Université Sidi Mohammed Ben Abdellah
Faculté des Sciences Dhar El Mahraz- Fès
Centre d'Etudes Doctorales
"Sciences et Technologies"

Formation Doctorale : SMPI

Discipline : Chimie Physique et Modélisation

Spécialité : **Chimie**

Laboratoire : LIMOM

THESE DE DOCTORAT

Présentée par

HADAJI El ghalia

Corrélation - (structure - activité anticancéreuse) par les méthodes QSAR des molécules hétérocycliques précurseurs de médicaments

Soutenue le 26/12/2018 devant le jury composé de :

Pr. **Mohammed LACHKAR**	**Faculté des sciences dhar mahraz, Fès**	***Président***
Pr. **Hamid TOUFIK**	**Faculté Polydisciplinaire, Taza**	***Rapporteur***
Pr. **Hamid MAGHAT**	**Faculté des sciences, Meknès**	***Rapporteur***
Pr. **Tahar LAKHLIFI**	**Faculté des sciences, Meknès**	***Rapporteur***
Pr. **Fouad KHALIL**	**Faculté des sciences et techniques Fès**	***Examinateur***
Pr. **Mohammed BOUACHRINE**	**Faculté des sciences, Meknès**	**Co-Directeur**
Pr. **Abdelkrim OUAMMOU**	**Faculté des sciences dhar mahraz, Fès**	**Directeur de thèse**

Année universitaire : 2018-2019

Résumé :

Le présent travail de thèse comporte trois applications de la relation quantitative structure-activité pour des séries de molécules anticancéreuses, qui sont réalisées à partir d'une base de données de 40, 32 et 29 molécules dérivées respectivement des sulfonamides, pyrazoliques et quinoléines.

La méthode de modélisation moléculaire DFT (B3LYP/6-31G(d)) a été utilisée dans notre travail afin de déterminer les paramètres structuraux, électroniques et énergétiques associés aux molécules étudiées. Divers descripteurs moléculaires sont calculés avec les logiciels GAUSSIAN03, CHEMSKETCH, et CHEMOFFICE. L'ensemble des données est soumis à des études statistiques : l'analyse en composantes principales ACP, la régression linéaire multiple RLM, la régression non linéaire multiple RNLM et la régression par les moindres carrés partiels PLS. Les modèles linéaires et non linéaires obtenus, ont été validés selon les cinq principes établis par l'Organisation de Coopération et de Développement Economique (OCDE). Une forte corrélation entre les valeurs des activités expérimentale et prédite a été observée, ce qui indique la fiabilité, la robustesse et la bonne qualité des modèles obtenu. Les derniers seront appliqués avec succès en vue de prédire des activités de nouveaux composés, dont les données expérimentales sont indisponibles.

***Mots-clés**: QSAR; Anti-cancer; RLM; PLS; RNLM; Validation croisée (CV); Y-randomisation; Règle de Lipinsky.*

Abstract:

The present study of my thesis defends involve the theorical investigation of the quantitative relationship structure-activity for series of anti-cancer molecules, which are made from a library of 40, 32, 30 molecules derived respectively from sulfonamides, pyrazole and quinoline.

The molecular modeling method DFT (B3LYP / 6-31G) was used in our work to determine the structural, electronic and energetic parameters associated with the molecules studied. Various molecular descriptors are calculated with the GAUSSIAN, CHEMSKETCH and CHEMOFFICE software. The data set is subjected to statistical studies: ACP main component analysis, multiple linear regression RLM, multiple nonlinear regression RNLM and regression by partial least squares PLS. The models obtained, linear and non-linear, have been validated according to the five principles established by the Organization for Economic Co-operation and Development (OECD). The strong correlation between the values of experimental and predicted activity was observed, indicating the reliability the robustness and good quality of the QSAR model construct. From which the developed models are applied successfully in order to predict activities / properties of new compounds that the experimental data of which are indispensible.

Keywords*: QSAR; Anti-cancer; MLR; PLS; MNRL; Cross validation (CV); Y-randomization; Lipinsky.*

Publications scientifiques à l'issu de ce travail

[1] E.G. Hadaji, M. Bourass, A. Ouammou, M. Bouachrine, 3D-QSAR models to predict the antiviral activities of a series of novel N-phenylbenzamide and N-phenylacetophenone compounds based on density functional theory using statistical methods, Moroccan Journal of Chemistry 4(2016) 204–214.

[2] E.G. Hadaji, M.Bourass, A. Ouammou, M. Bouachrine, 3D-QSAR models to predict anti-cancer activity on a series of protein P38 MAP kinase inhibitors, Journal of Taibah University for Science 11 (2017) 392–407.

[3] E.G. Hadaji, M. Bourass, A. Ouammou, M. Bouachrine, QSAR study of (E)-N-Aryl-2-ethene-sulfonamide analogues as microtubule targeted agents in prostate cancer based on density functional theory using statistical methods, Advances in Physical Chemistry, Volume 2017, Article ID 7629056, 14 pages https://doi.org/10.1155/2017/7629056

[4] H. Zaki, M. Bourass**, E.G. Hadaji**, , A. Ouammou, M. Benlyass, M. Bouachrine, QSAR analyses of Octahydroquinazolinone for insecticidal activity against spodoptera litura and its in-silico validation using molecular Docking study, Moroccan Journal of Chemistry. 5(2017) 202-211.

[5] E.G. Hadaji, A. Ouammou, M. Bouachrine, QSAR Study of Anthra[1,9-cd]pyrazol-6(2H)-one Derivatives as Potential Anticancer Agents Using Statistical Methods, Advances in Chemistry, Volume 2018, Article ID 3121802, 16 pages**.**

Présentations par affiches

(1) E.G. Hadaji, M.Bourass, A. Ouammou, M. Bouachrine. 3D-QSAR models to predict anti-cancer activity on a series of protein P38 MAP kinase inhibitors, 3^{éme} Edition de la Journée Internationale de Biotechnologie Médicale, organisée par le laboratoire de biotechnologie de la Faculté de Médecine et de Pharmacie de Rabat sous le thème: "Les applications biotechnologiques au Service du Biomédical" 17 Décembre 2015.

(2) E.G. Hadaji, M.Bourass, A. Ouammou, M. Bouachrine. 3D-QSAR models to predict the antiviral activities of a series of novel N-phenylbenzamide and N-phenylacetophenone compounds based on density functional theory using statistical methods, Journées doctorales organisées à la faculté des sciences de Meknès sous les thème: "Science de l'environnement, partager c'est pérenniser" 2,3 Juin 2016

(3) E.G. Hadaji, A. Ouammou, M. Bouachrine. QSAR study of (E)-N-Aryl-2-ethene-sulfonamide analogues as microtubule targeted agents in prostate cancer using statistical methods, 1ère Conférence internationale sur les matériaux et les sciences de l'environnement ICME 2016, 1,2,3 Décembre 2016, Oujda, Morocco.

(4) E.G. Hadaji, A. Ouammou, M. Bouachrine. QSAR study of (E)-N-Aryl-2-ethene-sulfonamide analogues as microtubule targeted agents in prostate cancer using statistical methods, 3éme édition du colloque international sur la valorisation des déchets pour un développement durable 2016 : Prévention de la pollution et gestion durable des rejets solides et liquides, Fculté des sciences Dhar Mahraz FES, 05 Novembre 2016.

(5) E.G. Hadaji, A. Ouammou, M. Bouachrine 3D-QSAR models to predict anti-cancer activity on a series of protein P38-MAP kinase inhibitors, 5éme édition internationale des journées jeunes chercheurs de chimie thérapeutique 2016 : place de la phytothérapie et la chimie médicinale dans le processus de « DRUG DISCOVERY », Faculté polydisciplinaire TAZA, 28-29 Novembre 2016.

(6) E.G. Hadaji, A. Ouammou, M. Bouachrine. Etudes QSAR des dérivés quinoleinique a activité anticancéreuse par utilisation des descripteurs électronique et physico-chimiques, 7th International Metting on Chemometrics and Quality, faculté des sciences et technique FST Fes, 24-25 October 2018.

(7) E.G. Hadaji, A. Ouammou, M. Bouachrine. Etude QSAR de l'activité anticancéreuse des pyrazoles à l'aide des descripteurs moléculaires, 7th International Metting on Chemometrics and Quality, faculté des sciences et technique FST Fes, 24-25 October 2018.

(8) R. Kacemi, E.G. Hadaji, A. Ouammou, H. Toufik. Synthesis and anticancer activity of new hétérocyclic, 7th International Metting on Chemometrics and Quality, faculté des sciences et technique FST Fes, 24-25 October 2018.

Présentations orales

(1) E.G. Hadaji, M. Bourass, A. Ouammou, M. Bouachrine. 3D-QSAR Models to Predict Anti-cancer Activity on a Series of Protein P38-MAP Kinase Inhibitors, 5ème Journée Nationale sur la Chimie Verte, Faculté des sciences Tétouan, 4 et 5 Novembre 2016.

(2) E.G. Hadaji, M. Bourass, A. Ouammou, M. Bouachrine. QSAR study of (E)-N-Aryl-2-ethene-sulfonamide analogues as microtubule targeted agents in prostate cancer using statistical, 5éme édition internationale des journées jeunes chercheurs de chimie thérapeutique 2016 : place de la

phytothérapie et la chimie médicinale dans le processus de « DRUG DISCOVERY », Faculté polydisciplinaire TAZA, 28-29 Novembre 2016.

Liste des figures

Chapitre I:

Chapitre II:

Chapitre III:

Chapitre IV:

Liste des tableaux

Chapitre II:

Chapitre III:

Chapitre IV:

Abréviations

ACP	Analyse en Composantes Principales
ANN	Artificial Neural Network
AG	Algorithme génétique
B3LYP	Becke 3-Parameter Lee-Yang-Parr
CLAO	Combinaison Linéaire d'Orbitales Atomiques
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Index Analysis
CV	Cross Validation
D	Density
DA	Domaine d'Applicabilité
DFT	Density Functional Theory
DS	Déviation Standard
ET	Energie Totale
ER	Energie de répulsion
HF	Hartree-Fock
HOMO	Highest Occupied Molecular Orbital
LDA	Local Density Approximation
LOO	Leave-One-Out
LUMO	Lowest Unoccupied Molecular Orbital
MR	Molecular Refractivity
MSE	Mean Squared Error
MAE	Mean Absolute Error
MV	Molecular Volume
MW	Molecular Weight
n	Index of refraction
OA	Orbitale Atomique
OECD	Organisation for Economic Cooperation and Development
OM	Orbitale Moléculaire
Pc	Parachor
PLS	Partial Least Squares regression
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Proprety Relationship
RLM	Régression Linéaire Multiple
RNLM	Régression Non Linéaire Multiple

SCE	Somme des Carrés des Ecarts
SCF	Self Consistent Field (Champ Self Consistant)
VIF	Variance Inflation Factor

Table de matières

Introduction générale

Le cancer est un problème majeur de santé publique dans le monde. Le nombre de nouveaux cas de cancer en 2012 est estimé à 14,1 millions et le nombre de décès à 8,2 millions[1]. Il a été estimé en 2018 que 70% des décès par cancer dans le monde survenaient dans les pays en développement. La fréquence des cancers pourrait augmenter de 50 % dans le monde, avec 15 millions de nouveaux cas par an en 2020. A l'horizon 2030, il est prévu que le nombre de décès par cancer dans le monde s'augmente à 13,1 millions[2-5]. Bien que son incidence soit en augmentation dans tout les régions du monde, les taux d'incidence demeurent les plus élevés dans les régions les plus développées, mais la mortalité est relativement beaucoup plus élevée dans les pays en développement, faute de détection précoce et d'accès aux traitements[6].

Les progrès scientifiques de ces dernières années permettent aujourd'hui de déchiffrer les codes génétiques du cancer et de comprendre à quel point cette maladie est liée aux mécanismes de la vie[7]. La recherche de nouvelles molécules thérapeutiques capables de stopper la prolifération tumorale constitue l'un des principaux thèmes de recherche en cancérologie. Ces molécules qu'elles soient naturelles ou synthétiques sont généralement sélectionnées pour leurs effets anti-prolifératifs sur des lignées cancéreuses en culture.

Découvrir de nouveaux médicaments de la façon la plus efficace et la moins coûteuse possible constitue un enjeu majeur pour les années à venir[8]. 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[9]. De plus, le développement d'un médicament demande généralement entre 10 et 15 ans de recherches. 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. Le prix de revient d'un médicament est essentiellement dû à ses synthèses longues, coûteuses et finalement inutiles[10]. Le développement d'outils informatiques fiables couplé à la croissance de la puissance informatique a permis la mise en place de techniques de modélisation moléculaire, qui sont devenues, actuellement des outils indispensables dans le domaine de la conception des médicaments. Pour cette raison, l'industrie pharmaceutique s'oriente vers de nouvelles méthodes de recherche, appelées modélisation moléculaire, qui consistent à prédire les propriétés et activités des molécules avant même que celles-ci ne soient synthétisées[11-12].

Les ordinateurs sont devenus des outils indispensables en chimie pharmaceutique moderne. Leur rôle est essentiel, tant au niveau de la découverte de nouveaux médicaments que du développement de ceux-ci. Les progrès rapides réalisés dans les logiciels et dans le matériel qui les accompagnent prouvent que la plupart des opérations qui étaient, jadis uniquement, réalisables par des informaticiens avertis peuvent maintenant être exécutées par des pharmaco-chimistes, avec des

ordinateurs couramment employés aux laboratoires, pour autant qu'ils possèdent les notions élémentaires de mécanique quantique et autres équations qui ont trait aux molécules[13].

Deux disciplines de la « chimie computationnelle » se sont développées en réponse à ce besoin : les relations structure-activité ou QSAR (Quantitative Structure-Activity Relationships), et les relations structure-propriété ou QSPR (Quantitative Structure-Property Relationships)[14-15]. Elles consistent essentiellement en 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, chimiques et l'activité biologique de composés, de développer de nouvelles théories ou de comprendre les phénomènes observés[16]. Elle permet également de guider les synthèses de nouvelles molécules, sans avoir à les réaliser, ou à analyser des familles entières de composés. La modélisation moléculaire est une application des méthodes théoriques et des méthodes de calcul pour résoudre des problèmes impliquant la structure moléculaire et la réactivité chimique[17-18].

Le principal objectif de ce travail est l'application de différentes méthodes de la modélisation moléculaire pour prédire les activités biologiques attendues dans des nouvelles molécules bioactives pour les trois types de molécules étudiées. Il s'agit donc, dans ces travaux de thèse, de développer et d'évaluer des modèles QSAR pour la prédiction de l'activité anticancéreuse reliant les activités expérimentales aux structures moléculaires, calculées à l'aide d'outils de la chimie quantique, en particulier la Théorie de la Fonctionnelle de la Densité (DFT)[19-20].

- Le **premier chapitre** vise à introduire les différents outils nécessaires à la mise en place des modèles QSAR. Le principe de la méthodologie QSAR, les descripteurs moléculaires ainsi que les outils d'analyse de données employés pour développer et évaluer les modèles, seront ainsi détaillés;
- Nous présenterons dans le **deuxième chapitre** de ce mémoire les résultats des études QSAR pour une série de 40 dérivés sulfoniques à l'aide de 16 descripteurs moléculaires; Dans le **troisième chapitre** nous effectuerons une étude QSAR de l'activité anticancéreuse des pyrazoles et prédiction de nouvelles molécules à l'aide des modèles développés; Dans le **quatrième chapitre** nous attaquerons une étude QSAR de l'activité anticancéreuse des dérivés de la 9-chloro-l H-indéno [1,2-c] quinoléine-11-one (dérivés de l'azafluorénone tétracyclique) en utilisant des méthodes statistiques. Enfin, nous terminerons ce manuscrit par une conclusion générale et des perspectives.

Références

1 E. Esteve, D. Bazin, C. Jouanneau , How to assess the role of Pt and Zn in the nephrotoxicity of Pt anti-cancer drugs?: an investigation combining XRF and statistical analysis. Part II:clinical application, Comptes Rendus Chimie, 19, 1580–1585, **2015**.

2 R. Siegel, D.Naishadham, A. Jemal, Cancer statistics 2013, CA Cancer Journal for Clinicians, 63, 11–30, **2013.**

3 R. L. Siegel, K. D.Miller, A. Jemal, Cancer statistics, CA Cancer Journal for Clinicians, 65, 5–29, **2015.**

4 E. Esteve., D. Bazin, C. Jouanneau, S. Rouziere, A. Bataille., A. Kellum, K. Provost, C. Mocuta, S. Reguer, D. Thiaudiere, K. Jorissen, A. Hertig, E. Rondeau, E. Letavernier, M. Daudon, P. Ronco, How to assess the role of Pt and Zn in the nephrotoxicity of Pt anti-cancer drugs. An investigation combiningm XRF and statistical analysis: Part I: On mice. C. R. Chimie, 19, 1580-1585, **2016**.

5 K. Morgans., C.Bommel, C. Stowell C, L. Abrahm, E. Basch, E.J. Bekelman, D. Berry., A. Bossi, I. Davis, T. Reijke, L. Denis, S. Evans, N. Fleshner, D. George, J. Kiefert, W. Lin, G. Matthe, R. McDermott, H. Payne, G. Roos, D.Schrag, T. Steuber, B.Tombal, J. Basten, M. Hoeven, F. Penson, Development of a Standardized Set of Patient-centered Outcomes for Advanced Prostate Cancer: An International Effort for a Unified Approach, European Urology, 68, 58-79, **2015**.

6 P. Wingo, C. Cardinez, S. Landis, Long-term trends in cancer mortality in the United States. Cancer, 97, 3133-3275, **2003**.

7 A. K. Morgans, A. C. M. Van Bommel, C. Stowell, Development of a standardized set of patient-centered outcomes for advanced prostate cancer: an international effort for a unified approach, EuropeanUrology,68-5, 891–898, **2015.**

8 N. Margossian, Le règlement REACH-La règlementation européenne sur les produits chimiques, Dunod/L'UsineNouvelle,Paris, **2008**.

9 S.P. Bradbury, Quantitative structure-activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research, Toxico. Lett., 79, 229-237, **1995.**

10 M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure-property relationships in pharmaceutical research–Part 2, Pharm. Sci. Tech. Today, 3, 50-57, **2000.**

11 R. Perkins; H. Fang; W. Tong, W.J. Welsh, Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. Environ. Toxicol. Chem, 22, 1666–1679, **2003**.

12 C.D. Selassie, History of quantitative structure-activity relationships. In: Abraham DJ (ed) Burger's medicinal chemistry and drug discovery, Drug DiscoveryWiley, New York, 1–48, **2003**.

13 M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure-property relationships in pharmaceutical research–Part 2, Pharm. Sci. Tech. Today, 3, 50-57, **2000**.

14 T. Puzyn, J. Leszczynski, M. T. D. Cronin, Recent advances in QSAR studies-Methods and applications , Springer Dordrecht Heidelberg London New York, **2010.**

15 E. Tenorio-Borroto, C. G. Rivas, J. C. Vásquez Chagoyán, N. Castañedo, F. J. Prado-Prado, X. García-Mera, G. Díaz. Computer-Aided Drug Design, Synthesis and Evaluation of New Anti-Cancer Drugs Bioorg. Med. Chem, 20, 6181–6194, **2012.**

16 V. Prachayasittikul V., Pingaew R., Worachartcheewan A., Nantasenamat C., Prachayasittikul S., Ruchirawat S., Prachayasittikul, Synthesis, anticancer activity and QSAR study of 1,4-naphthoquinone derivatives. European Journal of Medicinal Chemistry, 84, 247-263**, 2014**.

17 D.Winkler, A. Brief Association rule mining of cellular responses induced by metal and metal oxide nanoparticles. Bioinformatics, 57, 3-73, **2002.**

18 H. Gao, J.A. Katzenellenbogen, R. Garg, C. Hansch, Comparative QSAR Analysis of Estrogen Receptor Ligands, Chem. Rev., 37, 99-723, **1999.**

19 J. Taskinen, J. Yliruusi, Prediction of physicochemical properties based on neural network modeling, Adv. Drug Deliv. Rev. 55, 1163-1183, **2003**.

20 M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure-property relationships in pharmaceutical research-Part 1, Pharm. Sci. Tech. Today, 3, 28-35, **2000.**

Chapitre I

Méthodes QSAR et analyse des données

1. INTRODUCTION

La relation quantitative structure à activité QSAR (Quantitative Structure-Activity Relationship) est le procédé par lequel une structure chimique est corrélée avec un effet bien déterminé comme l'activité biologique ou la réactivité chimique[1-3]. Ses premiers essais ont commencé à la fin du 19èmesiècle, lorsque Crum -Brown et Frazer ont postulé que l'activité biologique d'une molécule dépend de sa constitution chimique. Mais ce n'est qu'au début des années 60 que les travaux de Corwin Hanch ont proposé un modèle mathématique qui relie l'activité biologique à la structure chimique. Ainsi, l'activité biologique s'exprime de manière quantitative, comme la concentration de substance nécessaire pour obtenir une certaine réponse biologique. Lorsque les propriétés ou structures physiochimiques sont exprimées par des chiffres, nous pouvons proposer une relation mathématique, ou relation quantitative structure à activité, entre l'activité étudiée et la structure moléculaire. L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de la réponse biologique pour des structures ayant des substituants différents[4-6].

Aujourd'hui, l'utilisation de QSAR n'a cessé de progresser. Elle est devenue indispensable en chimie pharmaceutique et pour la conception des médicaments, notamment dans le cas où la disponibilité des échantillons est limitée, ou les mesures expérimentales sont dangereuses, longues et chères[7].

Sans l'utilisation de grands instruments analytiques, les résultats des études QSAR peuvent fournir des informations utiles pour obtenir une meilleure connaissance des structures moléculaires et probablement le mode d'action au niveau moléculaire. Ces informations peuvent être alors utilisées dans la prédiction des activités biologiques de nouveaux composés, et dans la conception de nouvelles structures[8-9].

2. PRINCIPE DES METHODES 2D-QSAR

Comme son noms l'indiquent, le principe des méthodes 2D-QSAR est de mettre en œuvre une équation mathématique reliant de manière quantitative des propriétés moléculaires aussi bien électroniques que géométriques appelés descripteurs à une propriété bien déterminée comme l'activité biologique ou la réactivité chimique par utilisation des méthodes d'analyses des données[10].

Partant de ces relations, nous pouvons développer des modèles prédictifs de la forme générale suivante :

Activité = *f* (descripteurs moléculaires).

Ainsi, l'objectif de ces méthodes est d'analyser les données structurales afin de déterminer les facteurs influençant l'activité mesurée. Pour ce faire, différents types d'outils statistiques peuvent être employés :

- Régressions linéaires simples et multiples,

- Régressions des moindres carrées partiels (PLS),
- Régressions non linéaires multiples,
- les réseaux de neurone artificiels ANN.....

Une fois cette relation est établie et validée, elle peut alors être employée pour la prédiction de la propriété /activité de nouvelles molécules, pour lesquelles les valeurs expérimentales ne sont pas disponibles[11-13]. De tels modèles peuvent être également utilisés pour mieux comprendre les mécanismes et les modes d'action.

3. OBJECTIFS DE QSAR

L'objectif principal de toute analyse QSAR réside dans le développement rationnel d'un modèle mathématique accompagné de l'exploration de l'information chimique qui y est impliquée. Une telle modélisation utilise toujours moins de données de réponse chimique et permet la prédiction d'un nombre relativement important de composés à courte durée et sans mobilisation humaine. Ceci fournit une opportunité pour que cette technique soit utilisée dans divers domaines[14-15].

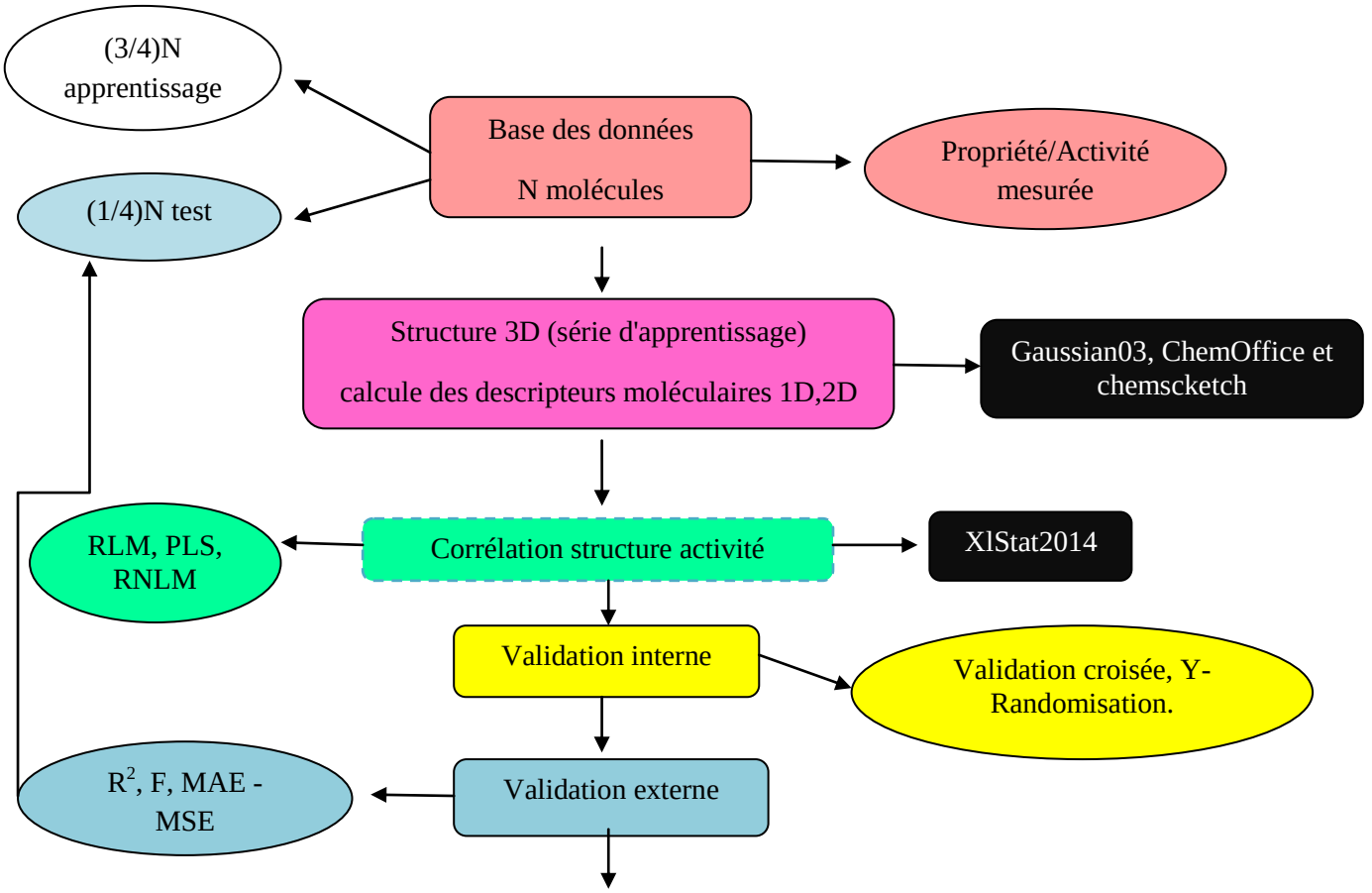
4. APPLICATIONS DE QSAR

Les produits chimiques représentent une partie indispensable de la nécessité humaine compte tenu des diverses applications allant des processus de laboratoire aux processus industriels, ainsi que l'utilisation du ménage. QSAR présente une option appropriée dans la surveillance rationnelle de l'activité / propriété / toxicité des produits chimiques et donc utile dans une grande variété d'applications. Étant donné que la mise au point de la nature comportementale des produits chimiques donne des résultats fructueux pour une classe importante de produits chimiques impliquant des produits pharmaceutiques, agrochimiques, parfumeries, réactifs analytiques, solvants, agents modificateurs de surface[16-17], etc.

Les domaines d'application de la technique QSAR sont énormes. Dans une perspective globale, les produits chimiques modélisés selon la méthode QSAR peuvent être visualisés en trois grands types, à savoir les produits chimiques bénéfiques pour la santé (médicaments, produits pharmaceutiques, ingrédients alimentaires, etc.), les produits chimiques industriels (solvants, réactifs, etc. .), et les produits chimiques qui en résultent et qui sont dangereux [polluants organiques persistants (POP), toxines, xénobiotiques, carcinogènes, composés organiques volatils (COV), etc.]. Il est intéressant de noter qu'en plus de modéliser l'activité biologique et les paramètres de toxicité, le paradigme de conception de médicaments implique la modélisation qui vise à surveiller le profil pharmacocinétique des médicaments candidats avant leur synthèses et ainsi améliorer l'efficacité des composés conçus dans le système biologique[18].

5. METHODOLOGIE GENERALE D'UNE ETUDE QSAR

Généralement une étude QSAR débute par la constitution de la base des données à partir des mesures expérimentales (propriété ou de l'activité de chaque composé) fiables et en nombre le plus important possible. Il s'agit ensuite de calculer un certain nombre de descripteurs car les paramètres qui décrivent l'activité ou la propriété étudiée sont mal connus, puis nous sélectionnons parmi les descripteurs calculés ceux qui sont pertinents par construction des modèles de RLM, PLS et RNLM et en divisant la base de données, aléatoirement, en une série d'apprentissage (training set) qui contient généralement les 3/4 de la base de données et une série de test (test set) constituée par le 1/4 restant. Ainsi les modèles mathématiques en utilisant la série d'apprentissage sont obtenus, et les modèles élaborés sont validés en utilisant la série de test et leurs paramètres statistiques sont calculés, et pour tester la fiabilité des résultats obtenus la méthode de validation interne est mise en jeu. Enfin les modèles élaborés peuvent être exploités pour comprendre les mécanismes et les modes d'action de l'activité étudiée et aussi pour prédire de nouvelles molécules avant même que celle-ci ne soient synthétisées[19-20].



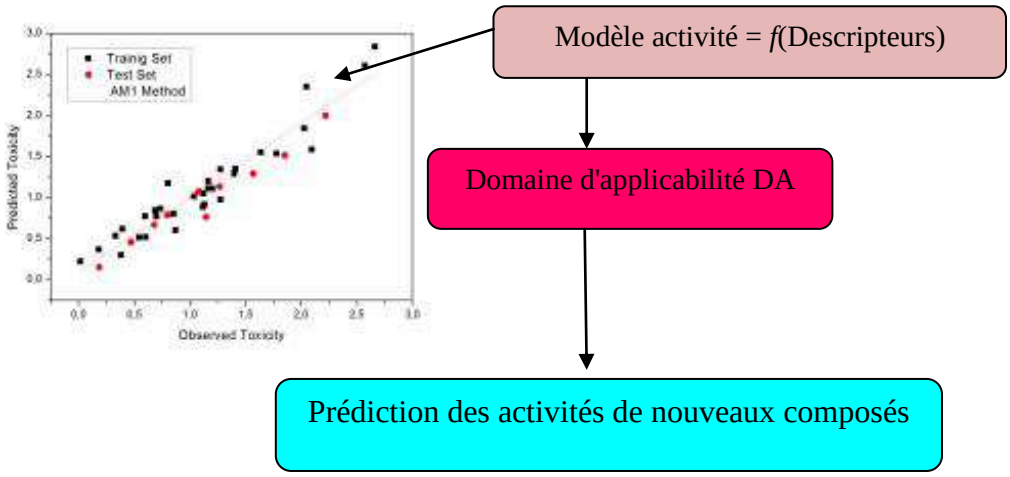


Figure 1: *Méthodologie générale d'une étude QSAR.*

6. BASES DE DONNEES ET PARAMETRES BIOLOGIQUES

Le choix de la base de données est une étape très importante dans le développement des modèles QSAR, puisque ses constructions sont dépendantes des données expérimentales utilisées.

Les données devraient, idéalement, être de grande qualité, ce qui signifie qu'elles devraient être fiables et cohérentes. Il est donc impératif de choisir ceux ayant des incertitudes faibles afin de limiter les barres d'erreurs expérimentales. De plus, le modélisateur doit s'assurer que les données expérimentales utilisées ont été obtenues selon le même protocole. En effet les conditions expérimentales ont, généralement, une forte influence sur les valeurs obtenues[21].

Il faut également que la distribution des données soit la plus homogène et normale que possible, car la plupart des méthodes statistiques sont basées sur ce type de distribution. L'efficacité d'un modèle QSAR dépend également du type de molécules qui y sont incluses, plus le modèle présentera des composés de structures proches et similaires, plus il aura de la chance d'être performant. Les données biologiques sont généralement exprimées en logarithmes inverses (log1/C) afin d'obtenir des valeurs mathématiques plus significatives lorsque les structures sont biologiquement très efficaces[22].

7. DESCRIPTEURS MOLECULAIRES

L'activité biologique ne peut pas être reliée directement à la structure moléculaire, cette dernière est codée par des grandeurs qui représentent d'une manière quantitative les informations contenues dans la structure moléculaire telles que les caractéristiques physico-chimiques et structurales. Ces grandeurs sont appelées descripteurs. Un descripteur moléculaire peut être considéré comme la conséquence d'un processus logique et mathématique, appliqué à l'information chimique codifiée à travers la représentation d'une molécule. Une fois les descripteurs sont disponibles, il est possible

d'établir des relations entre la structure moléculaire et une activité (ou propriété) à l'aide des outils de la modélisation. On en dénombre aujourd'hui des milliers de descripteurs, qui peuvent être calculés ou obtenus de manière empirique. L'information codée d'un descripteur moléculaire dépend du type de représentation moléculaire employée et de l'algorithme défini pour son calcul[23-24]. Il existe :

- Des descripteurs moléculaires simples dérivés du nombre d'atome-type ou de fragments structuraux de la molécule[25],
- Des descripteurs moléculaires dérivés d'une représentation géométrique, qui s'appellent géométriques ou descripteurs 3D[26].

De nombreux logiciels ont été développés pour calculer les différents descripteurs moléculaires, tels que : Gaussian 03[27], Chem Draw Office[28], Chem Scketch [29].....

7.1. Les descripteurs 1D :

Les descripteurs 1D sont obtenus 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.

Vu leur extrême simplicité, ces descripteurs sont couramment utilisés. Cependant, ils peuvent poser des problèmes d'interprétation des mécanismes, et d'interaction du fait qu'ils ne permettent pas de tenir en compte des effets stériques et d'isomérie[30].

Dans nos travaux nous avons utilisé :

- Le poids moléculaire, noté MW (appelé aussi le poids de formule), mesuré en daltons (Da). C'est la somme des poids atomiques des différents atomes constituant la molécule. 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 avoir un médicament. Pour les médicaments délivrés par voie orale, le poids moléculaire doit être inférieur ou égal à 500 daltons (optimum autour de 300 daltons) [31].

7.2. Les descripteurs 2D :

Les descripteurs 2D sont calculés à partir de la formule développée de la molécule. On distingue :

- Les indices 2D constitutionnels : qui caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles[32],…etc.
- Les indices 2D topologiques : peuvent être obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications.

Exemples: indice de Wiener [33], indice de Randić [34], indice de connectivité de valence de Kier-Hall[35], indice de Balaban[36], …etc

Ces descripteurs 2D permettent de prédire les propriétés physiques, mais sont insuffisants pour expliquer certaines propriétés et activités biologiques.

7.3. Les descripteurs 3D :

Les descripteurs 3D sont évalués à partir des positions relatives des atomes dans l'espace, et décrivent des caractéristiques plus complexes. Leurs calculs nécessitent donc de connaître la géométrie 3D de la molécule[37].

- Les descripteurs 3D géométriques : les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie[38].....
- Les descripteurs 3D électroniques : permettent de quantifier les différents types d'interactions inter- et intramoléculaires. Les descripteurs 3D ont une grande influence sur l'activité biologique des molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie est minimale, et fait souvent appel à la chimie quantique[39].

  Plusieurs descripteurs moléculaires sont classés en catégorie 3D, à savoir :

7.3.1 La polarisabilité

La polarisabilité c'est la facilité avec laquelle le nuage électronique peut se déformer sous l'effet d'un champ électrique (proportionnelle au volume atomique $4/3\pi r^3$ et au nombre d'électrons)[40].

$$\mathbf{P = \alpha.E}$$

P : dipôle qui est créé,

α : polarisabilité,

E : champ électrique,

r : rayon atomique.

