and computer sciences*, vol. 35, no. 6, pp. 1074–1080, 1995.
[34] L. H. Hall and L. B. Kier, "Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information," *Journal of Chemical Information and Computer Sciences*, vol. 35, no. 6, pp. 1039–1045, 1995.
[35] M. Randic, "Characterization of molecular branching," *Journal of the American Chemical Society*, vol. 97, no. 23, pp. 6609–6615, 1975.
[36] G. Cerruela García, I. Luque Ruiz, M. Á. Gómez-Nieto, J. A. Cabrero Doncel, and A. Guevara Plaza, "From Wiener index to molecules," *Journal of chemical information and modeling*, vol. 45, no. 2, pp. 231–238, 2005.
[37] J. S. Mason, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme, and R. F. Labaudiniere,

"New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures," *Journal of medicinal chemistry*, vol. 42, no. 17, pp. 3251–3264, 1999.

[38] L. Euler, "Solutio problematis ad geometriam situs pertinentis," *Commentarii academiae scientiarum Petropolitanae*, pp. 128–140, 1741.

[39] H. P. Schultz, "Topological organic chemistry. 1. Graph theory and topological indices of alkanes," *Journal of Chemical Information and Computer Sciences*, vol. 29, no. 3, pp. 227–228, 1989.

[40] R. A. Saunders and J. A. Platts, "Scaled polar surface area descriptors: development and application to three sets of partition coefficients," *New Journal of Chemistry*, vol. 28, no. 1, pp. 166–172, 2004.

[41] H. Kubinyi, G. Folkers, and Y. C. Martin, *3D QSAR in Drug Design: Volume 2: Ligand-Protein Interactions and Molecular Similarity*, vol. 2. Springer Science & Business Media, 1998.

[42] H. Hyperchem, "Molecular modeling system. Hyper Cube," *Inc. and Auto Desk, Inc*, 2002.

[43] R. Bosque, J. Sales, E. Bosch, M. Roses, M. C. García-Alvarez-Coque, and J. R. Torres-Lapasió, "A QSPR study of the p solute polarity parameter to estimate retention in HPLC," *Journal of chemical information and computer sciences*, vol. 43, no. 4, pp. 1240–1247, 2003.

[44] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, and R. K. Robins, "Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain ," *Journal of Chemical Information and Computer Sciences*, vol. 29, no. 3, pp. 163–172, Aug. 1989, doi: 10.1021/ci00063a006.

[45] C. Hansch, P. G. Sammes, and J. B. Taylor, "Comprehensive Medicinal Chemistry. 4. Volume (Quantitative Drug Design) Pergamon Press." Oxford, UK, 1990.

[46] E. H. Kerns and L. Di, "Drug-like properties: Concepts," *Structure Design and Methods. Elsevier, Amsterdam*, 2008.

[47] S. Belaidi, N. Melkemi, and D. Bouzidi, "Molecular geometry and structure-property relationships for 1, 2-dithiole-3-thione derivatives," *Int J Chem Res*, vol. 4, no. 2, pp. 134–139, 2012.

[48] H. A. Lorentz, "Ueber die Beziehung zwischen der Fortpflanzungsgeschwindigkeit des Lichtes und der Körperdichte," *Annalen der Physik*, vol. 245, no. 4, pp. 641–665, 1880.

[49] C. Hansen, B. R. Telzer, and L. Zhang, "Comparative QSAR in toxicology: examples from teratology and cancer chemotherapy of aniline mustards," *Critical reviews in toxicology*, vol. 25, no. 1, pp. 67–89, 1995.

[50] A. Cammarata, "An apparent correlation between the in vitro activity of chloramphenicol analogs and electronic polarizability," *Journal of medicinal chemistry*, vol. 10, no. 4, pp. 525–527, 1967.

[51] C. Hansch and E. Coats, "α-Chymotrypsin: A Case Study of Substituent Constants and Regression Analysis in Enzymic Structure—Activity Relationships," *Journal of pharmaceutical sciences*, vol. 59, no. 6, pp. 731–743, 1970.

[52] F. Neese, "A critical evaluation of DFT , including time-dependent DFT , applied to bioinorganic chemistry," pp. 702–711, 2006, doi: 10.1007/s00775-006-0138-1.

[53] R. G. I.I.R. Denning, T. Keith, J. Millam, K. Eppinnett, W.L. Hovell, "No Title."

KS,USA, 2003.
[54] K. Fukui, “Theory of orientation and stereoselection,” in *Orientation and Stereoselection*, Springer, 1970, pp. 1–85.
[55] R. Franke, “Theoretical drug design methods,” *Pharmacochemistry library*, vol. 7, 1984.
[56] P. W. Atkins, J. De Paula, and J. Keeler, *Atkins’ physical chemistry*. Oxford university press, 2018.
[57] F. Lamchouri *et al.*, “Quantitative structure–activity relationship of antitumor and neurotoxic β-carbolines alkaloids: nine harmine derivatives,” *Research on Chemical Intermediates*, vol. 39, no. 5, pp. 2219–2236, 2013.
[58] D. F. V Lewis, C. Ioannides, and D. V Parke, “Interaction of a series of nitriles with the alcohol-inducible isoform of P450: computer analysis of structure—activity relationships,” *Xenobiotica*, vol. 24, no. 5, pp. 401–408, 1994.
[59] Z. Zhou and R. G. Parr, “Activation hardness: new index for describing the orientation of electrophilic aromatic substitution,” *Journal of the American Chemical Society*, vol. 112, no. 15, pp. 5720–5724, 1990.
[60] O. Kikuchi, “Systematic QSAR procedures with quantum chemical descriptors,” *Quantitative Structure-Activity Relationships*, vol. 6, no. 4, pp. 179–184, 1987.
[61] A. Zeroual, M. El Idrissi, A. Benharref, and A. El Hajbi, “Etude theorique de la regioselectivite et la stereoselectivite de la condensation du [Beta]-himachalene avec le dichlorocarbene par la theorie de la fonctionnelle de la densite (DFT)[Theoretical study of regioselectivity and stereoselectivity of condensat,” *International Journal of Innovation and Applied Studies*, vol. 5, no. 2, p. 120, 2014.
[62] L. Kosychova *et al.*, “New 1-(3-Nitrophenyl)-5, 6-dihydro-4H-[1, 2, 4] triazolo [4, 3-a][1, 5] benzodiazepines: Synthesis and Computational Study,” *Molecules*, vol. 20, no. 4, pp. 5392–5408, 2015.
[63] R. G. Parr, R. A. Donnelly, M. Levy, and W. E. Palke, “Electronegativity: the density functional viewpoint,” *The Journal of Chemical Physics*, vol. 68, no. 8, pp. 3801–3807, 1978.
[64] R. S. Mulliken, “A new electroaffinity scale; together with data on valence states and on valence ionization potentials and electron affinities,” *The Journal of Chemical Physics*, vol. 2, no. 11, pp. 782–793, 1934.
[65] R. G. Parr and R. G. Pearson, “Absolute hardness: companion parameter to absolute electronegativity,” *Journal of the American chemical society*, vol. 105, no. 26, pp. 7512–7516, 1983.
[66] W. Yang and R. G. Parr, “Hardness , softness , and the fukui function in the electronic theory of metals and catalysis,” vol. 82, no. October, pp. 6723–6726, 1985.
[67] R. G. Pearson, “Absolute electronegativity and hardness: applications to organic chemistry,” *The Journal of Organic Chemistry*, vol. 54, no. 6, pp. 1423–1430, 1989.
[68] R. G. Parr, C. Hill, and N. Carolina, “Electrophilicity Index,” no. 10, pp. 1922–1924, 1999.
[69] R. S. Mulliken, “Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I,” vol. 1833, no. 1955, 2004, doi: 10.1063/1.1740588.
[70] N. Allard, “Design et synthèse de nouveaux polymères π - conjugués et optimisation de dispositifs photovoltaïques,” 2015.
[71] B. M, “etude physico-chimique par méthodes computionnelles des propriétés structurales et optoélectroniques des polymères, oligoméres et nano-système p-conjugués à la base de thiophène,” l’Université Moulay Ismail Faculté des sciences

etTechniques Errachidia, 2008.
[72] D. A. McQuarrie, “Statistical thermodynamics,” 1973.
[73] P. W. Atkins, “Physical Chemistry, W. H.” H. Freeman and Co., New York, 1982.
[74] A. R. Katritzky, V. S. Lobanov, and M. Karelson, “Normal boiling points for organic compounds: correlation and prediction by a quantitative structure− property relationship,” *Journal of chemical information and computer sciences*, vol. 38, no. 1, pp. 28–41, 1998.
[75] W. A. Wakeham, G. S. Cholakov, and R. P. Stateva, “Liquid density and critical properties of hydrocarbons estimated from molecular structure,” *Journal of Chemical & Engineering Data*, vol. 47, no. 3, pp. 559–570, 2002.
[76] A. R. Katritzky, U. Maran, M. Karelson, and V. S. Lobanov, “Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach,” *Journal of Chemical Information and Computer Sciences*, vol. 37, no. 5, pp. 913–919, Sep. 1997, doi: 10.1021/ci970027a.
[77] R. D. Cramer, D. E. Patterson, and J. D. Bunce, “Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins,” *Journal of the American Chemical Society*, vol. 110, no. 18, pp. 5959–5967, 1988, doi: 10.1021/ja00226a005.
[78] C. Navajas, A. Poso, K. Tuppurainen, and J. Gynther, “Comparative Molecular Field Analysis (CoMFA) of MX Compounds using different Semi-empirical Methods: LUMO Field and its Correlation with Mutagenic Activity,” *Quantitative Structure-Activity Relationships*, vol. 15, no. 3, pp. 189–193, Jan. 1996, doi: 10.1002/qsar.19960150302.
[79] F. Bonachera, “Les triplets pharmacophoriques ous Développement et applications.” Université de Lille 1, 2011.
[80] G. F. Bennett, “Lees’ Loss Prevention in the Process Industries: Hazard Identification, Assessment and Control, vol. II, Sam Mannan (Ed.), Elsevier, Butterworth, Heinemann, Burlington, MA (2005), three-volume set, US $476.00, 1071 pp., ISBN 0-7506-7555-1 (three-volume se.” Elsevier, 2005.
[81] M. J. Crawley, “Statistics: an introduction using RJ Wiley,” *Chichester, West Sussex, England*, 2005.
[82] P. Dagnelie, “Statistique théorique et appliquée volume 2, Tome 1. De Boeck et Larcier, Belgique, 508 p. CONSERVATION D’ARBRES SAUVAGES À FRUITS COMESTIBLES,” *LE POINT SUR*, 1998.
[83] A. R. Katritzky, V. S. Lobanov, and M. Karelson, “CODESSA: Reference Manual University of Florida Gainesville.” FL, 1994.
[84] S. M. Stigler, *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press, 2002.
[85] N. Trinajstič, S. Nikolič, S. C. Basak, and I. Lukovits, “Distance indices and their hyper-counterparts: Intercorrelation and use in the structure-property modeling,” *SAR and QSAR in Environmental Research*, vol. 12, no. 1–2, pp. 31–54, 2001.
[86] P. P. Roy, S. Paul, I. Mitra, and K. Roy, “Roy et al. On Two Novel Parameters for Validation of Predictive QSAR Models. Molecules, 2009, 14, 1660-1701,” *Molecules*, vol. 15, no. 1, pp. 604–605, 2010.
[87] J. G. Topliss and R. P. Edwards, “Chance factors in studies of quantitative structure-activity relationships Chance Factors in Studies of Quantitative Structure-Activity Relationships,” vol. 22, no. 10, pp. 1238–1244, 1979, doi: 10.1021/jm00196a017.
[88] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,”

Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100–108, 1979.
[89] D. Pan, M. Iyer, J. Liu, Y. Li, and A. J. Hopfinger, "Constructing Optimum Blood Brain Barrier QSAR Models Using a Combination of 4D-Molecular Similarity Measures and Cluster Analysis," pp. 2083–2098, 2004.
[90] K. Roy, S. Kar, and R. N. Das, *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic press, 2015.
[91] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators.," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.
[92] S. Haykin, "Self-organizing maps," *Neural networks-A comprehensive foundation, 2nd edition, Prentice-Hall*, 1999.
[93] W. Huang and S. Foo, "Neural network modeling of salinity variation in Apalachicola River," *Water Research*, vol. 36, no. 1, pp. 356–362, 2002.
[94] H. Demuth and M. Beale, "Neural Network Toolbox For Use with Matlab--User'S Guide Verion 3.0," 1993.
[95] A. J. Adeloye and A. De Munari, "Artificial neural network based generalized storage–yield–reliability models using the Levenberg–Marquardt algorithm," *Journal of Hydrology*, vol. 326, no. 1–4, pp. 215–230, 2006.
[96] J. Shao, "Bootstrap model selection," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 655–665, 1996.
[97] T. J. Archdeacon, "Evaluating the regression equation," *Correlation and Regression Analysis: a historian's Guide*. Univ of Wisconsin Press: USA, pp. 160–177, 1994.
[98] C. Rücker, G. Rücker, and M. Meringer, "y-Randomization and its variants in QSPR/QSAR," *Journal of chemical information and modeling*, vol. 47, no. 6, pp. 2345–2357, 2007.
[99] D. Laffly, "Régression multiple : principes et exemples d ' application Dominique Laffly Université de Pau et des Pays de l ' Adour Octobre 2006," p. 33, 2006.
[100] T. J. Archdeacon, "Regression and explained variance," *Correlation and Regression Analysis: a Historian's Guide*. Univ of Wisconsin Press: USA, pp. 178–196, 1994.
[101] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation.," *Encyclopedia of database systems*, vol. 5, 2009.
[102] R. Todeschini, V. Consonni, and R. Mannhold, "Methods and principles in medicinal chemistry," *Kubinyi H, Timmerman H (Series eds) Handbook of molecular descriptors. Wiley-VCH, Weinheim*, 2000.
[103] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, Feb. 1974.
[104] H. van der Voet, "Comparing the predictive accuracy of models using a simple randomization test," *Chemometrics and intelligent laboratory systems*, vol. 25, no. 2, pp. 313–323, 1994.
[105] S. Ekins *et al.*, "Three-and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors," *Drug Metabolism and Disposition*, vol. 28, no. 8, pp. 994–1002, 2000.
[106] I. V Tetko *et al.*, "Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection," *Journal of chemical information and modeling*, vol. 48, no. 9, pp. 1733–1746, 2008.
[107] J. Jaworska, N. Nikolova-Jeliazkova, and T. Aldenberg, "QSAR applicability domain

estimation by projection of the training set in descriptor space: a review," *Alternatives to laboratory animals*, vol. 33, no. 5, pp. 445–459, 2005.
[108] T. I. Netzeva *et al.*, "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52," *Alternatives to Laboratory Animals*, vol. 33, no. 2, pp. 155–173, 2005.
[109] L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, and P. Gramatica, "Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs.," *Environmental health perspectives*, vol. 111, no. 10, pp. 1361–1375, 2003.
[110] P. Gramatica, "Principles of QSAR models validation: internal and external," *QSAR & combinatorial science*, vol. 26, no. 5, pp. 694–701, 2007.
[111] J. C. Dearden, "The history and development of quantitative structure-activity relationships (QSARs): addendum," *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, vol. 2, no. 2, pp. 36–46, 2017.
[112] M. J. Frisch *et al.*, "Gaussian 09, Revision A. 02, 2009, Gaussian," *Inc., Wallingford CT*, 2009.
[113] W. E. Wagner III, *Using IBM® SPSS® statistics for research methods and social science statistics*. Sage Publications, 2019.

Chapitre III : Résultats et discussions

1 Introduction

Dans le cadre de la recherche d'un nouveau et puissant médicament anticancéreus (contre la tumeur d'ovarienne et cancer du poumon), une série de 28 dérivées d'indazole diversement substituées a été soumis à une analyse QSAR pour étudier, interpréter et prédire les activités et concevoir de nouveaux composés en utilisant des méthodes de régression linéaire multiple (RLM) et de réseaux neuronaux artificiels (RNN). Les descripteurs utilisés ont été calculés avec les programmes : Gaussian 09, ChemOffice et hyperChem. Les modèles QSAR développés ont été validés selon les principes établis par l'Organisation de Coopération et de Développement Economiques (OCDE). L'Analyse en Composantes Principales (ACP) a été utilisée pour sélectionner des descripteurs qui montrent une forte corrélation avec les activités.

La méthode de régression linéaire multiple (RLM) a montré une corrélation efficace de r =0,805 et r= 0,945 pour les activités biologique IC_{50} d'ovarienne et poumon, respectivement. Des validations internes et externes ont été utilisées pour déterminer la qualité statistique du modèle QSAR par RLM. La méthode des réseaux neuronaux artificiels (RNN), compte tenu des descripteurs pertinents obtenus par la méthode RLM, a montré des coefficients de corrélation de r=0,860 et r=0,945 pour A2780 et A549 respectivement. Le domaine d'applicabilité des modèles RLM a été étudié en utilisant des approches simples et à effet de levier pour détecter les valeurs aberrantes et les composés extérieurs.

Les effets de différents descripteurs d'activités ont été décrits et utilisés pour étudier et concevoir de nouveaux composés ayant une activité plus élevée que les composés existants.

2 Molécule-mère de ce travail indazole

L'indazole, aussi connu sous le nom de 1,2-benzopyrazole ou encore 1,2-benzodiazole (figure 9), fait partie des composés organiques hétérocycliques d'importance capitale en chimie organique. Cet hétérocycle constitue une classe de composés renfermant d'intéressantes activités tant chimiques[1] que biologiques [2].

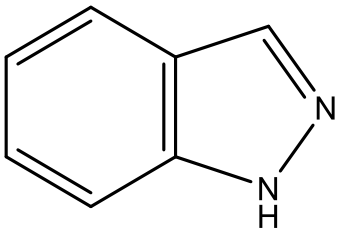


Figure 9:Molécule mère d'indazole [1].

Pour comprendre la réactivité chimique de l'indazole, il est important de connaître sa tautomérie et son aromaticité. L'indazole possède deux atomes d'azote et peut exister sous deux formes (1H-indazole et 2H-indazole) Schéma 1 qui résultent de la délocalisation du proton entre les deux atomes d'azote, un processus décrit comme une tautomérie prototropique annulaire. En raison de la différence d'énergie entre les tautomères, la forme 1H-indazole prédomine fortement en phase gazeuse, en solution, et en phase solide, et ses dérivés sont généralement plus stables thermodynamiquement que la forme 2H-indazole [3] .

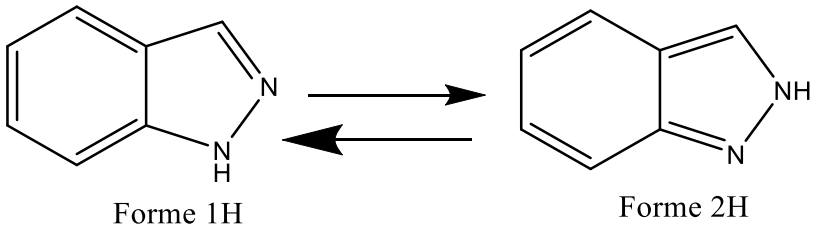


Schéma 1

Les structures 1H-indazole, et 2H-indazole peuvent exister séparément lorsque l'hydrogène pyrazolique est substitué. Plusieurs méthodes ont été développées pour accéder aux dérivés N-substitués de l'indazole. La réaction d'alkylation est une des méthodes utilisée pour préparer les dérivés indazoliques substitués en position N-1 et N-2 .

Le but de cette partie est d'étudier la corrélation entre les dérivées de la molécule indazole ou 1,2-benzodiazole [1][2][3][4][5][6], qui fait partie des composés organiques hétérocycliques d'importance capitale en chimie organique. Cet hétérocycle constitue une classe de composés renfermant d'intéressantes activités tant chimiques que

biologiques[7][8][9] , les propriétés physiques et chimiques dans le but d'optimiser les molécules testées pour construire un modèle QSAR de ces molécules.

2.1 Tumeurs étudiées

2.1.1 Tumeur d'ovarienne

Cancer de l'ovaire regroupe un ensemble de tumeurs pouvant toucher différents tissus de cet organe. En effet, il arrive que certaines **cellules de l'ovaire** subissent une transformation qui les rend cancéreuses. Dans certaines conditions, une cellule anormale peut se mettre à proliférer de manière anarchique et mener à la formation d'une **tumeur maligne** de l'ovaire[10].

2.1.2 Tumeurs cancéreuses du poumon

Cancer du poumon, appelé aussi cancer bronchique ou cancer broncho pulmonaire, est une maladie des cellules des bronches ou, plus rarement, des cellules qui tapissent les alvéoles pulmonaires. Il se développe à partir d'une cellule initialement normale qui se transforme et se multiplie de façon anarchique, jusqu'à former une masse appelée tumeur maligne[11].

3 Objectif

Le but de notre thèse c'est la création des modèles mathématiques qui lient entre les activités biologiques contre les deux tumeurs d'ovariennes et du poumon (A2780 et A549) et les propriétés physiques et chimiques de nos 28 molécules testées, la confirmation des règles de cinq (ROF) de Lipinski sur les 28 molécules et la création des nouvelles molécules avec d'activité biologique intéressante.

L'objectif de cette étude était de développer un modèle QSAR capable de corréler les caractéristiques structurelles des dérivés d'indazole avec leurs activités biologiques (A2780 et A549).

En général, la méthode QSAR est basée sur l'hypothèse que l'activité de certains composés chimiques est liée à leur structure par un certain algorithme mathématique. Cette relation peut être utilisée pour la prévision, l'interprétation et l'évaluation de nouveaux

composés dont l'activité est souhaitée, en réduisant et en rationalisant le temps, les efforts et les coûts de synthèse et de développement de nouveaux produits. L'hypothèse de base pour la conduite d'un modèle QSAR est présentée en raison d'une fonction mathématique des propriétés chimiques qui est liée à l'activité biologique.

Par conséquent, l'activité est similaire à la fonction "Y" des propriétés chimiques "x" : Y = f (x). Pour trouver cet algorithme, nous utilisons 28 composés chimiques similaires dont nous connaissons les valeurs de l'activité étudiée (Y). Pour chaque composé chimique, nous calculons une série de paramètres (appelés descripteurs chimiques et physiques). Ensuite, nous trouvons un algorithme qui fournit une valeur assez précise, similaire à la valeur expérimentale réelle. La dernière étape consiste à vérifier si l'algorithme obtenu est capable de prédire les valeurs d'activité d'autres substances chimiques non utilisées pour construire le modèle (validation externe).

En effet, il est très important de générer un modèle qui fonctionne non seulement pour les substances chimiques utilisées dans l'ensemble de formation, mais aussi pour d'autres substances chimiques similaires. Par conséquent, le défi consiste à définir les propriétés statistiques correctes du modèle.

Cependant, il est beaucoup plus difficile de prévoir : Absorption, Distribution, Métabolisme et Élimination des médicaments (ADME), qui nécessitent généralement une évaluation *in vivo*. Les études *in vivo* étant lentes et coûteuses, il est souhaitable de disposer de méthodes simples pour prédire les propriétés ADME des candidats médicaments. Une méthode largement acceptée pour prédire les propriétés de l'ADME est la règle des cinq proposée par Lipinski en 1997 [13]. Pour développer cette règle, Lipinski a effectué une analyse rétrospective de 2245 médicaments entrant en phase II, dont la plupart étaient des médicaments lipophiles actifs par voie orale et a identifié des propriétés physicochimiques communes. La corrélation qui en a résulté a permis d'identifier quatre paramètres physico-chimiques : le poids moléculaire (MM), le nombre de liaisons donneuses (NHD), le nombre de liaisons acceptrices (NHA) et le coefficient de partage octanol-eau (*Log P*).

