الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE DE MEDEA
FACULTE DE TECHNOLOGIE
Département du Génie des Procédés et
Environnement (G.P.E)

LBMPT
Laboratoire de Biomatériaux et Phénomènes de Transport

THESE DE DOCTORAT

Présentée par :

Said BITAM

à l'Université Yahia FARES de Médéa
pour l'obtention du titre de Docteur en Génie des Procédés

Modélisation de l'activité thérapeutique à partir de la structure moléculaire

Membres du jury :

Mr	**O. BENKORTBI**	**Professeur (UYF de Médéa)**	**Président**
Mme	**S. KOUADRI-MOUSTEFAI**	**Professeur (UHBBC de Chlef)**	**Examinateur**
Mr	**B. BENSEBIA**	**Maître de Conférences A (UHBBC de Chlef)**	**Examinateur**
Mr	**B. CHEKNANE**	**Maître de Conférences A (USD de Blida)**	**Examinateur**
Mr	**S. HANINI**	**Professeur (UYF de Médéa)**	**Co-Directeur de thèse**
Mr	**M. HAMADACHE**	**Maître de Conférences A (UYF de Médéa)**	**Directeur de thèse**

Année 2018

REMERCIEMENTS

En tout premier lieu, je tiens à rendre grâce à Allah, le tout puissant qui m'a donné la santé, la volonté et la force pour mener à bien ce travail.

Ce travail a été réalisé sous la direction de Messieurs ***Salah Hanini*** *et Dr* ***Mabrouk HAMADACHE****, Professeur et Maître de conférences à l'Université Yahia Farés de Médéa. Qu'ils trouvent dans ces quelques phrases l'expression de mes plus vifs remerciements et de ma reconnaissance pour leurs conseils, leurs encouragements, leur patience ainsi que leur disponibilité. Votre soutien indéfectible a été pour moi très précieux durant ces années que j'ai passées au sein de votre équipe.*

Je suis très reconnaissant envers les membres du jury : Professeur ***Othmane BENKORTBI*** *(UYF de Médéa), Professeur* ***Soumia KOUADRI-MOUSTEFAI*** *(UHBB de Chlef),* ***Dr BENSEBIA Bensaber*** *(UHBB de Chlef)* ***et Dr CHEKNANE Benamar*** *(USD de Blida) qui me font le plaisir d'examiner ce travail et de participer à ma soutenance de thèse.*

Enfin, je tiens à remercier tous ceux qui m'ont aidé de près ou de loin durant ces années de recherche. Je ne peux conclure sans un mot de remerciement à tous mes collègues enseignants de l'université Yahia Farés de Médéa.

Said BITAM

الملخص

الأنزيمات butyrylcholinesterase (BUCHE) وأستيل (أستيلكولينستراز) مسؤولة جزئيا عن ظهور مرض الزهايمر. تظهر العديد من الدراسات التجريبية أن مشتقات N-benzylpiperidine ومشتقات Tacrine لها نشاط تثبيطي متزايد لهذين الإنزيمين. ومع ذلك، فإن تركيب واختبار هذه المثبطات المتاحة حاليا يحتاج إلي الكثير من الوقت والمال لذا فمن المستحسن تطوير أساليب بديلة لتسهيل عملية الفرز لهذه المشتقات في المرحلة الأولى من مراحل اكتشاف الأدوية.

أيضا ، الهدف من هذا العمل هو تطوير نماذج تنبؤية قوية تفي بالتوصيات الصارمة لمنظمة التعاون الاقتصادي والتنمية (OECD)للتحقق من صحة نماذج QSAR .لهذا الغرض ، تم تطوير ثلاثة نماذج QSAR : MLR و MLP و SVR لمثبطات الـ BuChE وثلاثة نماذج أخرى MLR و MLP و GFA لمثبطاتAChE. واستخدمت استراتيجيات ومعايير التحقق المتاحة للتحقق من صحة هذه النماذج. هذه المصادقة كشفت عن قوتها وموثوقيتها. ومع ذلك ، فإن أداء نماذج SVR و GFA أفضل بكثير من النماذج الأخرى.

الكلمات المفتاحية : الزهايمر, Tacrine, N-benzylpiperidine, QSAR , AChE , BUCHE,

Résumé

Les enzymes Butyrylcholinestérase (BuChE) et Acétylcholinestérase (AChE) sont en partie responsables de l'apparition de la maladie d'Alzheimer. De nombreuses études expérimentales montrent que les dérivés de la Tacrine et ceux de la N-benzylpipéridine présentent une activité inhibitrice accrue contre ces deux enzymes. Cependant, la synthèse et les tests de dépistage de ces inhibiteurs actuellement disponibles sont coûteux et prennent du temps. Il est donc souhaitable de développer des méthodes alternatives pour faciliter le dépistage de ces dérivés dans la phase initiale de la découverte de médicaments.

Aussi, l'objectif de ce travail est de développer des modèles prédictifs robustes qui doivent satisfaire les recommandations rigoureuses de l'organisation de coopération et de développement économiques (OCDE) en matière de validation des modèles QSAR. Pour cela, trois modèles QSAR (MLR, MLP et SVR) ont été développés pour les inhibiteurs de la BuChE et trois autres (MLR, MLP et GFA) pour les inhibiteurs de l'AChE. Les stratégies et critères de validation disponibles ont été utilisés pour valider ces modèles. Cette validation a révélé leur robustesse et fiabilité. Néanmoins, la performance des modéles SVR et GFA est sensiblement meilleure que celle des autres modèles.

Mots clés : Alzheimer, AChE, BuChE, Tacrine, N-benzylpipéridine, QSAR

Abstract

The enzymes Butyrylcholinesterase (BuChE) and Acetylcholinesterase (AChE) are partially responsible for the appearance of Alzheimer's disease. Several experimental studies showed that the derivatives of Tacrine and those of N-benzylpiperidine present have an increased inhibitory activity against these two enzymes. However, the synthesis and screening tests for these currently available inhibitors are expensive and require more time. Therefore, it is desirable to develop alternative methods to facilitate detection of these derivatives in the initial phase of the drug discovery.

Also, the objective of this work is to develop robust predictive models that must satisfy the rigorous recommendations of the Organization for Economic Co-operation and Development (OECD) regarding the validation of QSAR models.

For this purpose, three QSAR models (MLR, MLP and SVR) have been developed for inhibitors of BuChE and three others (MLR, MLP and GFA) for AChE inhibitors. The available validation strategies and criteria were used to validate these models. This validation revealed their robustness and reliability. Nevertheless, the performance of the SVR and GFA models is significantly better than the other models.

Key words : Alzheimer, AChE, BuChE, Tacrine, N-benzylpipéridine, QSAR

ABREVIATIONS

En français

AChE	Enzyme Acetylcholinesterase
ADN	Acide Désoxyribonucléique
AFG	Algorithme d'approximation de la fonction génétique
A.M.M	Autorisation de Mise sur Le Marché
ARN	Acide Ribonucleique
BuChE	Enzyme Butyrylcholinestérase
CE_{50}	Concentration Efficace médiane
CI_{50}	Concentration Inhibitrice médiane
D.A	Domaine d'Applicabilité
FELASA	Fédération des Associations Européennes de Science des Animaux de Laboratoire
M.A	Maladie d'Alzheimer
M.P	Maladie de Parkinson
OCDE	Organisation de Coopération et de Développement Economiques
OMS	Organisation Mondiale de la Santé
PMC	Perceptron multicouche
RLM	Régression Linéaire Multiple
RNA	Réseaux de Neurones Artificiels

En Anglais

ANN	Artificial Neural Network
BNN	Backpropagation Neural Network
CETP	Cholesteryl Ester Transfer Protein
CID	PubChem Compound Identification
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Index Analysis
GA	Genetic Algorithm
GFA	Genetic Function Approximation

G/PLS	Genetic Partial Least Squares
GQSAR	Group-Based Quantitative Structure Activity Relationship
HQSAR	Hologram QSAR
HTS	High-Throughput Screening
IUPAC	International Union of Pure and Applied Chemistry
kNN-MFA	K Nearest Neighbor Molecular Field Analysis
MFA	Molecular Field Analysis
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
PSO	Particle Swarm Optimization
QSAR	Quantitative Structure-Activity Relationship
REACH	Registration, Evaluation, Authorization and Restriction of Chemicals
RMSE	Root Mean Square Error
S-MLR	Stepwise Multiple Linear Regression
SVM	Support Vector Machine
SVR	Support Vector Regression
VIF	Variance Inflation Factor
V.S	Virtual Screening

LISTE DES FIGURES

		Page
Chapitre I	**Etat de l'art sur l'utilisation des Approches "in silico" dans la conception des médicaments**	
Figure I.1	Répartition des molécules utilisées actuellement comme médicament en fonction de la nature de leurs cibles	6
Figure I.2	Croissance des publications liées à la modélisation QSAR basée sur le Google Scholar Search ("QSAR" comme mot clé + excluant les citations et les brevets – consulté le 23 février 2018)	10
Figure I.3	Evolution de la modélisation QSAR favorisée par la croissance des données expérimentales. Le graphique est généré par Google Ngram Viewer (http://books.google.com/ngrams) où l'axe des Y représente le pourcentage de tous les livres de la base de données	11
Figure I.4	Taux d'évolution de la cardiopathie coronarienne dans le monde entre 2005 et 2016 (sources : http://www.healthdata.org/algeria)	12
Figure I.5	Estimations de l'incidence du cancer en 2012 selon « International Agency For Research on Cancer »	13
Figure I.6	Maladies causant le plus de décès en Algérie entre 2005 et 2016. (Source : http://www.healthdata.org/algeria)	15
Chapitre II	**Approches QSAR et méthodologie de leur mise en œuvre**	
Figure II.1	Classification des approches QSAR	33
Figure II.2	Diagramme de dispersion des valeurs de l'activité observée en fonction de celles calculées par le modèle	40
Figure II.3	Neurone formel et neurone biologique	40
Figure II.4	Types de réseaux de neurones artificiels selon topologie des connexions entre les neurones	42

Chapitre II	**Approches QSAR et méthodologie de leur mise en œuvre**	
Figure II.5	Architecture d'un perceptron multicouche(PMC)	43
Figure II.6	Exemples d'une population de modèles représentée pour l'algorithme AFG	45
Figure II.7	Opération de croisement génétique	46
Figure II.8	Recherche d'un hyperplan de séparation optimal au sens de la marge	47
Figure II.9	Les paramètres de la fonction de perte pour un SVR	49
Figure II.10	Architecture d'une machine à vecteurs support	51
Chapitre III	**Développement de modèles QSAR pour la prediction de la concentration inhibitrice (CI_{50})**	
Figure III.1	capture d'écran du site www.bindingdb.com	58
Figure III.2	Structure moléculaire de la Tacrine	59
Figure III.3	Représentation graphique des concentrations inhibitrices (pIC50) expérimentale et celle prédite par le modèle MLP pour une série de dérivés de la Tacrine	71
Figure III.4	Contribution des descripteurs dans le cas du modèle MLP	73
Figure III.5	Représentation graphique des concentrations inhibitrices (pIC50) expérimentale et celle prédite par le modèle SVR pour une série de dérivés de la Tacrine	75
Figure III.6	Diagramme de Williams pour les ensembles d'apprentissage et de test du modèle SVR	78
Figure III.7	Structure chimique de la N-benzylpipéridine	80
Figure III.8	Courbe de corrélation entre les valeurs prédites de pIC50 en fonction des valeurs expérimentales pour les ensembles d'apprentissage et de test (Modèle MLR)	90
Figure III.9	Courbe de corrélation entre les valeurs prédites de pIC50 en fonction des valeurs expérimentales pour les ensembles d'apprentissage et de test (Modèle MLP)	94
Figure III.10	Contribution à l'effet inhibiteur par type de descripteurs	95
Figure III.11	Courbe de corrélation entre les valeurs prédites de pIC50 en fonction des valeurs expérimentales pour les ensembles d'apprentissage et de test (Modèle GFA)	97

LISTE DES TABLEAUX

Chapitre I	**Etat de l'art sur l'utilisation des Approches "in silico" dans la conception des médicaments**	
Tableau I.1	Coûts typiques des expériences	8
Tableau I.2	Croissance des publications liées à la modélisation QSAR basée sur le Google Scholar Search ("QSAR" and « maladie » mot clé + excluant les citations et les brevets – consulté le 25 Février 2018)	15
Chapitre II	**Approches QSAR et méthodologie de leur mise en œuvre**	
Tableau II.1	Exemples représentatifs des bases de données utiles pour la conception de médicaments	36
Tableau II.2	Différentes fonctions d'activation les plus utilisées	41
Chapitre III	**Développement de modèles QSAR pour la prediction de la concentration inhibitrice (CI_{50})**	
Tableau III.1	Dérivés de la tacrine utilisés et valeurs des concentrations inhibitrices expérimentales et prédites par MLR, MLP (ANN) et SVR	60
Tableau III.2	Description, VIF et MF des descripteurs sélectionnés	68
Tableau III.3	Paramètres statistiques du modèle MLP pour les ensembles d'apprentissage et de test	72
Tableau III.4	Paramètres statistiques et critères de validité du modèle SVR	74
Tableau III.5	Comparaison des paramètres statistiques des trois modèles	76
Tableau III.6	Domaine d'applicabilité du modèle SVR pour l'ensemble de test	77
Tableau III.7	Comparaison entre les modéles de la littérature et celui développé dans ce travail (enzyme BuChE)	80
Tableau III.8	Dérivés de la N-benzylpipéridine utilisés et valeurs des concentrations inhibitrices expérimentales et prédites par MLR, MLP (ANN) et GFA.	82
Tableau III.9	Descripteurs utilisés pour le développement des modèles QSAR	88

Tableau III.10	Matrice des coefficients de corrélation des descripteurs sélectionnés	90
Tableau III.11	Valeurs du coefficient VIF des descripteurs sélectionnés	90
Tableau III.12	Paramètres de validation interne et externe du modèle MLR	92
Tableau III.13	Valeurs du paramètre t-Test des descripteurs sélectionnés	93
Tableau III.14	Optimisation du nombre de neurones	94
Tableau III.15	Paramètres statistiques du modèle MLP	96
Tableau III.16	Paramètres statistiques du modèle GFA	98
Tableau III.17	Paramètres statistiques des modèles développés	99
Tableau III.18	Comparaison entre les modéles de la littérature et celui développé dans ce travail (enzyme AChE)	101

TABLE DES MATIERES

INTRODUCTION GENERALE 1

CHAPITRE I

Etat de l'art sur l'utilisation des Approches "in silico" dans la conception des médicaments

I.1 CONCEPTION DE MEDICAMENTS 4

I.1.1 Historique 4
I.1.2 Phases de conception actuelle d'un médicament 5
I.1.2.1 Première phase : recherche exploratoire 5
I.1.2.2 Seconde phase : études précliniques 7
I.1.2.3 Troisiéme phase : recherches cliniques 7

I.2 PROBLEMES LIES A LA CONCEPTION DES MEDICAMENTS 8

I.2.1 Le coût 8
I.2.2 Utilisation des animaux à des fins expérimentales 8

I.3 ACTIVITE INHIBITRICE D'UN MEDICAMENT 9

I.4 CONCEPTION DES MEDICAMENTS ET OUTIL INFORMATIQUE 9

I.5 APPLICATION DES APPROCHES IN SILICO A LA CONCEPTION DE MEDICAMENTS POUR DIVERSES MALADIES 10

I.6 APPLICATION DES APPROCHES IN SILICO A LA CONCEPTION DE MEDICAMENTS POUR LA MALADIE D'ALZHEIMER 14

I.7 CONCLUSION 24

REFERENCES 25

CHAPITRE II

Approches QSAR et méthodologie de leur mise en œuvre

II.1 UTILITES ET APPLICATIONS DES APPROCHES QSAR 32

II.2 CLASSIFICATION DES APPROCHES QSAR 32

II.3 VALIDITE D'UNE METHODOLOGIE QSAR 34

II.4 PRINCIPE ET METHODOLOGIE GENERALE D'UNE ETUDE QSAR 35

II.5 COLLECTE DE LA BASE DE DONNEES 35

II.6 CALCUL DES DESCRIPTEURS 36

II.6.1 Classification des descripteurs 37

II.6.2 Sélection des descripteurs 38

II.7 METHODES D'ANALYSE DES DONNEES 39

II.7.1 Régression linéaire multiple 39

II.7.2 Les réseaux de neurones artificiels 39

II.7.2.1 Types de réseau de neurones 42

II.7.2.2 Perceptron multicouche 42

II.7.3 L'algorithme d'approximation de la fonction génétique (AFG) 44

II.7.3.1 Fonctionnement de l'algorithme AFG 46

II.7.4 Machines à vecteurs supports pour la régression (SVR) 47

II.7.4.1 Machines à vecteurs supports (SVM) 47

II.7.4.2 SVM pour la régression (SVR) 48

II.8 OUTILS DE VALIDATION DE MODÉLES QSAR 52

II.9 DOMAINE D'APPLICABILITE D'UN MODELE QSAR 52

II.10 CONCLUSION 53

REFERENCES 54

CHAPITRE III

1ére partie : **Développement de modèles QSAR pour la prediction dela concentration inhibitrice (CI_{50}) d'une serie de dérives de la Tacrine sur l'enzyme butyrylcholinesterase (BuChE)**

III.1 MATERIELS ET METHODES 58

III.1.1 Collecte des données de l'activité inhibitrice 58

III.1.2 Calcul et selection des descripteurs 59
III.1.3 Développement des modèles QSAR 64
III.1.4 Validation des modèles QSAR 64
III.1.4.1 Validation interne 65
III.1.4.2 Validation externe 65
III.1.5 Domaine d'applicabilité 67

III.2 RESULTATS ET DISCUSSION 67

III.2.1 Calcul et selection des descripteurs 67
III.2.2 Modèle MLR 68
III.2.3 Modèle neuronal MLP 70
III.2.4 Modèle SVR 73
III.2.5 Comparaison des trois modèles MLR, MLP ET SVR 75
III.2.6 Domaine d'applicabilité des modèles 76
III.3 COMPARAISON AVEC D'AUTRES MODELES DE LA LITTERATURE 78
III.4 CONCLUSION 79

CHAPITRE III

2éme Partie : **Developpement de modeles qsar pour la prediction dela concentration inhibitrice (CI_{50}) d'une série de dérives de la N-benzylpipéridine sur l'enzyme acetylcholinesterase (AChE)**

III.4 MATERIELS ET METHODES 81

III.4.1 Collecte des données de l'activité inhibitrice 81
III.4.2 Calcul et sélection des descripteurs 81
III.4.3 Développement des modéles QSAR 87
III.4.4 Validation des modéles QSAR 88
III.4.5 Domaine d'applicabilité des modéles QSAR 88

III.5 RESULTATS ET DISCUSSION 88

III.5.1 Calcul et sélection des descripteurs pertinents 88
III.5.2 Modèle MLR 90
III.5.3 Modèle MLP 93
III.5.4 Modèle GFA 97
III.5.5 Comparaison des trois modèles 99

III.5.6 Domaine d'applicabilité 99

III.6 COMPARAISON AVEC LES MODELES DE LA LITTERATURE 99

III.7 CONCLUSION 100

REFERENCES : 102

CONCLUSION GENERALE ET PERSPECTIVES 107

PUBLICATIONS SCIENTIFIQUES

INTRODUCTION GENERALE

L'espérance de vie des humains continue à augmenter régulièrement suite aux progrès réalisés par la médecine moderne (cas des maladies infectieuses, de la santé maternelle et nutritionnelle, du paludisme, … etc.), mais aussi grâce à l'amélioration du niveau de vie. En conséquence, les humains vivent de plus en plus vieux. Néanmoins, un inconvénient majeur se dresse face à cette augmentation de l'espérance de vie. En outre, plus la population vieillit, plus le risque de contracter une des maladies neurodégénératives augmente également. Ces maladies, de plus en plus répandues, se caractérisent par une diminution des facultés cognitives, de la mémoire et par la survenue d'anomalies comportementales qui interfèrent avec la capacité de l'individu atteint à mener ses activités quotidiennes normales (bouger, parler et respirer). La démence, la maladie d'Alzheimer (M.A), la maladie de Parkinson (M.P), la maladie de Huntington et la sclérose en plaques sont des exemples de maladies neurodégénératives. Notons que la maladie d'Alzheimer est la maladie neurodégénérative la plus fréquente.

L'évolution démographique et le vieillissement progressif de la population laissent supposer que les maladies neurodégénératives vont constituer, dans les vingt prochaines années, un défi pour le système de santé et la politique de recherche de médicaments du fait qu'elles représentent une cause majeure de dépendance et ont un impact important sur la qualité de vie des personnes atteintes ainsi que sur celle de leurs proches. Selon l'organisation mondiale de la santé (OMS), 40 millions de personnes dans le monde sont actuellement touchées par la maladie d'Alzheimer alors que pour la maladie de Parkinson, le nombre de personnes atteintes dépasse les 10 millions.

Au cours des dernières années, la compréhension des facteurs responsables des dommages neuronaux dans diverses maladies neurodégénératives a considérablement progressé. Cependant, malgré toutes ces avancées, il n'existe à l'heure actuelle, aucun médicament qui soit efficace dans le traitement de ces maladies. Par conséquent, il est essentiel et urgent de concevoir et de développer de nouvelles classes de médicaments thérapeutiques, capables de traiter ce genre de maladies.

La conception et le développement de médicaments sont des processus très compliqués, couteux et qui durent dans le temps. Le coût de ces processus (phase de recherche exploratoire, phase préclinique, phase clinique et phase de mise sur le marché) est estimé entre 800 millions et 1,4 milliard de dollars et dure jusqu'à 14 ans entre le stade de la recherche initiale et la commercialisation réussie d'un nouveau médicament. Les phases de recherche exploratoire et préclinique sont parfois irréalisables au vu des coûts aussi bien financiers que temporels (environ 15% du budget total nécessaire au développement d'un médicament et une durée de 4 à 5ans) étant donné que des milliers de molécules sont synthétisées et ensuite testées sur des cibles. Pour pallier à ces nombreuses difficultés, la conception de médicaments assistée par ordinateur (utilisation de méthodes *in silico*) est une alternative très intéressante. Ces méthodes *in silico* permettent l'élaboration de modéles mathématiques de prédiction des propriétés d'une substance sans qu'elle ne soit synthétisée. Par conséquent, elles peuvent donc accélérer le processus de conception et réduire les coûts faramineux.

Ce mémoire se veut justement une contribution à l'utilisation d'une de ces méthodes *in silico* désignée sous le nom de "Relations Quantitatives Structure-Activité" (ou QSAR : Quantitative Structure-Activity Relationships). Les modéles QSAR quantifieront la relation et mettront en exergue l'impact de la structure moléculaire d'une substance sur une de ses propriétés. En se basant sur le modèle QSAR obtenu, on pourra faire la prédiction de la valeur de la propriété de toute molécule en un temps record et sans qu'on la synthèse ou qu'on ait recours aux essais sur la cible. Dans cette optique, on s'est assigné comme objectif, l'élaboration et l'application de modèles QSAR dans la conception de médicaments anti-Alzheimer. Pour ce faire, on a élaboré des modéles de prédiction du pouvoir inhibiteur d'une série de dérivés de la Tacrine sur l'enzyme butyrylcholinestérase (BChE) d'une part et d'une série de dérivés de la N-benzylpiperidine sur l'enzyme acétylcholinestérase (AChE) d'autre part.

Pour atteindre l'objectif assigné à cette thèse, trois chapitres ont été développés. La présente introduction générale de ce mémoire a été l'occasion de situer le contexte et les motivations de cette recherche. Il a été question de la problématique des maladies neurodégénératives et de celle de la conception des médicaments y afférant.

Dans le premier chapitre, nous proposerons tout d'abord, une description sommaire sur la conception du médicament, sur les difficultés qui y sont liées ainsi que sur l'importance de l'outil informatique durant les premières phases de cette conception. Par la suite, il est question d'une revue bibliographique succincte de travaux scientifiques entrepris dans le cadre de l'utilisation des approches QSAR à la conception de médicaments inhérents à certaines maladies. On clôturera ce chapitre par un état de l'art sur l'application des approches QSAR pour la conception de médicaments anti-Alzheimer.

Les approches QSAR étant la méthode *in silico* utilisée dans le présent travail, tout un chapitre (Chapitre II) leur a été consacré. Après un rappel succinct sur leur utilité, leur classification et leur validité, on y parlera longuement de la méthodologie de mise en

œuvre de ces approches et des outils nécessaires à leur élaboration (collecte de la base de données, calcul et sélection d'une série de descripteurs qui caractérisent les structures moléculaires, utilisation de méthodes d'analyse de données, validation du modèle et domaine d'applicabilité).

Le troisième chapitre est consacré, quant à lui, à l'élaboration de modèles QSAR dans la conception de médicaments anti-Alzheimer. Une première partie concernera les modéles QSAR (modèles MLR, GFA et ANN) de prédiction du pouvoir inhibiteur d'une série de dérivés de la N-benzylpiperidine sur l'enzyme AChE. Dans la seconde partie de ce chapitre, il sera question de l'élaboration de trois modèles QSAR (modèles MLR, SVR et ANN) de prédiction du pouvoir inhibiteur de dérivés de la Tacrine sur l'enzyme butyrylcholinestérase (BChE). Dans chacune des deux parties, on y relatera tout à tour la partie concernant les données utilisées, le calcul et la selection des descripteurs, le développement de trois modéles QSAR et leur validation ainsi que de leur domaine d'applicabilité. Ensuite, suivra une discussion sur les résultats trouvés et une comparaison entre les trois modèles élaborés.

Enfin, la conclusion générale aura pour objet de synthétiser l'ensemble des résultats de ce travail et de proposer des perspectives de recherche dans la continuité de ce thème de recherche prometteur.

CHAPITRE I

ETAT DE L'ART SUR L'UTILISATION DES APPROCHES "*IN SILICO*" DANS LA CONCEPTION DES MEDICAMENTS

De nos jours, une quantité considérable de travaux de recherche est orientée vers la conception de médicaments en utilisant l'approche "*in silico*". Ce chapitre dédié à l'état de l'art sur l'utilisation de cette approche, comprend les principaux thèmes suivants : (1) un rappel bref mais essentiel sur la conception des médicaments où sont mis en exergue l'historique et les différentes phases de cette conception ; (2) une rétrospective des problèmes (coûts et éthique) liés à cette conception ; (3) un rappel sur l'activité inhibitrice des médicaments ; (4) l'utilisation de l'outil informatique dans la conception des médicaments ; (5) une revue bibliographique sur l'utilisation de l'approche QSAR dans la conception de médicaments vis-à-vis de certaines maladies ; (6) une revue bibliographique sur l'utilisation de l'approche QSAR dans la conception de médicaments anti-Alzheimer comportant en amont le pourquoi du choix de cette maladie ; (7) une conclusion générale comportant des appréciations sur les approches QSAR relevées dans la partie (6).

I.1 CONCEPTION DE MEDICAMENTS

I.1.1 Historique

Les approches et les processus de recherche d'un médicament ont évolué depuis l'antiquité à nos jours. La première approche, appelée approche empirique, s'étale de l'antiquité jusqu'à la fin du XVIII^{ème} siècle. Elle fut caractérisée par la transmission par voie orale des savoirs médicinaux nés de l'utilisation des substances naturelles, principalement d'origine végétale mais aussi minérale et animale. Dès le début du XIX^{ème} siècle, le développement technologique et l'accumulation de connaissances en anatomie, en physiologie et en chimie ont donné naissance à l'approche fonctionnelle. Celle-ci avait

pour objectif de démontrer l'effet d'un composé sur une fonction de l'organisme tel que la respiration, la digestion, etc. Les composés utilisés étaient issus de l'extraction des principes actifs des substances naturelles précédemment utilisées. A partir du début des années 1970 et devant la nécessité de développer d'autres sources de médicaments, une multitude d'approches (approche par criblage à haut débit, approche par chimie combinatoire, approche rationnelle) ont été initiées.

I.1.2 Phases de conception actuelle d'un médicament

Les symptômes qui accompagnent une maladie sont dus au dysfonctionnement de processus biologiques clés. Le dérèglement initial d'un seul de ces processus au niveau de la cellule suffit le plus souvent par le jeu des interactions à entraîner le dérèglement progressif et souvent irréversible de tout un édifice de réactions biologiques. Il s'ensuit une perte progressive, parfois définitive, d'une fonction physiologique vitale (cardiaque, cérébrale, rénale, respiratoire,…) dont l'issue peut être fatale. C'est seulement en restaurant le fonctionnement normal des processus responsable de cet " effondrement " biologique qu'on peut espérer interrompre la progression d'une maladie. Il en résulte que seule la connaissance de l'origine d'une maladie et de ses facteurs (étiopathogénie)à un niveau moléculaire permettra de concevoir de façon rationnelle des médicaments capables d'en arrêter la progression (*drug design*) [1].

Le développement d'un nouveau médicament est un processus long (entre 12 et 14 ans), complexe avec un coût total de l'ordre de 800 millions à 1,8 milliard de dollars [2]. Ce processus est scindé en trois phases : une première phase dédiée à la recherche exploratoire, une seconde phase englobe les études précliniques et une troisième phase consacrée à la recherche clinique.

I.1.2.1 Première phase : recherche exploratoire

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

- **<u>Identification de la cible</u>**

Ces cibles sont nombreuses. On y dénombre plusieurs familles :

- Les enzymes visées par environ 25 % des molécules actives actuelles. Ces molécules sont essentiellement des inhibiteurs de l'activité enzymatiques.
- Les pompes, transporteurs et canaux ioniques, protéines membranaires qui régissent les équilibres transmembranaires des principaux ions, visés par environ 15 % des molécules actives, des inhibiteurs des mouvements ioniques;

- Les récepteurs membranaires visés par environ 40 % des molécules médicamenteuses, agonistes ou antagonistes. Leur diversité est illustrée sur la Figure I.1
- Les récepteurs nucléaires visés par environ 10 % des molécules actives, agonistes ou antagonistes

Ces quatre familles de cibles correspondent à près de 90 % des médicaments. Il faut y ajouter 5 % des molécules se liant à des cibles diverses, par exemple les protéines des microtubules qui lient la colchicine, et les alcaloïdes de la pervenche. Enfin, les cibles d'environ 5 % des molécules actives reconnues ne sont pas déterminées, par exemple la cible du paracétamol [3].

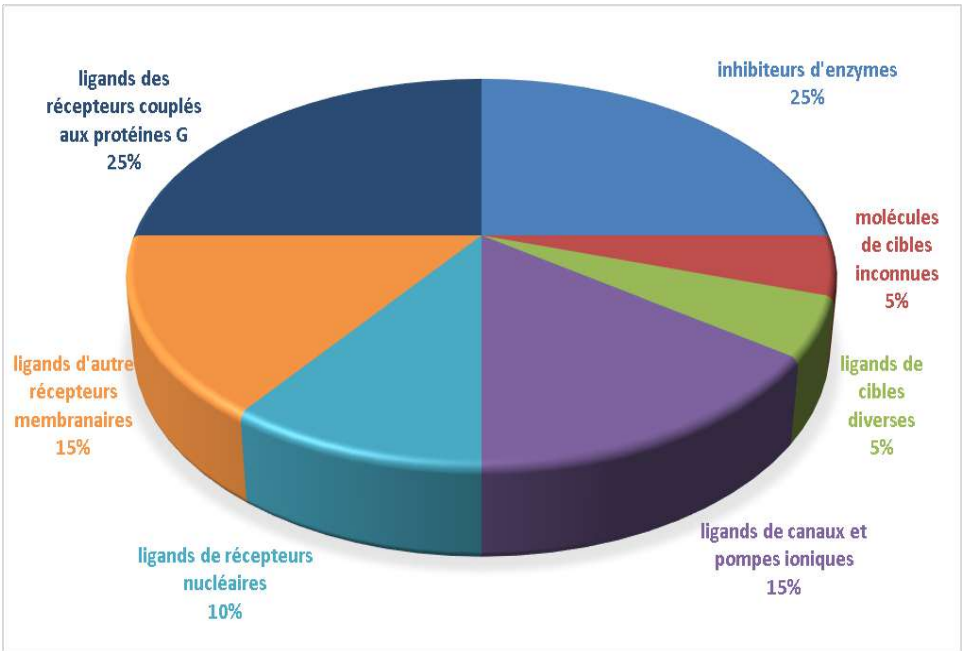


Figure I.1 Répartition des molécules utilisées actuellement comme médicament en fonction de la nature de leurs cibles [3].