7.3.2 L'hydrophobicité

Le caractère hydrophobe d'une drogue est crucial en ce qui concerne la facilité avec laquelle elle traverse les membranes cellulaires et peut également être un facteur important lors de ses interactions avec le récepteur (la protéine), il est donc important de pouvoir le quantifier[41].

7.3.3 Le coefficient de partage (P)

Le coefficient de partition est défini comme le rapport de la concentration du soluté dans la phase huileuse C` à la concentration du soluté non ionisé dans la phase aqueuse C, à l'équilibre. Il serait avantageux de ne pas être obligé de synthétiser chaque composé pour mesurer son coefficient de partage[42].

L'algorithme du coefficient de partage est le rapport :

P = Concentration de la drogue dans l'octanol / Concentration de la drogue dans l'eau

Permet d'estimer la biodisponibilité d'une molécule

0 < Log P < 3 : Activité biologique optimale (perméabilité, solubilité).

Log P < 0 : Composés trop hydrophiles (mauvaise perméabilité de bicouche lipidique).

Log P > 3 : Composés trop lipophiles (mauvaise solubilité aqueuse)[43].

7.3.4 L'énergie d'hydratation

L'énergie d'hydratation est la propriété physico-chimique qui est calculée pour chaque molécule est un facteur clé déterminant la stabilité de différentes conformations moléculaires. Il a été montré que l'hydratation représente jusqu'à 50% de l'affinité et de la spécificité dans les mécanismes de reconnaissance moléculaire. L'association entre une protéine et un ligand (ou principe actif) est généralement accompagnée par un départ de molécules d'eau. Celui-ci fournit une contribution à la fois entropique et enthalpique à l'énergie qui provient de la différence d'activité de l'eau entre la solution et la sphère d'hydratation de la protéine et du ligand. Par conséquent un changement de l'activité de l'eau peut modifier les propriétés de la liaison. Ces mécanismes ont un impact important dans la conception des principes actifs qui doivent posséder une affinité importante et spécifique pour une cible thérapeutique donnée [44].

7.3.5 Réfractivité molaire

Elle est généralement désignée comme une simple mesure du volume occupé par un individu atome ou un cluster d'atomes, cependant, la réfractivité molaire (RM) peut être obtenue à l'aide de l'expression suivante :

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

n : Indice de réfraction.

MM : Masse moléculaire.

d : Densité.

MM/ d : Volume.

$(n^2-1)/(n^2+1)$: Facteur de correction.

La réfraction molaire est particulièrement importante dans une situation dans laquelle le substituant possède une liaison π ou paires d'électrons célibataires[45].

7.3.6 Volume moléculaire

Le volume moléculaire est une fonction de la masse molaire (MM) et de la structure et tient compte de toutes les conformations existantes et accessibles à la molécule de les former. Ceci se rapporte réellement aux liens rotatifs et au nombre d'anneaux dans la molécule. Le compte en esclavage rotatif est maintenant un filtre employé couramment suivant la constatation que plus considérablement que dix corrélations rotatives de liens avec la disponibilité biologique orale diminuée de rat[46].

Le volume est défini par la relation : $\mathbf{V=\frac{MM}{d}}$

MM : Masse moléculaire.

d : Densité.

7.3.7 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 :

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

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 pour l'eau)[47].

7.4. Descripteurs de réactivité issus de la DFT conceptuelle

La DFT conceptuelle permet de caractériser les propriétés de réactivité des composés chimiques. Les descripteurs qui sont issus de la DFT conceptuelle représentent un moyen simple de rationaliser le comportement chimique des molécules, sur la base du principe HSAB (Hard and Soft Acids and Bases) de Pearson[48]. Leur fiabilité a d'ailleurs été démontrée via différentes analyses théoriques, dédiées principalement à la réactivité chimique. Les descripteurs calculés à partir de la méthode DFT son l'énergie totale, les énergies des orbitales frontières, le moment dipolaire et l'énergie de répulsion. On distingue plusieurs familles de descripteurs électroniques :

7.4.1. 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 détails dans l'annexe[49].

7.4.2. Le moment dipolaire :

Noté μ, mesuré en debye (D), mesure la polarité nette moléculaire, et décrit la séparation de charge dans une molécule où la densité d'électrons est partagée inégalement entre les atomes. L'existence d'un moment dipolaire dans une molécule a son origine dans la différence d'électronégativité entre les atomes. La densité électronique est plus élevée au voisinage de l'atome le plus électronégatif. Ceci entraîne une dissymétrie dans la répartition des électrons de liaison. Ainsi, plus le moment dipolaire d'une molécule est élevé, plus la dissymétrie dans la molécule est importante[50].

7.4.3. Les énergies des orbitales frontières:

elles jouent un rôle majeur dans de nombreuses réactions chimiques et dans les mécanismes réactionnels. Les énergies de ces orbitales sont des paramètres très populaires dans la chimie quantique et dans les études QSAR/QSPR [51] :

7.4.4. L'énergie de l'orbitale HOMO :

notée E_{HOMO}, mesurée en eV, est le niveau d'énergie le plus élevé dans la molécule qui contient des électrons, il est directement lié au potentiel d'ionisation. Lorsqu'une molécule agit comme une base de Lewis (un doublet d'électrons donneur) dans la formation d'une liaison, les électrons sont alimentés à partir de cette orbite. Elle mesure la nucléophilie d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par des électrophiles[52].

7.4.5. L'énergie de l'orbitale LUMO :

notée E_{LUMO}, mesurée en eV, est le niveau d'énergie le plus bas dans la molécule qui ne contient pas d'électrons, il est directement lié à l'affinité d'électrons.

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. Elle mesure l'électrophilicité d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par les nucléophiles[53].

Selon la théorie orbitale frontalière, l'attaque nucléophile se produit par écoulement d'électrons de l'HOMO du nucléophile dans le LUMO de l'électrophile. Dans les molécules stables, les électrons occupés résident toujours dans des orbitales avec des énergies négatives et les orbitales inoccupées ont des énergies positives. Les énergies de HOMO et LUMO sont liées à la réactivité de la molécule: les molécules avec des électrons aux niveaux HOMO accessibles (près de zéro) ont tendance à être de bons nucléophiles car elles ne demandent pas d'énergie pour passer à l'orbitale la plus haute. De même, les molécules avec des énergies LUMO inférieures ont tendance à être de bonnes électrophiles car elles ne demandent pas beaucoup d'énergie pour placer un électron dans une telle orbitale. Les distributions de charge des orbitales HOMO et LUMO sont similaires pour tous les dérivés pyrazoliques. Les auteurs remarquent que les HOMO de tous les composés sont presque délocalisés sur toute la molécule, en particulier pour les unités donneuses qui sont riches en électrons, tandis que pour l'orbitale LUMO, il existe une grande contribution sur les groupes déficients en électrons[54-55].

8. REDUCTION DU NOMBRE DE VARIABLES

Comme nous l'avons rappelé, lorsqu'un grand nombre de descripteurs différents sont collectés, certains d'entre eux peuvent contenir des informations redondantes, ce qui entraîne un problème de

colinéarité. De plus, les descripteurs calculés n'ont pas nécessairement une influence sur l'activité à modéliser. Aussi, le nombre de descripteurs, c'est-à-dire la dimension du vecteur d'entrée, détermine la dimension du vecteur des paramètres à ajuster. Si cette dimension est trop importante par rapport au nombre d'exemples de la base d'apprentissage, le modèle risque d'être surajusté, et incapable de prédire la grandeur modélisée sur de nouvelles observations[56].

Il est nécessaire donc d'éliminer les descripteurs dont l'influence est inférieure à celle de l'erreur, et de sélectionner uniquement les plus pertinents d'entre eux. De manière générale, pour qu'un descripteur soit retenu, il faut que son retrait entraîne une décroissance significative de performance du modèle. Il faut donc être attentif à ne pas perdre de l'information essentielle[57].

Il est donc nécessaire de réduire la dimensionnalité des variables d'entrées. Plusieurs approches sont possibles pour résoudre ce problème :

− Réduire la dimension de l'espace des entrées ;

− Remplacer les variables corrélées par de nouvelles variables synthétiques, obtenues à partir de leurs combinaisons ;

− Sélectionner les variables les plus pertinentes.

les méthodes de sélection et de réduction des descripteurs les plus fréquemment utilisées peuvent être effectuées en deux étapes[58]:

- Sélection objective
- Sélection subjective

8.1. Sélection objective

La sélection objective consiste à diminuer le nombre des variables en réduisant le nombre de descripteurs sans faire participer la variable dépendante (l'activité biologique). Cette sélection consiste en premier temps à exclure tous les descripteurs ayant un pourcentage élevé de valeurs identiques pour l'ensemble des composés (variance non significative). Cela permet de s'assurer que de tels descripteurs ne sont pas inclus par chance dans le modèle final. De même, lorsque deux descripteurs sont fortement corrélés et leur combinaison possède un coefficient de détermination supérieur au seuil requis ($R^2 \geq 0{,}96$), seul celui présentant la plus grande variance est retenu. Non seulement ces procédures évitent l'introduction, dans le modèle, de descripteurs inappropriés, mais elles rendent la suite de l'analyse moins coûteuse en terme de temps de calcul, puisqu'elles réduisent le nombre de descripteurs restant à traiter. parmi les méthodes utilisées pour cette sélection nous citons[59]:

8.1.1 L'analyse en composantes principales

L'analyse en composantes principales (ou ACP), est une technique d'analyse de données utilisée pour réduire la dimension de l'espace de représentation des données. Contrairement à d'autres

méthodes de sélection, celle-ci porte uniquement sur les variables, indépendamment des grandeurs que l'on cherche à modéliser. Les variables initiales sont remplacées par de nouvelles variables, appelées composantes principales, deux à deux non corrélées, et telles que les projections des données sur ces composantes soient de variance maximale. Elles peuvent être classées par ordre d'importance. Considérons un ensemble de *n* observations, représentées chacune par *p* données. Ces observations forment un nuage de *n* points dans *p*. Le principe de l'ACP est d'obtenir une représentation approchée des variables dans un sous-espace de dimension *k* plus faible, par projection sur des axes bien choisis ; ces axes principaux sont ceux qui maximisent l'inertie du nuage projeté, c'est-à-dire la moyenne pondérée des carrés des distances des points projetés à leur centre de gravité. La maximisation de l'inertie permet de préserver au mieux la répartition des points. Dès lors, les *n* composantes principales peuvent être représentées dans l'espace sous-tendu par ces axes, par une projection orthogonale des *n* vecteurs d'observations sur les *k* axes principaux. Puisque les composantes principales sont des combinaisons linéaires des variables initiales, l'interprétation du rôle de chacune de ces composantes reste possible. Il suffit en effet de déterminer quels descripteurs d'origine leur sont le plus fortement corrélés. Les variables obtenues peuvent ensuite être utilisées en tant que nouvelles variables du modèle. Par exemple, la régression sur composantes principales (ou PCR) est une méthode de modélisation dont la première étape est une analyse en composantes principales, suivie d'une régression linéaire multiple (dont le principe est présenté dans le paragraphe 9.1.)[60-61].

8.2. Sélection subjective

8.2.1 Introduction progressive

Cette méthode consiste à incorporer, une à une, les variables au modèle en sélectionnant à chaque étape la variable dont la corrélation partielle avec la grandeur modélisée est la plus élevée[62].

8.2.2 Elimination progressive

Cette méthode consiste en l'établissement du modèle avec l'ensemble des descripteurs pour ensuite ne garder que ceux qui permettent l'obtention d'un modèle ayant une bonne corrélation[63].

8.2.3 Sélection pas à pas

C'est la combinaison des deux méthodes citées précédemment. Les variables sont incorporées une à une dans le modèle par sélection progressive. Cependant, à chaque étape, on vérifie que les corrélations partielles des variables précédemment introduites sont encore significatives[64].

9. SELECTION DES VARIABLES PERTINENTES

La mise en place de modèles QSAR n'est pas facile. La première difficulté réside dans la différence d'échelles existant entre les données à corréler, la structure étant à une échelle moléculaire alors que les propriétés à prédire sont à une échelle macroscopique. De plus, il tient compte des problèmes d'incertitudes à la fois au niveau des structures moléculaires (liées au niveau de calcul) et des données expérimentales (protocoles de mesures).

Un des problèmes importants réside également dans le traitement de données en grande quantité. Un grand nombre de descripteurs et de molécules est testé à analyser, mais aucune règle stricte n'existe quant au choix des paramètres structuraux les plus importants parmi le jeu complet de ceux disponibles[65].

En fait, de nombreux outils existent et il s'agit de trouver le moyen le plus adapté pour obtenir un modèle fiable à partir des données disponibles. Selon les cas, plusieurs approches sont envisageables, il faut alors choisir celle permettant de caractériser au mieux le système.

Les différentes méthodes présentées dans la suite sont celles employées au cours de l'étude, pour développer des modèles (plus ou moins de paramétres, linéaires ou non linéaires , interprétables ou non), choisir les paramètres les plus pertinents, valider ces modèles (en interne et en externe) et déterminer leurs domaines d'applicabilité[66].

9.1. La régression linéaire multiple

La régression linéaire multiple (RLM) est la plus simple méthode statistique de modélisation et la plus appliquée dans les études de la relation structure-activité. La méthode a été popularisée par *Hanch* en reliant l'activité biologique aux propriétés expérimentales lipophiliques, électroniques et stériques pour des séries de composés[67].

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

$$Y= a_0+a_1X_1+a_2X_2+.....+a_nX_n$$

La régression linéaire est facile à mettre en œuvre, et les coefficients a_n obtenus peuvent être interprétés: ils mesurent l'influence de chacune des variables sur la grandeur étudiée. Cependant, il est souvent nécessaire d'avoir recours à des modèles qui prennent en compte la corrélation non linéaire[68-69].

9.2. La méthode de régression des moindres carrés partiels

La régression des moindres carrés partiels (MCP ou PLS) est également une méthode statistique utilisée pour construire des modèles prédictifs lorsque le nombre de variables est élevé et que celles-ci sont fortement corrélées. Cette méthode utilise à la fois des principes de l'analyse en

composantes principales et de la régression multilinéaire. Elle consiste à remplacer l'espace initial des variables par un espace de plus faible dimension, sous-tendu par un petit nombre de variables appelées « variable latentes », construites de façon itérative. Les variables retenues sont orthogonales (non corrélées), et sont des combinaisons linéaires des variables initiales. Les variables latentes sont obtenues à partir des variables initiales, mais en tenant compte de leur corrélation avec la variable modélisée, contrairement aux variables résultantes de l'analyse en composantes principales. Elles doivent ainsi expliquer le mieux possible la covariance entre les entrées et la sortie. Elles sont alors les nouvelles variables explicatives d'un modèle de régression classique, telles que la régression linéaire multiple[70-71].

9.3. Régression non linéaire multiple

La régression non linéaire multiple (RNLM) est une méthode non linéaire (exponentielle, logarithmique, polynomiale, …) qui permet de déterminer le modèle mathématique qui peut expliquer non-linéairement au mieux la variabilité d'une propriété ou d'une activité Y en fonction des descripteurs moléculaires. Dans l'ensemble de nos travaux nous avons utilisé le modèle polynomial en nous basant sur les descripteurs proposés par le modèle linéaire qui seront élevés à la puissance 2 selon l'équation suivante[72-73] :

$$\mathbf{Y = a_0 + \sum_{i=1}^{n} a_i X_i + b_i X_i^2}$$

Avec :y est la variable dépendante (à expliquer ou à prédire) ; x_i sont les variables indépendantes (explicatives) ; n est le nombre de variables explicatives ; a_0 est la constante de l'équation du modèle ; a_i et b_i sont les coefficients de descripteurs dans l'équation du modèle[74] ;

10. TEST DE LA SIGNIFICATION GLOBALE DES METHODES STATISTIQUES :

10.1. Coefficient de corrélation (R)

C'est le coefficient de corrélation de Bravais Pearson entre ŷ et Y, c'est à dire entre valeurs observées et prédites par le modèle de régression, il est noté R, sa valeur varie entre 0 et 1 [75].

$$R = \sqrt{1 - \frac{\Sigma(Y-\hat{y})}{\Sigma(\hat{y}-\bar{y})}}$$

avec: Yet ŷ sont, respectivement, les valeurs observées et calculées de la variable dépendante

10.2. Coefficient de détermination (R^2)

Le coefficient de détermination R^2 est la mesure du degré de liaison entre Yn et Xj. R^2 est déterminé par les différentes relations suivantes :

$$R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

avec:

TSS (total sum of squares) décrit la variation des valeurs de Y, et la somme de la différence au carré entre chaque valeur de Y et la moyenne de Y.

RSS (residual sum of squares) décrit la variation de Y observée à partir de Y (ajustée) estimée. Elle est dérivée de l'addition cumulative du carré de chaque résidu, où un résiduel est la distance d'un point de données au-dessus ou au-dessous de la ligne ajustée.

ESS (explained sum of squares) décrit la variation dans les valeurs ajustées de Y, et est la somme de la différence au carré entre chaque valeur ajustée de Y et la moyenne de Y.

Y_n la valeur prédite

X_j la valeur observée de la variable indépendante

Un bon ajustement correspondra à un R^2 proche de l'unité[76].

10.3. Test Fischer-Snedecor (F)

Le test Fischer permet de justifier la liaison globale entre Yn et Xj, une version dérivée de cet indicateur peut juger du degré de pertinence des variables du modèle. Il s'agit de vérifier pour chaque variable Xj si, lorsqu'on passe du modèle complet à p prédicateurs au modèle simplifié obtenu en t mesures, on fait l'apport marginal de variable Xj à l'explication de Yn. On peut définir quelques paramètres utilisés dans la régression multilinéaire[77].

Somme des carrés totaux : TSS = $\sum(\mathrm{Yobs} - \bar{Y})^2$

Somme des carrés expliqués : ESS = $\sum(\mathrm{Ycal} - \bar{Y})^2$

Somme des carrés résiduels : RSS = $\sum(\mathrm{Yobs} - \mathrm{Ycal})^2$

Ainsi, TSS = ESS + RSS

F peut s'écrire comme suit :

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

La forme de l'équation (9) représente le nombre de degrés de liberté associé avec chaque paramètre. Le ESS associé avec p degrés de liberté et le RSS associé à n-p-1 degrés de liberté.

Le test de Fischer mesure le rapport entre la variance de la variable dépendante expliquée et non expliquée par le modèle de régression. En d'autres termes le test de Fischer permet de tester l'hypothèse nulle selon laquelle chaque β est significativement différent de zéro, ce qui est signe d'une relation évidente entre la variable expliquée et les variables explicatives[78].

Intuitivement, nous rejetterons l'hypothèse nulle lorsque la somme des carrés expliquée par la régression est grande. En d'autres termes, la région critique de ce test est de la forme (F > seuil). Si

la quantité F observée dépasse le seuil, on rejette l'hypothèse H_0,dans le cas contraire, on conserve H_0.

Pour éviter de raisonner sur F, le programme fournit la p-value associée au F observé. La p-value est le niveau de significativité du test de Fischer-Snedecor, c'est-a-dire la probabilité de dépasser le F observé si l'hypothèse nulle est vraie. On compare la p-value au risque α choisi (par exemple α=0,05).

Si p-value ≤ α, alors on rejette l'hypothèse nulle $\beta_1 = \ldots = \beta p = 0$.

Ces résultats permettent d'interpréter les tables d'analyse de variance complètes fournies par tout logiciel mettant en œuvre la régression linéaire[79].

10.4. Ecart type (s)

L'écart type (s) est un autre paramètre habituellement rapporté; il indique dans quelle mesure la fonction de régression prédit les données observées, ce paramètre est donné par[80]:

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

Dont p est le nombre de variables indépendantes

10.5. $R^2_{ajusté}$.

Si l'on continue à augmenter le nombre de descripteurs dans un modèle pour un fixe nombre d'observations, les valeurs R^2 augmenteront toujours, mais cela conduira à une diminution du degré de liberté et à une faible fiabilité statistique. Ainsi, une valeur élevée de R^2 n'est pas nécessairement une indication d'un bon modèle statistique qui convient bien les données disponibles. Pour mieux refléter la variance expliquée (la fraction de la variance des données expliquée par le modèle), $R^2_{ajusté}$ a été défini de la manière suivante:

$$R_a^2 = \frac{(N-1)*R^2 - P}{N-1-P}$$

Le $R^2_{ajusté}$ quantifie la part du modèle expliquée par les variables explicatives en tenant compte du nombre de variables utilisées, privilégiant les modèles contenant peu de variables. On choisit le modèle dont le $R^2_{ajusté}$ est le plus élevé. Ce critère est beaucoup plus judicieux que le R^2 classique, qui lui privilégiera toujours le modèle contenant toutes les variables[81].

11. CRITERES DE PERFORMANCE ET DE VALIDATION DES MODELES QSAR :

La modélisation vise à fournir un modèle non seulement ajusté aux données expérimentales, mais pouvant également être généralisé à de nouveaux exemples. Pour ce faire, plusieurs méthodes de validation, telles que la validation interne, la validation externe et le test de randomisation, sont employés pour estimer la fiabilité du modèle QSAR et pour déterminer sa pertinence pour une

application donnée. Différents indicateurs statistiques sont également employés pour déterminer la qualité d'un modèle QSAR, nous citerons ci-dessous les plus répandus[82].

11.1. Validation interne

11.1.1 validation croisée

la validation croisée à n paquets est une technique qui consiste à découper aléatoirement un jeu de donnés à n paquets contenant sensiblement le même nombre de molécules. Pour un nombre *N* d'exemples d'apprentissage, on retire à chaque itération un exemple I de l'ensemble d'apprentissage initial. Une série d'apprentissage est réalisée pour les *N-1* molécules restantes et la molécule retirée est prédite par le modèle formé[83].

Un exemple de validation croisée à cinq paquets est représenté ci dessous:

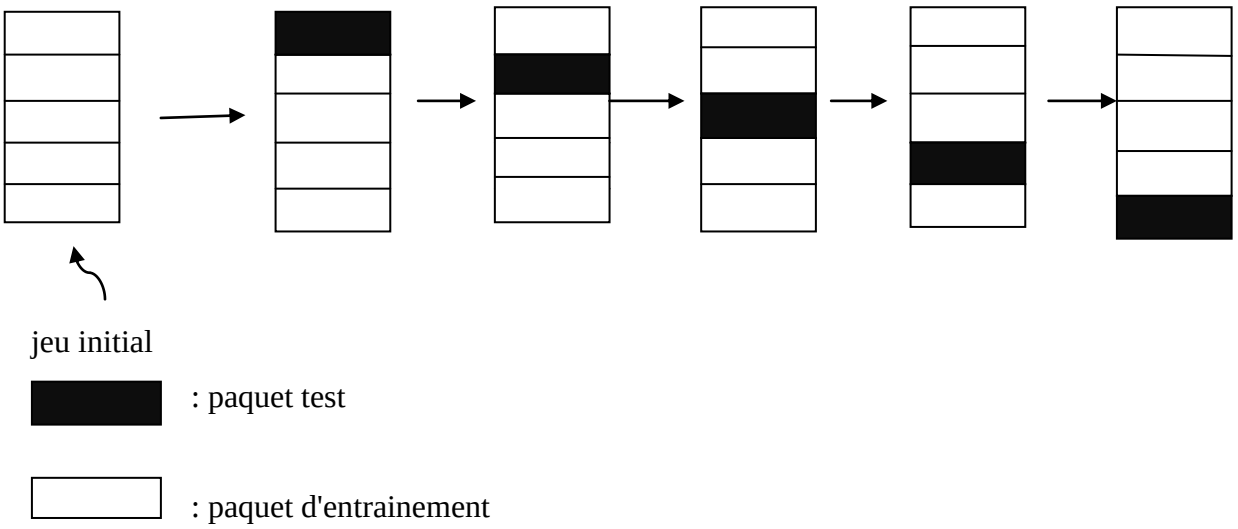


Figure 2: Principe de la validation leave-one-out.

11.1.2 Y-Randomisation

Afin de s'assurer qu'un modèle QSAR est fiable, les tests de Y-randomisation sont une des techniques les plus employées. En effet, il n'est pas rare d'obtenir des corrélations fortuites (ou « chance corrélation »), c'est-à-dire un modèle affichant de bons résultats statistiques (R^2, SD) pour l'apprentissage, mais impliquant des descripteurs qui ne sont pas reliés à la propriété/activité modélisée. Ces modèles aléatoires peuvent être détectés par la procédure Y-randomisation. Elle consiste à mélanger aléatoirement les activités expérimentales pour la série d'apprentissage en utilisant les mêmes descripteurs, de nouveaux modèles sont obtenus. Ces derniers doivent avoir des performances égales ou faibles à celles obtenues pour le premier modèle. Cependant, la validation interne est insuffisante pour étudier le pouvoir prédictif d'un modèle. Pour cette raison la validation externe du modèle est devenue une norme et une partie obligatoire dans la modélisation QSPR/QSAR[84].

11.2. Validation externe

Afin de tester de manière fiable le pouvoir prédictif du modèle QSAR, il est nécessaire d'employer un ensemble de validation externe, non employé pour le développement du modèle. Cette méthode consiste à prédire la propriété/activité d'une série de molécules appelée généralement série de test qui ne sont pas dans la série de développement du modèle, cette validation est caractérisée par les paramètres R^2_{test} , MSE_{test}, MAE_{test}. Une fois l'ensemble de validation mis en place, il suffit alors d'appliquer le modèle QSAR aux molécules qui le composent et de déterminer la corrélation existant entre les activités calculées et expérimentales. Plus cette corrélation est importante, plus le modèle est capable de prédire les activités pour des molécules hors l'ensemble d'apprentissage[85].

11.3. Domaine d'applicabilité :

Le domaine d'applicabilité est la région de l'espace chimique définie par les molécules de l'ensemble d'apprentissage du modèle. Les modèles QSAR ne peuvent pas prédire des propriétés de manière fiable pour l'intégralité des composés chimiques existants. En effet, un modèle QSAR n'est pas destiné à être employé en dehors de son domaine d'applicabilité, c'est-à-dire en dehors de l'espace chimique couvert par son ensemble d'apprentissage[86].

11.4. Règles de cinq de Lipinski

Idéalement, un médicament doit pouvoir être pris oralement sous la forme de petits comprimés. Un composé actif administré oralement doit être capable de résister à l'environnement acide du tube digestif pour être absorbé par l'épithélium intestinal. Donc, ce composé doit être capable de traverser les membranes de la cellule à une vitesse significative.

Drug-like apparaît comme un paradigme prometteur pour coder l'équilibre entre les propriétés moléculaires d'un composé qui influence ses pharmacodynamiques et pharmacocinétiques et optimise finalement, leur absorption, distribution, métabolisme et excrétion (ADME) dans le corps humain comme une drogue. Les conditions empiriques pour satisfaire la règle de Lipinski et manifester une bonne biodisponibilité orale impliquent un équilibre entre la solubilité aqueuse d'un composé et sa capacité à diffuser passivement à travers les différentes barrières biologiques[87].

Les règles de Lipinski prédisent que l'absorption sera probablement forte quand

- Le poids moléculaire ne dépasse pas 500g /mol.
- Le nombre des donneurs des liaisons hydrogène est plus petit que 5.
- Le nombre des accepteurs des liaisons hydrogène est plus petit que 10.
- La lipophilie [évalué par log (P)] ne dépasse pas 5.

Ces paramètres permettent la perméabilité orale d'absorption ou de membrane qui se produit quand la molécule évaluée suit la règle de Lipinski. Les molécules qui violent plusieurs de ces règles peuvent avoir des problèmes avec la biodisponibilité. Par conséquent, cette règle établit certains

paramètres structuraux pertinents pour la prédiction théorique du profil de biodisponibilité orale, et est largement utilisée dans la conception de nouveaux médicaments [88].

12- BIBLIOGRAPHIE SUR LES METHODES QSAR

R. Sabet et al [89] ont effectué en 2010 une étude QSAR des analogues de l'isatine en tant qu'agents anticancéreux *in vitro*. L'étude QSAR des dérivés anticancéreux de l'isatine a été réalisée par la méthode de régression linéaire multiple (RLM) et la méthode de l'algorithme génétique des moindres carrés partiels (GA-PLS), dont la base de ces méthodes étaient les descripteurs de groupes topologiques, chimiques, géométriques et fonctionnels, qui sont avérés être des paramètres efficaces de l'activité cytotoxique. Le nombre des atomes d'halogène et des carbones secondaires totaux a dévoilé des effets positifs , et les effets négatifs du nombre d'amides secondaires. L'activité anticancéreuse influencée par les cétones est en accord avec l'étude QSAR précédente. En comparant les deux méthodes de RLM et GA-PLS, et pour prédire l'activité de nouveaux composés, la RLM représentait des résultats supérieurs avec une haute qualité statistique (R^2 ¼ 0,92 et Q^2 ¼ 0,90).

Deux ans après, la découverte des agents anticancéreux des nouveaux et plus puissants est la base de l'étude de A. Speck-Planche et al [90]. La conception rationnelle de médicaments pour la chimiothérapie anticancéreuse (modèles QSAR multi-cibles pour la découverte *in silico* d'agents anti-cancer colorectaux) est l'un des domaines de recherche les plus actifs en chimiothérapie. Aussi le cancer colorectal (CRC) est le type de cancer le plus étudié en raison de sa forte prévalence et du nombre de décès. Cependant, jusqu'à présent, il n'existe pas de méthodologie capable de prédire l'activité anti-CRC des composés contre plus d'une lignée cellulaire CRC, ce qui constitue l'objectif principal. Pourtant pour tenter de surmonter ce problème, nous développons ici la première approche multi-cible (mt) pour le criblage virtuel et la découverte rationnelle *in silico* d'agents anti-CRC contre dix lignées cellulaires. Deux modèles de classification mt-QSAR ont été construits en utilisant une base de données importante et hétérogène de composés. Le premier modèle était basé sur l'analyse discriminante linéaire employant des descripteurs basés sur des fragments, tandis que le second modèle était obtenu en utilisant des réseaux neuronaux artificiels avec des descripteurs 2D globaux. Les deux modèles ont correctement classé plus de 90% des composés actifs et ceux inactifs dans l'ensemble d'entraînement. Certains fragments ont été extraits des molécules et leurs contributions à l'activité anti-CRC ont été calculées en utilisant le modèle1. Plusieurs fragments ont été identifiés comme étant des caractéristiques sous-structurales potentielles responsables de l'activité anti-CRC et de nouvelles molécules conçues à partir de ces fragments avec des contributions positives et celui ont été prédites par les deux modèles entant qu'agents anti-CRC puissants et polyvalents.

La même année Kunal Roy et al [91] ont achevé des recherches intitulés "Introduction de la métrique rm 2 (rang) intégrant les prédictions d'ordre de classement comme outil supplémentaire de validation des modèles QSAR / QSPR". Cependant des techniques in silico, impliquant le développement de modèles de régression quantitatifs, ont été largement utilisées pour prédire l'activité, la propriété et la toxicité de nouveaux produits chimiques. Par ailleurs l'acceptabilité et l'applicabilité ultérieure des modèles pour les prévisions sont déterminées à partir de plusieurs statistiques de validation internes et externes. Parmi les différentes techniques de validation, les paramètres Q^2 et R^2 représentent les valeurs de la validation interne et la validation externe. De plus, les métriques $r_m^{\ 2}$ introduites par Roy et ses collègues ont été largement utilisées pour assurer l'accord étroit des données de réponse prédites avec celles observées. De ce fait, aucune des méthodes de validation actuellement disponibles et couramment utilisées ne fournit d'informations concernant les prédictions d'ordre de classement pour l'ensemble de test. ainsi La capacité de cette nouvelle technique à effectuer la prédiction de classement est déterminée en fonction de son application pour le jugement de la qualité des prédictions de la régression quantitative. En conséquent les résultats ont surélevé que la mesure r_m^2 présentait la plus petite différence de classement par rapport à celle de la métrique de référence. Ainsi, la corrélation étroite de r_m^2 avec le coefficient de corrélation de rang de Spearman impliquait que la nouvelle mesure pouvait correctement prédire l'ensemble de données de test et pouvait être utilisée comme outil de validation supplémentaire, en plus des mesures conventionnelles , pour évaluer l'acceptabilité et la capacité prédictive d'un modèle QSAR / QSPR.

En 2013 S. CHTITA et al [92] ont combiné les résultats des méthodes DFT et QSAR dans le cadre de l'approche électron-topologique pour la prédiction de l'activité biologique des dérivés de l'imidazo [1,2-a] pyrazine. Cette étude de la relation structure-activité est réalisée contre deux différentes lignées cellulaires cancéreuses (HepG-2: lignée cellulaire de carcinome hépatocellulaire humain et HCF-7: lignée cellulaire d'adénocarcinome mammaire humain). Par la suite l'étude est effectuée sur une série de treize dérivés de l'imidazo [1,2-a] pyrazine en combinant les résultats DFT et QSAR, et en utilisant l'analyse en composantes principales (ACP), l'analyse de régression linéaire multiple (RLM), la régression des moindres carrés partiels (PLS), la régression non linéaire multiple (RNLM) et le réseau de neurones artificiel (NN). Ces méthodes sont basées sur le calcul des descripteurs de type topologiques (poids moléculaire, volume moléculaire, poids moléculaire, réfraction molaire, parachor, densité, indice de réfraction, tension superficielle et polarisabilité) et les descripteursde type électroniques (énergie totale (E), énergie orbitale moléculaire occupée la plus élevée (E_{HOMO}), le plus bas inoccupé énergie orbitale moléculaire, (E_{LUMO}) différence entre l'énergie LUMO et l'énergie HOMO (Gap), le moment dipolaire total des molécules, la dureté absolue, la négativité absolue des électrons et l'indice de réactivité) conçu par les programme ACD /

Chem Sketch et Gaussian 03. Le modèle quantitatif de l'activité des composés est proposé et interprété en s'appuyant sur l'analyse statistique multivariée.

Un an après, M. LARIF et al [93] ont prédit l'activité biologique de la cytotoxicité des dérivés de la chalcone (1,3-diphényl-2-propène-1-one) par rapport aux lignées cellulaires d'adénocarcinome du côlon HT-29 par les modèles DFT-QSAR, la dérivée du chalcone possède deux cycles aromatiques liés par un système carbonyle α, β-insaturé à trois atomes de carbone. Ils sont abondants dans les plantes comestibles et sont considérés comme des précurseurs des flavonoïdes et des isoflavonoïdes. L'objectif principal est d'étudier la relation entre les activités et la structure, une étude QSAR est appliquée à un ensemble de 20 molécules pour les dérivés de prédiction de l'activité biologique. Ainsi l'étude a été réalisée en utilisant la méthode ACP d'analyse en composantes principales; la méthode de régression linéaire multiple RLM et le réseau neuronal artificiel ANN. La procédure de validation croisée leave-one out a été utilisée pour valider le modèle ANN. Par conséquent les descripteurs pertinents obtenus de l'ANN ont montré un coefficient de corrélation de 0,949, ce qui est un bon résultat. À la suite de la relation quantitative structure-activité, des modèles ont été proposé comme des descripteurs majeurs pour décrire ces molécules. Les résultats obtenus suggèrent que la combinaison proposée de plusieurs paramètres calculés pourrait être utile pour prédire l'activité biologique des dérivés de la 1,3-diphényl-2-propène-1-one.

S. CHTITA et al [94] ont proposé en 2015 des études sous le thème "QSPR des dérivés de la 9-aniliioacridine [Ed1] pour leurs propriétés de liaison à l'ADN basées sur la théorie fonctionnelle de la densité à l'aide des méthodes statistiques: modèle, facteurs de validation et d'influence". Le développement et l'optimisation des activités / propriétés biologiques des dérivés de l'acridine, une série de 31 molécules à base de 9-aniliioacridines (25 pour le devellopement du modèle et 6 pour le testé) a été soumise à des analyses quantitatives de QSPR pour leurs propriétés de liaison médicament-ADN en utilisant la régression linéaire multiple (RLM) et la régression non linéaire multiple (RNLM). Des calculs chimiques quantiques utilisant les méthodes de la théorie de la densité fonctionnelle (B3LYP / 6-31G (d) DFT) ont été effectués sur les composés étudiés et utilisés aussi pour calculer les paramètres électroniques et quantiques chimiques. Les modèles ont été usagés pour prédire la constante d'association de la liaison du médicament d'ADN des composés de l'ensemble de test, et l'accord entre les valeurs expérimentales et prédites a été vérifié. Cependant les descripteurs déterminés par les études QSPR ont été le fondement pour l'étude et la conception de nouveaux composés. Ainsi les résultats statistiques indiquent que les valeurs prédites étaient en bon accord avec les résultats expérimentaux (R = 0,935 et R = 0,936 pour RLM et RNLM, respectivement). De ce fait pour valider la puissance prédictive des modèles résultants, les

coefficients de corrélation multiples de validation externe étaient respectivement de 0,932 et 0,939 pour la RLM et la RNLM. Ces résultats montrent que les deux modèles possèdent une stabilité d'estimation favorable et un bon pouvoir de prédiction.