Dans la première étape de la découverte de médicaments, il est tout à fait nécessaire d'appliquer des filtres de type médicamenteux pour éliminer les molécules non médicamenteuses des bases de données et de se concentrer ensuite uniquement sur les

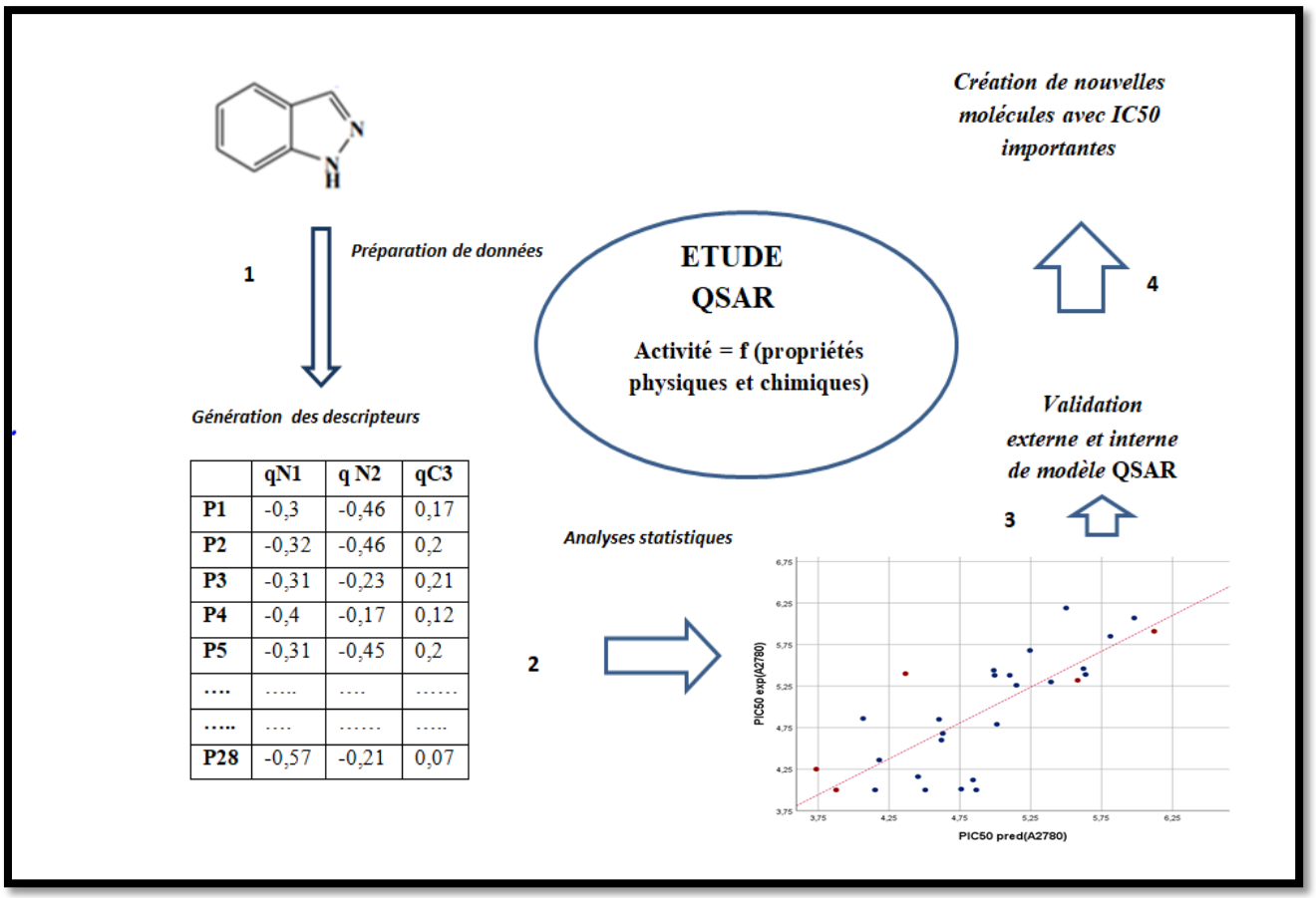
molécules de type médicamenteux. De nos jours, l'évaluation de la similarité des médicaments par exemple, la règle des cinq de Lipinski[14], les règles de l'*Opér*a de la similitude des médicaments[15], le filtre de *Rös*[16], etc. A déjà été, dans une certaine mesure, intégrée dans les pipelines de conception/découverte de médicaments par ordinateur. Au cours des dernières décennies, des efforts considérables ont été faits pour les approches informatiques visant à différencier les molécules de type médicament des réactifs, comme les filtres ou les règles simples basées sur les propriétés [17][18], et l'index de type médicament pour classer les molécules [16][18].

Nos recherches actuelles visent à décrire les relations structure-propriété sur l'indazole et un modèle QSAR sur ces composés en ce qui concerne leur activité inhibitrice [19].

4 Méthodes et matériels

4.1 Méthodologie de ce travail

Étude QSAR actuelle porte sur la prédiction et l'interprétation des composés étudiés et a également été utilisée pour concevoir de nouveaux composés proposés en utilisant des méthodes linéaires. Elle comprend quatre étapes figure 10 sélection de l'ensemble de données et génération de descripteurs moléculaires, analyse descriptive, analyse statistique et suggestion de nouveaux composés ou molécules avec des propriétés biologiques très importantes.



	qN1	q N2	qC3
P1	-0,3	-0,46	0,17
P2	-0,32	-0,46	0,2
P3	-0,31	-0,23	0,21
P4	-0,4	-0,17	0,12
P5	-0,31	-0,45	0,2
….	…..	….	……
…..	….	……	…..
P28	-0,57	-0,21	0,07

Figure 10 : Organigramme développement du modèle QSAR dans le cadre de ce travail

4.2 Descripteurs moléculaires

4.2.1 Sélection de l'ensemble de données

Au cours de la première étape de notre étude, les ensembles de données sur les activités IC_{50} (contre la tumeur d'ovarienne et poumon) (A2780 et A549) de dérivés d'indazole **(**figure11) sont :

- N-(6-indazolyl)-arylsulfonamides ,
- N-(5-indazolyl)-arylsulfonamides,
- N-(7-indazolyl)-arylsulfonamides,
- 2-(7(4)-hydroxyimino-N-alkyl-1,7-dihydro-indazol7(4)-ylidene)-2-arylacetonitriles

- 3-aryl-isoxazolo[4,3-h]quinolines,2-(aryl)-2-(7(4)-(arylsulfonyl)oxime-1-éthyl-1Hindazol-4-ylidene)acétonitriles) [20].

Les structures moléculaires et leurs activités IC_{50} (A2780 et A549) sont présentées dans le tableau 3. Toutes les valeurs expérimentales de l'activité de la IC_{50} (μM) ont été converties en logarithme négatif de la IC_{50} (pIC_{50} =-log ($CI_{50}x10^{-6}$) μM).

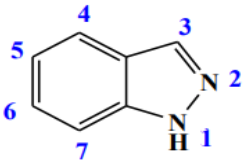
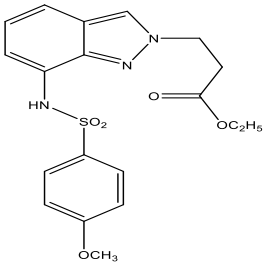
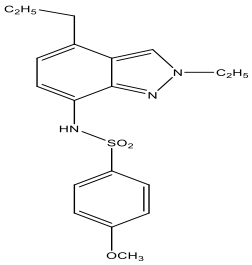


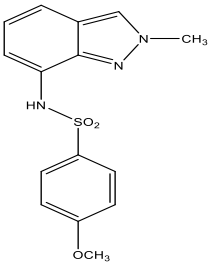
Figure 11 : indazole [1]



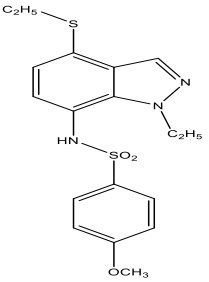
M1



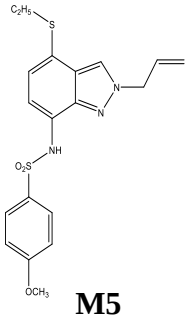
M2



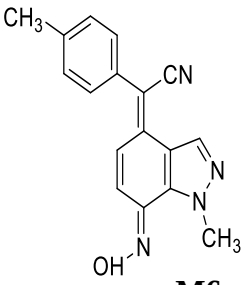
M3



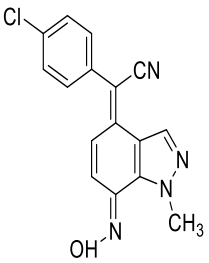
M4



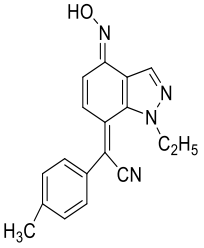
M5



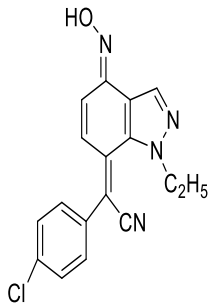
M6



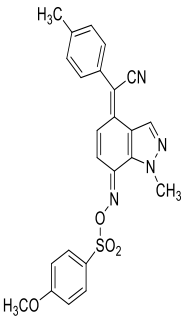
M7



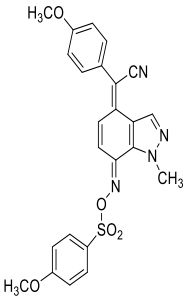
M8



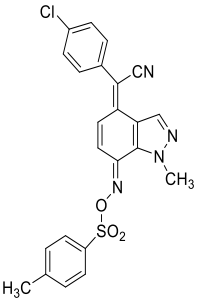
M9



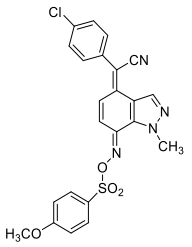
M10



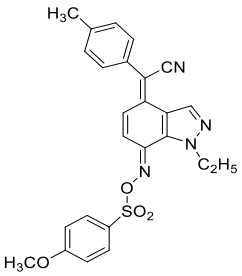
M11



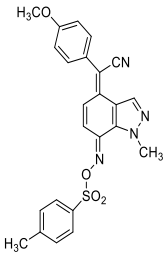
M12



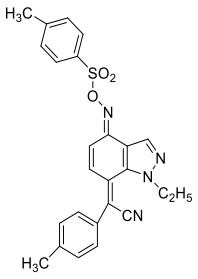
M13



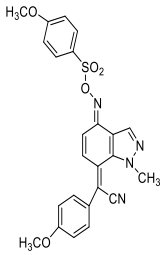
M14



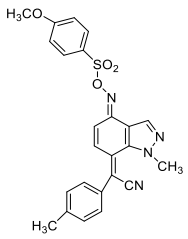
M16



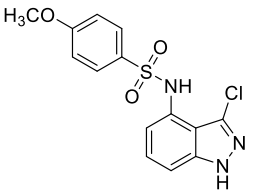
M15



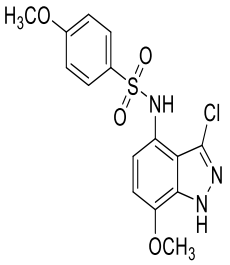
M17



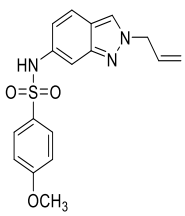
M18



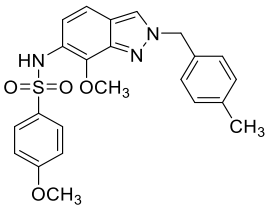
M19



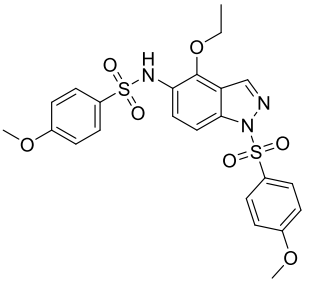
M20



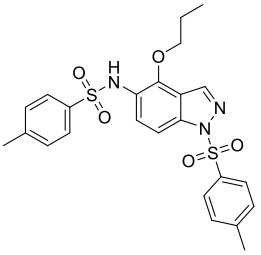
M21



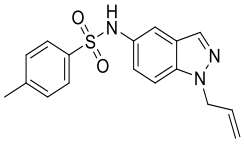
M22



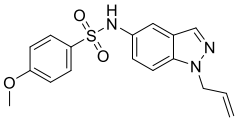
M23



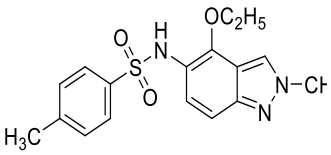
M24



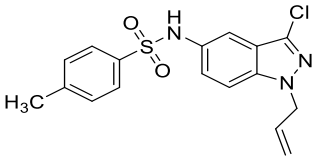
M25



M26



M27



M28

Figure 12 : Vingt-huit molécules dérivées de l'indazole.

Tableau 3 : Activites biologiques expérimentales

Molecules	*pIC_{50}(OVAR)	*pIC_{50}(POUM)
M1	4,00	4,00
M2	4,00	4,00
M3	4,00	4,00
M4	5,40	5,32
M5	4,00	4,06
M6	4,85	4,86
M7	4,68	4,86
M8	4,36	4,14
M9	4,25	4,10
M10	5,32	5,22
M11	5,46	5,34
M12	5,68	5,69
M13	5,30	5,28
M14	5,39	5,41
M16	6,19	5,46
M15	5,44	5,38
M17	5,85	5,42
M18	5,38	5,46
M19	6,07	5,74
M20	5,91	5,23
M21	5,38	5,26
M22	5,26	5,38
M23	4,12	Non
M24	4,01	Non
M25	4,16	4,55
M26	4,04	4,24
M27	4,38	4,28
M28	4,80	4,18

*pIC_{50}(OVAR) : pIC50 (µM) (A2780) qui inhibe la tumeur d'ovarienne
*pIC_{50}(POUM) : pIC_{50} (µM) (A549) qui inhibe la tumeur du poumon

4.2.2 Génération des descripteurs moléculaire

Une grande variété de descripteurs moléculaires a été calculée et déterminée à l'aide des logiciels Gaussian 09, ChemOffice et hyperchem [21][22][23], pour prédire la corrélation entre les descripteurs des molécules étudiées avec l'activité IC_{50} et les modèles de développement (régression linéaire multiple (RLM)) et non linéaire (réseau neuronal artificiel (RNN)). Le tableau 5 présent les descripteurs sélectionnés à utiliser dans cette étude.

4.2.3 Logiciels utilisés pour la génération

Dans ce travail on a utilisé des logiciels pour développer tableau 4, évaluer l'étude QSAR de dérivées de l'indazole : Logiciel pour générer la structure 3D, tracer, calculer et optimiser les structures chimiques des molécules testées, et une analyse de régression linéaire multiple des descripteurs moléculaires a été réalisée à l'aide de la stratégie par étapes de SPSS version 25 pour Windows[24]. Le tableau 4 présent les logiciels et les applications utilisées :

Tableau 4 : Logiciels utilisés dans ce travail

Trace de structures chimiques	ChemeDraw
Génère structures 3D	Gauss View 5.0
Calcule des descripteurs chimiques et physiques des molécules	Gaussian 09 , HyperChem
Développé le modèle QSAR	SPSS version 25

Nous avons utilisé deux logiciels pour déterminer les descripteurs physiques et chimiques des 28 molécules testées (voir le tableau ci-dessous) :

Tableau 5 : Logiciels utilisés pour générer les descripteurs

Logiciel	**Descripteurs**
HyperChem	-Polarisabilité moléculaire (**P**) $Å^3$, -Réfraction moléculaire (**MR**) A^3, -Coefficient de partage octanol/eau (***logP***), - Énergie d'hydratation (**HE) (**kcal/Mol), -Volume molaire (**MV**) $\mathbf{A^3}$, -Grille de surface (**SAG**) $\mathbf{A^2}$, -Masse moléculaire (**MM**) **uam.**
Gaussian 09	-E_{HOMO}, E_{LUMO} **(eV)** -DM **en (D)** -Charges atomiques nettes (**qN1, qN2, qC3,**

	qC4, qC5, qC6, qC7, qC8, qC9). **C** -Electronégativité $\chi = -\frac{(E_{LUMO}+E_{HOMO})}{2}(eV)$ -Dureté **η**, et de sa douceur inverse **S (eV)** - Egap =E_{HOMO}-E_{LUMO} **(eV)** -Indice d'électrophilicité : $\omega = \chi 2 / \eta\ 2$ (**eV**)

5 Résultats et discussion

5.1 Descripteurs générés

les vingt-huit molécules ont été pré-optimisées en utilisant le champ de force de la mécanique moléculaire (MM^+) inclus avec HyperChem la version 8.07 [21].Ensuite, les structures minimisées ont été affinées par des méthodes semi-empiriques. Les PM3 Hamiltoniennes également implémentées dans HyperChem, nous avons choisi une limite standard de gradient de 0.01 kcal/A pour l'optimisation de la géométrie, puis, pour les dérivées réoptimisées du 1, 2 benzodiazole à l'aide du logiciel Gaussian 09 [22], au niveau de la théorie de densité fonctionnelle DFT en utilisant les trois paramètres de Lee Becke-Parr (B3LYP), avec l'ensemble de base 6-31G . Cette théorie a été utilisée pour calculer un certain nombre de descripteurs électroniques : moment dipolaire (DM), E_{HOMO}, E_{LUMO} et charges atomiques nettes (qN1, qN2, qC3, qC4, qC5, qC6, qC7, qC8, qC9). Le module des propriétés QSAR de HyperChem 8.07 a été utilisé pour calculer : la polarisabilité moléculaire (P), la réfraction moléculaire (RM), le coefficient de partage octanol/eau (*LogP*), l'énergie d'hydratation (EH), le volume molaire (VM), surface moléculaire (SAG) et la masse moléculaire (MM). Dans le tableau ci-dessous, vous trouvez les paramètres qui sont calculés ou générés pour construire le modèle QSAR de deux activités biologiques étudiées A2780 et A549.

Tableau 6 : Descripteurs générer et calculer par gaussian et HyperChem

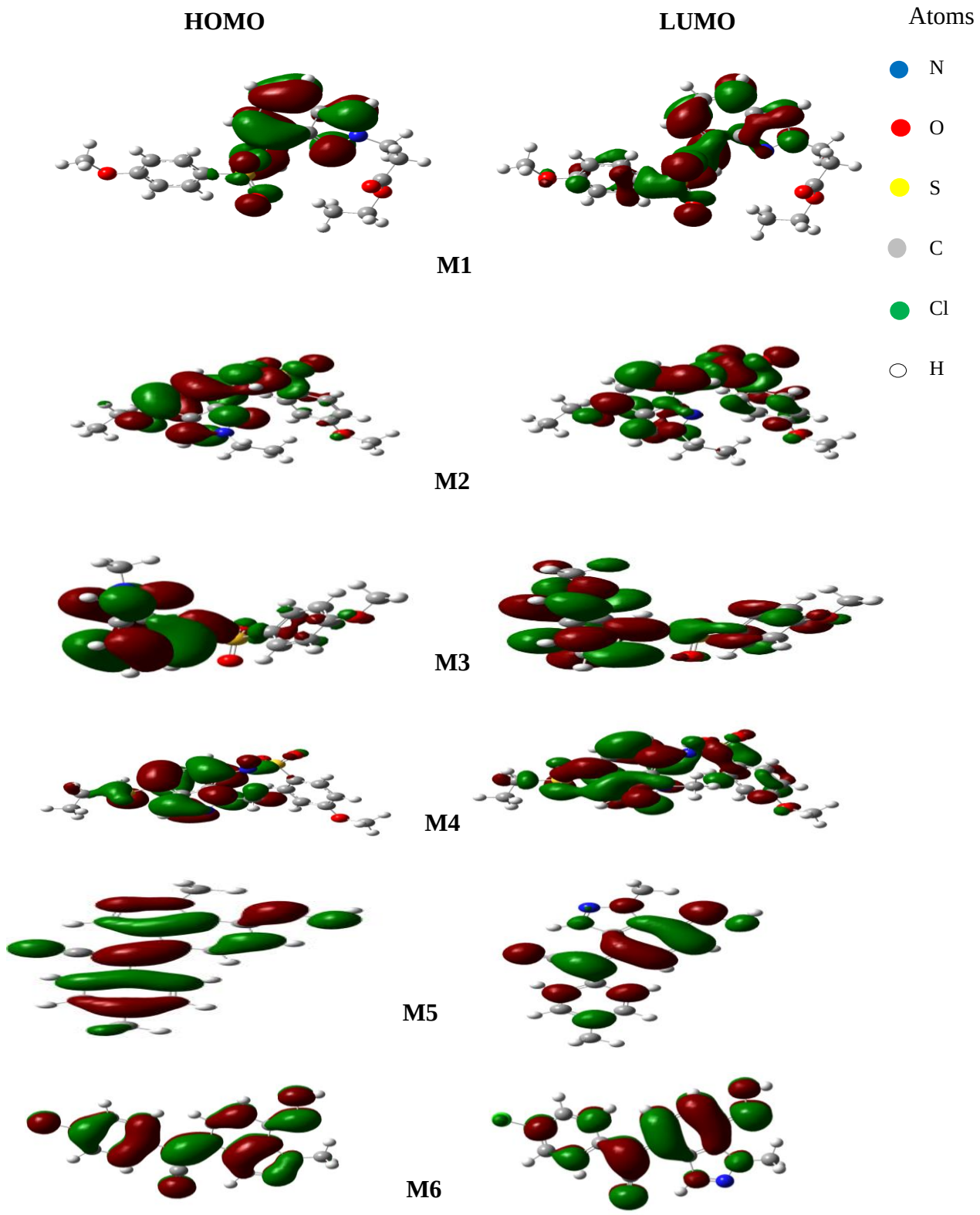
	EH	*LogP*	RM	P	MW	SAG	VM	DM
M1	-10,49	-0,89	112,04	38,30	403,45	683,32	1130,20	5,83
M2	-9,50	-1,29	107,61	36,38	375,44	631,32	1058,86	10,97
M3	-9,58	-0,98	91,74	30,24	317,36	535,63	878,52	9,07
M4	-8,55	-2,70	110,30	36,91	377,48	586,61	1006,27	4,02
M5	-10,50	-0,24	100,90	33,71	343,40	587,92	969,24	6,49
M6	-11,86	1,24	91,39	32,37	290,32	508,74	849,97	5,58
M7	-12,70	0,86	91,83	32,46	310,74	493,07	826,30	5,87
M8	-9,94	1,58	96,14	34,20	304,35	537,26	898,39	4,56
M9	-13,73	1,21	96,58	34,30	324,77	511,72	870,35	1,82
M10	-12,72	0,44	125,77	45,64	472,52	738,29	1282,28	10,22
M11	-15,98	-0,71	137,86	46,28	476,51	732,49	1261,51	10,63
M12	-11,17	1,21	134,11	45,09	464,93	702,73	1209,09	8,87
M13	-14,00	0,06	136,20	45,73	480,93	711,69	1230,42	9,56
M14	-12,45	0,78	140,51	47,47	474,53	729,43	1280,62	10,70
M15	-15,98	-0,71	137,86	46,28	476,51	730,79	1262,11	10,06
M16	-10,50	1,93	138,42	46,84	458,53	720,88	1254,74	9,74
M17	-16,53	-0,71	137,86	46,28	476,51	728,35	1262,92	13,02
M18	-13,80	0,44	135,77	45,64	460,51	717,62	1239,28	9,47
M19	-11,18	-1,41	92,63	30,33	337,78	479,14	809,50	8,32
P20	-14,27	-2,41	99,00	32,80	367,81	562,74	931,02	8,34
M21	-8,98	-0,32	97,52	32,36	328,39	564,89	933,69	8,62
M22	-7,41	-0,76	127,84	42,85	422,50	679,40	1173,05	8,45
M23	-15,50	-3,88	143,46	45,98	517,57	769,80	1325,50	8,15
M24	-9,11	-1,11	143,80	46,54	499,60	775,46	1331,18	7,76
M25	-8,52	-0,51	99,81	33,08	327,40	570,15	948,28	4,76
M26	-11,36	-1,65	101,90	33,71	347,40	590,22	971,78	5,53
M27	-6,89	-0,61	100,69	33,91	345,42	520,83	905,62	8,14
M28	-7,86	0,38	104,59	35,01	361,85	598,23	989,38	3,19