- **Criblage**

Une fois que la cible est identifiée, on recherche des composés capables de l'inhiber de façon spécifique, importante et durable. Ces composés sont appelés "*hits*". La recherche de hits se fait par criblage *in vitro* (ou *screening*) de larges banques de composés appelées chimiothèques et par la réalisation de tests biologiques. Les coûts aussi bien financiers que temporels de ce criblage, où des milliers de molécules sont utilisées, s'avèrent énormes [4]. Aussi, dans le but de minimiser les coûts de cette approche classique *in vitro*, le développement d'approches *in silico* pour le criblage virtuel des composés chimiques est une alternative prometteuse.

Ces approches *in silico* font appel à des outils informatiques. Ce terme "*in silico*" fait référence au silicium, matériau principal retrouvé dans les puces informatiques de tous les ordinateurs. Ces approches permettent de simuler ou de modéliser un phénomène biologique à l'aide de l'outil informatique. Une des plus prometteuses approches *in silico* est celle désignée sous l'acronyme anglais QSAR (Quantitative Structure Activity Relationship). Elle se base sur des relations quantitatives entre structures et propriétés des composés.

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

I.1.2.2 Seconde phase : études précliniques

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

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

I.1.2.3 Troisième phase : recherche clinique

Si les tests précliniques sont satisfaisants, on procède alors aux tests cliniques, c'est-à-dire qu'on passe aux tests sur l'être humain. Beaucoup de candidats médicaments sont éliminés avant ce stade atteint par seulement un médicament sur quinze. Cette recherche se déroule chez l'humain en trois phases principales (phase I, II, III) dans des conditions bien réglementées et dans un cadre juridique. Les essais de la phase I ont pour but de rechercher la dose toxique au-delà de laquelle le candidat médicament ne pourrait pas être utilisé, tandis que pour la phase II il est question de rechercher les doses ayant une activité pharmacologique favorable. La phase III, quant à elle, est dédiée à la mesure précise de l'efficacité du candidat médicament. Après ces trois phases, un dossier d'Autorisation de Mise sur le Marché (A.M.M) est déposé pour un seul candidat médicament parmi ceux ayant été soumis aux tests cliniques [5].

I.2 PROBLEMES LIES A LA CONCEPTION DES MEDICAMENTS

I.2.1 Le coût

La conception de médicaments est un processus multidisciplinaire extrêmement long (10-20 ans) et coûteux [6]. On trouve dans la littérature des estimations assez divergentes concernant son coût global, évalué entre 300 millions et plus de 1,7 milliard de dollars [7]. L'étude de DiMasi [8, 9], l'une des plus citées, estime ce coût à environ 802 millions de dollars, tandis qu'une autre étude le situe autour de 1,2 milliard[10]. Récemment, une autre étude faite par DiMasi [11] montre que le coût estimatif moyen des dépenses par nouveau composé approuvé est de 1 395 millions de dollars (cours du dollars en 2013). La capitalisation des coûts hors poche jusqu'au point d'approbation de la commercialisation à un taux d'actualisation réel de 10,5% donne un coût total pré-approuvé estimé à 2588 millions de dollars (cours du dollar de 2013). Pour ne rien arranger, le taux de succès des projets de l'industrie pharmaceutique reste faible [12] et le nombre de nouveaux médicaments mis sur le marché stagne [13]. Certains coûts typiques pour divers types d'expériences sont énumérés dans le tableau I.1 [14] ci-dessous.

Tableau I.1 Coûts typiques des expériences [14]

Expérience	Coûts ($) /composé
Modélisation informatique	10
Dosage biochimique	400
Test de culture cellulaire	4000
Toxicité aiguë du rat	12000
Structure cristalline de protéine	100000
Essai d'efficacité animale	300000
Rat Toxicité orale chronique de 2 ans	500000
Essai clinique humain	500000000

I.2.2 Utilisation des animaux à des fins expérimentales

Bien que l'expérimentation sur les animaux a joué un rôle essentiel dans presque toutes les percées médicales au cours de la dernière décennie, il n'en demeure que des dizaines d'associations ayant vu le jour aux Etats unis et en Europe, dénoncent l'horreur de l'expérimentation animale et militent pour son abolition. C'est ainsi que plusieurs pays se dotent de la réglementation la plus stricte en termes de protection des animaux de laboratoire. A titre d'exemple, selon Smith et al. [15], la Fédération des associations européennes de science des animaux de laboratoire (FELASA) a affirmé que " l'examen éthique devrait viser à garantir, à toutes les étapes du travail scientifique impliquant des animaux, une justification éthique adéquate et clairement expliquée de l'utilisation des animaux ". FELASA a non seulement souligné la nécessité d'analyses avantages-dommages avant de se lancer dans des projets de recherche impliquant des animaux, mais a

également souligné l'importance des processus éthiques de recherche animale "normatifs" pour refléter diverses perspectives éthiques [15].

Par conséquent, la tendance actuelle est de ne recourir au modèle animal que si aucun autre modèle ne peut lui être substitué. Parmi ces modèles substitutifs ou méthodes non animales, on cite les méthodes *in vitro* (culture de cellules) et les modèles *in silico* (modélisation informatique).

I.3 ACTIVITE INHIBITRICE D'UN MEDICAMENT

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

L'inhibition de l'activité enzymatique est une stratégie majeure utilisée dans la conception de médicaments [17]. Aussi, la concentration inhibitrice médiane (CI_{50}) est la mesure in vitro la plus pertinente utilisée pour évaluer le potentiel thérapeutique (ou efficacité) d'un candidat médicament. La CI_{50} est définie comme étant la concentration d'un médicament nécessaire pour 50% d'inhibition d'un processus biologique ou d'un composant d'un processus comme une enzyme, une cellule, un récepteur cellulaire ou un microorganisme. Pour les candidats médicaments qui sont des inhibiteurs, des définitions analogues, CI_{25}, CI_{90}, CI_{95}, et CI_{99}, sont souvent utilisées pour décrire la concentration d'un inhibiteur requise pour réduire la réponse de 25%, 90%, 95% et 99% de sa valeur en absence d'inhibiteur, respectivement [18].

I.4 CONCEPTION DES MEDICAMENTS ET OUTIL INFORMATIQUE

Des méthodes informatiques (*in silico*) ont été développées et largement appliquées au développement et aux tests d'hypothèses pharmacologiques. Ces méthodes comprennent des bases de données, relations quantitatives structure-activité, recherche de similarités, pharmacophores et autres modélisations moléculaires, méthodes d'apprentissage, exploration de données.

De telles méthodes ont été fréquemment utilisées dans la découverte et l'optimisation de nouvelles molécules ayant une affinité pour une cible, avec des propriétés d'absorption, de distribution, de métabolisme, d'excrétion et de toxicité ainsi que des caractérisation physicochimique [19].

Les simulations physiopathologiques peuvent maintenant être examinées à l'aide de programmes de modélisation informatique de haute technologie (modélisation *in silico*) [20, 21]. Le criblage de la toxicité [22] et les événements pharmacocinétiques fondamentaux tels que l'absorption intestinale, la liaison protéique peuvent également

être effectués rapidement *in vitro* selon la disponibilité du programme de modélisation *in silico* [23].

Il existe d'autres techniques basées sur des logiciels (relations quantitatives structure-activité ou QSAR en anglais) [24] qui utilisent des estimations sophistiquées de la capacité inductrice d'une molécule, basée sur sa similarité avec les molécules existantes et la physiologie humaine existante. La modélisation QSAR s'est développée, diversifiée et a évolué de l'application à de petites séries de composés congénères en utilisant des méthodes de régression relativement simples à l'analyse de très grands ensembles de données comprenant des milliers de structures moléculaires variées (Figure I.2).

Notons qu'une recherche, effectuée sur le nombre de livres publiés, indique que la croissance continue des publications sur la conception des médicaments, a favorisé une croissance simultanée des publications QSAR (Figure I.3).

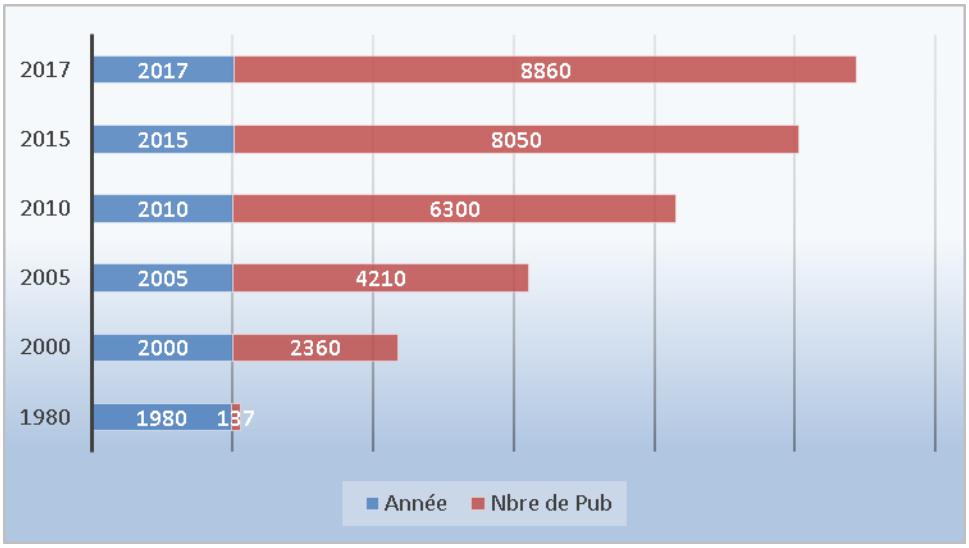


Figure I.2 Croissance des publications liées à la modélisation QSAR basée sur le Google Scholar Search ("QSAR" comme mot clé + excluant les citations et les brevets/ consulté le 23 février 2018).

I.5 APPLICATION DES APPROCHES *IN SILICO* A LA CONCEPTION DE MEDICAMENTS POUR DIVERSES MALADIES

Ce paragraphe se veut une revue bibliographique succincte de travaux scientifiques entrepris dans le cadre de l'utilisation des approches in silico à la conception de médicaments inhérents à certaines maladies (paludisme, tuberculose, maladie de Chagas, maladies parasitaires, Cardiopathie Coronarienne et cancer). Ces travaux couronnés de succès [25, 26] ont été obtenus par suite de l'utilisation du criblage virtuel, des méthodes QSAR, de l'amarrage moléculaire, de la modélisation du pharmacophore et de la dynamique moléculaire.

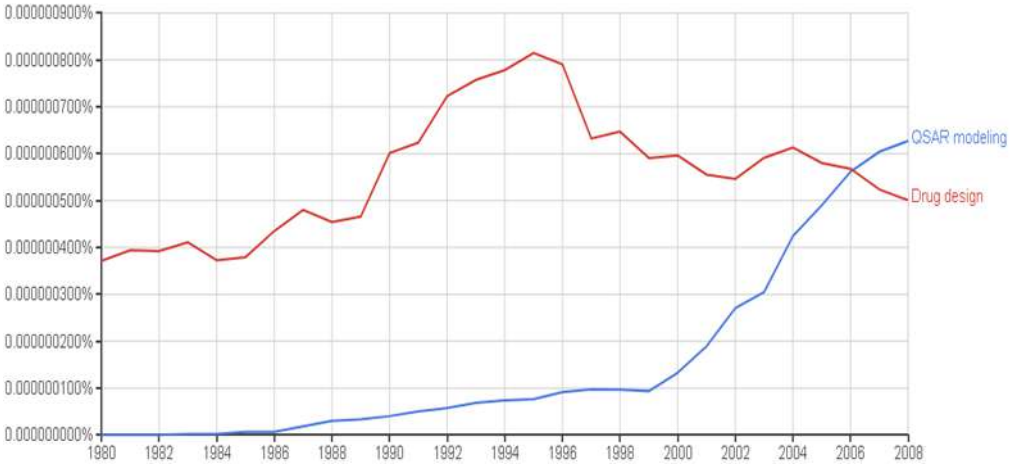


Figure I.3 Evolution de la modélisation QSAR favorisée par la croissance des données expérimentales. Le graphique est généré par Google Ngram Viewer (http://books.google.com/ngrams) où l'axe des Y représente le pourcentage de tous les livres de la base de données Google Ngram et l'axe des X celui des années (consulté le 24/02/2018).

Une série de composés présentant une activité puissante contre *Plasmodium falciparum* (parasite qui cause le paludisme) a été découverte [27]. Les auteurs ont par la suite développé plusieurs modèles QSAR basés sur l'activité *in vitro* de près de 3000 composés contre les cellules de *P. falciparum*. Puis, la base de données ZINC (http://zinc.docking.org/) a été filtrée pour les propriétés de ressemblance médicamenteuse (*drug-like*). Les composés sélectionnés ont été soumis à une procédure d'analyse du domaine d'applicabilité (D.A) qui délimite l'espace chimique dans lequel un modèle QSAR fournit des prédictions fiables pour les ensembles de validation internes et externes. Des prédictions fiables ont été fournies par les modèles QSAR pour des composés appartenant au D.A, et un sous-ensemble de composés a été testé expérimentalement. Plusieurs d'entre eux ont inhibé la croissance du parasite dans les cultures cellulaires. Cette technique a ainsi permis la conception de nouveaux agents antipaludiques.

L'enzyme 2-trans-énoyl-ACP-(CoA)-réductase (InhA) est une cible clé de la synthèse des acides gras de Mycobacterium tuberculosis, agent causal de la tuberculose. Un sous-ensemble de la base de données ZINC a été utilisé dans une stratégie combinant le virtual screening (V.S) et 3D pharmacophores [28]. Cette étude a permis la découverte des oxadiazols et thiadiazols en tant qu'inhibiteurs de *M.tuberculosis* InhA. Dans un travail intéressant, une stratégie combinée comprenant des études d'amarrage (docking) et un test de criblage à haut débit (high-throughput screening : H.T. S) d'environ 20 000 composés a été développée en utilisant l'enzyme cruzain de T. cruzi. Cette dernière est une cible validée pour la maladie de Chagas [29]. Un ensemble initial avec près de 1000 touches " *hit* " a été identifié à partir du H.T.S. Des études d'amarrage moléculaire ont été effectuées sur ces résultats, et les composés ayant les meilleurs scores ont été évalués par

rapport à l'enzyme cible. Cette procédure a permis la sélection d'un sous-ensemble d'inhibiteurs puissants avec des valeurs de K_i (constante de dissociation pour le complexe enzyme-inhibiteur) acceptables. Ces inhibiteurs, appartenant à cinq classes structurelles différentes, sont des points de départ intéressants pour le développement futur.

La combinaison de la cristallographie des protéines et de l'amarrage moléculaire a conduit à l'étude d'une série d'inhibiteurs puissants de l'enzyme purine nucléoside phosphorylase (P.N.P) de *Schistosoma mansoni*, agent responsable d'une maladie parasitaire appelée "schistosomiase" [30]. Les données cristallographiques de l'enzyme ont été utilisées pour une étude de criblage d'une série de dérivés de la déazaguanine et d'autres bases puriques. Les composés ont montré un excellent profil inhibiteur, avec des valeurs de CI_{50} acceptable. La ligase d'édition d'ARN (REL1) est une enzyme essentielle de la voie d'édition de l'ARN. Elle représente une cible de médicaments pour un groupe de protozoaires qui comprend les agents responsables de la maladie du sommeil africaine. Des inhibiteurs de type *drug-like* ont été découverts grâce à la combinaison d'un criblage virtuel basé sur la structure et la dynamique moléculaire [31].

La mortalité par cardiopathie coronarienne a diminué en Europe de l'Ouest et en Amérique du Nord au cours des dernières décennies (Figure I.4). Cependant, elle demeure l'une des principales causes de mortalité humaine [32, 33].

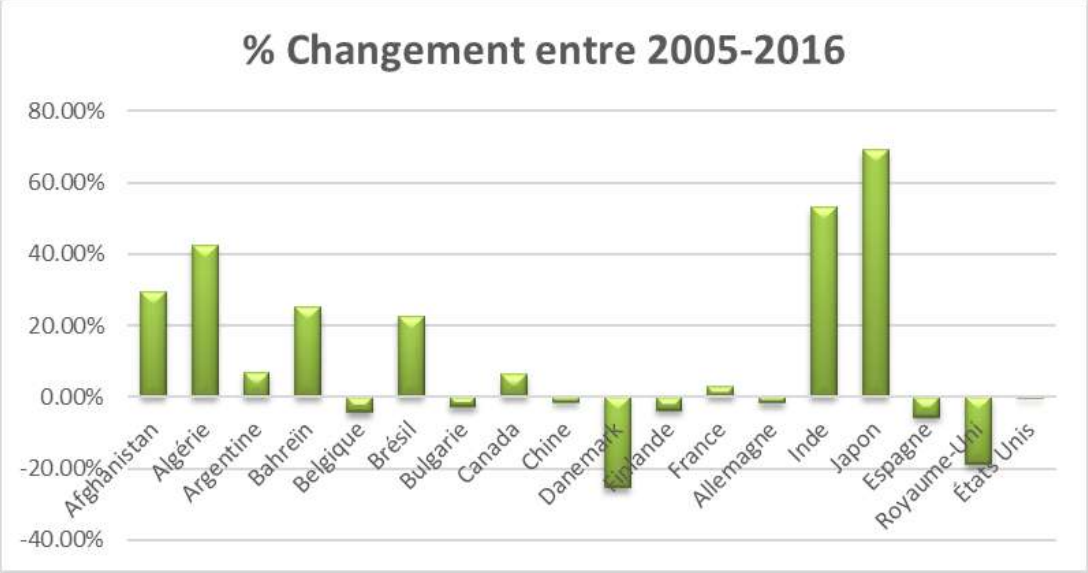


Figure I.4 Taux d'évolution de la cardiopathie coronarienne dans le monde entre 2005 et 2016 (sources : http://www.healthdata.org/algeria consulté le 25/02/2018).

Dans le cadre de la conception de médicaments pour la cardiopathie coronarienne, l'une des stratégies utilisées comme approche thérapeutique efficace est basée sur l'inhibition de la protéine de transfert du cholestérol ester (CETP). Castilho et al.[34] ont utilisé des modèles QSAR 2D robustes classiques (coefficients r^2 = 0,76 et q^2 = 0,72) et

hologrammes QSAR (coefficients r^2 = 0,88 et q^2= 0,70) pour une série de 85 inhibiteurs (NN-disubstitué trifluoro-3-amino- Les dérivés de 2-propanol) de la CETP. Ces modèles sont de nature complémentaire et mettent en évidence des caractéristiques structurelles importantes pour la conception de nouveaux inhibiteurs de la CETP dotés d'une puissance améliorée. Politi et al. [35] ont utilisé les techniques CoMFA (Comparative Molecular Field Analysis) et CoMSIA (Comparative molecular similarity index analysis) pour dériver des modèles QSAR 3D stables pour 30 inhibiteurs de la rénine [5 (S) amino-4 (S) -hydroxy-8-phényl-octanecarboxamides substitués par 2,7-dialkyl-substitué] utilisés comme ensemble d'entraînement (training set). Les modèles QSAR produits ont fourni pour les coefficients statistiques les valeurs suivantes : r^2=0.895 et 0.971 et r^2_{cv} =0.628 et 0.666 respectivement pour les techniques CoMFA et CoMSIA. Par ailleurs, les 2-arylbenzoxazoles se sont révélés être une classe potentielle d'inhibiteurs de la CETP. Dans ce cadre, une étude QSAR 2D réalisée par Jana et al. [36] sur une série de 2-arylbenzoxazoles en utilisant des techniques PCR (Principal component regression), PLS (Partial least square) et MLR (Multiple linear regression) a été validée de manière externe pour déterminer des modèles significatifs. De même, les techniques kNN-MFA (k-Nearest Neighbour Molecular Field Analysis 3D QSAR ont été réalisées sur la même série pour corréler les effets des paramètres électrostatiques, stériques et hydrophobes avec l'activité inhibitrice CETP.

Les radicaux libres jouent un rôle important dans la physiologie et la pathologie de l'homme en raison de leur capacité à endommager les membranes, les protéines, les enzymes ou l'ADN [37]. Ainsi, la concentration accrue des radicaux libres augmente le risque de maladies telles que le cancer, la maladie d'Alzheimer et la maladie de Parkinson [38]. Le cancer est la deuxième cause de décès la plus fréquente après les maladies du cœur. Notons que les cancers du sein, de la prostate, du poumon et colorectal sont les types de cancer les plus fréquents [39-42] dans le monde (Figure I.5) (http://gco.iarc.fr/today).

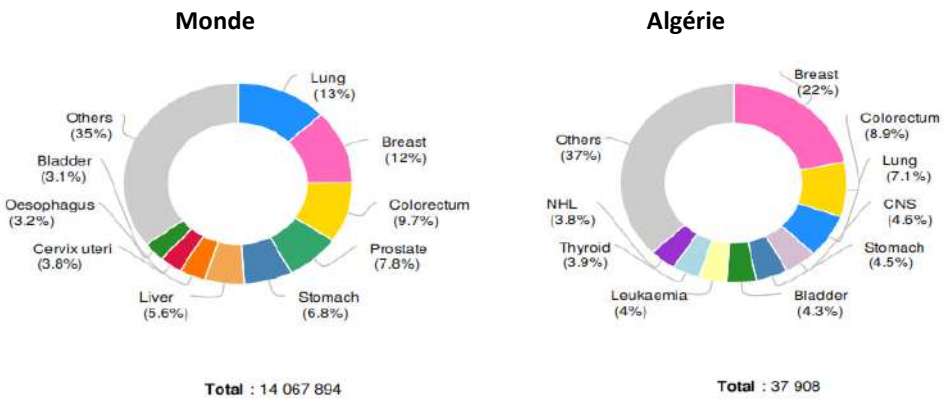


Figure I.5 Estimations de l'incidence du cancer en 2012 selon " International Agency for Research on Cancer: http://gco.iarc.fr/today consulté le 25/02/2018".

Développer un modèle pour prédire l'activité anticancéreuse basée sur la structure moléculaire est un objectif très important pour les chimistes médicinaux. Différents descripteurs moléculaires peuvent être utilisés pour résoudre ce problème. Les descripteurs moléculaires stochastiques, appelés MARCHE-INSIDE, se sont révélés très efficaces dans la conception de médicaments. González et al.[43] ont élaboré un modèle non linéaire de réseaux neuronaux artificiels (RNA) appelé Smart MLP-RNA et ce pour modéliser l'activité anticancéreuse des composés organiques. Ce modèle a montré une précision moyenne élevée de 93,79% (performance du train) et une prévisibilité de 90,88% (performance de validation) pour la topologie 8:3-MLP avec différentes séries d'entraînement et de prédiction. Afin d'illustrer l'utilisation pratique du modèle en chimie médicinale, ils ont rapporté la prédiction *in silico* et l'évaluation *in vitro* de six nouveaux analogues synthétiques ayant des valeurs de la CI_{50} comprises entre 37,1 et 138 µg mL^{-1}. Ils ont constaté que les prédictions théoriques coïncident convenablement avec les résultats expérimentaux. Par ailleurs, une approche combinée de modélisation QSAR validée et de criblage virtuel a été élaboré par Zhang et al.[44] et a contribué à la découverte de nouveaux dérivés de la tylophrine en tant qu'agents anticancéreux. Des modèles QSAR utilisant des descripteurs topologiques chimiques et des méthodes de sélection par variables voisines (kNN) ont été développés au moyen des CE_{50} expérimentales de 52 dérivés de la tylophrine à base de phénanthrine chimiquement divers (PBT : phenanthrine-based tylophrine derivatives). Plusieurs protocoles de validation ont été appliqués pour obtenir des modèles QSAR robustes. Les dix meilleurs modèles ont ensuite été utilisés pour extraire une base de données, ce qui a abouti à 34 résultats avec des activités prédites modérées à élevées. Dix résultats structurellement divers ont été testés expérimentalement et huit ont été confirmés actifs avec la CE_{50} expérimentale la plus élevée de 1,8 µM. Cet état de fait implique un taux de succès exceptionnellement élevé (80%). Les dix précédents modèles ont été appliqués pour prédire la CE_{50} pour quatre nouveaux PBT. Le coefficient de corrélation (R^2) entre les CE_{50} expérimentale et prédite dépasse 0,57. Cette étude suggère que l'approche combinant le criblage virtuel et la modélisation QSAR validée pourrait être utilisée avec succès en tant qu'outil général pour la découverte de nouveaux composés biologiquement actifs.

I.6 APPLICATION DES APPROCHES *IN SILICO* A LA CONCEPTION DE MEDICAMENTS POUR LA MALADIE D'ALZHEIMER

Dans ce paragraphe, il sera question d'une revue bibliographique sur l'application de l'approche QSAR dans la conception de médicaments anti-Alzheimer. Nous avons choisi de travailler sur la maladie d'Alzheimer du fait qu'elle représente un véritable enjeu de santé publique. A titre d'exemple, au niveau mondial, on estime que 100 millions de personnes seront touchés par cette maladie d'ici l'année 2050 [45].

Outre qu'elle se caractérise par une perte de mémoire progressive et un déclin des compétences langagières [46], la maladie d'Alzheimer (M.A) présente d'autres signes de diagnostic comme la désorientation dans le temps et dans l'espace, le changement d'humeur et de la personnalité. Elle est fréquente chez les personnes âgées, affectant environ 7% de la population âgée de plus de 65 ans [47]. En Algérie, la maladie d'Alzheimer est devenue l'une des maladies mortelles (figure 1.6) selon les statistiques du "Institute for Health Metrics and Evaluation" (I.H.M.E).

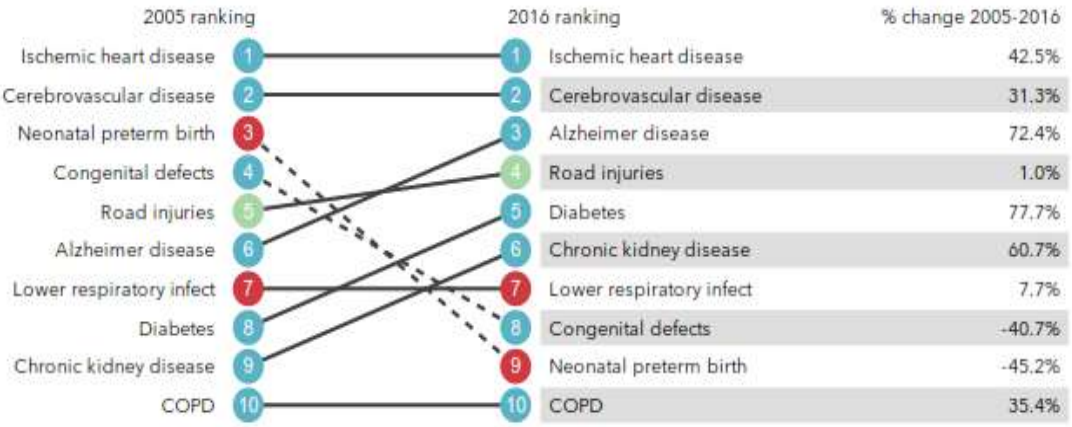


Figure I.6 Maladies causant le plus de décès en Algérie entre 2005 et 2016. (Source : http://www.healthdata.org/algeria consulté le 25/02/2018)

Il s'agit d'une maladie incurable sans approche thérapeutique efficace [48], et il n'est pas surprenant que des recherches aient été menées tant sur son traitement palliatif que sur sa guérison potentielle entre 2012 et 2017 (Tableau I.2).

Tableau I.2 Croissance des publications liées à la modélisation QSAR basée sur le Google Scholar Search ("QSAR" and " maladie " mot clé + excluant les citations et les brevets – consulté le 25 Février 2018)

Maladie	Année					
	2012	2013	2014	2015	2016	2017
Alzheimer	731	786	766	811	925	1 050
Parkinson	485	520	526	514	587	601
La cardiopathie ischémique	402	408	372	355	389	403
Les agents du VIH	1 530	1 610	1 580	1 580	1 670	1 470
Cancer	3 030	3 410	3 420	3 530	3 800	4 040
Grippe	332	376	365	369	399	336
Diabète	270	334	363	398	389	526
Tuberculose	505	591	649	689	709	715
Hypertension	324	326	329	369	352	392

Les causes profondes de la maladie d'Alzheimer (M.A) sont encore incertaines, alors que plusieurs hypothèses existent comme l'hypothèse cholinergique, l'hypothèse amyloïde et l'hypothèse tau [49]. Sur la base de l'hypothèse tau, l'hyper phosphorylation de la protéine tau conduit à la formation d'enchevêtrements neurofibrillaires responsables de la M.A. Selon l'hypothèse amyloïde, le dépôt extracellulaire de plaques bêta-amyloïdes est la cause fondamentale de la maladie [49].

Selon l'hypothèse cholinergique, cette maladie est causée par un taux réduit d'acétylcholine (ACh), neurotransmetteur important qui augmente l'attention et facilite l'apprentissage [50]. Cette réduction est due à une enzyme appelée acétylcholinestérase (AChE), qui catalyse l'hydrolyse rapide de l'acétylcholine en choline et acide acétique, entraînant ainsi son inactivation [51]. Par conséquent, une des stratégies adoptées pour améliorer la transmission cholinergique est de procéder à l'inhibition de l'AChE considérée dans ce cas comme une cible [52]. Une autre cible participant au métabolisme de l'ACh est désignée sous le nom de butyrylcholinestérase (BuChE). Néanmoins, la BuChE est prise comme cible dans les stades ultérieurs du traitement de la maladie d'Alzheimer [53].

La dernière décennie a été marquée par un intérêt grandissant sur la gravité de la maladie d'Alzheimer qui joue un rôle important en raison de son incidence extrêmement élevée et de sa fatalité. Plusieurs stratégies pharmacologiques ont été essayées mais l'Alzheimer continue toujours d'être une maladie incurable. Dans le cadre de ce Doctorat, notre travail a été axé sur l'élaboration de modèles QSAR pour les inhibiteurs des enzymes AChE et BuChE. Par conséquent, dans cette revue bibliographique, l'accent sera mis uniquement sur ces deux enzymes. De nombreux modèles QSAR ont été conçus par utilisation de différentes approches de modélisation. Cette section résume ces travaux, lesquels sont présentés par ordre chronologique de publication.

Dans une étude datant de 1996, Recanatini et al. [54] ont réalisé des études comparatives 2D-QSAR sur trois catégories d'inhibiteurs de l'AChE : les analogues de la physostigmine, les 1,2,3,4-tétrahydroacridines (analogues de la tacrine) et les benzylamines. Les auteurs ont constaté que : i) l'hydrophobicité joue un rôle crucial à la fois dans la physostigmine et dans les classes dérivées de la benzylamine; ii) les effets électroniques sont vitaux pour les interactions montrées par la partie variable des dérivés de benzylamine; et iii) les facteurs stériques sont également importants. Cependant, la colinéarité entre les paramètres stériques et hydrophobes n'a pas permis aux auteurs de tirer toute conclusion finale. Durant la même année, une série de dérivés de 1-benzyl-4- [2-(N-benzoylamino) éthyl] pipéridine et de N-benzylpipéridine benzisoxazoles utilisés comme inhibiteur de l'AChE a fait l'objet d'une étude par Tong et al.[55]. Comme technique de modélisation, ils ont utilisé l'analyse comparative du champ moléculaire (CoMFA). L'étude a permis aux auteurs de suggérer ce qui suit : i) les substitutions avec des groupes volumineux et/ou lipo-philiques sur les fractions benzisoxazole et benzoyle sont importantes pour l'activité; ii) l'oxygène dans l'anneau isoxazole, s'il est remplacé par un atome moins électronégatif comme l'azote ou le soufre, diminue la puissance; iii) la

basicité de l'atome de nitrogène dans l'anneau N-pipéridine est importante pour contribuer à l'activité; et iv) occuper la position ortho de la fraction benzoyl avec un volume stérique a un effet négatif sur l'activité.