Dans la même année, M. GHAMALI et al [95] ont proposé l'activité inhibitrice de l'aldose réductase des composés flavonoïdes par combinaison des calculs DFT et QSAR. Toutefois, la méthode DFT-B3LYP, avec l'ensemble de base 6-31G (d), est utilisée pour calculer les descripteurs chimiques quantiques des 44 flavonoïdes substitués. En effet les meilleurs descripteurs ont été sélectionnés pour établir la relation quantitative structure activité (QSAR) de l'activité inhibitrice contre l'aldose réductase par analyse en composantes principales (ACP), la régression linéaire multiple (RLM), en régression non linéaire multiple (RNLM) et en réseau de neurones artificiels (ANN). De même les modèles quantitatifs ont été proposés et interprétés en s'appuyant sur l'analyse statistique multivariée. Cette étude montre que le RLM et le RNLM ont servi à prédire l'activité, mais comparé aux résultats donnés par le modèle ANN, les prédictions obtenues par cette dernière sont plus efficaces et bien meilleures que les autres modèles. Ainsi les résultats statistiques indiquent que le modèle ANN est statistiquement significatif et montre une très bonne stabilité vis-à-vis de la variation des données dans la méthode de validation, et la contribution de chaque descripteur à la relation structure-activité est évaluée.

En 2016, HMAMOUCHI et al [96] ont exécuté l'étude de la relation quantitative structure activité basée sur la théorie fonctionnelle de la densité des dérivés du cycloguanil agissant comme Plasmodium falciparum. Ce travail présente un recueil de la relation quantitative structure-activité (QSAR) sur les dérivés du cycloguanil qui sont rapportés comme inhibiteurs de croissance du clone de Plasmodium (T9 / 94 RC17) qui contient l'enzyme mutante A16V + S108T dihydrofolate réductase (DHFR). Un ensemble de 24 cycloguanil dérivés de molécules a été modélisé en utilisant la méthode DFT B3LYP, 6-31G(d) comme fonction de base, par le logiciel Gaussian (03). Les descriptions obtenues sont purement électroniques, et l'ensemble constitue l'activité inhibitrice et les descripteurs électroniques calculés ont été statistiquement traités avec l'analyse en composantes principales (ACP), la régression linéaire multiple (RLM), les régressions non linéaires multiples (RNLM) et le réseau de neurones artificiels (ANN). En conséquence les résultats obtenus par le réseau neuronal artificiel (ANN) montrent que les activités attendues sont en bon accord avec les résultats expérimentaux, avec un coefficient de corrélation égal R = 0, 912.Pour déterminer l'architecture de ce réseau, nous avons varié le nombre de couches cachées , le nombre de neurones dans les couches cachées, les fonctions de transfert et les paires de fonctions de transfert. Les meilleurs résultats ont été acquis avec une architecture de réseau [3-3-1], des fonctions d'activation (Tansig-Purelin) et un algorithme d'apprentissage de Levenberg-Marquardt.

L'étude QSAR sur les inhibiteurs PIM1 et PIM2 utilisant la stratégie rustique pour dépister les analogues de 5- (1H-indol-5-yl) -1,3,4-thiadiazol et prédire leurs activités inhibitrices de PIM, la méthode de la relation structure activité quantitative a été réalisée par A. AOUIDAT et al en 2017 [97] afin d'étudier la série d'inhibiteurs de PIM1 et PIM2. La présente étude a été réalisée sur vingt-cinq 5- (1H-indol-5-yl) -1,3,4-thiadiazols dérivés en tant qu'inhibiteurs de PIM1 et PIM2 ayant pIC_{50} allant de 5,55 à 9 μM et de 4,66 à 8,22 μM, respectivement , en utilisant l'algorithme de la fonction génétique pour la sélection de variables et l'analyse de régression linéaire multiple (RLM) pour établir des modèles QSAR non ambigus et simples basés sur des descripteurs moléculaires topologiques. Les résultats ont montré que les RLM prédisent l'activité de manière satisfaisante pour les deux activités. En effet, le but de la présente étude est double : premièrement, un modèle QSAR linéaire simple a été développé, qui pourrait être facilement manipulé pour filtrer des bases de données chimiques, ou concevoir de nouveaux inhibiteurs puissants PIM1 et PIM2. Deuxièmement, les résultats extraits de l'étude actuelle ont été exploités pour prédire l'activité inhibitrice PIM de certains analogues des composés étudiés. Le but de cette étude est de développer un modèle QSAR simple et interprétable qui pourrait être manipulé pour toute bases de données chimiques, ou pour concevoir de nouveaux inhibiteurs PIM1 et PIM2 dérivés du 5- (1H-indol-5-yl) -1,3,4-thiadiazol.

Il s'agit donc, dans ces travaux de thèse, de développer et d'évaluer des descripteurs de plusieurs catégories pour la prédiction des activités anticancéreuses de composés Hétérocycliques, tels que les sulfonamides, les pyrazoles et les quinoléines, par le biais de la méthodologie de type QSAR reliant les propriétés expérimentales aux structures moléculaires, calculées à l'aide d'outils de chimie quantique, en particulier la Théorie de la Fonctionnelle de la Densité (DFT) basée sur les études citées précédemment.

Références

1 C. Hansch, Quantitative approach to biochemical structure-activity relationships, Acc. Chem. Res, 2, 232-239, **1969**.

2 A. Crum-Brown, T.R. Fraser, On the connection between chemical constitution and physiological action. Part 1, On the physiological action of salts of the ammonium bases, derived from strychnia, brucia, thebia, codeia, morphia and nicotia, Trans. Roy. Soc. Edinburgh, 74, 151-203, **1868**.

3 C. Hansch, A. Leo, R.W. Taft, A Survey of Hammett Substituent Constants and Resonance and Field Parameters, Chem. Rev, *91*, 165-195, **1991**.

4 C. Hansch, T. Fujita, The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients, J. Am. Chem. Soc, 86, 1616-1626, **1964**.

5 S.M. Free, J.W. Wilson, A Mathematical Contribution to Structure-Activity Studies, J. Med. Chem.. 7, 395-399, **1964**.

6 D.R. Lowis, HQSAR: a new, highly predictive QSAR technique. Tripos Tech. Notes**.** 1, 1–15, **1997.**

7 G.E.P. Box and D.R Cox., An analysis of distributions, Journal of the royal statistical society, Series B. 26, 211-243, **1964**.

8 C.D. Selassie, History of quantitative structure-activity relationships. In: Abraham DJ (ed) Burger's medicinal chemistry and drug discovery, vol 1., Drug DiscoveryWiley, New York, 37, 1–48, **2003.**

9 D.R. Lowis, DR HQSAR a new, highly predictive QSAR technique. Tripos Tech Notes, 1,1–10, **1997.**

10 P. A. Cornillon, E.M.atzner-Lober, Regression theorie et Applications , Springer-Verlag France, Paris, **2007**.

11 R. Todeschini, V. Consonni, and R. Mannhold Molecular Descriptors for Chemoinformatics, Drug Discovery & Development, 8, 41-72, **2009**.

12 J.G.Topliss, R.J. Costello, Chance correlation in structure-activity studies using multiple regression analysis. J Med Chem,. 15, 1066–1068, **1972**.

13 K. Hasegawa, K. Funatsu, GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. J. Mol. Struct.**,**425, 255–262, **1998.**

14 S. Ajmani S, K. Jadhav, S.A. Kulkarni, Group-based QSAR (G-QSAR): mitigating interpretation challenges in QSAR. QSAR Comb Sci,. 28, 36–51, **2009.**

15 O. Nicolotti, V. Gillet, P.J. Fleming, D.V. Green, Multiobjective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs. J. Med. Chem**.**. 45, 5069–5080, **2002.**

16 F. Bonachera, ―Les triplets pharmacophoriques flous : développement et applications‖, Thèse de Doctorat, Université Lille 1 Sciences et Téchnologies, France, **2011**.

17 K. Hasegawa, Y. Miyashita, K. Funatsu, Strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists-. J. Chem. Inf. Comput. Sci**.**, 37, 306–310, **1997**.

18 M.P. Freitas, S.D. Brown, J.A.Martins, QSAR: a simple 2D image-based approach for quantitative structure–activity relationship analysis. J Mol Struct, 738, 149–154, **2005.**

19 M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley, New York, **2000**.

20 P.Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb Sci, 26, 694–701, **2007**.

21 A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'Min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, QSAR modeling: where have you been? Where are you going to?, J Med Chem, 57, 4977–5010, **2014.**

22 S.P. Niculescu, A. Atkinson, G. Hammond, M. Lewis, SAR and QSAR in Environmental Research, J Chem Inf Comput Sci, 15, 293–309, **2004.**

23 R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley, Weinheim, **2000**.

24 R. Todeschini, V. Consonn, Molecular descriptors for chemoinformatics, Wiley-VCH, New York, **2009**.

25 SC Basak, BD Gute, GD Grunwald, Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach. J Chem Inf Comput Sci, 37, 651–655, **1997**.

26 G. Wang, Y. Li, X. Liu, and Y. Wang, The QSAR and Combinatorial Science, 28, 1418–1431, **2009.**

27 M.J. Frischl, Gaussian 03, Revision B.01, Gaussian, Inc., Pittsburgh, PA, **2003**.

28 CambridgeSoft Desktop Software –ChemDraw (Windows/Mac). http:// www. cambridgesoft .com/.

29 ACD/Labs.com:: Freeware:: ACD/ChemSketch. http://www.acdlabs.com/ resources/ freeware/ chemsketch/.

30 V.Y. Nalimov, the Application of Mathematical Statistics to Chemical Analysis‖, Addison-Wesley, Reading, MA, **1962**.

31 M.R. Doddareddy, Y.J. Lee, Y.S.Cho, K.I.Choi, H.Y.Koh, Pae AN Hologram quantitative structure activity relationship studies on 5-HT6 antagonists. Bioorg Med Chem, 12, 3815–3824, **2004.**

32 G. Fayet, P. Raybaud, H. Toulhoat, de Bruin, J. Mol. Struct. (Theochem),12,903-100, **2009**.

33 H. Wiener, Structural determination of paraffin boiling points, Journal of Chemical Information and Computer Sciences, 69, 17-20, **1947.**

34 C. Hansch, A. Leo., D. Hoekmann,. Exploring QSAR : hydrophobic, electronic and steric constants. Washington, DC , American Chemical Society, **1995**.

35 M. Randić, On characterization of molecular branching. Journal of the American Chemical Society, 97, 6609-6614, **1975**.

36 A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22, 69-77, **2003**.

37 M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-Chemical Descriptors in QSAR/QSPR Studies, Chem. Rev.. 96, 1027-1044, **1996**.

38 F.Luan, X. Xu, H. Liu, Cordeiro, M.N.D.S. Prediction of the baseline toxicity of non-polar narcotic chemical mixtures by QSAR approach. Chemosphere, 90, 1980–1986, **2013.**

39 M. Srivastava, H.Singh, P.K. Naik, Quantitative structure–activity relationship (QSAR) of artemisinin: the development of predictive in vivo antimalarial activity models. J. Chemometr., 23, 618–635**, 2009.**

40 R. Todeschini , V. Consonni, P. Gramatica, Chemometrics in QSAR. In: Brown S, Tauler R, Walczak R (eds) Comprehensive chemometrics, vol 4. Elsevier, Oxford, pp, 6,129–172, **2009.**

41 K. Roy, R.N.Das, A review on principles, theory and practices of 2D-QSAR. Current Drug Metabol, 15, 346–379, **2014**.

42 M. Stone, P.Jonathan, Statistical thinking and technique for QSAR and related studies. Part I: General theory. J. Chemom.**,**7, 455–475, **1993.**

43 R. Khosrokhavar, J. Ghasemi,F. Shiri, 2D quantitative structure-property relationship study of Mycotoxins by multiple linear regression and support vector machine. Int. J. Mol. Sci., 11, 3052–3068, **2010.**

44 S. Borman, New QSAR techniques eyed for environmental assessments. Chem. Eng. News**,**68, 20–23, **1990.**

45 Q. Du, R.Huang, Y. Wei, Z.Pang, L.Du, K.C. Chou, Fragment-based quantitative structure–activity relationship (FB-QSAR) for fragment-based drug design. J. Comput. Chem.**,**30, 295–304, **2009.**

46 P.P. Roy, Paul, S. Mitra, I. Roy, On two novel parameters for validation of predictive QSAR models, Molecules, 14, 1660–1701, **2009**.

47 R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley, **2000**.

48 F. Neese, A critical evaluation of DFT, including time-dependent DFT, applied to bioinorganic chemistry, Journal of Biological Inorganic Chemistry, 11, 702–711, **2006**.

49 I.I.R. Denning, T. Keith, J. Millam, K. Eppinnett, W.L. Hovell, R. Gilliland, GaussView Version 3.09, Semichem Shawnee Mission, KS, USA, **2003**.

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

51 R. Franke, Theoretical Drug Design Methods‖, Elsevier Amsterdam, 8, 115–123, **1984** .

52 P.W. Atkins and J. de Paula, Atkins' Physical Chemistry‖, 7th ed., Oxford University Press, Oxford, **2002**.

53 D.F.V. Lewis, C. Ioannides, and D.V. Parke, Interaction of a series of nitriles with the alcoholinducible isoform of P450: Computer analysis of structure activity relationships, Xenobiotica, 24, 401–408, **1994**.

54 Z. Zhou and R.G. Parr, Activation hardness: new index for describing the orientation of electrophilic aromatic substitution‖, Journal of the American Chemical Society, 112, 5720–5724, **1990**.

55 O. Kikuchi, Systematic QSAR Procedures with Quantum Chemical Descriptors, Molecular Informatics, 6, 179–184, **1987**.

56 S. Chtita, R. Hmamouchi, M. Larif, M. Ghamali, M. Bouachrine and T. Lakhlifi,, QSPR studies of 9-aniliioacridine derivatives for their DNA drug binding properties based on density functional theory using statistical methods: Model, validation and influencing factors, Journal of Taibah University for Science, 11, 392–407,**2016**.

57 J.G. Topliss and R.P. Edwards, Chance factors in studies of quantitative structure-activity relationships, Journal of Medicinal Chemistry. 22, 1238-1244, **1979**.

58 L. Zhang, H. Zhu, T.I. Oprea, A. Golbraikh, A. Tropsha, QSAR Modeling of the Blood-Brain Barrier Permeability for Diverse Organic Compounds, Pharm. Res. 25, 1902–1914, **2008**.

59 L. He, P.C. Jurs, Assessing the reliability of a QSAR model's predictions, J. Mol. Graph. Model. 23, 503-523, **2005**.

60 A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22, 69-77, **2003**.

61 P.P. Roy, S. Paul, I. Mitra, K. Roy, On Two Novel Parameters for Validation of Predictive QSAR models, Molecules. 14, 1660-1701, **2009**.

62 I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection, J. Chem. Inf. Model. 48, 1733-1746, **2008**.

63 S. Kar, O. Deeb, K.Roy, Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor. Ecotox. Environ. Saf., 82, 85–95, **2012**.

64 C.I. Cappelli, S.Manganelli, A.Lombardo, A.Gissi, E. Benfenati, Validation of quantitative structure-activity relationship models to predict water-solubility of organic compounds. Sci. Total Environ., 46, 781–789, **2013.**

65 D.Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J. Chem. Inf. Comput. Sci.**,** 34, 854–866, **1994.**

66 P. Goodford, Multivariate characterization of molecules for QSAR analysis. J. Chemom.**,** 10, 107–117, **1996.**

67 E. Besalu, J.V. De Julian-Ortiz, L. Pogliani, Trends and plot methods in MLR studies.J. Chem. Inf. Model, 47, 751–760, **2007**.

68 V.Consonni, D.Ballabio, R.Todeschini, Evaluation of model predictive ability by external validation techniques. *J Chemometrics* 24, 194–201, **2010.**

69 A.O. Aptula,N.G. Jeliazkova, T.W.Schultz, M.T. Cronin, The better predictive model: High Q2 for the training set or low root mean square error of prediction for the test set? *QSAR Comb. Sci.*, 24, 385–396, **2005**.

70 S. Wold, M. Sjöström, L. Eriksson PLS-regression: a basic tool of chemometrics, Chemom Intell Lab Syst, 58, 109–130, **2010**.

71 N. Chirico, P. Gramatica, Real External predictivity of QSARmodels: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model, 51,2320–2335, **2011.**

72 S.P. Niculescu, M.A. Lewis, J. Tigner, SAR and QSAR in Environmental Research, 19, 735-750, **2008.**

73 OECD, Guidance Document on the Validation of (Quantitative) Structure–Activity Relationships Models,Organisation for Economic Co-operation and Development **2007**.

74 I. Mitra , A.Saha, K.Roy, Exploring quantitative structure-activity relationship (QSAR) studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. Mol Simult, 36, 1067–1079, **2010**.

75 K.Roy, I.Mitra, S.Kar, P.K.Ojha, R.N.Das, H.Kabir, Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model, 52, 396–408, **2012.**

76 K.Roy, P.Chakraborty, I.Mitra, P.K.Ojha, S.Kar , R.N.Das, Some case studies on application of "rm 2 " metrics for judging quality of QSAR predictions: emphasis on scaling of response data. J Comput Chem, 34, 1071–1082, **2013**.

77 I.T. Jolliffe, ―Principal Component Analysis‖, New-York, NY: Springer, 2ème édition, **2002**.
78 I. Mitra; P.P. Roy, S. Kar, P. Ojha, K. Roy, On further application of rm2 as a metric for validation of QSAR models. J. Chemometr.**,** 24**,** 22–33, **2010**..
79 K. Roy; I. Mitra; S. Kar; P.K. Ojha, R.N. Das, H. Kabir, Comparative studies on some metrics for external validation of QSPR models. J. Chem. Inf. Model, 52, 396–408**, 2012**.
80 P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring rm2 metrics for validation of QSPR models. Chemom. Intell. Lab. Sys., 107, 194–205, **2011**.
81 K. Roy, I. Mitra, P.K. Ojha, S. Kar, R.N. Das, H. Kabir, Introduction of rm2(rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/ QSPR models. Chemom. Intell. Lab. Sys., 118, 200–210, **2012**.
82 R. Guha, R, P.C. Jurs, Determining the validity of a QSAR model-a classification approach. J. Chem. Inf. Model., 45, 65–73, **2005**.
83 a OECD Principles for the Validation of (Q)SARs, http://www.oecd.org/datao ecd/33/37/37849783.pdf (Accessed December 29, 2016), **2016**.
b K. Roy, On some aspects of validation of predictive QSAR models. Expert Opin. Drug Discov., 2, 1567–1577, **2007**.
c D.M. Hawkins, S.C. Basak, D. Mills, Assessing model fit by cross-validation. J. Chem. Inf. Comput. Sci., 43, 579–586, **2003.**
84 N. Chirico, P. Gramatic, Real External predictivity of QSAR models: How to evaluateit? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, J Chem Inf Model, 51, 2320–2335, **2011**.
85 G. Schuurmann, R.U. Ebert, J. Chen, B. Wang, R. Kuhne, External validation and prediction employing the predictive squared correlation coefficient-Test-set activity mean vs training set activity mean. J Chem Inf Model, 48, 2140–2145, **2008**.
86 A. Golbraikh, M. Shen, Z. Xiao, Y.D. Xiao, K.H. Lee, A.J. Tropsha., Rational selection of training and test sets for the development of validated QSAR models. Comput. Aided Mol. Design, 17, 241–253, **2003**.
87 A. Chikhi «Calcule et modelisation des interactions peptide defomylase-substances antibactériennes a l'aide de technique "docking" (arrimage) moléculaire» Doctorat d'état en microbiologie **2007.**
88 C. Hansch, A. Leo, S. B. Mekapati, A. Kurup, Drug-like Properties: Concepts, Structure Design and Methods, Bioorganic & Medicinal Chemistry, 12, 3391-3400, **2004**.
89 S. Razieh, M. Mehrdad, A. Sadeghi a, A. Fassihi, QSAR study of isatin analogues as in vitro anti-cancer agents European Journal of Medicinal Chemistry, 45, 1113–1118, **2010.**

90 A. Speck-Planche , V. Kleandrova , A.Feng Luan, M. Natália, D.S. Cordeiro, Chemoinformatics in anti-cancer chemotherapy: Multi-target QSAR model for the in silico discovery of anti-breast cancer agents European Journal of Pharmaceutical Sciences, 47, 273–279, **2012.**

91 K. Roy, I. Mitra, P. Kumar Ojha, S. Kar, R. Narayan Das, K.Humayun, Introduction of rm 2 (rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models Chemometrics and Intelligent Laboratory Systems, 118, 200–210, **2012**.

92 S. Chtita, M. Ghamali, M. Larif, A. Adad, R.Hmamouchi, M. Bouachrine and T. Lakhlifi, Prediction of biological activity of imidazo[1,2-a]pyrazine derivatives by combining DFT and QSAR results International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization), 204, 2-12, **2013**.

93 M. Larif, S. Chtita, A. Adad, R. Hmamouchi, M. Bouachrine and T. Lakhlifi, Predicting biological activity of chalcone (1,3-diphenyl-2-propen-1-one) derivatives cytotoxicity against HT-29 human colon adenocarcinoma cell linesby by DFT-QSAR models Journal of Computational Methods in Molecular Design, 4, 121-130, **2014**.

94 S. Chtita, R. Hmamouchi, M. Larif, M. Ghamali, M. Bouachrine, T. Lakhlifi, QSPR studies of 9-aniliioacridine derivatives for their DNA drug binding properties based on density functional theory using statistical methods: Model, validation and influencing factors, Journal of Taibah University for Science, 9,165–190,**2015.**

95 M. Ghamali, S. Chtita, R. Hmamouchi, A. Adad, M. Bouachrine, T. Lakhlifi, The inhibitory activity of aldose reductase of flavonoids compounds. Combining DFT and QSAR calculations, , Journal of Taibah University for Science, 11, 292-315,**2015.**

96 R. Hmamouchi, M. Larif, S. Chtita, M. Bouachrine and T. Lakhlifi, Density Functional Theory Based Quantitative Structure-Activity Relationship Study of Cycloguanil Derivatives Acting as Plasmodium falciparum. Mor. J. Chem., 4 , 1061-1075,**2016.**

97 A. Aouidate, A. Ghaleb, M. Ghammali, S. Chtita, M. Choukrad, A. Sbai, M. Bouachrine and T. Lakhlifi, QSAR studies on PIM1 and PIM2 inhibitors using statistical methods: a rustic strategy to screen for 5-(1H-indol-5-yl)-1,3,4-thiadiazol analogues and predict their PIM inhibitory activity Chemistry Central Journal, 11, 269-290,**2017.**

Chapitre II

Etude QSAR de l'activité anticancéreuse des dérivés sulfonamides à l'aide des descripteurs moléculaires

Résumé

Pour prédire l'activité anticancéreuse d'une série de molécules sulfonamides, les analyses quantitatives de la relation structure-activité (QSAR), y compris l'analyse en composantes principales (ACP), la régression linéaire multiple (RLM), et la régression non linéaire multiple (RNLM), ont été appliquées sur un ensemble de 40 dérivés de (E) -N-Aryl-2-éthylène-sulfonamide comme agents anticancéreux. La théorie de la fonctionnelle de la densité (DFT) avec la méthode hybride B3LYP et la base étendue 6-31G, a été utilisée pour déterminer les paramètres structuraux, les propriétés électroniques et l'énergie associée aux molécules étudiées afin de les relier à l'activité étudiée. L'analyse en composantes principales a pour but de minimiser la matrice d'entrée de la méthode de régression linéaire multiple (RLM). La méthode de la régression linéaire multiple (RLM), et celui de la régression non linéaire multiple (RNLM) sont utilisées pour concevoir les relations entre l'activité anticancéreuse des molécules étudiées et leurs structures.

Les validations des modèles RLM et RNLM ont été effectuées en divisant l'ensemble de données en deux séries, la série de test et la série d'apprentissage, les valeurs de la validation externe des coefficients de corrélation de l'ensemble du test sont respectivement 0,81 et 0,91 pour la RLM et la RNLM. La RNLM, compte tenu des descripteurs retenus à partir de la RLM, a montré un coefficient de la corrélation de 0,91. Nous avons conclu que les prévisions obtenues par cette dernière sont plus efficaces et bien meilleures que les autres modèles. Les modèles de prédiction obtenus ont été confirmés par deux méthodes de validation, la validation croisée (leave-one-out) et le test de Y-randomisation. La forte corrélation entre les valeurs des activités expérimentale et prédite, prouve la validité et la fiabilité du modèle QSAR obtenu.

1. INTRODUCTION

De nombreux dérivés sulfonamides ont été signalés pour montrer dans l'industrie pharmaceutique des propriétés antitumorales importantes. Les sulfanilamides sont les premiers médicaments largement et systématiquement utilisés comme agents préventifs et chimiothérapeutiques contre diverses maladies[1,2]. Plus de 30 médicaments contenant cette caractéristique sont utilisés cliniquement, y compris les antagonistes anti-hypertensifs, antibactériens, antiprotozoaires, antifongiques, anti-inflammatoires et non-peptidiques des récepteurs de la vasopressine. Les sulfonamides sont des composés qui ont une structure générale représentée sur la Figure 1[3].

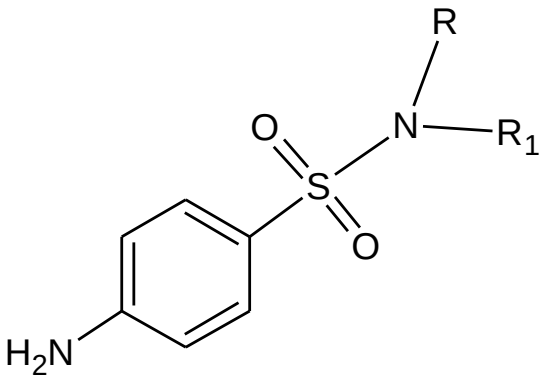


Figure 1 : Structure de base des dérivés sulfanilamides.

Après la découverte de la sulfonamide, des milliers de changements chimiques ont été étudiés et les meilleurs résultats thérapeutiques ont été obtenus à partir de composés dans lesquels l'anneau d'hydrogène (SO_2NH_2) a été remplacé par un hétérocycle. À ce jour, plus de vingt mille dérivés de sulfanilamide ont été synthétisés. Ces synthèses ont abouti à la découverte de nouveaux composés ayant des propriétés pharmacologiques qui varient dans la structure principale (Figure 1) où R et R1 peuvent être l'hydrogène, un alkyle, un aryle ou un hétéroaryle, et ainsi de suite[4-6].

La découverte de médicaments est un processus long et complexe Il faut consacrer plus de 12 années et plus d'un milliard d'euros en moyenne aux activités de recherche et de développement avant qu'un nouveau médicament ne soit disponible pour les patients. Le développement de médicaments est une aventure très risquée[7-9]. Grâce à un système dynamique de découvertes de médicaments, les molécules candidates aux médicaments acquièrent un haut contenu informationnel qui est ensuite traité par des méthodes sophistiquées d'analyses de données qui font partie du curriculum des chimistes informaticiens. Parmi les techniques de chimie-informatique, nous pouvons citer les techniques QSAR qui permettent de trouver une corrélation entre l'activité biologique mesurée pour un panel de composés et certains descripteurs moléculaires[10-12]. Les techniques QSAR sont basées sur le concept postulant que des structures similaires ont des

propriétés similaires, plus les molécules sont différentes, plus il est difficile de corréler les propriétés physico-chimiques et l'activité biologique, alors que le contraire est plus simple[13]. Le résultat de l'intégration de l'information à haut contenu informationnel dans le processus de découverte de médicaments est la constitution d'une capacité technologique pouvant prédire avec un très haut niveau de confiance les propriétés de biodisponibilités optimales et ce en vue de réduire le taux élevé de conception des nouveaux médicaments lors des études en phases cliniques[14].

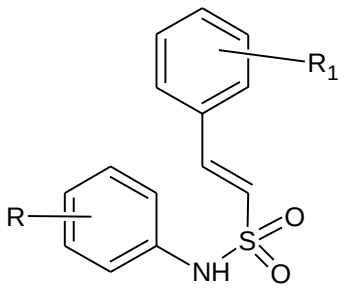
En guise d'application, nous avons proposé d'élaborer un modèle statistique (QSAR) pour coder l'information chimique sous forme d'équation mathématique d'une série de molécules sulfoniques, pour son activité anticancéreuse. Dans cette étude, l'analyse en composantes principales, la régression linéaire multiple (RLM), la Régression Non-Linéaire Multiple (RNLM), l'analyse de validation croisée et le test de Y- Randomisation ont été appliqués à une série sulfonique afin de développer le modèle QSAR pour prédire de nouveaux composés ayant la même activité biologique .

2. METHODOLOGIE

2.1. Base de données :

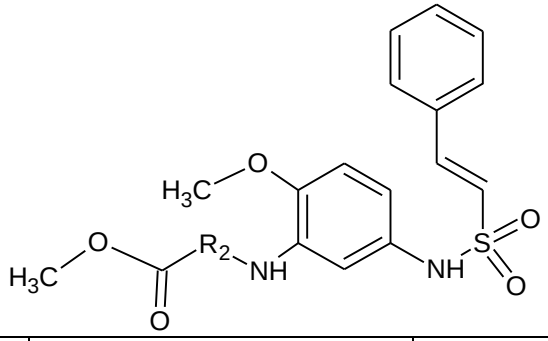
Dans cette étude, nous avons sélectionné 40 dérivés de (E) -N-Aryl-2-éthylène-sulfonamide contenant chacun un halogène, un groupe méthyle, un groupe nitro et d'autres substituants, qui sont rapportés de la base des données par F. Shiri et al [15]. Pour exercer la méthode QSAR de manière opportune, les valeurs déclarées de IC_{50} ont été converties en pIC_{50} en calculant le logarithme négatif ($pIC_{50} = -\log_{10} IC_{50}$) et ensuite les utiliser comme variables dépendantes pour le développement du modèle QSAR. Le tableau 1 montre les substituants des composés étudiés et leurs activités expérimentales pIC_{50}.

Tableau 1 : *Structures et activités expérimentales des composés étudiés.*

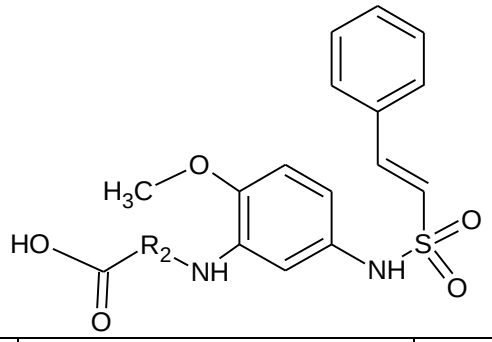


Composés	R	R_1	pIC_{50}
1	H	H	5,00
2	4-Cl	H	4,70

3	4-F	4-Br	5,00
4	4-F	4-OCH_3	5,00
5	4-OCH_3	4-OCH_3	5,30
6	4-OCH_3	2,4-$(OCH_3)_2$	4,82
7	4-OCH_3	2,6-$(OCH_3)_2$	6,43
8	4-OCH_3	2,4,6-$(OCH_3)_3$	6,7
9	4-OCH_3	3,4,5-$(OCH_3)_3$	4,46
10	2,4,6-$(OCH_3)_3$	4-OCH_3	5,12
11	4-OCH_3	2,6-$(OCH_3)_2$,4-OH	5,00
12	4-OCH_3	2,4 ,6-F_3	4,12
13	3-F , 4-OCH_3	2,4,6-$(OCH_3)_3$	7,52
16	3-NH_2, 4-OCH_3	2,4,6-$(OCH_3)_3$	4,12
17	3-NO_2, 4-OCH_3	3,4,5-$(OCH_3)_3$	4,46
18	3-NH_2, 4-OCH_3	3,4,5-$(OCH_3)_3$	4,00
22	3-NO_2, 4-F	2,4,6-$(OCH_3)_3$	5,00
23	3-NH_2, 4-F	2,4,6-$(OCH_3)_3$	5,00
24	3,5-$(NO_2)_2$, 4-OCH_3	2,4,6-$(OCH_3)_3$	5,00
25	3,5-$(NH_2)_2$, 4-OCH_3	2,4,6-$(OCH_3)_3$	5,60
26	3-F , 4-OCH_3	4-OCH_3	5,30
27	3-F , 4-OCH_3	2,3,4,5,6-F_5	5,00
28	3-NO_2, 4-OCH_3	2,3,4,5,6-F_5	4,00
29	3-NH_2, 4-OCH_3	2,3,4,5,6-F_5	4,12
30	2,3,4,5,6-F_5	3-NO_2, 4-OCH_3	4,00
31	2,3,4,5,6-F_5	3-NH_2, 4-OCH_3	4,46
32	2,3,4,5,6-F_5	2,3,4,5,6-F_5	4,46



Composés	**R_2**	***p*IC_{50}**
33	CH_2	6,46
34	$CH(CH_3)$	7,00
35	$C(CH_3)_2$	6,70
36	$CH(C_6H_5)$	5,60
37	$CH(C_6H_4$4-F)	5,60
38	$CH(C_6H_4$4-Br)	5,12



Composés	R_2	pIC_{50}
39	CH_2	6,40
40	$CH(CH_3)$	7,40
41	$C(CH_3)_2$	7,15
42	$CH(C_6H_5)$	7,12
43	$CH(C_6H_4$4-F)	7,12
44	$CH(C_6H_4$4-Cl)	7,12
45	$CH(C_6H_4$4-Br)	6,60

2.2. Calculs des descripteurs :

Avant toute modélisation, il est nécessaire de calculer un certain nombre de descripteurs, en raison que les paramètres qui décrivent l'activité anticancéreuse des sulfonamides sont mal connus. Une part du succès de tout modèle QSAR réside dans le choix des descripteurs moléculaires employés.

En général, les descripteurs classiques utilisés pour une telle analyse sont des descripteurs constitutionnels, topologiques voire géométriques. Or, il est souvent difficile de relier ces paramètres aux phénomènes de réactivité des inhibiteurs avec les cellules cibles.

L'emploi des descripteurs issus de la chimie quantique est moins fréquent en QSAR, alors qu'ils présentent l'avantage d'être reliés directement aux propriétés de réactivité des systèmes moléculaires. Ainsi, il est essentiel de sélectionner parmi les descripteurs calculés ceux qui sont pertinents pour coder l'information chimique sous forme d'équation mathématique. Cette sélection est effectuée en utilisant l'analyse RLM. Les quarante molécules ont été optimisées au moyen de la mécanique quantique utilisant la méthode DFT et la fonction B3LYP associée à l'ensemble de base 6-31G(d) par utilisation du logiciel GAUSSIAN 03[16]. Un certain nombre de descripteurs électroniques ont ensuite été calculés à partir des molécules optimisées, y compris le moment dipôlaire (DM), l'énergie des orbitales frontières (E_{HOMO}, E_{LUMO}), l'énergie totale (E_{total}), et l'energie de répulsion (RE).

Nous avons utilisé aussi le logiciel CHEMBIO OFFICE (2015)[17] pour calculer les paramètres suivants: Poids moléculaire (MW), la lipophilie (logP), les accepteurs de liaison hydrogène (HA) et les donneurs de liaison hydrogène (HD).

Les descripteurs: Volume Molaire (MV (cm^3)), Réfractivité molaire (MR (cm^3)), Parachor (Pc (cm^3)), Densité (g / cm^3), Indice de réfraction, Tension superficielle (Dyne / Cm) et Polarisabilité (cm^3) ont été conçus à l'aide du programme CHEMSCKETCH[18].