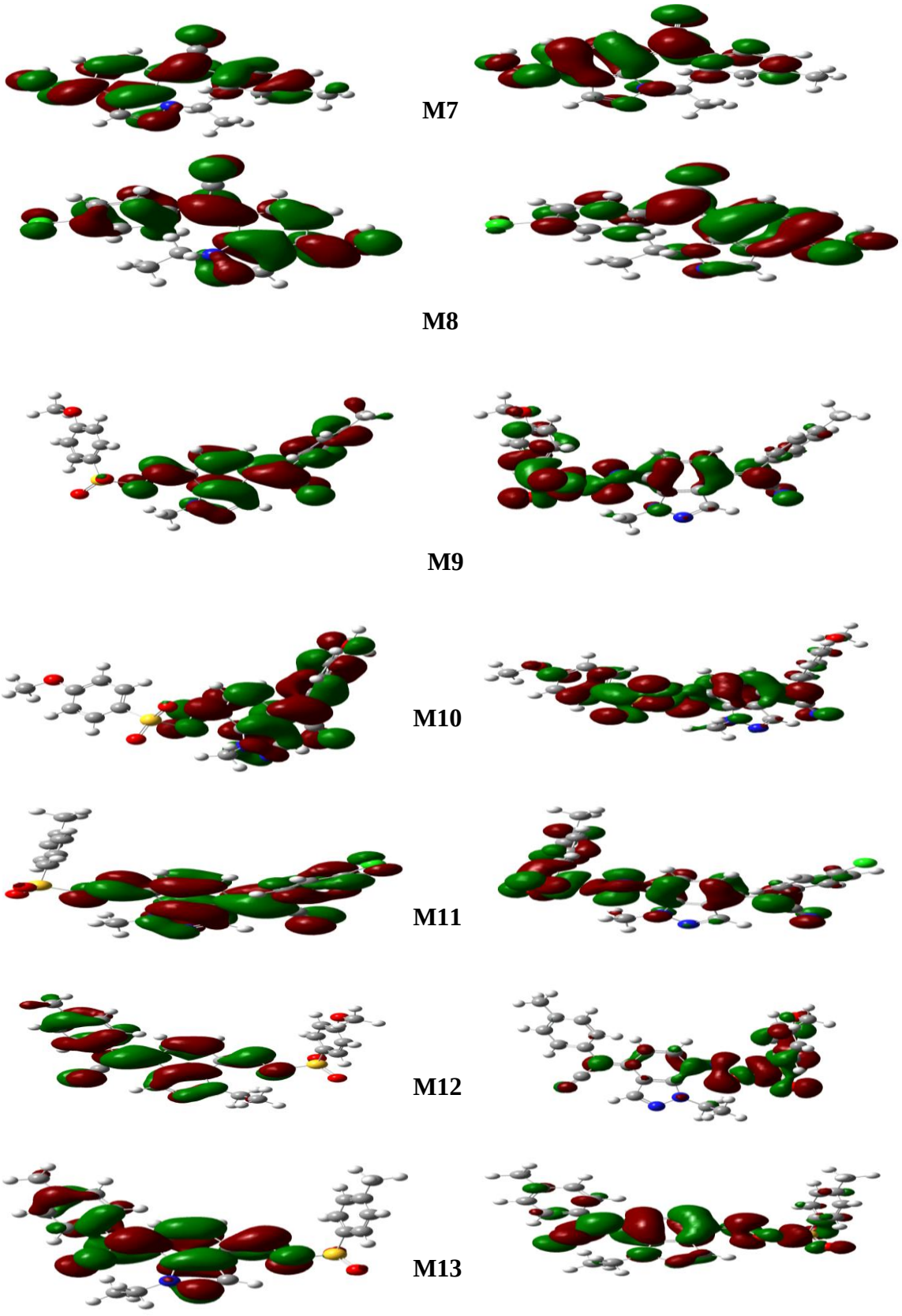
Tableau 7 : Charge des atomes de notre ensemble des molécules

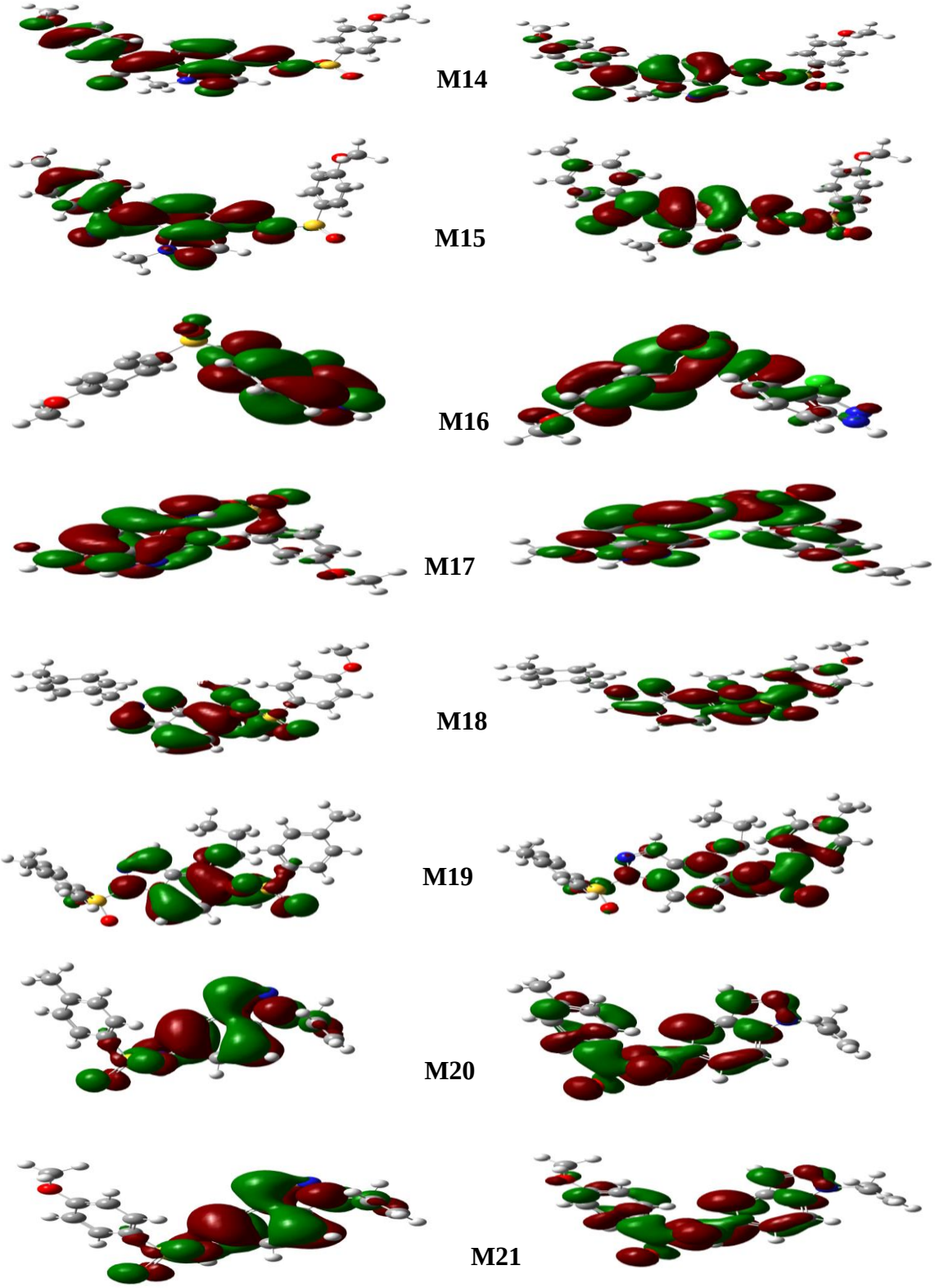
	qN1	q N2	qC3	qC4	qC5	qC6	qC7	qC8	qC9
M1	-0,3	-0,46	0,17	-0,06	-0,11	-0,15	-0,11	0,24	0,18
M2	-0,32	-0,46	0,2	-0,08	0,28	-0,16	-0,11	0,18	0,19
M3	-0,31	-0,23	0,21	-0,3	-0,07	-0,11	-0,07	0,19	0,23
M4	-0,4	-0,17	0,12	-0,24	-0,23	-0,1	0,01	0,07	0,43
M5	-0,31	-0,45	0,2	-0,09	-0,26	-0,15	-0,12	0,22	0,27
M6	-0,59	-0,22	0,06	-0,16	0,13	-0,14	-0,11	0,27	0,41
M7	-0,59	-0,22	0,07	-0,16	0,14	-0,14	-0,1	0,27	0,42
M8	-0,57	-0,21	0,07	-0,08	0,15	-0,1	-0,16	0,17	0,29
M9	0,29	-0,57	0,07	-0,06	0,14	0,14	-0,09	-0,16	0,17
M10	-0,6	-0,2	0,07	-0,2	0,15	-0,16	-0,11	0,23	0,48
M11	-0,6	-0,2	0,07	-0,2	0,15	-0,16	-0,11	0,23	0,48
M12	-0,6	-0,2	0,07	-0,2	0,15	-0,17	-0,11	0,23	0,48
M13	-0,6	-0,2	0,07	-0,2	0,16	-0,17	-0,11	0,23	0,48
M14	-0,6	-0,21	0,07	-0,21	0,15	-0,17	-0,11	0,23	0,49
M15	-0,6	-0,21	0,07	-0,2	0,15	-0,17	-0,11	0,23	0,48
M16	-0,59	-0,21	0,1	-0,07	0,15	-0,07	-0,2	0,22	0,33
M17	-0,59	-0,21	0,09	-0,06	0,15	-0,08	-0,18	0,21	0,33
M18	-0,59	-0,21	0,09	-0,06	0,15	-0,07	-0,2	0,23	0,32
M19	-0,65	-0,17	-0,09	-0,06	0,34	-0,14	-0,17	-0,05	0,36
M20	-0,64	-0,16	-0,11	0,01	0,15	-0,1	-0,17	0,36	0,33
M21	-0,3	-0,44	0,18	-0,05	-0,12	-0,11	0,11	0,38	0,16
M22	-0,29	-0,47	0,18	-0,07	-0,13	-0,1	0,14	-0,03	0,15
M23	-0,75	-0,2	0,1	-0,12	0,29	0,15	-0,13	-0,04	0,38
M24	-0,75	-0,2	0,1	-0,12	0,29	0,15	-0,13	-0,04	0,38
M25	-0,61	-0,2	0,06	-0,11	-0,09	0,13	-0,1	-0,07	0,39
M26	-0,61	-0,2	0,06	-0,11	-0,09	0,13	-0,1	-0,07	0,39
M27	-0,29	-0,47	0,23	-0,11	0,34	0,12	-0,11	-0,06	0,2
M28	-0,61	-0,17	-0,08	-0,08	-0,07	0,13	-0,1	-0,07	0,4

Tableau 8 : Résultats descripteurs électronique obtenus en (eV)

	E_{HOMO}	E_{LUMO}	EGap	χ	η	ω
M1	-5,714	-2,177	3,537	3,946	-3,946	7,784
M2	-5,442	-1,905	3,537	3,674	-3,674	6,747
M3	-5,714	-1,361	4,354	3,537	-3,537	6,257
M4	-5,987	-1,361	4,626	3,674	-3,674	6,747
M5	-5,442	-2,177	3,265	3,810	-3,810	7,257
M6	-5,714	-2,993	2,721	4,354	-4,354	9,478
M8	-5,987	-3,265	2,721	4,626	-4,626	10,700
M9	-5,987	-2,993	2,993	4,490	-4,490	10,080
M10	-6,259	-3,265	2,993	4,762	-4,762	11,338
M11	-5,987	-2,993	2,993	4,490	-4,490	10,080
M12	-5,714	-2,993	2,721	4,354	-4,354	9,478
M13	-6,259	-3,265	2,993	4,762	-4,762	11,338
M14	-5,987	-3,265	2,721	4,626	-4,626	10,700
M15	-5,987	-2,993	2,993	4,490	-4,490	10,080
M16	-5,987	-3,265	2,721	4,626	-4,626	10,700
M17	-5,987	-2,993	2,993	4,490	-4,490	10,080
M18	-5,714	-2,993	2,721	4,354	-4,354	9,478
M19	-5,987	-2,993	2,993	4,490	-4,490	10,080
M19	-5,987	-1,633	4,354	3,810	-3,810	7,257
M20	-5,714	-2,177	3,537	3,946	-3,946	7,784
M21	-5,170	-1,905	3,265	3,537	-3,537	6,257
M22	-5,714	-1,905	3,810	3,810	-3,810	7,257
M23	-5,987	-2,449	3,537	4,218	-4,218	8,895
M24	-5,987	-2,449	3,537	4,218	-4,218	8,895
M25	-5,987	-1,905	4,082	3,946	-3,946	7,784
M26	-5,987	-1,905	4,082	3,946	-3,946	7,784
M27	-5,442	-2,721	2,721	4,082	-4,082	8,330
M28	-6,259	-2,177	4,082	4,218	-4,218	8,895


HOMO
LUMO
Atoms
N
O
S
C
Cl
H
M1
M2
M3
M4
M5
M6


M7
M8
M9
M10
M11
M12
M13


M14
M15
M16
M17
M18
M19
M20
M21

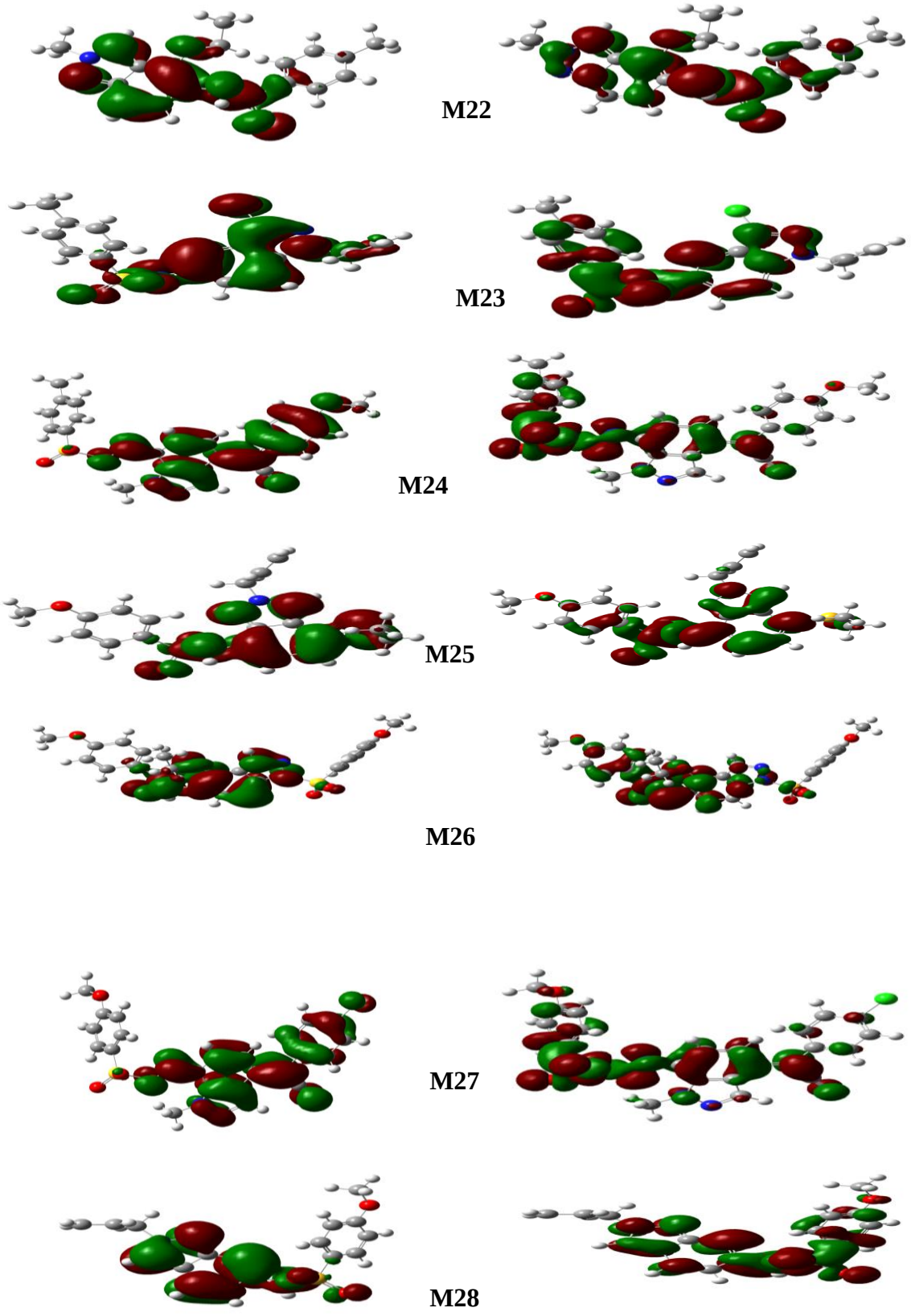


Figure 13 : Présentation spatiale d'HOMO et LUMO des molécules étudiées

5.2 Interprétation des résultats obtenus

La polarisabilité moléculaire d'une molécule caractérise la capacité de son système électronique à être déformé par le champ extérieur, et joue un rôle important dans la modélisation moléculaire de nombreuses propriétés et activités biologiques. La partie intéressante de l'interaction de Van Der Waals est une bonne mesure de la polarisabilité. La molécule de haute polarisabilité devrait avoir de fortes attirances avec d'autres molécules. La polarisabilité d'une molécule peut également améliorer la solubilité aqueuse. La réfractivité molaire (RM) est un critère important pour mesurer le facteur de stress. Elle est généralement considérée comme une simple mesure du volume occupé par un atome individuel ou un groupe d'atomes [36]. La polarisabilité et la réfractivité molaire augmentent relativement pour étudier la taille et le poids moléculaire tableau 7. Ce résultat est en accord avec la formule de Lorentz-Lorenz qui donne une relation entre la polarisabilité, la réfractivité molaire et le volume [24]. Cette relation montre que la polarisabilité et la réfraction molaire augmentent avec le volume et le poids moléculaire. Exemple d'indazole. Par exemple, pour le composé 6, la petite molécule de la série étudiée a une faible valeur de polarisabilité (32,37) et de réfractivité molaire (91,39) ; en revanche le compose 14 a des valeurs élevées de polarisabilité (47,47) et de réfractivité molaire (140,51).

La présence de groupes hydrophobes dans la structure d'indazole (inhibition de la prolifération des cellules A2780 et A549) induit une diminution de l'énergie d'hydratation, tandis que la présence de groupes hydrophiles augmente l'énergie d'hydratation (tableau 7).

L'énergie d'hydratation la plus élevée en valeur absolue (16,53 kcal/Mol) est celle du composé M17, mais la plus faible (6,89 kcal/Mol) a été atteinte pour le composé M27 (tableau 7). Dans le milieu biologique, les molécules d'eau entourent les molécules polaires où des liaisons hydrogène peuvent être établies entre la molécule d'eau et les molécules étudiées, l'eau et le complexe ayant la liaison hydrogène la plus forte. Au moins, ces molécules hydratées sont partiellement déshydratées avant leur interaction ; ces interactions à faible énergie sont généralement réversibles, en particulier entre les messagers et les récepteurs. La lipophilie est une propriété qui a un effet majeur sur la solubilité, l'absorption, la distribution, les propriétés de métabolisme et d'excrétion ainsi que sur l'activité pharmacologique. *Hansch et Leorayed* montrent que les molécules hautement lipophiles se répandent à l'intérieur des

membranes lipidiques et y restent pour une bonne biodisponibilité orale ,pour un *log* P plus élevé, le médicament a des difficultés à pénétrer les membranes lipidiques[25] . Contrairement à l'énergie d'hydratation, la présence de groupes hydrophobes (apolaires) dans la structure d'indazole induit une augmentation de la lipophilie (polaires) . Le composé M13 a un faible coefficient de partition (0,06), ce qui se traduit par une meilleure tolérance gastrique. Les composés M23 qui ont une valeur plus élevée (3,88), ont des capacités dépendantes des protéines plasmatiques.

5.3 Analyse descriptive

5.3.1 Analyse en composantes principales (ACP)

Analyse en composantes principales (ACP) a été utilisée pour déterminer la non linéarité des descripteurs et pour sélectionner les descripteurs qui sont en corrélation avec l'activité et dépendant à l'activité biologique soit A2780 et A549.

Le tableau 9 présente les descripteurs avec un coefficient de corrélation avec l'activité supérieure, en valeur absolue, à 0,1. L'absence de colinéarité sérieuse entre les descripteurs présents dans le modèle a été confirmée par la matrice de corrélation.

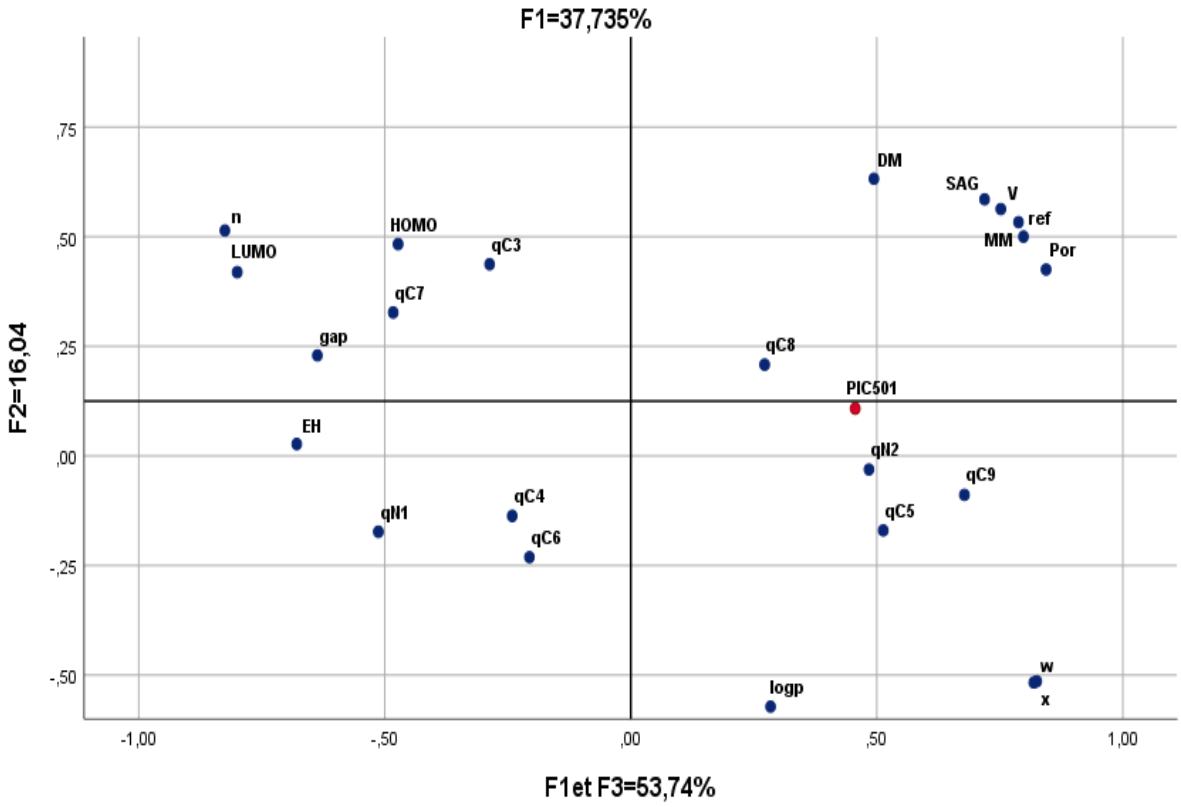
Tableau 9 : Coefficient de corrélation

pIC50(OVR) A2780		**pIC_{50}(POUM)** A549	
Descripteurs	**Coefficient de corrélation**	**Descripteurs**	**Coefficient de corrélation**
EH	0,380	**EH**	0,411
Log P	0,060	***Log P***	0,003
RM	0,316	**RM**	0,603
P	0,333	**P**	0,579
MW	0,328	**MW**	0,608
SAG	0,184	**SAG**	0,454
VM	0,231	**VM**	0,503
DM	0,464	**DM**	0,526
qC5	0,195	**qC5**	0,268
qC4	0,021	**qC4**	0,120
qC3	0,446	**qC3**	0,339
qN2	0,390	**qN2**	0,470
qN1	0,312	**qN1**	0,488
qC9	0,406	**qC9**	0,501
qC8	0,338	**qC8**	0,321
qC7	0,026	**qC7**	0,026
qC6	0,435	**qC6**	0,450
E_{HOMO}	0,105	**E_{HOM}**	0,176
E_{LUMO}	0,257	**E_{LUMO}**	0,269
EGap	0,226	**Egap**	0,208
χ	0,249	**χ**	0,283
η	0,249	**η**	0,283
Ꞷ	0,249	**Ꞷ**	0,285

- ACP pour pIC_{50} d'A2780 activité biologique contre tumeur d'ovarienne

La ACP de ces données peut être représentée par une matrice X composée de p =23 variables descripteurs et n = 28 molécules (les composés synthétisés). Les analyses statistiques ont été effectuées en utilisant SPSS.

La ACP montre que 53.74 % des informations contenues dans le tableau 5 peuvent être représentés en deux composantes principales (PC). Dans la figure 14 présente le graphique des scores de PC1 (37.73 % de la variance) et PC2 (16.35 % de variance), et montre aussi 50,74 % des informations représentés en deux composantes prancipales PC1par 37,73% et PC3 par 13,01%.