Dans une approche 2D-QSAR élaborée par utilisation de divers composés (dérivés de la 2-amino-4,6-dime-thylpyridine, des aryl (alkyl) carboxamides, des thiocarbamides et des amidrazones) inhibiteurs des enzymes AChE et BuChE, Debord et al.[56] ont constaté que l'affinité de liaison était améliorée par l'augmentation du volume moléculaire, l'insertion d'un groupe méthylène entre l'amide carbonyle et le noyau aromatique et par le remplacement de l'oxygène de l'amide par du soufre.

Kaur et Zhang [57] ont élaboré un modèle basé sur des données de modélisation moléculaire et des analyses 2D-QSAR de toutes les classes thérapeutiquement importantes d'inhibiteurs de l'AChE (AChEI), comme les dérivés de physostigmine, tacrine, donepezil et huperzine A . Il a été conclu que tous les inhibiteurs étaient de nature hydrophobe comme suggéré par la présence de logP dans la majorité des modèles QSAR. De plus, il a été observé que toutes les classes d'inhibiteurs contenaient de l'azote ionisable. Durant la même année, Recanatini et al. [58] ont synthétisé et analysé une série d'analogues de la tétrahydroacridine (tacrine) substitués en sixième et septième position du noyau acridine et portant des groupes sélectionnés sur la fonction 9-amino. Les analyses 2D-QSAR et CoMFA ont fourni deux modèles qui ont mis en évidence l'effet stérique négatif des substituants en septième position et la liberté stérique relative de la sixième position.

Dans une publication de 2001, Yazal et al. [59] ont appliqué une combinaison d'analyse conformationnelle et de modèles 3D-pharmacophore sur une collection d'AChEI organophosphorés. Les modèles de pharmacophores 3D résultants sont caractérisés par au moins un site accepteur de liaison hydrogène et 2-3 sites hydrophobes. Les modèles démontrent un degré élevé de corrélation entre les puissances inhibitrices calculées et expérimentalement mesurées. Durant la même année, Sippl et al. [60] ont réalisé la construction, la validation et l'application d'un modèle 3D-QSAR basé sur la structure de 42 composés d'aminopyridazine comme AChEI. Les résultats ont été pris en compte pour la conception de nouveaux composés.

En 2005, une analyse QSAR pour l'inhibition de l'AChE et de la BuChE par la lovastatine, la simvastatine, bésylate d'amlodipine, la nifédipine, le chlorhydrate d'hydralazine, le diazépam et le chlorhydrate de chlordiazépoxide a été entreprise par Chiou et al. [61]. Du fait que les valeurs de pKi pour les deux inhibitions de la cholinestérase sont corrélées linéairement les unes avec les autres, les auteurs suggérèrent que les deux inhibitions d'enzymes passent par un mécanisme commun. Comme le bésylate d'amlodipine est un inhibiteur très puissant des deux cholinestérases, il peut, être utile dans le traitement de la M.A. Une année après, Akula et al. [62] ont publié des études 3D-QSAR sur des composés de bis-tacrine en utilisant des scores d'amarrage moléculaire en plus de " *Comparative Molecular Field Analysis* " (CoMFA). Les scores d'amarrage sont utilisés comme descripteurs moléculaires avec les valeurs de champ

stérique et électrostatique obtenues à partir de CoMFA, et 16 molécules sont mises de côté dans l'ensemble d'apprentissage. Le modèle structure-activité est validé sur un ensemble de test ne comportant que 3 molécules. Par conséquent, la puissance prédictive du modèle QSAR développé ne peut être fiable. Dans une autre étude de la même année, Fernández et al. [63] ont appliqué des réseaux de neurones génétiques régularisés bayésiens (BRGNN) à 136 analogues de la tacrine. Le modèle résultant est évalué en faisant la moyenne de plusieurs ensembles de validation générés en tant que membres d'ensembles de réseaux neuronaux (NNE en anglais : Neural Network ensembles) d'ensembles divers. Des statistiques fiables ont été obtenues.

La recherche bibliographique effectuée nous a permis de constater que l'année 2007 a été prolifique en publications traitant l'inhibition des enzymes AChE et BuChE. Jung et al. [64] ont construit des modèles QSAR en utilisant des sélections variables basées sur la régression linéaire multiple (MLR), l'algorithme génétique (GA-MLR) et " *Simulated Annealing* " (SA-MLR). Ils ont utilisé 68 molécules (Tacrines) dans l'ensemble d'apprentissage et 12 dans celui de test. Le meilleur modèle est obtenu par SA-MLR avec une plus grande capacité d'explication et de prédiction, et donc un plus petit écart-type (S). Sur la base de leurs modèles, les auteurs suggèrent des rôles importants pour les interactions hydrophobes et électrostatiques. Une étude 3D-QSAR publiée par Sheng et al. [65] durant la même année a concerné des dérivés de 2-phénoxy-indan-1-one ayant des activités inhibitrices de l'AChE. Les résultats montrent que les contributions à l'activité des champs stériques sont supérieures à celles des champs électrostatiques. L'ajout de CoMSIA a permis d'élucider le rôle de la liaison hydrophobe et de l'hydrogène ainsi que l'effet des propriétés stériques et électrostatiques révélées par le CoMFA. Par ailleurs, un modèle 3D-QSAR élaboré par Liu et al. [66] par utilisation de dérivés de la phénylpenténone, a montré une capacité à prédire les activités de nouveaux inhibiteurs, et serait utile donc pour la conception future de nouveaux inhibiteurs de l'AChE. Une autre étude QSAR pour explorer le mode de liaison des dérivés du 1-indanone substitués en 2 avec l'AChE a été menée par Shen et al. [67]. Des modèles 3D-QSAR hautement fiables et prédictifs sont établis par des méthodes CoMFA et CoMSIA, qui sont ensuite validées avec succès avec un ensemble de test externe. A la suite de ces modèles mathématiques, une meilleure compréhension de l'interaction entre les inhibiteurs et l'AChE est fournie.

Trois des publications parues en 2008 ont retenu notre attention. Zaheer-ul et al. [68] ont développé des modèles 3D-QSAR (CoMFA et CoMSIA) pour une série d'alcaloïdes stéroïdiens en tant qu'inhibiteurs de la BuChE. Il est apparu que l'effet hydrophobe joue un rôle principal dans l'activité inhibitrice de BuChE. En outre, le groupe volumineux en position R2 augmente l'activité, mais en même temps, le substituant plus volumineux en R3 diminue l'activité. Saracoglu et al. [69] ont effectué des analyses QSAR des inhibiteurs de l'AChE liés à la Tacrine et aux analogues de la Tacrine 11 H-Indeno - [1,2-b] -quinoline-10-ylamine tétracyclique sur un ensemble d'apprentissage de 44 composés. Certains fragments moléculaires spécifiques des composés actifs et inactifs sont révélés. Pour leur

part, Roy et al. [70] ont réalisé les analyses CoMFA et CoMSIA sur une série d'inhibiteurs de l'AChE constituée par 78 dérivés de carbamates. Les modèles générés ont suggéré que les interactions stériques, électrostatiques et hydrophobes jouent un rôle important dans la description de la variation de l'affinité de liaison. Pour améliorer l'activité, ils ont suggéré que l'azote carbamoyle devrait être plus électropositif, les substitutions sur cet azote devraient avoir un encombrement stérique et une hydrophobicité élevée.

Une étude de 2009 de Kuzmin et al. [71] portant sur une modélisation QSAR de l'inhibition de l'AChE par divers composés organophosphorés a été rapportée. Les approches par effet de levier et par domaine d'applicabilité ont été utilisées pour évaluer la qualité du modèle. Il a été constaté que la stéréochimie de l'environnement chiral de l'atome asymétrique du phosphore est essentielle pour l'inhibition de l'AChE. Ainsi, les isomères (R) sont toujours moins actifs que les isomères (S) et le racémate. Pour leur part, Solomon et al. [72] ont étudié les QSAR d'une série de 88 dérivés de N-aryle présentant une activité inhibitrice variée à la fois de l'acétylcholinestérase et de la butyrylcholinestérase et ce en utilisant des descripteurs topologiques, de formes moléculaires, électroniques et structurales. Le pouvoir prédictive du modèle QSAR a été évalué en utilisant un ensemble de test de 26 composés pour AChE (r^2_{pred} = 0,857 et q^2 = 0,803) et 20 composés pour BuChE (r^2_{pred} = 0,882 et q^2 = 0,857). Les résultats ont montré qu'AlogP98, Wiener, Kappa-1-AM, Dipole-Mag et CHI-1 sont les descripteurs importants décrivant efficacement l'activité des composés. Une analyse 3D QSAR a été réalisée par Sammi et al. [73] sur une série de 67 analogues de benzodiazépines rapportés comme inhibiteurs de la γ-secrétase en utilisant l'analyse de champ moléculaire (MFA), avec G/PLS. L'étude MFA a été réalisée en utilisant un ensemble de 54 composés. Le pouvoir prédictive du modèle développé a été évalué en utilisant un ensemble de test de 13 composés (r^2_{pred} aussi élevé que 0,729). Le modèle MFA analysé a démontré un bon ajustement, ayant une valeur r^2 de 0,858 et r^2_{cv} de 0,790. L'analyse du meilleur modèle de MFA a fourni un aperçu de la modification possible des molécules pour une meilleure activité.

Dans leur travail publié en 2010, Sharma et al. [74] ont appliqué une approche QSAR sur les dérivés de l'indanone et de l'aurone en utilisant divers paramètres physico-chimiques. La relation structure-activité est établie au moyen d'une analyse de régression linéaire multiple séquentielle pour un ensemble de 23 composés utilisé dans l'ensemble d'apprentissage, alors que la validation est effectuée avec 9 composés. Le meilleur modèle trouvé, statistiquement significatif, inclut l'énergie orbitale moléculaire (LUMO), le diamètre et l'énergie libre de Gibbs comme descripteurs moléculaires. Takahashi et al. [75] ont synthétisé une série de dérivés de norcymsérine N1-substitués et évalué leurs activités inhibitrices sur l'AChE. L'étude 2D-QSAR a montré que log*S* pourrait être la caractéristique clé de l'activité améliorée et que l'azote ionisable du substituant contribue également à cette amélioration.

Gupta et al. [76] ont développé des modèles QSAR comparatifs pour 42 inhibiteurs (pipéridines, tétrahydroacridines, tétrahydroazépines et les carbamates) de l'AChE se liant

au site anionique catalytique et périphérique, identifiés sur la base de l'approche de l'amarrage moléculaire. Les modèles QSAR sont développés à l'aide des techniques RNA (ANN en anglais : Artificial Neural Network), GA (Genetic Algorithms), SVM (Support Vector Machine) et PLS génétique. L'ensemble d'apprentissage comprend 31 molécules alors que celui du test externe en compte 11 molécules. En conclusion, les modèles générés utilisant des descripteurs thermodynamiques, électrotopologiques et électroniques montrent que les méthodes non linéaires sont plus robustes que les méthodes linéaires. Dans leur travail, Lu et al. [77] ont développé des modèles 3D-pharmacophores qualitatifs et quantitatifs basés sur des inhibiteurs (polyamines, cétones, inhibiteurs à base de Donezil, composés hybrides à base de Tacrine et de benzofurane) de l'AChE. 62 sont sélectionnés dans le cadre de l'ensemble d'apprentissage et 26 dans celui du test. Le meilleur modèle de pharmacophore à cinq caractéristiques inclut un donneur de liaison hydrogène et quatre caractéristiques hydrophobes, et est appliqué pour identifier neuf nouveaux inhibiteurs. Durant la même année, Araújo et al. [78] ont élaboré plusieurs modèles 3D-QSAR en utilisant 60 structures de benzylpipéridine. Une approche combinant les GA et les moindres carrés partiels (PLS) est utilisée comme méthodologie de calcul pour développer et valider de tels modèles. L'ensemble d'apprentissage comprend 47 molécules et l'ensemble de test englobe les 13 autres. Les descripteurs utilisés sont des énergies d'interaction stérique (Lennard-Jones) et électrostatique (Coulomb), calculées entre chaque ligand et les résidus HuAChE. Selon les deux meilleurs modèles calculés, la contribution de Lennard-Jones et la somme des contributions de Lennard-Jones et Coulomb sont plus importantes que celles de Coulomb sur la relation entre structure et activité.

Un ensemble de données de 91 composés constitués de 9 classes différentes d'inhibiteurs de l'AChE (hybrides tacrine-8-hydroxyquinoline, tacripyrines, dérivés de donepezil-tacrine, tétrahydroacridine, hybrides à base de benzofuran, analogues de rivastigmine, hybrides tacrine-mélatonine, carbamates de tétrahydrofurobenzofuran et de methanobenzodioxepine) ont été considérés pour le développement de modèles 3D-QSAR par Deb et al. [79]. Ces modèles de prédiction ont été élaborés en utilisant le S-MLR basé sur le consensus de différents docking et leurs méthodes de notation (GOLD, Ligand Fit et GLIDE). Yan et Wang [80] ont développé quatre modèles 2D-QSAR en utilisant les techniques MLR et SVM (Support vector machine) sur divers inhibiteurs (au nombre de 404) de l'AChE. Seize descripteurs reflétaient la structure et les caractéristiques fondamentales dont la charge, l'électronégativité, la polarisabilité et la capacité de liaison H de ces molécules. Il s'est avéré que ces propriétés chimiques contribuaient toutes à l'inhibition de l'AChE. Bitencourt et al. [81] ont réalisé une analyse d'image multi variée-QSAR (MIA-QSAR) sur un ensemble de 34 composés, dont des amines quaternaires et des carbamates ayant une activité anti-AChE connue. Cette analyse, qui met en corrélation les images de structures chimiques avec leurs activités biologiques, a mis en évidence l'importance d'un groupe phénol avec un échafaudage carbamate en position méta du

cycle benzène, pour améliorer l'activité inhibitrice de l'AChE. Durant la même année, De Souza et al. [82] ont développé des modèles HQSAR pour une série de 36 inhibiteurs de 4-(diéthylamino) méthyl-phénol en utilisant la méthode HQSAR. Les cartes de contribution HQSAR ont été utilisées pour expliquer l'importance du fragment structurel dans l'activité globale de cette série et il a été trouvé que les fragments structuraux contenant des fragments aromatiques et de longues chaînes latérales augmentent la puissance.

Chitranshi et al. [83] ont développé en 2013, un modèle 2D-QSAR basé sur 44 dérivés de 1-indanone comme inhibiteurs de l'AChE. Le modèle a été développé en utilisant cinq descripteurs riches en informations, à savoir, l'accepteur de liaison hydrogène, logP, la chaleur de formation, l'énergie électronique et le moment dipolaire, qui a joué un rôle important dans la détermination de l'activité inhibitrice. En utilisant une combinaison de modélisation pharmacophore, de criblage virtuel, de docking moléculaire et de QSAR, ces auteurs ont identifié avec succès plusieurs autres nouveaux échafaudages comme benzothiazole, benzylpipérazine, benzylpipéridine, pyrazole, pipéridine, pyridine et dérivés de thiazolidine, qui n'ont pas été rapportés précédemment dans la littérature comme inhibiteurs de l'AChE pour le traitement de la M.A. Au cours de la même année, dans le but de comparer la quercetine avec les inhibiteurs classiques de l'AChE, Islam et al. [84] ont réalisé une étude QSAR. Comme premier résultat, cette étude a montré de manière concluante la supériorité de la quercetine naturelle par rapport aux médicaments conventionnels en tant qu'inhibiteur de l'AChE. Pour découvrir un inhibiteur encore meilleur, les auteurs ont procédé à une modification de la structure de base de la quercetine avec différents groupes fonctionnels. L'analyse QSAR des composés dérivés a montré que la méthylation de l'O-4 était spécifiquement responsable d'une meilleure affinité que celle de la quercetine mère.

Dans la littérature parue au cours des cinq dernières années (de 2014 à 2018), le nombre de tentatives de conception de médicaments liés aux approches QSAR concernant de nouveaux traitements contre la maladie d'Alzheimer est consistant. Ambure et al. [85] ont utilisé le même ensemble de données qu'ont utilisé Yan et Wang [80] pour développer un modèle de pharmacophore à large spectre, puis ont examiné la base de données InterBioScreen (consistant en des composés naturels uniquement) en utilisant le pharmacophore développé. De plus, des études d'amarrage ont été menées qui ont conduit à la sélection des hits finaux constitués de potentiels inhibiteurs naturels de l'AChE. L'étude des analogues de la Tacrine présente un intérêt continu, et pour cette raison, Wong et al. [86] ont établi des modèles QSAR sur leur activité inhibitrice de l'acétylcholinestérase. Pour ce faire, dix groupes de nouveaux inhibiteurs liés à la Tacrine ont été explorés. Les modèles QSAR développés dans ce travail ont une capacité prédictive satisfaisante, et sont obtenus en sélectionnant les descripteurs moléculaires les plus représentatifs de la structure chimique, représentés par plus d'un millier de types de descripteurs constitutionnels, topologiques, géométriques, quantiques et électroniques.

D'autre part, un modèle QSAR (GQSAR) robuste et très prédictif basé sur les groupes a été développé par Goyal et al. [87] sur la base des descripteurs calculés pour les fragments de 20 dérivés de 1,4-dihydropyridine (DHP). Cette étude ouvre la voie à la considération de ces molécules en tant que médicaments potentiels pour l'inhibition efficace de l'AChE. Le modèle GQSAR fournit des indices spécifiques au site sur les molécules où certaines modifications peuvent entraîner une augmentation de l'activité biologique. Cette information pourrait être d'une grande valeur pour la conception et le développement de médicaments multifonctionnels pour lutter contre la maladie d'Alzheimer. Au cours de cette même année 2014, Bautista-Aguilera et al. [88] ont travaillé sur la conception, la synthèse et l'évaluation pharmacologique d'amines à base de donépézil-indolyle, d'amides et de dérivés d'acides carboxyliques, capables d'inhiber la cholinestérase (ChE) des enzymes pour le traitement potentiel de la maladie d'Alzheimer (AD). Des études théoriques utilisant la relation structure-activité quantitative (3D-QSAR) ont été utilisées pour définir les pharmacophores 3D pour l'inhibition des enzymes AChE et BuChE. Ces auteurs ont trouvé qu'en général, et pour le même substituant, les amines sont des inhibiteurs de ChE plus puissants que les amides correspondants. Cependant, ils suggèrent que certains dérivés N-allyliques et N-morpholine analogues méritent également une attention car ils présentent un profil d'inhibition efficace. En résumé, le donépézil-indolyle hybride15 est un médicament prometteur pour la poursuite du développement de la prévention et du traitement de la maladie d'Alzheimer.

Dans un article paru en 2015, des modèles 3D-QSAR ont été réalisés par Zhou et al. [89] afin d'étudier 60 dérivés de la Tacrine et leurs activités inhibitrices sur l'AChE. La modélisation 3D-QSAR a abouti à un modèle CoMFA optimal avec q^2 = 0,552 et r^2 = 0,983 et un modèle CoMSIA optimal avec q^2 = 0,581 et r^2 = 0,989. Ces modèles QSAR ont également montré que les champs de liaisons stériques et liaisons hydrogènes de ces composes ont une influence importante sur leurs activités. Les auteurs ont conclu que les résultats de cette étude améliorent la compréhension des mécanismes des inhibiteurs de l'AChE et fournissent des informations précieuses qui devraient aider à la conception de nouveaux inhibiteurs potentiels de l'AChE. En 2016, un grand ensemble de valeurs de CI_{50} non redondantes de 2570 composés inhibiteurs de l'AChE est utilisé dans une étude QSAR par Simeone et al. [90]. Les inhibiteurs de l'AChE ont été décrits par un ensemble de 12 descripteurs et des modèles prédictifs ont été construits à partir de 100 différentes répartitions des données. Le meilleur modèle construit en utilisant le comptage des substructures a été sélectionné selon les directives de l'OCDE et a donné des valeurs R^2, Q^2_{CV} et Q^2_{ext} de 0,92 ± 0,01, 0,78 ± 0,06 et 0,78 ± 0,05, respectivement. Les auteurs ont suggéré que toutes les informations obtenues peuvent être utilisées comme lignes directrices pour la conception de nouveaux inhibiteurs de l'AChE robustes.

Zhang et al. [91] ont réalisé des études QSAR comparatives sur certaines huprines inhibitrices vis-à-vis de l'AChE en utilisant l'analyse comparative de champs moléculaires (CoMFA), l'analyse des indices de similarité moléculaire comparative (CoMSIA) et

l'hologramme QSAR (HQSAR). Trois modèles QSAR hautement prédictifs ont été construits avec succès en fonction de l'ensemble d'apprentissage. Les modèles CoMFA, CoMSIA et HQSAR ont des valeurs respectives de r^2= 0,988, q^2 = 0,757, ONC = 6 ; r^2= 0,966, q^2 = 0,645, ONC = 5 ; et r^2= 0,957, q^2 = 0,736, ONC = 6. Le pouvoir prédictif a été évalué en utilisant un ensemble de tests externes, et les valeurs prédites de r^2obtenues pour les trois modèles étaient de 0,984, 0,973 et 0,783, respectivement. Sur la base de cette étude QSAR, 14 nouvelles molécules puissantes ont été conçues et six d'entre elles sont plus actives que le meilleur composé actif rapporté auparavant dans la littérature. Aussi, les auteurs conclurent que les modèles QSAR finaux pourraient être utiles dans la conception et le développement de nouveaux inhibiteurs actifs de l'AChE. Ahmadi et Ganji [92] ont établi en 2016 des modèles QSPR avec 5 descripteurs par utilisation de 88 dérivés N-aryle comme inhibiteurs de la butyrylcholinestérase (BChE). Les valeurs expérimentales pour les constantes d'inhibition ont été obtenues à partir de la littérature. Les résultats de cette étude ont montré qu'une GA-MLR fonctionne généralement mieux que la Stepwise MLR. L'équation QSPR peut être utile dans la conception de nouveaux dérivés N-aryliques comme inhibiteurs de la butyrylcholinestérase avec une activité d'inhibition améliorée.

Dans une étude très récente menée par Pang et al. [93], une série de nouveaux dérivés de DL0410 (identifié comme inhibiteur de l'AChE et de la BuChE) contenant des squelettes de diphényle et de pipéridine ont été évalués pour leurs activités inhibitrices sur AChE et BuChE. Des modèles de pharmacophores et 3D-QSAR ont été élaborés. Pour l'ensemble d'apprentissage, le coefficient R^2 du modèle 3D-QSAR de l'AChE et de la BuChE s'est révélé être de 0,925 et 0,883, alors que pour l'ensemble de test les valeurs trouvées étaient de 0,850 et 0,881, respectivement. Les auteurs ont conclu que ces études fournissent un meilleur aperçu des comportements inhibiteurs des dérivés de DL0410, ce qui est bénéfique pour la conception rationnelle des inhibiteurs de l'AChE et de la BuChE dans le futur. Durant la même année 2017, Das et al. [94] ont réalisé une étude QSAR pour prédire le flavonoïde le plus efficace des classes prénylées et pyrano-flavonoïdes pour l'inhibition de l'AChE. Trois flavonoïdes isolés d'Artocarpus anisophyllus ont été sélectionnés pour l'étude. D'après l'analyse QSAR, un ces trois flavonoïdes présentait une CI_{50} de 1659,59 nM. Les résultats obtenus indiquent que les flavonoïdes pourraient être des inhibiteurs efficaces de l'AChE et pourraient donc être utiles dans la prise en charge de la maladie d'Alzheimer. D'autre part, deux ensembles différents d'inhibiteurs de l'AChE, un ensemble de données I (30 composés) et un ensemble de données II (20 composés) ont été utilisés par Pulikkal et al. [95] pour développer des modèles QSAR linéaire (M.L.R) et non linéaires (S.V.M) pour déterminer les descripteurs chimiques responsables de l'activité sur l'AChE. Les modèles QSAR obtenus ont été trouvés statistiquement ajustés, stables et prédictifs sur les échelles de validation. Les descripteurs MATS5m, RDF045m, MATS5e, HATSe et Mor17e se retrouvent pour les deux ensembles de données. Les auteurs pensent que ces résultats pourraient être utilisés pour concevoir de nouveaux inhibiteurs de l'AChE avec une activité biologique améliorée.

Dans une autre étude élaborée par Gurung et al. [96] et parue en 2017, diverses méthodes de construction de modèles 3D-QSAR telles que la régression linéaire multiple (M.L.R), les moindres carrés partiels et la régression des composantes principales ont été utilisées moyennant l'utilisation des descripteurs de champs stériques et électrostatiques. Les dérivés d'Isoalloxazine ont été amarrés contre l'AChE humaine, ce qui a révélé des résidus critiques impliqués dans les liaisons hydrogène ainsi que des interactions hydrophobes. Un modèle statistiquement significatif a été obtenu en utilisant M.L.R couplé avec une méthode de sélection par étapes ayant r^2 = 0,9405, q^2 = 0,6683, r^2_{pred} = 0,666 et erreur standard valant 0,2491. Selon les auteurs, leur étude peut être utile dans la conception de dérivés d'Isoalloxazine plus puissants en tant qu'inhibiteurs de l'AChE.

Durant cette année 2018, deux articles ont retenu notre attention. Faisant la constatation que les traitements actuels sont limités du fait qu'ils sont basés sur une seule cible, Kumar et al. [97] ont énuméré un certain nombre de cibles les plus importantes pour la conception de médicaments anti-Alzheimer. De plus, ils ont préconisé une approche qui module simultanément plus d'une cible pouvant être réalisée par la pharmacologie de réseau ou le terme plus récemment proposé "médicaments conçus multi-cibles". Ils ont ainsi conclu que les conceptions de médicaments computationnelles jouent un rôle important dans l'optimisation de la découverte de médicaments multi-cibles. Le second article est l'œuvre de Jiang et Gao [98], qui ont utilisé des flavonoïdes avec des activités inhibitrices sur l'AChE et une faible toxicité pour développer de nouveaux agents anti-Alzheimer. Dans cette étude, un modèle de pharmacophore 3D-QSAR a été élaboré sur la base d'un ensemble de flavonoïdes choisi par dépistage virtuel à partir de bases de données de médecine traditionnelle chinoise. A partir des analyses d'amarrage, trois candidats potentiellement inhibiteurs de l'AChE ont finalement été obtenus.

I.7 CONCLUSION

Les travaux publiés sur l'approche QSAR appliquée à la recherche de médicaments anti-Alzheimer au cours des 40 dernières années ont été répertoriés. En premier lieu, eu égard au nombre de publications, il y'a lieu de relever cette importante quête des scientifiques à chercher sans cesse de nouveaux médicaments vu que ceux actuellement prescrits pour cette maladie se manifestent par des effets secondaires. En second lieu, malgré que l'enzyme AChE demeure une cible privilégiée et unique dans la stratégie thérapeutique pour le traitement de la maladie d'Alzheimer, il faut compter sur la nouvelle approche qui consiste à considérer non seulement une seule mais plusieurs cibles à la fois. La troisième remarque est inhérente à l'évolution de l'approche in silico d'années en années. Après l'utilisation des modèles 2D-QSAR simples avant l'année 2000, les années suivantes (2000-2018) ont vu l'utilisation d'autres approches plus sophistiquées comme la 3D-QSAR, la HQSAR, la conception moléculaire, la modélisation du pharmacophore ou encore la dynamique moléculaire.

Références

1. Brunko, P. and M. Forum International sur la Recherche et le, *Recherche et innovation pharmaceutique à l'approche du troisième millénaire : 4e Forum internationale sur la recherche et le médicament, Bordeaux, 7-13 octobre 1993*. John Libbey Eurotext ed. 1994, Montrouge. 210.
2. Morgan, S., et al., *The cost of drug development: a systematic review.*Health policy, 2011. **100**(1): p. 4-17.
3. Landry, Y. and J.P. Gies, *Pharmacologie - 3e édition: Des cibles à la thérapeutique*. 2014: Dunod. 544.
4. Bajorath, J., *Integration of virtual and high-throughput screening.*Nature Reviews Drug Discovery, 2002. **1**(11): p. 882.
5. Mouly, S., et al., *Évaluation de l'efficacité d'un médicament: de la découverte à la mise sur le marché.* Journal Français d'Ophtalmologie, 2008. **31**(1): p. 75-79.
6. Le Guilloux, V., *Développement de méthodes et d'outils chémoinformatiques pour l'analyse et la comparaison de chimiothèques*. 2013, Université d'Orléans.
7. Adams, C.P. and V.V. Brantner, *Estimating the cost of new drug development: is it really $802 million?* Health affairs, 2006. **25**(2): p. 420-428.
8. DiMasi, J.A., *The value of improving the productivity of the drug development process.* Pharmacoeconomics, 2002. **20**(3): p. 1-10.
9. DiMasi, J.A., R.W. Hansen, and H.G. Grabowski, *The price of innovation: new estimates of drug development costs.* Journal of health economics, 2003. **22**(2): p. 151-185.
10. Adams, C.P. and V.V. Brantner, *Spending on new drug development.* Health economics, 2010. **19**(2): p. 130-141.
11. DiMasi, J.A., H.G. Grabowski, and R.W. Hansen, *Innovation in the pharmaceutical industry: new estimates of R&D costs.* Journal of health economics, 2016. **47**: p. 20-33.
12. Kola, I. and J. Landis, *Can the pharmaceutical industry reduce attrition rates?*Nature reviews Drug discovery, 2004. **3**(8): p. 711.
13. Ghemtio, L., *Simulation numérique et approche orientée connaissance pour la découverte de nouvelles molécules thérapeutiques*. 2010, Université Henri Poincaré-Nancy I.
14. Young, D.C., *Computational Drug Design: A Guide for Computational and Medicinal Chemists*. 2009: Wiley.
15. Smith, J.A., et al., *Principles and practice in ethical review of animal experiments across Europe: summary of the report of a FELASA working group on ethical evaluation of animal experiments.*Laboratory Animals, 2007. **41**(2): p. 143-160.
16. Universalis, E. and L.G. Articles, *Pharmacologie*. Les Grands Articles d'Universalis. 2015: Encyclopaedia Universalis. 60.
17. W Caldwell, G., et al., *The IC50 concept revisited.* Current topics in medicinal chemistry, 2012. **12**(11): p. 1282-1290.
18. Mayers, D.L., et al., *Antimicrobial Drug Resistance: Clinical and Epidemiological Aspects*. 2017: Springer International Publishing. 1627.
19. Ekins, S., J. Mestres, and B. Testa, *In silico pharmacology for drug discovery: applications to targets and beyond.* British journal of pharmacology, 2007. **152**(1): p. 21-37.