Tableau 2 : valeurs des descripteurs utilisées pour l'analyse QSAR des dérivés sulfonamides

Composés	MD	Rep E	MW	HA	HD	LogP	polar	dent	surf ten	Parc	MR	R Ind	MV	E_{HOMO}	E_{LUMO}	E_{TOTAL}
1	6,55	1313,26	259,32	2	1	3,33	28,98	1,3	57,6	547,8	73,1	1,66	198,7	-6,42	-2,24	-31352,99
2	6,68	1517,44	293,76	2	1	3,95	30,89	1,39	58,8	583,7	77,93	1,66	210,7	-6,56	-2,43	-43945,49
3	4,77	1846,46	256,22	3	1	4,29	32,08	1,63	58,2	605,4	80,93	1,66	219,1	-6,68	-2,57	-104516,2
4	7,19	1661,17	307,34	4	1	3,4	31,54	1,35	52,7	611	79,58	1,62	226,9	-6,49	-2,27	-37208,38
5	6,01	1786,57	319,37	4	1	3,14	34,02	1,29	51,5	661,1	85,83	1,61	246,7	-5,99	-2,05	-37626,78
6	7,03	2037,67	349,4	5	1	3,03	36,55	1,29	49,3	717,8	92,19	1,6	270,7	-5,67	-1,89	-40763,80
7	5,72	2250,6	349,4	5	1	3,03	36,55	1,29	49,3	717,8	92,19	1,6	270,7	-5,87	-1,91	-40763,80
8	5,51	2522,44	379,43	6	1	2,94	39,07	1,29	47,6	774,5	98,56	1,58	294,7	-5,66	-2,03	-43900,55
9	4,70	2384,79	379,43	6	1	2,68	39,07	1,29	47,6	774,5	98,56	1,58	294,7	-5,98	-2,05	-43900,28
10	6,69	2452,79	379,43	6	1	2,94	39,07	1,29	47,6	774,5	98,56	1,58	294,7	-5,47	-1,8	-43900,55
11	8,98	2298,19	365,43	6	2	2,63	37,15	1,36	54,9	732,8	93,72	1,61	269,2	-5,40	-2,33	-42824,28
12	9,82	2162,67	343,32	6	1	3,72	31,63	1,46	49,9	625,9	79,8	1,59	235,4	-6,27	-2,86	-42643,44
13	2,65	2109,79	397,42	7	1	3,1	39,11	1,33	46,7	781,6	98,76	1,57	299	-4,77	-1,17	-32198,01
16	9,91	2606,21	394,42	7	2	2,12	40,5	1,33	52,6	800,3	102,18	1,60	297	-4,98	-1,46	-45416,87
17	9,48	2825,23	424,42	7	1	2,12	40,96	1,32	44,9	831,8	103,32	1,56	321,1	-6,16	-3,17	-49501,11
18	7,81	2553,51	394,44	7	2	1,86	40,5	1,33	52,6	800,3	102,18	1,60	297	-5,21	-1,98	-45416,6
22	8,06	2772,74	412,39	7	1	2,9	38,6	1,36	44,6	781,7	97,38	1,56	302,3	-6,13	-3,72	-49082,99
23	11,44	2459,85	382,41	7	2	2,38	38,02	1,38	53,7	750,8	95,92	1,61	277,2	-5,46	-1,74	-44998,47
24	9,03	3522,25	469,42	8	1	2,08	43,05	1,42	50	876,4	108,61	1,57	329,5	-5,99	-3,69	-55102,22
25	12,16	3073,97	409,46	8	3	1,3	41,94	1,37	58	826,2	105,79	1,62	299,3	-5,07	-1,9	-46931,54
26	4,88	1946,17	337,37	5	1	3,29	34,07	1,34	50,2	668,3	85,94	1,6	250,9	-6,14	-2,1	-40339,65
27	4,03	2455,9	379,29	9	1	4,2	31,77	1,6	46,3	647,2	80,14	1,56	248	-6,47	-2,98	-50799,05
28	6,77	2692,35	424,29	9	1	3,48	33,79	1,56	40,3	681,9	85,23	1,54	270,5	-6,37	-3,39	-53681,81
29	6,22	2450,84	394,32	9	2	3,22	33,16	1,6	53,6	665,9	83,64	1,6	246,1	-5,48	-2,78	-49597,29
30	7,01	2727,73	424,29	9	1	3,48	33,79	1,56	40,3	681,9	85,23	1,54	270,5	-6,65	-3,96	-53681,81
31	9,76	2426,24	394,32	9	2	3,22	33,16	1,6	53,6	665,9	83,64	1,6	246,1	-5,66	-2,67	-49597,29
32	4,34	2860,76	439,23	12	1	4,94	29,42	1,82	43,6	619,1	74,22	1,53	240,8	-7,52	-3,45	-58534,89
33	9,27	2333,89	376,43	5	2	2,6	39,19	1,35	57,5	767,7	98,86	1,63	278,7	-5,24	-2,01	-43324,61
34	6,32	2675,05	390,45	5	2	3,12	41,01	1,32	54,9	804,8	103,47	1,62	295,6	-5,17	-2,03	-44401,43
35	5,85	2707,64	404,48	5	2	3,24	42,86	1,3	53,2	842,3	108,13	1,61	311,7	-5,07	-2,05	-45478,52
36	5,83	3193,46	452,53	5	2	4,16	48,98	1,33	57,9	938,5	123,55	1,65	340,1	-5,17	-2,01	-49654,01
37	3,93	3308,33	470,52	6	2	4,32	49,02	1,37	56,8	945,6	123,66	1,64	344,3	-5,28	-2,10	-52372,36
38	6,62	3797,02	531,42	5	2	4,96	52,04	1,49	59,3	989	131,27	1,66	356,3	-5,36	-2,01	-12009,6
39	7,03	2254,82	362,4	5	3	2,29	37,27	1,43	67	724,9	94,02	1,66	253,3	-5,31	-2,07	-42248,06

40	9,09	2392,44	376,43	5	3	2,81	39,1	1,39	63,2	762,1	98,63	1,65	270,2	-5,46	-2,04	-43324,88
41	5,28	2537,25	390,45	5	3	2,93	40,94	1,36	60,7	799,1	103,29	1,64	286,4	-5,17	-2,00	-44401,7
42	6,32	3059,09	438,5	5	3	3,85	47,06	1,39	65,5	895,8	118,71	1,68	314,8	-5,61	-2,1	-48576,91
43	10,00	3197,74	456,49	6	3	4,01	47,1	1,43	64,1	902,9	118,83	1,67	319	-5,44	-2,05	-51295,81
44	10,25	3320,18	472,94	5	3	4,47	48,97	1,45	66	931,6	123,54	1,68	326,7	-5,48	-2,06	-61169,40
45	10,00	3628,51	517,394	5	3	4,64	50,12	1,56	66,7	946,3	126,43	1,69	331	-5,46	-2,06	-119022,1

2.3. Analyse statistique

Nous avons commencé cette étude par la méthode d'analyse en composantes principales dans l'intention de minimiser la matrice d'entrée de la RLM, à l'aide du logiciel XLSTAT. Pour étudier la relation entre une variable dépendante et plusieurs variables indépendantes il faut se référer à la régression linéaire qui est une technique mathématique qui minimise la différence entre les valeurs réelles et celles prédites. Elle sert également à sélectionner les descripteurs utilisés comme paramètres d'entrée dans la régression non linéaire multiple pour prendre en compte la corrélation non linéaire entre l'activité et la structure. La technique de validation croisée est l'une des méthodes les plus efficaces et connues pour la sélection des modèles de régressions, basée sur le critère « leave-one-out ». La procédure « Leave-One-Out » retire successivement une molécule du jeu d'apprentissage contenant *32* molécules. Un modèle QSAR est construit sur un ensemble de 31 composés et la molécule retirée est prédite par le modèle. Cette procédure est répétée 32 fois afin de prédire les propriétés de toutes les molécules. Pour s'assurer qu'un modèle QSAR est fiable, les tests de Y-randomisation ont été déployés pour prouver ce constat. En effet, il n'est pas rare d'obtenir des corrélations fortuites (ou « *chance correlation* »), c'est-à-dire un modèle affichant de bons résultats statistiques (R^2, MAE) pour l'apprentissage, mais impliquant des descripteurs qui dans la réalité ne sont pas reliés à la propriété modélisée. Ces modèles aléatoires peuvent être détectés par la procédure Y-randomisation. Cette méthode consiste à mélanger aléatoirement les propriétés expérimentales pour le jeu d'apprentissage, en utilisant les mêmes descripteurs et à entraîner à nouveau l'algorithme d'apprentissage pour tenter d'obtenir un modèle de validation. Normalement, les modèles obtenus doivent contenir des résultats proches de ceux des résultats de la validation croisée. Ainsi, on peut choisir les modèles qui ont au plus de 1% de chance d'être confondus avec un modèle fortuit.

2.4. Evaluation des modèles

La stabilité et la robustesse du modèle développé sont évaluées à l'aide du coefficient de détermination (R^2), le coefficient de corrélation ajusté $R^2_{ajusté}$, la valeur MES (racine du carré moyen de l'erreur), la valeur des déviations standard SD et des critères de Fisher F. De plus, le choix des descripteurs a été appuyé par un test de Student à un niveau de confiance de 95%. Tous les modèles ont été validés par la validation croisée, selon la procédure de Leave-One-Out(LOO) ainsi pour vérifier que les résultats obtenus par validation croisée ne sont pas dûs à la chance, une procédure de Y-randomisation est mise en jeu. Aussi le modèle a été évalué par une validation externe, cette dernière est effectuée à partir de données qui ne font pas partie du jeu d'entraînement, le pouvoir prédictif est alors caractérisé par le coefficient de corrélation pour le jeu de validation (R^2_{test}).

3. RESULTATS

Afin de bien cibler les paramètres affectant l'activité anticancéreuse et de déterminer la relation quantitative existant entre la structure et l'activité des 40 substitués de (E) -N-Aryl-2-éthylène-sulfonamide, 16 descripteurs on été utilisés comme données d'entrées des analyses statistiques. les méthodes directrices de notre étude sont l'ACP, la RLM, et la RNLM.

3.1. Analyse en composantes principales

La totalité des 16 descripteurs codant les 40 molécules est soumise à une analyse en composantes principales (ACP). 16 composantes principales ont été obtenues.

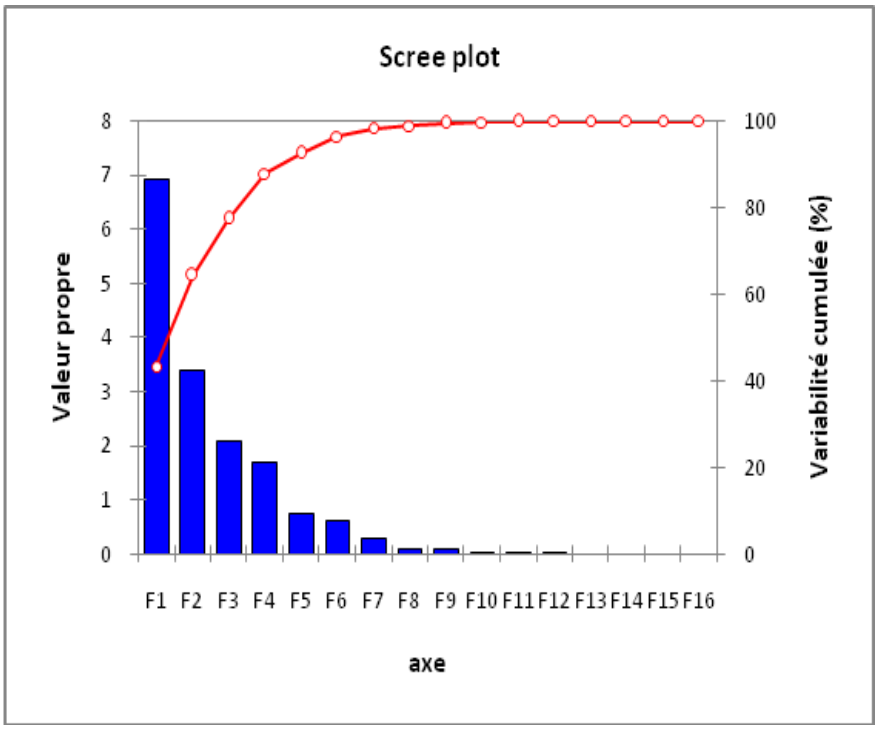


Figure 2 : Les composantes principales et leurs écarts.

Les trois premiers axes principaux sont suffisants pour décrire les informations fournies par la matrice de données. En effet, les pourcentages de variances sont de 43,24%, 21,12% et 13,00% pour les axes F1, F2 et F3 respectivement. L'information totale est estimée à un pourcentage de 77,36%. Le Tableau 3 régénère la matrice de corrélation (Pearson (n)).

Tableau 3 : Matrice de corrélation

Desc	MD	Rep E	MW	HA	HD	LogP	polarb	densité	surf ten	parac	MR	R Ind	MV	E_{HOMO}	E_{LUMO}	E T
MD	**1**															
Rep E	0,259	**1**														
MW	0,199	0,958	**1**													
HA	0,066	0,384	0,44	**1**												
HD	0,435	0,476	0,44	-0,087	**1**											
LogP	-0,375	0,212	0,25	-0,095	-0,026	**1**										
polrb	0,209	0,813	0,78	-0,122	0,609	0,124	**1**									
densit	-0,037	0,228	0,23	0,551	0,026	0,532	-0,241	**1**								
surf tn	0,294	0,196	0,13	-0,542	0,817	0,174	0,492	-0,058	**1**							
parac	0,212	0,856	0,83	-0,002	0,569	0,063	**0,989**	-0,226	0,391	**1**						
MR	0,209	0,813	0,78	-0,122	0,609	0,124	**1,000**	-0,241	0,492	**0,989**	**1**					
R Ind	0,206	0,096	0,02	-0,714	0,676	0,257	0,470	-0,154	0,959	0,347	0,470	**1**				
MV	0,156	0,864	0,85	0,126	0,409	0,012	0,937	-0,235	0,174	**0,974**	0,937	0,136	**1**			
E_{HOMO}	0,222	0,320	0,33	-0,156	0,622	-0,360	0,652	-0,503	0,443	0,647	0,653	0,386	0,587	**1**		
E_{LUMO}	-0,022	-0,132	-0,11	-0,495	0,388	-0,136	0,351	-0,553	0,488	0,285	0,351	0,539	0,182	0,697	**1**	
E T	-0,361	-0,164	-0,06	-0,156	-0,272	0,359	-0,094	0,037	-0,105	-0,104	-0,094	-0,056	-0,084	-0,183	-0,11	**1**

Comme indiqué dans le tableau ci-dessus, les descripteurs ayant des valeurs de corrélation supérieure ou égale à 0,96 sont les suivants :

- La polarité et le parachoc sont parfaitement corrélés (r = 0,989),
- La polarisation et le MR sont parfaitement corrélés (r = 1),
- Le parachoc et le MR sont parfaitement corrélés (r = 0,989).

Les deux variables redondantes, Le parachoc et le MV sont parfaitement corrélés (r = 0,974). Les descripteurs suivants seront supprimés: Parachoc et MR.

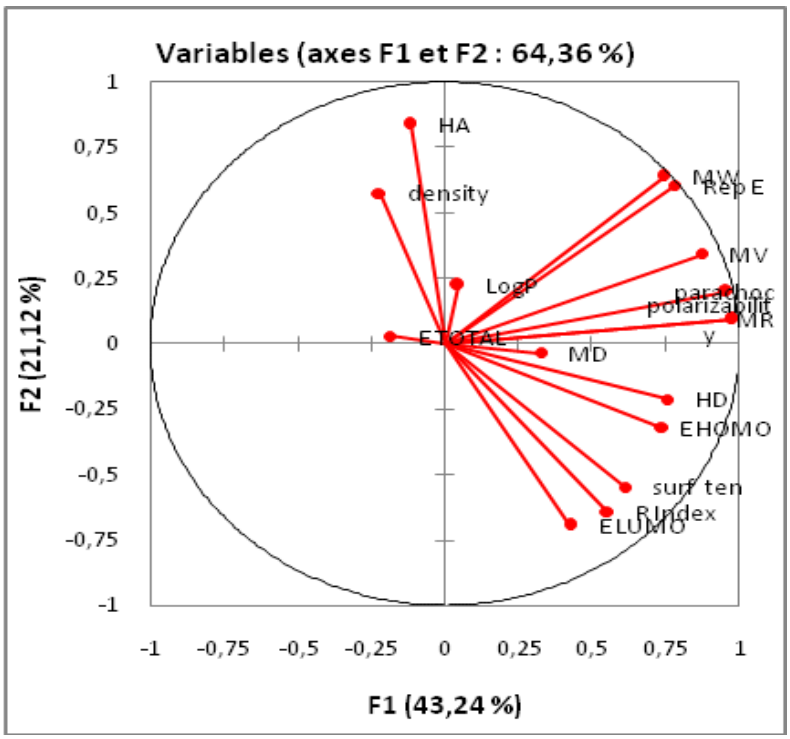


Figure 3 : Cercle de corrélation entre descripteurs

Les coefficients de corrélation de Pearson sont résumés dans le tableau 3. La matrice obtenue fournit des informations sur la corrélation négative ou positive entre les variables. L'analyse des composantes principales (ACP) a été menée pour identifier le lien entre les différentes variables. Les corrélations entre les 16 descripteurs sont présentées dans le tableau3 comme une matrice de corrélation et, dans la figure 3, ces descripteurs sont représentés dans un cercle de corrélation.

3.2. Régression Linéaire Multiple (RLM)

Pour proposer un modèle mathématique et évaluer quantitativement les effets physico-chimiques des substituants sur pIC_{50} pour une série de 40 composés des dérivés des sulfonamide, nous avons soumis la matrice de données contenant 14 variables qui correspondaient aux 40 molécules à une analyse de régression multiple descendante.

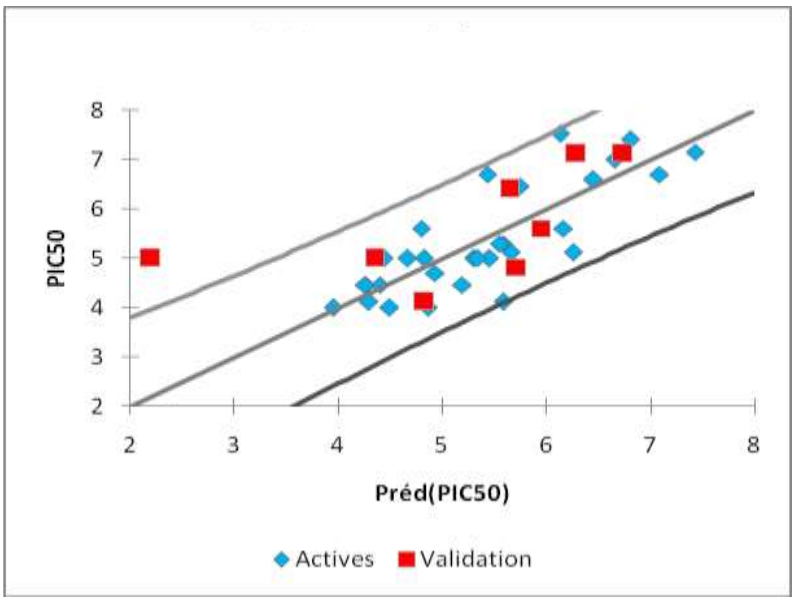


Figure 4 : Représentation graphique de l'activité calculée et observée par RLM.

D'autre part, nous avons utilisé l'étude décroissante de RLM pour éliminer les descripteurs aberrants jusqu'à la validation du modèle (y compris la probabilité critique: p-value <0,05 pour tous les descripteurs et le modèle complet). Cependant, cette méthode utilise les coefficients R, R^2, $R^2_{ajusté}$, MSE, MAE et F-values afin de sélectionner la meilleure performance de régression, où R est le coefficient de corrélation; R^2 est le coefficient de détermination, MSE est l'erreur quadratique moyenne, MAE est l'erreur absolue moyenne et F est la statistique Fisher F. Le modèle QSAR obtenu en utilisant la méthode de régression linéaire multiple (RLM) est représenté par les équations suivantes:

$$\boldsymbol{pIC_{50} = 84{,}233 - 0{,}703\ HA + 2{,}024\ HD + 0{,}857\ LogP - 48{,}758\ R\ Index + 0{,}956\ E_{LUMO}}$$

Les caractères statistiques obtenus par l'équation sont :

N=32 R=0,81 R^2=0,652 $R^2_{Ajusté}$= 0,585 F = 9,742 MSE= 0,418 MAE=0,528 P<0,0001

t_{HA} = -4,119 t_{HD} = 4,793 t_{logP} = 3,916 t_{RI} = -4,053 t_{ELUMO} = 3,273

Comme indiqué dans l'équation ci-dessus, les descripteurs les plus significatifs affectant l'activité anticancéreuse des dérivés sulfoniques étudiés sont des descripteurs électroniques (E_{LUMO}) et des descripteurs stériques (HA, HD, LogP, RIndex). La valeur p est beaucoup plus petite que 0,05, nous prenons un risque inférieur à 0,01% en supposant que l'hypothèse nulle est erronée. Les valeurs du coefficient de corrélation multiple (R) et du coefficient de détermination (R^2) sont supérieures à 0,6, ce qui prouve la capacité du modèle obtenu à prédire l'activité de nouvelles molécules.

La figure 4 montre la répartition régulière des valeurs observées et expérimentales des activités anticancéreuses. Comme indiqué dans l'équation de régression pIC_{50} change avec les valeurs des coefficients des descripteurs: E_{LUMO}, LogP et HA qui sont directement proportionnels au pIC_{50}, tandis que HA et RIndex sont inversement proportionnels au pIC_{50}. La corrélation entre les valeurs expérimentales et les données prédites en utilisant la régression linéaire multiple donnée dans les tableaux 3 et 4 montre que les valeurs prédites par ce modèle sont proches de celles obtenues expérimentalement. Ce résultat a démontré que le modèle développé dans ce travail peut être utilisé avec succès pour prédire de nouveaux composés sulfoniques.

3.3. Régression Non Linéaire Multiple (RNLM)

Pour augmenter la probabilité d'une bonne caractérisation des composés étudiés, le procédé de régression non linéaire est mis en jeu afin de générer des modèles prédictifs non linéaires de la relation quantitative structure-activité (QSAR) entre un ensemble de descripteurs moléculaires obtenus à partir de la RLM. Le modèle QSAR obtenu en utilisant la méthode de RNLM est représenté par l'équation suivante:

$$\mathbf{pIC_{50} = 1,086 - 0,925\ HA - 0,164\ HD + 2,062\ LogP + 52,785\ RIndex + 3,673\ E_{LUMO} + 0,0245\ HA^2 + 0,626\ HD^2 - 0,194\ LogP^2 - 29,65\ R\ Index^2 + 0,526\ E_{LUMO}^2}$$

N=32 R=0,91 R^2=0,828 MSE= 0,398 MAE=0,476

La corrélation entre les activités observées et celles calculées à l'aide de la RNLM est illustrée dans la figure 5. Le coefficient de corrélation (R = 0,91) et l'écart-type (SD = 0,211), obtenus indiquent que les descripteurs sélectionnés par RLM sont performants.

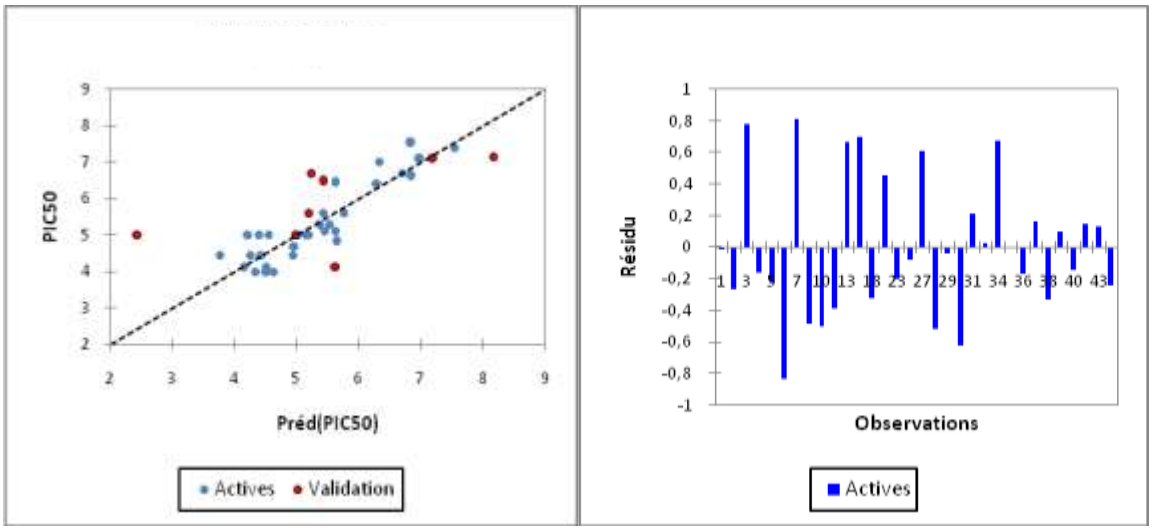


Figure 5 : Activités anticancéreuses prédites par la RNLM en comparaison avec les valeurs expérimentales et les valeurs des résidus.

Tableau 4 : Comparaison entre les activités observées et prédites des modèles statistiquement significatifs obtenus par la série d'apprentissage.

Composés	pIC_{50}	pIC_{50} RLM	pIC_{50} RNLM
1	5,000	4,819	5,016
2	4,700	4,931	4,965
3	5,000	4,435	4,221
4	5,000	5,296	5,156
5	5,300	5,582	5,526
7	6,430	5,696	5,647
8	6,700	5,433	5,620
9	4,460	5,191	4,943
10	5,120	5,658	5,616
11	5,000	5,449	4,503
13	7,520	6,132	6,849
16	4,120	5,580	3,758
17	4,460	4,259	4,323
18	4,000	4,858	4,539
23	5,000	5,332	5,199
25	5,600	4,798	5,373
26	5,300	5,543	4,389
27	5,000	4,667	4,513
28	4,000	4,489	4,160
29	4,120	4,293	4,622
30	4,000	3,948	4,249
31	4,460	4,398	4,435
32	4,460	4,256	6,324
33	6,460	5,744	6,696
34	7,000	6,657	5,766
35	6,700	7,084	5,432
36	5,600	6,161	5,451
38	5,120	6,258	6,301
40	7,400	6,802	7,544
41	7,150	7,429	6,970
42	7,120	6,272	6,984
45	6,600	6,452	6,840

Tableau 5 : Comparaison entre les activités observées et prédites des modèles statistiquement significatifs obtenus par la série de tests.

Composés	pIC_{50}	pIC_{50}RLM test	pIC_{50} RLNM test
6	4,820	5,712	5,226
12	4,120	4,826	4,971
22	5,000	4,350	5,634
24	5,000	2,195	2,436
37	5,600	5,942	5,187
39	6,400	5,643	5,428
43	7,120	6,290	8,165
44	7,120	6,735	7,199

3.4. Validation interne

3.4.1 Validation Croisée

Pour valider nos résultats, nous avons utilisé la procédure "leave-one-out", qui implique l'élimination d'une seule molécule à partir de l'ensemble contenant 32 molécules et faire une prédiction pour la molécule sélectionnée. Cette procédure est répétée 32 fois afin de prédire les propriétés de toutes les molécules.

N=32 R= 0,95 MSE=0,133 MAE=0,231 SD=0,371 P<0,0001

La cohérence et la fiabilité du modèle de RLM est validé à l'aide de la technique de validation croisée. Une bonne corrélation a été obtenue avec un coefficient de corrélation R_{vc} = 0,95 de validation croisée, ce qui indique que le modèle de RLM a une puissance prédictive significative.

3.4.2. Y-Randomisation

La randomisation est largement utilisée dans les études QSAR afin de garantir la robustesse des modèles obtenus. Cette méthode est utilisée après que le modèle de régression soit sélectionné pour s'assurer qu'il n'y a pas de corrélations possibles. Y-randomisation valide le modèle QSAR en comparant la performance du modèle original à celui des modèles construits pour les réponses permutées (randomisées) et basées sur les descripteurs d'origine et la procédure d'origine utilisée pour construire le modèle. Si le coefficient de corrélation des modèles construits pour les réponses permutées est inférieur de celui obtenu en appliquant le modèle de validation croisée, ce résultat indique qu'il existe une indépendance entre les molécules, car les points de mesure du point cible les plus proches n'obscurcissent pas les autres données expérimentales et ne sont pas presque exclusivement impliqués dans l'estimation. Les données utilisées dans cette étude sont réparties uniformément dans l'espace. Par conséquent, le modèle de production peut être extrapolé à l'ensemble de la série d'apprentissage.

N=32 R= 0,65 MSE=0,1043 MAE = 0,214 SD= 0,305 P<0,0001

3.5. Règle Cinq de Lipinski

Les résultats de calcul (Tableau 6) prouvent que tous les composés sont soumis aux règles de Lipinski, suggérant que ces composés théoriquement prédits n'ont pas de problèmes avec la disponibilité biologique orale sauf pour les molécules 38 et 45 ayant le poids moléculaire qui dépasse 500g/mol et la molécule 32 ayant des accepteurs de liaison hydrogène qui dépasse 10.

Tableau 6 : Violation des règles de Lipinski

Composés	MW	HA	HD	Log(P)	Nbr. de violation
1	259,32	2	1	3,332	0
2	293,76	2	1	3,954	0
3	256,22	3	1	4,289	0
4	307,34	4	1	3,396	0
5	319,37	4	1	3,137	0
6	349,4	5	1	3,03	0
7	349,4	5	1	3,03	0
8	379,43	6	1	2,94	0
9	379,43	6	1	2,68	0
10	379,43	6	1	2,94	0
11	365,43	6	2	2,63	0
12	343,32	6	1	3,72	0
13	397,42	7	1	3,1	0
16	394,42	7	2	2,12	0
17	424,42	7	1	2,12	0
18	394,44	7	2	1,86	0
22	412,39	7	1	2,9	0
23	382,41	7	2	2,38	0
24	469,42	8	1	2,08	0
25	409,46	8	3	1,3	0
26	337,37	5	1	3,9	0
27	379,29	9	1	4,2	0
28	424,29	9	1	3,48	0
29	394,32	9	2	3,22	0
30	424,29	9	1	3,48	0
31	394,32	9	2	3,22	0
32	439,23	**12**	1	4,94	1
33	376,43	5	2	2,6	0
34	390,45	5	2	3,12	0
35	404,48	5	2	3,24	0
36	452,53	5	2	4,16	0
37	470,52	6	2	4,32	0
38	**531,42**	5	2	4,96	1
39	362,4	5	3	2,29	0
40	376,43	5	3	2,81	0
41	390,45	5	3	2,93	0
42	438,5	5	3	3,85	0
43	456,49	6	3	4,01	0
44	472,94	5	3	4,47	0
45	**517,394**	5	3	4,64	1

4. DISCUSSION STATISTIQUE ET MECANISTIQUE DU MODELE QSAR OBTENU

La technique ACP est utilisée pour obtenir une vue d'ensemble des composés examinés pour les similitudes et les dissimilarités, afin d'éliminer les descripteurs indépendants qui sont fortement corrélés en examinant la multicollinéarité entre les descripteurs. L'ACP est une méthode factorielle de réduction de dimension pour l'exploration statistique de données quantitatives complexes. Par conséquent, le parachor et la réfraction molaire seront supprimés dans la suite des études statistiques. La présence de la multicolinéarité entre les descripteurs a été confirmée à partir de la matrice de corrélation et les valeurs de VIF.

Dans le cadre de l'étude statistique, nous présentons dans un premier temps la construction d'un modèle QSAR linéaire (RLM), décrivant la relation quantitative structure-activité anticancéreuse des dérivés sulfonamides pour les 40 molécules dérivées du (E) -N-Aryl-2-éthylène-sulfonamide, et par utilisation de la régression non linéaire, en tant que méthode d'apprentissage, le modèle non linéaire est élaboré. Dans un second temps nous procéderons à une comparaison entre les résultats obtenus par le modèle linéaire utilisant la RLM et celui obtenu par le modèle non linéaire utilisant la RNLM. La régression non linéaire employée dans cette étude a été générée en utilisant les cinq descripteurs apparus dans le modèle RLM. En effet, nous avons divisé de façon aléatoire, dans une première étape, l'ensemble total de molécules en deux sous ensembles : un sous ensemble d'apprentissage et un sous ensemble de test. Le sous ensemble d'apprentissage, composé de 32 molécules, a été utilisé pour la génération du modèle QSAR, tandis que le sous ensemble de validation, composé de 8 molécules, a été utilisé pour la validation externe du modèle. Par la suite, nos résultats affirment que la régression non linéaire multiple est le meilleur fondement sur lequel la relation quantitative structure-activité est construite, de même l'activité anticancéreuse dépend des paramètres non linéaireS d'où le modèle proposé dans cette étude a un pouvoir prédictif élevé (R_{RNLM} = 0,91). Les modèles QSAR élaborés (Equation de la RLM et équation de la RNLM) révèlent que l'activité anticancéreuse des dérivés sulfonamides pourrait s'expliquer par un certain nombre de facteurs électroniques et stériques. L'électrophilie, explique la valeur de l'énergie LUMO, est importante dans la description de l'interaction électronique et la réactivité des sulfonamides, tandis que l'hydrophobicité, telle qu'elle est exprimée par log P est importante pour décrire le transport vers le site d'action. Les résultats obtenus par le calcul de log P des dérivés sulfonamides, montrent que tous les composés présentent un log P> 0, ces composés ont le caractère hydrophobe. Les valeurs de log P des composés (8-9-10-11-16-17-18-22-23-24-25-33-39-40-41) sont dans le domaine des valeurs optimales (0< log P <3) donc on peut dire que ces composés ont une activité biologique optimale généralement citée dans la perméabilité et la solubilité.

En effet, il s'avère que la combinaison des trois paramètres E_{LUMO}, HD, et logP augmentent considérablement le pouvoir prédictif du modèle QSAR donné par l'equation de la RLM (N=32, R=0,81, R^2=0,652, $R^2_{Ajusté}$= 0,585, F = 9,742, MSE= 0,418, MAE=0,528, P<0,0001). Le modèle QSAR obtenu peut expliquer environ 81% de la variance expérimentale de la variable dépendante (IC_{50}), en plus il présente un F élevé de Fischer (F = 9,742) et une faible déviation standard (SD = 0,418) ce qui confirme que le modèle RLM explique l'activité anticancéreuse (variable dépendante) d'une manière statistiquement significative satisfaisante. La distribution des résidus autour de la ligne zéro montre qu'aucune erreur systématique n'existe dans les modèles construits de RLM et RNLM.

Selon les valeurs du test Student t (| t |), l'importance des descripteurs impliqués dans ce modèle est dans l'ordre suivant:

HD> HA >R_{index}> LogP>E_{LUMO}. Le descripteur le plus important selon le t- test est le donneur de liaison hydrogène. Le second descripteur est l'accepteur de la liaison hydrogène et le dernier est l'énergie LUMO. E_{LUMO} est directement liée à l'affinité électronique d'une molécule et elle caractérise la sensibilité de la molécule d'être attaquée par les nucléophiles.

Pour montrer que les résultats du modèle de validation croisée ne sont pas obtenus par hasard, un test de randomisation a été effectué. Les valeurs de pIC_{50} de l'ensemble d'apprentissage a été mélangées de façon aléatoire en gardant les paramètres retenus de la régression linéaire multiple inchangés (cette opération est répétée trois fois). les résultats montrent que les valeurs des coefficients de détermination de l'ensemble d'apprentissage calculées par le modèle généré sont les mêmes par rapport à celles de notre modèle. Ceci confirme que les modèles de la RLM et de la RNLM n'ont pas été obtenus par hasard.

Parmi les applications de notre étude le domaine de la pharmacocinétique et la pharmacodynamique est également important pour l'efficacité d'un médicament candidat, ces propriétés doivent toutes deux être optimisées pour aboutir à un médicament d'intérêt médical. D'après les principes empiriques suivants, énoncés par Christopher Lipinski et regroupés sous le nom de «règle des cinq», cette règle est la plus utilisée pour l'identification des composés « drug-like », une substance sera mieux absorbée ou pénétrée si :

1. Sa masse moléculaire est inférieure ou égale à 500 Da.

2. Elle possède moins ou 5 donneurs de liaisons hydrogène (somme de OH et NH).

3. Elle possède moins ou 10 accepteurs de liaisons hydrogène (somme de O et N).

4. Sa valeur de log P est inférieure ou égale à 5.

Les Molécules qui violent plusieurs de ces règles peuvent avoir des problèmes avec la biodisponibilité. Par conséquent, cette règle établit certains paramètres structurels pertinents pour la prédiction théorique du profil de biodisponibilité orale, et largement utilisée dans la conception de nouveaux médicaments.