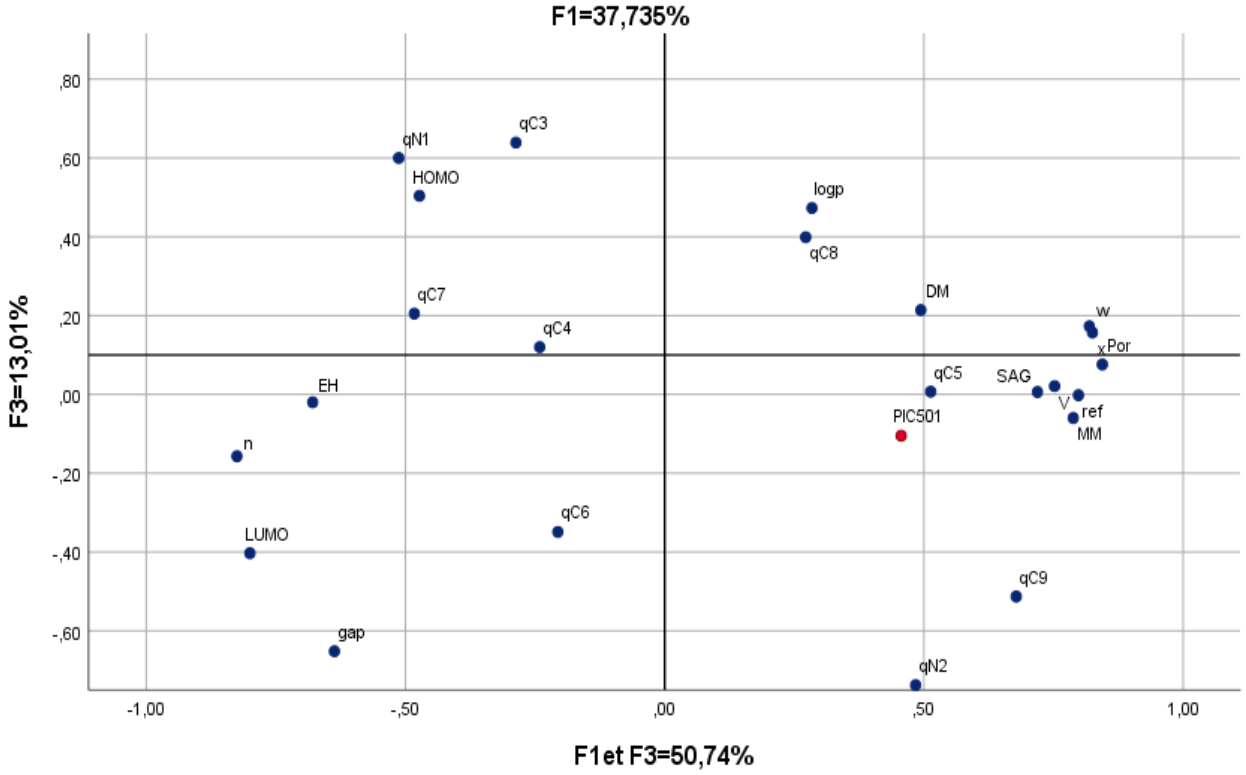
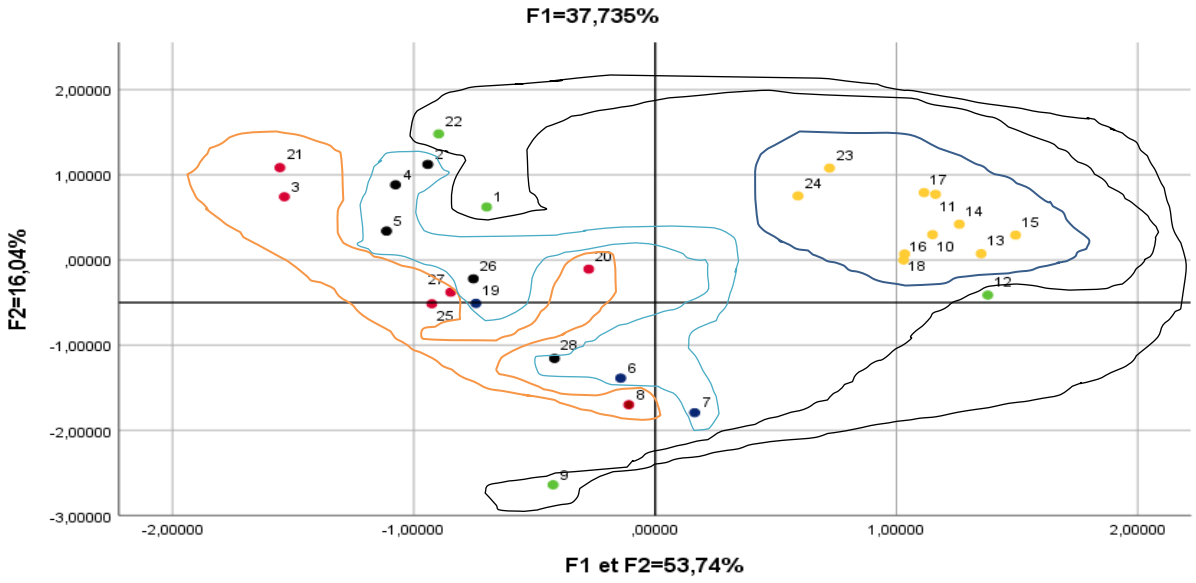


Figure 14 : Représentation des descriptions en cercles de corrélation pour A2780

Dans la figure 15, la projection des composés dans les trois premiers axes, F1, F2 et F3 pour A2780 montre le regroupement des molécules à 4 groupes.



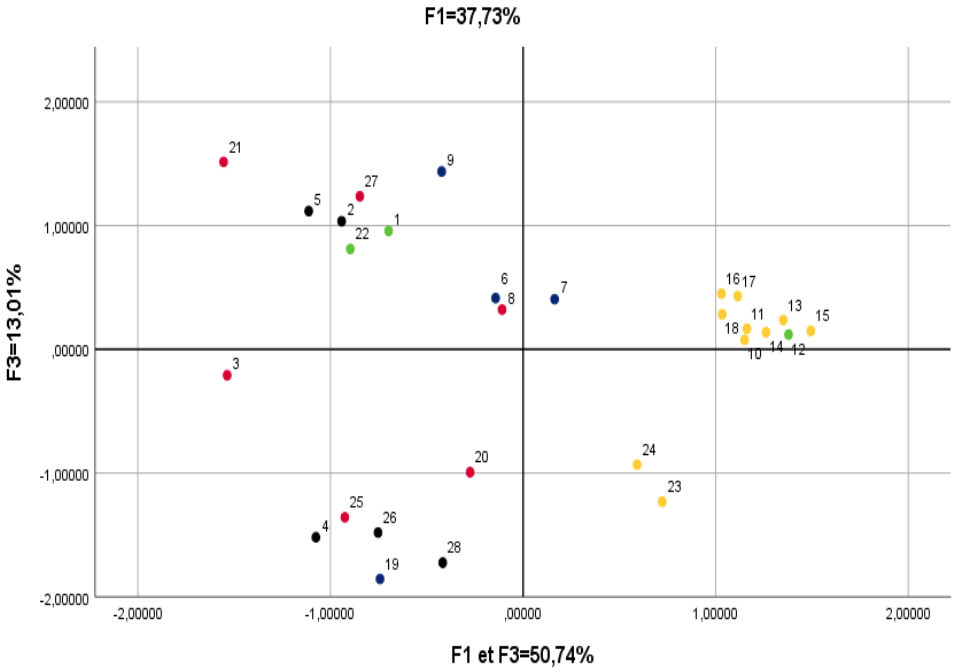


Figure 15 : Projection des composés dans les trois premiers axes, F1, F2 et F3 pour A2780

- ACP pour pIC_{50} d'A549 l'activité biologique contre cancer de poumon

Les projections des composés (ensembles d'entraînement et de test) dans les trois premiers axes, F1, F2 et F3, sont illustrées dans la figure 16. La ACP montre que 57.19 % des informations contenues dans le tableau 5 peuvent être représentés en deux composantes principales (PC). Dans la figure 16 présente le graphique des scores de PC1 (40.01 % de la variance) versus PC2 (17.18 % de variance), et montre aussi 52,96 % des informations représentés en deux composantes prancipales PC1 par 40,01% et PC3 par 12,18%.

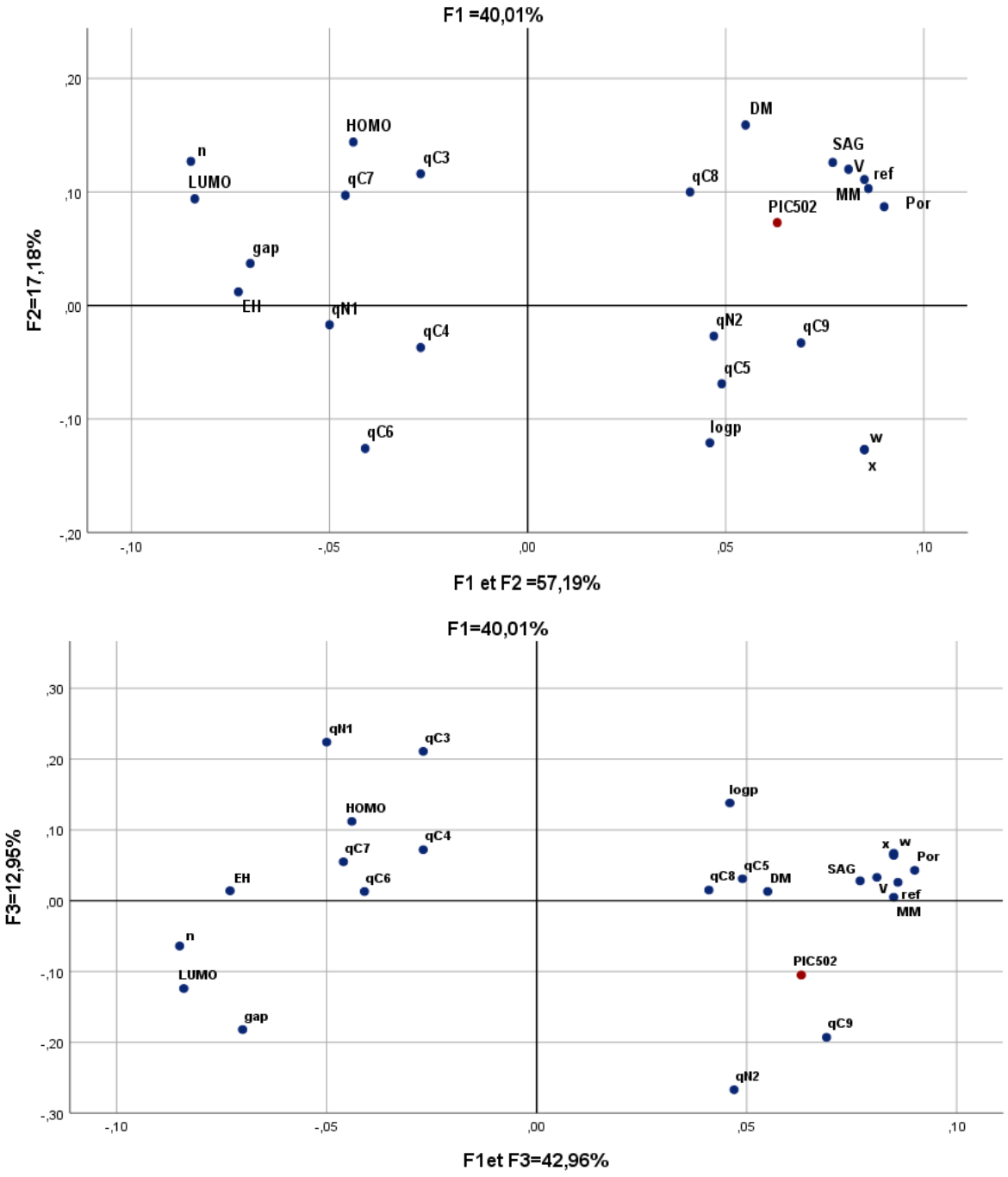


Figure 16 : Projection des composés dans les trois premiers axes, F1, F2 et F3 pour A549.

Dans la figure 17, la projection des composés dans les trois axes, F1, F2 et F3 pour A2780 represente 5 groupes homogènes des molecule etudiees.

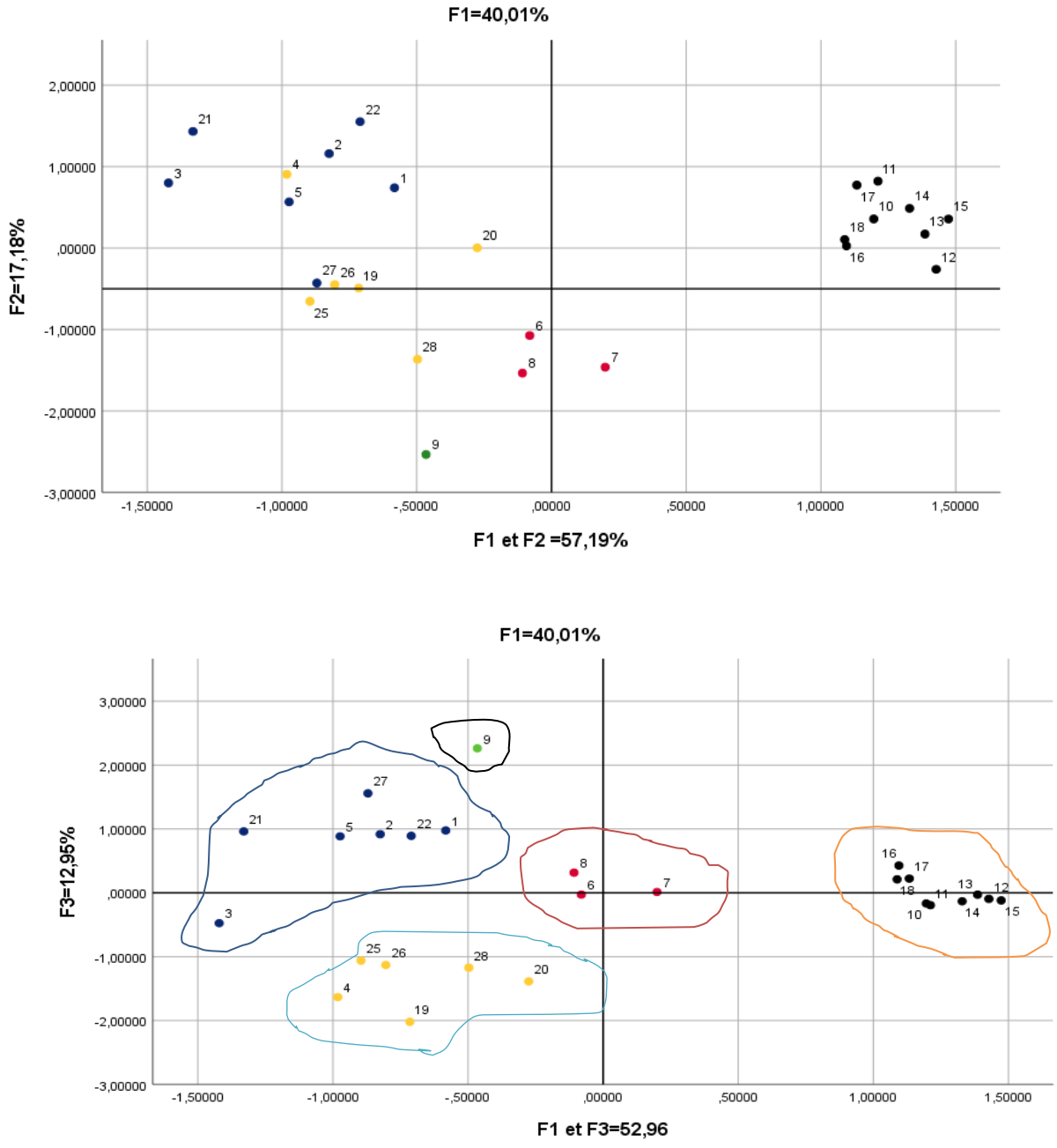


Figure 17 : Projection des composés dans les trois premiers axes, F1, F2 et F3 pour A549.

5.3.2 Classification des données (K-means)

- K-means pour les deux activités

La méthode k-means est utilisée, dans cette étude, pour diviser les observations en groupes homogènes, en fonction de leur description par les descripteurs utilisés. Pour la division de l'ensemble de données en sous-ensembles d'un ensemble d'entraînement (80 % des composés) et d'un ensemble de test (20 % des composés), on a utilisé une combinaison des résultats de la méthode des k-means (tableau 10). L'ensemble de tests est constitué de cinq molécules sont :1 ;10 ;9 ;4 ;20 pour A2780 et 9 ;20 ;3 ;13 ;8 pour A549, les molécules restantes (28 molécules) sont l'ensemble de formation.

Tableau 10 : Classes sélectionnées

	pIC_{50} (tumeur d'ovrainne) A2780 **1 ;10 ;9 ;4 ;20**	**pIC_{50} (cancer Poumon)** A549 **9 ;20 ;3 ;13 ;8**
1	**1** ;12;22	**9**
2	**10**;11;13;14;15;16;17;18;23;24	4;19;**20**;25;26;28
3	6;7;**9**;19	1;2;**3**;5;21;22;27
4	2;**4**;26;5;28	10;11;12;**13**;14;15;16;15;18
5	3;8;**20**;21;25;27	6;7;**8**

5.4 Elaboration et évaluation des modèles

Au cours de cette étape, des modèles QSAR linéaires et non linéaires ont été développés et évalués pour prédire les activités des composés d'essai ou tester .la régression linéaire (RLM) disponible dans le logiciel SPSS [24] et le réseau neuronal artificiel (RNN) disponible dans le logiciel lui-même logiciel .

Afin de proposer des modèles et d'évaluer quantitativement les effets physico-chimiques des composées sur les activités IC_{50} biologiques des molécules, on a soumis la matrice de données constituée évidemment à partir des descripteurs correspondant aux données des molécules à l'analyse de la RLM et à l'RNN. Les coefficients des corrélations $\boldsymbol{R}$, $\boldsymbol{R^2}$ et $\mathbf{R^2_{adj}}$ est la valeur permettant de sélectionner la meilleure performance de régression [29], ce qui donne une

indication de la probabilité qu'un QSAR soit corrélé. Afin d'évaluer l'importance des modèles et la précision des prédictions pour les nouveaux composés :

1. Nous utilisons une procédure de validation interne (validation croisée "leave one out"), dans laquelle un composé est éliminé et le modèle reconstruit avec les molécules restantes est utilisé pour prédire la réponse du composé éliminé, qui revient ensuite et un deuxième est éliminé, et le cycle est répété, et ainsi de suite, jusqu'à ce que tous les composés aient été éliminés un par un, et qu'un coefficient de corrélation global $\boldsymbol{R_{cv}}$ soit calculé ;

2. Après la construction du modèle, une prédiction externe est nécessaire. Dans celui-ci, le modèle obtenu a été utilisé pour prédire les activités d'un ensemble de test comprenant des composés similaires à ceux qui ne sont pas utilisés dans l'ensemble de formation. Cette opération est généralement réalisée en divisant un ensemble de données en un ensemble d'entraînement et un ensemble de test, généralement dans un rapport de 1/5. En outre, avant de procéder à la validation externe du modèle.

3. Si le modèle présente une erreur systématique élevée (biais), il doit être écarté et les tests de validation externes doivent être effectués à partir d'un modèle de souris non biaisé. *Xternal Validation Plus* est un outil qui vérifie la présence d'erreurs systématiques dans le modèle et calcule ensuite tous les paramètres de validation externe nécessaires.

5.4.1 Régression linéaire multiple (RLM)

a) Régression linéaire multiple (RLM) d'A2780

Tout d'abord, différentes molécules de derivés du 1,2-benzodiazole (tableau 3) ont été évaluées pour leur activité inhibitrice. Le paramètre biologique IC_{50} de A2780 a été introduit dans cette recherche et les résultats sont illustrés dans le tableau 6 ,7 et 8**.** Afin de déterminer le rôle des caractéristiques structurelles, une série de 28 molécules de dérivés du 1,2-benzodiazole, a été étudiée par la méthode QSAR. Ces composés ont été utilisés pour générer des modèles de régression multilinéaire. Différents descripteurs physico-chimiques ont été utilisés comme variables indépendantes et corrélées avec l'activité biologique. Le développement d'un modèle QSAR nécessite un ensemble de données diversifiées, et donc un grand nombre de

descripteurs doivent être pris en compte. Les descripteurs sont des valeurs numériques qui codent différentes caractéristiques structurelles des molécules.

La sélection d'un ensemble de descripteurs appropriés qui forment un grand nombre d'entre eux nécessite une méthode, qui est capable de discriminer entre les paramètres. La matrice de corrélation de *Pearson* a été réalisée sur tous les descripteurs à l'aide du logiciel SPSS. L'analyse de la matrice a révélé 23 descripteurs pour le développement du modèle RLM. La valeur des descripteurs sélectionnés pour le modèle MLR est présentée dans les tableaux 6,5,4.

- La corrélation entre l'activité biologique (pIC_{50}) A2780 de tumeur d'ovarienne et les descripteurs s'exprime par la relation suivante :

(40)

$$\mathbf{pIC_{50}} = 4{,}177 - 1{,}753\ \mathbf{qC6} - 6{,}637\mathbf{qC3} + 0{,}187\mathbf{*DM} + 2{,}630\ \mathbf{qC7}$$

Avec r = 0.860 r^2=0,740 et n = 23 EQM =0, 6948

Équation 39 : Relation de la première activité étudiée

Le modèle QSAR ayant R^2<0,6 ne sera pris en compte que pour la validation. Par exemple, la valeur r = 0,860 et r^2 = 0,740 nous a permis d'indiquer fortement la corrélation entre différents paramètres (variables indépendantes) avec IC_{50} de A2780 des composés.

Dans le modèle QSAR, les facteurs positifs de DM et qC7 et les facteurs négatifs qC3 et qC6, augmentation de qC3 et qC6 des composés entraînent une diminution de l'activité biologique A2780. Pour tester la validité prédictive d'un privilège du modèle RLM sélectionné (eq. PIC_{50}), la technique leave-one-out (LOO) a été utilisée. Les modèles développés ont été validés en calculant les paramètres statistiques suivants : la somme résiduelle estimée des carrés (PRESS), la somme totale de l'écart quadratique (SSY) et le coefficient de corrélation croisée validé (R^2_{adj}) (tableau 11).

Tableau 11 : Coefficients de corrélation pour modèle 1

Modéle	PRESS	SSY	PRESS/SSY	R^2_{CV}	R^2_{adj}
Coefficients	3,098	11.8	0,260	0.740	0.682

PRESS est un paramètre de validation croisée important, car il constitue une bonne approximation de l'erreur de prédiction réelle du modèle. Sa valeur inférieure à SSY indique que le modèle prédit mieux que le hasard est peut-être considéré comme statistiquement significatif, plus la valeur de PRESS est faible, plus le modèle est prévisible. Sur la base des résultats décrits dans le tableau 11, le modèle est statistiquement significatif.

En outre, pour un modèle QSAR raisonnable, le rapport PRESS / SSY devrait être inférieur à 0,4. Les données présentées dans le tableau 11 indiquent que pour le mode développé, ce rapport est de 0,260. Notre résultat R^2_{CV} pour ce modèle QSAR est de 0,682. La valeur élevée de R^2_{CV}= 0,740 et de R^2_{adj} = 0,682 est un critère essentiel pour la quantification optimale du modèle QSAR.

La figure 18 montre les courbes de régression linéaire prédites par rapport à la valeur expérimentale de l'activité biologique du 1,2-benzodiazole décrite ci-dessus. Les graphiques de ce modèle montrent qu'elles sont plus pratiques avec R^2 = 0,740. Il indique que le modèle peut être appliqué avec succès pour prédire l'activité pIC_{50} de ces composés.

b) Molécules de test d'A2780

Tableau 12 : Valeurs des descripteurs chimiques et les activités observées de (A2780) et prévues à l'aide des modèles RLM pour l'ensemble d'essai.

Test validation (ovarienne) (A2780)						
N°	**qC3**	**qC7**	**qC6**	**DM**	**Obsrv**	**RLM**
1	0,17	-0,11	-0,15	5,83	4	4,11
10	0,12	0,01	-0,1	4,02	5,4	4,33

9	0,07	-0,09	0,14	1,82	4,25	3,57
4	0,07	-0,11	-0,16	10,22	5,32	5,61
20	-0,11	-0,17	-0,1	8,34	5,91	6,19

En comparant l'importance de chaque descripteur sur la pIC_{50} des 1,2-benzodiazole, nous devons connaître le coefficient normalisé et les valeurs t-test de celles-ci dans les équations de la RLM. Plus la valeur absolue de la valeur t-test est grande, plus l'influence du descripteur est importante.