20. Aguda, B.D., et al., *An in silico modeling approach to understanding the dynamics of sarcoidosis.* PloS one, 2011. **6**(5): p. e19544.
21. Martonen, T., et al., *In silico modeling of asthma.* Advanced drug delivery reviews, 2003. **55**(7): p. 829-849.
22. Golbamaki, A., et al., *Comparison of in silico models for prediction of Daphnia magna acute toxicity.* SAR and QSAR in Environmental Research, 2014. **25**(8): p. 673-694.
23. Raunio, H., et al., *In vitro methods in the prediction of kinetics of drugs: focus on drug metabolism.* Alternatives to laboratory animals: ATLA, 2004. **32**(4): p. 425-430.
24. Van Leeuwen, K., et al., *Using chemical categories to fill data gaps in hazard assessment.* SAR and QSAR in Environmental Research, 2009. **20**(3-4): p. 207-220.
25. R Munteanu, C., et al., *Drug discovery and design for complex diseases through QSAR computational methods.* Current pharmaceutical design, 2010. **16**(24): p. 2640-2655.
26. Ferreira, L.G. and A.D. Andricopulo, *Medicinal chemistry approaches to neglected diseases drug discovery.* Journal of Modern Medicinal Chemistry, 2014. **2**(1): p. 20-30.
27. Zhang, L., et al., *Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening.* Journal of chemical information and modeling, 2013. **53**(2): p. 475-492.
28. Pauli, I., et al., *Discovery of new inhibitors of Mycobacterium tuberculosis InhA enzyme using virtual screening and a 3D-pharmacophore-based approach.* Journal of chemical information and modeling, 2013. **53**(9): p. 2390-2401.
29. Ferreira, R.S., et al., *Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors.* Journal of medicinal chemistry, 2010. **53**(13): p. 4891-4905.
30. Castilho, M.S., et al., *Structural basis for selective inhibition of purine nucleoside phosphorylase from Schistosoma mansoni: kinetic and structural studies.* Bioorganic & medicinal chemistry, 2010. **18**(4): p. 1421-1427.
31. Amaro, R.E., et al., *Discovery of drug-like inhibitors of an essential RNA-editing ligase in Trypanosoma brucei.* Proceedings of the National Academy of Sciences, 2008. **105**(45): p. 17278-17283.
32. Abildstrom, S., et al., *Trends in incidence and case fatality rates of acute myocardial infarction in Denmark and Sweden.* Heart, 2003. **89**(5): p. 507-511.
33. Jemal, A., et al., *Trends in the leading causes of death in the United States, 1970-2002.* Jama, 2005. **294**(10): p. 1255-1259.
34. Castilho, M.S., R.V. Guido, and A.D. Andricopulo, *2D Quantitative structure–activity relationship studies on a series of cholesteryl ester transfer protein inhibitors.* Bioorganic & medicinal chemistry, 2007. **15**(18): p. 6242-6252.
35. Politi, A., et al., *Application of 3D QSAR CoMFA/CoMSIA and in silico docking studies on novel renin inhibitors against cardiovascular diseases.* European journal of medicinal chemistry, 2009. **44**(9): p. 3703-3711.
36. Jana, D., et al., *Chemometric modeling and pharmacophore mapping in coronary heart disease: 2-arylbenzoxazoles as cholesteryl ester transfer protein inhibitors.* MedChemComm, 2011. **2**(9): p. 840-852.

37. Cooke, M.S., et al., *Oxidative DNA damage: mechanisms, mutation, and disease.* The FASEB Journal, 2003. **17**(10): p. 1195-1214.
38. Halliwell, B., *Drug antioxidant effects.* Drugs, 1991. **42**(4): p. 569-605.
39. Kirkegaard, T., et al., *Amplified in breast cancer 1 in human epidermal growth factor receptor–positive tumors of tamoxifen-treated breast cancer patients.* Clinical Cancer Research, 2007. **13**(5): p. 1405-1411.
40. Welsh, J.B., et al., *Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer.* Cancer research, 2001. **61**(16): p. 5974-5978.
41. Rivera, M.P. and D.E. Stover, *Gender and lung cancer.* Clinics in chest medicine, 2004. **25**(2): p. 391-400.
42. Jain, K.K., *Recent advances in clinical oncoproteomics.* Journal of BU ON.: official journal of the Balkan Union of Oncology, 2007. **12**: p. S31-8.
43. González-Díaz, H., et al., *ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds.* European journal of medicinal chemistry, 2007. **42**(5): p. 580-585.
44. Zhang, S., et al., *Antitumor agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents.* Journal of computer-aided molecular design, 2007. **21**(1-3): p. 97-112.
45. Alzheimer's, A., *2015 Alzheimer's disease facts and figures.* Alzheimer's & dementia: the journal of the Alzheimer's Association, 2015. **11**(3): p. 332.
46. Goedert, M. and M.G. Spillantini, *A century of Alzheimer's disease.* science, 2006. **314**(5800): p. 777-781.
47. McDowell, I., *Alzheimer's disease: insights from epidemiology.* Aging Clinical and Experimental Research, 2001. **13**(3): p. 143-162.
48. Schmidt, R., et al., *Therapy of Alzheimer's disease: current status and future development.* Neuropsychiatrie: Klinik, Diagnostik, Therapie und Rehabilitation: Organ der Gesellschaft Osterreichischer Nervenarzte und Psychiater, 2008. **22**(3): p. 153-171.
49. Ambure, P. and K. Roy, *Advances in quantitative structure–activity relationship models of anti-Alzheimer's agents.* Expert opinion on drug discovery, 2014. **9**(6): p. 697-723.
50. del Mar Alcalá, M., et al., *Characterisation of the anticholinesterase activity of two new tacrine–huperzine A hybrids.* Neuropharmacology, 2003. **44**(6): p. 749-755.
51. Quinn, D.M., et al., *Delineation and decomposition of energies involved in quaternary ammonium binding in the active site of acetylcholinesterase.* Journal of the American Chemical Society, 2000. **122**(13): p. 2975-2980.
52. Bartus, R.T., et al., *The cholinergic hypothesis of geriatric memory dysfunction.* Science, 1982. **217**(4558): p. 408-414.
53. Greig, N.H., D.K. Lahiri, and K. Sambamurti, *Butyrylcholinesterase: an important new target in Alzheimer's disease therapy.* International Psychogeriatrics, 2002. **14**(S1): p. 77-91.
54. Recanatini, M., A. Cavalli, and C. Hansch, *A comparative QSAR analysis of acetylcholinesterase inhibitors currently studied for the treatment of Alzheimer's disease.* Chemico-biological interactions, 1997. **105**(3): p. 199-228.

55. Tong, W., et al., *A comparative molecular field analysis study of N-benzylpiperidines as acetylcholinesterase inhibitors.* Journal of medicinal chemistry, 1996. **39**(2): p. 380-387.
56. Debord, J., et al., *Cholinesterase inhibition by derivatives of 2-amino-4, 6-dimethylpyridine.* Journal of enzyme inhibition, 1997. **12**(1): p. 13-26.
57. Kaur, J. and M. Zhang, *Molecular modelling and QSAR of reversible acetylcholinesterase inhibitors.* Current medicinal chemistry, 2000. **7**(3): p. 273-294.
58. Recanatini, M., et al., *SAR of 9-amino-1, 2, 3, 4-tetrahydroacridine-based acetylcholinesterase inhibitors: synthesis, enzyme inhibitory activity, QSAR, and structure-based CoMFA of tacrine analogues.* Journal of medicinal chemistry, 2000. **43**(10): p. 2007-2018.
59. Yazal, J.E., et al., *Prediction of organophosphorus acetylcholinesterase inhibition using three-dimensional quantitative structure-activity relationship (3D-QSAR) methods.* Toxicological Sciences, 2001. **63**(2): p. 223-232.
60. Sippl, W., et al., *Structure-based 3D QSAR and design of novel acetylcholinesterase inhibitors.* Journal of Computer-Aided Molecular Design, 2001. **15**(5): p. 395-410.
61. Chiou, S.-Y., et al., *QSAR for acetylcholinesterase and butyrylcholinesterase inhibition by cardiovascular drugs and benzodiazepines.* Medicinal Chemistry Research, 2005. **14**(5): p. 297-308.
62. Akula, N., et al., *3D QSAR studies of AChE inhibitors based on molecular docking scores and CoMFA.* Bioorganic & medicinal chemistry letters, 2006. **16**(24): p. 6277-6280.
63. Fernandez, M., et al., *Modeling of acetylcholinesterase inhibition by tacrine analogues using Bayesian-regularized Genetic Neural Networks and ensemble averaging.* Journal of enzyme inhibition and medicinal chemistry, 2006. **21**(6): p. 647-661.
64. Jung, M., et al., *Quantitative structure–activity relationship (QSAR) of tacrine derivatives against acetylcholinesterase (AChE) activity using variable selections.* Bioorganic & medicinal chemistry letters, 2007. **17**(4): p. 1082-1090.
65. Sheng, R., et al., *3D-QSAR studies on AChE inhibitory activities of 2-phenoxy-indan-1-one derivatives.* Chinese Journal of Medicinal Chemistry, 2007. **17**(6): p. 348-353.
66. Liu, A., et al., *3D-QSAR analysis of a new type of acetylcholinesterase inhibitors.* Science in China Series C: Life Sciences, 2007. **50**(6): p. 726-730.
67. Shen, L.-l., G.-x. Liu, and Y. Tang, *Molecular docking and 3D-QSAR studies of 2-substituted 1-indanone derivatives as acetylcholinesterase inhibitors.* Acta Pharmacologica Sinica, 2007. **28**(12): p. 2053.
68. Uddin, R., et al., *Receptor-based modeling and 3D-QSAR for a quantitative production of the butyrylcholinesterase inhibitors based on genetic algorithm.* Journal of chemical information and modeling, 2008. **48**(5): p. 1092-1103.
69. Saracoglu, M. and F. Kandemirli, *The investigation of structure-activity relationships of tacrine analogues: electronic-topological method.* The open medicinal chemistry journal, 2008. **2**: p. 75.
70. Roy, K.K., A. Dixit, and A.K. Saxena, *An investigation of structurally diverse carbamates for acetylcholinesterase (AChE) inhibition using 3D-QSAR analysis.* Journal of Molecular Graphics and Modelling, 2008. **27**(2): p. 197-208.

71. Kuz'min, V.E., et al., *Consensus QSAR Modeling of Phosphor-Containing Chiral AChE Inhibitors.* QSAR & combinatorial science, 2009. **28**(6-7): p. 664-677.
72. Solomon, K.A., S. Sundararajan, and V. Abirami, *QSAR studies on N-aryl derivative activity towards Alzheimer's disease.* Molecules, 2009. **14**(4): p. 1448-1455.
73. Sammi, T., O. Silakari, and M. Ravikumar, *Three-dimensional quantitative structure-activity relationship (3D-QSAR) studies of various benzodiazepine analogues of γ-secretase inhibitors.* Journal of molecular modeling, 2009. **15**(4): p. 343-348.
74. Y Wong, K., et al., *QSAR applications during last decade on inhibitors of acetylcholinesterase in Alzheimer's disease.* Mini reviews in medicinal chemistry, 2012. **12**(10): p. 936-946.
75. Takahashi, J., et al., *Design, synthesis, evaluation and QSAR analysis of N 1-substituted norcymserine derivatives as selective butyrylcholinesterase inhibitors.* Bioorganic & medicinal chemistry letters, 2010. **20**(5): p. 1718-1720.
76. Gupta, S., et al., *Molecular docking guided comparative GFA, G/PLS, SVM and ANN models of structurally diverse dual binding site acetylcholinesterase inhibitors.* Molecular informatics, 2011. **30**(8): p. 689-706.
77. Lu, S.-H., et al., *The discovery of potential acetylcholinesterase inhibitors: a combination of pharmacophore modeling, virtual screening, and molecular docking studies.* Journal of Biomedical Science, 2011. **18**(1): p. 8.
78. Araújo, J.Q., et al., *Receptor-dependent (RD) 3D-QSAR approach of a series of benzylpiperidine inhibitors of human acetylcholinesterase (HuAChE).* European journal of medicinal chemistry, 2011. **46**(1): p. 39-51.
79. Deb, P.K., et al., *Molecular docking and receptor-specific 3D-QSAR studies of acetylcholinesterase inhibitors.* Molecular diversity, 2012. **16**(4): p. 803-823.
80. Yan, A. and K. Wang, *Quantitative structure and bioactivity relationship study on human acetylcholinesterase inhibitors.* Bioorganic & medicinal chemistry letters, 2012. **22**(9): p. 3336-3342.
81. Bitencourt, M., M.P. Freitas, and R. Rittner, *The MIA-QSAR Method for the Prediction of Bioactivities of Possible Acetylcholinesterase Inhibitors.* Archiv der Pharmazie, 2012. **345**(9): p. 723-728.
82. de Souza, S.D., et al., *Hologram QSAR models of 4-[(diethylamino) methyl]-phenol inhibitors of acetyl/butyrylcholinesterase enzymes as potential anti-Alzheimer agents.* Molecules, 2012. **17**(8): p. 9529-9539.
83. Chitranshi, N., et al., *New molecular scaffolds for the design of Alzheimer's acetylcholinesterase inhibitors identified using ligand-and receptor-based virtual screening.* Medicinal Chemistry Research, 2013. **22**(5): p. 2328-2345.
84. Islam, M.R., et al., *In silico QSAR analysis of quercetin reveals its potential as therapeutic drug for Alzheimer's disease.* Journal of Young Pharmacists, 2013. **5**(4): p. 173-179.
85. Ambure, P., S. Kar, and K. Roy, *Pharmacophore mapping-based virtual screening followed by molecular docking studies in search of potential acetylcholinesterase inhibitors as anti-Alzheimer's agents.* Biosystems, 2014. **116**: p. 10-20.
86. Wong, K.Y., et al., *QSAR analysis on tacrine-related acetylcholinesterase inhibitors.* Journal of biomedical science, 2014. **21**(1): p. 84.

87. Goyal, M., et al., *Development of dual inhibitors against Alzheimer's disease using fragment-based QSAR and molecular docking.* BioMed research international, 2014. **2014**.
88. Bautista-Aguilera, O.M., et al., *Design, synthesis, pharmacological evaluation, QSAR analysis, molecular modeling and ADMET of novel donepezil–indolyl hybrids as multipotent cholinesterase/monoamine oxidase inhibitors for the potential treatment of Alzheimer's disease.* European journal of medicinal chemistry, 2014. **75**: p. 82-95.
89. Zhou, A., et al., *Combined 3D-QSAR, molecular docking, and molecular dynamics study of tacrine derivatives as potential acetylcholinesterase (AChE) inhibitors of Alzheimer's disease.* Journal of molecular modeling, 2015. **21**(10): p. 277.
90. Simeon, S., et al., *Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking.* PeerJ, 2016. **4**: p. e2322.
91. Zhang, S., et al., *Design and prediction of new acetylcholinesterase inhibitor via quantitative structure activity relationship of huprines derivatives.* Archives of pharmacal research, 2016. **39**(5): p. 591-602.
92. Ahmadi, S. and S. Ganji, *Genetic Algorithm and Self-Organizing Maps for QSPR Study of Some N-aryl Derivatives as Butyrylcholinesterase Inhibitors.* Current drug discovery technologies, 2016. **13**(4): p. 232-253.
93. Pang, X., et al., *Evaluation of novel dual acetyl-and butyrylcholinesterase inhibitors as potential anti-Alzheimer's disease agents using pharmacophore, 3D-QSAR, and molecular docking approaches.* Molecules, 2017. **22**(8): p. 1254.
94. Das, S., et al., *Prediction of Anti-Alzheimer's Activity of Flavonoids Targeting Acetylcholinesterase in silico.* Phytochemical analysis, 2017. **28**(4): p. 324-331.
95. Pulikkal, B.P., et al., *Common SAR Derived from Linear and Non-linear QSAR Studies on AChE Inhibitors used in the Treatment of Alzheimer's Disease.* Current neuropharmacology, 2017. **15**(8): p. 1093-1099.
96. Gurung, A.B., et al., *Identification of molecular descriptors for design of novel Isoalloxazine derivatives as potential Acetylcholinesterase inhibitors against Alzheimer's disease.* Journal of Biomolecular Structure and Dynamics, 2017. **35**(8): p. 1729-1742.
97. Kumar, A., A. Tiwari, and A. Sharma, *Changing Paradigm from one Target one Ligand Towards Multi-target Directed Ligand Design for Key Drug Targets of Alzheimer Disease: An Important Role of In Silico Methods in Multi-target Directed Ligands Design.* Current neuropharmacology, 2018. **16**(6): p. 726-739.
98. Jiang, Y. and H. Gao, *Pharmacophore-based drug design for potential AChE inhibitors from Traditional Chinese Medicine Database.* Bioorganic chemistry, 2018. **76**: p. 400-414.

CHAPITRE II

APPROCHES QSAR ET METHODOLOGIE DE LEUR MISE EN ŒUVRE

Les relations quantitatives structure-activité/propriété (QSAR/QSPR) sont des méthodes pour estimer les propriétés d'un produit chimique à partir de sa structure moléculaire, tout en réduisant les délais, les coûts et les essais sur animaux actuellement nécessaires. Elles sont de plus en plus utilisées, du fait du développement des moyens de calculs. Récemment, un nouvel essor au développement de tels approches a été remarqué du fait que la législation européenne sur les substances chimiques, l'enregistrement, l'évaluation et l'autorisation des substances chimiques (REACH) a suggéré leur utilisation dans un but de limiter le recours à l'expérience [1, 2].

Lors de ces dernières années, l'utilisation des méthodes QSAR n'a cessé de progresser. Elle est même devenue indispensable en chimie pharmaceutique et pour la conception de médicaments [3]. L'approche QSAR rend la conception de médicaments plus rationnelle en minimisant le nombre d'expériences coûteuses.

Toutes ces approches partent de l'hypothèse que des composés structurellement similaires ont des activités similaires. Par conséquent, ces méthodes ont des capacités de prédiction et de diagnostic. Elles peuvent être utilisées pour prédire l'activité biologique (par exemple la CI_{50}) ou la classe (par exemple, inhibiteur par rapport à non inhibiteur) de composés avant le test biologique réel.

Dans ce chapitre, on abordera tout d'abord l'utilité, les diverses applications et la classification des méthodes QSAR. Par la suite, on présentera la méthodologie de mise en œuvre des modèles QSAR ainsi que les différents outils employés pour leur mise en place et leur évaluation : bases de données expérimentales, descripteurs, outils d'analyse de

données. Ensuite, on parlera respectivement de la procédure de validation des modèles QSAR établis ainsi que de leur domaine d'applicabilité.

II.1 UTILITES ET APPLICATIONS DES APPROCHES QSAR

Initiées sur la base de relations empiriques durant la seconde partie du 19éme siècle, les QSAR ont été mathématiquement formalisée dans les années 1930 et leur acceptation, en tant que discipline à part entière, a été définitivement acquise au début des années 1960 suite aux travaux de Hansch et Fujita qui ont proposé un modèle mathématique pour corréler l'activité biologique et la structure chimique. Depuis lors, un grand nombre de modèles ont été conçus pour la prédiction de diverses activités d'intérêt biologique.

Il y'a de nombreuses utilités dans l'utilisation des méthodes QSAR. Nous donnons ci-dessous, à titre d'exemple, quelques-unes :

- prédire l'activité biologique et les propriétés physico-chimiques.
- comprendre et rationaliser les mécanismes d'action dans une série de produits chimiques.
- Economiser le coût de développement de produits (par exemple dans l'industrie pharmaceutique, pesticides, produits d'hygiène personnelle, etc.).
- Réduire ou même remplacer les tests longs et coûteux sur les animaux.

La capacité à prédire une activité biologique est utile dans un certain nombre de situations. Il existe un grand nombre d'applications de ces modèles QSAR au sein de l'industrie, du milieu universitaire et des organismes gouvernementaux chargés de la réglementation. Nous résumons ci-après quelques cas d'utilisations de ces modèles:

- L'optimisation de l'activité pharmacologique, biocide ou pesticide.
- La conception rationnelle de nombreux autres produits tels que des agents tensio-actifs, des parfums, des colorants ou des produits chimiques.
- L'identification des composés dangereux à des stades précoces de développement.
- La conception de la toxicité et des effets secondaires pour les nouveaux composés.
- La prédiction de la toxicité pour les humains pour les multiples expositions (délibérée, occasionnelle ou professionnelle) et de l'écotoxicité.
- la prédiction d'une variété de propriétés physico- chimiques des molécules (qu'il s'agisse de produits pharmaceutiques, de pesticides, de produits d'hygiène personnelle, la chimie fine, etc.).

II.2 CLASSIFICATION DES APPROCHES QSAR

Les méthodes QSAR sont classées selon trois axes : l'un est inhérent à la façon dont les valeurs des descripteurs sont calculées, le second a trait à la méthode chimiométrique

d'analyse de données et le troisième axe correspond au nombre de cible utilisées (Figure II.1).

Le modèle 1D-QSAR corrèle l'activité ou la propriété d'un composé avec des propriétés moléculaires telles que le pKa, la solubilité, le logP, les groupes fonctionnels, etc. Le modèle 2D-QSAR corrèle quant à lui l'activité avec les indices de connectivité ou l'indice de Wiener. Le modèle 3D-QSAR traite pour sa part de l'orientation des molécules dans l'espace, et corrèle ainsi l'activité ou la propriété avec des champs d'interaction non-covalents (champ stérique et électrostatique) entourant les molécules. D'autres modèles comme le 4D-QSAR, le 5D-QSAR, le 6D-QSAR prennent en compte la représentation de chaque molécule dans différentes conformations (stéréoisomères, tautomères) et différents modèles de solvatation.

Selon le type de méthodes chimiométriques d'analyse de données utilisées, les méthodes QSAR sont classées comme linéaires et non linéaires (Figure II.1). Les méthodes linéaires comprennent la régression linéaire multiple (MLR), la MLR par étapes (Stepwise MLR ou S-MLR), l'analyse en composantes principales (PCA), la méthode des moindres carrés partiels (PLS) et l'approximation de la fonction génétique (GFA). Les méthodes non linéaires comprennent par exemple des techniques comme la machine à vecteurs de support (SVM) et les réseaux de neurones artificiels (ANN).

Enfin, selon le troisième axe de classification, le modèle QSAR peut également être classé en tant que QSAR à cible unique ou QSAR à cibles multiples (Figure II.1). Des études récentes indiquent que la conception de médicaments multi-cibles est un sujet de recherche contemporain dans le domaine de la conception de médicaments [4]

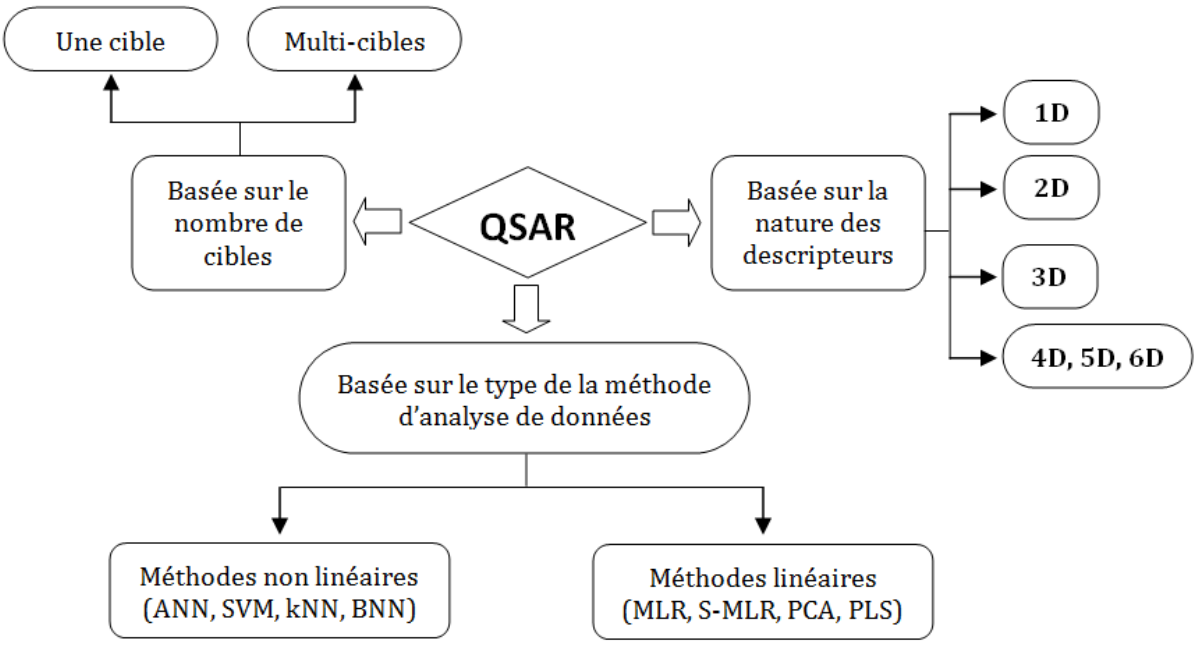


Figure II.1 Classification des approches QSAR

II.3 VALIDITE D'UNE METHODOLOGIE QSAR

Pour faciliter l'application pratique des approches QSAR dans des contextes réglementaires par les autorités et l'industrie et pour améliorer leur acceptation réglementaire, l'Organisation de Coopération et de Développement Économiques (OCDE) a développé des principes de validité de modèles QSAR. Cette validité du modèle devait être établie en appliquant les principes de validation QSAR de l'OCDE [5].Ces principes, énumérés ci-dessous, constituent le meilleur aperçu possible des points essentiels à traiter lors de l'élaboration de modèles QSAR fiables et reproductibles [6].

1er principe – effet bien défini

L'objectif de ce principe est d'assurer la clarté de l'effet prévu par un modèle donné. Ceci est dû au fait qu'un effet donné pourrait être déterminé par différents protocoles expérimentaux et dans différentes conditions expérimentales. Il est donc important d'identifier le système expérimental qui est modélisé par l'approche QSAR.

2éme principe – un Algorithm non ambigu

L'objectif de ce principe est d'assurer la transparence dans l'algorithme du modèle qui génère des prédictions d'un effet à partir d'informations sur la structure chimique et/ou les propriétés physicochimiques. Il est reconnu que, dans le cas de modèles développés commercialement, cette information n'est pas toujours rendue publique.

3éme principe – un domaine d'applicabilité bien défini

La nécessité de définir un domaine d'applicabilité exprime le fait que les QSAR sont des modèles qui sont inévitablement associés à des limitations en termes de structures chimiques, de propriétés physicochimiques et de mécanismes d'action pour lesquels les modèles peuvent générer des prédictions. Aussi, le domaine d'applicabilité d'un modèle QSAR est défini comme l'espace d'activité et de structure chimique dans lequel le modèle fait des prédictions avec une fiabilité donnée.

4éme principe – une évaluation statistique du modèle

Le but de ce principe est d'établir la performance interne d'un modèle (représentée par la qualité de l'ajustement et la robustesse) et sa prédictivité (déterminée par une validation externe) et ce par les mesures de la corrélation et de la robustesse des modèles à partir du jeu d'apprentissage et de leur pouvoir prédictif à partir d'un jeu de validation (ou test).

5éme principe – une interprétation mécanistique (si possible)

Le cinquième principe de validation stipule qu'un modèle QSAR doit être associé à une interprétation mécanistique chaque fois que possible. Cependant, l'absence de base mécanique clairement identifiée pour un modèle ne signifie pas nécessairement que le modèle n'est pas potentiellement utile pour une application réglementaire donnée [5].

II.4 PRINCIPE ET METHODOLOGIE GENERALE D'UNE ETUDE QSAR

Le principe des QSAR consiste à mettre en place une relation mathématique qui tente de relier de manière quantitative les caractéristiques dérivées de la structure d'un composé avec une observable macroscopique (activité biologique, toxicité, propriété physico-chimique, etc.) pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données. Les caractéristiques quantitatives des molécules sont appelées descripteurs. Ils sont actuellement calculés par utilisation de diverses ressources disponibles. Ainsi, la forme générale d'un tel modèle est la suivante :

Propriété = *f* (Descripteurs)

Une fois que cette relation est mise en place et est validée sur un ensemble de validation, elle peut alors être employée pour la prédiction de la propriété de nouvelles molécules, pour lesquelles la valeur expérimentale n'est pas disponible, ou pour des molécules non encore synthétisées.

La méthodologie de l'élaboration d'un modèle QSAR repose, quant à elle, sur l'utilisation des outils suivants : (1) Collecte de données expérimentales fiables et en nombre conséquent qui constitueront la base de données de travail ; (2) Développer et sélectionner une série de descripteurs qui caractérisent les structures moléculaires électroniques et géométriques des composés de la base de données en vue de les relier à la propriété expérimentale étudiée ; (3) Des méthodes d'analyse de données (ou méthodes d'apprentissage statistique) sont alors employées pour choisir les descripteurs adéquats et mettre en place le modèle QSAR ; (4) Une fois développé, le modèle doit alors être validé en termes de corrélation (sur le jeu de données d'apprentissage). Sa robustesse, c'est-à-dire l'influence des composés du jeu d'entraînement sur le modèle, est estimée par des méthodes de validation interne. Pour estimer son pouvoir prédictif, il est ensuite nécessaire de disposer de données expérimentales supplémentaires afin de déterminer la capacité du modèle à prédire ces valeurs ; (5) Enfin, pour tout modèle, il est important de savoir pour quel type de molécules il est utilisable ou non, c'est-à-dire connaître son domaine d'applicabilité. Dans ce qui suit, nous expliciterons en détail ces différents outils inhérents à la méthodologie d'élaboration des modèles QSAR.

II.5 COLLECTE DE LA BASE DE DONNEES

En pratique, le développement de modèles QSAR débute par la première étape qui consiste en la collecte de données expérimentales fiables et en nombre conséquent. Il s'agira de faire une sélection des données expérimentales de référence afin de construire une base de données de départ et ce pour une propriété désirée. Comme les modèles QSAR sont dépendants des données expérimentales de référence, le choix de la base de données expérimentale de référence est décisif. Elle doit être composée de données expérimentales fiables obtenues en suivant un protocole expérimental unique. Par

conséquent, il est nécessaire d'exclure les données de faible qualité car elles réduiront la robustesse du modèle final.

Par ailleurs, deux conditions doivent être satisfaites : (1) l'intervalle des valeurs de la propriété à étudier doit être le plus grand possible, car il détermine le domaine d'applicabilité du modèle. Par conséquent, plus cet intervalle sera grand, plus des modèles prédictifs sur une large gamme de valeurs pourront être attendus. (2) La distribution des valeurs doit être normale pour la simple raison que la plupart des méthodes statistiques sont basées sur des distributions normales et sont donc plus performantes sur ce type de distributions.

Dans la plupart des cas, les données expérimentales sont issues de la littérature. Il existe plusieurs bases de données "online" regroupant un grand nombre de molécules (Tableau II.1). Cependant, la collecte de données implique la revue manuelle des publications de référence dans les fichiers PDB.

Tableau II.1 Exemples représentatifs des bases de données utiles pour la conception de médicaments.

Base Données	Dates [a]	Contenu	Site web
Ligand Expo	2004*	> 13000 ligands différents	ligand-expo.rcsb.org
PDBeChem	2005*	> 13000 ligands différents	www.ebi.ac.uk/pdbe/
HiC-Up	1997–2008[b]	7870 ligands différents	xray.bmc.uu.se/hiccup
SuperLigands	2005–2009[b]	10085 ligands différents dans 401300 complexes	bioinformatics.charite.de/superligands/
PDBBind	2004*	Données d'affinité pour 7986 complexes PDB	http://www.pdbbind-cn.org/
Binding MOAD	2005*	Données d'affinité pour 4782 complexes PDB	http://www.bindingmoad.org/
BindingDB	2001-2017	1200000 données d'affinité pour 5500 protéines	www.bindingdb.org/bind
ChEMBL	2008*	> 5 millions de données d'affinité pour 8603 protéines	http://www.ebi.ac.uk/chembldb

a) L'année de création de la base de données est celle de la publication primaire relative. Il est suivi de l'année de la dernière mise à jour de la base de données (*indique que la base de données est toujours mise à jour).
b) Valeurs cités par Hoffmann et al [7]

II.6 CALCUL DES DESCRIPTEURS

La seconde étape de la méthodologie QSAR consiste en la quantification de la structure moléculaire en utilisant un ensemble de descripteurs moléculaires appropriés,

qui sont des nombres portant des informations importantes sur les aspects constitutionnels, topologiques, géométriques, hydrophobes et/ou électroniques de la structure chimique.

Un descripteur moléculaire est le résultat final d'une procédure mathématique et logique qui transforme l'information chimique encodée dans la représentation symbolique d'une molécule en une valeur numérique utile. Par la suite, un ensemble de descripteurs peut ensuite être corrélé statistiquement à différentes activités biologiques expérimentales, ce qui donne un modèle qui peut être utilisé à des fins de prédiction.