Les résultats de notre étude ont montré que tous les composés satisfont aux règles de Lipinski, donc ces composés théoriquement n'ont pas de problème avec la disponibilité biologique orale sauf pour les composés 38 et 45 ayant leur poids moléculaire qui dépasse 500 et le composé 32 qui possède des accepteurs de liaison hydrogène qui dépasse 10.

5. CONCLUSION

Au cours de cette première application les études QSAR ont été établies en utilisant les modèles de régression linéaire multiple (RLM), et la régression non linéaire multiple (RNLM) pour 40

molécules provenant des sulfonamides pour leur activité anticancéreuse. La présente étude montre que les descripteurs de la chimie quantique, à savoir, l'énergie LUMO en combinaison avec l'indice d'hydrophobicité log P, indice de réfraction, donneurs et accepteurs de liaison hydrogène sont utiles pour la prédiction de l'activité anticancéreuse des dérivés sulfonamides. L'évaluation de la qualité des modèles RLM et RNLM a révélé que la capacité prédictive de la RNLM était considérablement meilleure que celle des autres méthodes. Le pouvoir prédictif du modèle obtenu a été confirmé par la méthode de la validation croisée. Le meilleur modèle QSAR obtenu est capable de décrire environ 81% de la variance de l'activité anticancéreuse expérimentale et pourrait être utilisé efficacement pour estimer l'activité anticancéreuse des dérivés sulfonamides pour lesquels les données expérimentales sont indisponibles. Une forte corrélation a été observée entre les valeurs expérimentales et prédites de l'activité biologique, ce qui a révélé la validité, la forte robustesse et la qualité du modèle QSAR développé dans ce travail.

Enfin, cette étude représente une tentative d'élaborer des modèles QSAR pour la prédiction de l'activité anticancéreuse des dérivés d'une autre molécule à savoir les pyrazoles en tenant compte du mécanisme d'action et en utilisant un nombre réduit de descripteurs simples et pertinents.

Références

1 E. Esteve ,D. Bazin, C.Jouanneau, S. Rouziere, A. Bataille, A. Kellum, K. Provost , C.Mocuta, S. Reguer, D. Thiaudiere, K. Jorissen, A. Hertig, E. Rondeau, E.Letavernier, M. Daudon, P.Ronco, "How to assess the role of Pt and Zn in the nephrotoxicity of Pt anti-cancer drugs? An investigation combiningm XRF and statistical analysis: Part I: On mice," Comptes Rendus Chimie, 19, 1580–1585, **2016**.

2 K. Morgans, C. Bommel, C. Stowell , L. Abrahm, Ethan Basch, Justin E. Bekelman, Donna L. Berry, Alberto Bossi, Ian D. Davis,Theo M. de Reijke, Louis J. Denis, Sue M. Evans, Neil E. Fleshner, Daniel J. George,Jim Kiefert, W. Lin , G. Matthe, Ray McDermott, H. Payne, G. Roos, D.Schrag, T. Steuber, B.Tombal, J. van Basten, M. van der Hoeven,F. Penson, "Development of a Standardized Set of Patient-centered Outcomes for Advanced Prostate Cancer: An International Effort for a Unified Approach," European urology, 68, 891-898, **2015.**

3 R.Siegel, D. Naishadham, A. Jemal, "Cancer Statistics," CA: a cancer journal for clinicians, 63, 11-30, **2013.**

4 Rebecca L. Siegel, Kimberly D. Miller, "Cancer Statistics," CA: a cancer journal for clinicians, 65, 5-29, **2015.**

5 J.Price, C.Travis,N.Appleby,D.Albanes, B.Gurrea ,T.Bjørge,H.Bueno-de Mesquita, J.Donova, R.Gislefoss, G.Goodman,M.Gunter, C.Hamdy , " Circulating Folate and VitaminB12 and Risk of Prostate Cancer: A Collaborative Analysis of Individual Participant Data from Six Cohorts Including 6875 Cases and 8104 Controls," European urology, 29, 196-215,**2016**.

6 A.Kołaczek, I. Fusiarz, J. Ławecka, D. Branowska , " Biological activity and synthesis of sulfonamide derivatives: a brief review," CHEMIK, 68, 620–628, **2014**.

7 C. T. Supuran, A. Casini, A. Scozzafava, "Protease inhibitors of the sulfonamide type: anticancer, antiinflammatory, and antiviral agents," Medicinal research reviews, 23, 535-558, **2003.**

8 A. Scozzafava, T. Owa, A. Mastrolorenzo, C. T. Supuran , "Anticancer and antiviral sulfonamides," Current medicinal chemistry, 10, 925-953, **2003.**

9 M. Nassir, A.Yatimah Alias , A. Zanariah, M.S. Raied, M.T. Ekhlass, M. Taha, A.H. Aidil , "Synthesis and antibacterial evaluation of some novel imidazole and benzimidazole sulfonamides," Molecules, 18, 11978-11995, **2013**.

10 N. Özbek, H. Katırcıoğlu, N. Karacan, T. Baykal, Bioorg, "Synthesis, characterization and antimicrobial activity of new aliphatic sulfonamide," Bioorganic and medicinal chemistry,15, 5105-5109, **2007**.

11 L. De Luca, G. Giacomelli, An easy microwave-assisted synthesis of sulfonamides directly from sulfonic acids, J. organic chemistry, 73, 3967-3969, **2008**.

12 S. S.Stokes, R.Albert, E. T.Buurman, B. Andrews, A. B. Shapiro, O. M. Green, A. R.McKenzie, L. R. Otterbein, Inhibitors of the acetyltransferase domain of N-acetylglucosamine-1-phosphate-uridylyltransferase/glucosamine- 1-phosphate acetyltransferase (GlmU). Part 2: Optimization of physical properties leading to antibacterial aryl sulfonamides, Bioorganic and medicinal chemistry letters, 22, 7019-7023, **2012**.

13 J. C. Dearden, The History and Development of Quantitative Structure-Activity Relationships (QSARs) ,International Journal of Quantitative Structure-Property Relationships (IJQSPR), 1, 1-44, **2016**

14 N. Adhikari, M. K. Maiti, T. Jha, "Exploring structural requirements of 1-N-substituted thiocarbamoyl-3-phenyl-2-pyrazolines as antiamoebic agents using comparative QSAR modelling," Bioorganic and Medicinal Chemistry Letters, 20, 4021-4026, **2010**

15 F. Shiri, S.M. Bakhshayesh, J.B. Ghasemi, "Computer-aided molecular design of (E)-N-Aryl-2-ethene-sulfonamide analogues as microtubule targeted agents in prostate cancer, "Arabian Journal of Chemistry , 11, 306-331,2015.

16 M. Parac and S. Grimme, "All calculations were done by GAUSSIAN 03 W software,"The Journal of Physical Chemistry A, 106, 6844–6850, **2003**.

17 a ACD/Labs Extension for ChemDraw Version 8.0, Advanced Chemistry Development, Inc., Toronto, ON, Canada, **2015**, http://www.acdlabs.com.

b ChemBioOffice, PerkinElmer Informatics, **2010**, http://www.cambridgesoft.com.

c ACDLABS 10, Advanced Chemistry Development, Inc.,Toronto, Canada, **2015**, http://www.acdlabs.com/.

18 MarvinSketch 5.11.4, Chem Axon, **2012**, http://www.chemaxon.com.

19 XLSTAT **2015** software (XLSTAT Company), http://www.xlstat.com.

Chapitre III

Etude QSAR de l'activité anticancéreuse des pyrazoles à l'aide des descripteurs moléculaires

Résumé

Les dérivés pyrazoliques sont des agents anticancéreux puissants, souvent utilisés en raison de leur implication dans la signalisation des cellules immunitaires pour traiter les tumeurs malignes, pour inhiber la prolifération des cellules cancéreuses et la croissance tumorale ainsi que les maladies auto-immunes. Dans cette étude, une série de 32 molécules à base de anthra [1,9-cd] pyrazol-6 (2H) -one (24 série d'apprentissage et 8 série de test) a été soumise à l'étude quantitative structure-activité (QSAR) : la régression linéaire multiple (RLM), la méthode des moindres carrés partiels (PLS), et la régression non linéaire multiple (RNLM). Les modèles QSAR obtenus sont validés par la méthode de la validation croisée et par la méthode Y-randomisation. Afin de déterminer les paramètres structurels, les propriétés électroniques et l'énergie associée aux molécules, la modélisation moléculaire utilisant la théorie de la fonctionnelle de la densité (B3LYP / 6-31G DFT), est mise en jeu. Compte tenu des descripteurs obtenus à partir de la RLM, la RNLM a montré un coefficient de corrélation proche de 0,91 donc un bon pouvoir prédictif. Les résultats montrent que les modèles sont statistiquement significatifs et présentent une bonne stabilité vis-à-vis de la variation des données de la méthode de validation. Le modèle QSAR obtenu pourrait être appliqué pour l'estimation par prédiction des activités de 15 nouveaux composés ayant le même motif de base (pyrazole) pour lesquels les mesures expérimentales ne sont pas disponibles.

1.INTRODUCTION

Le noyau pyrazole est un isomère structural de l'imidazole, le nom pyrazole provient du noyau pyrrole auquel on a ajouté un atome d'azote : « azole ». Les deux atomes d'azote ayant des propriétés différentes : l'un se comportant comme celui de la pyridine peut subir une protonation en milieu acide; l'autre possède la propriété de l'azote du pyrrole, le doublet participant à l'aromaticité du cycle. La nomenclature officielle du motif pyrazole est le 1,2-diazole. Le pyrazole est un hétérocycle aromatique plan π-excédentaire. Les réactions de substitutions électrophiles se font préférentiellement en position 4 et les attaques nucléophiles en positions 3 et 5. Durant ces dernières années, la synthèse de composés hétérocycliques fusionnés a fait l'objet de recherches scientifiques intenses [1-5].

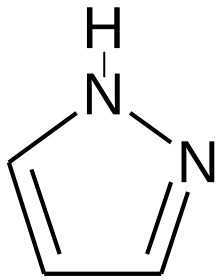


Figure 1: Structure générale des dérivés pyrazoles.

Les pyrazoles sont des composés aromatiques très stables, cette stabilité est due à l'association de dipôles résultant de la séparation permanente très marquée des charges dans le noyau, ce qui confère son caractère amphotère (ou donneur-accepteur). Pour le pyrazole non substitué en position 1, on a la formation de ponts « hydrogène » intermoléculaires qui est très soluble dans l'eau. Cette grande hydrosolubilité est le résultat de la formation de liaison N-H---OH_2 avec les molécules d'eau. A l'état solide, ces composés sont associés sous forme de chaines très structurées qui composent un système fibreux dans les cristaux. En biologie, au pH physiologique, le pyrazole fonctionne à la fois comme accepteur et donneur de proton au site actif de toute une variété d'enzymes. Les groupes aromatiques et hétéroaromatiques variés de pyrazoles ont de nombreuses activités biologiques, ce qui les rend particulièrement intéressants[6-8].

Parmi les diverses applications des dérivés pyrazoles, de très grands progrès ont été accomplis notamment dans le domaine médicinal et thérapeutique. La structure particulière du pyrazole confère à ce composé et à ses dérivés non seulement une grande stabilité thermique, mais aussi des propriétés physico-chimiques uniques, et un grand nombre de molécules contenant cette entité se sont montrés d'une efficacité avérée dans le traitement de divers types de pathologies[9-10]. Cette particularité architecturale du noyau pyrazole confère à ce dernier, et par la même à ses dérivés, le pouvoir de former des liaisons avec toute une variété d'enzymes et de récepteurs dans les systèmes biologiques par l'intermédiaire de liaisons de type hydrogène, Van der Waals, coordination, ion-

dipôle, etc. induisant de ce fait une large gamme d'activités biologiques. Le noyau pyrazole est également présent dans les produits naturels et synthétiques bioactifs, et joue un rôle essentiel dans le métabolisme humain[11-13]. De nombreuses molécules biologiques telles la purine, l'histidine, la vitamine B12, l'acide désoxyribonucléique (ADN), et autres protéines et hormones associées telles que l'hémoglobine par exemple, montrent que le noyau pyrazole est essentiel dans l'action physiologique de plusieurs activités biologiques importantes. Ces propriétés physiologiques spécifiques, et le rôle important dans le processus vital font que l'incorporation du noyau pyrazole est d'une importance stratégique dans la synthèse de médicaments à spectre pharmacologique aussi large que diversifié[14-15].

Les techniques QSAR sont utilisées pour chercher la corrélation entre l'activité biologique mesurée pour un panel de composés des dérivés pyrazoles et certains descripteurs moléculaires. Selon l'application des modèles QSAR, le nombre de composés qui doivent être synthétisés par un chimiste médical peut être considérablement réduit. Ainsi, le temps et le coût de la découverte et du développement de médicaments peuvent également être réduits. Par conséquent, l'objectif de la présente étude est de développer une hypothèse de pharmacophore et de construire un modèle QSAR à base d'atomes robustes pour trouver des caractéristiques responsables de l'activité anticancéreuse des dérivés pyrazoles. La relation quantitative d'activité de structure (QSAR) est l'outil le plus pratique en chimie-informatique. Dans cette deuxième application notre objectif principal est de développer un modèle novateur pour étudier la relation entre la structure et l'activité anticancéreuse des pyrazoles et de ses dérivés[16-19].

Dans cette étude, la régression linéaire multiple (RLM), les moindres carrés partiels (PLS), la régression non linéaire multiple (RNLM) et les analyses de validation croisée et Y-Randomisation, ont été appliquées à une série de dérivés pyrazoles, afin de développer un modèle QSAR fiable et capable de prédire l'activité anticancéreuse pour des molécules dont les activités sont mal connues.

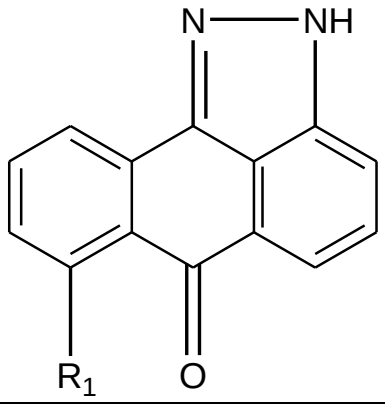
2. METHODOLOGIE

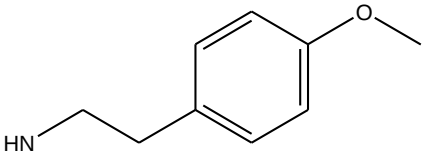
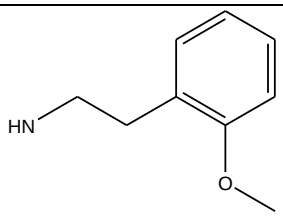
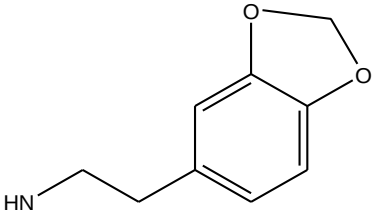
2.1. Base de données:

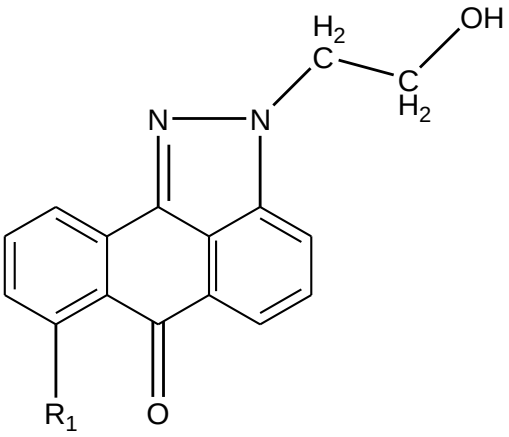
Dans la présente étude, nous avons choisi 32 substituants à l'anthra [1,9-cd] pyrazol-6 (2H) -one pour lesquels les activités anti-cancer sont rapportées dans la littérature par T.C. Chen et et al[20]. Les valeurs déclarées de IC_{50} ont été converties en pIC_{50} en prenant un logarithme négatif ($pIC_{50} = -\log_{10} IC_{50}$) et ensuite utilisées comme variables dépendantes pour développer le modèle QSAR.

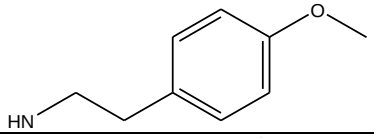
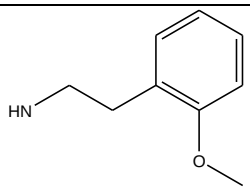
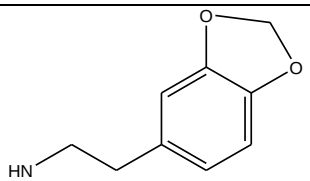
La figure 1 illustre la structure de base des molécules étudiées et le tableau 1 montre les substituants de ces composés et leurs activités expérimentales correspondantes pIC_{50}.

Tableau 1 : Structures et activités expérimentales des composés étudiés.



Composés	R_1	pIC_{50}
1	Cl	1,258
2	$NHCH_2CH(CH_3)_2$	1,251
3	$NHCH_2CH_2CH_3$	1,491
4	$NHCH_2CH_2CH_2CH_2CH_2OH$	1,000
5	$NHCH_2CH_2C_6H_5$	1,340
6	NH-Cyclohexane	1,109
7	NH-Cyclopentane	1,200
8	$NHCH_2$-Cyclohexane	1,302
9	$NHCH_2CH_2CH_2OH$	1,160
10	$NHCH_2CH_2CH_2CH_3$	1,386
11	NH CH_3	1,556
12	$NHCH_2C_6H_5$	0,946
13	$NHCH_2CH_2$-Cyclohexane	1,505
14		1,532
15		0,926
16		0,740



17	Cl	1,137
18	$NHCH_2CH(CH_3)_2$	1,455
19	$NHCH_2CH_2CH_3$	1,484
20	$NHCH_2CH_2CH_2CH_2CH_2OH$	1,486
21	$NHCH_2CH_2C_6H_5$	1,489
22	NH-Cyclohexane	1,490
23	NH-Cyclopentane	1,346
24	$NHCH_2$-Cyclohexane	1,254
25	$NHCH_2CH_2CH_2OH$	1,272
26	$NHCH_2CH_2CH_2CH_3$	1,519
27	NH CH_3	1,487
28	$NHCH_2C_6H_5$	1,493
29	$NHCH_2CH_2$-Cyclohexane	1,324
30		1,505
31		1,504
32		1,497

2.2. Calcul des descripteurs moléculaires

Le calcul des descripteurs électroniques a été effectué à l'aide de la méthode basée sur la théorie de la fonctionnelle de la densité DFT B3LYP / (6-31G (d)) : l'énergie de l'orbitale moléculaire la plus haute occupée E_{HOMO} (eV), l'énergie de l'orbitale moléculaire la plus basse inoccupée E_{LUMO} (eV), l'énergie totale E_{totale} (eV), le moment dipolaire μ (D), et l'énergie de répulsion $E_{répulsion}$ (eV).

D'autres descripteurs ont été obtenus à l'aide de l'application CHEM BIO OFFICE 2015, à savoir le poids moléculaire (MW), le coefficient de partage (logP), les éléments accepteurs de la liaison hydrogène (HA) et les éléments donneurs de la liaison hydrogène (HD).

Les valeurs du Volume Molaire (MV (cm^3)), la Réfractivité molaire (MR (cm^3)), le Parachor (Pc (cm^3)), la Densité (g / cm^3), l'indice de réfraction IR, la tension superficielle TS (Dyne / Cm) et la Polarisabilité PLR (cm^3) sont mesurées par le programme CHEMSKETCH.

2.3. Analyses statistiques

Les bases de l'étude QSAR, représentées par la régression linéaire multiple, la régression des moindres carrées partiels et la régression non linéaire, sont appliquées par utilisation du logiciel XLSTAT 2014. Les résultats obtenus sont validés à l'aide de la validation externe en divisant 32 molécules en 24 molécules pour former les modèles mathématiques et pour tester les modèles 8 molécules sont utilisées (24 jeu d'entrainement et 8 pour le jeu de test).

La validation croisée est mise en jeu afin de tester la fiabilité et la robustesse des modèles obtenus, et la procédure supplémentaire Y-randomisation a été utilisée pour s'assurer que les modèles obtenus ne sont pas dus à la chance.

2.4. Validation des modèles QSAR

Le meilleur modèle QSAR est sélectionné sur la base de la valeur des paramètres statistiques citons R^2 (coefficient de détermination pour la série d'apprentissage), $R^2_{ajusté}$ coefficient de détermination ajusté et R le coefficient de corrélation. Dans de nombreux cas, un modèle avec des valeurs R^2 et R élevées peuvent s'avérer inexactes, donc la seule façon d'estimer le vrai pouvoir prédictif d'un modèle est de le tester par une validation externe. Dans notre étude le modèle QSAR a été validé et testé son pouvoir extrapolable en élaborant la série de test de huit composés. Aussi nous avons utilisé la validation externe qui est le seul moyen d'établir un modèle QSAR fiable. Des tests statistiques sont utilisés pour tester la fiabilité des méthodes QSAR tels que le coefficient de détermination (R^2), le coefficient de corrélation (R), le coefficient de corrélation ajusté (R^2_{adj}), l'erreur moyenne carrée (MSE), l'erreur absolue moyenne (MAE), La valeur de Fischer (F) et la valeur de p-value (p) <0,05. Un modèle QSAR que l'on qualifiera de bon possède une grande valeur de F, une petite valeur de MSE, une très petite valeur de p, et des valeurs R et R^2 proches de l'unité.

3. RESULTATS

3.1. Ensemble de données pour analyse

L'étude QSAR a été réalisée pour une série de 32 substituants à l'anthra [1,9-cd] pyrazol-6 (2H) - one, afin de déterminer une relation quantitative entre leurs structures et leurs activités anticancéreuses. Les valeurs des 16 descripteurs calculés sont indiquées dans le tableau 2. Les résultats obtenus pour QSAR en utilisant ACP, RLM, RNLM, PLS sont représentés dans les Tableaux 3 et 4.

Tableau 2 : Valeurs des descripteurs utilisées pour l'analyse QSAR des dérivés pyrazoles

	E_{LUMO}	E_{HOMO}	E_{totale}	RE ev	MD	MR	MV	Parac	Indi R	S T	Desty	Polariss	Log P	D H	A H	MW
1	-2,6	-6,265	-32173,3	35590,20	5,70	69,18	162,4	488,5	1,80	81,8	1,57	27,42	4,18	1	3	254,67
2	-2,08	-5,261	-25451,8	46521,92	1,76	87,75	219,8	636,1	1,73	70	1,33	34,79	4,30	2	4	291,35
3	-2,19	-5,318	-24382,3	43034,30	1,87	83,16	202,9	598	1,76	75,4	1,37	32,97	3,98	2	4	277,33
4	-2,48	-5,362	-28567,7	54475,02	3,49	93,96	233,4	695,2	1,74	78,6	1,38	37,25	3,53	3	5	321,38
5	-2,27	-5,242	-29598,9	57352,68	1,55	103,02	247,1	731,8	1,77	76,8	1,37	40,48	5,10	2	4	339,40
6	-2,23	-5,187	-27558,6	52684,48	1,02	94,93	232,4	686,7	1,75	76,1	1,37	37,63	4,98	2	4	317,39
7	-2,17	-5,224	-25508,7	49729,89	1,74	90,32	214,7	646,7	1,78	82,3	1,41	35,8	4,42	2	4	303,37
8	-2,18	-5,265	-28627,1	56732,18	1,79	99,58	252,4	726,8	1,72	68,7	1,31	39,47	5,59	2	4	331,42
9	-2,01	-5,197	-26427,9	46287,76	2,01	84,7	200,4	615,1	1,79	88,6	1,46	33,57	3,28	3	5	293,33
10	-2,08	-5,259	-25451,8	46211,30	1,83	87,8	219,4	638,1	1,73	71,4	1,33	34,8	4,54	2	4	291,35
11	-2,09	-5,293	-22243,3	36510,24	1,82	73,9	169,9	517,9	1,82	86,2	1,47	29,29	3,22	2	4	249,27
12	-2,26	-5,374	-28528,8	53453,30	2,21	98,39	230,6	691,7	1,80	80,9	1,41	39	4,94	2	4	325,37
13	-2,18	-5,268	-29696,8	59558,41	1,71	104,21	268,9	766,9	1,70	66,1	1,28	41,31	6,35	2	4	345,45
14	-2,08	-5,266	-32713,4	63731,66	1,16	109,7	271,1	790,4	1,74	72,2	1,36	43,48	5,00	2	5	369,42
15	-2,23	-5,230	-31644,6	61133,18	0,78	105,06	254,6	750,4	1,76	75,4	1,40	41,65	5,00	2	5	369,42
16	-2,30	-5,315	-33657,2	63325,51	1,68	104,55	241,8	739,4	1,81	87,4	1,53	41,44	4,85	2	6	383,41
17	-2,4	-5,967	-36357,9	46089,73	7,72	79,4	195,5	546,9	1,75	61,2	1,52	31,47	3,24	1	4	298,73
18	-1,91	-5,125	-29669,1	57802,61	4,09	95,91	253,4	677,3	1,68	50,9	1,32	38,02	3,36	2	5	335,41
19	-1,91	-5,128	-28598,4	54046,80	4,14	91,49	238,2	646,2	1,69	54,1	1,34	32,27	3,04	2	5	321,38
20	-1,94	-5,159	-32787,5	63911,53	4,50	101,75	268,5	736,6	1,68	56,6	1,36	40,33	2,59	3	6	365,43
21	-2,04	-5,199	-33820,1	70238,20	4,30	112,17	290,9	791,9	1,70	54,9	1,31	44,47	4,16	2	5	383,45
22	-2,01	-5,116	-31777,5	65294,20	4,18	102,95	258,8	714,7	1,73	58,1	1,39	40,81	4,04	2	5	361,45
23	-2	-5,098	-30706,6	61164,80	4,11	98,35	242,7	676,1	1,74	60,1	1,43	38,98	3,48	2	5	347,42
24	-2,02	-5,142	-32847,8	68582,30	4,17	107,56	274,8	753,3	1,71	56,4	1,36	42,64	4,65	2	5	375,47
25	-2,05	-5,185	-30646,2	57276,36	5,66	92,53	236,4	659,4	1,71	60,5	1,42	36,68	2,34	3	6	337,38
26	-2,01	-5,137	-29669,0	57252,92	4,14	96,1	254,3	684,8	1,70	52,5	1,31	38,09	3,60	2	5	335,41
27	-1,92	-5,151	-26457,0	47005,42	4,17	82,27	206,1	569	1,73	58	1,42	32,61	2,28	2	5	293,33
28	-1,95	-5,167	-32749,5	65020,66	4,18	107,56	274,8	753,3	1,71	56,4	1,34	42,64	4,00	2	5	369,42
29	-2,09	-5,096	-33919,4	71855,38	3,36	112,17	290,9	791,9	1,70	54,9	1,33	44,47	5,41	2	5	389,50
30	-2,08	-5,081	-36939,3	76895,79	2,90	117,99	312,5	842,2	1,70	52,7	1,32	46,77	4,06	2	6	413,48
31	-1,93	-4,998	-35868,0	73878,45	3,17	113,38	296,5	803,6	1,69	53,9	1,34	44,94	3,90	2	6	399,45
32	-2,15	-5,215	-37883,5	76645,08	3,86	112,4	275,3	775,5	1,75	62,9	1,5	44,56	3,75	2	7	413,43

3.2. Analyse en composantes principales

La totalité des 16 descripteurs codant les 32 molécules est soumise à une analyse des composants principales (ACP), 16 composants principales ont été obtenues (Figure 2).

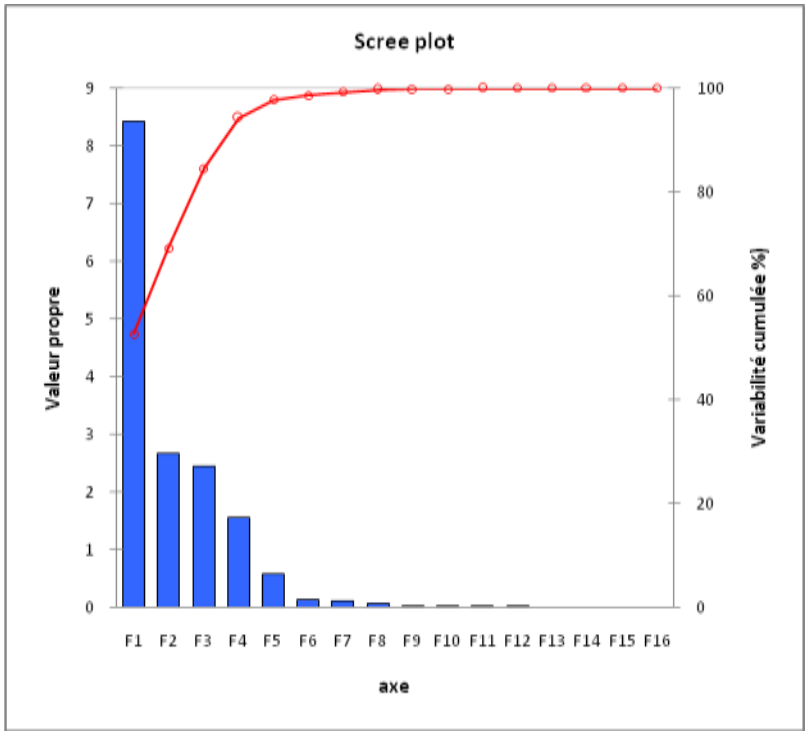


Figure 2 : Les composantes principales et leurs variances

Les trois premiers axes principaux sont suffisants pour décrire les informations fournies par la matrice de données. En effet, les pourcentages de variances sont de 52,6%, 16,62% et 15,25% pour les axes F1, F2 et F3, respectivement. L'information totale est estimée à un pourcentage de 84,47%. Le tableau 3 montre la matrice de corrélation (Pearson (n)) ainsi obtenue entre différents descripteurs.

Tableau 3 : La matrice de corrélation

Desc.	E_{LUMO}	E_{HOMO}	E_{TOT}	RE	MD	MR	MV	Parac	IndiR	S T	Desty	Plris	LogP	DH	AH	M W
E_{LU}	**1**															
E_{HO}	0,78	**1**														
E_{TOT}	0,04	0,065	**1**													
RE	0,36	0,555	-0,76	**1**												
MD	0,03	-0,37	-0,45	0.047	**1**											
MR	0,31	0,590	-0,65	**0,963**	-0,17	**1**										
MV	0,42	0,614	-0,65	0,958	-0,05	**0,968**	**1**									
Para	0,28	0,593	-0,59	0,929	-0,25	**0,989**	**0,962**	**1**								
InR	-0,57	-0,45	0,352	-0,54	-0,34	-0,47	-0,68	-0,486	**1**							
S T	-0,61	-0,39	0,494	-0,58	-0,55	-0,45	-0,63	-0,401	0,904	**1**						
Dest	-0,50	-0,62	-0,11	-0,35	0,285	-0,47	-0,59	-0,532	0,734	0,504	**1**					
PolS	0,27	0,563	-0,65	0,950	-0,18	**0,990**	0,951	**0,980**	-0,44	-0,42	-0,44	**1**				
LoP	-0,34	-0,02	-0,06	0,199	-0,6	0,388	0,284	0,437	0,137	0,268	-0,29	0,406	**1**			
DH	0,33	0,563	0,210	0,173	-0,19	0,182	0,195	0,247	-0,17	0,046	-0,25	0,182	-0,30	**1**		
AH	0,45	0,535	-0,59	0,744	0,190	0,616	0,620	0,575	-0,38	-0,44	-0,01	0,599	-0,36	0,464	**1**	
MW	0,3	0,504	-0,79	**0,989**	0,020	**0,962**	0,937	0,929	-0,47	-0,52	-0,29	0,950	0,225	0,152	0,754	**1**

L'analyse en composantes principales (ACP) a été réalisée pour identifier le lien entre les différentes variables et fournit des informations sur la corrélation négative ou positive entre eux. Les corrélations entre les 16 descripteurs sont présentées dans le tableau 3 comme une matrice de corrélation, et les descripteurs sont représentés dans un cercle de corrélation (Figure 3).

La matrice obtenue fournit des informations sur l'interrelation haute ou basse entre les variables. En général, une bonne co-linéarité (r> 0,5) a été observée entre la plupart des variables, dans notre étude une corrélation élevée a été observée entre :

MW et RE, MR et RE, MR et MW, MR et Pola, MR et Parachor, MR et MV, Parachor et MV, polrt et le parachor.

Pour diminuer la redondance dans notre matrice de données, les descripteurs fortement corrélés (R ≥ 0,96) ont été exclus. Les variables qui seront retirées sont: Parachoc, RE et MR.

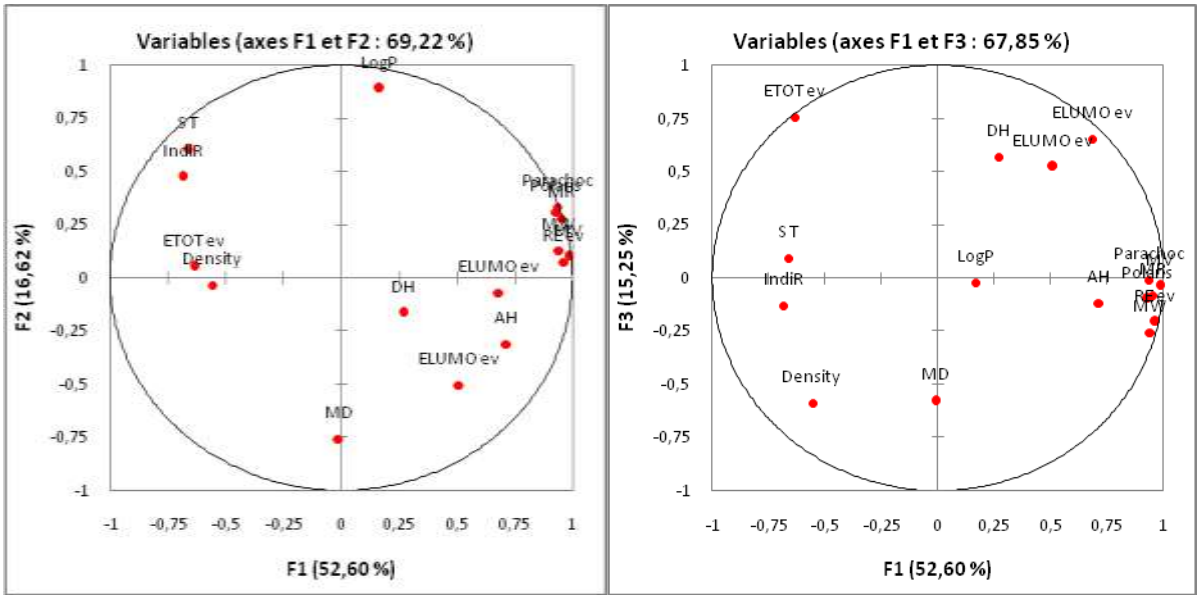


Figure 3 : Cercles de corrélations entre les descripteurs

3.3. Régression linéaire multiple (RLM)

Afin de sélectionner les descripteurs prédominants affectant l'activité anticancéreuse des dérivés pyrazoles, l'analyse de corrélation a été réalisée à l'aide du logiciel XLSTAT 2014 prenant chaque descripteur calculé comme variable indépendante et $p\text{IC}_{50}$ comme variable dépendante. Sur la base de l'analyse de corrélation, la méthode de régression linéaire multiple a été utilisée pour établir le modèle QSAR. Dans cette étude, tous les modèles QSAR développés sont statistiquement significatifs avec un niveau de signification étant (p) $<10^{-3}$. Étant donné que la valeur p est beaucoup plus petite que 0,05, nous prenons un risque inférieur à 0,01% en supposant que l'hypothèse nulle est erronée. Les valeurs du coefficient de corrélation (R) et du coefficient de

détermination (R^2) qui sont supérieures à 0,87 et 0,75 respectivement, prouvent la capacité estimée des modèles QSAR.