Les valeurs de t-test qui sont calculés à l'aide de SPSS sont respectivement de 3,647 ; -4,261 ; 1,799 et -1,305 pour DM ; qC3 ; qC7 et qC6 respectivement. Cela signifie que les valeurs t-test de DM, qC3 sont toutes supérieures à celles des autres descripteurs, ce qui indique que, dans ce modèle, l'influence de ces descripteurs sur l'activité est plus forte que celle des autres. Cela montre également l'importance de la ramification dans l'atome C3 et DM dans la prédiction de la pIC_{50}.

En conclusion, ces résultats illustrent que, pour augmenter l'activité pIC_{50} A2780 contre le cancer de poumon on va augmenter la ramification de C3.

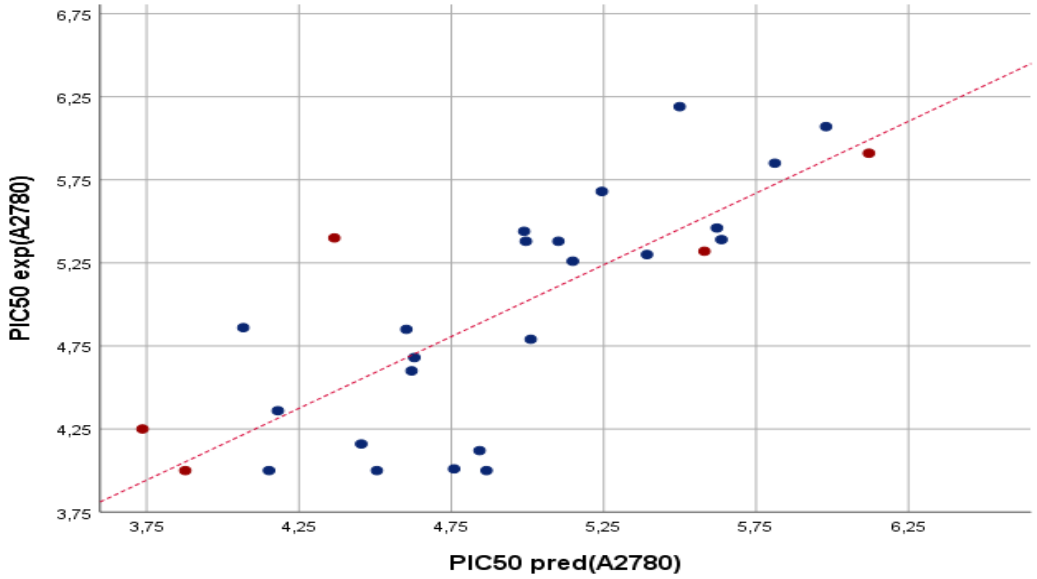


Figure 18 : Courbe de corrélation de l'activité biologique observée d'ovarienne en fonction de l'activité biologique perdit d'ovarienne

Une comparaison des valeurs du pIC_{50} (test) et du pIC_{50} (observe) montre que le modèle a fait de bonnes prédictions pour les 5 composés : pour la RLM : N= 5 ; r test= 0,827 ; r^2 test = 0,683.

Les résultats obtenus par la RLM nous permettent de conclure que le modèle fonctionne bien, ce qui est confirmé par les résultats obtenus lors des tests des 5 composés testés. Même si ce bon pouvoir prédictif est le fruit du hasard, on peut affirmer qu'il s'agit d'un résultat positif. En conséquence, ce modèle pourrait être appliqué à tous les dérivés de 1,2-benzodiazole du figure 19 et ajouter des connaissances supplémentaires pour améliorer la recherche de médicaments anti cancer d'ovarienne.

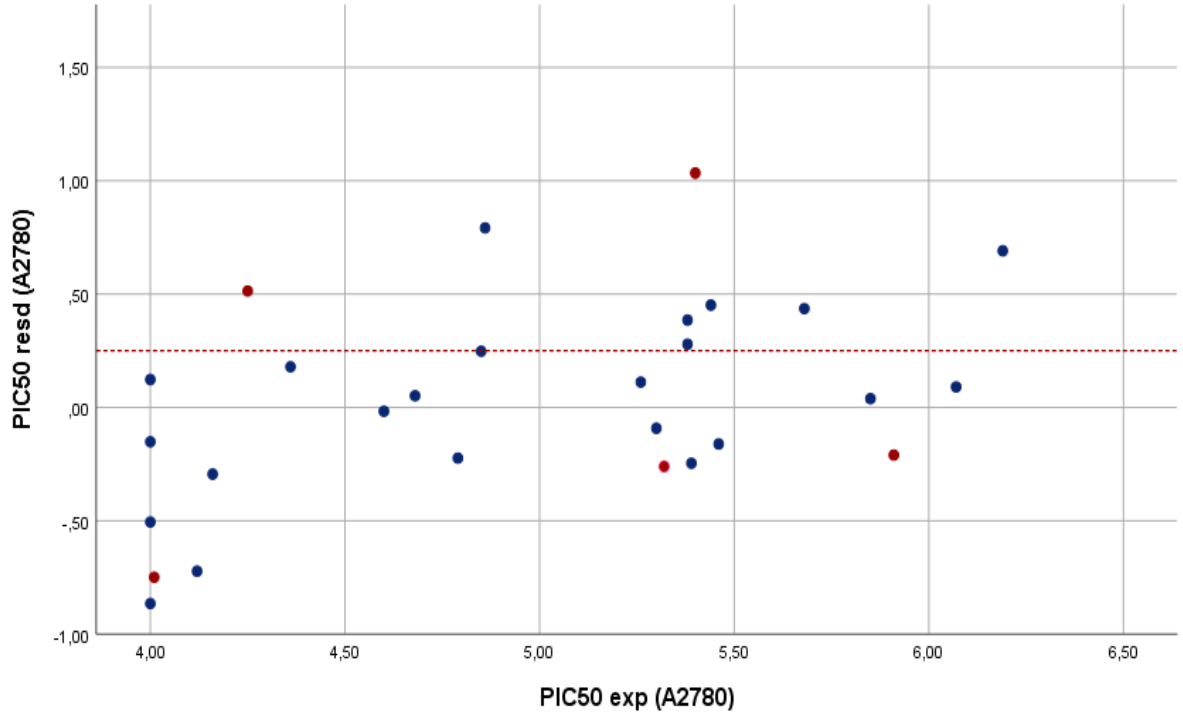


Figure 19 : Courbe de la corrélation entre l'activité biologique prédite d' A2780 et les résiduels

c) Régression linéaire multiple (RLM) d'A549

La corrélation entre l'activité biologique pIC_{50} de de tumeur poumon (A549) et les 24 descripteurs ou paramètres physiques et chimiques de 26 molécules d'indazole de s'exprime par équation mathématique par méthode étape par étape et logiciel SPSS suivante :

(41)

$$\mathbf{pIC_{50}=}\ 7{,}323\text{-}0{,}031\mathbf{*SAG}\ \text{-}2{,}638\mathbf{*\ qC6}\ +3{,}351\mathbf{*qN2}+3{,}351\mathbf{*qC7}+4{,}332\mathbf{*qC4}+0{,}018\mathbf{V}$$

Avec r= 0,945 r^2 = 0,893 et n=21 EQM=0, 613

Équation 40 : Relation de la deuxième activité étudiée

Le modèle QSAR ayant $R^2 < 0{,}6$ ne sera pris en compte que pour la validation. Par exemple, la valeur r=0.945 et r^2=0,893 nous a permis d'indiquer fortement la corrélation entre différents paramètres (variables indépendantes) avec l'inhibition de A549 des composés.

Dans le modèle QSAR, les coefficients négatifs de SAG, qC6, les coefficients positifs de qN2, qC7, V et qC4, expliquent que toute augmentation de qN2, qC7, V et qC4 des composés entraîne une diminution de l'activité biologique et augemantation de PIC_{50}. Pour tester la validité prédictive d'un privilège du modèle RLM sélectionné (eq. pIC_{50}), la technique leave-one-out (LOO) a été utilisée. Les modèles développés ont été validés en calculant les paramètres statistiques suivants : la somme résiduelle estimée des carrés (PRESS), la somme totale de l'écart quadratique (SSY) et le coefficient de corrélation croisée validé (R^2adj) (tableau 13).

Tableau 13 : Coeficients de corrélation de modèle 2

Modèle	**PRESS**	**SSY**	**PRESS/SSY**	$\mathbf{R^2_{CV}}$	$\mathbf{R^2_{adj}}$
Coefficients	0,775	7,270	0,1066	0,893	0,848

PRESS est un paramètre de validation croisée important car il constitue une bonne approximation de l'erreur de prédiction réelle du modèle. Sa valeur inférieure à SSY indique que le modèle prédit mieux que le hasard est peut-être considéré comme

statistiquement significatif, plus la valeur de PRESS est faible, plus le modèle est prévisible. Sur la base des résultats décrits dans le tableau 13, le modèle est statistiquement significatif.

En outre, pour un modèle QSAR raisonnable, le rapport PRESS / SSY devrait être inférieur à 0,4. Les données présentées dans le tableau 13 indiquent que pour le mode développé, ce rapport est de 0,1066. Notre résultat R^2_{CV} pour ce modèle QSAR est de. La valeur élevée de R^2_{CV} =0,893 et de R^2adj = 0,848 est un critère essentiel pour la quantification optimale du modèle QSAR.

d) Molécules de test (validation externe)

Tableau 14 : Valeurs des descripteurs chimiques et les activités observées d'A549 et prévues à l'aide des modèles RLM pour l'ensemble d'essai

Test validation (POUM)(A549)								
C	**SAG**	**V**	**qC4**	**qN2**	**qC7**	**qC6**	**Obsrv**	**RLM**
9	511,72	870,35	-0,06	-0,57	-0,09	0,14	4,1	4,28508
3	535,63	878,52	-0,3	-0,23	-0,07	-0,11	4	4,51711
8	537,26	898,39	-0,08	-0,21	-0,16	-0,1	4,14	5,51633
20	562,74	931,02	0,01	-0,16	-0,17	-0,1	5,23	5,83771
13	711,69	1230,42	-0,2	-0,2	-0,11	-0,17	5,28	5,95142

En comparant l'importance de chaque descripteur sur la pIC_{50} (A549) des 1,2-benzodiazole, nous devons connaître le coefficient normalisé et les valeurs t-test de celles-ci dans les équations de la RLM. Plus la valeur absolue de la valeur t-test est grande, plus l'influence de descripteur est importante.

Les valeurs de t-test qui sont calculés à l'aide de SPSS sont respectivement de -5,628 ; -4,775 ; 6,424 ; 4,573 ; 3,907 ; 5,908, pour SAG, qC6, qN2, qC7 et qC4 respectivement. Cela signifie que les valeurs t-test de SAG, qN2 et qC4 sont toutes deux supérieures à celles des autres descripteurs, ce qui indique que, dans ce modèle, l'nfluence de

ces descripteurs sur l'activité est plus forte que celle des autres. Cela montre également l'importance de la ramification dans l'atome qN2 dans la prédiction de la pIC50.

En conclusion, ces résultats illustrent que, pour augmenter l'activité pIC_{50} A549 contre le cancer de poumon on va diminuer la surface des molécules et la ramification de N2.

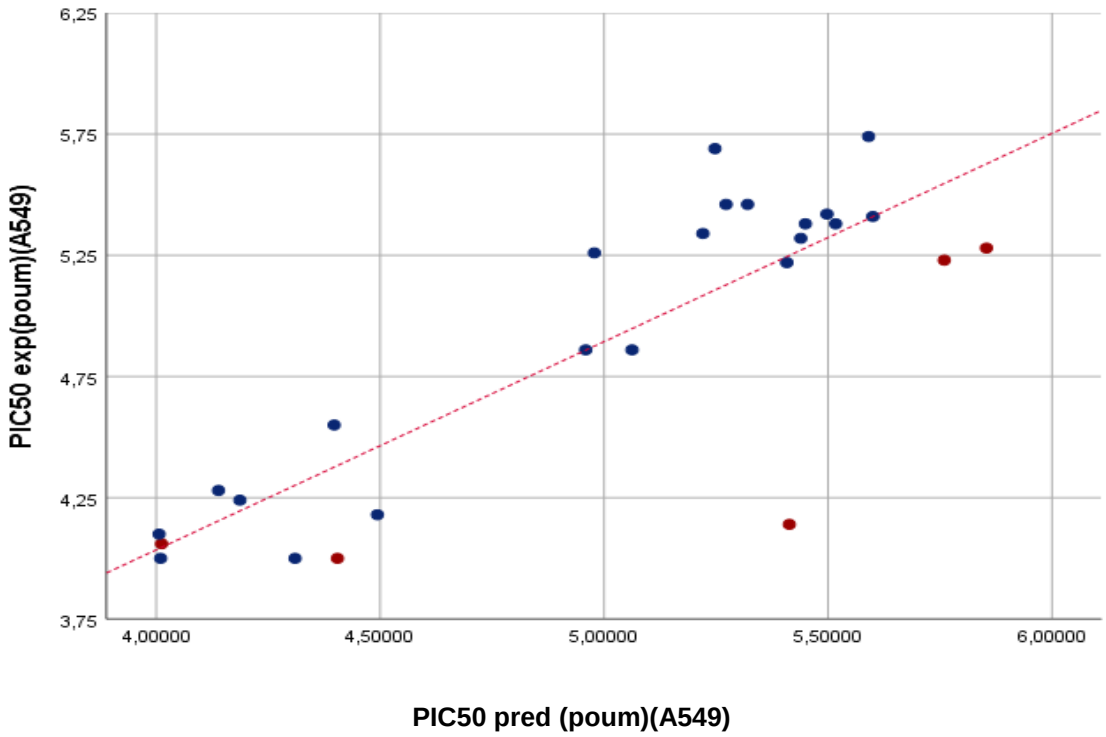


Figure 20 : Courbe de corrélation de l'activité biologique observée A549 en fonction de l'activité biologique prédite (A549) rouge pour les données de tests et Bleu pour les molécules de traitement.

Le véritable pouvoir prédictif de ces modèles est de tester leur capacité à prédire parfaitement la pIC_{50} des composés à partir d'un ensemble de tests externes (composés qui n'ont pas été utilisés pour le modèle développé) ; les pIC_{50} de l'ensemble restant de 5 composés sont déduites des modèles quantitatifs proposés avec les composés utilisés dans l'ensemble de formation par la RLM. Ces modèles seront en mesure de prédire les activités des molécules de l'ensemble de test en accord avec la valeur déterminée expérimentalement. Les valeurs de pIC_{50} observées et calculées sont indiquées dans le tableau 14. La capacité prédictive des modèles a été jugée, la valeur la plus élevée de r_{test} et r^2_{test} : r^2_{test}=0,679 r=0,824.

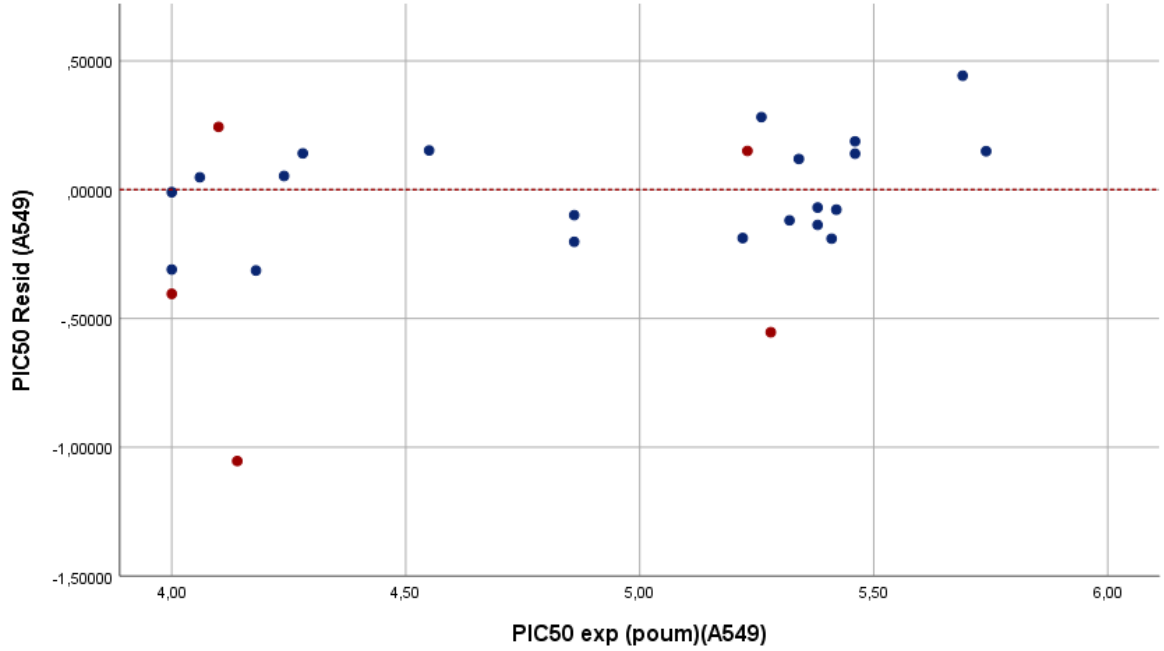


Figure 21 : Courbe de corrélation de résidus de l'activité biologique en fonction de l'activité biologique observée de poumon (A549) rouge pour les données de tests et Bleu pour les molécules de traitement.

5.4.2 Réseau de neurones artificiels (RNN)

Le résultat du calcul et les performances des modèles établis sont enregistrés dans la couche de sortie. Pour vérifier la qualité prédictive des modèles, les données totales sont réparties de manière aléatoire en deux groupes. Le premier groupe (80% des données totales) est utilisé pour piloter le système, les 20 % restants qui n'ont pas participé aux modèles d'apprentissage seront utilisés comme test indépendant de généralisation du réseau. La répartition des données totales entre les ensembles de formation, de validation et de test est présentée dans le tableau 15. La corrélation entre les valeurs expérimentales et calculées à l'aide des modèles de réseaux neuronaux artificiels est très significative, comme l'illustre la Figure 22 et comme l'indiquent les meilleures valeurs de r et r^2 et les petites valeurs d'EQM pour les deux phases : validation et test.

a) Pour activité biologique pIC_{50} A2780 contre tumeur d'ovarienne.

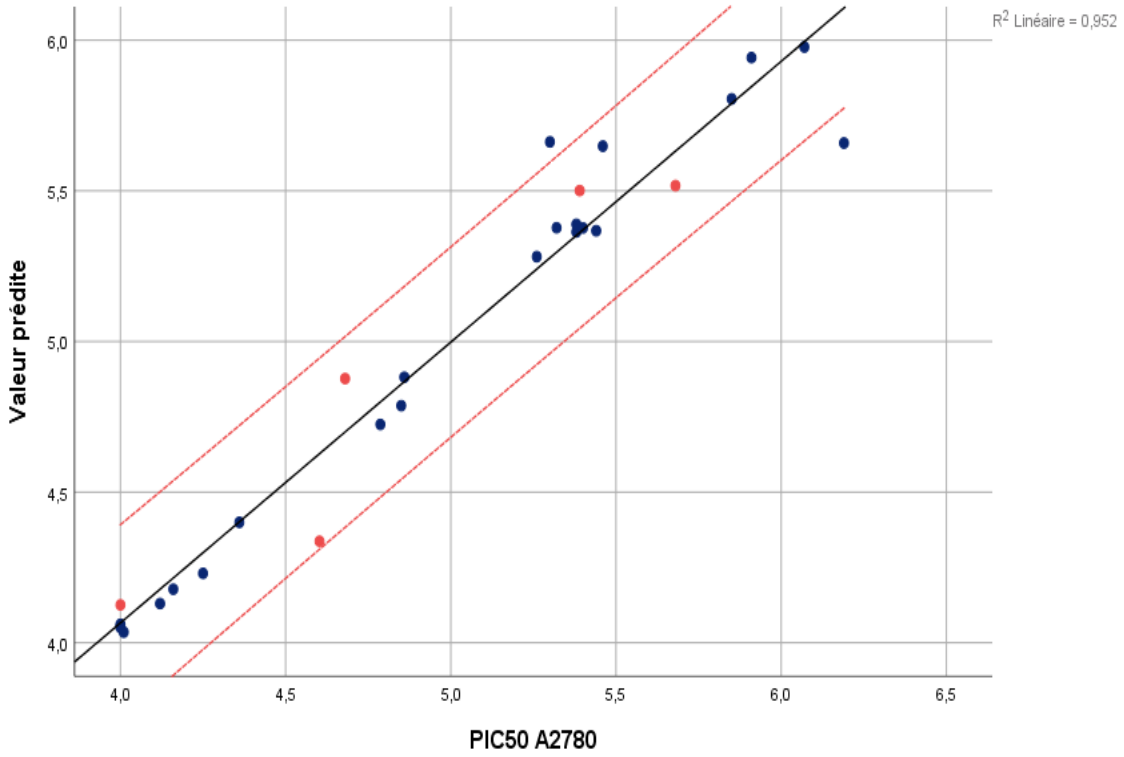


Figure 22 : Corrélations entre les valeurs d'activités observées d'A2780 et prédites calculées à l'aide de modèles RNN (ensemble validation en bleu, ensemble de test)

Tableau 15 : Coefficients générés par RNN d'A2780

Modèle r=0,886 r^2=0,952	
Données de validation 1.2.4.5.6.8.9.10.11.13.15.16.17.18.19.20.21.22.23.24.25.27.28.	EQM=0,141 R^2=0,952
Données de test 12,14,7,26,3	EQM=0,175 R^2=0,909

b) Pour la 2^{emme} activité pIC$_{50}$ (A549)

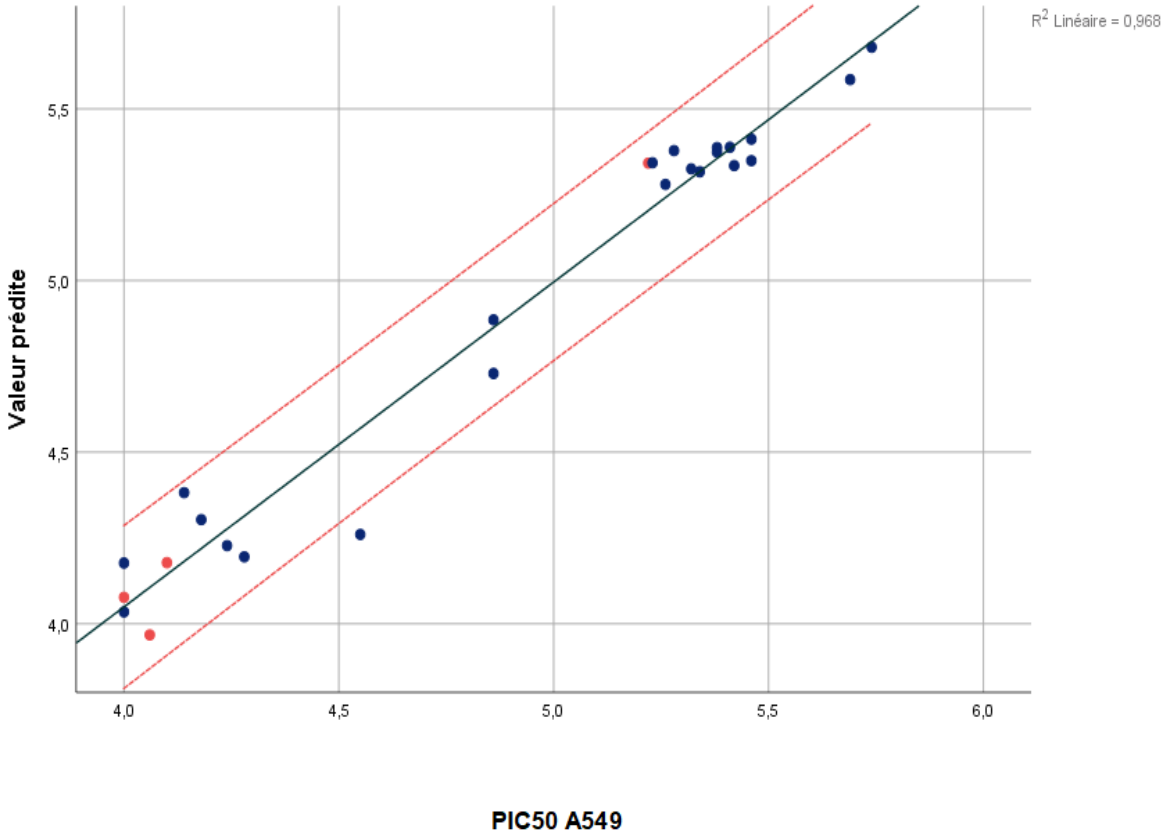


Figure 23 : Corrélations entre les valeurs d'activités observées d'A549 et prédite calculées à l'aide de modèles RNN (ensemble d'entraînement en bleu, ensemble de test.