II.6.1 Classification des descripteurs

Les descripteurs moléculaires peuvent être classés de plusieurs façons. La majorité des descripteurs sont basés sur des atomes plutôt que sur des champs. Les descripteurs basés sur les atomes sont généralement générés par l'analyse de tables de connexion 2D ou 3D et peuvent inclure des informations 1D, 2D ou 3D sur la molécule. Les descripteurs basés sur l'atome comprennent des atomes individuels, des comptages de caractéristiques, des fragments sous-structuraux, des indices topologiques, des propriétés atomiques, des pharmacophores et des propriétés physicochimiques calculées.

- **Les descripteurs constitutionnels (1D)**: sont accessibles à partir de la formule brute de la molécule (par exemple C_6H_6O pour le phénol), et décrivent des propriétés globales du composé. Il s'agit par exemple du nombre d'atomes (C, H, N…), du nombre de groupes fonctionnels (NO_2, COOH, OH…), du nombre de liaisons (simples, doubles, aromatiques…), du nombre de cycles (aromatiques ou non), de la masse moléculaire [8].
- **Les descripteurs 2D**: Ces descripteurs 2D reflètent bien les propriétés physiques dans la plupart des cas, mais sont insuffisants pour expliquer de façon satisfaisante certaines propriétés ou activités, telles que les activités biologiques. Ils sont calculés à partir de la formule développée de la molécule. On y distingue les indices topologiques qui peuvent être obtenus à partir de la structure 2D de la molécule. Le calcul de centaines de descripteurs topologiques est expéditif même pour un grand ensemble de données. les descripteurs topologiques les plus fréquemment utilisés sont l'indice de Wiener [9, 10] (permet de caractériser le volume moléculaire et le taux de ramification d'une molécule), l'indice de Randic [9, 11] qui peut être interprété comme une mesure de l'aire de la molécule accessible au solvant et l'indice de Balaban [9, 12].
- **Les descripteurs 3D**: Ces descripteurs tiennent compte de la position des atomes entre eux dans l'espace. Leurs calculs nécessitent donc de connaître la modélisation moléculaire avec les méthodes de calculs telle que les méthodes semi-empiriques et les méthodes *ab initio*. Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la

modélisation de propriétés ou d'activités qui dépendent de la structure 3D. On distingue plusieurs familles importantes de descripteurs 3D :

- Les descripteurs géométriques : les plus importants sont le volume moléculaire, la surface accessible au solvant et le moment principal d'inertie.
- Les descripteurs électroniques : De nombreuses propriétés physicochimiques calculées peuvent être utilisées comme descripteurs telles que les densités électroniques atomiques, les valeurs propres des orbitales moléculaires (HOMO et LUMO), le volume et la surface de Van der Waals.
- Les pharmacophores : Percevoir un pharmacophore est l'étape la plus importante pour comprendre l'interaction entre un récepteur et un ligand. Selon la définition de l'IUPAC (1998), un pharmacophore est " un ensemble de propriétés stériques et électroniques défini à partir d'une interaction entre deux entités moléculaires et nécessaire pour induire la réponse biologique souhaitée" [13]. Autrement dit, les pharmacophores se basent sur les propriétés physico-chimiques et les interactions non covalentes potentielles ou existantes des molécules et non sur leurs structures.

II.6.2 Sélection des descripteurs

Plus de 3000 descripteurs moléculaires sont disponibles et sont principalement calculés par des logiciels commerciaux tels que QSARIS [14], Cerius2 [15], Vol Surf [16] ou Dragon [17].

Le modèle QSAR généré en utilisant tous les descripteurs disponibles ne produit pas le modèle le plus prédictif. La redondance dans les descripteurs peut conduire au développement d'un modèle avec des problèmes statistiques, tels que le sur-apprentissage ou la corrélation aléatoire. La sélection de descripteurs est nécessaire pour éviter de tels problèmes et pour construire un modèle QSAR fiable. Aussi, la sélection de descripteurs est impérative car ils doivent inclure une information maximale des structures et une colinéarité minimale entre eux. En conséquence, un modèle très prédictif développé avec un nombre relativement faible de descripteurs est simple et facile à interpréter.

Plusieurs techniques sont disponibles pour cette opération de sélection des descripteurs les plus pertinents. Parmi ces techniques, on peut citer la régression par étapes (Stepwise regression), le recuit simulé (simulated annealing) [18, 19], l'algorithme génétique (Genetic Algorithm) [20, 21], l'algorithme évolutif (EA) [22], l'algorithme ACO (Ant Colony Optimization) [23] et la technique PSO (Particle Swarm Optimization) [24]. Notons que ces techniques de sélection représentent les techniques d'optimisation simulant les systèmes biologiques.

L'erreur de prédiction du critère carré (PRESS), le critère d'information Akaike (AIC) [25] et les statistiques de Kolmogorov-Smirnov (KS) [26] sont utilisés comme fonction ou critères de fitness au cours de la sélection des variables.

II.7 METHODES D'ANALYSE DES DONNEES

La relation entre les descripteurs sélectionnés et la base de données est déterminée grâce à des outils d'analyse de données comme les régressions linéaires multiples (MLR en anglais), les régressions aux moindres carrés partiels (PLS), les réseaux de neurones, et les algorithmes génétiques. Dans ce qui suit, nous résumerons les rappels théoriques des méthodes d'analyse de données que nous avons utilisées au cours de notre travail pour établir des modèles QSAR linéaire et non linéaire. Ces méthodes d'analyse de données sont les suivantes : la régression linéaire multiple (RLM), les réseaux de neurones artificiels (RNA), l'algorithme d'approximation de la fonction génétique (AFG) et les machines à vecteurs supports pour la régression (SVR)

II.7.1 Régression linéaire multiple

La régression linéaire multiple (MLR en anglais : Multiple linear regression) est la méthode la plus simple et la plus communément employée pour le développement de modèles prédictifs et ce en raison de sa simplicité [27]. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante Y (la propriété) et une série de n variables indépendantes Xi (différents descripteurs moléculaires). L'objectif est d'obtenir une équation de la forme suivante :

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \cdots + a_nX_n \qquad \textbf{(II.1)}$$

Où $a_1, a_2, \ldots, a_n$, sont les coefficients de régression correspondants. Le coefficient a_0 est un terme constant du modèle.

Cependant, La régression linéaire multiple présente un inconvénient de taille vue qu'elle est défaillante dans la mise en évidence de dépendances non linéaires. Néanmoins, l'interprétation de la contribution des descripteurs $X_1, X_2, \ldots, X_n$ est directe en fonction de la valeur du coefficient correspondant et de son signe algébrique.

Les descripteurs présents dans un modèle MLR ne devraient pas être très inter-corrélés. Pour un modèle statistiquement fiable, le rapport "nombre de composés/descripteurs" soit > 5. Un bon modèle MLR représenté avec un diagramme de dispersion (activité observée *vs* activité calculée) montre une déviation minimale des points de la ligne d'ajustement (Figure II.2).

II.7.2 Les réseaux de neurones artificiels

L'idée des réseaux neuronaux artificiels (RNA) a été inspirée de la manière par laquelle les neurones biologiques traitent les informations [28]. Ce concept est utilisé pour

mettre en œuvre des logiciels de simulations des processus massivement parallèles qui impliquent des éléments de traitement connectés dans une architecture de réseau.

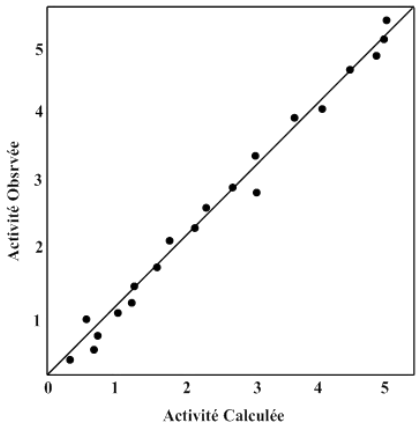


Figure II.2 Diagramme de dispersion des valeurs de l'activité observée en fonction de celles calculées par le modèle.

L'apprentissage dans le cerveau humain se fait à travers un réseau des neurones (Figure II.3). De façon très réductrice, un neurone biologique est une cellule qui se caractérise par des synapses (points de connexion avec les autres neurones, fibres nerveuses ou musculaires), des dentrites (entrées des neurones), l'axone (sortie du neurone vers d'autres neurones ou fibres musculaires) et le noyau qui active la sortie en fonction des stimuli en entrée.

Par analogie, Un neurone formel est une fonction algébrique paramétrée, non linéaire en ses paramètres, et à valeurs bornées. Ses entrées peuvent être les sorties d'autres neurones ou des entrées de signaux extérieurs. Sa sortie est une fonction non linéaire f d'une combinaison linéaire v des entrées (x_i). Le potentiel v le plus fréquemment utilisé est la somme pondérée des entrées x_i pondérées par les coefficients (w_i) également appelés poids de connexions. Le calcul de la valeur de cette fonction peut se décomposer en deux étapes :

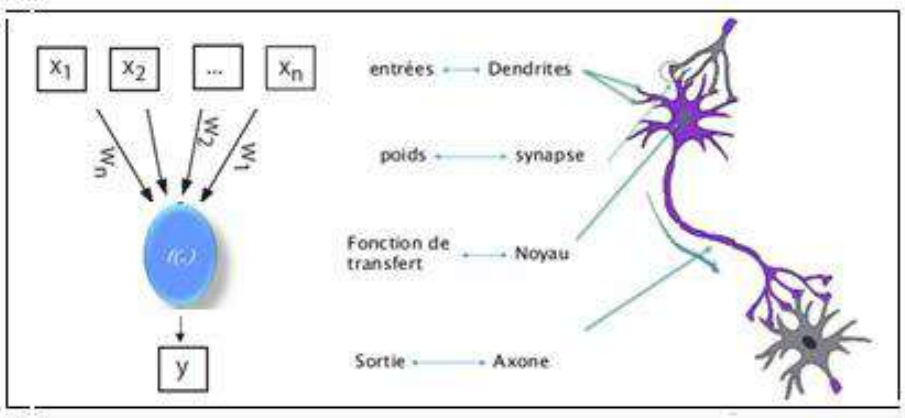


Figure II.3 Neurone formel et neurone biologique (https://fr.slideshare.net/Oussama Werfelli/réseaux-de-neurones-artificiels consulté le 02/03/2018).

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

$$\boldsymbol{v = w_0 + \sum_{i=1}^{n}(w_i\, x_i)} \qquad \textbf{(II.2)}$$

où $(\boldsymbol{x_1, x_2, \dots, x_n})$ sont les neurones d'entrée (descripteurs). $\boldsymbol{w_0}$ est appelé biais. Le biais peut être considéré comme la pondération de l'entrée 0 fixée à 1. $\boldsymbol{v}$ est appelé potentiel du neurone.

- ✓ Dans la seconde étape, une fonction *f*, appelée fonction d'activation, est appliquée à ce potentiel *v*. Cette fonction est la plupart du temps une fonction en "S", bornée en ses extrema, continue et dérivable. Le choix d'une fonction d'activation se révèle dans certains cas être un élément constitutif important des réseaux de neurones. Il en existe plusieurs types, dont les plus souvent utilisées sont regroupées dans le tableau II.2. Ainsi, le neurone formel peut être défini mathématiquement comme étant une fonction algébrique, a priori non linéaire (suivant la fonction d'activation *f*) et bornée, avec des entrées x_i et paramétrée par les poids w_i. Un neurone formel, réalise donc l'opération suivante :

$$y = \mathrm{f}(v) = \mathrm{f}\,(w_0 + \sum_{i=1}^{n}(w_i\, x_i) \qquad \textbf{(II.3)}$$

Tableau II.2 Différentes fonctions d'activation les plus utilisées.

Fonction	Définition	Description
Identité	$\boldsymbol{x}$	L'activation du neurone est transmise directement en sortie
Tangente hyperbolique	$\boldsymbol{\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}}$	Produit généralement de meilleurs résultats que la fonction logistique en raison de sa symétrie. Idéale pour les perceptrons multicouches, en particulier, pour les couches cachées
Exponentielle	$\boldsymbol{e^{-x}}$	La fonction exponentielle négative
Sinus	$\boldsymbol{\sin(x)}$	S'utilise éventuellement si les données sont distribuées radialement. N'est pas utilisé par défaut
Gaussienne	$\boldsymbol{\frac{1}{\sqrt{2\pi\sigma}}\, exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}$	Ce type de fonction d'activation n'est utilisé que par les unités cachées d'un réseau de neurones RBF également connu sous le nom de fonctions à base radiale.

II.7.2.1 Types de réseau de neurones

Les types de réseau de neurones diffèrent selon plusieurs paramètres dont la topologie des connexions entre les neurones (voir figure II.4). Ainsi, on distingue deux grandes classes : les réseaux de neurones non bouclés et les réseaux de neurones bouclés.

- Réseaux de neurones non bouclés (*feed forward*) : un réseau de neurones non bouclés peut être imaginé comme un ensemble de neurones connectés entre eux, l'information circulant des entrées vers les sorties sans retour dans le sens sortie-entrée. On peut alors représenter le réseau par un graphe *acyclique* dont les nœuds sont les neurones et les arêtes les " connexions " entre ceux-ci. Si l'on se déplace dans le réseau, à partir d'un neurone quelconque, en suivant les connexions et en respectant leurs sens, on ne peut pas revenir au neurone de départ [29].
- Réseaux de neurones bouclés ou récurrents (*feed back*) :l'architecture la plus générale, pour un réseau de neurones, est celle des " réseaux bouclés ", dont le graphe des connexions est cyclique : lorsque l'on se déplace dans le réseau en suivant le sens des connexions, il est possible de trouver au moins un chemin qui revient à son point de départ. Un tel chemin est désigné sous le terme de " cycle ".

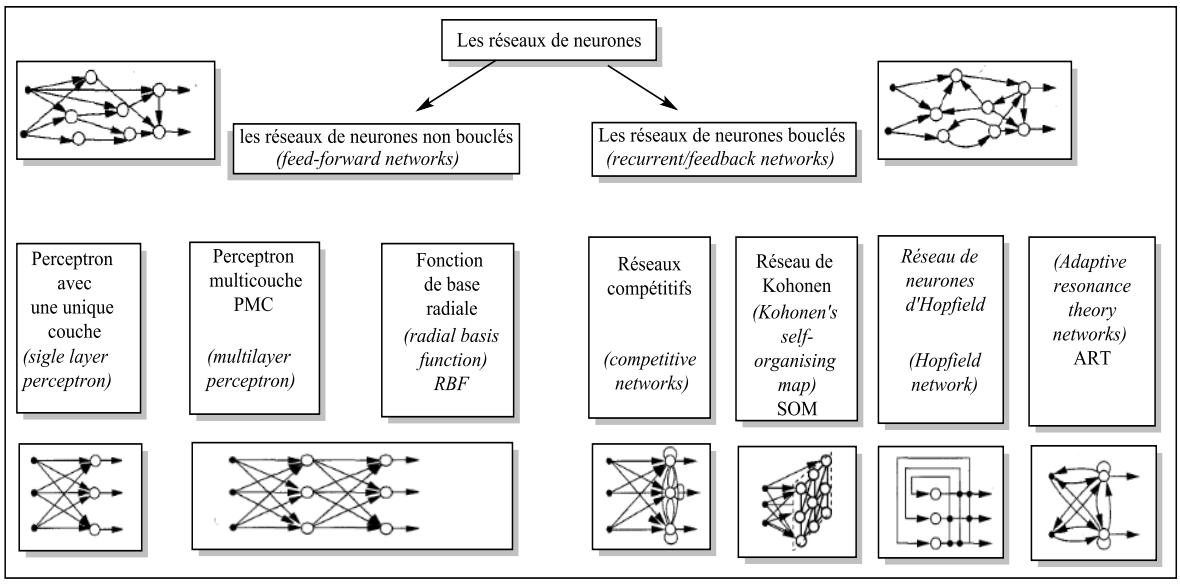


Figure II.4 Types de réseaux de neurones artificiels selon la topologie des connexions entre les neurones (https://slideplayer.fr/slide/1582517 (consulté le 02/03/2018).

II.7.2.2 Perceptron multicouche

Dans cette section du chapitre II, on s'intéressera précisément au type de réseau de neurones non bouclés appelé "perceptron multicouche (PMC)" du fait que tout notre travail a été réalisé avec ce type de réseau. Par ailleurs, notons que de nombreux chercheurs ont exploré les capacités d'approximation universelle des PMC (feed-

forward)[30-32]. De plus, Huang et Babri [33] ont montré qu'un réseau neuronal avec au plus N neurones cachés et avec presque toute fonction d'activation non linéaire, peut apprendre N observations distinctes avec une erreur nulle.

Le perceptron multicouche (ou désigné en anglais par l'abréviation MLP : multilayer perceptron) est un réseau composé de couches successives (figure II.5) :

- Une couche d'entrée : constituée des neurones d'entrée. Leur nombre est égal au nombre de variables d'entrée (descripteurs) plus un (biais). Chaque neurone est connecté aux neurones cachés.
- Une couche cachée : constituée d'un nombre variable de neurones. Pour chaque neurone caché, le réseau effectue une opération de somme pondérée avec les différents poids de chaque neurone d'entrée.
- Une couche de sortie : où le nombre de neurones de sortie est égal au nombre de propriétés (ou activités) à modéliser.

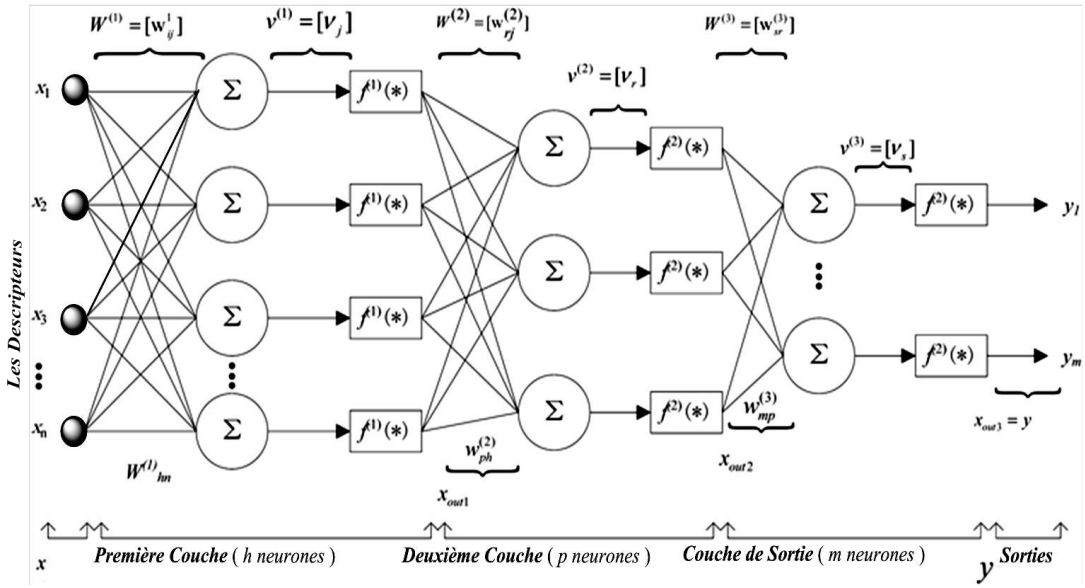


Figure II.5 Architecture d'un perceptron multicouche (PMC)

La fonction de sommation calcule un total pour les signaux d'entrée combinés selon l'équation :

$$v_i^p = \sum_{j=1}^{m_{p-1}} w_{ij}^p \, y_j^{p-1} \qquad \textbf{(II.4)}$$

Où $\boldsymbol{v_i^p}$ est l'entrée dans le neurone *j* (dans la couche *p*), alors que $\boldsymbol{y_j^{p-1}}$ est la sortie du neurone *i* dans la couche précédente (*p-1*); et $\boldsymbol{w_{ij}^p}$ est le poids associé aux nœuds *i* et *j* (dans la couche *p*). Le résultat de la sommation passe par une autre fonction appelée la fonction d'activation (ou de transfert).

La fonction de transfert la plus couramment utilisée est la fonction logistique (sigmoïdale), qui incorpore la caractéristique de non-linéarité. Cependant, il peut y avoir un

petit avantage pratique à utiliser une fonction d'activation '*tanh*' bornée à valeurs réelles comprises entre -1 et +1 de la forme :

$$y_j^p = f(v_i^p) = \frac{(e^{v_i^p} - e^{-v_i^p})}{(e^{v_i^p} + e^{-v_i^p})} \quad \textbf{(II.5)}$$

II.7.3 L'approximation de la fonction génétique (AFG)

L'approximation de la fonction génétique (Genetic approximation function : GFA) a été initialement conçu en s'inspirant de deux algorithmes: l'algorithme génétique de Holland [34] et l'algorithme MARSplines (*Multivariate Adaptive Regression Splines*) de Friedman [35].

Friedman a proposé l'algorithme MARS en tant que membre le plus récent d'une classe d'algorithmes de modélisation statistique bien utilisés tels que CART [36]: il utilise des splines comme fonctions de base pour partitionner l'espace de données en construisant ses modèles de régression. La construction de modèles de régression à base de splines d'ensembles de données avec un nombre modéré de caractéristiques MARS donne de hauts niveaux de performance et rivalise bien avec de nombreuses approches de réseaux neuronaux. Cependant, elle est trop coûteuse à utiliser avec plus d'une vingtaine de caractéristiques. De plus, étant donné que MARS construit son modèle de façon incrémentielle, il ne peut pas découvrir des modèles contenant des combinaisons de caractéristiques qui prédisent bien en tant que groupes. Roger et ses collaborateurs ont reconnu que Friedman effectuait une recherche sur un très grand "espace de fonction" et qu'une meilleure recherche pouvait être faite en utilisant un algorithme génétique plutôt que son approche incrémentale. Le remplacement des chaînes binaires de Holland par des chaînes de fonctions de base a conduit au travail publié sur G / SPLINES [37, 38], qui a évolué plus tard vers l'approximation de la fonction génétique (AFG).

L'algorithme AFG a été appliqué récemment à la génération de modèles QSAR [39, 40]. Dans la plupart des cas, les modèles QSAR sont représentés sous la forme de sommes de termes linéaires ou non linéaires, comme c'est montré dans l'équation (II.1).

$$y = F(X) = \beta_0 + \sum_{m=1}^{M} \beta_m H_{km}(X_{V(k,m)}) \quad \textbf{(II.6)}$$

$$H_{km}(X_{V(k,m)}) = \prod_{k=1}^{K} h_{km} \quad \textbf{(II.7)}$$

Les termes $\{\boldsymbol{h_{km}}\}$ sont appelés les fonctions de base; ce sont des fonctions d'une ou de plusieurs caractéristiques, telles que $(X_5 - 10)^2$, $SIN(X_1) \times SIN(X_2)$, ou X_{10}, où les $\{\boldsymbol{X_{V(k,m)}}\}$ sont les mesures des caractéristiques (descripteurs). Les coefficients $\{\boldsymbol{\beta_m}\}$sont déterminés en utilisant la régression par les moindres carrés ou une autre technique d'ajustement appropriée.

De nombreuses techniques, y compris MARS, CART et PCA, développent un modèle de régression unique par addition ou suppression de fonctions de base. En revanche, l'algorithme AFG utilise une population de nombreux modèles et ne teste que le modèle final entièrement construit.

Des modèles améliorés sont construits en effectuant l'opération de croisement génétique pour recombiner les termes des modèles les plus performants. Un algorithme génétique exige qu'un individu soit représenté comme une chaîne linéaire, qui joue le rôle de l'ADN pour l'individu. Lors de l'utilisation de l'AFG, la chaîne est la série de fonctions de base, comme illustré dans la Figure II.6.

En utilisant l'information contenue dans la chaîne, il est possible de reconstruire le modèle d'activité en utilisant la régression des moindres carrés pour régénérer les coefficients$\{\beta_m\}$. Les modèles initiaux sont générés en sélectionnant de manière aléatoire un certain nombre de caractéristiques (descripteurs) de l'ensemble de données d'apprentissage, en construisant des fonctions de base à partir de ces caractéristiques en utilisant les types de fonctions de base spécifiées par l'utilisateur, puis en construisant les modèles génétiques à partir des séquences aléatoires de ces fonctions de base.

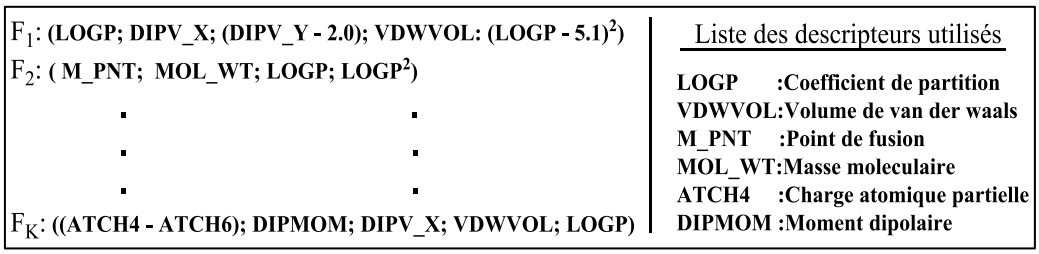


Figure II.6 Exemples d'une population de modèles représentée pour l'algorithme AFG [37].

La fonction d'évaluation (*fitness*) utilisée durant l'évolution est dérivée de la fonction de score LOF (*lack-of-fit*) (manque d'ajustement) de Friedman [41]qui est une mesure d'erreur des moindres carrés pénalisée. Cette mesure équilibre la diminution d'erreur au fur et à mesure que d'autres fonctions de base s'ajoutent. En se basant sur la faible erreur de l'ensemble d'entraînement, la fonction de dérivée de la LOF guide le système vers des modèles qui prédisent bien et résistent au sur-ajustement [42]. Cette fonction est donnée par :

$$LOF = \frac{LSE}{\left(1-\frac{c+dp}{M}\right)^2} \qquad \textbf{(II.8)}$$

où$\{\boldsymbol{c}\}$ est le nombre de fonctions de base ; $\{d\}$ est le paramètre de lissage (*smoothing parameter*) ; $\{p\}$ est le nombre total de descripteurs contenus dans toutes les fonctions de base ;$\{M\}$ est le nombre d'échantillons dans l'ensemble d'apprentissage.

Contrairement à l'erreur des moindres carrés (Least Squares Error en anglais : LSE) la plus couramment utilisée, la mesure LOF ne peut pas toujours être réduite en ajoutant

plus de termes au modèle de régression. Alors que le nouveau terme peut réduire le LSE, il augmente également les valeurs de c et p. qui tend à augmenter le score LOF.

II.7.3.1 Fonctionnement de l'algorithme AFG

Une fois que tous les modèles de la population ont été évalués en utilisant le score LOF, l'opération de croisement génétique est effectuée à plusieurs reprises. À ce stade, nous effectuons à plusieurs reprises l'opération de recombinaison génétique ou de croisement:

- Deux bons modèles sont choisis comme "parents" proportionnels à leur forme physique.
- Chaque modèle est divisé au hasard en deux sections. Les coupures se produisent entre les fonctions de base. Un nouveau modèle est créé en utilisant les fonctions de base extraites d'une section de chaque parent.
- Les opérateurs de mutation optionnels peuvent modifier le modèle nouvellement créé.
- Le modèle avec le pire fitness est remplacé par ce nouveau modèle.
- Le processus global est terminé lorsque l'aptitude moyenne des modèles dans la population cesse de s'améliorer. Pour une population de 300 modèles, 3000 à 10 000 opérations génétiques sont généralement suffisants pour atteindre la convergence [40, 42].
- Une description d'une opération de croisement est montrée sur la figure II.7.

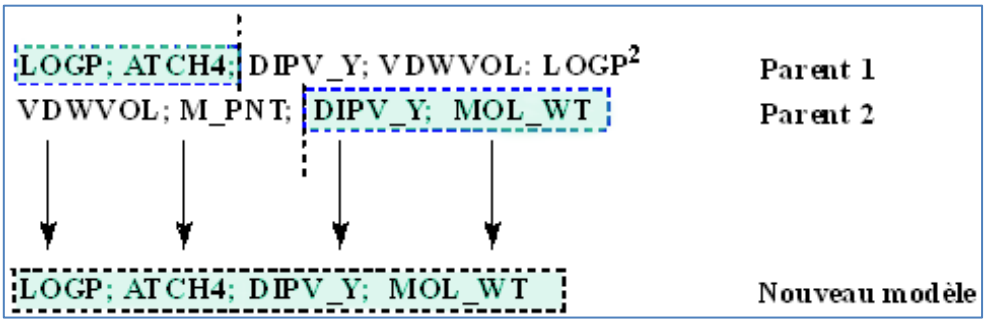


Figure II.7 Opération de croisement génétique [37]

À la fin, l'utilisateur peut simplement sélectionner le modèle de la population ayant le meilleur score d'évaluation, bien qu'il soit généralement préférable d'inspecter les différents modèles et de sélectionner en fonction de la pertinence des caractéristiques, des fonctions de base et des combinaisons de caractéristiques [42].

II.7.4 Machines à vecteurs supports pour la régression (SVR)

II.7.4.1 Machines à vecteurs supports (SVM)

Les Machines à Vecteurs Supports (Support Vector Machines en anglais : SVM) souvent traduit par l'appellation de Séparateur à Vaste Marge (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination ; c'est-à-dire la prévision d'une variable qualitative initialement binaire. Ils ont été ensuite généralisés à la prévision d'une variable quantitative.

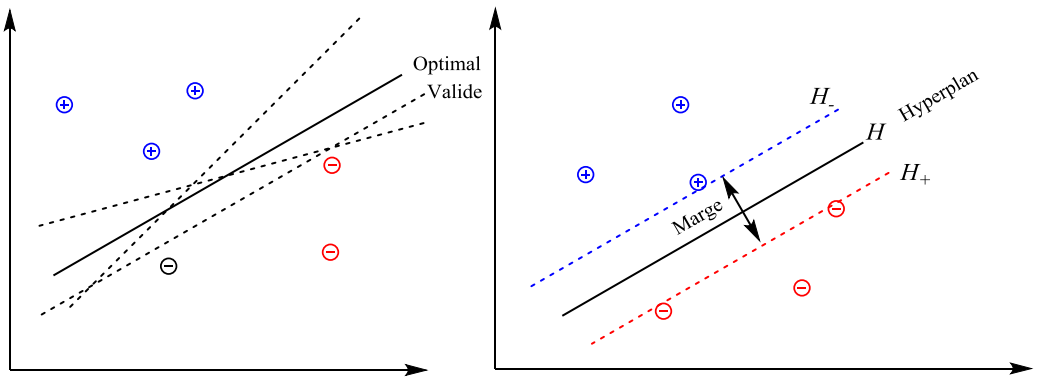
Cette approche découle directement des travaux de Vapnik en théorie de l'apprentissage à partir de 1995. Elle s'est focalisée sur les propriétés de généralisation (ou prévision) d'un modèle en contrôlant sa complexité [43]. Les SVM ont rapidement été adoptés pour leur capacité à travailler avec des données de grandes dimensions, avec un faible nombre de paramètres, leurs garanties théoriques, et leurs bons résultats en pratique. Les SVM ont été appliqués à de très nombreux domaines (bio-informatique, recherche d'information, vision par ordinateur, finances, etc.).

Principe général

Cette technique est une méthode de classification à deux classes qui tente de séparer les exemples positifs des exemples négatifs dans l'ensemble des exemples. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale.

Marge d'un hyperplan séparateur

Il existe en effet une infinité d'hyperplans séparateurs, dont les performances en apprentissage sont identiques (le risque empirique est le même), mais dont les performances en généralisation peuvent être très différentes. Pour résoudre ce problème, il a été montré [44] qu'il existe un unique hyperplan optimal, défini comme l'hyperplan qui maximise la marge entre les échantillons et l'hyperplan séparateur (Figure II.8).