L'énergie LUMO, le volume Molaire (MV), la densité et le poids moléculaire (MW) sont les descripteurs qui dépendent de l'activité anticancéreuse des dérives pyrazoles. Le modèle QSAR construit à l'aide de la méthode de régression linéaire multiple (RLM) est représenté par l'équation suivante :

pIC_{50} = -5,834 + 0,767*E_{LUMO} + 0,04005*MV + 6,238*Densité - 0,0286*MW (équation 1)

Les caractéristiques statistiques de l'équation obtenue sont :

N=24 R=0,87 R^2=0,755 $R^2_{Ajusté}$= 0,704 F = 14,661 MSE= 0,011 MAE=0,0818 P<0,0001 N_{test}=8 R_{test}=0,72 MSE_{test}=0,035 MAE_{test}=0,132

t_{ELUMO} = 4,247 t_{MV} = 4,81 $t_{Density}$ = 3,962 t_{MW} = -4,939

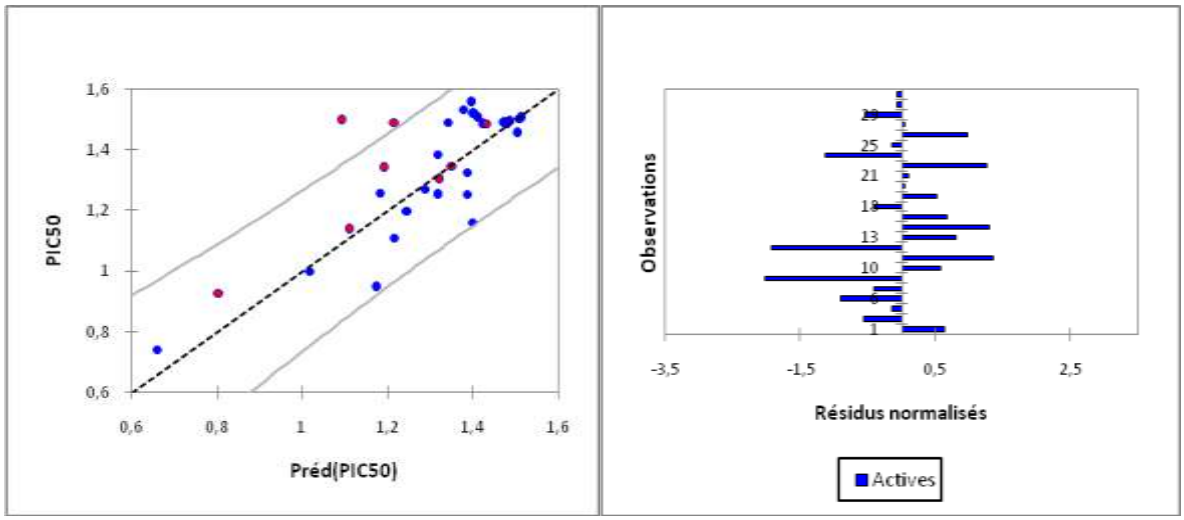


Figure 4 : Activités anticancéreuses prédites pIC_{50} par (RLM) par rapport aux valeurs expérimentales (la série d'apprentissage en bleu et la série de test en rouge).

La corrélation entre les valeurs expérimentales et les données prédites à partir de la régression linéaire multiple est donnée dans le tableau3. Les résultats obtenus montrent une grande corrélation entre les valeurs expérimentales et prédites. Le modèle développé pourront appliqué avec succès pour prédire l'activité d'autres dérivés pyrazoliques.

3.4. Méthode des moindres carrés partiels (PLS)

La base de données de l'étude des moindres carrés partiels est les descripteurs retenus par la RLM correspondant aux 32 molécules. Cette méthode utilise les coefficients R, R^2, $R^2_{ajusté}$ et les valeurs F, pour sélectionner la meilleure performance de régression. L'équation obtenue à partir de l'étude PLS est la suivante :

pIC_{50} = -3,991 + 0,757 E_{LUMO} + 0,0328 MV + 5,113 Densité – 0,0241MW (equation 2)

N=24 ; R^2=0,69 ; R=0,83 ; MAE=0,0687 , MSE=0,0078 N_{test}=8 R_{test}=0,71 MAE_{test}=0,159 MSE_{test}= 0,043

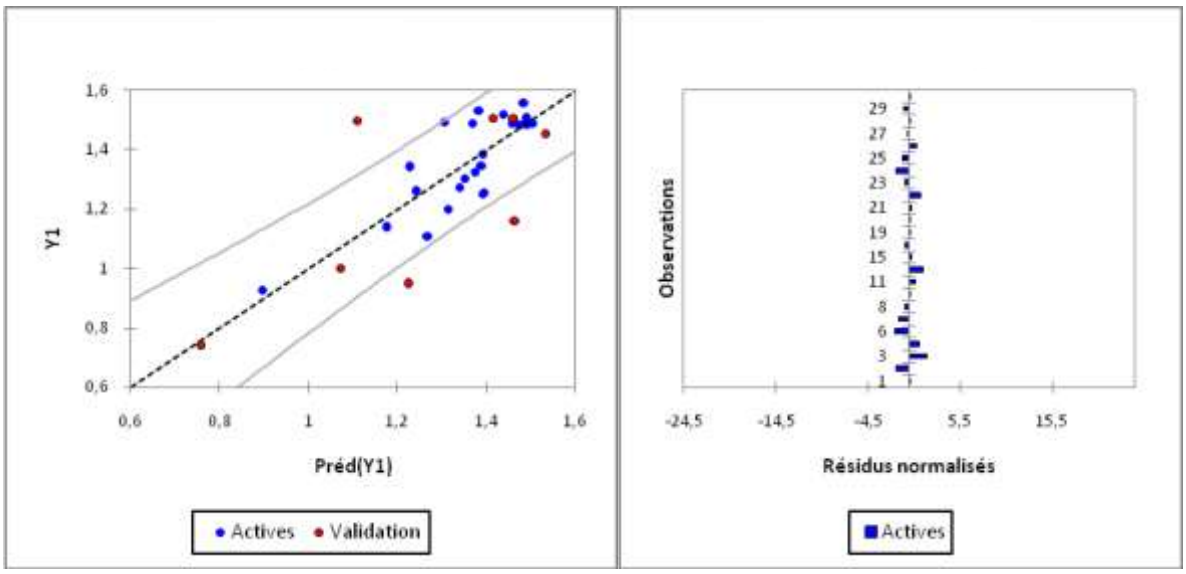


Figure 5 : Représentation graphique des activités calculée et observée par la méthode PLS.

Le coefficient de corrélation obtenu dans l'équation (2) est assez intéressant (0,69). Pour améliorer et prendre en compte la corrélation non linéaire entre la structure et l'activité anticancéreuse de manière quantitative, compte tenu de plusieurs paramètres nous avons utilisé la technique du modèle de régression non linéaire.

3.5. Régression non linéaire multiple (RNLM)

Les descripteurs de base de la RNLM sont ceux retenus par la RLM. Les coefficients R, R^2, erreur moyenne absolue MAE et l'erreur quadratique moyenne MSE sont utilisés pour sélectionner la meilleure performance de la régression.

L'équation résultante est :

$$pIC_{50} = 9{,}561 + 4{,}178\ E_{LUMO} + 0{,}0955\ MV - 2{,}935\ Densité - 0{,}0977\ MW + 0{,}7983\ (E_{LUMO})^2 - 0{,}0001134\ (MV)^2 + 3{,}0171\ (Density)^2 + 0{,}000103(MW)^2 \quad \text{(équation 3)}$$

N=24 R=0,905 R^2=0,82 MSE= 0,0091 MAE=0,0818 N_{test}=8 R_{test}=0,75 MSE_{test}=0,0203 MAE_{test}=0,1234

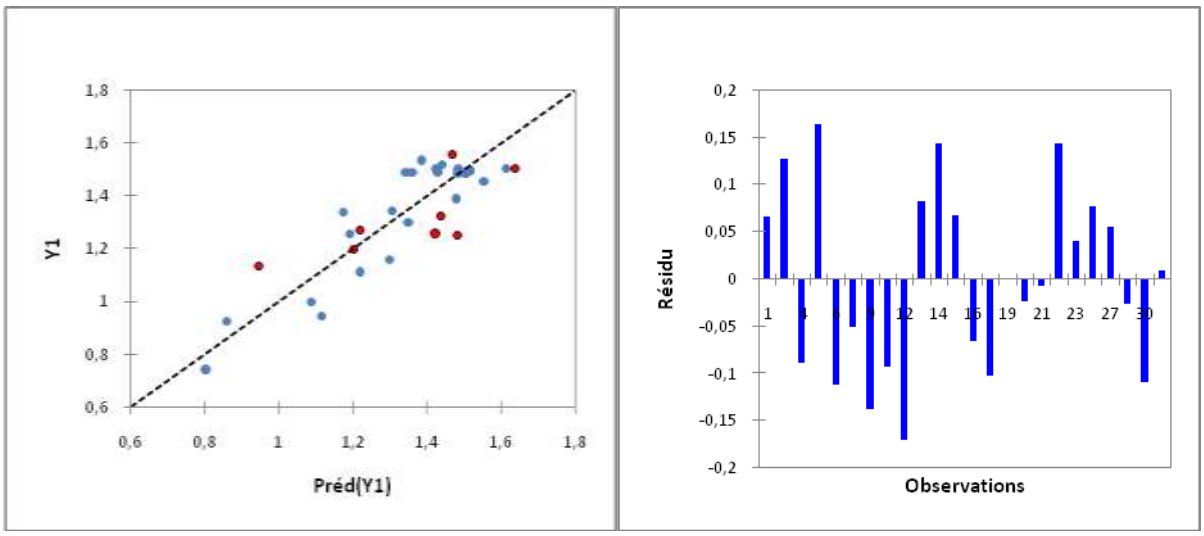


Figure 6 : Activité anticancéreuses prédites par la méthode RNLM en comparaison avec les valeurs expérimentales

La corrélation entre les activités de valeurs prédites et observées est indiquée dans le tableau 3 et la figure 6. Le coefficient de corrélation obtenu dans l'équation 3 est très intéressant (0,91) pour prédire l'activité anticancéreuse. Nous pouvons dire que les valeurs obtenues à partir de la régression non linéaire sont fortement corrélées avec celles de l'activité observée en comparant les résultats obtenus par les méthodes RLM et PLS.

Tableau 4: Comparaison entre les activités observées et prédites des modèles statistiquement significatifs obtenus par l'ensemble des formations.

Composés	pIC_{50}	Préd(pIC_{50}) RLM	Résidu	Préd(pIC_{50}) RNLM	Résidu	Préd(pIC_{50}) PLS	Résidu
1	1,258	1,183	0,075	1,192	0,066	1,242	0,016
2	1,251	1,318	-0,067	1,363	0,128	1,390	-0,139
4	1,000	1,017	-0,018	1,089	-0,09	1,305	0,186
6	1,109	1,215	-0,106	1,175	0,165	1,228	0,112
7	1,200	1,247	-0,047	1,220	-0,11	1,268	-0,160
9	1,160	1,398	-0,238	1,352	-0,05	1,313	-0,113
10	1,386	1,318	0,068	1,298	-0,14	1,350	-0,048
11	1,556	1,395	0,161	1,478	-0,09	1,390	-0,005
12	0,946	1,173	-0,227	1,116	-0,17	1,485	0,072
13	1,505	1,411	0,094	1,422	0,083	1,384	0,147
14	1,531	1,379	0,153	1,387	0,145	0,895	0,031
16	0,740	0,660	0,079	0,858	0,068	1,176	-0,039
18	1,455	1,503	-0,048	0,805	-0,07	1,476	0,008
19	1,484	1,422	0,062	1,557	-0,10	1,488	-0,002
20	1,486	1,480	0,005	1,484	0,001	1,457	0,031
21	1,489	1,475	0,013	1,508	-0,02	1,368	0,122
22	1,490	1,341	0,149	1,496	-0,01	1,390	-0,044
24	1,254	1,387	-0,133	1,345	0,144	1,394	-0,140
25	1,272	1,289	-0,017	1,305	0,040	1,338	-0,067
26	1,519	1,404	0,115	1,441	0,077	1,437	0,081
28	1,493	1,486	0,006	1,432	0,055	1,505	-0,017

29	1,324	1,387	-0,063	1,518	-0,03	1,488	0,004
30	1,505	1,514	-0,009	1,615	-0,11	1,374	-0,050
31	1,504	1,512	-0,008	1,488	0,009	1,490	0,014

Tableau 5: Comparaison entre les activités observées et prédites des modèles statistiquement significatifs obtenus par l'ensemble des modèles tests.

Composés	pIC_{50}	pIC_{50} RLM test	pIC_{50} RLNM test	pIC_{50} PLStest
3	1,491	1,217	1,483	1,071
5	1,340	1,192	1,201	1,465
8	1,302	1,321	1,468	1,225
15	0,926	0,805	0,945	1,414
17	1,137	1,110	1,426	0,759
23	1,346	1,353	1,220	1,531
27	1,487	1,432	1,439	1,462
32	1,497	1,093	1,639	1,110

3.6. Validation

Nous avons utilisés la procédure «Leave-One-Out» qui supprime successivement une molécule de la série d'apprentissage du jeu contenant 24 molécules. Cette procédure est répétée 24 fois pour prédire les propriétés de toutes les molécules.

La cohérence et la fiabilité des modèles RLM, RNLM et PLS sont validées à l'aide de la technique de validation croisée avec une bonne corrélation dont le coefficient de corrélation est Rcv = 0,86. Donc le pouvoir prédictif de ce modèle est très important.

N=24 R= 0,86 MSE = 0,0126 MAE = 0,078 SD=0,08002 P<0,0001

Le coefficient de corrélation obtenu est très intéressant (0,86) ce qui prouve la fiabilité du modèle obtenu par RLM.

3.7. Scrambling ou Y-randomisation

La méthode Randomisation est largement utilisée dans les études QSAR pour assurer la fiabilité des résultats de la validation croisée. Cette méthode est utilisée après que le modèle de régression "meilleur" est sélectionné pour s'assurer qu'il n'y a pas de corrélations possibles. Scrambling valide le modèle LOO en comparant les performances du modèle original à celles des modèles construits pour les réponses permutées (réparties au hasard). Si le coefficient de corrélation des modèles obtenues pour les réponses permutées est inférieur de celui obtenu en appliquant le modèle LOO, ce résultat indique qu'il existe une indépendance entre les molécules, car les points de mesure du point cible les plus proches n'obscurcissent pas les autres données expérimentales et ne sont pas presque exclusivement impliqués dans l'estimation, et les données utilisées dans cette validation sont réparties uniformément dans l'espace. Par conséquent, le modèle résultant peut être extrapolé à tout l'ensemble de la série d'apprentissage.

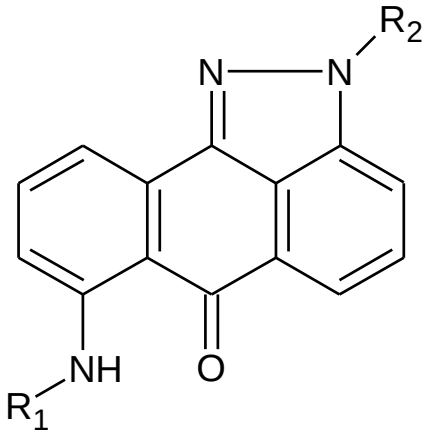
N = 24 ; R = 0,68 ; SD =0,076 P < 0,0001 MSE=0,0126 MAE=0,0854

La valeur du coefficient de corrélation des molécules en désordre est inférieure de celle obtenue en appliquant le modèle initial (LOO). Ce résultat démontre l'absence de dépendance entre les descripteurs inclus dans le modèle. D'où le modèle de RLM n'est pas obtenu par hasard.

3.8. Proposition de nouveaux composés

Les valeurs des paramètres obtenus par calculs DFT pour les composés proposés sur la base des informations dérivées des équations 1, 2 et 3 sont citées dans le tableau 6. Nous avons observés que les PLS conçus ont des valeurs de $p\text{IC}_{50}$ plus élevées que le modèle RLM et RNLM. En outre, les composés (X1 et X15) ont des valeurs $p\text{IC}_{50}$ plus élevées que les composés existants dans le cas des 32 composés étudiés.

Tableau 6 : Proposition de nouveaux composés



Composés	R_1	R_2	E LUMO	MW	MV	Density	$p\text{IC}_{50}$ RLM	$p\text{IC}_{50}$ RNLM	$p\text{IC}_{50}$ PLS
X1	H	H	-2,277	237,26	147,4	1,6	1,52	1,44	1,58
X2	OH	H	-2,421	253,26	152,6	1,65	1,48	1,3	1,51
X3	COOH	H	-2,555	281,27	166	1,69	1,36	1,14	1,38
X4	CN	H	-2,901	262,27	168,2	1,55	0,85	1,17	0,93
X5	CH_2OH	H	-2,238	267,28	168,7	1,58	1,42	1,23	1,49
X6	CH_2Cl	H	-2,741	285,73	180,7	1,58	1	1,05	1,05
X7	CCl_3	H	-3,01	354,62	204,7	1,73	0,71	1,26	0,75
X8	H	CH_2CH_2OH	-2,096	281,31	182,9	1,53	1,38	1,23	1,47
X9	OH	CH_2CH_2OH	-2,235	297,31	188,1	1,57	1,29	1,06	1,35
X10	COOH	CH_2CH_2OH	-2,266	325,32	201,5	1,61	1,25	1,04	1,29
X11	CN	CH_2CH_2OH	-2,365	306,32	203,7	1,5	1,2	1,01	1,19
X12	CH_2OH	CH_2CH_2OH	-2,059	311,34	204,2	1,52	1,33	1,19	1,39

X13	CH_2Cl	CH_2CH_2OH	-2,554	329,78	216,2	1,52	0,93	0,94	0,99
X14	CCl_3	CH_2CH_2OH	-2,817	398,67	240,2	1,65	0,53	1,23	0,59
X15	$(C_6H_{10})_2$	CH_2CH_2OH	-2,065	445,6	332,6	1,33	1,52	1,91	1,42

3.9. Règle de cinq de Lipinski

Les résultats du calcul (tableau 7) montrent que tous les composés satisfont les règles de Lipinski, ce qui suggère que ces composés n'ont pas de problèmes de biodisponibilité orale sauf les molécules 5, 8, 13 et 29 qui ont des valeurs de Log P qui dépassent 5.

Tableau 7 : Violations de la règle de lipinski

Comp.	Log(P)	DH	AH	MW	Nbr. Of Violation
1	4,18	1	3	254,67	0
2	4,30	2	4	291,35	0
3	3,98	2	4	277,33	0
4	3,53	3	5	321,38	0
5	**5,10**	2	4	339,40	1
6	4,98	2	4	317,39	0
7	4,42	2	4	303,37	0
8	**5,59**	2	4	331,42	1
9	3,28	3	5	293,33	0
10	4,54	2	4	291,35	0
11	3,22	2	4	249,27	0
12	4,94	2	4	325,37	0
13	**6,35**	2	4	345,45	1
14	5,00	2	5	369,42	0
15	5,00	2	5	369,42	0
16	4,85	2	6	383,41	0
17	3,24	1	4	298,73	0
18	3,36	2	5	335,41	0
19	3,04	2	5	321,38	0
20	2,59	3	6	365,43	0
21	4,16	2	5	383,45	0
22	4,04	2	5	361,45	0
23	3,48	2	5	347,42	0
24	4,65	2	5	375,47	0
25	2,34	3	6	337,38	0
26	3,60	2	5	335,41	0
27	2,28	2	5	293,33	0
28	4,00	2	5	369,42	0
29	**5,41**	2	5	389.50	1
30	4,06	2	6	413,48	0
31	3,90	2	6	399,45	0
32	3,75	2	7	413,43	0

4. DISCUSSION DES RESULTATS

Dans cette deuxième application, la technique ACP à été utilisée afin d'éliminer les descripteurs indépendants qui sont fortement corrélés entre eux en examinant la multicollinéarité entre les

descripteurs. Par conséquent, le Parachor, la réfraction molaire et l'énergie de répulsion seront supprimés dans la suite des études statistiques. La présence de la multicolinéarité entre les descripteurs a été confirmée à partir de la matrice de corrélation.

Pour le modèle de RLM, la valeur p est inférieure à 0,0001, cela signifie que nous prendrions un risque inférieur à 0,01% en supposant que l'hypothèse nulle (sans effet des variables explicatives) est erronée et que l'équation de régression est statistiquement importante. Par conséquent, nous concluons que les variables sélectionnées contiennent des informations importantes.

Les valeurs statistiques du modèle de la RLM sont (R=0,87, R^2=0,755, $R^2_{Ajusté}$= 0,704, F = 14,661, MSE= 0,011, MAE=0,0818, P<0,0001, N_{test}=8, R_{test}=0,72, MSE_{test}=0,035, MAE_{test}=0,132) de MNLR (R=0,905, R^2=0,82, MSE=0,0091, MAE=0,0818, N_{test}=8, R_{test}=0,75, MSE_{test}=0,0203, MAE_{test}=0,1234) et de PLS (R^2=0,69, R=0,83, MAE=0,0687 , MSE=0,0078 N_{test}=8 R_{test}=0,71 MAE_{test}=0,159 MSE_{test}= 0,04) aussi la valeur des erreurs (MAE et MSE) qui s'approche de la valeur zéro, indiquent que les trois modèles proposés sont prédictifs et statistiquement fiables.

Les modèles obtenus ont été validés en interne par la technique de validation croisée, le coefficient de validation croisée R^2_{cv} a été déterminé en fonction de la capacité prédictive du modèle obtenu par RLM. La valeur de R^2_{cv} est supérieure à 0,5, indique une meilleure prédiction des modèles.

La comparaison de la qualité des modèles RLM, PLS et RNLM montre que les trois approches possèdent la bonne capacité prédictive ; ce qui est suffisant pour conclure la performance de ces modèles pour établir une relation satisfaisante entre les descripteurs sélectionnés et l'activité anticancéreuse. En outre, les résultats obtenus par RNLM sont relativement meilleurs que ceux obtenus par RLM et PLS, mais cette dernière approche (PLS) est plus fiable et donne des résultats plus interprétables. Par conséquent, avec les approches RLM, PLS et RNLM, nous pouvons concevoir de nouveaux composés avec des valeurs d'activité améliorées que les composés étudiés. Compte tenu des résultats trouvés, nous avons pu ajouter des substitutions appropriées et calculés les activités de nouveaux composés en utilisant les modèles proposés (éq1, éq2 et éq3).

Selon les valeurs du test t (| t |), l'importance des descripteurs impliqués dans ce modèle est dans l'ordre suivant : MW> MV > E_{LUMO} >densité . Le descripteur le plus important selon le t- test est le poids moléculaire. Le second descripteur est le volume molaire suivi de l'énergie LUMO et le dernier est la densité. L'énergie LUMO est directement liée à l'affinité électronique d'une molécule et elle caractérise la sensibilité de la molécule d'être attaquée par les nucléophiles.
Le coefficient (équation 1) positif obtenu pour le MV, E_{LUMO} et la densité comme descripteurs de la chimie quantique dans le meilleur modèle, appuie le concept que l'activité anticancéreuse des

dérivés pyrazoliques augmente avec l'augmentation de leur capacité d'accepter des électrons, indiquant, que le transfert d'électrons a lieu de l'organisme aux cellules cibles. tant dis qu'une diminution du poids moléculaire des dérivés pyrazoliques entraine l'augmentation de l'activité, donc le MW varie en sens inverse que l'activité.

Pour montrer que les résultats du modèle de validation croisée ne sont pas obtenus par hasard, un test de randomisation a été effectué. Les valeurs de pIC_{50} de la série d'apprentissage ont été mélangé de façon aléatoire en gardant les paramètres retenus de la régression linéaire multiple inchangés. Cette opération est répétée trois fois. Les résultats montrent que les valeurs des coefficients de détermination de la série d'apprentissage calculés par le modèle généré sont inférieures par rapport à celles de notre modèle. Ceci confirme que le modèle LOO n'est pas obtenu par hasard.

La plus grande énergie de l'orbitale moléculaire occupée E_{LUMO} a un signe positif dans le modèle de RLM, ce qui suggère que la substitution des dérivés pyrazoles par un groupe de capacité électronique donateur plus fort (comme le phényl) peut conduire à des valeurs élevées de l'activité. L'énergie de l'orbitale moléculaire inoccupée . Le coefficient de partage Octanol / Eau Log P a un signe positif dans le modèle de RLM, ce qui suggère qu'une valeur plus élevée du Log P (capacité hydrophile lipophile / plus faible) et une substitution des pyrazoles a un groupe non polaire peut entraîner des valeurs élevées de l'activité.

Les équations développées peuvent être utilisées pour la conception de nouveaux dérivés pyrazoliques avec une activité anticancéreuse améliorée (pIC_{50}). Si nous développons un nouveau composé avec des valeurs d'activités plus élevées que les composés existants, cela peut donner lieu à l'apparition de composés plus actifs que ceux actuellement utilisés.

Les conditions empiriques pour satisfaire les règles de Lipinski ont démontrés une bonne biodisponibilité orale impliquant un équilibre entre la solubilité aqueuse d'un composé et sa capacité à se diffuser passivement à travers diverses barrières biologiques. Ces paramètres permettent aux membranes la perméabilité qui se produit lorsque la molécule évaluée suit la règle cinq de Lipinski lors de l'absorption orale.

Les molécules qui violent de nombreuses règles de Lipinski peuvent avoir des problèmes de biodisponibilité. Par conséquent, cette règle établit certains paramètres structurels pertinents pour la prévision théorique du profil de biodisponibilité orale et elle est largement utilisée dans la conception de nouveaux médicaments.

D'après notre étude (Résultats de calcul du tableau 7) nous avons conclu que les molécules 5, 8, 13 et 29 ne répondent pas aux règles de Lipinski (log P dépasse la valeur 5) suggérant que ses composés ont des problèmes de biodisponibilité orale.

Selon les discussions et le tableau 6 ci-dessus, les modèles de RLM, RNLM et PLS pourraient être appliqués à d'autres dérivés pyrazoliques et pourraient apporter des connaissances supplémentaires dans l'amélioration de nouvelles méthodes de recherche sur les médicaments anticancéreux. Si nous développons un nouveau composé avec de meilleures valeurs que celles existantes, cela peut entraîner le développement de composés plus actifs que ceux actuellement utilisés. De cette façon, nous avons effectué une modification structurale à partir de composés ayant les valeurs les plus élevées de pIC_{50} comme modèle (numéro 13). Les structures des composés conçus et leurs valeurs de paramètres calculées par les mêmes méthodes, ainsi que les valeurs de pIC_{50} théoriquement prédites par les modèles de RLM, RNLM et PLS sont énumérées dans le tableau 6. A partir des activités prédites, il a été observé que les quinze composés conçus ont des valeurs de pIC_{50} supérieures à celles des composés existants dans le cas des composés étudiés (tableau 1). Nous suggérons tous les autres composés en tant que candidats qui seront synthétisés et évalués en tant que médicaments anticancéreux.

5. CONCLUSION

Dans ce troisième chapitre, nous avons utilisé les méthodes d'analyses statistiques pour élaborer des modèles QSAR fiables, capables de prédire l'activité anticancéreuse d'une série constituée de 32 molécules dérivées de pyrazoles dont les valeurs expérimentales des activités sont comprises entre 0,7 et 1,5 afin de mettre en place le modèle le plus performant possible. L'évaluation de la qualité des modèles RLM, PLS et RNLM a révélé que la capacité prédictive de la RNLM était nettement supérieure à celle des autres méthodes. Le pouvoir prédictif du modèle obtenu a été confirmé par la validation croisée de la procédure LOO.

Cette étude montre que ce modèle est défini avec quatre descripteurs moléculaires non corrélés entre eux et qui sont essentiellement de type électroniques et topologiques. Par conséquent, le modèle pourrait être utilisé pour prédire l'activité anticancéreuse de nouvelles molécules à base des pyrazoles pour lesquelles les valeurs expérimentales sont indisponibles dans la littérature. La conclusion la plus importante de cette recherche est que nous ayant conçu et proposé quinze nouveaux composés avec des valeurs d'activités plus élevées que celles existantes en ajoutant des substituants appropriés et en calculant leurs activités en utilisant les équations de régression. Par conséquent, les modèles proposés réduiront le temps et le coût de la synthèse, ainsi que la détermination des activités anticancéreuses des dérivés pyrazoliques.

Références

1 C. Mowbray, S. Braillard, W. Speed, P. Glossop, G. Whitlock, K. Gibson, J.E.Mills, A.D. Brown, J. Gardner, Y. Cao, W. Hua, G.L.Morgans, P. Feijens, A. Matheeussen, and L.J. Maes, "Novel Amino-pyrazole Ureas with Potent In Vitro and In Vivo Antileishmanial Activity", Journal of Medicinal Chemestry, 58, 9615−9624, **2015**.

2 V. K. Aggarwal , J.Vicente, and R.V. Bonnert, "A Novel One-Pot Method for the Preparation of Pyrazoles by 1,3-Dipolar Cycloadditions of Diazo Compounds Generated in Situ", J. Org. Chem., 2003, 68, 5381–5383, **2003**.

3 A. Jamwal, A. Javed, V. Bhardwaj "A review on Pyrazole derivatives of pharmacological potential", Journal of Pharmaceutical and BioScience, 24, 2321-0125, **2013**.

4 G. Nitulescu, C. Draghici and O.T. Olaru, "New Potential Antitumor Pyrazole Derivatives: Synthesis and Cytotoxic Evaluation", International Journal of Molecular Sciences, 14, 21805-21818, **2013**.

5 A. M. Dar, and A.Shamsuzzaman, A Concise Review on the Synthesis of Pyrazole Heterocycles J Nucl Med Radiat Ther , 6, 6720-6739, **2015.**

6 D. Pal, S. Saha, S. Singh, "Importance Of Pyrazol Moiety in the Field of Cancer", International Journal of Pharmacy and Pharmaceutical Sciences, 7, 41-72, **2012**.

7 R. Aggarwal, V. Kumar, R. Kumar and S. P. Singh, Approaches towards the synthesis of 5-aminopyrazoles, Beilstein J. Org. Chem., 7, 179–197, **2011**.

8 K. Du, Y. Mei, X. Cao, P. Zhang, and H. Zheng, The Synthesis of Pyrazole Derivatives Based on on amino derivatives of indole as potent isoprenylcysteine carboxyl methyltransferase (Icmt) inhibitors, Journal of Molecular Structure , 1081, 466–476, **2015**.

11 A. Alafeefy, A. Ashour, O. Prasad, L. Sinha, S. Pathak, F. Alasmari, A. Rishi, H. Abdel-Aziz, "Development of certain novel N-(2-(2-(2-oxoindolin-3-ylidene) hydrazinecarbonyl)phenyl)-benzamides and 3-(2-oxoindolin-3-ylideneamino)-2-substituted quinazolin-4(3H)-ones as CFM-1 analogs: Design, synthesis, QSAR analysis and anticancer activity", European Journal of Medicinal Chemistry, 92, 191-201, **2015**.

12 N. Hernández, R. Kiralj, M. Ferreira , I. Talavera, Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors, Chemometrics and Intelligent Laboratory Systems, 98, 65–77, **2009**.

13 K. Roy, I. Mitra, P. Ojha, S. Kar, R. Das, H. Kabir,Introduction of rm 2 (rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models Chemometrics and Intelligent Laboratory Systems, 118, 200–210, **2012**.

14 A. Worachartcheewan , P. Mandi, V. Prachayasittikul, A. Toropova, A. Toropov, C. Nantasenamat, Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors, Chemometrics and Intelligent Laboratory Systems,138, 120–126, **2014**.

15 D. Dimić, G. Mercader, E. Chalcone,"derivative cytotoxicity activity against MCF-7 human breast cancer cell QSAR study", Chemometrics and Intelligent Laboratory Systems, 146, 378–384, **2015**.

16 R. Sabet, M. Mohammadpour, A. Sadeghi, A. Fassihi, "QSAR study of isatin analogues as in vitro anti-cancer agents" European Journal of Medicinal Chemistry, 45, 1113–1118, **2010**.

17 V. Prachayasittikul, R. Pingaew, A. Worachartcheewan, C. Nantasenamat, S. Prachayasittikul, S. Ruchirawat, V. Prachayasittikul, Synthesis, anticancer activity and QSAR study of 1,4-naphthoquinone derivatives, European Journal of Medicinal Chemistry,84-, 247-263, **2014**.

18 M. Irfan, B. Aneja, U. Yadava, S. Khan, N. Manzoor, C. Daniliuc, M. Abid, Synthesis, QSAR and anticandidal evaluation of 1,2,3-triazoles derived from naturally bioactive scaffolds, European Journal of Medicinal Chemistry, 93, 246-254, **2015**.

19 B. Chena, T.. Zhanga, T. Bondb, Y. Gan, Development of quantitative structure activity relationship (QSAR) model for disinfection byproduct (DBP) research: A review of methods and resources, Journal of Hazardous Materials, 299, 260–279, **2015**.

20 T. Chen, J. Guh, H. Hsu, C. Chen, C. Lee, C. Wu, Y. Lee, J. Lin, D. Yu, H. Huang, Synthesis and biological evaluation of anthra [1,9-cd]pyrazol-6(2H)-one scaffold derivatives as potential anticancer agents, Arabian Journal of Chemistry, 91, 205-235,2015.

Chapitre IV

Etude QSAR des dérivés de la 9-chloro-llH-indéno [1,2-c] quinoléine-11-one (dérivés de l'azafluorénone tétracyclique)

Résumé

Nous avons procédé à l'étude des structures chimiques de 29 dérivés 9-chloro-11H-indeno[1,2-c]quinoleine-11-one au moyen des descripteurs électroniques et physico-chimiques. La théorie fonctionnelle de la densité (DFT) avec les trois paramètres de Beck se basant sur les calculs fonctionnels de corrélation LYP (B3LYP / 6-31G) a été réalisée dans l'objectif de nous informer sur la relation structure activité anticancéreuse pour les 29 composés étudiés.

La présente étude a été réalisée par le biais des méthodes suivantes : la méthode d'analyse en composantes principales (ACP), la méthode de régression linéaire multiple (RLM), la méthode des moindres carrés partiels (PLS) et la régression non linéaire multiple (RNLM). La qualité statistique des modèles RLM, PLS et RNLM a été jugée efficace pour la prédiction de l'activité anticancéreuse, or le modèle de RLM est le meilleur et le plus pertinent car il a fourni des résultats statistiquement significatifs et a montré une bonne stabilité interne et une puissante prévisibilité de nouveaux composés. cette efficacité est démontrée par les coefficients de corrélation qui sont respectivement de 0,923, 0,89 et 0,91 pour les modèles RLM, PLS et RNLM. Les modèles de prédiction obtenus ont été confirmés par la méthode de validation croisée et le test de randomisation (ou Y-randomisation). La forte relation entre les valeurs des activités expérimentales et prédites par le modèle est discernable, de ce fait la bonne qualité du modèle QSAR a été développée. A partir des trois modèles QSAR obtenus, nous avons pu prédire dix nouveaux composés ayant le même motif de base la même activité et des données expérimentales indisponibles.

1. INTRODUCTION

Après une période plus longue ou plus courte, certaines cellules cancéreuses peuvent échapper à leur tumeur originale et se déplacer vers d'autres parties du corps via des vaisseaux sanguins ou des lymphatiques. Ces colonies «secondaires» portent le nom de métastases. Le processus de cancérologie est habituellement très important. Il peut s'étendre sur plusieurs années ou même des dizaines après le premier dommage cellulaire[1-2].

Les progrès scientifiques des dernières années permettent maintenant de déchiffrer le code génétique du cancer et de comprendre comment cette maladie est liée aux mécanismes de la vie elle-même. Pour cette raison, la maladie du cancer ne sera probablement jamais complètement éradiquée, et la recherche de base redirige vers l'industrie pharmaceutique qui fournit une pléthore de molécules de plus en plus ciblées, de plus en plus efficaces ... et toujours plus coûteuses, afin de les combiner dans la fabrication des médicaments[3-6].