Tableau 16 : Coefficients générés par RNN d'A549

Modèle r=0,991 r^2=0,968 EQM=0,10	
Données de validation **2.3.4.5.6.7.** **11.14.15.16.17.18.19.20.21.23.24.26.27.28.12.13.22**	EQM=0,0968 R^2=0,981
Données de test **1.10.25.9.8.**	EQM=0,217 R^2=0,852

Les résultats obtenus par RLM et RNN sont très suffisants pour conclure la qualité de la représentation.des modèles.

Une comparaison de la qualité de ces modèles montre que l'RNN a une capacité prédictive significativement meilleure. L'RNN a pu établir une relation plus satisfaisante entre les descripteurs et l'activité des composés analysés par rapport à la RLM ; mais cette méthode présente l'inconvénient d'être faiblement transparente, alors que la transparence de la méthode RLM donne des résultats plus interprétables et permet de bien expliquer les activités avec les descripteurs. Par conséquent, nous pouvons concevoir de nouveaux composés avec des valeurs améliorées par rapport aux composés étudiés à l'aide des modèles de RLM. En tenant compte de là au-dessus des résultats, nous avons ajouté des substitutions appropriées, puis nous sommes passés au calcul de leurs activités en utilisant les équations du modèle proposé par RLM. Par conséquent, les modèles proposés réduire le temps et le coût de la synthèse ainsi que la détermination de l'activité biologique dans notre cas A2780 et A549.

5.4.3 Domaine d'applicabilité (AD)

Le domaine d'applicabilité (AD) de ces modèles a été évalué par une analyse de l'effet de levier exprimée sous forme de diagramme de Williams (figures 24 et 25), dans lequel les résidus normalisés et les valeurs seuils de l'effet de levier (h^* = 0,52 et 0,85 pour la pIC50 (A2780) et la pIC_{50} (A549), respectivement) ont été tracés.

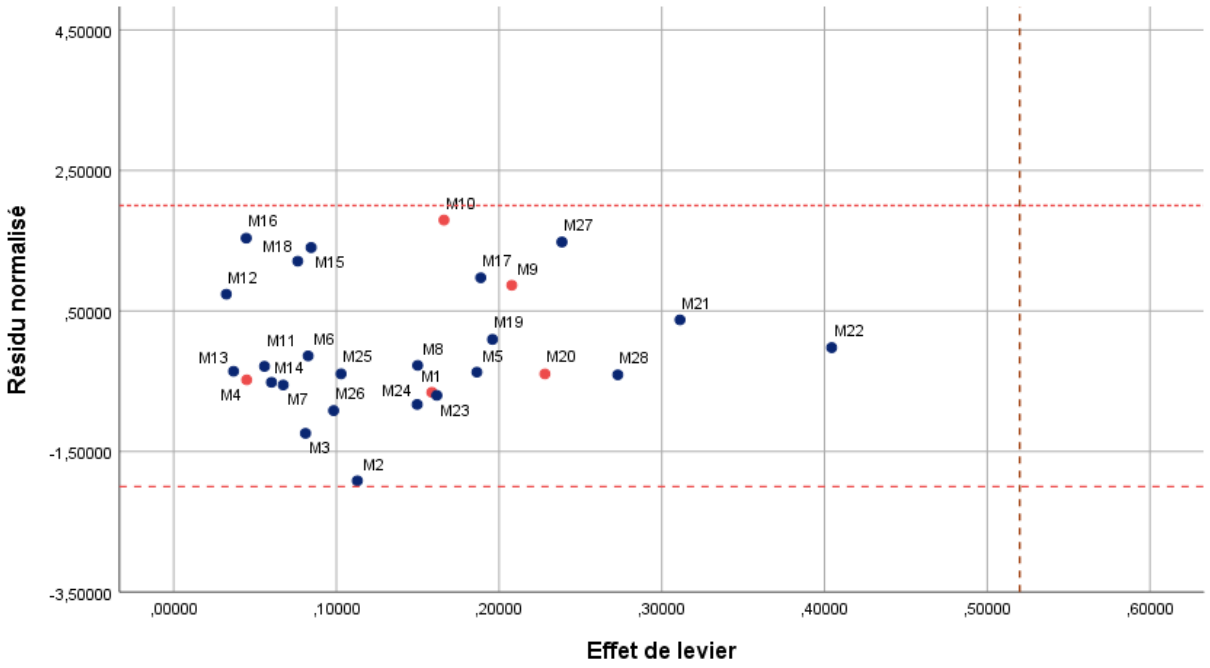


Figure 24 : Tracé du résidu normalisé en fonction de l'effet de levier pour le modèle RLM pour A2780

D'après la figure 24, il y a aucun composé est plus grand que l'effet de levier h* = 0,521 ou plus grand que la valeur maximale ou minimale de l'écart-type. Ces résultats montrent qu'on a aucune erreur expérimentale ou bien dans la structure de ces valeurs aberrantes.

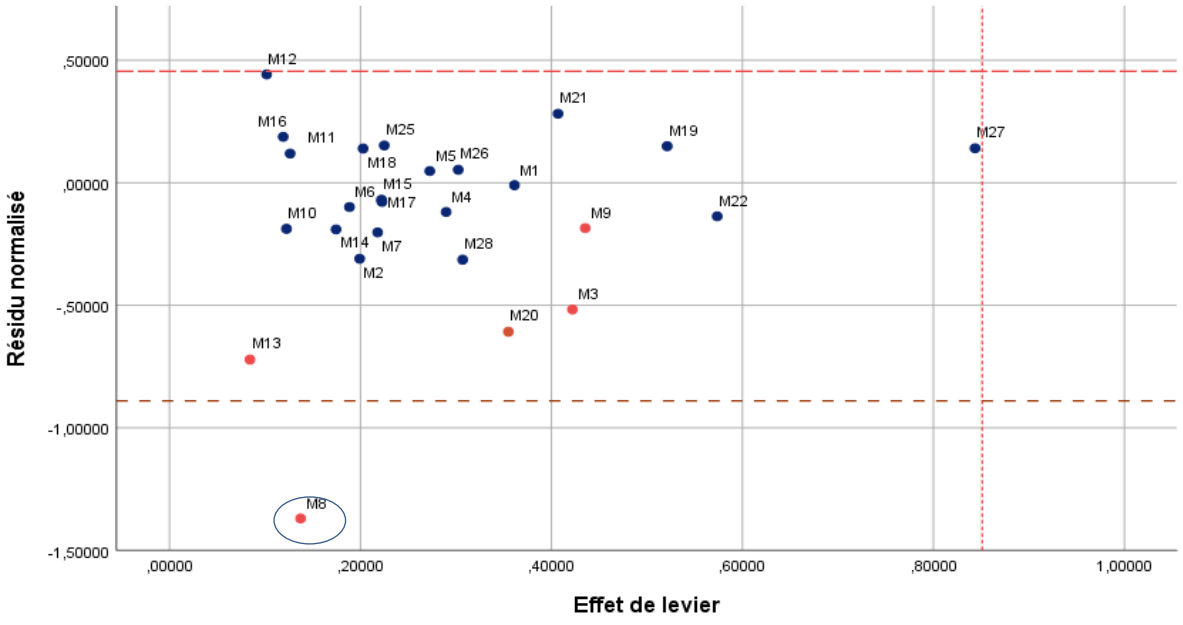


Figure 25 : Tracé du résidu normalisé en fonction de l'effet de levier pour le modèle RLM pour A549

Il est évident de remarque dans la figure 25, qu'il y a deux réponses aberrantes dans l'ensemble de formation et aucune réponse en dehors de l'ensemble de test ; et aucun composé n'a d'écart-type dans l'intervalle ±x. Seule substance chimique est identifiée comme étant des valeurs aberrantes pour le modèle RLM ; cette valeur aberrante est le composé M8 dans l'ensemble d'entraînement qui a un effet de levier inferieur à la valeur h* de 0,852 mais avec un résidu plus petit que la valeur minimale de l'écart-type. Ces résultats montrent qu'on a une erreur expérimentale ou bien dans la structure de ces valeurs aberrantes.

6 Nouveaux composés ayant des valeurs d'activité anticancéreuse plus élevées

Selon les discussions ci-dessus, les modèles RLM pourraient être appliqués à d'autres dérivées de 1,2-benzodiazole selon le tableau 3 et pourraient ajouter des connaissances supplémentaires dans l'amélioration de nouvelles méthodes de recherche sur les médicaments anti cancer. Si nous développons un nouveau composé avec de meilleures valeurs que les molécules existantes, cela pourrait donner lieu au développement de composés plus actifs que ceux actuellement utilisés. De cette façon, nous avons procédé à une modification structurelle en partant des composés ayant les valeurs de pIC_{50} les plus élevées comme modèle (numéro 15 et numéro 19). Les structures des composés conçus et leurs valeurs de paramètres calculées par les mêmes méthodes ainsi que les valeurs de pIC_{50} théoriquement prévues par les modèles RLM sont énumérées dans les tableaux 17 et 18. D'après les activités prédites, il a été observé que les composés conçus ont des valeurs de pIC_{50} plus élevées que les composés existants dans le cas des 28 molécules étudiés (tableau 3). Nous suggérons tous les composés comme candidats qui seront synthétisés et évalués comme médicaments anti cancer.

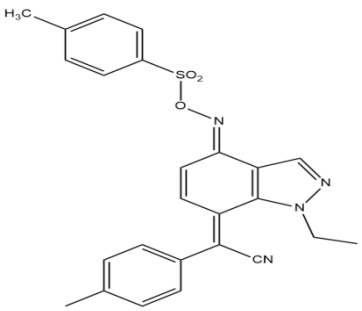
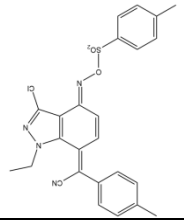
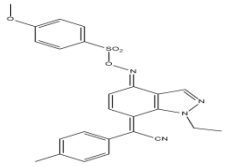
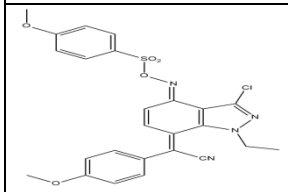
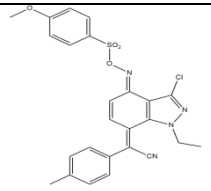
Tableau 17 : Valeurs théoriques de l'activité biologique A2780 par RLM

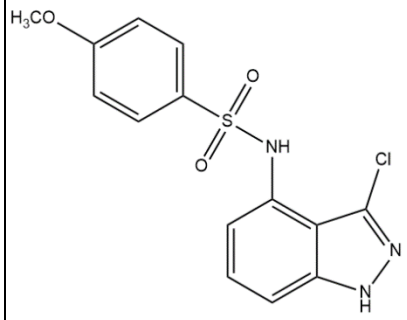
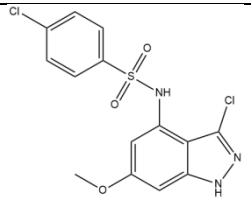
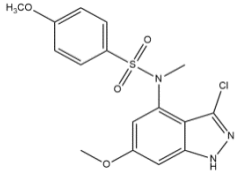
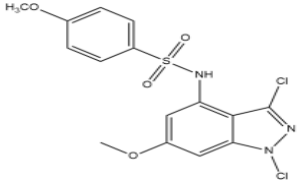
	pIC50 exp A2780 (µM)	pIC_{50} pred A2780 (µM)	résidu
M1	4	4,11	-0,11
M2	4	4,89	-0,89
M3	4	4,48	-0,48
M4	5,4	5,61	-0,29
M5	4	4,00	-0,01
M6	4,85	4,77	0,07
M7	4,68	4,79	-0,11
M8	4,36	4,31	0,04
M9	4,25	3,57	0,68
M10	5,32	4,33	1,07
M11	5,46	5,68	-0,23
M12	5,68	5,37	0,30
M13	5,3	5,50	-0,20
M14	5,39	5,72	-0,33
M15	6,19	5,60	0,58
M16	5,44	4,93	0,51
M17	5,85	5,67	0,17
M18	5,38	4,94	0,43
M19	6,07	6,12	-0,06
M20	5,91	6,19	-0,28
M21	5,38	5,07	0,31
M22	5,26	5,10	0,15
M23	4,12	4,43	-0,31
M24	4,01	4,35	-0,34
M25	4,16	4,17	-0,02
M26	4,04	4,32	-0,28
M27	4,38	3,67	0,70
M28	4,8	4,81	-0,01

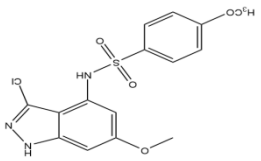
Tableau 18 : Valeurs théoriques de l'activité biologique A549 par RLM

	pIC_{50}exp A549(µM)	pIC_{50} pred A549 (µM)	résidu
M1	4	4,01	-0,01
M2	4	4,31	-0,31
M3	4	4,52	-0,52
M4	5,32	5,44	-0,12
M5	4,06	4,01	0,04
M6	4,86	4,95	-0,09
M7	4,86	5,06	-0,20
M8	4,14	5,51	-1,37
M9	4,10	4,28	-0,18
M10	5,22	5,41	-0,19
M11	5,34	5,22	0,12
M12	5,69	5,24	0,44
M13	5,28	5,95	-0,67
M14	5,41	5,60	-0,19
M15	5,46	5,27	0,19
M16	5,38	5,44	-0,07
M17	5,42	5,49	-0,07
M18	5,46	5,32	0,14
M19	5,74	5,59	0,15
M20	5,23	5,84	-0,61
M21	5,26	4,98	0,28
M22	5,38	5,52	-0,14
M23	-	4,53	-
M24	-	4,53	-
M25	4,55	4,40	0,15
M26	4,24	4,19	0,05
M27	4,28	4,14	0,14
M28	4,18	4,49	-0,31

Tableau 19 : Data des molécules créées avec d'activité très intéressante

		pIC_{50} pour l'activité biologique A2780 (µM)	**pIC_{50} pour l'activité biologique A549 (µM)**
		qC3=-0,058 qC6=-0,073 qC7=-0,190 DM=11,32 **pIC_{50}=6,315**	qC4=-0,069 qN2=-0,207 qC7=-0,202 **pIC_{50}=6,326** qC6=-0,069 V=1280,42 SAG=727.64
		qC3= 0,062 qC6=-0,069 qC7=-0,198 DM=9,367 **pIC_{50}=5,19**	qC4=-0,070 qN2=-0,207 qC7=-0,198 **pIC50=6,11** qC6=-0,069 V=1282.51 SAG=736,02
		qC6=-0,074 qC7=-0,195 DM=11,14 qC3= -0,062 **pIC_{50}=6,29**	qC4=-0,003 qN2=-0,176 qC7=-0,195 **pIC_{50}=6,50** qC6=-0,074 V=1363.62 SAG=785,93
		qC3= -0,062 qC6=-0,074 qC7=-0,192 DM=10,98 **pIC_{50}=6,27**	qC4=-0,003 qN2=-0,175 qC7=-0,191 **pIC_{50}=6,46** qC6=-0,074 V=1363.62 SAG=785,93

		pIC_{50} pour l'activité biologique A2780	**pIC_{50} pour l'activité biologique A459**
		qC3=- 0,087 qC6=-0,152 **pIC50= 6,5533** qC7=0,249 DM=7,5428	V=842,28 SAG=497,09 qC6=-0,152 **pIC50=7,73** qC7=0,249 Qc4=-0,065
		qC3=-0,099 qC6=-0,105 **pIC50=7,296** qC7=0,246 DM=8,725	V=956,97 SAG=558,54 qC4=0,014 qN2=-0,171 **pIC50=7,819** qC7=0,245 qC6=-0,105
		qC3=- 0,075 qC6=-0,149 **pIC50=7,355** qC7=0,248 DM=8,2413	V=1011,94 SAG=598.75 qC4=0,065 qN2=-0,149 **pIC50=6,94** qC7=0,248 qC6=0,248

	qC3=- 0,088 qC6=-0,153 **pIC50=7,4057** qC7=0,248 DM=9,2207	V=968.08 SAG=576.95 qC4=-0,064 qN2=-0,176 **pIC50=7,23** qC7=0,248 qC6=-0,153

7 Application de la règle ROF

Le médicament ressemble à un modèle prometteur pour quantifier l'équilibre entre les propriétés moléculaires d'un composé qui influencent sa pharmacodynamique et sa pharmacocinétique et qui, en fin de compte, optimise leur absorption, leur métabolisme et leur excrétion (ADME) dans le corps humain en tant que médicament. Les conditions empiriques pour satisfaire à la règle de *Lipinski* et démontrer une bonne biodisponibilité orale impliquent un équilibre entre la solubilité aqueuse d'un composé et sa capacité à se diffuser passivement à travers différentes barrières biologiques. Ces paramètres permettent de déterminer l'absorption orale ou la perméabilité membranaire lorsque la molécule évaluée suit la règle de *Lipinski* de cinq à partir du poids moléculaire (MM) ≤ 500 DA, d'un coefficient de partage octanol-eau log P ≤ 5, de donneurs de liaisons H, d'atomes d'azote ou d'oxygène un ou plusieurs atomes d'hydrogène (HBD) ≤ 5, d'accepteurs de liaisons H, d'atomes d'azote (HBA) ≤ 10 et la réfraction molaire doit être comprise entre 40 et 130.

Les molécules qui enfreignent plus d'une de ces règles peuvent présenter des problèmes de biodisponibilité. Cette règle établit donc certains paramètres structurels pertinents pour la prédiction théorique du profil de biodisponibilité orale, et est étroitement utilisée dans la création de nouveaux médicaments. Toutefois, les classes de composés qui sont des substrats pour des transporteurs biologiques tels que les antibiotiques, les antifongiques, les vitamines et les glycosides cardiaques font exception à la règle. Le nombre total de violations est le ROF-Score, entre 0 et 4 [14].

Les résultats du calcul (tableau 19) montrent que là plus part des composés étudiés suf M23 sont en accord avec les règles de *Lipinski* avec un ROF-Score =<1, ce qui suggère que ces composés n'auraient théoriquement pas de problèmes de biodisponibilité orale. Les molécules dont le score ROF est supérieur à 1 sont considérées comme marginales pour un développement ultérieur. Bien que, comme Lipinski et ses collègues l'ont souligné. Enfin, il est bien connu que de nombreux médicaments violent le ROF, mais ce n'est pas un problème grave car il n'a pas été conçu à l'origine comme un outil d'évaluation de la similarité des médicaments.

Tableau 20 : Confirmation de règle ROF sur les 28 molécules organique de l'indazole

Molécules	LogP*	MM	HBD (OH et NH)	HBA (N ou O)	MR*	Confirmation de ROF
M1	-0.89	403.45	1	8	112.04	0
M2	-1.29	375.44	1	7	107.61	0
M3	-0.98	317.36	0	6	91.74	0
M4	-2.7	377.48	1	6	110.3	0
M5	-0.24	343.4	1	6	100.9	0
M6	1.24	290.32	1	5	91.39	0
M7	0.86	310.74	1	5	91.83	0
M8	1.58	30435	1	4	96.14	0
M9	1.21	324.77	1	4	96.58	0
M10	0.44	472.52	0	8	125.77	0
M11	-0.71	476.51	0	9	137.86	1
M12	1.21	464.93	0	7	134.11	1
M13	0.06	480.93	0	8	136.2	1
M14	0.78	474.53	0	8	140.51	1
M15	-0.71	476.51	0	8	137.86	1
M16	1.93	458.53	0	7	138.42	1
M17	-0.71	476.51	0	9	137.86	1
M18	0.44	460.51	0	8	135.77	1
M19	-1.41	337.78	2	6	92.63	0
M20	-2.41	367.81	2	7	99	0
M21	-0.32	328.39	2	7	97.52	0
M22	-0.76	422.5	1	6	127.84	0
M23	-3.88	517.57	1	10	143.46	2
M24	-1.11	499.6	1	8	143.8	1
M25	-0.51	327.4	1	5	99.81	0
M26	-1.65	347.4	1	6	101.9	0
M27	-0.61	345.42	1	6	100.69	0
M28	0.38	361.85	1	5	104.59	0

*Le calcule de Logp et MR par utilisation de HyperChem

Conclusion

Pour la validation des modèles RQSA, et l'innovation des nouvelles molécules avec des activités biologiques très importantes, les deux méthodes statistiques de modélisation différentes, la régression linéaire multiple (RLM) et le réseau neuronal artificiel (RNN), ont été utilisées dans la construction des modèles QSAR pour les activités IC_{50} contre cancer ovarien et carcinome du poumon des dérivées d'indazole . Les bons résultats obtenus grâce aux validations internes et externes montrent que les modèles proposés dans cet article sont capables de prédire des activités avec une grande performance et que les descripteurs sélectionnés sont pertinents. Le domaine d'applicabilité (DA) du modèle RLM a été défini. Les modèles qui en résultent ont montré qu'on a établi une relation entre certains descripteurs et les activités chez les clients satisfaits. Les résultats de l'RNN ont une capacité prédictive nettement meilleure que celle du RLM, mais ce dernier leur donne des résultats d'interprétation très importants. Les résultats obtenus montrent que, pour augmenter l'activité A2780 contre tumeur d'ovarienne, on va augmenter qC6 et qC3 et diminuer DM et qC7. De plus, pour augmenter l'activité A549 contre carcinome du poumon, nous augmenterons SAG et qC6 et diminuerons MV, qC7 et qC4. Le résultat le plus important de cette recherche est que nous avons conçu et proposé de nouveaux composés ayant des valeurs d'activité plus élevées que les composés existants en ajoutant des substituts appropriés et en calculant leur activité à l'aide des équations de régression. En conséquence, les modèles proposés réduiront le temps et le coût de la synthèse ainsi que la détermination des activités biologique contre les deux cancers de poumon et d'ovarienne des dérivées de l'indazole.

References

[1] N. Abbassi *et al.*, "Synthesis, antiproliferative and apoptotic activities of N-(6 (4)-indazolyl)-benzenesulfonamide derivatives as potential anticancer agents," *European journal of medicinal chemistry*, vol. 57, pp. 240–249, 2012.

[2] D. D. Gaikwad *et al.*, "Synthesis of indazole motifs and their medicinal importance: An overview," *European journal of medicinal chemistry*, vol. 90, pp. 707–731, 2015.

[3] N. Abbassi *et al.*, "Synthesis and antitumor activity of some substituted indazole derivatives," *Archiv der pharmazie*, vol. 347, no. 6, pp. 423–431, 2014.

[4] L. Boiani *et al.*, "In vitro and in vivo antitrypanosomatid activity of 5-nitroindazoles," *European journal of medicinal chemistry*, vol. 44, no. 3, pp. 1034–1040, 2009.

[5] J. Rodríguez *et al.*, "Study of 5-nitroindazoles' anti-Trypanosoma cruzi mode of action: Electrochemical behaviour and ESR spectroscopic studies," *European Journal of Medicinal Chemistry*, vol. 44, no. 4, pp. 1545–1553, 2009, doi: 10.1016/j.ejmech.2008.07.018.

[6] B. Muro *et al.*, "New perspectives on the synthesis and antichagasic activity of 3-alkoxy-1-alkyl-5-nitroindazoles," *European journal of medicinal chemistry*, vol. 74, pp. 124–134, 2014.

[7] A. Schmidt, A. Beutler, and B. Snovydovych, "Recent advances in the chemistry of indazoles," *European Journal of Organic Chemistry*, vol. 2008, no. 24, pp. 4073–4095, 2008.

[8] M. Pordel, S. A. Beyramabadi, and A. Mohammadinejad, "Synthesis, DFT calculations and cyclic voltammetry analysis of new heterocyclic green dyes: 2-(5-Hydroxyimino-1-alkyl-4, 5-dihydro-1H-4-indazolyliden)-2-arylacetonitriles," *Dyes and Pigments*, vol. 102, pp. 46–52, 2014.

[9] J. Catalán, "On the solvatochromism, dimerization and tautomerism of indazole," *Arkivoc*, vol. 2, pp. 57–70, 2014.

[10] "Cancer de l'ovaire _ symptôme, dépistage, traitement." .