FigureII.8 Recherche d'un hyperplan de séparation optimal au sens de la marge

Un hyperplan séparateur $\mathcal{H}$ de marge maximale γ définit une zone, un " *no man's land* " en quelque sorte, délimitée par deux hyperplans $\mathcal{H}_+$ et $\mathcal{H}_-$ parallèles à $\mathcal{H}$, situés chacun à une distance $\frac{\gamma}{2}$ de $\mathcal{H}$, et entre lesquels il n'y a aucun point du jeu de données. Ce "*no man's land*" est souvent appelé zone d'indécision car on manque d'information pour classifier les points dans cette zone.

Les points positifs les plus proches de $\mathcal{H}$ sont situés sur $\mathcal{H}_+$, et de même, les points négatifs les plus proches de $\mathcal{H}$ sont situés sur $\mathcal{H}_-$. Ainsi, ces points *supportent* $\mathcal{H}_-$ et $\mathcal{H}_+$ et s'appellent les vecteurs de support, d'où le nom de machines à vecteurs support (SVM).

II.7.4.2 SVM pour la régression (SVR)

Le principe de base de la régression par Machines à Vecteurs Support (Support Vector Regression : SVR) [43] est similaire à celui du classement par SVM : minimiser une fonction de perte pénalisées, la pénalisation étant l'inverse du carré de la marge, et la fonction de perte pouvant être en régression la fonction « ε –insensible » .

Cette fonction *f(x)* qui a au plus une déviation ε par rapport aux exemples d'apprentissage (*xi, yi*), pour *i* = 1,…, *N*, et qui est la plus plate possible. Cela revient à ne pas considérer les erreurs inférieures à ε et à interdire celles supérieures à ε [45]. Maximiser la platitude de la fonction permet de minimiser la complexité du modèle qui influe sur ses performances en généralisation. En effet, la théorie de l'apprentissage [43]permet de borner l'erreur de généralisation par une somme de deux termes : l'un dépendant de la complexité du modèle et l'autre dépendant de l'erreur sur les données d'apprentissage [46, 47]. Les méthodes SVMs sont basées sur le contrôle de la complexité du modèle lors de l'apprentissage.

Dans ce qui suit, on va présenter l'algorithme SVR pour un modèle linéaire, puis pour le cas non-linéaire.

<u>SVR pour un modèle linéaire</u>

La fonction linéaire *f(x)*est de la forme :

$$f(x) = \langle w, x_i \rangle + b \qquad \textbf{(II.9)}$$

où $w \in \mathbb{R}^p$ est un vecteur des paramètres (ou poids), orthogonal au plan, $x \in \mathbb{R}^n$ le vecteur d'entrée $\langle w, x_i \rangle$ désigne le produit scalaire entre *w* et x_i et *b* une constante à déterminer.

Pour assurer la platitude de la fonction *f(x)*, la norme des poids est minimisée (contrainte sur la dérivée de la fonction *f(x)*). Le problème revient donc à minimiser cette norme en garantissant que les erreurs sont inférieures à ε et peut s'écrire :

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ avec\ |y_i - \langle w, x_i \rangle - b| \leq \varepsilon,\ i = 1, \dots, N \end{cases} \qquad \textbf{(II.10)}$$

L'objectif de cette formulation n'est pas de minimiser l'erreur d'apprentissage comme dans les réseaux de neurones ou la plupart des algorithmes de régression, mais d'assurer la platitude de la fonction. La minimisation de l'erreur n'apparaît que sous forme de contraintes (aucune erreur n'est autorisée). Cette description du problème considère donc qu'une fonction linéaire *f(x)* qui approxime tous les exemples avec une précision $\boldsymbol{\varepsilon}$ existe.

En présence de valeurs aberrantes, il est plus important d'autoriser certaines erreurs. De manière analogue à la fonction de perte "marge souple" (soft margin) [48] introduite dans les machines SV par Cortes et Vapnik [49], on peut introduire des variables de relâchement (*slack variables*) $\boldsymbol{\xi_i, \xi_i^*}$ d'optimisation pour faire face aux contraintes du problème d'optimisation ci-dessus. Nous arrivons donc à la formulation suivante :

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \qquad \textbf{(II.11)}$$

Sous les contraintes :

$$\begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, \dots, N \end{cases}$$

où $\boldsymbol{\xi_i}$ *et* $\boldsymbol{\xi_i^*}$ représentent respectivement les erreurs ($|\boldsymbol{y_i - f(x)}|$) positives et négatives. La constante C > 0 détermine le compromis entre la platitude de *f(x)* et la quantité d'erreurs autorisées (les écarts supérieurs à ε autorisés). Cela correspond à traiter une fonction de perte dite ε-insensible $|\boldsymbol{\xi}|_{\boldsymbol{\varepsilon}}$décrite par :

$$|\xi|_\varepsilon = \begin{cases} 0 & si\ |\xi_i| \le \varepsilon \\ |\xi_i| - \varepsilon & autrement \end{cases} \qquad \textbf{(II.12)}$$

Les paramètres de la fonction de perte ε-insensible sont présentés sur la figure II.9. Seuls les points en dehors de la région ombrée contribuent au coût dans la mesure où les écarts sont pénalisés de manière linéaire.

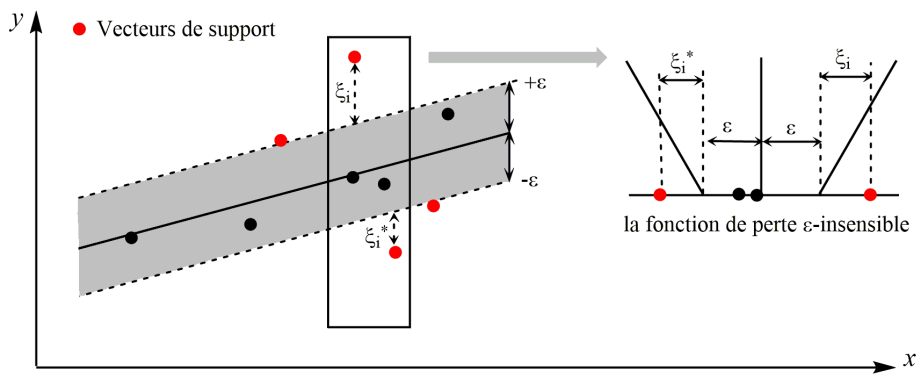


Figure II.9 Les paramètres de la fonction de perte pour un SVR

Il s'avère que, dans la plupart des cas, le problème d'optimisation (équation II.11) peut être résolu plus facilement dans sa formulation duale. De plus, la formulation duale fournit la clé pour étendre SV aux fonctions non linéaires. Par conséquent, nous utiliserons une méthode de dualisation standard utilisant des multiplicateurs de Lagrange. Le problème (équation II.11) se résout en minimisant le Lagrangien *L* donné par :

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{N}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{N}\alpha_i^*(\varepsilon + \xi_i^* - y_i - \langle w, x_i\rangle - b) \quad \textbf{(II.13)}$$

où η_i, η_i^*, α_i, α_i^* représentent les multiplicateurs de Lagrange. La solution de la minimisation du lagrangien $\boldsymbol{L}$ peut s'obtenir en annulant le gradient du lagrangien par rapport aux variable primales ($\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\xi_i}, \boldsymbol{\xi_i^*}$) :

$$\partial L_b = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0 \quad \textbf{(II.14)}$$

$$\partial L_w = w - \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)x_i = 0 \quad \textbf{(II.15)}$$

$$\partial L_{\xi_i} = C - \alpha_i - \eta_i = 0 \quad \textbf{(II.16)}$$

$$\partial L_{\xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \quad \textbf{(II.17)}$$

Les variables $\eta_i$et η_i^* peuvent être éliminés selon (équations II.16 et II.17) pour avoir le Lagrangien dual qui doit être maximisé :

$$L_D = -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle - \varepsilon\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \quad \textbf{(II.18)}$$

Le modèle finale s'écrit :

$$f(x) = \sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \quad \textbf{(II.19)}$$

Le paramètre *b* peut se calculer grâce aux conditions de Karush-Kuhn-Tucker (KKT).

On peut remarquer que seuls les exemples avec des α_i et α_i^* non nuls sont nécessaires pour implémenter le modèle (équation II.18). Ces exemples sont appelés Vecteurs de Support.

SVR pour un modèle non-linéaire

Même lorsque les populations à discriminer sont relativement bien séparées, la séparation peut ne pas être linéaire. On cherche alors, au moyen d'une transformation non linéaire Φ, à passer de l'espace d'origine dans un espace de dimension plus grande doté d'un produit scalaire (espace de Hilbert) et où existe une séparation linéaire (figure II.10). Le problème devient linéaire et le modèle est donné par :

$$f(x) = \sum_{i,j=1}^{N} (\alpha_i - \alpha_i^*) k\langle x_i, x\rangle + b \qquad \textbf{(II.20)}$$

où, $\Phi(x_i)$ est la projection de x_i dans l'espace étendu $\mathcal{F}$.

La quantité $k\langle x_i, x\rangle = \langle \Phi(x_i), \Phi(x)\rangle$est appelée fonction noyau ou noyau (*kernel function* ou *kernel*), qui doit vérifier les conditions de Mercer [50] qui se résument à vérifier que la fonction *k* correspond bien à un produit scalaire définissant une norme dans un espace étendu $\mathcal{F}$. Dans l'algorithme, c'est le noyau *k* et non Φ qui est choisi, et en s'y prenant bien, on peut calculer $k\langle x_i, x\rangle$ sans faire apparaitre Φ. Les calculs sont alors fait dans l'espace de départ, et deviennent beaucoup plus simples et plus rapides ; pour cette raison, on parle de machine à noyau ou " *kernel machine* ".

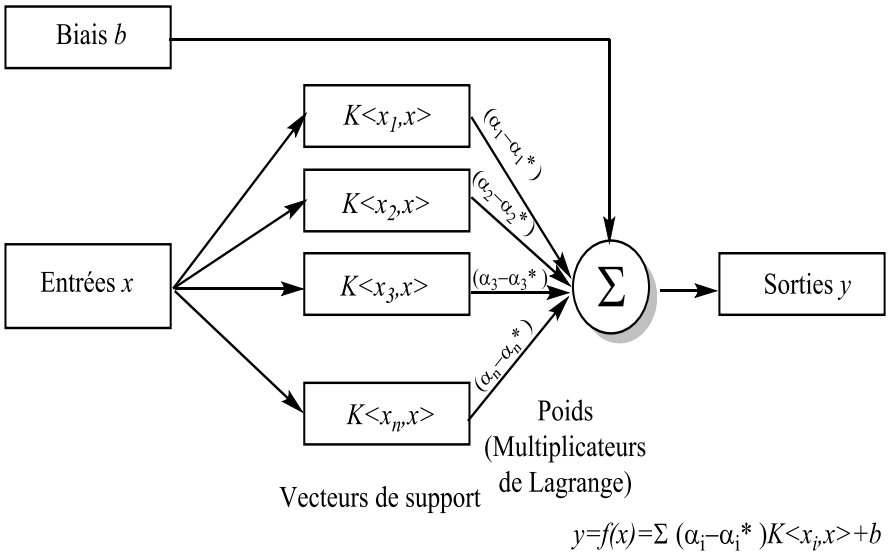


Figure II.10 Architecture d'une machine à vecteurs support

Les noyaux les plus couramment utilisés pour les SVMs sont :

- Linéaire: $k(x, x') = \langle x, x'\rangle$; le plus simple, utilisé de grands volumes de données, notamment en « *textmining* » ;
- Polynomial: $k(x, x') = (\gamma\langle x, x'\rangle + c)^d$; où *d* est le degré du polynôme, utilisé en traitement de l'image ;
- Sigmoïdal: $k(x, x') = tanh(\gamma\langle x, x'\rangle + c)$; souvent utilisée pour le perceptron multicouches ;
- Gaussien: $k(x, x') = exp\left(\frac{-\|x-x\prime\|^2}{2\sigma^2}\right)$; l'une des plus couramment employés avec le noyau radial gaussien ;
- Radial gaussien (RBF): $k(x, x') = exp(-\gamma\|x - x'\|^2)$.

II.8 OUTILS DE VALIDATION DE MODÉLES QSAR

Les modéles QSAR étant utiles à diverses fins, y compris la prédiction des activités de produits non testés, ils doivent être validés sinon ils peuvent conduire à une fausse prévision de l'activité biologique. La validation des modèles QSAR, après le développement du modèle, est donc la partie la plus importante des études QSAR.

Durant cette décennie, la validation des modèles QSAR a reçu une attention considérable [51]. Les outils d'évaluation de la validité des modèles QSAR sont :

- Les coefficients et tests statistiques standard : comme le coefficient de correlation (R^2), l'erreur absolue moyenne (EAM) ou la déviation standard (s) et l' indice de Fischer (F).
- La validation interne : elle sert à déterminer la stabilité du modèle élaboré et concerne les données qui ont créé le modèle (échantillon d'apprentissage). Elle englobe à titre d'exemple la méthode de la validation croisée (Q^2), celle du R^2 ajusté (R^2_{adj}), l'erreur quadratique moyenne (EQM).
  Globalement la validation interne permet l'évaluation de la robustesse du modèle, c'est à dire la stabilité du modèle QSAR vis-à-vis des molécules de l'échantillon d'apprentissage. Par conséquent, elles ne permettent en aucun cas de juger du pouvoir prédictif des modèles.
- La validation externe : le pouvoir prédictif d'un modèle QSAR est déterminé par l'utilisation d'un échantillon de validation (ou de test) non employé pour le développement du modèle. Aussi, avant le développement du modèle, la base de données initiale est scindée en deux échantillons: un pour l'apprentissage (ou entraînement) avec lequel le modèle est élaboré et un second échantillon est utilisé pour caractériser son pouvoir prédictif.

De plus amples informations sur ces outils de validation des modèles QSAR seront abordées dans le chapitre III.

II.9 DOMAINE D'APPLICABILITE D'UN MODELE QSAR

Le domaine d'applicabilité (D.A) est la région de l'espace chimique définie par les molécules de l'échantillon d'apprentissage du modèle. Il peut être caractérisé de différentes manières. Dans la suite de cette étude (chapitre III), seront rapportées les méthodes qu'on a utilisées pour caractériser ce domaine.

Conformément au 3éme principe de l'OCDE, la détermination du domaine d'applicabilité d'un modèle QSAR est d'une grande importance, car il n'est pas destiné à être employé en dehors de son D.A, autrement dit en dehors de l'espace couvert par son échantillon d'apprentissage. Seules les prédictions concernant des molécules dans ce domaine peuvent être considérées comme fiables.

II.10 CONCLUSION

Comme nous venons de le voir tout au long de ce chapitre, les approches QSAR impliquent l'étude des corrélations entre la structure chimique et l'activité biologique ou les propriétés physico-chimiques ou d'autres propriétés, et ce dans le but de prédire l'activité ou les propriétés de substances. Avec le nombre croissant de bases de données expérimentales aussi bien chimiques que pharmaceutiques, la recherche axée sur les QSAR a considérablement progressé ces dernières années. Par conséquent, un bon nombre d'applications réussies des approches QSAR dans le processus de conception de médicaments ont montré l'efficacité de cette méthode.

Néanmoins, la mise au point de modèles QSAR fiables n'est pas une chose aisée car elle dépend d'un certain nombre de facteurs, dont les plus primordiaux sont la disponibilité de données de bonne qualité, le recours aux tests de validation et la définition du domaine d'applicabilité.

References

1. Lilienblum, W., et al., *Alternative methods to safety studies in experimental animals: role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH).* Archives of toxicology, 2008. 82(4): p. 211-236.
2. Worth, A., et al., *The role of the European Chemicals Bureau in promoting the regulatory use of (Q) SAR methods.* SAR and QSAR in Environmental Research, 2007. 18(1-2): p. 111-125.
3. Grover, M., et al., *Quantitative structure–property relationships in pharmaceutical research–Part 2.* Pharmaceutical science & technology today, 2000. 3(2): p. 50-57.
4. Ambure, P., S. Kar, and K. Roy, *Pharmacophore mapping-based virtual screening followed by molecular docking studies in search of potential acetylcholinesterase inhibitors as anti-Alzheimer's agents.* Biosystems, 2014. 116: p. 10-20.
5. Co-operation, O.f.E. and Development, *Guidance document on the validation of (quantitative) structure-activity relationship [(Q) SAR] models*. 2014: OECD Publishing.
6. Jaworska, J.S., et al., *Summary of a workshop on regulatory acceptance of (Q) SARs for human health and environmental endpoints.* Environmental Health Perspectives, 2003. 111(10): p. 1358.
7. Hoffmann, R.Ã.D., et al., *Data Mining in Drug Discovery*. 2013: Wiley.
8. Goulon-Sigwalt-Abram, A., *A new approach to learning from structured data and its applications to computer-aided drug design*. 2008, Université Pierre et Marie Curie - Paris VI.
9. Todeschini, R., et al., *Handbook of Molecular Descriptors*. 2008: Wiley.
10. Wiener, H., *Structural determination of paraffin boiling points.* Journal of the American Chemical Society, 1947. 69(1): p. 17-20.
11. Randic, M., *Characterization of molecular branching.* Journal of the American Chemical Society, 1975. 97(23): p. 6609-6615.
12. Balaban, A.T., *Highly discriminating distance-based topological index.* Chemical Physics Letters, 1982. 89(5): p. 399-404.
13. Wermuth, C., et al., *Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998).* Pure and Applied Chemistry, 1998. 70(5): p. 1129-1143.
14. QSARIS, *www.scivision.com/qsaris.html.*
15. Cerius2, *www.accelrys.com/products/cerius2.*
16. VolSurf, *www.moldiscovery.com/softvolsurf.php.*
17. DRAGON, *http://www.talete.mi.it/index.htm.*
18. Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi, *Optimization by simulated annealing.* science, 1983. 220(4598): p. 671-680.
19. Kirkpatrick, S., C. Gelatt, and M. Vecchi, *Optimization by simulated annealing IBM Research Report RC p 9355.* 1982.
20. Hasegawa, K., Y. Miyashita, and K. Funatsu, *GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists.* Journal of Chemical Information and Computer Sciences, 1997. 37(2): p. 306-310.
21. Lucasius, C.B. and G. Kateman, *Understanding and using genetic algorithms Part 1. Concepts, properties and context.* Chemometrics and intelligent laboratory systems, 1993. 19(1): p. 1-33.

22. Kubiny, H., *Variable selection in QSAR studies. I. An evolutionary algorithm.* Molecular Informatics, 1994. 13(3): p. 285-294.
23. Bonabeau, E., M. Dorigo, and G. Theraulaz, *Inspiration for optimization from social insect behaviour.* Nature, 2000. 406(6791): p. 39.
24. Eberhart, R. and J. Kennedy. *A new optimizer using particle swarm theory*. in *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*. 1995. IEEE.
25. Akaike, H., *Factor analysis and AIC*, in *Selected Papers of Hirotugu Akaike*. 1987, Springer. p. 371-386.
26. Wilcox, R., *Kolmogorov–smirnov test.* Encyclopedia of biostatistics, 2005.
27. Todeschini, R., V. Consonni, and P. Gramatica, *Chemometrics in QSAR.* 2009.
28. McCulloch, W.S. and W. Pitts, *A logical calculus of the ideas immanent in nervous activity.*The bulletin of mathematical biophysics, 1943. 5(4): p. 115-133.
29. Dreyfus, G., et al., *Réseaux de neurones - Méthodologie et applications*. 2002.
30. Hornik, K., *Approximation capabilities of multilayer feedforward networks.* Neural networks, 1991. 4(2): p. 251-257.
31. Ito, Y., *Approximation of continuous functions on Rd by linear combinations of shifted rotations of a sigmoid function with and without scaling.* Neural Networks, 1992. 5(1): p. 105-115.
32. Leshno, M., et al., *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function.* Neural networks, 1993. 6(6): p. 861-867.
33. Huang, G.-B. and H.A. Babri, *Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions.* IEEE Transactions on Neural Networks, 1998. 9(1): p. 224-229.
34. Holland, J., *Adaption in natural and artificial systems.* Ann Arbor MI: The University of Michigan Press, 1975.
35. Friedman, J., *Multivariate Adaptive Regression Splines (Tech. Rep. No. 102)*. 1988, Stanford, CA: Laboratory of Computational Statistics-Dept. of Statistics.
36. Breiman, L., et al., *Classification and Regression Trees, Wadsworth International Group, Belmont, California (1984).* Google Scholar, 1984.
37. Rogers, D. *Data analysis using G/SPLINES*. in *Advances in Neural Information Processing Systems*. 1992.
38. Rogers, D., *G/SPLINES: A hybrid of Friedman's multivariate adaptive regression splines (MARS) algorithm with Holland's genetic algorithm.* 1991.
39. Rogers, D. *Genetic function approximation: A genetic approach to building quantitative structureactivity relationship models. In?, editor*. in *Proceedings of the 10th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modeling, page*. 1994.
40. Rogers, D. and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships.* Journal of Chemical Information and Computer Sciences, 1994. 34(4): p. 854-866.
41. Friedman, J.H., *Fitting functions to noisy data in high dimensions*. 1988.
42. Devillers, J., *Genetic Algorithms in Molecular Modeling*. 1996: Elsevier Science.
43. Vapnik, V.N., *The Nature of Statistical Learning Theory*. 1995: Springer New York.

44. Vapnik, V.N. and S. Kotz, *Estimation of dependences based on empirical data*. Vol. 40. 1982: Springer-Verlag New York.
45. Smola, A.J. and B. Schölkopf, *A tutorial on support vector regression.* Statistics and computing, 2004. 14(3): p. 199-222.
46. Cristiani, N. and S.J. Taylor, *An introduction to support vector machines.*2000.
47. Lauer, F. and G. Bloch. *Méthodes SVM pour l'identification*. in *Journées Identification et Modélisation Expérimentale (JIME'2006)*. 2006.
48. Bennett, K.P. and O.L. Mangasarian, *Robust linear programming discrimination of two linearly inseparable sets.* Optimization methods and software, 1992. 1(1): p. 23-34.
49. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. 20(3): p. 273-297.
50. J Mercer, B., *XVI. Functions of positive and negative type, and their connection the theory of integral equations.* Phil. Trans. R. Soc. Lond. A, 1909. 209(441-458): p. 415-446.
51. Veerasamy, R., et al., *Validation of QSAR models-strategies and importance.* International Journal of Drug Design & Discovery, 2011. 3: p. 511-519.

CHAPITRE III

DEVELOPPEMENT DE MODELES QSAR POUR LA PREDICTION DE LA CONCENTRATION INHIBITRICE (CI_{50})

Comme rapporté dans le paragraphe I.6 du chapitre I, notre travail est axé sur la maladie d'Alzheimer (MA). Actuellement, plus de 35 millions de personnes âgées dans le monde souffrent de cette maladie. Cette dernière est apparue comme la principale cause de la démence et la cause la plus fréquente de perte de mémoire et de déficience cognitive [1]. Le seul traitement symptomatique prouvé à ce jour est l'utilisation d'inhibiteurs de la cholinestérase (ChE) pour augmenter l'activité cholinergique.

Il existe deux types de ChE qui sont omniprésentes dans tout le corps. L'acétylcholinestérase (AChE) est principalement localisée dans les neurones du système nerveux central. Le second type de ChE désigné sous le nom de butyrylcholinestérase (BuChE) est principalement associé aux cellules gliales et est sécrété par elles [2]. L'inhibition des cholinestérases entraîne une augmentation de la concentration synaptique de l'acétylcholine favorisant ainsi son action sur les récepteurs nicotiniques et muscariniques. Selon la littérature [2-5], les effets bénéfiques des inhibiteurs sont liés à leur action anticholinestérasique au niveau cérébral et en particulier pour l'AChE. En d'autres termes, la conception de nouveaux inhibiteurs de BuChE et AChE puissants et sélectifs est d'une grande importance dans la découverte de médicaments.

Les N-benzylpipéridines sont un exemple d'inhibiteur de l'acétylcholinestérase (AChEI). Il a été démontré expérimentalement que cette famille présente une activité inhibitrice accrue (efficacité exceptionnelle *in vitro* et *in vivo*, effets secondaires minimes et sélectivité élevée) vis-à-vis de l'AChE par rapport aux autres inhibiteurs [3-8]. D'autre part, la Tacrine, premier inhibiteur de l'AChE commercialisé, a été le composé le plus étudié et demeure une structure de référence dans le développement de nouveaux inhibiteurs de la ChE en tant que médicaments potentiels pour la MA. C'est l'un des inhibiteurs de la butyrylcholinestérase (BuChE) les plus puissants connus à ce jour [1].

Comme constaté dans la revue bibliographique (voir chapitre I), beaucoup de recherches ont été orientées vers des études de modélisation moléculaire des inhibiteurs de l'AChE et de BuChE. Cependant, parmi cette abondante littérature, les études QSAR spécifiquement dédiées aux N-benzylpipéridines et aux dérivés de la Tacrine apparaissent plutôt limitées. Par conséquent, l'objectif assigné à ce travail de Doctorat est d'établir des modèles QSAR pour la prédiction du pouvoir inhibiteur des dérivés de la Tacrine et des N-benzylpipéridines vis-à-vis de la BChE et l'AChE.

1ére Partie

MODELES QSAR DE PREDICTION DU POUVOIR INHIBITEUR D'UNE SERIE DE DERIVES DE LA TACRINE SUR L'ENZYME BUTYRYLCHOLINESTERASE (BuChE)

III.1 MATERIELS ET METHODES

III.1.1 Collecte des données de l'activité inhibitrice

La concentration inhibitrice (CI_{50}) de dérivés de la Tacrine, extraite de la littérature [3, 6, 9-13], a été utilisée comme ensemble de données pour la présente étude. Les données CI_{50} correspondent à la concentration (mol / l) d'un médicament nécessaire pour une inhibition de 50% *in vitro* et seules celles mesurées par la méthode d'Ellman [14] ont été sélectionnées. Les concentrations inhibitrices (CI_{50}) des dérivés de la Tacrine ont été récupérées à partir de la base de données BindingDB [15] (Figure III.1)

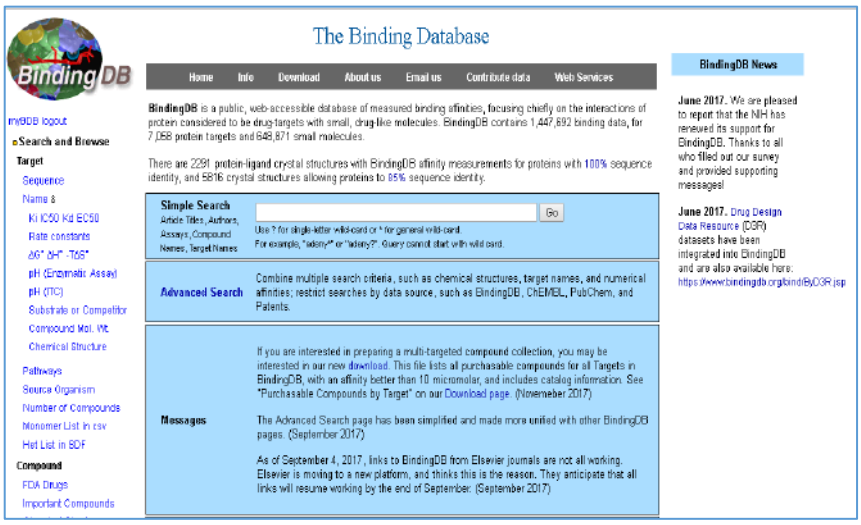


Figure III.1 capture d'écran du site www.bindingdb.com

Afin de récupérer toutes les structures des dérivés de la Tacrine ayant une activité inhibitrice sur la BuChE, nous avons procédé comme suit:

- Utiliser le mot-clé " butyrylcholinestérase" pour chercher dans la base de données BindingDB tous les inhibiteurs de BuChE.
- Sélectionner les inhibiteurs dont la structure possédant le fragment de la Tacrine.
- Éliminer les composés dupliqués.
- Vérifier que les dosages du cholinestérase étaient effectués en utilisant du sérum humain avec la méthode colorimétrique décrite par Ellman [14].

Grâce à cette recherche, nous avons pu récupérer 151 dérivés de la Tacrine (Figure III.2) ayant une activité inhibitrice sur BuChE. Les valeurs CI_{50} sont converties en échelle logarithmique négative [pCI_{50} = -log CI_{50} (µmol/l)].

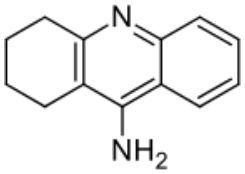


Figure III.2 Structure moléculaire de la Tacrine

Les 151 dérivés de la Tacrine composant la base de données ont été sauvegardés dans un fichier (*.sdf), puis ont été optimisés en utilisant la méthode de la mécanique moléculaire (MM +) pour générer des structures initiales. Afin d'obtenir un minimum d'énergie, une deuxième optimisation géométrique a été réalisée en utilisant une méthode semi-empirique au niveau AM1 implémentée dans le logiciel MOPAC 2012 [16].

III.1.2 Calcul et selection des descripteurs

Une étape importante dans l'obtention d'un modèle QSAR est la représentation numérique des caractéristiques structurelles des molécules, appelées descripteurs moléculaires. Les descripteurs moléculaires jouent un rôle fondamental dans le développement de modèles QSAR. 2430 descripteurs ont ainsi été obtenus pour chaque composé à partir du programme en ligne E-Dragon 1.0 (www.vcclab.org) et du logiciel MOE [17].Toutes les informations nécessaires sur ces descripteurs sont données dans la littérature [17, 18].

Compte tenu de ce grand nombre de descripteurs et pour éviter le phénomène du sur-apprentissage, il était nécessaire de procéder à une présélection efficace pour ne garder que ceux qui ont la capacité de représenter efficacement la propriété étudiée (pouvoir d'inhibition dans notre cas). La méthode utilisée pour sélectionner les descripteurs les plus significatifs a été décrite précédemment par Hamadache et al.[19] et a lieu comme suit : dans une première étape, les descripteurs invariants, les descripteurs avec des valeurs absentes (représenté par le code "999"), ont été enlevés manuellement.

Tableau III.1 Dérivés de la tacrine utilisés et valeurs des concentrations inhibitrices expérimentales et prédites par MLR, MLP (ANN) et SVR.