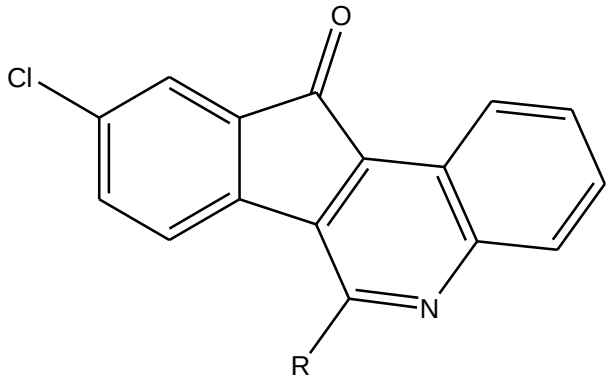


Figure 1 : Structure générale de l'azafluorénone tétracyclique

De nombreux dérivés d'azafluorénones tétracycliques ont été signalés pour montrer finalement dans l'industrie pharmaceutique des propriétés antitumorales importantes (figure 1 ci-dessus). La quinoléine est un échafaudage pharmacologiquement précieux qui prévaut dans une variété de composés synthétiques et naturels biologiquement actifs. Tout au long du 20^{éme} siècle, la chimie des quinoléines a fait l'objet d'études intensives et de bioactivité intéressante comme les activités antibactériennes[7], antifongiques[8], anti-inflammatoires[9], antipaludiques[10] et anticancéreuses[11]. L'activité anticancéreuse est assez large, les dérivés de la quinoléine ayant été utilisés contre de nombreux cancers, tels que ceux du sein, de la prostate, du tractus gastro-intestinal, du côlon et du foie. Il est important de noter qu'un certain nombre de médicaments anticancéreux à base de quinoléine ont également été utilisés cliniquement, y compris la camptothécine et ses analogues (irinotecan et topotecan)[12-15].

Des études QSAR ont été rapportées dans le choix des caractéristiques structurales importantes de l'activité anticancéreuse. Les relations quantitatives structure-activité (QSAR)

sont certainement des facteurs importants dans la conception des médicaments contemporains[16]. Par conséquent, il est claire pourquoi de nombreux utilisateurs QSAR sont situés dans des unités de recherche industrielle[17]. Les QSAR classiques sont donc des domaines de recherche très actifs dans la conception de médicaments. La base de différentes méthodes quantitatives de relation structure-activité (QSAR) est la «description» des structures moléculaires en utilisant des nombres importants de descripteurs afin de coder l'information chimique sous forme d'équation mathématique[18-19]. À l'heure actuelle, il existe un grand nombre de descripteurs moléculaires qui peuvent être utilisés dans les études QSAR. Dans cette étude, les analyses RLM, RNLM et PLS sont appliquées à une série de structures de l'azafluorenone tétracyclique, pour la conception d'un modèle QSAR fiable qui prédit l'activité anticancéreuse des dérivés de la quinoléine.

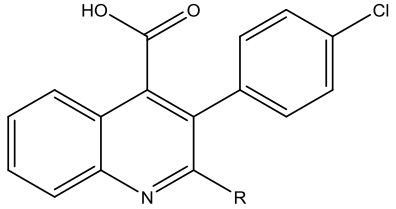
2. MATERIEL ET METHODES

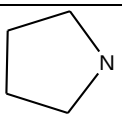
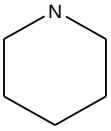
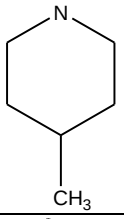
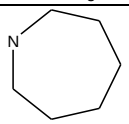
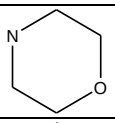
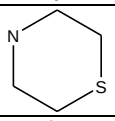
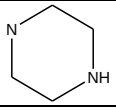
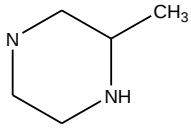
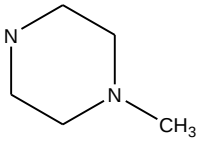
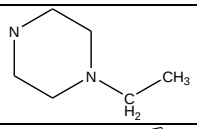
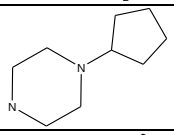
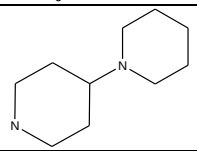
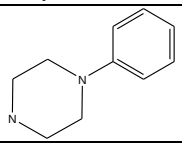
2.1. Données expérimentales

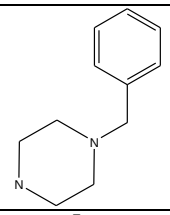
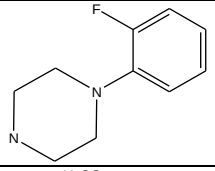
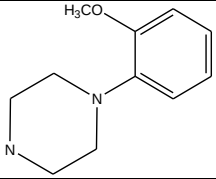
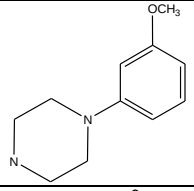
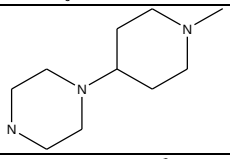
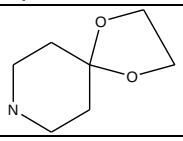
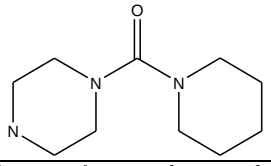
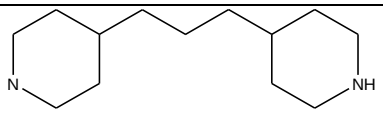
Dans la présente étude, nous avons opté pour 29 substituants de dérivés d'azafluorénone tétracyclique ayant des propriétés anticancéreuses potentielles, leurs activités sont rapportées dans la littérature par T.-C. Chen et al[20].

Pour l'étude QSAR, les valeurs déclarées de IC_{50} ont été converties en pIC_{50} en prenant un logarithme négatif ($pIC_{50} = - \log_{10} IC_{50}$), ensuite utilisées comme variables dépendantes pour le développement du modèle 2D-QSAR. La figure 1 ci-dessus représente la structure basique des dérivés de l'azafluorénone tétracyclique et le tableau 1 montre les substituants des composés étudiés et leurs activités expérimentales correspondantes pIC_{50}.

Tableau 1: Structures des composés étudiés et leurs activités anticancéreuses

Composés	R	pIC_{50}
	-	1,5
2	Cl	1,5
3	$NHCH_3$	1,37
4	$N(CH_3)_2$	1,14
5	$NHCH_2CH_2N(C_2H_5)_2$	0,59

6		1,52
7		1,51
8		1,52
9		1,47
10		0,86
11		0,96
12		0,18
13		0,22
14		0,91
15		0,99
16		1,29
17		0,77
18		0,56

19		1,49
20		1,03
21		1,52
22		1,11
23		0,26
24		0,78
25		1,5
26		1,34
27	SCH_2CH_2OH	0,3
28	OH	1,5
29	OCH_3	1,52

2.2. Méthodes de calculs

Dans cette recherche, nous avons utilisé la méthode quantique DFT(Théorie Fonctionnelle de la Densité) pour mettre en œuvre les différentes caractéristiques physico-chimiques des molécules en question, et rechercher leur géométrie la plus stable et les mettre en corrélation avec l'activité étudiée. Les structures 3D des molécules ont été générées à l'aide de "GAUSS VIEW 4.1" et tous les calculs quantiques de tous les composés ont été effectués en faisant appel au logiciel "GAUSIAN 03". La géométrie d'optimisation et les propriétés physico-

chimiques des 29 composés ont été prédites en utilisant la fonction B3LYP associée à six ensembles de base 31G.

2.3. Calcul des descripteurs moléculaires

Les molécules optimisées ont été utilisées pour calculer un certain nombre de descripteurs électroniques : le moment dipolaire (DM), l'énergie des orbitales frontières (E_{HOMO}, E_{LUMO}), l'énergie totale (E_{totale}) et l'énergie de répulsion RE. En outre nous avons calculé les descripteurs suivants: Poids moléculaire (MW), coefficient de partage (logP), l'accepteur de liaison hydrogène (HA) et le donneur de liaison hydrogène(HD) en nous servant du programme CHEMBIO OFFICE (2015).

Par ailleurs le programme ChemSketch nous a permis de calculer les paramètres suivants: Volume Molaire (MV (cm^3)), Réfractivité Molaire (MR (cm^3)), Parachor (Pc (cm^3)), Densité (g / cm^3), Indice de Réfraction, Tension Superficielle (Dyne / Cm) et la Polarisabilité (cm^3).

2.4. Analyses statistiques :

L'interprétation de l'activité structurale des 29 molécules étudiées nécessite l'utilisation de 16 descripteurs qui sont calculés à l'aide des logiciels Gaussian 03, chemoffice 2015 et chemscketch. Les démarches statistiques mises en œuvre dans cette étude sont :

- L'analyse en composantes principales (ACP) qui a pour but d'examiner la multicolinéarité entre les descripteurs, est conçue par utilisation du logiciel XLSTAT, version 2014.
- La technique statistique de régression linéaire multiple (RLM) a pour objectif l'étude de la relation entre une variable dépendante et plusieurs variables indépendantes (descripteurs). Elle sert également à sélectionner les descripteurs utilisés comme paramètres d'entrée pour les méthodes PLS et RNLM.
- La méthode de régression des moindres carrés partiels (PLS) est une méthode efficace pour identifier les critères basés sur la covariance.
- La technique de validation croisée suivie du test de Y-randomisation ont été utilisés pour valider les modèles obtenus.

3. RESULTATS

3.1. Base des données

L'étude QSAR a été focalisée sur la série de 29 substituants aux dérivés de l'azafluorénone tétracyclique, afin de déterminer une relation quantitative entre la structure et les activités anticancéreuses.

Les valeurs des 16 descripteurs sont indiquées dans le tableau 2 et les résultats de QSAR obtenus par l'intermédiaire des modèles RLM, PLS, RNLM sont représentés dans les tableaux 3 et 4.

Tableau 2 : valeurs des descripteurs calculés pour les molécules étudiées

Composés	E_{LUMO}	E_{HOMO}	E_{TOTALE}	RepE	MD	MW	LogP	DH	AH	MR	MV	Pc	IR	ST	Dsty	Pls
1	-2,19	-6,52	-36899,5	45212,14	2,891	299,7	4,62	2	3	80,48	204,9	591,6	1,71	69,4	1,462	31,9
2	-3,47	-6,65	-45279,4	43424,44	1,823	300,1	4,95	0	2	79,65	196	564,3	1,75	68,6	1,53	31,6
3	-3,06	-5,81	-35349,6	44006,73	4,108	294,7	4,61	1	3	84,36	203,5	593,7	1,767	72,3	1,447	33,4
4	-3,05	-5,85	-36418,7	48262,38	3,896	308,8	5,17	0	3	89,06	222,1	631,8	1,734	65,4	1,39	35,3
5	-2,91	-5,57	-42202,2	67646,78	6,153	379,9	5,34	1	4	111,3	293,3	820,4	1,683	61,1	1,294	44,1
6	-2,98	-5,68	-38524,7	55252,63	4,808	334,8	5,66	0	3	94,95	238,6	682,5	1,727	66,8	1,402	37,6
7	-3,1	-6,11	-39594,1	59257,99	3,780	384,8	6	0	3	99,57	256,3	722,5	1,704	63	1,36	39,5
8	-3,1	-6,11	-40663,7	63061,9	3,826	362,9	6,52	0	3	104,3	276,3	760,6	1,678	57,3	1,312	41,3
9	-3,03	-5,70	-40663,4	64280,90	4,072	362,9	6,34	0	3	104,2	274,1	762,6	1,684	59,9	1,323	41,3
10	-3,21	-6,34	-40570,4	59383,10	2,079	350,8	4,8	0	4	96,56	274,1	702,8	1,708	65,1	1,417	38,3
11	-3,23	-6,18	-49358,7	62808,84	2,043	366,9	5,94	0	3	103,1	255,8	734,9	1,739	68	1,433	40,9
12	-3,04	-5,81	-40030	59584,46	4,121	349,8	4,6	1	4	9,3	252,4	710,7	1,706	62,8	1,385	39
13	-3,03	-5,8	-41099,6	63538,07	4,183	363,8	5,07	1	4	103,0	272,4	748,8	1,68	57	1,335	40,8
14	-3,05	-5,77	-41099,4	63470,05	3,913	363,8	4,85	0	4	103,2	267,4	748,9	1,698	61,4	1,36	40,9
15	-3,05	-5,76	-42169	67252,24	4,030	377,9	5,28	0	4	107,9	285	788,9	1,681	58,7	1,325	42,8
16	-3,03	-5,54	-45344,1	78626,02	4,240	417,9	6,06	0	4	119,6	309,8	877,6	1,699	64,4	1,348	47,4
17	-3,05	-5,79	-46849,1	83063,97	4,285	431,9	6,4	0	4	124,2	327,5	917,7	1,683	61,6	1,318	49,3
18	-3,12	-5,37	-46315,5	79265,45	2,811	425,9	7,03	0	4	123	311,6	882,7	1,719	64,3	1,366	48,5
19	-3,05	-5,85	-47384,9	83357,84	3,843	439,9	6,57	0	4	127,7	325,7	922,7	1,712	64,3	1,35	50,6
20	-3,10	-5,61	-49015	84201,35	4,075	443,9	7,19	0	5	123	315,8	890	1,706	63	1,405	48,7
21	-3,02	-5,21	-49430,5	64354,37	4,515	455,9	6,93	0	5	129,6	335,6	941,3	1,699	61,8	1,358	51,4
22	-3,095	-5,32	-49430,5	87278,8	4,136	455,9	6,93	0	5	129,6	335,6	941,3	1,699	61,8	1,358	51,4
23	-3,049	-5,34	-47918,7	86878,81	3,808	446,9	5,25	0	5	127,9	338,6	944	1,679	60,4	1,32	50,7
24	-3,023	-5,85	-45792	75578,50	5,365	406,9	5,26	0	5	110,0	276	810,2	1,728	74,1	1,47	43,6
25	-3,123	-6,01	-49933,3	91085,48	5,609	460,9	5,22	0	4	127,7	332,2	951,3	1,695	67,1	1,387	50,6
26	-3,024	-5,54	-49622,3	91700,42	4,721	474	7,87	1	4	137,9	388,9	1038	1,627	50,6	1,218	54,7
27	-3,213	-6,31	-47792,7	54243,14	3,423	341,8	5,79	1	4	93,14	225,1	674,9	1,765	80,8	1,51	36,9
28	-3,206	-6,39	-34820,9	40415,01	3,470	281,7	4,99	1	4	76,63	182,5	542,3	1,78	77,9	1,543	30,4
29	-3,124	-6,31	-35889,7	44101,97	4,085	295,7	5,51	0	4	81,43	208,1	585,7	1,711	62,7	1,42	32,3

3.2. Analyse en composantes principales

La totalité des 16 descripteurs codant les 29 molécules est soumise à une analyse en composantes principales (ACP), 16 composantes principales ont été obtenus (figure 2).

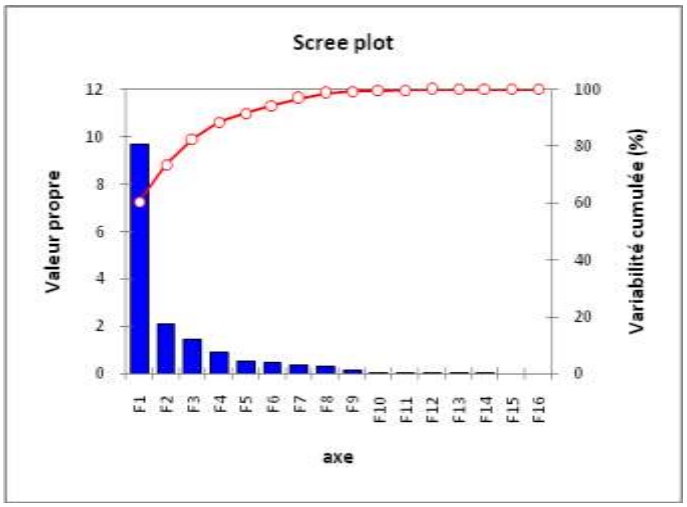


Figure 2 : Composantes principales et leurs variances

Les trois premiers axes principaux sont suffisants pour décrire les informations fournies par la matrice de données. En effet, les pourcentages de variances sont respectivement de 60,37% 13,09% et 9,09% pour les axes F1, F2 et F3. L'information totale est estimée à un pourcentage de 82,55%. Le tableau 3 montre la matrice de corrélation (Pearson (n)) ainsi obtenue entre les différents descripteurs.

Tableau 3: La matrice de corrélation

Variab.	E_{LUMO}	E_{HOMO}	E_{TOTALE}	RepE	MD	MW	LogP	DH	AH	MR	MV	Parc	IR	ST	Dsity	Pl
E_{LUM}	**1**															
E_{HOM}	0,065	**1**														
E_{TOTA}	0,257	-0,41	**1**													
RepE	-0,04	0,647	-0,79	**1**												
MD	0,218	0,519	-0,08	0,390	**1**											
MW	-0,05	0,696	-0,84	0,954	0,382	**1**										
LogP	-0,13	0,508	-0,61	0,604	0,126	0,685	**1**									
DH	0,565	-0,26	0,333	-0,34	0,037	-0,37	-0,32	**1**								
AH	0,017	0,582	-0,48	0,599	0,413	0,637	0,294	-0,13	**1**							
MR	-0,03	0,757	-0,81	0,954	0,405	**0,987**	0,698	-0,35	0,622	**1**						
MV	-0,01	0,724	-0,76	0,942	0,391	**0,967**	0,678	-0,32	0,596	**0,984**	**1**					
Parc	-0,01	0,739	-0,80	0,959	0,429	**0,986**	0,683	-0,32	0,619	**0,997**	**0,989**	**1**				
IR	-0,23	-0,48	0,330	-0,64	-0,39	-0,62	-0,43	0,120	-0,28	-0,65	-0,75	-0,69	**1**			
ST	-0,09	-0,5	0,178	-0,47	-0,23	-0,47	-0,43	0,216	-0,12	-0,53	-0,62	-0,54	0,911	**1**		
Dsity	-0,17	-0,67	0,270	-0,63	-0,45	-0,62	-0,5	0,178	-0,23	-0,69	-0,76	-0,71	0,906	0,897	**1**	
Pl	-0,03	0,756	-0,81	0,954	0,407	**0,987**	0,697	-0,35	0,622	**1,000**	**0,984**	**0,998**	-0,65	-0,53	-0,69	**1**

L'analyse en composantes principales (ACP) a été menée pour identifier le lien entre les différentes variables. Les corrélations entre les seize descripteurs sont présentées dans le tableau 3 en tant que matrice de corrélation.

La matrice obtenue fournit des informations sur l'interrelation haute ou basse entre les variables. En

général, la co-linéarité (r> 0,5) a été observée entre la plupart des variables, et l'activité *p*IC_{50}. Une corrélation élevée a été observée entre :

MV et MW (r = 0,967), Le parachoc et le MW (r = 0,986), La polarisation et le MW(r = 0,974), MR et Parachoc (r = 0,997), MR et MV (r = 0,984), La MR et la polarisabilité(r = 1)

MV et Parachoc (r = 0,989), MV et Polarisability (r = 0,984), Parachoc et Polarisability (r = 0,998), MR et MW (r = 0,987).

De plus, pour diminuer la redondance existante dans notre matrice de données, les descripteurs fortement corrélés (R ≥ 0,96) ont été exclus.

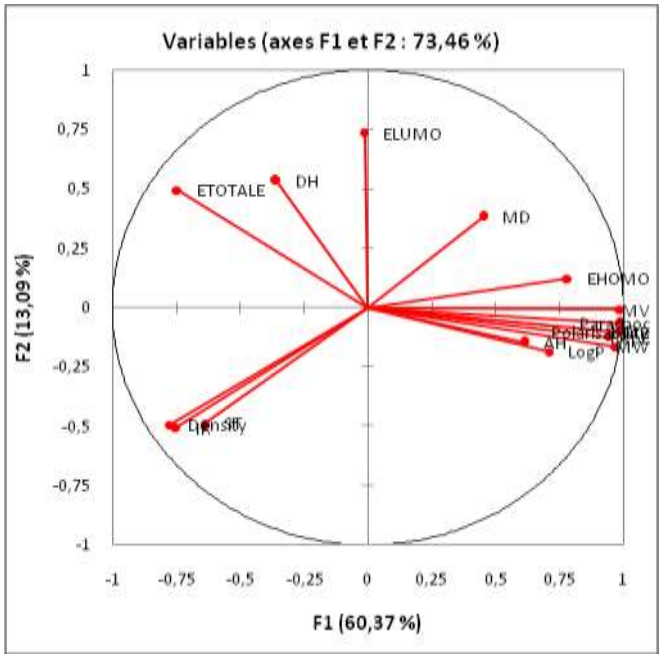


Figure 3 : cercle de Corrélation entre les descripteurs

3.3. Régression linéaire multiple

Afin de sélectionner les descripteurs prédominants qui affecteront les activités anticancéreuses des composés étudiés, l'analyse de corrélation a été effectuée avec le logiciel statistique XLSTAT 2014 prenant chaque descripteur calculé en tant que variable indépendante et *p*IC_{50} comme variable dépendante. Sur la base de l'analyse de corrélation, la méthode de régression linéaire multiple a été utilisée pour établir le modèle QSAR. Cependant, cette méthode utilise les coefficients R, R^2, $R^2_{ajusté}$, MSE, MAE et F-values afin de sélectionner la meilleure performance de régression, où R est le coefficient de corrélation; R^2 est le coefficient de détermination, MSE est l'erreur quadratique moyenne, MAE est l'erreur absolue moyenne et F est la statistique Fisher F. Etotale, l'accepteur de liaison d'hydrogène (HA), la densité et (LogP) sont les descripteurs qui dépendent de l'activité anticancéreuse des dérivés d'azafluorénone tétracycliques.

Le modèle QSAR obtenu en utilisant la méthode de régression linéaire multiple (RLM) est représenté par l'équation suivante :

$$\mathbf{\textit{p}IC_{50}= -3,8241 + 0,00004*E_{TOTALE} + 0,50342*LogP - 0,10244*AH + 3,246*Densité}$$

(équation1)

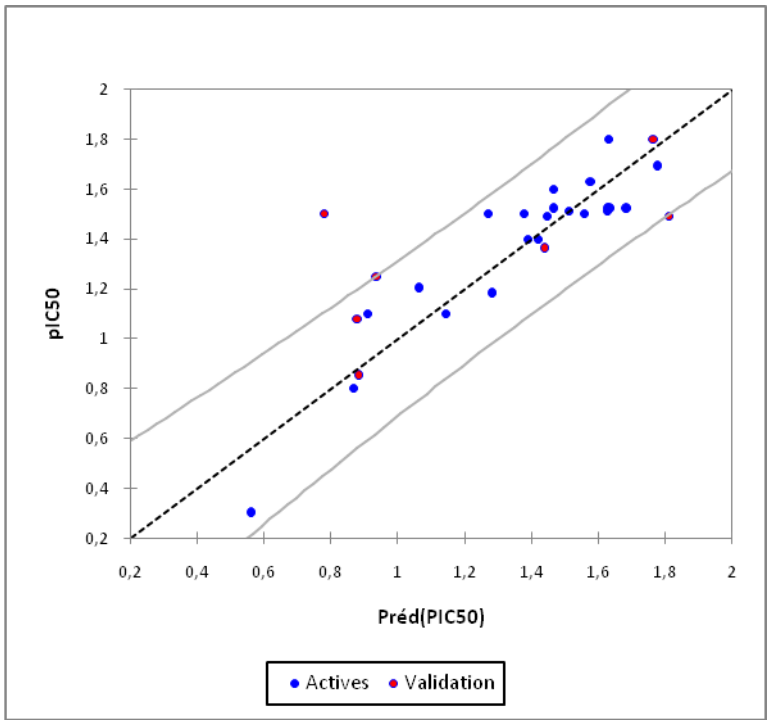


Figure 4 : Activités anticancéreuses prédites *p*IC$_{50}$ par la méthode RLM par rapport aux valeurs expérimentales (série d'apprentissage en bleu et la série de test en rouge).

Les caractéristiques statistiques de l'équation obtenue sont :

N = 22, R = 0,923, R^2 = 0,852, $R^2_{ajusté}$ = 0,82, MSE= 0,038, MAE=0,042, F = 24,434

$t_{Etotale}$ = 4,401 t_{logP} = 8,79 t_{AH} = -2,228 $t_{Density}$ = 5,781

Comme indiqué dans l'équation ci-dessus, les descripteurs les plus significatifs qui affectent l'activité anticancéreuse et font partie dans l'établissement du modèle QSAR sont : les descripteurs électroniques (Etotale) et les descripteurs stériques (densité, AH et LogP).

3.4. Méthode des moindres carrés partiels

La PLS a deux objectifs : approximer la matrice X des descripteurs de la structure moléculaire à la matrice Y des variables dépendantes et maximiser la corrélation entre elles. Nous avons proposé la matrice de données constituée clairement à partir des descripteurs retenus par la RLM correspondant aux 29 molécules, pour la méthode des moindres carrés partiels (PLS).

Cette méthode utilise les coefficients R, R^2 et les valeurs F pour sélectionner la meilleure performance de régression. l'équation obtenue par la méthode PLS est:

pIC_{50} = -2,0223 + 0,00003*E_{totale} + 0,401*LogP – 0,149*AH + 2,135*Density (équation 2)

N = 22 ; R = 0,89 ; R^2 = 0,79 ; MSE= 0,119 ; MAE= 0,23

La figure 5 montre une répartition très régulière des valeurs des activités anticancéreuses prédites par le modèle PLS en fonction des valeurs expérimentales.

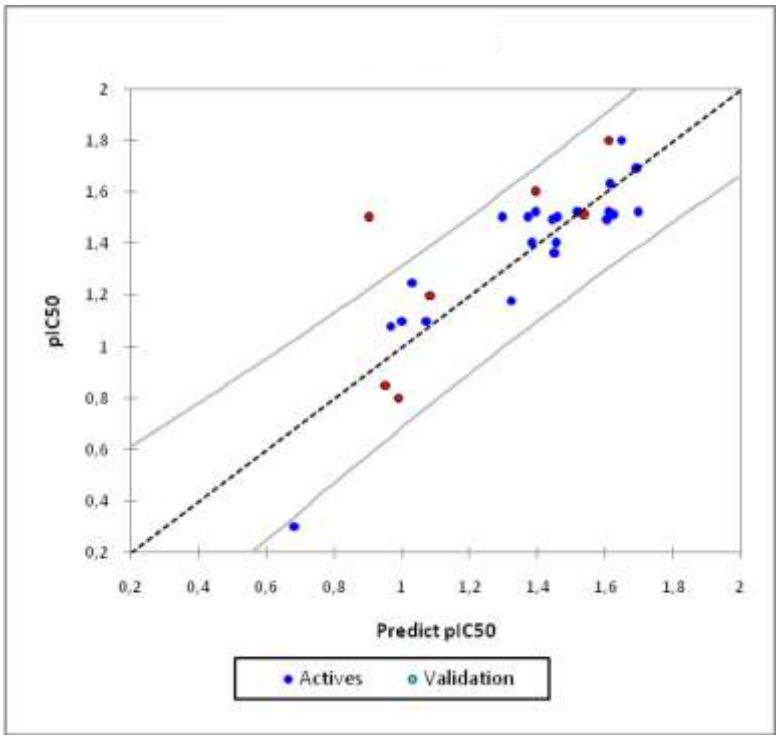


Figure 5 : Représentation graphique des pIC_{50} calculées et prédites par la méthode PLS.

Le coefficient de corrélation obtenu dans l'équation (2) est assez intéressant (0,89). Pour améliorer l'activité anticancéreuse de manière quantitative, compte tenu de plusieurs paramètres, nous avons utilisé la technique du modèle de régression non linéaire.

3.5. Régression non linéaire multiple (RNLM)

Les descripteurs de base correspondant aux 29 composés appliqués à la matrice de données pour la RNLM sont ceux retenus par la RLM. Les coefficients R, R^2, l'erreur absolue moyenne MAE et l'erreur quadratique moyenne MSE sont utilisés pour sélectionner la meilleure performance de la régression non linéaire. L'équation résultante est :

$$pIC_{50} = -23,72 + 0,00006*E_{totale} - 1,086*LogP + 0,571*AH + 36,99*Densité + 5,66\ 10^{-10}*(E_{totale})^2 +0,121*(LogP)^2 - 0,0972*(AH)^2 - 12,153*(Densité)^2 \quad \text{(équation 3)}$$

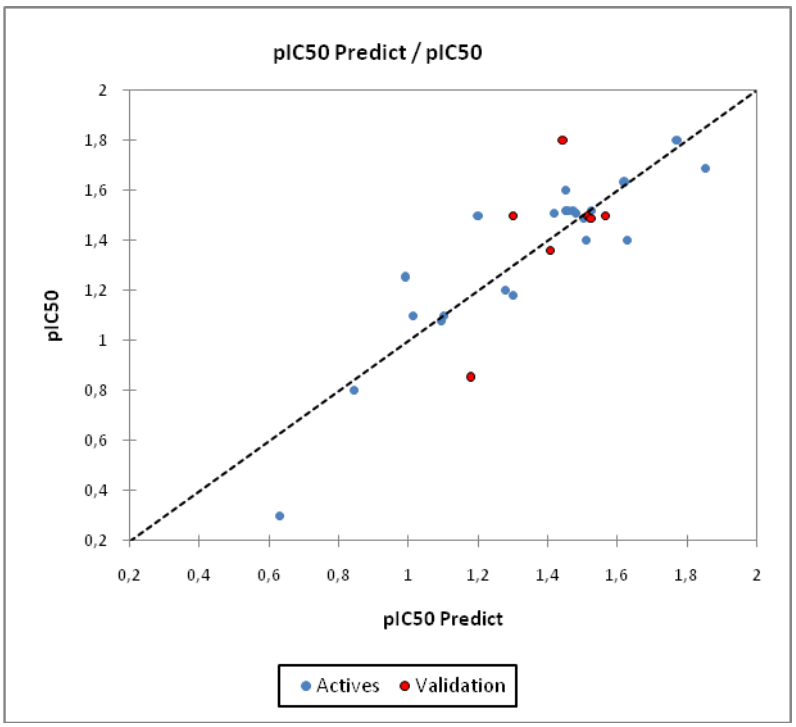


Figure 6 : Représentation graphique des pIC_{50} calculés et observés avec RNLM

N = 22 ; R = 0,91 ; R^2 = 0,82 ; MSE= 0,341 ; MAE= 0,497

L'importance du coefficient de corrélation obtenu dans l'équation (0,91) confirme l'efficacité du modèle de la RNLM à décrire l'activité anticancéreuse. Nous pouvons dire que les valeurs obtenues à partir de la régression non linéaire sont fortement corrélées avec celles de l'activité observée. La Figure 6 montre une répartition régulière des valeurs prédites en fonction des valeurs expérimentales.

Tableau 4 : Comparaison entre les activités observées et prédites des modèles statistiquement significatifs obtenus par un ensemble de formation des modèles 2D-QSAR.

Comp.	pIC_{50}	Préd(pIC_{50}) RLM	Résidu	Préd(pIC_{50}) RNLM	Résidu	Préd(pIC_{50}) PLS	Résidu
1	1,50	1,376	0,124	1,419	0,091	1,371	0,129
2	1,51	1,511	-0,001	1,512	-0,112	1,383	0,017
3	1,4	1,388	0,012	0,844	-0,044	1,453	-0,093
5	0,8	0,866	-0,066	1,526	-0,006	1,610	-0,090
6	1,5	1,636	-0,116	1,481	0,029	1,624	-0,114
7	1,51	1,626	-0,116	1,476	0,044	1,697	-0,177
8	1,52	1,686	-0,166	1,280	-0,080	1,649	0,151
9	1,8	1,632	0,168	1,629	-0,229	1,456	-0,056
10	1,2	1,063	0,137	1,013	0,087	1,003	0,097
11	1,4	1,419	-0,019	1,094	-0,014	0,968	0,112
13	1,100	0,910	0,190	0,992	0,258	1,033	0,217

16	1,500	1,271	0,229	1,301	-0,121	1,297	0,203
17	1,180	1,281	-0,101	1,853	-0,163	1,323	-0,143
18	1,690	1,777	-0,087	1,775	0,025	1,695	-0,005
19	1,490	1,448	0,042	1,452	0,068	1,443	0,047
21	1,520	1,466	0,054	1,452	0,148	1,393	0,127
22	1,600	1,466	0,134	0,631	-0,331	0,685	-0,385
23	0,300	0,561	-0,261	1,102	-0,002	1,075	0,025
24	1,100	1,143	-0,043	1,203	0,297	1,614	0,016
26	1,630	1,579	0,051	1,624	0,006	1,459	0,041
27	1,500	1,557	-0,057	1,503	-0,013	1,608	-0,118
29	1,520	1,629	-0,109	1,461	0,059	1,521	-0,001

Tableau 5:Comparaison entre les activités observées et prédites de modèles statistiquement significatifs obtenus par un ensemble de tests des modèles 2D-QSAR.

COMP.	***p*IC$_{50}$**	***p*IC$_{50}$RLM test**	***p*IC$_{50}$ RNLM test**	***p*IC$_{50}$ PLStest**
4	1,360	1,440	1,514	1,540
12	0,850	0,881	1,407	0,990
14	1,080	0,881	1,446	1,086
15	1,250	0,938	1,180	0,954
20	1,800	1,767	1,300	1,611
25	1,500	0,780	1,528	1,393
28	1,490	1,811	1,565	0,902

3.6. Validation

La détermination de la capacité prédictive du modèle développé nécessite une confirmation au moyen de la procédure de la validation interne, et l'approche de la validation croisée (Leave-One-Out).

N=22 ; R=0,942 ; MSE= 0,24 ; MAE=0,291 ; SD=0,1071 ; P<0,0001

Une corrélation parfaite a été constaté à l'aide de la validation croisée R_{vc} = 0,91 donc le pouvoir prédictif de ce modèle est très important.

Le résultat le plus important de cette recherche est que l'activité anticancéreuse peut être prédite en utilisant des modèles QSAR obtenues par les trois modèles. Les résultats de la validation croisée affirment que la RLM est la meilleure méthode pour construire les modèles de la relation quantitative structure d'activité, aussi le modèle proposé dans cette étude indique une capacitée prédictive élevée et significative.

3.7.Y-Randomisation

Cette procédure consiste à réorganiser de manière aléatoire la propriété du modèle dans le jeu d'origine et à recréer des modèles de validation croisée. Si la valeur du coefficient de corrélation des molécules mélangées est inférieure à celle obtenue en appliquant la validation croisée, il existe donc une indépendance entre les molécules, car les points de mesure du point cible les plus proches

n'obscurcissent pas les autres données expérimentales et ne sont pas impliqués dans l'estimation. D'où les données utilisées dans cette validation sont réparties uniformément dans l'espace et le modèle résultant peut être extrapolé à l'ensemble de la série.

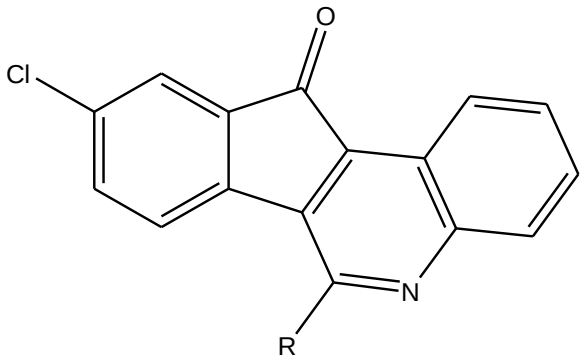
N = 22 ; R = 0,67 ; MSE= 0,248 ; MAE=0,321 ; SD = 0,1013 ; P <0,0001

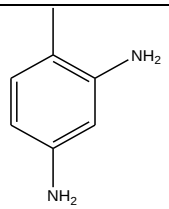
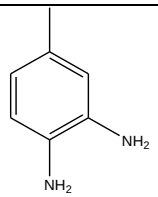
3.8. Molécules proposées

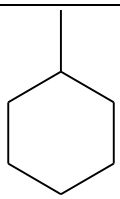
Les valeurs des paramètres obtenus par les calculs DFT pour les composés proposés sur la base des informations dérivées des équations 1, 2 et 3 sont représentées dans le tableau 6. Nous avons observé que le RNLM conçu avait des valeurs de pIC_{50} plus élevées que les modèles RLM et le PLS (Tableau6).

En outre, les composés (X1 et X10) ont des valeurs de pIC_{50} plus élevées que les composés étudiés.