[11] É. Terrat, "Traitements des cancers pulmonaires," *Aide Soignante*, vol. 27, no. 146, pp. 15–17, 2013, doi: 10.1016/j.aidsoi.2013.02.007.

[12] Société canadienne du cancer, "Tumeurs cancéreuses du poumon - Société canadienne du cancer." .

[13] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced drug delivery reviews*, vol. 23, no. 1–3, pp. 3–25, 1997.

[14] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery permeability in drug discovery and development settings," no. August, 2014.

[15] T. I. Oprea, "Property distribution of drug-related chemical databases," *Journal of computer-aided molecular design*, vol. 14, no. 3, pp. 251–264, 2000.

[16] D. Biswas, S. Roy, and S. Sen, "A simple approach for indexing the oral druglikeness of a compound: discriminating druglike compounds from nondruglike ones," *Journal of chemical information and modeling*, vol. 46, no. 3, pp. 1394–1401, 2006.

[17] K. H. Bleicher, H.-J. Böhm, K. Müller, and A. I. Alanine, "Hit and lead generation: beyond high-throughput screening," *Nature reviews Drug discovery*, vol. 2, no. 5, pp. 369–378, 2003.

[18] J. Xu and J. Stevenson, "Drug-like index: a new approach to measure drug-like compounds and their diversity," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 5, pp. 1177–1187, 2000.
[19] A. Kouakou *et al.*, "SnCl2/RSH: a versatile catalytic system for the synthesis of 4-alkylsulfanyl-indazole derivatives," *Journal of Sulfur Chemistry*, vol. 36, no. 1, pp. 86–95, 2015.
[20] L. E. M. Indazole, A. K. Synthese, and E. P. E. T. A. Bi-, "Systemes Heterocycliques Comportant," 2017.
[21] H. Hyperchem, "Molecular modeling system. Hyper Cube," *Inc. and Auto Desk, Inc*, 2002.
[22] M. J. Frisch *et al.*, "Gaussian 09, Revision A. 02, 2009, Gaussian," *Inc., Wallingford CT*, 2009.
[23] P. Informatics, "ChemOffice." 2010.
[24] W. E. Wagner III, *Using IBM® SPSS® statistics for research methods and social science statistics*. Sage Publications, 2019.
[25] J. Wang, X.-Q. Xie, T. Hou, and X. Xu, "Fast approaches for molecular polarizability calculations," *The Journal of Physical Chemistry A*, vol. 111, no. 20, pp. 4443–4448, 2007.
[26] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, and R. K. Robins, "Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics," *Journal of Chemical Information and Computer Sciences*, vol. 29, no. 3, pp. 163–172, Aug. 1989, doi: 10.1021/ci00063a006.
[27] A. K. Ghose and G. M. Crippen, "Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions," *Journal of chemical information and computer sciences*, vol. 27, no. 1, pp. 21–35, 1987.
[28] N. Bodor, Z. Gabanyi, and C. K. Wong, "A new method for the estimation of partition coefficient," *Journal of the American Chemical Society*, vol. 111, no. 11, pp. 3783–3786, 1989.
[29] A. Gavezzotti, "The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state organic reactivity," *Journal of the American Chemical Society*, vol. 105, no. 16, pp. 5220–5225, 1983.
[30] R. G. Parr, R. A. Donnelly, M. Levy, and W. E. Palke, "Electronegativity: the density functional viewpoint," *The Journal of Chemical Physics*, vol. 68, no. 8, pp. 3801–3807, 1978.
[31] R. S. Mulliken, "A new electroaffinity scale; together with data on valence states and on valence ionization potentials and electron affinities," *The Journal of Chemical Physics*, vol. 2, no. 11, pp. 782–793, 1934.
[32] R. G. Parr and R. G. Pearson, "Absolute hardness: companion parameter to absolute electronegativity," *Journal of the American chemical society*, vol. 105, no. 26, pp. 7512–7516, 1983.
[33] W. Yang and R. G. Parr, "Hardness, softness, and the fukui function in the electronic theory of metals and catalysis.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 20, pp. 6723–6726, 1985, doi: 10.1073/pnas.82.20.6723.
[34] Z. Zhou and R. G. Parr, "Activation hardness: new index for describing the orientation

of electrophilic aromatic substitution," *Journal of the American Chemical Society*, vol. 112, no. 15, pp. 5720–5724, 1990.

[35] R. G. Pearson, "Absolute electronegativity and hardness: applications to organic chemistry," *The Journal of Organic Chemistry*, vol. 54, no. 6, pp. 1423–1430, 1989.

[36] K. J. Miller, "Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships: 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics," *J. Am. Chem. Soc*, vol. 112, no. 23, pp. 8533–8542, 1990.

Annexe : Aspects théoriques

1 Introduction

L'élaboration des modèles QSAR repose sur le calcul des structures (descripteurs), ce dernier est assuré en faisant appel aux outils de la modélisation moléculaire. Différentes approches sont envisageables dans le cadre des outils de modélisation moléculaire. Les méthodes quantiques en l'occurrence les méthodes ab initio, la théorie de la fonctionnelle de la densité et les méthodes semi-empiriques sont capables de calculer plusieurs propriétés des systèmes. C'est pour cette raison que ces approches ont été employées dans le cadre de cette étude. Dans cette partie, il s'agit d'expliciter les méthodes de chimie quantique, utilisées non seulement pour le calcul des structures moléculaires (descripteurs) ou (les paramètres physiques et chimiques) nécessaires à la mise en place de modèles, prédictives, mais aussi pour l'explication des mécanismes de toxicité des séries de composés étudiées.

2 Méthodes de la modélisation moléculaire

2.1 Mécanique quantique (MQ)

La chimie quantique applique le principe de la mécanique quantique aux systèmes moléculaires pour tenter de résoudre l'équation de *Schrödinger* [1]en effet, le comportement électronique et nucléaire des molécules, étant responsable des propriétés chimiques qui peuvent être décrites à partir de cette équation .En particulier, différentes méthodes de résolution ont alors été développées [2]nous distinguons trois approches :

- ✓ **Les méthodes ab initio** : elles visent à résolution de l'équation électronique d'un système d'électronique de *Schrödinger* pour déterminer la fonction d'onde approchée étudié ;

- ✓ **La théorie de la fonctionnelle de la densité (DFT)** : elle recherche de la densité électronique la plus proche possible en partant du principe que la densité électronique d'un système d'électron détermine toutes les propriétés de ce système ;

- ✓ **Les méthodes semi-empiriques** : elles sont une simplification des méthodes *ab initio* et sont paramétrées de façon à reproduire des résultats expérimentaux.Les méthodes semi-empiriques sont surtout utilisées pour des systèmes moléculaires[3] .

2.2 Equation de *Schrödinger*

Chimie quantique a fait l'objet de nombreux développements de logiciels permettant de réaliser des calculs plus ou moins compliqués sur des systèmes moléculaires, en se basant sur le formalisme de la mécanique quantique et certains niveaux d'approximations. Un calcul de chimie quantique permet d'obtenir des données importantes sur les propriétés du système étudié. Des propriétés d'ordre structural, énergétique, magnétique, électronique. Ce qui mène à bien définir, comprendre et même anticiper la réactivité du système considéré. Le comportement du système peut être décrit complètement par l'équation de *Schrödinger* [4].

$$\mathbf{H\Psi=E\Psi}$$

$$\textbf{Avec :}\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{T}} + \widehat{\boldsymbol{V}}$$

Ψ :Sont les fonctions propres de H

E : Sont les valeurs propres de H

L'hamiltonien complet non relativiste et indépendant du temps décrivant un système isolé constitué de N noyaux et n électrons est constitué de cinq termes : (1)

$$H = -\frac{\hbar^2}{2m_e}\sum_{i}^{n}\Delta_i \; -\frac{\hbar^2}{2M_K}\sum_{K}^{N}\Delta_K + \sum_{i>j}^{n}\frac{e^2}{r_{ij}} + \sum_{K>L}^{N}\frac{Z_K Z_L e^2}{r_{KL}} - \sum_{K=1}^{N}\sum_{i=1}^{n}\frac{Z_K e^2}{R_{Ki}}$$

Dans cette équation, les deux premiers termes de l'hamiltonien sont respectivement les opérateurs énergie cinétique des N électrons (indexés i) et des A noyaux atomiques (indexés I). Les trois autres termes sont des termes de corrélation. Ces derniers représentent les différents potentiels d'interaction électron-noyau (attraction colombienne), électron-électron (répulsion électronique) et noyau-noyau (répulsion nucléaire) respectivement.

Parmi les propriétés moléculaires calculées par la résolution de l'équation de Schrödinger se trouve : la géométrie moléculaire, les stabilités relatives entre systèmes moléculaires, les spectres de vibrations, les moments dipolaires et multipolaires.

En général, pour les systèmes moléculaires, les spectres électroniques et aussi les fonctions descriptives de la réactivité telles que les charges atomiques ne peuvent être résolues de manière exacte. En conséquence, un certain nombre d'approximations s'imposent pour remédier à cet obstacle

a) Approximation de Born-oppenheimer

L'approximation de Born-Oppenheimer [5]. Considère que l'on traite uniquement le cas des électrons et supposez les noyaux fixes dans l'espace. Cette approximation est justifiée du fait qu'elle s'appuie sur la grande différence de masse de l'électron 1880 fois plus faible que celle du proton (un électron voit le noyau immobile). Les noyaux sont considérés comme fixes dans l'espace (T_N =0) et le traitement se fera uniquement sur les électrons.

$$H=-\frac{\hbar^2}{2m_e}\sum_i^n \Delta_i -\sum_{K=1}^N \sum_{i=1}^n \frac{Z_K e^2}{R_{Ki}}+\sum_{i>j}^n \frac{e^2}{r_{ij}}$$

La résolution exacte de l'équation (1) n'est possible que pour l'atome d'hydrogène et les systèmes hydrogénoïdes. Pour les systèmes poly-électroniques, il est nécessaire de faire appel aux méthodes d'approximation pour résoudre l'équation de Schrödinger d'une manière approchée.

Les propriétés moléculaires qui peuvent être calculées par la résolution de l'équation de Schrödinger sont multiples. On peut citer entre autres :

- Structures et énergies moléculaires
- Energies et structures des états de transition
- Fréquences de vibration
- Spectres IR et Raman
- Propriétés thermochimiques
- Energies de liaison
- Chemins réactionnels
- Orbitales moléculaires
- Charges atomiques

- Moments multipolaires
- Déplacements chimiques RMN et susceptibilités magnétiques
- Affinités électroniques et potentiels d'ionisation
- Polarisabilités et hyperpolarisabilités
- Potentiels électrostatiques et densités électroniques
- etc.

b) Approximation orbitalaire

La fonction d'onde électronique Ψe (que nous désignerons dorénavant uniquement par la lettre Ψ) est une fonction des coordonnées de tous les électrons du système. Si 2n est le nombre d'électrons (2n est choisi ici par commodité), Ψ est une fonction à (2n) ×3 variables que l'on note communément Ψ (1,2,... 2n). L'approximation orbitale introduite par Hartree en 1928 [6] consiste à découpler les 2n électrons en développant la fonction Ψ (1,2,...,2n) en un produit de 2n fonctions mono électroniques de sorte que :

$$\boldsymbol{\Psi}(\mathbf{1},\mathbf{2},\ldots\mathbf{2n}) = \prod_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{2n}} \boldsymbol{\phi_i(i)}$$

Où l'indice i désigne l'orbital i cette situation correspond physiquement à un modèle de particules indépendantes dans lequel chaque électron se déplace dans un champ moyen créé par les noyaux et la densité électronique moyenne des autres électrons. Cela signifie que chaque électron ressent les autres en moyenne, ce qui constitue naturellement une approximation. La fonction d'onde ainsi obtenue ne satisfait plus le principe de Pauli. Ce problème est alors résolu en écrivant la fonction d'onde comme un déterminant de Slater construit sur la base de n spin-orbitales (où n/2 orbitales spatiales sont combinées à deux fonctions de spin possibles). Le problème réside alors dans l'obtention des meilleures spin-orbitales pour obtenir la fonction d'onde du système à n électrons. La résolution exacte d'un tel Hamiltonien est hors de portée de toutes les méthodes numériques. Il faut donc ajouter des approximations supplémentaires à celle de Born- Oppenheimer et les méthodes se scindent en deux catégories Hartree- fock ou la théorie de fonctionnelle de la densité (DFT) [7].

c) Approximation de LCAO – MO

La méthode L.C.A.O. (Linear Combinaison of Atomic Orbitals) montre que chaque orbitale moléculaire peut se développer en une combinaison linéaire d'orbital atomique (A.O.). Pour cela, on choisit une base d'A.O.(χ_μ) de dimension M, les M orbitales moléculaires doublement occupés sont de la forme :

$$\Psi_i = \sum_{\mu=1}^{M} C_{\mu i}\, \chi_\mu$$

i=1,2,3……M,

Le calcul de O.M. se ramène donc à la détermination des coefficients $C_{\mu i}$ Le déterminant de Slater, solution de l'équation a N électrons, est construit à partir des N/2 orbitales de plus basses énergies.[8] [9]

2.3 Méthode quantique : ab initio

Ab initio les méthodes ab-initio sont caractérisées par l'introduction d'une base arbitraire pour étendre les orbitales moléculaires et alors le calcul explicite toutes les intégrales exigées qui impliquent cette base. La théorie de ces méthodes est basée sur les considérations suivantes :

- les interactions électroniques sont traitées de manière explicite et quantique ;

- les interactions des noyaux sont calculées de manière classique (énergie d'interaction coulombienne).

a) Méthode de Hartree-Fock

La méthode de Hartree-Fock [6][10] est une approximation de champ moyen à particules indépendantes appelée principe du champ auto-cohérent. Chaque électron est représenté par une spin-orbitale. Les électrons étant des fermions, leurs fonctions d'onde doivent respecter le principe d'Antisymétrie. Tenant compte que les électrons sont indiscernables de la fonction d'onde qui s'écrit sous la forme d'un déterminant de Slater [11] .

$$\boldsymbol{\Psi}\,(\mathbf{1},,\ldots\mathbf{n}) = \frac{\mathbf{1}}{\sqrt{\boldsymbol{n}!}}\begin{bmatrix} \emptyset_{\mathbf{1}}\boldsymbol{a}(\mathbf{1})\emptyset_{\mathbf{1}}\boldsymbol{\beta}(\mathbf{1}) & \ldots & \emptyset_{\boldsymbol{n}}\boldsymbol{a}(\mathbf{1})\emptyset_{\boldsymbol{n}}\boldsymbol{\beta}(\mathbf{1}) \\ & . & \\ & . & \\ & . & \\ \emptyset_{\mathbf{1}}\boldsymbol{a}(\boldsymbol{n})\emptyset_{\mathbf{1}}\boldsymbol{\beta}(\boldsymbol{n}) & \ldots & \emptyset_{\boldsymbol{n}}\boldsymbol{a}(\boldsymbol{n})\emptyset_{\boldsymbol{n}}\boldsymbol{\beta}(\boldsymbol{n}). \end{bmatrix}$$

$$\frac{\mathbf{1}}{\sqrt{\boldsymbol{n}!}}\ \boldsymbol{est\ le\ facteur\ de\ normalisation}$$

Par construction, le déterminant de Slater respecte la propriété d'antisymétrie de la fonction d'onde à condition que tous les spin-orbitales occupées soient différentes. Dans le cas contraire, le déterminant s'annule, il s'en suit donc que dans un déterminant, deux spin-orbitales ne peuvent être égales et doivent donc différer par au moins un nombre quantique, c'est le principe de Pauli [12]. Les équations HF ne sont pas toujours faciles à résoudre. Aussi exprime-t-on les orbitales moléculaires OM comme des combinaisons linéaires de jeux prédéfinis de fonctions mono électroniques (χμ), d'où le qualificatif de cette approximation : LCAO pour « Linear Combinaison of Atomic Orbitals ». À partir de l'équation des orbitales moléculaires :

$$\boldsymbol{\Psi_i} = \sum_{\mu=\mathbf{1}}^{\boldsymbol{k}} \boldsymbol{C}_{\mu\boldsymbol{i}}\,\boldsymbol{\chi}_{\mu}\boldsymbol{i} = \mathbf{1},\mathbf{2},\mathbf{3}\ldots\ldots\boldsymbol{K}$$

Il s'agira de déterminer les coefficients Cμi. Le déterminant de Slater, solution de l'équation à N électrons, est construit à partir des N/2 orbitales de plus basses énergies. La méthode HF possède deux variantes : l'approche Hartree-Fock restreinte ou RHF de l'anglais *Restricted Hartee-Fock* et l'approche Hartree-Fock non restreinte ou UHF de l'anglais *Unrestricted Hartee-Fock* [13] [14].

Le premier formalisme qui concerne les systèmes à couches dites "fermées" contraint les spin-orbitales appariées de spins différents à avoir la même partie spatiale. Le second formalisme concerne les systèmes à couches dites « ouvertes » et consiste à traiter indépendamment les orbitales de spin α et β. Cette approche est plus coûteuse en temps de calcul, car elle double le nombre d'intégrales à calculer, les orbitales n'étant plus doublement occupées. Il faut également remarquer que dans le cadre de la méthode HF, les électrons sont considérés comme indépendants les uns des autres et se déplace chacun dans un potentiel

moyen créé par l'ensemble des noyaux et des autres électrons. Il n'y a donc pas d'interaction instantanée électron-électron d'où le développement de certaines méthodes pour tenter de remédier à ce problème de manque de corrélation. La résolution de l'équation de Hartree-Fock se fait par une procédure itérative dite : procédure du champ auto-cohérent ou SCF «Self Consistant Field » [15]. La minimisation de l'énergie est effectuée par la méthode SCF, tout en respectant la contrainte d'ortho-normalité des orbitales.

3 Théorie de la fonctionnelle de la densité

3.1 Aperçu historique

La théorie de la fonctionnelle de la densité a pour objet de décrire un système en considérant la densité ρ(r) comme variable de base. Ainsi le problème à n électrons est étudié dans l'espace de ρ(r) qui est de dimension 3 au lieu de l'espace de dimension 3n de la fonction d'onde Ψ. Les premiers à avoir exprimé l'énergie en fonction de la densité furent L. H. Thomas et E. Fermi en 1927. Dans leur modèle, les interactions électroniques sont traitées classiquement et l'énergie cinétique est calculée en supposant la densité électronique homogène. Ce modèle a été amélioré par P. A. Dirac en 1930 avec un terme d'échange. Un peu plus tard, en 1951 J. C. Slater [16] proposa un modèle basé sur l'étude d'un gaz uniforme améliorée avec un potentiel local. Cette méthode, appelée Hartree-Fock-Slater ou Xα, fut essentiellement utilisée en physique du solide dans les années 70. Les premières applications de la DFT pour la recherche sur la structure électronique moléculaire a commencé à apparaître dans les années 90 avec le développement des fonctionnels d'échange et de corrélation [16], les plus précises et les plus rapides pour le calcul des propriétés électroniques de grands systèmes moléculaires. Enfin, il est à signaler qu'un prix Nobel a été attribué à Kohn et à Pople [16][17] en 1998 dans le cadre de développement de cette méthode.

3.2 Definition

Toutes les méthodes à fonction d'onde explicite, vues plus haut, ne sont applicables qu'aux petits systèmes avec une dizaine d'atomes, et sont très couteuses. Pour contourner cette difficulté et étudier des systèmes de plus grande taille, on peut faire appel aux méthodes de la fonctionnelle de la densité DFT, qui introduisent la corrélation par l'intermédiaire d'une fonctionnelle de la densité électronique. La démarche de la DFT semble

particulièrement avantageuse car la densité électronique ρ (x, y, z) ne dépend que de 3 variables (4, avec spin) et peut être considérée comme une observable. Par contre, le nombre de variables d'espace entrant dans la fonction d'onde, qui n'est pas une observable, est de 3N (N étant le nombre d'électrons du système). De plus, la précision des résultats obtenus ainsi que les performances des calculs de toutes les méthodes DFT permettent d'avoir un outil très efficace pour le calcul des propriétés moléculaires.

L'idée d'exprimer l'énergie totale d'un système multi-électronique par une fonctionnelle de la densité électronique totale a été introduite par Thomas et Fermi. Mais, ce n'est qu'en 1964 que Hohenberg et Kohn [18] proposèrent la formulation exacte de ce modèle, appelé théorie de la fonctionnelle de la densité. Elle est fondée sur deux théorèmes.

Le but de ces méthodes est de produire des fonctionnelles mettant en relation la densité avec l'énergie de l'état électronique fondamental.

3.3 Quelques définitions essentielles

a) Densité électronique

Pour un système à N électrons se trouvant dans un état représenté par la fonction d'onde Ψ, la probabilité de trouver n'importe lequel de ces N électrons dans l'élément de volume $d\overrightarrow{r_1}$ quelque soit son spin et quelques soient les positions et les spins des N-1 autres électrons est donné par :

$$\boldsymbol{\rho(\vec{r}) = N \ldots \int |\varphi(\overrightarrow{r_1}, \overrightarrow{r_2}, \overrightarrow{r_3}, \overrightarrow{r_4} \ldots\ldots, \overrightarrow{r_n}|}$$

où $\boldsymbol{\rho(\vec{r})}$ est la densité de probabilité ou en terme plus courant la « densité électronique ». Dans cette équation, l'intégrale multiple représente la probabilité qu'un seul électron se trouve dans

$\boldsymbol{d\overrightarrow{r_1}}$. Mais comme les électrons sont indiscernables, la probabilité de trouver n'importe lequel des électrons à cette position est tout simplement N fois la probabilité d'un seul électron.

La densité électronique possède deux propriétés fondamentales : elle s'annule à l'infini et son intégral donne le nombre total des électrons :

$$\boldsymbol{\rho(\vec{r} \to \infty) = 0}$$

$$\boldsymbol{\int \rho(\vec{r}) d\vec{r_1} = N}$$

Reste à mentionner que contrairement à la fonction d'onde, la densité électronique est une grandeur observable et peut être mesurée expérimentalement par

b) Théorie

Une fois les différentes quantités définies, il est maintenant nécessaire de poser les fondements de la DFT. Ils ont été exprimés pour la première fois en 1964 par Hohenberg et Kohn [18] et se déclinent en deux théorèmes.

- Premier théorème

Le premier théorème de Hohenberg et Kohn[18] montre très simplement que la densité $\rho(r)$estla seule fonction nécessaire pour obtenir toutes les propriétés électroniques d'un système dans son état fondamental. La densité électronique fixe également le nombre d'électrons n du système via la condition :

$$\boldsymbol{n = \int \rho(r) d(r)}$$

Où ρ (r) est la densité électronique et r les coordonnées des électrons. Elle est définie par :

$$\boldsymbol{\rho(r) = \int |\Psi(r)|^2}$$

Avec Ψ la fonction d'onde électronique solution de l'équation de Schrödinger électronique (en s'affranchissant du terme de répulsion entre les noyaux V_{NN}) :

$$\boldsymbol{H\Psi = [T + V_{Ne} + V_{ee}]\Psi = E\Psi}$$

Dans la pratique, le terme d'attraction électron-noyau VNe est souvent remplacé par un potentiel extérieur Vext regroupant, en plus de VNe, les différentes perturbations externes (champs électriques, etc…).

La densité électronique totale peut être donnée en fonction des densités de spin ρ_α et ρ_β :

$$\boldsymbol{\rho(r) = \rho_\alpha(r) + \rho_\beta(r)}$$

L'énergie électronique est donc une fonctionnelle de la densité et sera notée E[α] où ρ = (ρα, ρ_β). Les calculs effectués seront donc similaires pour les systèmes à couches ouvertes et les systèmes à couches fermées. Cette énergie, exprimée en termes de fonctionnelle de la densité, se décompose en trois parties :

$$\mathbf{E[\rho] = T[\rho] + E_{Ne}[\rho] + E_{ee}[\rho]}$$

où
T[ρ] est l'énergie cinétique.
$E_{Ne}[\rho]$ Ne est l'énergie provenant de l'interaction électron-noyau.
E_{ee} ee est celle provenant de l'interaction électron-électron.