N°	PubChem CID(*)	pCI_{50}(µmol/l) expérimentale	pCI_{50} predite par MLR	pCI_{50} predite par MLP	pCI_{50} predite par SVR
Ensemble d'apprentissage					
1	16094876	3.857	3.206	3.526	3.715
2	16094860	3.851	3.648	4.030	3.709
3	16094859	3.646	3.490	3.747	3.548
4	90654210	3.62	3.560	3.650	3.480
5	90654209	3.602	3.110	3.243	3.328
6	54768920	1.420	1.807	1.483	1.420
7	71653762	1.420	0.646	0.628	1.069
8	1935	1.398	1.192	1.383	1.258
9	101886789	1.387	1.070	1.061	1.246
10	14372818	1.387	1.141	1.307	1.246
11	54769157	1.387	0.606	0.601	1.247
12	23644776	1.347	1.267	1.151	1.207
13	11304963	1.323	1.150	0.944	1.464
14	205772	1.284	1.849	1.385	1.214
15	10625415	1.276	1.202	1.340	1.136
16	11753295	1.268	1.875	1.956	1.410
17	71654239	1.260	1.169	1.217	1.173
18	46855186	1.260	1.233	1.019	1.119
19	71653619	1.244	1.264	1.227	1.154
20	11455410	1.229	1.481	1.527	1.219
21	11329824	1.222	1.453	1.523	1.190
22	71653919	1.208	0.864	0.863	1.066
23	11169647	1.196	1.263	1.002	1.152
24	54577394	1.194	0.507	0.484	0.510
25	71653918	1.174	1.122	1.088	1.110
26	71654077	1.114	1.180	1.129	1.034
27	11191030	1.102	1.303	1.331	0.962
28	122199174	1.097	0.474	0.624	0.804
29	22023	1.097	0.655	0.748	0.956
30	118732688	1.097	0.884	0.823	0.980
31	10883396	1.097	0.664	0.591	0.956
32	71653620	1.086	1.031	1.026	1.113
33	1593398	1.081	1.031	1.189	1.137
34	44342760	1.046	0.630	0.534	0.618
35	54768918	1.041	1.378	1.149	1.085
36	71653761	1.036	0.886	0.933	1.136
37	23644778	1.032	1.278	1.268	1.173

Tableau III.1 (suite)

N°	PubChem CID(*)	pCI_{50}(µmol/l) expérimentale	pCI_{50} predite par MLR	pCI_{50} predite par MLP	pCI_{50} predite par SVR
Ensemble d'apprentissage					
38	23644783	1.009	1.355	1.224	1.150
39	54768917	0.996	1.407	1.328	1.137
40	54768922	0.996	0.876	0.672	0.927
41	71653617	0.963	0.834	0.834	1.104
42	71653760	0.921	1.321	1.274	1.061
43	52948017	0.921	0.224	0.164	0.424
44	118734712	0.907	0.281	0.247	0.766
45	44342769	0.886	0.457	0.330	0.263
46	118732691	0.886	1.017	0.805	0.547
47	24949525	0.863	1.146	1.204	0.924
48	118732692	0.796	1.175	0.936	0.654
49	9850331	0.785	1.390	1.174	1.076
50	24949524	0.724	1.059	1.132	0.864
51	118732690	0.721	0.990	0.824	0.677
52	9851127	0.712	0.564	0.469	0.570
53	101886794	0.706	0.699	0.604	0.555
54	16077296	0.699	0.621	0.799	0.839
55	122199187	0.699	0.272	0.412	0.684
56	24949523	0.676	0.856	0.904	0.817
57	118709901	0.642	1.077	0.664	0.783
58	101886788	0.592	0.525	0.472	0.467
59	9828345	0.590	1.232	1.033	0.977
60	101886796	0.588	0.912	0.830	0.580
61	102369422	0.573	0.338	0.211	0.431
62	118721872	0.542	0.554	0.456	0.682
63	9861757	0.538	0.422	0.372	0.391
64	23644777	0.506	0.350	0.233	0.490
65	102369423	0.484	0.941	0.680	0.526
66	118721629	0.483	0.552	0.493	0.343
67	72549439	0.456	0.399	0.344	0.219
68	52946279	0.444	0.447	0.193	0.584
69	44342592	0.398	0.739	0.573	0.429
70	1933	0.398	0.717	1.030	0.540
71	11067083	0.398	0.278	0.236	0.120
72	101886793	0.300	-0.053	0.100	0.159
73	90655199	0.292	-0.245	0.093	0.432
74	44342795	0.268	0.922	0.696	0.408
75	78350462	0.266	0.709	0.660	0.425
76	118732689	0.252	0.697	0.548	0.493

Tableau III.1 (suite)

N°	PubChem CID(*)	pCI_{50}(µmol/l) expérimentale	pCI_{50} predite par MLR	pCI_{50} predite par MLP	pCI_{50} predite par SVR
Ensemble d'apprentissage					
77	44224105	0.232	-0.189	0.126	0.092
78	56955111	3.097	2.745	3.014	2.726
79	71478328	3.056	2.543	2.443	2.869
80	16094856	3.014	2.323	2.512	2.874
81	11713155	3.000	2.733	3.038	2.860
82	56955109	3.000	2.889	3.184	2.859
83	90654212	2.921	2.980	3.004	2.864
84	71478270	2.883	3.190	3.193	2.868
85	16094874	2.860	2.508	2.725	2.719
86	56955100	2.824	2.359	2.507	2.491
87	90654211	2.785	2.918	2.738	2.926
88	90654203	2.690	2.925	3.016	3.001
89	56955199	2.602	2.418	2.564	2.461
90	9926968	0.208	0.218	0.210	0.314
91	118721626	0.197	0.199	0.229	0.200
92	90655197	0.190	0.738	0.512	0.332
93	11130919	0.161	0.532	0.417	0.281
94	44342503	0.155	0.772	0.575	0.319
95	9926905	0.155	0.580	0.434	0.295
96	10433216	0.148	0.322	0.241	0.139
97	118720935	0.130	-0.349	0.090	0.145
98	118709900	0.116	0.284	0.206	0.256
99	9882480	0.114	0.422	0.333	0.597
100	118709916	0.082	0.422	0.291	0.223
101	118721630	0.013	-0.048	0.086	0.130
102	118720934	0.008	0.021	0.154	0.148
103	52941983	0.000	-0.361	0.092	0.141
104	10049496	-0.004	0.556	0.284	0.137
105	44224109	-0.032	-0.082	0.126	0.109
106	72549683	-0.041	0.630	0.456	0.276
107	118721628	-0.107	0.296	0.322	0.235
108	44357449	-0.114	-0.471	-0.168	0.027
109	122199179	-0.146	-0.176	-0.176	-0.005
110	118709917	-0.152	0.324	-0.137	-0.011
111	70692913	-0.182	0.477	0.339	0.109
112	122199188	-0.204	-0.511	-0.440	-0.345
113	54577395	-0.220	0.260	0.128	0.174
114	44224101	-0.286	-0.556	0.053	-0.145

Tableau III.1 (suite)

N°	PubChem CID(*)	pCI_{50}(µmol/l) expérimentale	pCI_{50} predite par MLR	pCI_{50} predite par MLP	pCI_{50} predite par SVR
115	122199176	-0.380	-0.553	-0.436	-0.282
116	11352229	2.489	2.083	1.985	2.349
117	90654202	2.369	2.360	2.348	2.510
118	122199180	-0.623	-0.543	-0.464	-0.682
119	122199177	-0.716	-0.092	-0.042	-0.414
120	118709918	-0.750	-0.999	-0.531	-0.609
121	122199190	-0.833	-0.757	-0.596	-0.692
Ensemble de test					
1	90654201	3.602	2.961	3.111	3.149
2	71478203	3.553	3.347	3.455	3.498
3	90654204	3.420	2.604	2.656	2.552
4	24800299	1.387	1.488	1.420	1.312
5	54768919	1.367	1.627	1.352	1.275
6	118721875	1.319	1.163	1.069	1.224
7	9549196	1.268	0.978	1.069	1.078
8	71654078	1.237	0.918	1.040	1.227
9	23644779	1.229	1.269	1.255	1.092
10	23644784	1.125	1.502	1.359	1.199
11	10322257	1.046	0.745	0.315	0.178
12	118721874	1.018	0.827	0.737	1.047
13	118732693	0.959	1.421	1.132	0.903
14	118734713	0.947	0.181	0.203	0.638
15	24881944	0.866	0.054	0.183	0.834
16	11341105	0.866	0.997	0.903	0.737
17	70684451	0.676	-0.086	0.079	0.154
18	44224107	0.662	-0.226	0.114	0.023
19	118721873	0.631	0.896	0.755	0.867
20	44224108	0.631	-0.231	0.111	0.059
21	24949376	0.529	0.355	0.284	0.365
22	90654213	3.108	3.131	2.896	2.787
23	56955099	2.824	2.691	2.917	2.414
24	90654208	2.684	2.598	2.794	2.599
25	118721871	0.130	0.336	0.288	0.229
26	122199184	-0.279	0.175	0.134	0.366
27	11504245	2.523	1.945	2.293	2.459
28	56955198	2.301	2.566	2.740	2.600
29	11385572	2.298	2.287	2.640	2.472
30	10711007	-0.602	-0.031	-0.018	-0.260

(*) CID : une notation utilisée par PubChem pour Identifier un composé " *PubChem Compound Identification*".

Ensuite, tout descripteur qui a une valeur identique pour plus de 75% des échantillons et des descripteurs avec l'écart type relatif inférieur à 0,05 sont enlevés. Enfin, l'un des deux d'un couple de descripteurs avec une valeur absolue du coefficient de corrélation de Pearson supérieur à 0,75 est retiré. Par ailleurs, les descripteurs sélectionnés ne doivent pas être multi-colinéaires. Aussi, afin de les détecter et d'identifier les descripteurs qui y sont impliqués, on détermine le paramètre VIF (Variance Inflation Factor) qu'on calcule selon l'équation III.1. La valeur du paramètre VIF, calculée à l'aide du logiciel XLSTAT, doit se situer dans la gamme de valeur de 1 à 5 [20].

$$VIF = \frac{1}{1-R^2} \qquad \textbf{(III.1)}$$

III.1.3 Développement des modèles QSAR

La base de données contenant 151 composés a été divisée en deux ensembles: un ensemble d'apprentissage englobant 80% et un ensemble de avec 20% du nombre total de composés. L'ensemble d'apprentissage avec 121 dérivés est dédié au développement du modèle QSAR. Par contre, l'ensemble de test, qui comprend 30 dérivés de la tacrine qui n'avaient pas été utilisés pour le développement des modèles QSAR, a été laissé pour la validation externe.

Pour le développement des modèles QSAR, trois méthodes d'analyse de données ont été utilisées: la régression linéaire multiple (RLM ou MLR en anglais), les réseaux de neurones artificiels (RNA ou ANN en anglais) et les machines à vecteurs supports pour la régression (SVR).

L'approche MLR a été réalisée à l'aide du logiciel MLRplusValidation 1.3 (http://dtclab.webs.com/softwaretools ou http://teqip.jdvu.ac.in/QSAR_Tools/). Les approches ANN et SVR ont, quant à elles, été réalisées à l'aide du logiciel STATISTICA (STATISTICA 8.0, StatSoft, Inc., Tulsa, OK, USA).

III.1.4 Validation des modèles QSAR

Comme indiqué au paragraphe II.8 du chapitre II, la validation est un aspect crucial et important pour la détermination de la fiabilité des modèles. Il existe plusieurs approches de validation, dont la validation interne et externe. Des études récentes [21] ont indiqué que la validation interne est jugée nécessaire pour la validation du modèle. De plus, la validation externe est une méthode de validation importante et nécessaire utilisée pour déterminer à la fois la généralisabilité et la capacité prédictive réelle des modèles QSAR [22]. Récemment, Roy et al. [23] ont proposé des paramètres (r_m^2) comme outils de validation supplémentaires dont le calcul se fait au niveau du site suivant : http://aptsoftware.co.in/rmsquare/.

III.1.4.1 Validation interne

En plus de l'erreur quadratique moyenne (RMSE) et du coefficient de détermination (R^2), les principaux paramètres statistiques les plus importants utilisés dans ce travail pour vérifier les performances des modèles sont les suivants :

- le coefficient de validation croisée (Q^2_{appr} ou Q^2_{LOO}) qui doit être > 0,5
- les métriques $\overline{r_m^2}$ et Δr_m^2 introduits par Roy et al. [23]. Il a été démontré que la valeur de $\Delta r_{m(LOO)}^2$ doit être de préférence inférieure à 0,2 à condition que la valeur de $\overline{r_{m(LOO)}^2}$ soit supérieure à 0,5.

III.1.4.2 Validation externe

Les paramètres et critères de validation externe les plus couramment utilisés sont donnés ci-dessous :

- RMSE et R^2
- Le coefficient Q^2_{test} (ou R^2_{pred}) qui reflète le degré de corrélation entre les données d'activité observées et prédites de l'ensemble de test. Les modèles avec des valeurs de R^2_{pred} supérieures à la valeur stipulée de 0.5 sont considérés comme bien prédictifs.
- Critères de Golbraikh et Tropsha : ces derniers [24] ont proposé un ensemble de paramètres pour déterminer la prévisibilité externe du modèle QSAR. Selon eux, les modèles sont considérés comme satisfaisants, si toutes les conditions suivantes sont remplies :
- $Q^2_{train} > 0.5.$
  $Q^2_{test} > 0.6.$
  $\frac{r^2 - r_0^2}{r^2} < 0.1 \quad et \quad 0.85 \leq k \leq 1.15$ ou $\frac{r^2 - r'^2_0}{r^2} < 0.1 \quad et \quad 0.85 \leq k' \leq 1.15.$
  $|r_0^2 - r'^2_0| < 0.3.$
- Le paramètre CCC peut être calculé afin de vérifier la fiabilité du modèle, en détectant la distance des observations de la ligne d'ajustement et le degré d'écart de la ligne de régression par rapport à celui qui passe par l'origine, respectivement. Tout écart de la ligne de régression par rapport à la ligne de concordance (ligne passant par l'origine) donne une valeur CCC inférieure à1 [25, 26].

Les équations de ces différents paramètres de validation interne et externe et la signification des différents termes sont donnés ci-dessous.

$$RMSE = \sqrt{\frac{\sum \left(Y_{\exp} - Y_{pred}\right)^2}{n}} \quad \textbf{(III.2)}$$

$$R^2 = 1 - \frac{\sum \left(Y_{\exp} - Y_{pred}\right)^2}{\sum (Y_{\exp} - \bar{Y}_{\exp})^2} \quad \textbf{(III.3)}$$

$$Q^2_{app} = 1 - \frac{\sum \left(Y_{\exp(appr)} - Y_{pred\ (appr)}\right)^2}{\sum (Y_{\exp(appr)} - \bar{Y}_{\exp(appr)})^2} \quad \textbf{(III.4)}$$

$$Q^2_{test} = 1 - \frac{\sum \left(Y_{\exp\ (test)} - Y_{pred\ (test)}\right)^2}{\sum (Y_{\exp\ (test)} - \bar{Y}_{\exp(appr)})^2} \quad \textbf{(III.5)}$$

$$r^2_m = r^2 \left(1 - \sqrt{r^2 - r^2_0}\right) \quad \textbf{(III.6)}$$

$$r'^2_m = r^2 \left(1 - \sqrt{r^2 - r'^2_0}\right) \quad \textbf{(III.7)}$$

$$\bar{r}^2_m = \frac{\left(r^2_m + r'^2_m\right)}{2} \quad \textbf{(III.8)}$$

$$\Delta r^2_m = \left| r^2_m - r'^2_m \right| \quad \textbf{(III.9)}$$

$$CCC = \frac{2 \sum \left(Y_{exp(test)} - \bar{Y}_{exp(test)}\right)\left(Y_{pred(test)} - \bar{Y}_{pred(test)}\right)}{\sum \left(Y_{exp(test)} - \bar{Y}_{exp(test)}\right)^2 + \sum \left(Y_{pred(test)} - \bar{Y}_{pred(test)}\right)^2 + n \sum \left(\left(\bar{Y}_{exp(test)} - \bar{Y}_{pred(test)}\right)\right)} \quad \textbf{(III.10)}$$

où $Y_{\exp(appr)}$ et $Y_{\exp(test)}$ sont les valeurs expérimentales de la propriété étudiée pour les ensembles d'apprentissage et de test, $Y_{\mathrm{pred}(appr)}$ et $Y_{\mathrm{pred}(test)}$ sont les valeurs prédites de la propriété étudiée pour les ensembles d'apprentissage et de test, n est le nombre de composes (ou d'observations) de la base de données(apprentissage ou test), $\bar{Y}$ est la valeur moyenne de la propriété dans l'ensemble d'apprentissage ou de test et dans le cas expérimental ou prédit.

III.1.5 Domaine d'applicabilité

Le domaine d'applicabilité (DA) est défini comme la "justification qu'un modèle dans son domaine d'applicabilité possède une plage de précision satisfaisante dans l'application prévue du modèle". En d'autres termes, c'est une région théorique dans l'espace défini par les descripteurs du modèle et la réponse modélisée pour laquelle un QSAR donné devrait faire des prévisions fiables.Les modèles QSAR ne sont valides que dans le domaine où ils ont été formés et validés. L'extrapolation est dangereuse et peut conduire à des prédictions de modèle grossièrement erronées [27]. La détermination du DA est donc d'une grande importance [28].

Il existe différentes approches pour déterminer la DA des modèles QSAR. Chaque méthode a ses propres avantages et inconvénients. Dans le cadre de notre travail, nous avons utilisé à des fins de comparaison différentes approches: (1) approche par effet de levier (graphique de Williams) qui a été largement utilisée pour identifier les valeurs aberrantes et les composés résidant en dehors du DA ; (2) approche de standardisation (approche statistique simple) pour définir le DA d'un modèle QSAR. Cette approche, qui a été rapportée par Roy et al. [29] est très facile; (3) bounding box ; (4) Euclidean distance (95 percentile) ; (5) classical kNN et (6) modified kNN. La théorie de fond, l'algorithme de calcul, la méthodologie et les avantages des approches proposées sont disponibles dans la littérature [29 - 31].

III.2 RESULTATS ET DISCUSSION

III.2.1 Calcul et selection des descripteurs

Le nombre de descripteurs obtenus après la sélection effectuée était de 21. Etant donné les critères rapportés au paragraphe II.7.1 du chapitre II, il est nécessaire et important de réduire le nombre de descripteurs jusqu'à ce que le rapport "nombre de dérivés des Tacrine**/**descripteurs" soit **>** 5 et qu'ils ne soient pas corrélés entre eux [32]. Après avoir utilisé l'algorithme génétique implémenté dans le logiciel BuildQsar [36], 8 descripteurs(ATSC5e, MATS5s, GATS7c, SpMax3_Bhm, SpMin1_Bhi, RotBFrac, SRW9 et vsa_don) ont été sélectionnés pour le développement des modèles prédictifs.

Afin de vérifier l'inter-corrélation de huit descripteurs, l'analyse du facteur d'inflation de la variance (VIF) a été effectuée. De plus, pour examiner l'importance relative et la contribution de chaque descripteur dans les modèles QSAR, la valeur de l'effet moyen (Mean effect en anglais : MF) a été calculée pour chaque descripteur. Les valeurs du facteur VIF et de l'effet MF sont présentées dans le tableau III.2.

D'après le tableau III.2, tous les descripteurs ont des valeurs VIF inférieures à 5, ce qui indique que la selection effectuée est bonne et que les descripteurs obtenus ne sont pas corrélés entre eux.

Tableau III.2 Description, VIF et MF des descripteurs sélectionnés

Ref	Descripteurs	Description	VIF	MF
[18]	ATSC5e	Centered Broto-Moreau autocorrelation - lag 5.	1.759	-0.002
[18]	MATS5s	Moran autocorrelation - lag 5.	2.495	0.010
[18]	GATS7c	Geary autocorrelation - lag 7.	1.836	0.090
[18]	SpMax3_Bhm	Largest absolute eigenvalue of Burden modified matrix - n 3.	1.665	0.133
[18]	SpMin1_Bhi	Smallest absolute eigenvalue of Burden modified matrix - n 1.	1.313	0.860
[18]	RotBFrac	Fraction of rotatable bonds, excluding terminal bonds.	1.679	-0.062
[18]	SRW9	Self-returning walk count of order 9 (ln(1+x).	1.368	-0.002
[34]	vsa_don	Approximation to the sum of VDW surface areas of pure hydrogen bond donors.	1.308	-0.027

III.2.2 Modèle MLR

Le modèle MLR obtenu pour la prédiction de la concentration inhibitrice des dérivés de la Tacrine, en utilisant 151 composés, est donné par la relation linéaire ci-dessous. Les valeurs de pCI_{50} prédites par le modèle MLR pour l'ensemble des données (151 composés) sont présentées dans le Tableau III.1.

pCI_{50} = 30.20351 (+/-4.94621) - 0.10371 (+/-0.05156) ATSC5e + 3.48641 (+/-0.48069) MATS5s - 3.50392 (+/-0.25099) GATS7c - 1.05831 (+/-0.30975) SpMax3_Bhm - 12.42265 (+/-2.6109) SpMin1_Bhi + 7.56893 (+/-0.5521) RotBFrac + 0.02208 (+/-0.01261) SRW9 + 0.0691 (+/-0.00657) vsa_don

(III.11)

N_{appr} = 121; R^2 = 0.879, R^2_{adj} = 0.870, F = 101.462, p < 0.0001, Q^2 = 0.857, PRESS = 16.125, N_{test} = 30, R^2 = 0.847, $r_0^2 = 0.828$, $\overline{r_m^2} = 0.787$, ${}^cR_p^2 = 0.849$

Les erreurs-types des coefficients de régression sont données entre parenthèses. D'après les résultats statistiques, les huit descripteurs pertinents (variables) de l'équation III.10 pourraient expliquer 87% de la variance (coefficient de variation ajusté) de la concentration inhibitrice. La différence entre R^2 et Q^2 est égale à 0,022. Cette différence étant inférieure à 0.3, cela signifie que ce modèle est robuste [35]. D'autre part, pour vérifier que l'obtention du modèle MLR n'est pas dû au hasard ou à une chance, on a eu recours au coefficient ${}^cR_p^2$ suggéré par Mitra et al. [36]**.** La valeur de ${}^cR_p^2$ (0,847) est supérieure à 0,5 (valeur seuil), ce qui signifie que le modèle obtenu n'est pas dû au hasard.

Une valeur acceptable de $\overline{r_m^2}$ (0,787) a été également obtenue pour l'ensemble de test, ce qui indique que ce modèle a un pouvoir prédictif satisfaisant.

Par ailleurs, comme le montre le tableau III.2, la valeur du MF du descripteur SpMin1_Bhi est supérieure à celle des autres descripteurs. Cela stipule que, dans ce modèle QSAR, l'influence de SpMin1_Bhi sur le pouvoir inhibiteur est la plus forte. Par contre, le descripteur RotBFrac est celui qui a la plus faible influence sur le pouvoir inhibiteur.

Les signes des coefficients de régression suggèrent la direction de l'influence des descripteurs dans un modèle donné. D'après les coefficients de régression de l'équation III.10, il convient de noter que les coefficients de régression des descripteurs MATS5s, RotBFrac, SRW9 et vsa_don ont des signes positifs. Ces descripteurs ont eu une influence positive et les valeurs les plus élevées vont contribuer à améliorer le pouvoir inhibiteur des dérivés de la Tacrine. Le descripteur MATS5s est une autocorrélation Moran 2D pondérée par les indices de pseudo-connectivité moléculaire à l'état intrinsèque. Ainsi, augmenter la taille de la molécule augmente la valeur de MATS5s, ce qui conduira à une augmentation du pouvoir inhibiteur des dérivés de la Tacrine. Le descripteur RotBFrac est synonyme du nombre de liaisons dans la molécule ayant des rotations considérées comme significatives pour la mécanique moléculaire. Tous les atomes d'hydrogène terminaux sont ignorés. Le descripteur SRW9 peut être associé à d'éventuels mouvements d'électrons. Aussi, une augmentation de la valeur de SRW9 conduit à une augmentation du pouvoir inhibiteur de la molécule. Le descripteur vsa_don appartient aux descripteurs 2D MOE décrivant les caractéristiques pharmacophoriques et la surface moléculaire totale des donneurs de liaisons hydrogène pures. Une augmentation de la surface moléculaire totale des donneurs de liaison hydrogène pure dans une molécule conduit à une augmentation inhibitrice de la molécule.

Par contre, les coefficients de régression des descripteurs ATSC5e, GATS7c, SpMax3_Bhm, SpMin1_Bhi, présentent des signes négatifs, ce qui aura un impact négatif sur le pouvoir inhibiteur des dérivés de la Tacrine. Le descripteur ATSC5e est un descripteur d'autocorrélation Moreau-Broto 2D, défini pour le chemin de cinq liaisons et pondéré par des charges partielles. C'est un indicateur d'association de charge partielle spatiale. Ainsi, une augmentation de la charge partielle spatiale d'une molécule entraîne une diminution de son pouvoir inhibiteur. Le descripteur GATS7c est une autocorrélation 2D Geary pondérée par les charges. Une augmentation de la valeur de GATS7c conduira à une diminution de l'activité inhibitrice. Les valeurs propres de Burden modifiées, sont les valeurs propres d'une matrice de connectivité associée aux propriétés atomiques telles que la masse atomique relative, la polarisabilité, l'électronégativité de Sanderson et le volume de Van der Waals. De petites valeurs pour les descripteurs SpMax3_Bhm et SpMin1_Bhi contribueront à l'amélioration de l'activité inhibitrice des dérivés de la Tacrine. Par contre une augmentation aura pour conséquence une altération du pouvoir inhibiteur.

En conclusion, on peut affirmer que le pouvoir inhibiteur des dérivés de la Tacrine est principalement influencé par la forme et les masses atomiques des molécules (MATS5s, SpMax3_Bhm et RotBFrac), ainsi que par la charge négative relative, la polarisabilité atomique et l'électronégativité atomique des molécules (ATSc5e, GATS7c, SpMin1_Bhi, SRW9 et vsa_don).

III.2.3 Modèle neuronal MLP

L'utilisation des réseaux de neurones artificiels (ANN) est nécessaire et est devenue une technique de modélisation importante. Elle est largement utilisée pour les études de QSAR non linéaires et est un outil puissant pour construire des modèles prédictifs. Dans le cadre de ce travail, les réseaux de neurones artificiels (ANN) ont été utilisés pour construire un modèle non linéaire sur la base des mêmes descripteurs utilisés dans le cas du modèle MLR. Un réseau typique à trois couches de type perceptron multicouche (MLP) avec une couche d'entrée, une couche cachée et une couche de sortie a été adopté dans cette étude. Des travaux théoriques ont montré qu'une seule couche cachée suffit pour que l'ANN se rapproche d'une fonction non linéaire complexe et que de nombreux résultats expérimentaux semblent confirmer qu'une couche cachée peut suffire à la plupart des problèmes de prédiction [37].

Il n'y a pas de principes théoriques rigoureux pour choisir la topologie du réseau appropriée. Ainsi, l'utilisation d'une régression neuronale nécessite l'optimisation de l'architecture du réseau neuronal. Dans cette étude, la fonction sigmoïde a été utilisée comme fonction de transfert de la couche cachée et la fonction identité comme fonction de transfert de la couche de sortie. Le réseau a été formé à l'aide de l'algorithme des méthodes quasi-Newton BFGS. Un neurone de sortie a été utilisé pour représenter la pCI_{50} observée. La couche d'entrée est formée des 8 descripteurs (ou variables) sélectionnés auparavant. L'ensemble de données (151 composés) des dérivés de la Tacrine a été divisé en deux lots: un ensemble d'apprentissage et un ensemble de test (ou de validation) composés respectivement de 121 et 30 composés.

La performance optimale du modèle MLP a été évaluée en termes d'erreur quadratique moyenne (RMSE) et de Q^2 [41, 42]. Pour optimiser le nombre de nœuds dans la couche cachée, plusieurs calculs ont été effectués avec différents nombres de nœuds cachés (de 1 à 10). C'est ainsi que le meilleur modèle choisi présentait une RMSE minimum et une valeur de Q^2 maximum. Par conséquent, un réseau MLP avec l'architecture 8-2-1 a été choisi comme modèle final dans ce travail. Les résultats prédictifs du modèle MLP pour l'ensemble de données complet (151 composés) sont présentés dans le Tableau III.1**.**

Les concentrations inhibitrices prédites par le modèle MLP pour l'ensemble des données (151 composés) ont été obtenus; elles sont rassemblées dans le Tableau III.1.

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

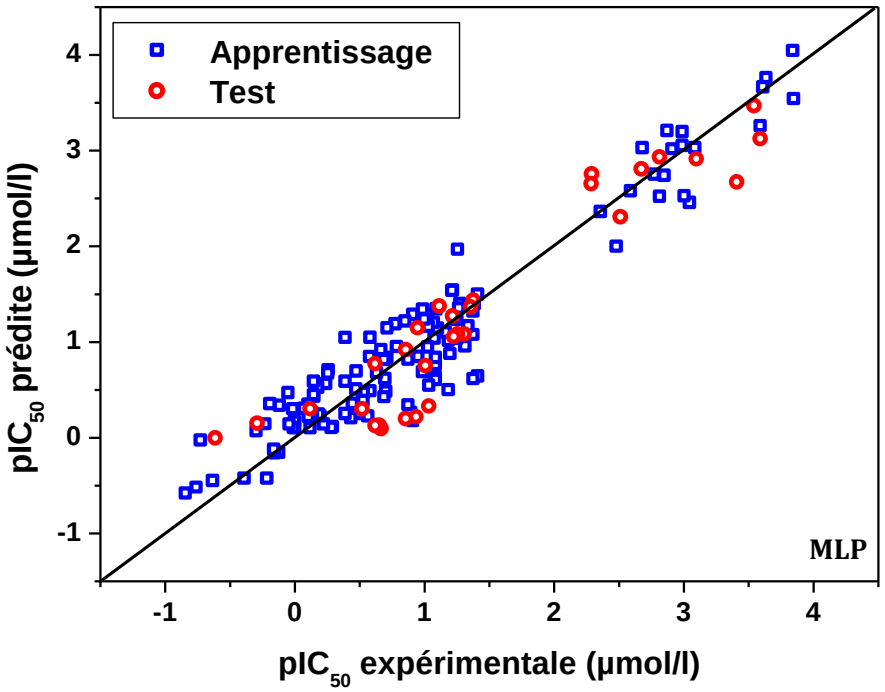


Figure III.3 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentales et celle prédite par le modèle MLP pour une série de dérivés de la Tacrine

D'après la figure III.3, une bonne corrélation entre les valeurs prédites par le modèle MLP et les valeurs expérimentale a été constatée. Comme on peut le voir dans le tableau III.3, le modèle non linéaire des réseaux de neurones (modèle MLP) a donné de bons résultats avec des coefficients de corrélation R^2 élevés, ainsi qu'une meilleure robustesse (Q^2) aussi bien lors de l'apprentissage que lors du test. En plus de ces paramètres de validation classiques, différentes coefficients r_m ont également été vérifiés pour les ensembles d'apprentissage et de test. Les valeurs de $\overline{r_m^2}$ pour l'ensemble d'apprentissage (0.882) et l'ensemble de test (0.835) sont supérieurs à 0.5. De plus, les valeurs de Δr_m^2 pour les ensembles d'apprentissage (0,058) et de test (0,092) sont toutes inférieures à 0,2 [25]. Par ailleurs, le graphique des résidus pour les valeurs expérimentales de pCI_{50} pour les ensembles d'apprentissage et de test par rapport à leurs valeurs prédites ont été étudiés. Le modèle n'a pas montré d'erreur systématique, car la distribution des résidus des deux côtés de la ligne médiane (ligne zéro) est aléatoire.

Selon la recommandation de Tropsha et Golbraikh [24], si la différence entre R^2 et Q^2_{appr} ne dépasse pas 0.3, il n'y'a pas de sur-apprentissage dans le modèle. Dans le présent travail, cette différence est égale à 0. En conséquence, il y'a absence du sur-apprentissage dans le modèle MLP.

Tableau III.3 Paramètres statistiques du modèle MLP pour les ensembles d'apprentissage et de test

Paramètres statistiques	**Apprentissage**	**Test**	**Critéres de validité du modèle** [28]
N	121	30	
R^2	0.911	0.888	0.6
RMSE	0.312	0.394	
Q^2_{appr}	0.911	---	> 0.5
Q^2_{test}	---	0.895	> 0.6
$\overline{r_m^2}$	0.882	0.835	> 0.5
Δr_m^2	0.058	0.092	< 0.2
k	---	1.036	0.85 < k < 1.15
k'	---	0.920	0.85 < k' < 1.15

Une analyse de sensibilité utilisant la méthode proposée par Goh [40], a été également réalisée pour déterminer l'importance de chaque variable dans la prédiction de l'activité inhibitrice des dérivés de la Tacrine. La contribution de chacun des descripteurs dans le modèle MLP obtenu est reproduite sur la figure III.4 ci-après.

En conclusion, le modèle MLP non linéaire a donné de bons résultats pour tous les paramètres statistiques de l'ensemble d'apprentissage. Les paramètres statistiques obtenus pour le test satisfont évidemment aux critères d'acceptabilité et démontrent ainsi le pouvoir prédictif du modèle développé. Ces résultats indiquent que le modèle MLP a non seulement bien fonctionné dans le développement du modèle, mais possède également un excellent pouvoir de prédiction. Ceci laisse supposer l'existence d'une corrélation non linéaire entre la concentration inhibitrice et les descripteurs sélectionnés.