Tableau 6 : Proposition de nouveaux composés



	R	**Etotale**	**AH**	**Density**	**LogP**	***p*IC$_{50}$ RLM**	***p*IC$_{50}$ RNLM**	***p*IC$_{50}$ PLS**
X1	$N(OH)_2$	-38411,48	5	1,616	2,98	0,873	1,113	0,725
X2	NH_2	-34318,15	3	1,411	2,88	0,526	1,599	0,668
X3	COOH	-37945,05	3	1,49	2,95	0,672	1,64	0,755
X4	NHOH	-36364,58	4	1,519	2,81	0,657	1,626	0,66
X5	$NHC(CH_3)_3$	-38600,28	3	1,253	4,06	0,436	0,499	0,675
X6		-42117,94	4	1,394	3,46	0,348	1,125	0,481
X7		-42117,94	4	1,394	3,46	0,348	1,125	0,481

X8		-39201,36	2	1,236	5,31	1,088	0,35	1,271
X9	$C(NH_2)_2C(CH_3)_2NH_2$	-41614,28	5	1,304	2,18	-0,671	09,66	-0,359
X10	$C(CH_3)_2C(CH_3)_3$	40304,3	2	1,129	6,59	4,566	4,74	3,941

4. DISCUSSIONS

Dans le cadre de l'étude statistique, nous présentons d'abord la construction d'un modèle 2D-QSAR linéaire (RLM) décrivant la relation quantitative structure-activité anticancéreuse de 29 molécules dérivées de la quinoléine et par utilisation de la méthode de régression non linéaire. Dans un deuxième temps, nous avons comparé les résultats obtenus par le modèle linéaire de la RLM, et ceux obtenus par le modèle non linéaire et le PLS. La régression non linéaire et la méthode PLS utilisées dans cette étude ont été générées en utilisant les quatre descripteurs qui sont apparus dans le modèle de la RLM. En effet, dans une première étape, nous avons divisé l'ensemble des molécules en deux sous-ensembles : un sous-ensemble d'apprentissage et un sous-ensemble de test. Le sous-ensemble d'apprentissage de 22 molécules a été utilisé pour générer les modèles 2D-QSAR. Le sous-ensemble de validation, composé de 7 molécules, a été utilisé pour la validation externe du modèle. Nos résultats montrent que la régression linéaire multiple est la meilleure base sur laquelle la relation quantitative structure-activité se construit et que l'activité anticancéreuse dépend des paramètres non-linéaires, ainsi le modèle proposé dans cette étude a un pouvoir prédictif élevé (R_{RLM} = 0,92). Les modèles 2D-QSAR élaborés (Eqs.1, 2 et 3) révèlent que l'activité anticancéreuse des dérivés de quinoléine pourrait s'expliquer par un certain nombre de facteurs électroniques et stériques. En effet, il semble que la combinaison des trois paramètres Etotale, densité et LogP augmente considérablement la puissance prédictive du modèle 2D-QSAR (Eq.1)

(N = 22 ; R = 0,92 ; R^2 = 0,852 ; $R^2_{Ajusté}$ = 0,82 ; F = 24,434 ; MSE = 0,0154 ; MAE = 0,104 ; P <0,0001). Le modèle QSAR obtenu peut représenter environ 92% de la variance expérimentale de la variable dépendante (IC_{50}) et a un Fischer F élevé (24,434) et un faible écart type (SD = 0,015), ce qui confirme que le modèle de la RLM explique l'activité anticancéreuse d'une manière satisfaisante et statistiquement significative. La distribution des résidus autour de la ligne zéro montre qu'il n'y a pas d'erreur systématique dans les modèles construits par la RLM, PLS et la RNLM.

D'après les valeurs du test t (| t |), l'importance des descripteurs impliqués dans ce modèle est dans l'ordre suivant : log P> densité> E_{totale}> AH. Le descripteur le plus important selon l'essai est le

coefficient de partage log P, ensuite le deuxième descripteur est la densité suivis par les accepteurs de liaisons hydrogène et enfin l'énergie totale.

Pour montrer que les résultats du modèle de validation croisée ne sont pas obtenus au hasard, un test de randomisation a été effectué. Les valeurs pIC_{50} de l'ensemble d'apprentissage ont été réorganisées de manière aléatoire en conservant les paramètres retenus de la régression linéaire inchangés. Cette opération est répétée trois fois, les résultats obtenus sont récapitulés dans la Figure 5. Les résultats montrent que les valeurs des coefficients de détermination de l'ensemble d'apprentissage calculés par le modèle généré sont les mêmes par rapport à celles de notre modèle, donc le modèle n'est pas obtenu par hasard.

5. CONCLUSION

Dans ce chapitre, les techniques RLM, RNLM et PLS ont été utilisées pour élaborer des modèles QSAR linéaires et non-linéaires afin de prédire l'activité anticancéreuse de 29 molécules dérivées de la quinoléine. En outre, les modèles obtenus se caractérisent par la bonne stabilité et le pouvoir prédictif élevé, vérifié par la validation interne qui est évidente à partir du coefficient de détermination R^2 et le coefficient de validation croisée R^2_{cv} et plus précisément de sa validation externe. Notre étude montre que, les descripteurs étudiés, qui sont suffisamment riches en informations chimiques, électroniques et topologiques pour coder la caractéristique structurelle, peuvent être utilisés avec d'autres descripteurs pour le développement des modèles 2D-QSAR prédictifs et peuvent être efficacement appliqués pour évaluer les activités anticancéreuses des dérivés quinoleines pour lesquels les données expérimentales ne sont pas disponibles. Nous concluons que le résultat le plus important de cette recherche est que nous avons pu concevoir et proposer de nouveaux composés avec des valeurs plus élevées ou plus faibles que les composés existants (tableau 6) en ajoutant des substituants appropriés en calculant leurs activités à l'aide des équations de régression. Par conséquent, les modèles proposés réduiront le temps et le coût de la synthèse.

Références

1 E. Esteve, D. Bazin, C. Jouanneau., S. Rouziere, A. Bataille, A. Kellum, K. Provost, C. Mocuta, S. Reguer, D. Thiaudiere, K. Jorissen, A. Hertig, E. Rondeau, E. Letavernier, M. Daudon, P. Ronco, How to assess the role of Pt and Zn in the nephrotoxicity of Pt anti-cancer drugs. An investigation combiningm XRF and statistical analysis: Part I: On mice. C. R. Chimie., 19, 1560-1585, **2016.**

2 K. Morgans, C. Bommel, C. Stowell, L. Abrahm, E. Basch, E.J. Bekelman, D. Berry, A. Bossi, I.Davis, T. Reijke, L. Denis, S. Evan, N. Fleshner, D. George, J. Kiefert, W. Lin, G. Matthe, R. McDermott, H. Payne, G. Roos, D. Schrag, T. Steuber, B. Tombal, J. Basten, M. Hoeven, F. Penson, Development of a Standardized Set of Patient-centered Outcomes for Advanced Prostate Cancer: An International Effort for a Unified Approach. Europran Urology, 6 8, 871 – 898, **2015.**

3 R. Siegel, D. Naishadham, A. Jemal.Cancer Statistics 2013. Cancer Journal for Clinicians., 63, 11–30, **2013**.

4 R. Siegel, K. Miller, A. Jemal, Cancer Statistics, Cancer Journal for Clinicians, 65,1-29, **2015**.

5 P. Wingo, C. Cardinez, S. Landis, Long-term trends in cancer mortality in the United States. Cancer, 97, 3133-3275, **2003.**

6 K. Morgans, C. Bommel, C. Stowell, L. Abrahm, E. Basch, J. Bekelman, D. Berry, A. Bossi, Davis I., Reijke T., Denis L., Evans S., Fleshner N., George D., Kiefert J., Lin W., Matthe G., McDermott R., Payne H., Roos G., Schrag D., Steuber T., Tombal B.,Basten J.,der Hoeven M., Penson F., Development of a Standardized Set of Patient-centered Outcomes for Advanced Prostate Cancer: An International Effort for a Unified Approach. European urology, 68,891-898, **2015.**

7 S. Pine, B. Ryan, L. Varticovski, A. Robles, C. Harris C., Microenvironmental modulation of asymmetric cell division in human lung cancer cells, Proceedings of the National Academy of Sciences, 107, 2175–2200, **2010**.

8 C. Lagadec, E. Vlashi, D. Donna, Y. Meng, C. Dekmezian, K. Kim, F. Pajonk, Survival and self-renewing capacity of breast cancer initiating cells during fractionated radiation treatment, Breast Cancer Res, 1, 112-139, **2010.**

9 J. McLaughlin, Paw paw and cancer: annonaceous acetogenins from discovery to commercial products, Rev. J. Nat. Prod., 71, 1311–1321, **2008.**

10 Z. Yuan, S. Chen, C. Chen, J. Chen, Q. Dai, C. Gao, Y Jiang, Structure-based hybridization, synthesis and biological evaluation of novel tetracyclic heterocyclic azathioxanthone analogues as potential antitumor agents. Rev. Eur. J. Med. Chem., 103, 605–627, **2015.**

11 A. Aoyagi, T. Kobunai, T. Utsugi, K. Wierzba, Y. Yamada, Establishment and characterization of 6-[[2- (dimethylamino) ethyl]amino]-3-hydroxy-7H-indeno[2,1-c]quinolin- 7-one dihydrochloride (TAS-103)-resistant cell lines. Rev. Jpn. J. Cancer Res., 91, 543–550, **2000.**

12 T. Chen, D. Yu, S. Chen, C. Chen, C. Lee, Y. Hsieh, L. Chang, J. Guh, J. Lin, H. Huang, Structure-based hybridization, synthesis and biological evaluation of novel tetracyclic heterocyclic azathioxanthone analogues as potential antitumor agents. Rev. Eur. J. Med. Chem. 103, 615–627**, 2015.**

13 N. Hernández, R. Kiralj, M. Ferreira, I. Talavera, Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors Chemometrics and Intelligent Laboratory Systems, 98, 65–77, **2009.**

14 K. Roy, I. Mitra, P. Ojha, S. Kar, R. Das, H. Kabir, Introduction of r_m^2 (rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models, Chemometrics and Intelligent Laboratory Systems, 118, 186–210, **2012.**

15 A. Worachartcheewan, P. Mandi, V. Prachayasittikul, A. Toropova, A. Toropov, QSAR study of aromatase inhibitors using SMILES-based descriptors, Chemometrics and Intelligent Laboratory Systems, 138, 101–126, **2014.**

16 D. Dimić, A. Mercader, A. Eduardo, C. Chalcone, Derivative cytotoxicity activity against MCF-7 human breast cancer cell QSAR study, Chemometrics and Intelligent Laboratory Systems, 146, 358–384, **2015.**

17 R. Sabet, M. Mohammadpoura, A. Sadeghi, A. Fassihi, QSAR study of isatin analogues as in vitro anti-cancer agents. European Journal of Medicinal Chemistry, 45, 1103–1118, **2010.**

18 V. Prachayasittikul, R. Pingaew, A. Worachartcheewan, C. Nantasenamat, S. Prachayasittikul, S. Ruchirawat, V. Prachayasittikul, Synthesis, anticancer activity and QSAR study of 1,4-naphthoquinone derivatives. European Journal of Medicinal Chemistry, 84, 247-263, **2014.**

19 M. Irfan, B. Aneja, U. Yadava, S. Khan, N. Manzoor, C. Daniliuc, M. Abid, Synthesis, QSAR and anticandidal evaluation of 1,2,3-triazoles derived from naturally bioactive scaffolds. European Journal of Medicinal Chemistry., 93, 236-254, **2015.**

20 T. Chen, D. Shyong, S. Chen, C. Chen, C. Lee, Y. Hsieh, L. Chang, J. Guh, J. Lin, H. Huang, Design, synthesis and biological evaluation of tetracyclic azafluorenone derivatives with topoisomerase I inhibitory properties as potential anticancer agents. Arabian Journal of Chemistry., 37, 254-272, 2016.

21 Advanced Chemistry Development Inc., Toronto, Canada **2009**. www.acdlabs.com/resources/freeware/chemsketch/).

22 XLSTAT 2014 software (XLSTAT Company), http://www.xlstat.com.

23 D.K. Asgaonkar, G. Mote, T. Chitre, QSAR and Molecular Docking Studies of Oxadiazole-Ligated Pyrrole Derivatives as Enoyl-ACP (CoA) Reductase Inhibitors. *Sci. Pharm., 82*, 61-86, **2014**.

24 S. Singh, S. Love, K. Yenamandra, S. Prabhakarb, S.G. Kaskhedikar, QSAR studies on benzoylaminobenzoic acid derivatives as inhibitors of β-ketoacyl-acyl carrier protein synthase III, European Journal of Medicinal Chemistry, 43, 1071–1080, **2008.**

Conclusion générale:

Notre travail de thèse a été consacré à la modélisation de la relation quantitative structure activité pour le développement de modèles QSAR fiables , robustes, stables et précis, capables de prédire efficacement l'activité anticancéreuse de quelques composés variés et appartenant aux trois familles chimiques: sulfonamides, pyrazoles et quinoléines.

Nous avons calculé dans un premier temps, un grand nombre de descripteurs moléculaires (descripteurs électroniques, topologiques, géométriques, physicochimiques, ...). Diverses méthodes statistiques ont été utilisées dans la construction de ces modèles (ACP, RLM, RNLM, PLS,...). Les principales techniques de validation ont été utilisées (les tests statistiques standards, la validation interne, la validation externe..). Ces modèles ont été développés en accord avec les cinq principes de l'OCDE pour la validation des modèles QSAR.

La première application a montré que les descripteurs de la chimie quantique, à savoir l'énergie LUMO et l'indice de réfraction, en combinaison avec l'indice d'hydrophobicité « log P », et les donneurs et accepteurs de la liaison hydrogène, sont utiles pour la prédiction de l'activité anticancéreuse des dérivés sulfoniques. Le modèle 2D-QSAR obtenu est capable de décrire environ 81% de la variance de l'activité expérimentale et pourrait être utilisé efficacement pour estimer l'activité anticancéreuse des dérivés sulfonamides pour lesquels les données expérimentales sont indisponibles.

Dans la deuxième application, nous avons utilisé la RLM, la PLS et la RNLM pour construire des modèles QSAR des dérivés pyrazoles pour leurs activités anticancéreuses. Les méthodes utilisées ont été comparées, et parmi elles, la RNLM basée sur des descripteurs proposés par la RLM (Stepwise) qui a présenté la meilleure capacité prédictive que les autres, même que la RLM Stepwise donne les résultats les plus simplement interprétables. Les résultats de cette application montrent que les modèles proposés peuvent prédire l'activité anticancéreuse avec une bonne précision et que les paramètres sélectionnés (ELUMO, MW...) sont pertinents. Nous avons conçu et suggéré, sur la base des résultats obtenus, quelques nouveaux composés ayant des activités anticancéreuses théoriquement supérieures à celles des composés étudiés. Ces composés peuvent être synthétisés et testés dans le cadre des recherches des médicaments anticancéreux à base de la pyrazole.

Dans la troisième application, l'activité anticancéreuse des dérivés quinoléinique est modélisée avec succès par l'énergie totale, log(P), la densité et les accepteurs de la liaison hydrogène, codée par différentes méthodes statistiques RLM, RNLM et PLS. Les modèles QSAR développés sont simples, interprétables et transparents en utilisant un nombre réduit de descripteurs. En outre, ils sont caractérisés par la stabilité, la robustesse et le pouvoir prédictif élevé vérifié par la validation

interne et externe. Ainsi, les modèles sont considérés comme validés et applicables pour l'exploitation de la base de données. Les modèles résultants ont été utilisés pour prédire les activités des 10 nouvelles molécules des dérivés d'azafluorénone tétracycliques. Il a également été démontré que les méthodes proposées sont une aide utile pour réduire le temps et le coût de la synthèse et la détermination de l'activité anticancéreuse des dérivés quinoléiniques.

Les perspectives de ce travail nous semblent diverses. D'une part, nous avons l'intention de reprendre les mêmes bases de données et élaborer des modèles en utilisant d'autres méthodes d'analyse de données telles que CoMFA, CoMSIA, EVA, les algorithmes génétiques...D'autre part, nous projetons élaborer des modèles QSAR pour d'autres bases de données de molécules anticancéreuses vis-à vis de leurs cellules cibles par utilisation du Docking Moléculaire.

Annexe : Méthodes de la chimie quantique

1 INTRODUCTION

Dans ce travail, une méthodologie combinant les approches de chimie quantique et les outils QSAR a été employée. L'élaboration des modèles QSPR/QSAR repose sur le calcul des paramètres (descripteurs), qui sont assurés en faisant appel aux outils de la modélisation moléculaire.

La chimie quantique est une science qui se base sur la résolution de l'équation de Schrödinger. Avec la puissance des ordinateurs courants, la rigueur et même l'exactitude de la théorie peuvent être mises à profit pour obtenir des réponses précises à toutes sortes de questions : géométrie d'une molécule, d'un intermédiaire ou d'un état de transition, ou faisabilité d'une réaction… Dans cette partie, il s'agit de représenter les méthodes de chimie quantique, utilisées pour le calcul des structures moléculaires nécessaires à la mise en place de modèles prédictifs[1,2].

2 BASES DE LA CHIMIE QUANTIQUE

La chimie quantique est une branche de la chimie dans laquelle les phénomènes chimiques sont élucidés déductivement sur la base de la mécanique quantique. La chimie couvre une large gamme d'échelles, des atomes et des petites molécules aux grands systèmes tels que biomolécules et des solides, et comprend leurs structures, leurs propriétés et leurs réactions. En effet, la structure électronique de la matière est déterminée en résolvant l' équation de Schrödinger. L'état d'un système à N noyaux et n électrons s'écrit en mécanique quantique par une fonction d'onde Ψ satisfaisant l'équation de Schrödinger[3-6].

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

Ψ: sont les fonctions propres de H

E : sont les valeurs propres de H

Après le développement de l'équation de Schrödinger (Schrödinger 1926) diverses théories fondamentales significatives de la mécanique quantique ont été produites dans une période remarquablement courte. Grâce au principe d'incertitude (Heisenberg 1927) et au principe de complémentarité onde-particule de Bohr (exposé à Côme, Italie, 1927), l'équation de Schrödinger relativiste (équation de Dirac) (Dirac 1928) a été développée 2 ans plus tard. Un tas d'expériences ont ensuite été menées pour soutenir ces théories fondamentales, formant les concepts de la mécanique quantique. Les théories fondamentales de la chimie quantique ont également été rapidement développées au cours de cette période comme approches pour clarifier la chimie basée sur la mécanique quantique[7].

L'hamiltonien H total d'une molécule comportant N noyaux et n électrons, est défini par la somme de cinq termes (terme cinétique des électrons, terme cinétique des noyaux, terme de répulsions électrons-électrons, terme de répulsions noyaux-noyaux et terme d'attractions electrons-noyaux)[8].

$$H= \frac{-\hbar}{2m_e}\sum_i^n \Delta_i - \frac{-\hbar}{2M_k}\sum_i^n \Delta_k + \sum_{j<i}^n \frac{e^2}{r_{ij}} + \sum_{l<k}^n \frac{Z_k Z_l e^2}{r_{kl}} - \sum_{k=1}^N \sum_{i=1}^n \frac{Z_k e^2}{R_{ki}}$$

Born et Oppenheimer ont proposé l'approximation des noyaux fixes qui consiste à séparer l'hamiltionien électronique de l'hamiltonien nucléaire. Dans le cadre de cette approximation (et en se plaçant dans le cadre non relativiste), l'hamiltonien H peut se réduire à la forme suivante[5]:

$$H= \frac{-\hbar}{2m_e}\sum_i^n \Delta_i + \sum_{j<i}^n \frac{e^2}{r_{ij}} - \sum_{k=1}^N \sum_{i=1}^n \frac{Z_k e^2}{R_{ki}}$$

Il n'est pas possible de résoudre cette équation pour des systèmes d'intérêt chimique (au-delà de H_2), de manière exacte. Il faut alors introduire différentes approximations[9-12].

3 METHODE DE HARTREE-FOCK

L'approximation de Hartree-Fock représente un point de départ de presque toutes les méthodes ab initio, soit pour faire des approximations supplémentaires comme dans le cas des méthodes semi-empiriques, soit pour ajouter des déterminants supplémentaires générant des solutions qui convergent vers une solution aussi proche que possible de la solution exacte de l'équation de Schrödinger électronique[13].

Cependant, lorsque nous résolvons réellement l'équation de Schrödinger pour ces atomes, nous sommes confrontés à un problème sérieux: le problème des trois corps. C'est-à-dire que l'état de mouvement ne peut pas être résolu analytiquement pour des systèmes dans lesquels trois masses distinctes ou plus interagissent. Ce problème à trois corps n'est pas propre à la mécanique quantique mais constitue un problème classique en physique analytique[14-17].

En 1928, deux ans après la publication de l'équation de Schrödinger, Hartree a proposé une méthode pour résoudre cette équation pour les systèmes à électrons multiples, basée sur des principes physiques fondamentaux: la méthode de Hartree (Hartree 1928). Considérons le mouvement électronique d'un atome d'hélium[18].

l'approximation de Hatree permet de remplacer l'interaction d'un électron avec les autres électrons par l'interaction de celui-ci avec un champ moyen créé par la totalité des électrons de la molécule; ce qui permet de remplacer le potentiel biélectronique $\sum_j e^2/r_{ij}$ qui exprime la répulsion instantanée entre l'électron i et les autres électrons j≠i par un potentiel monoélectronique moyen de l'électron i de la forme U(i). Par conséquent et en se basant sur le théorème des électrons indépendants, nous pouvons écrire la fonction d'onde totale comme le produit de fonctions d'onde mono-électroniques[19]:

$$\Psi= \Psi_1(1).\ \Psi_2(2).\ \Psi_3(3)..... \Psi_n(n)$$

3.1. Méthode de Hartree-Fock-Roothaan

Les expressions analytiques des orbitales moléculaires ϕ_i n'ont pas été définies dans le cadre de la méthode de Hartree-Fock. C'est Roothaan [20] qui a utilisé la technique OMCLOA pour construire les OM. Cette méthode consiste à exprimer l'orbitale moléculaire ϕ_i par une combinaison linéaire d'orbitales atomiques ϕ_μ :

$$\phi_i = \sum_{\mu=1}^{N} C_{i\mu}\phi_\mu$$

$C_{i\mu}$ sont les coefficients à faire varier. N étant le nombre d'OA combinées.

Les meilleurs coefficients sont ceux qui minimisent l'énergie. En procédant par la méthode des variations et après certaines manipulations algébriques, on aboutit aux équations de Roothaan définies par le système séculaire suivant [21]: P_{qp}

$$\sum_{r=1}^{N} C_{kr}\,(F_{rs} - £_k S_{rs}) = 0 \qquad S = 1,2,.....N$$

avec

$$\begin{cases} F_{rs} = h_{rs}^{c} + \sum_{p=1}^{n}\sum_{q=1}^{n} P_{qp}\left\{2\langle rs|pq\rangle - \langle rq|ps\rangle\right\} \\ S_{rs} = \langle \phi_r|\phi_s\rangle \\ h_{rs}^{c} = \int \phi_r^*(i)\, h^c\, \phi_i(i)\, dr \end{cases}$$

Où r, s, p et q symbolisent les OA. Ppq est l'élément de la matrice densité. Les termes<rs pq> et <rq ps> représentent les intégrales

3.2. Méthodes Post-Hartree-Fock

La méthode Hartree-Fock-Roothaan présente l'inconvénient majeur de ne pas tenir compte de la corrélation électronique qui existe entre le mouvement des électrons. Ceci rend cette méthode relativement restreinte dans le calcul quantitatif des propriétés thermodynamiques telles que l'enthalpie d'activation, l'énergie de Gibbs de réactions, énergies de dissociation[22].

Ces propriétés peuvent être calculées d'une manière efficace par les méthodes Post-HF en tenant compte de la corrélation électronique. L'énergie de corrélation d'un système correspond à la différence entre l'énergie Hartree-Fock et l'énergie exacte non-relativiste du système est :

$E_{corr} = E_{exacte} - E_{HF}$

Les techniques Post-HF sont en général très efficaces pour retrouver l'énergie de corrélation, mais cependant à l'heure actuelle elles sont, pour la majeure partie d'entre-elles, trop lourdes pour être applicables à des systèmes dont le nombre d'atomes est grand. Il s'est ainsi parallèlement développé à ces techniques un modèle alternatif qui a atteint le statut de théorie à la fin des années 60. La théorie de la fonctionnelle de la densité (DFT) qui est actuellement la seule permettant l'étude de systèmes

chimiques de grande taille avec la prise en compte des effets de la corrélation électronique de manière satisfaisante[23].

4 THEORIE DE LA FONCTIONNELLE DE LA DENSITE

Jusqu'à présent, les approches présentées, Hartree-Fock ou semi-empiriques, sont toutes fondées autour d'une fonction mathématique : la fonction d'onde. Même si cette dernière peut être reliée à l'énergie de la molécule, il n'en reste pas moins que cette grandeur n'a pas en soi de signification physique[24].

La théorie de la fonctionnelle de la densité (DFT) diffère des approches précédentes en prenant pour propriété fondamentale la densité électronique, qui est, quant à elle, une observable. Cette approche a pour origine le postulat de Hohenberg- Kohn [16], qui établit que l'énergie d'un système dans son état fondamental est une fonctionnelle de la densité électronique de ce système, ρ(r), et que toute densité, ρ'(r), autre que la densité réelle conduit nécessairement à une énergie supérieure. Ainsi contrairement aux méthodes précédentes, la théorie de la fonctionnelle de la densité ne consiste pas à chercher une fonction d'onde complexe, ψ, à 3N-dimensions décrivant le système à étudier, mais plutôt une simple fonction à trois dimensions : la densité électronique totale ρ [25].

Il existe trois types de fonctionnelles énergies d'échange-corrélation : les fonctionnelles locales, les fonctionnelles à correction du gradient et les fonctionnelles hybrides.

Ces méthodes de calcul de la structure électronique des molécules et des solides reposent sur une approche assez différente des méthodes du type SCF. Leur fondement se trouve dans un théorème du a Hohenberg et Kohn qui ont démontre que toutes les propriétés d'un système dans un état fondamental non dégénère sont complètement déterminées par sa densité électronique *p (r).* Formellement, 1'énergie apparait comme une fonctionnelle de la densité, fonctionnelle qui demeure inconnue du fait de 1'impossibilité de résoudre exactement un problème a plusieurs électrons.

Par ailleurs, Kohn et Sham18 ont étendu a la densité le principe variationnel, en montrant que la fonction *p (r)* exacte correspond au minimum de 1'énergie, ce qui permet la recherche de solutions approchées, sous réserve que 1'on sache évaluer 1'énergie. Moyennant une décomposition de celle-ci en un terme d'énergie cinétique, un terme d'interaction coulombienne des électrons entre eux et avec les noyaux et un terme complémentaire rassemblant les contributions liées aux effets d'échange et de corrélation, le problème se ramène a la recherche d'une expression approchée pour évaluer le terme d'échange-corrélation a partir de la densité qui devient une *fonctionnelle* de la densité. La démarche la plus simple consiste a se référer au traitement statistique du gaz homogène d'électrons. En posant : $\rho(\vec{r}) = \rho\ (\vec{r}) + \rho\ (\vec{r})$ ou $\rho\ (\vec{r})$ désigne la densité électronique de spin $-\frac{1}{2}$ et $\rho\ (\vec{r})$ la même fonction relative au spin - $-\frac{1}{2}$, on obtient deux fonctions d'échange-corrélation de la forme[26] :

$V_{xc}(\vec{r}) = -3\alpha\left[-\frac{3}{4\pi}\rho(\vec{r})\right]$

Le coefficient *a*, qui prend la valeur 2/3 dans le système considéré, est généralement traite comme un paramètre dont la valeur a été ajustée pour retrouver des résultats de référence, souvent atomiques. Cette valeur optimale

est toujours un peu supérieure a 2/3.

Si nous faisons 1'hypothèse que la densité électronique peut s'écrire sous la forme du carre du module d'un déterminant de Slater construit a 1'aide de spin orbitales *,-(!), celles-ci sont obtenues en recherchant le minimum de 1'energie[28] .

Fonctionnelle hybride B3LYP :

La fonctionnelle hybride B3LYP (Becke 3-parameters Lee-Yang-Parr) consiste à une hybridation (mélange) de plusieurs fonctionnelles de différentes méthodes comme le montre l'expression[30] suivante[29]:

$$E_{xc}^{B3LYP} = (1-a_0-a_x)\,E_X^{LSDA} + a_0\,E_x^{exact} + a_x\,E_x^{B88} + (1-a_c)\,E_C^{VWN} + a_c\,E_C^{LYP}$$

Les valeurs des 3 paramètres d'ajustement sont:

$a_0 = 0.20$

$a_x = 0.72$

$a_c = 0.81$

5 METHODES SEMI-EMPIRIQUES

Les calculs semi-empiriques sont eux développés sur la même structure générale que les calculs HF, mais certaines parties de l'information sont sujettes à approximation ou même complètement omises, afin de les rendre moins exigeants en termes de temps de calcul. En particulier, dans le cadre de ces approches, trois types de simplifications sont principalement réalisés :

- non-considération des électrons de cœur, puisqu'ils ne contribuent pas à la réactivité chimique du système. Ils seront alors considérés avec le noyau au sein d'une fonction paramétrée (ex : méthode de Hückel étendu [30]) ;
- utilisation d'un jeu réduit de fonctions de base (typiquement STO-3G) ;
- réduction du nombre d'intégrales bi-électroniques grâce à l'introduction de paramètres empiriques.

En effet, l'étape la plus exigeante en termes de temps de calcul dans la résolution des équations HF réside dans le traitement des intégrales bi-électroniques de la matrice de Fock. Pour une base de dimension *n*, *n4* intégrales bi-électroniques doivent être traitées. Aussi, le traitement de systèmes de taille importante nécessite des temps de calcul eux aussi importants. Afin de pouvoir traiter des systèmes moléculaires de plus grande taille, les méthodes semi-empiriques introduisent des approximations supplémentaires dans les équations HF.

Il s'agit de négliger certains recouvrements orbitalaires et d'estimer les autres à partir de considérations empiriques. L'approche CNDO (*Complete Neglect of Differential Overlap*), par exemple, néglige tous les recouvrements[32]. L'approche NDDO (*Neglect of Diatomic Differential Overlap*) néglige, quant à elle, uniquement les recouvrements mettant en jeu des orbitales centrées sur des noyaux différents[33].

L'avantage des calculs semi-empiriques est qu'ils sont donc plus rapides que les autres méthodes quantiques. Leur inconvénient réside dans le fait qu'elles sont soumises à de nombreuses approximations. En fait, peu de propriétés peuvent être prédites de manière fiable, en particulier, pour des molécules de structures trop éloignées de celles utilisées pour la paramétrisation des méthodes. En général, celle-ci est réalisée afin de reproduire différents propriétés : géométries, énergies de formation, ou encore énergies de réaction, moments dipolaires, potentiels d'ionisation voire des propriétés spécifiques telles que des spectres électroniques ou des déplacements chimiques RMN.

L'une des méthodes semi-empiriques les plus utilisées est le modèle AM1 (*Austin Model 1*) [34]. Cette approche emploie un schéma de type NDDO dans lequel les recouvrements des intégrales bi-électroniques mono-centrées sont paramétrés sur des données spectroscopiques pour des atomes isolés, les autres considérant des interactions entre multipôles. Si cette méthode est en particulier largement utilisée pour les composés organiques, elle présente quelques limitations reconnues dans l'estimation des énergies d'activation, stabilité de certains composés ou enthalpies de liaison[35].

Références

1 R.G. Parr, W. Yang, Density Functional Theory of Atoms and Molecules, Oxford University Press, New-York, **1989**.

2 L. Pauling, E.B.J. Wilson, Introduction to Quantum Mechanics with Applications to Chemistry,

3 McGraw-Hill Book Company Inc., New York, **1935**.

4 K.S. Pitzer, Quantum Chemistry, Prentice-Hall Inc., Englewood Cliffs, New Jersey, **1953**.

5 A. Szabo, N.S. Ostlund, Modern Quantum Chemistry - Introduction to Advanced Electronic Structure Theory, Macmillan Publishing Co. Inc., New York, **1982**.

6 R.G. Parr, W. Yang, Density Functional Theory of Atoms and Molecules, Oxford University Press, New-York, **1989**.

7 J.-L. Rivail, Elements de chimie quantique à l'usage des chimistes, CNRS Editions, Paris, **1999**.

8 D.C. Young, Computational Chemistry: A Practical Guide for Applying Techniques to Real-World Problems, John Wiley & Sons Inc., New York, **2001**.

9 C.J. Cramer, Essentials of Computational Chemistry - Theories and Models, Wiley, Chicherster, U.K., **2004**.

10 C. Audouze, Efficient methods of approximations of solutions in molecular quantum chemistry, Ph.D. Thesis, University Paris XI Orsay, Department of Mathematics, **2004**.

11 T. Clark, A Handbook of Computational Chemistry: A Practical Guide to Chemical Structure and Energy Calculations, Edition, wiley, London, **1985.**

12 R.G. Parr, R.A. Donnelly, M. Levy, and W.E. Palke, Electronegativity: the density functional viewpoint‖, The Journal of Chemical Physics, 68, 3801–3807, **1978**.

13 J.L. Rivail, Eléments de chimie quantique à l'usage des chimistes‖, 2éme édition, CNRS Editions, **1999**.

14 R.S. Mulliken, The Assignment of Quantum Numbers for Electrons in Molecules. II. Correlation of Molecular and Atomic Electron States‖, Physical Review, 32, 761–772, **1928**.

15 L.H. Thomas, The calculation of atomic fields‖, Proceedings of the Cambridge Philosophical Society, 23, 542–548, **1927**.

16 E. Fermi, "A Statistical Method for the Determination of Some Atomic Properties", Rendiconti Lincei, 6, 602–607, **1927**.

17 P.A.M. Dirac, ―The Quantum Theory of the Electron. Part II, Proceedings of the Royal Society of London A, 118, 351–361, **1928**.

18 P. Hohenberg and W. Kohn, Inhomogeneous Electron Gas‖, Physical Review, 136, 864–871, **1964**.

19 C. Corminboeuf, F.Tran, J. Weber, The role of density functional theory in chemistry: Some historical landmarks and applications to zeolites, J. Mol. Struct.: Theochem.762, **2006.**

20 A.R. Leach, Quantum Mechanical Models, In: Molecular Modelling: Principles and Applications. Addison Wesley Longman Ltd., Harlow, **1996.**

21 H. Chermette, Density functional theory: A powerful tool for theoretical studies in coordination chemistry, Coord. Chem. Rev. 178-180 () 699-721, **1998.**

22 P.J. Stephens, F.J. Devlin, C.F. Chabalowski, M.J. Frisch, Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields, J. Phys. Chem. 98, 11623-11627, **1994.**

23 L. Landau, E.ifchiz, « Mécanique quantique », Editions Mir, moscau, **1967.**

24 J.P. Doucet; J. Weber, J. Computer-aided moleculardesign: Theory and applications, Academic Press, London, 266, **1996.**

25C.K. Skylaris, thèse de doctorat, The Computing Modelling of Heavy Atom Chemistry, université de Cambridge, **1999**.

26 D.R. Hartree, The wave mechanics of an atom with a non-coulomb central field, I. Theory and methods, Proc. Cambridge Philos. Soc. 24, 89-110, **1928**.

27 Rivail. J.L, Eléments de chimie quantique à l'usage des chimistes, 2ième éd., CNRS Edition **1999.**

28 J.C. Slater, The Theory of Complex Spectra, Phys. Rev. 34, 1293-1322, **1929.**

29 D.R. Hartree, The wave mechanics of an atom with a non-coulomb central field, I. Theory and methods, Proc. Cambridge Philos. Soc. 24, 89-110, **1928**.

30 V. Fock, Approximation method for solving the quantum mechanical multibody problem, Physc. 61,126-148, **1930**.

31 C. Lee, W. Yang, and R.G. Parr, Development of the colle-salvetti correlation-energy formula into a functional of the electron density, Physical Review, 37, 785–789, **1988**.

32 A.D. Becke, Density-functional thermochemistry.3: The role of exact exchange, The Journal of Physical Chemistry, 98, 5648–5652, **1993**,.

33 P.J. Stephens, J.F. Devlin, C.F. Chabalowski, and M.J. Frisch, ab initio calculations of vibrational absorption and circular dichroïsm spectra using SCF, MP2 and density functional theory force fields‖, The Journal of Physical Chemistry, 98, 11623–11627, **1994**.

34 E. Fermi, A Statistical Method for the Determination of Some Properties of the Atom and Its Application to the Theory of the Periodic System of the Elements, Zeitschrift für Physik, 48, 73–75, **1928**.

35 E. Fermi, On the Statistical Deduction of Some Atomic Properties. Application to the Theory of the Periodic System of the Elements Rendiconti Lincei, 7(6), 342–346, **1928**.