- Second théorème

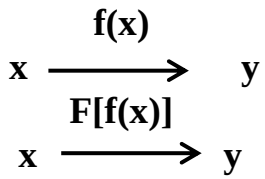
Le second théorème de Hohenberg et Kohn [18] démontre que seule la densité électronique vraie permet de calculer l'énergie exacte. N'importe quelle autre densité différente de la densité exacte conduit à une énergie plus grande. L'énergie d'interaction électron-électron s'écrit comme la somme d'un terme coulombien J et d'un terme d'échange K. L'expression de l'énergie devient :

$$\mathbf{E[\rho] = T[\rho] + E_{Ne}[\rho] + J[\rho] + K[\rho]}$$

Les expressions analytiques de E_{Ne} et J sont connues, en revanche les termes cinétiques T et d'échange K ne peuvent être exprimés analytiquement. Diffraction de rayon X.

c) Fonctionnelle

Le concept mathématique principal de la théorie de la fonctionnelle de la densité est, comme le nom l'indique bien, la notion de la fonctionnelle. Ainsi contrairement au concept familier de fonction qui associe un nombre à un autre nombre, la fonctionnelle associe une fonction à un nombre. En d'autres termes, on peut dire qu'une fonctionnelle est une fonction dont l'argument est lui-même une fonction. Par convention on note entre crochets l'argument de la fonctionnelle. Le schéma suivant illustre plus clairement ce que l'on vient de dire :



Où f(x) est la fonction qui associe x à y et F[f(x)] est la fonctionnelle qui associe la fonction f(x) au scalaire y.

Dans la théorie de la fonctionnelle de la densité, la fonction principale est la densité électronique. L'idée principale étant d'exprimer l'énergie du système en fonction de la seule densité électronique rendant de cette façon l'énergie une fonctionnelle de la densité électronique : $[\rho(\vec{r_1})]$

d) Différents types de fonctionnelles

Selon l'approximation utilisée, Il existe différentes classes de fonctionnelles énergies d'échange et corrélation :

- Fonction d'échanges et de corrélation

La partie la plus critique d'un calcul utilisant la fonctionnelle de la densité et l'énergie de Kohn-Sham est de déterminé l'énergie d'échange et de corrélation. Celle-ci est la correction nécessaire aux approximations et effets ignorés dans le calcul de l'énergie.

Une première correction à tenir compte est l'erreur sur le potentiel de Hartree. Le facteur 1/ 2 tient compte que l'attraction est calculée deux fois entre chaque paire d'électrons, mais l'intégrale implique qu'un électron a un effet d'attraction sur lui-même.

D'autres corrections doivent aussi être considérées dans le calcul de la fonction d'échange et de corrélation. Il n'existe pas de formulation exacte qui permet de corriger tous les effets.

- Approximation locale (Local Density Approximation(LDA)

Dans cette approche, on suppose la densité comme étant localement constante. On

Définit alors l'énergie d'échange-corrélation par :

$$\boldsymbol{E_{xc}^{LDA}[\rho] = \int e_{xc}(\rho)dr}$$

Ou exc(ρ) représente l'énergie d'échange-corrélation, énergie qui peut etre partitionnée en une partie d'échange, ex(ρ), et une partie de corrélation, ec(ρ) :

$$\mathbf{Exc(\rho) = e_x(\rho) + e_c(\rho)}$$

On utilise alors l'énergie d'échange donnée par Dirac [19] pour approximer $e_x(\rho)$:

$$\mathbf{K_D[\rho(r)] = C_x \int \rho(r)^{4/3}\, dr}$$

$$\mathbf{Avec \quad C_x = \frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}}}$$

Quant aux fonctionnelles de corrélation, la plus utilisée est celle développée par Vosko,Wilk et Nusair [20], connue sous le nom de VWN et ajustée de facon analytique sur la base de simulation Monte Carlo.

Bien que plus performante que la méthode HF, l'utilisation de cette méthode conduit bien souvent à de mauvaises descriptions géométriques du système, c'est à dire une mauvaise description du SEP. Les barrières de réactions ont de mieux souvent tendance à etre surestimées par l'utilisation de telles méthodes.

C'est pourquoi depuis 1985, un grand nombre de travaux se sont attachés à une nouvelle description des fonctionnelles d'échanges et corrélations en prenant en compte l'inhomogénéité de la densité électronique, c'est à dire en considérant la densité ainsi que son gradient : ceci constitue l'approximation du gradient généralisé.

- Approximation du gradient généralisée (GGA)

Les fonctionnelles de ce type sont conseillées pour l'étude des systèmes moléculaires généralement caractérisés par une densité électronique fortement inhomogène selon les 3 dimensions de l'espace. Dans ce cas, les énergies d'échange et de corrélation apparaissent comme des fonctionnelles de la densité ρ (r) mais aussi du gradient de la densité $\nabla\rho(r)$(équation (1.72)).

$$E_{XC}^{GGA} = \int_{\varepsilon XC}^{GGA} (\rho, \nabla\rho) dr$$

où E_{XC}^{GGA} est la densité d'énergie d'échange-corrélation. Le développement des fonctionnelles GGA s'est organisé autour de deux idées motrices :

la première due à Becke [21][22]repose sur l'introduction de formalismes empiriques dans lesquels certains paramètres sont ajustés sur la base d'un ensemble de valeurs expérimentales déterminées pour des molécules modèles

la seconde défendue par Perdew [23][24] est de s'assurer du respect de principes et résultats fondamentaux issus de la mécanique quantique (limites correctes pour les densités élevées ou faibles, recouvrer le comportement LSDA quand la densité varie lentement…).

En général, les méthodes GGA [25] représentent une amélioration significative par rapport aux méthodes LSDA: elles ont tendance à mieux décrire les énergies totales, les énergies d'atomisation et les barrières énergétiques de réaction. Toutefois, la précision des méthodes GGA n'est pas toujours suffisante pour obtenir une description correcte de nombreuses propriétés des molécules. Par exemple, bien que donnant usuellement des résultats fiables pour la description des liaisons hydrogène, covalentes, ioniques, métalliques, elles échouent généralement lors de la description des interactions de van der Waals [26][27]. En outre, les différences observées lors de l'utilisation de différentes fonctionnelles GGA sont souvent aussi grandes que celles observées entre une fonctionnelle GGA et une fonctionnelle LSDA.

- Approche hybride(H-GGA)

Les fonctionnels hybrides sont des méthodes qui combinent, à l'énergie d'échange corrélation issue d'une méthode GGA conventionnelle, un certain pourcentage d'échange (parfois appelé exact) de Hartree-Fock. Un certain degré d'empirisme est utilisé pour optimiser le facteur de pondération pour chacune des composantes et dans ce cas les fonctionnelles sont mixtes. Une façon de procéder est d'ajuster ces coefficients à partir de valeurs d'énergies d'atomisation, de potentiels d'ionisation, d'affinités protoniques, et

d'autres paramètres expérimentaux d'un ensemble représentatif de molécules [28]. En général les méthodes hybrides représentent une amélioration significative par rapport aux méthodes antérieures pour l'étude de nombreuses propriétés moléculaires. On peut citer comme exemple de fonctionnelle H-GGA (hybrid-GGA functional), La fonctionnelle d'échange corrélation hybride B3LYP.

- Fonctionnelle B3LYP

La fonctionnelle B3LYP signifie Becke- 3 paramètres – Lee Yang Parr. Elle a été introduite par l'équipe de Becke en 1993 [29][30]. Elle appartient à la famille des fonctionnelles GGA hybrides (H-GGA). La particularité de cette fonctionnelle est de présenter une combinaison linéaire entre des fonctionnelles d'échange-corrélation GGA et de l'échange Hartree-Fock. Cette fonctionnelle utilise 20 % d'échange Hartree-Fock et 80% de corrélation. L'énergie d'échange-corrélation de la fonctionnelle B3LYP s'écrit donc sous la forme

$$\boldsymbol{E_{xc}^{B3LYP} = (1 - a_0 - a_x)E_x^{LSDA} + a_0E_x^{EXACT} + a_xE_x^{B88} + +(1 - a_0)E_c^{VWN} + +a_cE_c^{EXACT}}$$

- Les indices X et C désignent l'énergie d'échange et de corrélation respectivement.
- LDA et GGA désignent les termes énergétiques calculées par la DFT.
- HF désigne la contribution calculée par la théorie HF.

avec les valeurs optimisées suivantes pour les coefficients : a_0 =0,2 ; a_x=0,72 et a_c=0,81 [28][31].

Cette fonctionnelle s'est révélée jusqu'à présent relativement efficace pour traiter la « plupart » des systèmes moléculaires, cette robustesse expliquant la très grande popularité de la méthode (plus de 80 % des calculs de DFT de par le monde utilisent B3LYP). Néanmoins, cette méthode présente quelques limitations telle que :

- la sous-estimation des hauteurs de barrières énergétiques [32].
- L'absence de prise en compte des interactions non-covalentes :
- Fonctionnelle B3LYP est incapable de décrire des liaisons de van der Waals pour des composés liés par des interactions de portée moyenne.

Malgré ces problèmes, cette fonctionnelle reste la base de calculs pour la plupart des composés chimiques et l'outil le plus utilisé en modélisation moléculaire.

3.4 Méthode DFT dépendante du temps (TD-DFT)

A l'origine, la DFT a été développée dans le cadre de la théorie quantique non-relativiste (équation de Schrödinger indépendante du temps) et dans le cadre de l'approximation de BornOppenheimer. La théorie fut par la suite étendue au domaine de la mécanique quantique dépendante du temps (TD-DFT pour time- Dependent Densité Fonctional Theory).

La TD-DFT [33] [34] [35]est une théorie de la mécanique quantique appliquée en physique et en chimie pour étudier les propriétés et la dynamique des systèmes à plusieurs corps dans la présence des potentiels dépendant du temps, tel que les champs électriques ou magnétiques. Les calculs basés sur la méthode TD-DFT permettent d'avoir accès aux spectres UV–Visible et à différents paramètres optiques (longueur d'onde maximale λ_{max} d'adsorption, valeur approximative du gap entre les orbitales HOMO et LUMO (ΔE_{H-L}).

3.5 Limites de la méthode DFT

Depuis le début des années 90, le nombre de publications scientifiques dans différents domaines de la chimie et de la physique utilisant la DFT a connu une ascension véritable.

Donnant des résultats comparables à ceux obtenus au moyen des méthodes HF et post-HF à un coût en temps de calcul nettement moindre (dans un rapport de 1 à 5 en moyenne), les méthodes DFT sont de plus en plus utilisées. Cependant, la DFT souffre encore d'un certain nombre de faiblesses. Étant une méthode mono-déterminantale, elle ne permet pas la description correcte des systèmes multiconfigurationnels des états excités. En raison de sa limitation par l'approximation de la fonctionnelle d'échange-corrélation, l'énergie du système peut varier dans de très larges limites selon la fonctionnelle utilisée. De plus, il n'existe pas de critère pour choisir une fonctionnelle plutôt qu'une autre ; comme il est difficile de trouver des critères permettant l'amélioration d'une fonctionnelle donnée. Néanmoins, les travaux se poursuivent pour corriger ces défauts. Les développements récents

utilisent un formalisme dépendent du temps (TD-DFT de l'anglais Time Dependant Density Functional Theory) qui permet de décrire les états excités.

3.6 Calcule de DFT

Tout calcul de structure électronique vise avant tout la détermination des orbitales moléculaires, de leur énergie et de leur schéma d'occupation dans l'état fondamental. Des propriétés moléculaires découlant de cette structure orbitalaire peuvent ensuite être calculées. Tous les calculs d'optimisation réalisés ont été exécutés sur un PC utilisant GUAUSSIAN 03 [23] qui est un programme de chimie computationnelle publié en 1970 par John Pople et son groupe de recherche à l'Université Carnegie-Mellon. Ce programme fait partie d'une série de programmes gaussian de la compagnie Gaussian, Inc. qui a été installée dans les années 80 pour distribuer le programme. C'est un logiciel utilisé par des chimistes, des ingénieurs chimistes, des biochimistes, des physiciens et des autres, permettant de faire des calculs de modélisation moléculaire basés sur les principes de la chimie quantique. C'est un volume de programmes de calculs d'orbitales moléculaires. Ces calculs utilisent une vaste gamme de méthodes, allant des méthodes s'approchant d'une résolution numérique exacte de l'équation de Schrödinger, aux méthodes semi empiriques, dans lesquelles certaines intégrales sont remplacées par des valeurs empiriques ajustées pour reproduire certaines observations expérimentales.

3.7 Méthodes de calcules accessibles sur Gaussian

A partir de la base des lois de la mécanique quantique, Gaussian 03 peut être utilisé pour modéliser un grand nombre de propriétés d'atomes et de molécules et aussi des réactions chimiques en phase gazeuse et en solution à l'état solide:

- ✓ Energies en utilisant un grand nombre de méthodes, incluant Hartree-Fock, Théorie Fonctionnelle de la Densité.
- ✓ Géométries d'équilibres ou d'états de transition (optimisée en coordonnées internes redondantes pour la vitesse). Spectres de vibration, incluant IR, intensités Raman non résonnantes et pré résonance, couplage de vibration-rotation.
- ✓ Propriétés magnétiques, incluant déplacements chimiques et constantes de couplage RMN.

✓ Spectres de molécules chirales : rotations optiques, VCD et ROA. un examen de la réactivité et des spectres de grosses molécules (plus particulièrement avec la méthode ONIOM).

4 GaussView

Pour rendre l'utilisation de Gaussian 09 plus intuitive et visualiser les divers résultats une interface graphique complète a été utilisée : Gauss View 05[36]. Avec Gauss View, nous pouvons construire les systèmes moléculaires rapidement et efficacement, en utilisant la fonction de construction des molécules et lancer nos calculs sur Gaussian 09[37]. Cette interface comprend un excellent constructeur de molécules, permettant :

- Construction rapide, même pour des grosses molécules.
- Construction de molécules par atomes, cycle, groupe et acide aminé ·
- Importation de molécules d'autres sources en les ouvrant tout simplement..
- Ajouter automatiquement des hydrogènes aux structures provenant de fichiers PDB, avec une excellente fiabilité.
- Rotation en 3 dimensions même pour de très grosses molécules. Pour la visualisation des résultats des calculs d'optimisation effectués, l'introduction des coordonnées internes ou des coordonnées cartésiennes a été nécessaire.

4.1 Nomenclature de bases usuelles

Outre la base minimale STO-3G, un jeu de bases très utilisé est symbolisé par

n-n'n''...(++)G(**)

n désigne le nombre de gaussiennes de la couche interne.

n'n''... indiquent le nombre de gaussiennes utilisé dans chaque couche de valence.

++ (facultatif) désigne un (+) ou deux (++) ensembles de diffuses

****** (facultatif) désigne pour la première * des fonctions d sur les atomes de la deuxième période et des fonctions p sur H. Une notation équivalente est (...) G (d, p).

Par exemple, la base très utilisée 6-31G** comporte, pour le carbone, 6 gaussiennes pour l'orbitale 1s, un double ensemble de valence, 2s 2p décrit par 3 gaussiennes

et 2s' 2p' décrit par 1 gaussienne, avec des orbitales de polarisation d (p sur les hydrogènes). Ce code est reconnu par le programme GAUSSIAN.

Une autre famille de bases de bonne qualité est celle de Dunning. Elles sont codées cc-PVDZ, cc-PVTZ, cc-PVQZ, cc-PV5Z, cc-PV6Z : - cc signifie corrélation consistent - PV pour Polarisation Valence - XZ, pour Double, Triple, Quadruple ... Zêta. Ces méthodes offrent maintenant pour la plupart des complexes des métaux de transition, une description satisfaisante et cohérente des systèmes moléculaires et de leurs observables physiques associés [38].

5 Domaine d'application de la modélisation moléculaire

On peut diviser l'application de la MM en trois catégories :

- Soit pour obtenir une géométrie à laquelle on attache de l'intérêt. Cette situation se présente lorsque la modélisation guide l'interprétation des résultats provenant des études de structure par rayons X ou par diffraction électronique, ou lorsqu'il s'agit de modéliser une molécule pour les besoins de l'infographie.
- Dans l'interprétation des effets stériques sur la réactivité ou bien de la stabilité relative des isomères en tant qu'énergie stérique ou de tension.

- Quand aucune liaison n'est rompue, ni formée et qu'aucun intermédiaire chargé n'intervient, l'interconversion conformationnelle se prête particulièrement bien à une description par la MM. On peut obtenir grâce à cette analyse des informations structurales sous forme d'un profil énergétique (en fonction d'un angle dièdre par exemple) ou des cartes énergétiques 3D.

En conclusion, on peut dire que la mécanique moléculaire aujourd'hui est à la porte de tous les chercheurs. La mécanique moléculaire ne peut pas encore rivaliser avec la mécanique quantique dans beaucoup de domaine qui lui sont propre, mais elle reste une méthode de choix dans l'interprétation de phénomènes sous contrôle stérique et dans le calcul de structure. La mécanique moléculaire permet de passer en revue de grosses molécules (produits pharmaceutiques, colorants, etc.) pour établir des relations entre structure et réactivité et ainsi faire un tri avant de passer au stade expérimental. La différence du temps de calculs par la mécanique moléculaire par rapport aux autres méthodes quantiques est

d’environ de quelque puissance de dix, cette différence augmente en fonction de la taille de la molécule.

6 Limitation de la modélisation moléculaire la modélisation moléculaire

S’adresse surtout à des organiciens intéressés par des problèmes de réactivité et de structure de molécules comportant déjà un nombre significatif d’atomes, elle s’adresse aussi au biochimistes et pharmaciens préoccupés par la relation structure-activité. Si l'on veut exploiter intelligemment les programmes disponibles pour le calcul et la visualisation, certains principes de base doivent être retenus, il est nécessaire de connaître les origines de la méthode, ses potentialités et ses limites.

Cette méthode empirique, ne s’applique bien que lorsqu'on étudie des molécules voisines de celles qui ont servi à établir le champ de force. Plus on sophistique le champ de force de la mécanique moléculaire (MM2----->MM3) plus on a besoin de paramètres; il est difficile d'avoir un champ de force général et on s'oriente plutôt vers des champs de force spécifiques, sur les hydrocarbures conjugués, les protéines, les peptides [39]et les polymères,...etc. Enfin il faut toujours valider une étude en Modélisation Moléculaire par confrontation avec l’expérience (RX, RMN....) sur des molécules types.

References

[1] E. Cancès, C. Le Bris, and Y. Maday, "Convergence des algorithmes SCF," *Méthodes mathématiques en chimie quantique Une introduction*, pp. 201–227, 2006.

[2] J. Hladik, M. Chrysos, P.-E. Hladik, and L. U. Ancarani, *Mécanique quantique*. Masson, 1997.

[3] D. E. S. Sciences, D. E. L. A. Nature, and E. T. D. E. La, "Remerciements."

[4] E. Schrödinger, "SCHRÖDINGER 1926C," *Annalen der Physik*, vol. 79, p. 734, 1926.

[5] M. Born and R. Oppenheimer, "Zur Quantentheorie der Molekeln," *Annalen der Physik*, vol. 389, no. 20, pp. 457–484, 1927, doi: 10.1002/andp.19273892002.

[6] D. R. Hartree, "The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, no. 1, pp. 89–110, 1928.

[7] B. Nadia, "Modélisation de la structure cristalline d'un nouveau composé à propriétés optiques non linéaires." 2013.

[8] F. Jensen, *Introduction to computational chemistry*. John wiley & sons, 2017.

[9] D. U. D. Eterminisme, G. En, C. Eres, Q. Chez, V. Eg, and D. E. Q. T. L. Et, "Table des mati ` eres," *Revue des tudes Armniennes*, vol. 28, pp. 527–529, 2005.

[10] V. Fock, "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems," *Zeitschrift für Physik*, vol. 61, no. 1–2, pp. 126–148, 1930.

[11] J. C. Slater, "Atomic Shielding Constants," *Physical Review*, vol. 36, no. 1, pp. 57–64, Jul. 1930, doi: 10.1103/PhysRev.36.57.

[12] W. Pauli, "Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren," *Zeitschrift für Physik*, vol. 31, no. 1, pp. 765–783, 1925.

[13] J. A. Pople and R. Nesbet, "Self-consistent orbitals for radicals," *The Journal of Chemical Physics*, vol. 22, no. 3, pp. 571–572, 1954.

[14] G. Berthier, "Configurations électroniques incomplètes-Partie I. La Méthode du Champ Moléculaire Self-Consistent et l'Etude des Etats à Couches Incomplètes," *Journal de Chimie Physique*, vol. 51, pp. 363–371, 1954.

[15] A. Hinchliffe, *Modelling molecular structures*. J. Wiley, 1996.

[16] W. Kohn, "Nobel Lecture: Electronic structure of matter—wave functions and density functionals," *Reviews of Modern Physics*, vol. 71, no. 5, p. 1253, 1999.

[17] J. A. Pople, "Nobel lecture: Quantum chemical models," *Reviews of Modern Physics*, vol. 71, no. 5, p. 1267, 1999.

[18] P. Hohenberg and W. Kohn, "Density functional theory (DFT)," *Phys. Rev*, vol. 136, p. B864, 1964.

[19] J. HrdliCka, "Diary of Societies.," in *Proceedings of the Cambridge Philosophical Society*, 1931, vol. 27, no. Part 3.

[20] S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis," *Canadian Journal of physics*, vol. 58, no. 8, pp. 1200–1211, 1980.

[21] A. D. Becke, “Density functional calculations of molecular bond energies,” *The Journal of Chemical Physics*, vol. 84, no. 8, pp. 4524–4529, 1986.
[22] A. D. Becke, “Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals,” *The Journal of chemical physics*, vol. 107, no. 20, pp. 8554–8560, 1997.
[23] J. P. Perdew, “Density-functional approximation for the correlation energy of the inhomogeneous electron gas,” *Physical Review B*, vol. 33, no. 12, p. 8822, 1986.
[24] S. Kurth, J. P. Perdew, and P. Blaha, “Molecular and solid-state tests of density functional approximations: LSD, GGAs, and meta-GGAs,” *International journal of quantum chemistry*, vol. 75, no. 4-5, pp. 889–909, 1999.
[25] E. Clementi and S. J. Chakravorty, “A comparative study of density functional models to estimate molecular atomization energies,” *The Journal of chemical physics*, vol. 93, no. 4, pp. 2591–2602, 1990.
[26] D. C. Patton and M. R. Pederson, “Erratum: Application of the generalized-gradient approximation to rare-gas dimers [Phys. Rev. A 56 R2495 (1997)],” *Physical Review A*, vol. 71, no. 1, p. 19906, 2005.
[27] J. C. REINHARDT, “2005,” in *The Winning Cars of the Indianapolis 500*, 2019.
[28] A. D. Becke, “A half-half theory of density functionals,” *J. Chem. Phys*, vol. 98, p. 1372, 1993.
[29] A. D. Becke, “P hys. Rev. A 1988, 38, 3098. b) C. Lee, W. Yan, RG Parr,” *Phys. Rev. B*, vol. 37, p. 785, 1988.
[30] A. D. Becke, “Density-functional thermochemistry. IV. A new dynamical correlation functional and implications for exact-exchange mixing,” *The Journal of chemical physics*, vol. 104, no. 3, pp. 1040–1046, 1996.
[31] P. J. Stephen, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, “Ab Initio Calculation of Vibrational Absorption,” *J. Phys. Chem*, vol. 98, pp. 11623–11627, 1994.
[32] Y. Zhao, N. González-García, and D. G. Truhlar, “Benchmark database of barrier heights for heavy atom transfer, nucleophilic substitution, association, and unimolecular reactions and its use to test theoretical methods,” *The Journal of Physical Chemistry A*, vol. 109, no. 9, pp. 2012–2018, 2005.
[33] E. Runge and E. K. U. Gross, “Density-functional theory for time-dependent systems,” *Physical Review Letters*, vol. 52, no. 12, p. 997, 1984.
[34] E. K. U. Gross and W. Kohn, “Time-dependent density-functional theory,” *Adv. Quantum Chem*, vol. 21, no. 255, pp. 287–323, 1990.
[35] E. K. U. Gross and W. Kohn, “Local density-functional theory of frequency-dependent linear response,” *Physical review letters*, vol. 55, no. 26, p. 2850, 1985.
[36] C. G. and J. A. P. M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, T. Vreven, K. N. Jr., Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Peterss, “Gaussian 03, Revision B.04.” Gaussian, 2003.
[37] M. J. Frisch *et al.*, “Gaussian 09, Revision A. 02, 2009, Gaussian,” *Inc., Wallingford CT*, 2009.
[38] T. Ziegler, “Density functional theory study of vibrational spectra of small molecules,” *Chem Rev*, vol. 91, p. 651, 1991.
[39] D. F. Mierke, O. E. Said-Nejad, P. W. Schiller, and M. Goodman, “Enkephalin analogues containing β-naphthylalanine at the fourth position,” *Biopolymers: Original Research on Biomolecules*, vol. 29, no. 1, pp. 179–196, 1990.