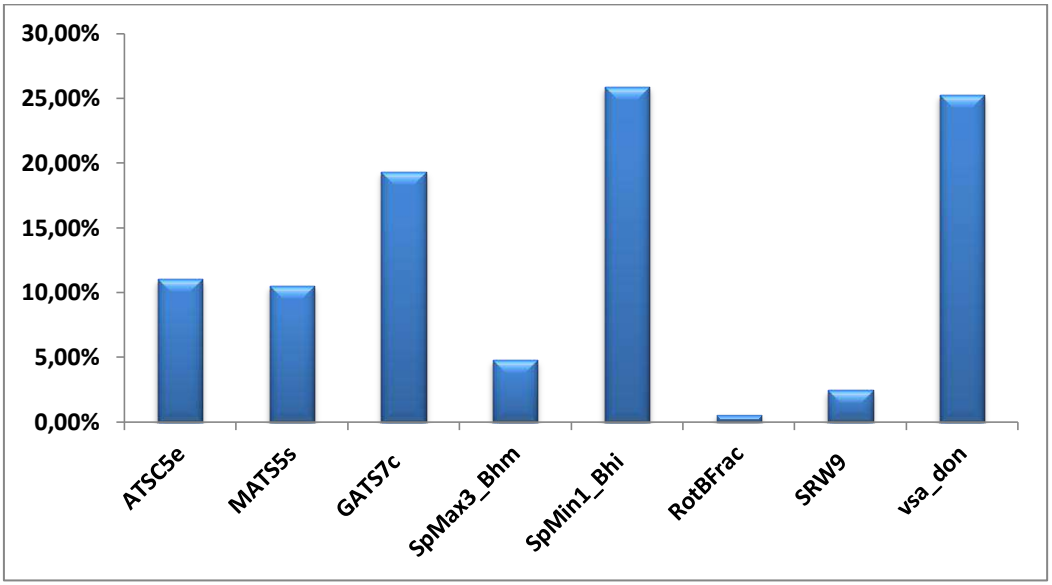


Figure III.4 Contribution des descripteurs dans le cas du modèle MLP

III.2.4 Modèle SVR

L'algorithme de régression du vecteur de support (SVR) comprend trois paramètres à optimiser : le paramètre de capacité C, ε de la fonction de perte ε-insensible et le type de fonction du noyau. Tout d'abord, la fonction du noyau devrait être décidée, ce qui détermine la distribution de l'échantillon dans l'espace de mappage. Comme fonctions d'activation, on a utilisé des fonctions de base radiales (Radial Basis Function : RBF) qui sont couramment utilisées dans de nombreuses études en raison de leur bonne performance générale et de quelques paramètres à ajuster [41]. Dans ce travail, le RBF a été utilisé.

Deuxièmement, le γ de la fonction du noyau affecte grandement le nombre de vecteurs de support, ce qui a une relation étroite avec les performances du SVM et le temps d'apprentissage. Trop de vecteurs de support pourraient produire un sur-apprentissage et augmenter le temps d'entraînement. De plus, γ contrôle l'amplitude de la fonction RBF.

Le paramètre de ε-insensible empêche tout l'ensemble d'apprentissage de répondre aux conditions aux limites et permet donc la possibilité de rareté dans la solution de formulation double. La valeur optimale de ε dépend du type de bruit présent dans les données, qui est généralement inconnu.

Finalement, l'effet du paramètre de capacité C a été testé. Il contrôle le compromis entre la maximisation de la marge et la minimisation de l'erreur d'entraînement. Si "C" est trop petit, le stress lié à l'ajustement des données d'entraînement sera insuffisant. Si "C" est trop grand, l'algorithme sur-adaptera les données d'apprentissage. Pour déterminer les paramètres optimaux, une recherche de grille a été effectuée sur la base de la validation croisée «leave-one-out» (LOOCV) sur l'ensemble d'apprentissage original pour toutes les

combinaisons de paramètres de C de 10 à 100 avec incréments de 1, γ entre 0.01 et 2,5 avec des incréments de 0,01 et ε de 0,01 à 0,5 avec des incréments de 0,01.

Le meilleur modèle SVR est obtenu avec γ égal à 2 (avec C = 10 et ε = 0.06). Les paramètres statistiques trouvés pour ce modèle sont reportés sur le tableau III.4. Les résultats prédictifs du modèle SVR pour l'ensemble de données complet (151compounds) sont obtenus et présentés dans le Tableau III.1**.**

Tableau III.4 Paramètres statistiques et critères de validité du modèle SVR

Paramètres statistiques	**Apprentissage**	**Test**	**Critéres de validité du modèle**
N	121	30	
R^2	0.969	0.907	> 0.6
RMSE	0.197	0.362	
Q^2_{appr}	0.964	---	> 0.5
Q^2_{test}	---	0.891	> 0.6
$\overline{r_m^2}$	0.938	0.868	> 0.5
Δr_m^2	0.023	0.022	< 0.2
k	---	1.070	0.85 < k < 1.15
k'	---	0.901	0.85 < k' < 1.15

Sur la figure III.5 ci-dessous, sont représentées les concentrations inhibitrices prédites par le modèle SVR en fonction des concentrations inhibitrices expérimentales pour l'ensemble d'apprentissage et de test.

D'après la figure III.5, une bonne corrélation entre les valeurs prédites par le modèle SVR et les valeurs expérimentale a été constatée. Comme on peut le voir dans le tableau III.4, le modèle SVR a donné de bons résultats avec des coefficients de corrélation R^2 élevés, ainsi qu'une meilleure robustesse (Q^2) aussi bien lors de l'apprentissage que lors du test. En plus de ces paramètres de validation classiques, différentes coefficients r_m ont également été vérifiés pour les ensembles d'apprentissage et de test. Les grandes valeurs de $\overline{r_m^2}$ obtenues pour les ensembles d'apprentissage et celui du test sont supérieurs à 0.5. De plus, les valeurs de Δr_m^2 pour les ensembles d'apprentissage et de tests ont toutes inférieures à 0,2 [25]. Par ailleurs, le graphique des résidus pour les valeurs expérimentales de pCI_{50} pour les ensembles d'apprentissage et de test par rapport à leurs valeurs prédites ont été étudiés. Le modèle n'a pas montré d'erreur systématique, car la distribution des

résidus des deux côtés de la ligne médiane (ligne zéro) est aléatoire. D'un autre côté, la différence entre R^2 et Q^2_{appr} est de 0.005, ce qui confirme qu'il n'y'a pas eu de sur-apprentissage lors du développement du modèle [24].

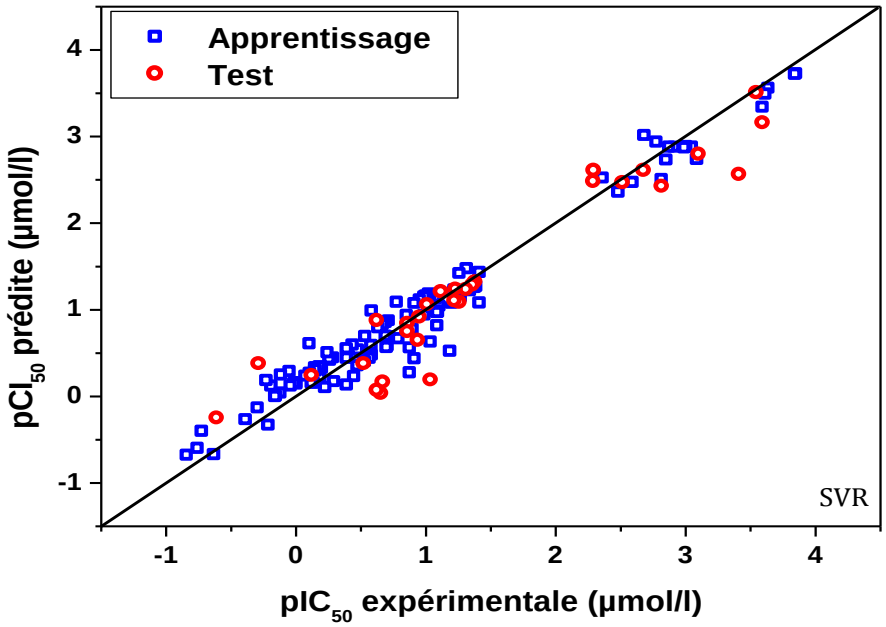


Figure III.5 Représentation graphique des concentrations inhibitrices (pCI_{50}) expérimentale et celle prédite par le modèle SVR pour une série de dérivés de la Tacrine

III.2.5 Comparaison des trois modèles MLR, MLP et SVR

Pour comparer à la fois les performances et la qualité de la prédiction des trois modèles développés dans ce travail (MLR, MLP et SVR), une évaluation statistique est donnée dans le tableau III.5. Les coefficients statistiques de la validation interne pour les 3 modèles sont tous acceptables et satisfaisants. En conséquence, ces modèles sont robustes. La qualité des modèles a également été jugée en termes de différents paramètres de validation externes. Là aussi, les 3 modéles se distinguent par un excellent pouvoir de prédiction.

La grande précision et l'exactitude sont vérifiées avec les valeurs de CCC; ce coefficient mesure à la fois la précision (dans quelle mesure les observations proviennent de la ligne d'ajustement) et l'exactitude (dans quelle mesure la droite de régression dévie de la ligne de pente passant par l'origine). On peut noter que les points dans les trois modèles sont presque parfaitement alignés sur les lignes d'ajustement alors que la précision augmente graduellement de MLR à SVR, suggérant que CCC reflète bien la distribution des points de données dans le modèle SVR.

Tableau III.5 Comparaison des paramètres statistiques des trois modèles

	Modèle MLR	Modèle MLP	Modèle SVR	Critéres de validité
Validation interne				
R^2	0.879	0.911	**0.969**	> 0.6
Q^2_{appr}	0.857	0.911	**0.964**	> 0.5
$\overline{r_m^2}$	0.840	0.882	**0.938**	> 0.5
Δr_m^2	0.025	0.058	**0.023**	<0.2
RMSE	0.365	0.312	**0.197**	
Validation externe				
R^2	0.847	0.888	**0.907**	> 0.6
Q^2_{test}	0.805	**0.895**	0.891	> 0.6
RMSE	0.461	0.394	**0.362**	
$\overline{r_m^2}$	0.787	0.835	**0.868**	> 0.5
Δr_m^2	0.085	0.092	**0.022**	<0.2
CCC	0.910	0.935	**0.942**	
k	1.000	1.036	1.070	0.85 < k < 1.15
k'	0.999	0.920	0.901	0.85 < k' < 1.15

Toutes ces constatations nous permettent d'affirmer que les modèles obtenus permettent l'établissement, dans une large mesure, d'une relation non linéaire entre l'activité thérapeutique des dérivés de la Tacrine et l'information structurale (descripteurs moléculaires sélectionnés). Cependant, une amélioration substantielle des paramètres statistiques pour le modèle SVR peut être notée. Par conséquent, nous pouvons conclure que le modèle SVR semble légèrement meilleur, tant en ce qui concerne la robustesse que le pouvoir de prédiction.

III.2.6 Domaine d'applicabilité des modèles

Le troisième principe des lignes directrices de l'OCDE [42] exige un domaine d'applicabilité défini. Dans ce présent travail, le domaine d'applicabilité a été analysé en

utilisant différentes approches: " *bounding box* ", " *leverage approach* ", " *Euclidean distance (95 percentile)* ", "*classical kNN* ", " *modified kNN* " et approche de standardisation. Sur le Tableau III.6sont reportés les résultats trouvés par application de ces approches sur l'ensemble de test (30 composés) du modèle.

Tableau III.6 Domaine d'applicabilité du modèle SVR pour l'ensemble de test

Approche utilisée	Nombre de composés à l'intérieur du DA	Nombre de composés à l'extérieure du DA
Bounding box	30	0
Leverage	28	2
Euclidean distance (95 percentile)	30	0
Classical kNN (Euclidean dist., *k* = 5)	29	1
kNN (Euclidean dist., *k* = 23)	30	0
Standardization approach	30	0

Pour les méthodes : " *bounding box* ", " *Euclidean distance (95 percentile)* ", " *modified kNN* (distance euclidienne et k = 23) " et approche de standardisation, aucun des composés de l'ensemble test n'est en dehors du domaine d'applicabilité, tandis que la méthode " *classical kNN* " indique que 97% des composés sont à l'intérieur (un seul composé à l'extérieur) du domaine d'applicabilité (Tableau III.6). Par contre, l'approche du leverage nous montre que 2 composés sont à l'extérieur du domaine d'applicabilité, soit 93% de composés sont inclus dans ce domaine.

A titre d'exemple, nous reproduisons le diagramme de Williams (Figure III.6) obtenu par l'approche du leverage pour les ensembles d'apprentissage et de test du modèle SVR. Cette approche est basée sur la variation des résidus r_i de prédiction *standardisés* (équation III.12) en fonction des valeurs des leviers h_i pour chacun des composés. La valeur du levier critique h^* est donnée par l'équation III.13 ci-dessous :

$$r_i = \frac{y_{exp} - y_{pred}}{s} = \frac{y_{exp} - y_{pred}}{\sqrt{\frac{\sum_{i=1}^{n}(y_{exp} - y_{pred})^2}{n-k-1}}} \qquad \textbf{(III.12)}$$

$$h^* = \frac{3(k+1)}{n} \qquad \textbf{(III.13)}$$

Où *n*, *k* et *s* représentent respectivement le nombre d'observation (ou produits) de l'ensemble d'apprentissage ou de test, le nombre de descripteurs et l'écart type.

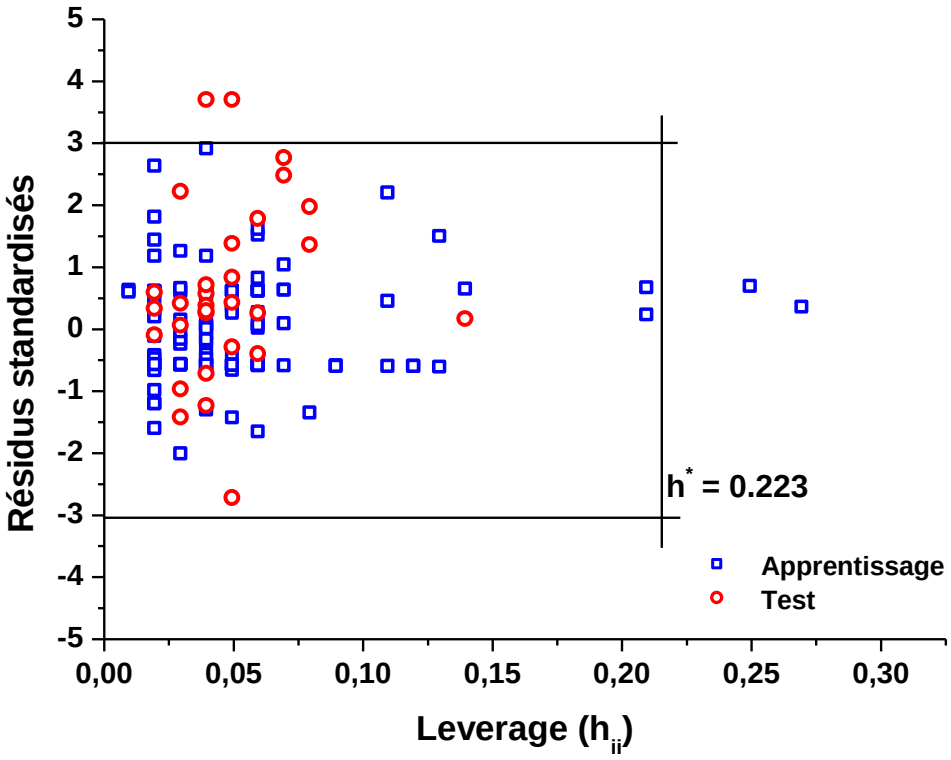


Figure III.6 Diagramme de Williams pour les ensembles d'apprentissage et de test du modèle SVR

On remarque que 2 composés appartenant à l'ensemble de test sont en dehors du domaine d'applicabilité (aire limitée par les deux lignes horizontales et la ligne verticale). Ces résultats montrent que les modèles développés sont conformes au troisième principe de l'OCDE. Ils peuvent être ainsi utilisés pour prédire l'activité inhibitrice de dérivés de la Tacrine, en particulier pour ceux qui n'ont pas été testés ainsi que pour de nouveaux dérivés.

III.3 COMPARAISON AVEC D'AUTRES MODELES PARUS DANS LA LITTERATURE

Après les étapes de validation des modèles développés dans ce présent travail, nous nous sommes fixé comme objectif de les comparer à un nombre limité de modèles QSAR disponibles dans la littérature (tableau III.7). Ces modéles sont consacrés à l'inhibition de l'enzyme BuChE au moyen de divers composés (dérivés de la tacrine, alcaloïdes stéroïdiens, dérivés de la Berbérine et dérivés de la DL0410). Notons que l'évaluation de leurs avantages et inconvénients est assez difficile, car chaque étude publiée utilisait différents ensembles de données et une approche de modélisation différente (descripteurs chimiques, méthodes d'analyse de données, algorithmes, etc.) .

En outre, il convient de noter que la plupart des modèles QSAR rapportés dans la littérature ont été obtenus avec des bases de données assez petites et dont le nombre ne dépasse pas la cinquantaine de composés. Sur le tableau III.7, il est loisible d'observer que, contrairement à notre modèle, aucune approche de la validation externe n'a été réalisée dans les autres modèles, à l'exception du modèle de Castilho et al. [43] qui ont utilisé un seul paramètre (Q^2_{ext}). Ainsi, la comparaison était limitée aux résultats obtenus pour les statistiques de la validation interne. Là encore, le nombre de paramètres statistiques utilisés pour la validation interne de ces modèles QSAR est limité en comparaison de ceux utilisés pour notre modèle. Il est possible d'observer que tous ces modèles ont donné des valeurs élevées du coefficient de corrélation R^2. De plus, le domaine d'applicabilité des modèles de la littérature n'ont pas été établis. Selon ces résultats, le modèle établi par non soins peut être utilisé de manière prometteuse pour prédire la concentration inhibitrice de nouveaux composés, contribuant ainsi à la conception de médicaments anti-Alzheimer, tout en réalisant des économies substantielles d'argent et de temps.

III.4 CONCLUSION

Trois modèles QSAR ont été développés pour la prédiction de la concentration inhibitrice sur l'enzyme BuChE de dérivés de la tacrine. Dans un premier temps, une base de données de la CI_{50} de 151 dérivés a été établie. Le calcul et la selection des descripteurs nous ont permis de dégager 8 descripteurs pertinents qui ne sont pas inter-corrélés entre eux. La signification de ces descripteurs nous a amené à affirmer que le pouvoir inhibiteur des dérivés de la Tacrine est principalement influencé par la forme et les masses atomiques des molécules, la charge négative relative, la polarisabilité atomique et l'électronégativité atomique des molécules.

Au niveau de l'analyse des données, l'utilisation de l'approche MLR, des réseaux de neurones (Perceptron multicouche) et de l'approche SVR nous a permis de développer 3 modèles de prédictions. Une fois développés, ces modèles ont été soumis aux tests de validation aussi bien interne qu'externe. A la lecture des paramètres de validation, on a conclu que les 3 modèles sont robustes et sont dotés d'un pouvoir prédictif intéressant. Néanmoins, le modèle SVR s'est avéré être légèrement meilleur que les deux autres. Enfin, l'établissement du domaine d'applicabilité du modèle SVR selon six approches différentes a montré que 93 à 100% de composés du test appartiennent à ce domaine.

Ces modèles, et en particulier le modèle SVR, représentent des outils tout-à-fait intéressants pour une utilisation dans la prédiction de l'activité inhibitrice sur BuChE de dérivés de la Tacrine, surtout pour ceux dont la CI_{50} n'a pas été déterminée ainsi que pour de nouveaux dérivés.

Tableau III.7 Comparaison entre les modéles de la littérature et celui développé dans ce travail (enzyme BuChE)

Modèles	**Validation interne**						**Validation externe**					
	n	**R^2**	**RMS/SEE**	Q^2_{Loo}	$\overline{r^2_m}$	Δr^2_m	**n**	**R^2_{ext}**	**RMS/SEE**	Q^2_{ext}	$\overline{r^2_m}$	Δr^2_m
Castilho et al. [43]	40	0.84		0.75			10	0.93		0.72		
Zaheer-ul-Haq et al. [44]	34	0.98	0.07	0.70			5	0.68	0.27			
Fang et al. [45]	48	0.88	0.02	0.78			19	0.77				
Pang et al. [46]	35	0.88					10	0.88				
Présent travail	**121**	**0.97**	**0.20**	**0.96**	**0.94**	**0.02**	**30**	**0.91**	**0.36**	**0.89**	**0.87**	**0.02**

2éme Partie

MODELES QSAR DE PREDICTION DU POUVOIR INHIBITEUR D'UNE SERIE DE DERIVES DE LA N-BENZYLPIPERIDINE SUR L'ENZYME ACETYLCHOLINESTERASE (AChE)

III.4 MATERIELS ET METHODES

III.4.1 Collecte des données de l'activité inhibitrice

99 dérivés de la N-benzylpipéridine ont été utilisés comme ensemble de données pour la présente étude. Le squelette structural commun des inhibiteurs de l'AChE est donné sur la figure III.7. Les données de la CI_{50} correspondent à la concentration (mol/l) d'un médicament nécessaire pour une inhibition de 50% in vitro et uniquement celles mesurées par la méthode d'Ellman et al. [14] ont été sélectionnées. Les valeurs de la concentration inhibitrice (CI_{50}) extraites de la littérature [3, 6, 9-13] sont reportées sur le Tableau III.8. Les valeurs des CI_{50} ont été converties en échelle logarithmique négative (pCI_{50} = -logCI_{50} (M)). Tous les composés ont été optimisés en utilisant la méthode de mécanique moléculaire (MM +) pour générer les structures initiales. Afin d'obtenir le minimum d'énergie, une deuxième optimisation géométrique a été réalisée en utilisant une méthode semi empirique au niveau AM1 implémentée dans le logiciel MOPAC 2012 [47].

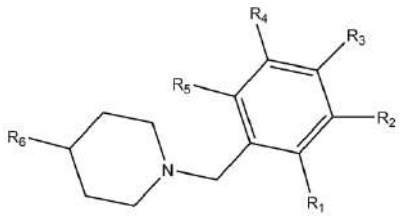


Figure III.7 Structure chimique de la N-benzylpipéridine

III.4.2 Calcul et sélection des descripteurs

2489 descripteurs ont été obtenus pour chaque composé à partir du programme en ligne E-Dragon 1.0 (www.vcclab.org) et du logiciel MOE [17]. Toutes les informations nécessaires sur ces descripteurs sont fournies dans la littérature [17-18]. Le même procédé de sélection des descripteurs pertinents utilisé lors de la 1ére partie (voir paragraphe III.1.2) a été utilisé dans cette seconde partie.

Tableau III.8 Dérivés de la N-benzylpipéridine utilisés et valeurs des concentrations inhibitrices expérimentales et prédites par MLR, MLP (ANN) et GFA.

R1	R2	R3	R4	R5	R6	pCI_{50} (µmol/l)			
						Exp.	Prédite MLR	Prédite GFA	Prédite MLP
H	H	H	H	H	propan-1-ol	1.000	1,496	1.557	1.341
H	H	H	H	H	ethylpropionate	1.523	1,088	1.281	1.587
H	H	H	H	H	ethyl acrylate	1.000	0,899	0.991	1.183
H	H	H	H	H	ethanol	2.097	1,835	1.703	1.977
H	H	H	H	H	methylacetate	1.602	1,802	1.770	1.841
H	H	H	H	H	8-(ethoxymethyl)-1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione	3.000	2,632	2.901	2.928
H	H	H	H	H	8-ethyl-1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione	2.602	2,399	2.480	2.498
H	H	H	H	H	1,3-dimethyl-7-propyl-3,7-dihydro-1H-purine-2,6-dione	2.699	2,628	2.720	2.620
H	H	H	H	H	9-ethyl-6-(methylthio)-9H-purine	1.456	1,506	1.619	1.555
H	H	H	H	H	9-ethyl-6-methoxy-9H-purine	1.377	1,414	1.464	1.367
H	H	H	H	H	7-ethyl-1-methyl-3,7-dihydro-1H-purine-2,6-dione	1.699	1,796	1.822	1.810
H	H	H	H	H	3-phenyl-6-propoxypyridazine	1.523	1,563	1.281	1.587
H	H	H	H	H	N-propyl-6,7-dihydro-5H-benzo[6,7]cyclohepta[1,2-c]pyridazin-3-amine	3.921	4,189	3.967	4.051
H	H	H	H	H	N-propylbenzo[h]cinnolin-3-amine	3.721	4,678	4.343	3.984
H	H	H	H	H	N-propyl-5,6-dihydrobenzo[h]cinnolin-3-amine	3.721	4,356	4.080	4.076
H	H	H	H	H	N-propyl-5H-indeno[1,2-c]pyridazin-3-amine	4.131	4,681	4.362	4.097
H	H	H	H	H	5-ethyl-6-phenyl-N-propylpyridazin-3-amine	4.180	3,742	3.471	4.202
H	H	H	H	H	5-methyl-6-phenyl-N-propylpyridazin-3-amine	4.409	4,035	3.729	4.104
H	H	H	H	H	6-phenyl-N-propylpyridazin-3-amine	3.854	3,762	3.490	3.675
H	H	H	H	H	(E)-3-(3,4-dihydroxyphenyl)-N-propylacrylamide	1.000	1,165	0.991	1.183

Tableau III.8 (suite)

R1	R2	R3	R4	R5	R6	pCI_{50} (µmol/l) Exp.	Prédite MLR	Prédite GFA	Prédite MLP
H	H	H	H	H	(E)-3-(2,4-dihydroxyphenyl)-N-propylacrylamide	3.004	2,814	2.600	2.727
H	H	H	H	H	(E)-3-(4-hydroxyphenyl)-N-propylacrylamide	2.097	1,901	1.703	1.977
H	H	H	H	H	(E)-3-(3-hydroxyphenyl)-N-propylacrylamide	1.602	1,795	1.770	1.841
H	H	H	H	H	(E)-3-(4-hydroxy-3-methoxyphenyl)-N-propylacrylamide	3.409	3,23	3.125	3.018
H	H	H	H	H	(E)-3-(2,4-dimethoxyphenyl)-N-propylacrylamide	2.602	2,51	2.480	2.498
H	H	H	H	H	(E)-3-(4-methoxyphenyl)-N-propylacrylamide	2.699	2,78	2.720	2.620
H	H	H	H	H	(E)-3-(3-methoxyphenyl)-N-propylacrylamide	3.125	2,744	2.720	2.619
H	H	H	H	H	(E)-3-(3,4-dimethoxyphenyl)-Npropylacrylamide	3.854	3,762	3.490	3.675
H	H	H	H	H	2-ethyl-5,6-dimethoxy-2,3-dihydro-1H-inden-1-one	5.000	4,621	4.459	4.415
H	H	H	H	H	4,5-bis(4-chlorophenyl)-N-methylthiazol-2-amine	2.860	2,824	2.853	2.756
H	H	H	H	H	N-methyl-4,5-di-p-tolylthiazol-2-amine	2.893	2,834	2.927	2.878
H	H	H	H	H	4-methyl-7-((propylamino)methyl)-2H-chromen-2-one	4.174	3,891	3.826	3.702
H	H	H	H	H	3-(2-(methylamino)ethoxy)-7,8,9,10-tetrahydro-6H-benzo[c]chromen-6-one	2.863	3,138	3.105	3.067
H	H	H	H	H	3-chloro-4-methyl-7-(2-(methylamino)ethoxy)-2H-chromen-2-one	2.740	3,049	2.955	2.851
H	H	H	H	H	3,4-dimethyl-7-(2-(methylamino)ethoxy)-2H-chromen-2-one	2.772	2,97	2.886	2.853
H	H	H	H	H	4-chloro-7-(2-(methylamino)ethoxy)-2H-chromen-2-one	2.690	3,295	3.139	3.044
H	H	H	H	H	3-(4-hydroxy-3-methoxyphenyl)-N-propylpropanamide	2.097	1,899	1.703	1.977

Tableau III.8 (suite)

R1	R2	R3	R4	R5	R6	pCI_{50} (µmol/l) Exp.	Prédite MLR	Prédite GFA	Prédite MLP
H	H	H	H	H	(E)-3-(4-nitrophenyl)-N-propylacrylamide	3.472	3,151	3.097	3.102
H	H	H	H	H	(E)-3-(4-chlorophenyl)-N-propylacrylamide	2.863	3,138	3.105	3.067
H	H	H	H	H	(E)-N-propyl-3-(3,4,5-trimethoxyphenyl)acrylamide	2.772	2,97	2.886	2.853
H	H	H	H	H	(E)-3-(4-methoxyphenyl)-N-propylacrylamide	3.161	3,106	2.778	2.672
H	H	H	H	H	(E)-3-(3,4-dimethoxyphenyl)-N-propylacrylamide	3.279	3,61	3.517	3.556
H	H	H	H	H	(E)-2-methoxy-4-(3-oxo-3-(propylamino)prop-1-en-1-yl)phenylacetate	3.470	3,333	3.315	3.276
H	H	H	H	H	N-propylcinnamamide	3.386	3,166	3.169	3.167
H	H	H	H	H	(E)-3-(2-hydroxyphenyl)-N-propylacrylamide	2.572	3,391	3.270	3.140
H	H	H	H	H	(E)-3-(3,4-dihydroxyphenyl)-N-propylacrylamide	2.690	3,295	3.139	3.044
H	H	H	H	H	(E)-3-(4-hydroxyphenyl)-N-propylacrylamide	2.658	3,246	3.172	3.070
H	H	H	H	H	(E)-3-(4-hydroxy-3,5-dimethoxyphenyl)-N-propylacrylamide	3.493	3,738	3.625	3.711
H	H	H	H	H	(E)-3-(3-hydroxy-4-methoxyphenyl)-N-propylacrylamide	3.000	2,934	2.634	2.542
H	H	H	H	H	(E)-3-(4-hydroxy-3-methoxyphenyl)-N-propylacrylamide	3.000	2,674	2.901	2.928
R1	**R2**	**R3**	**R4**	**R5**	**R6**				
H	H	H	H	CN	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	4.260	4,911	4.773	4.835
H	H	H	H	Ome	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	4.194	4,386	4.319	4.424
H	H	H	H	Cl	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	5.292	4,966	4.844	4.928

Tableau III.8 (suite)

R1	R2	R3	R4	R5	R6	pCI_{50} (µmol/l) Exp.	Prédite MLR	Prédite GFA	Prédite MLP
R1	**R2**	**R3**	**R5**	**R4**	**R6**				
H	H	H	H	Me	8-(ethoxymethyl)-1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione	3.000	2,498	2.634	2.542
H	H	H	H	CN	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	5.167	4,855	4.733	4.817
H	H	H	H	NO2	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	5.538	5,116	4.999	5.036
H	H	H	H	OH	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	5.060	4,878	4.754	4.868
H	H	H	H	Ome	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	3.903	4,63	4.548	4.640
H	H	H	H	Cl	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	5.310	4,801	4.708	4.826
R1	**R2**	**R4**	**R5**	**R3**	**R6**				
H	H	H	H	CN	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	4.495	4,897	4.758	4.787
H	H	H	H	NO2	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	4.367	5,109	4.982	4.978
H	H	H	H	F	8-butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	5.337	4,924	4.804	4.972
H	H	H	H	CF3	4,5-bis(4-methoxyphenyl)-N-methylthiazol-2-amine	3.125	2,48	2.575	2.726
H	H	H	H	Ome	4,5-bis(4-methoxyphenyl)-N-methylthiazol-2-amine	3.114	3,197	3.082	3.142
H	H	H	H	Me	4,5-bis(4-methoxyphenyl)-N-methylthiazol-2-amine	3.081	2,833	2.804	2.667
H	H	H	H	CN	4,5-bis(4-chlorophenyl)-N-methylthiazol-2-amine	2.991	3,105	3.084	2.948
H	H	H	H	Ome	4,5-bis(4-chlorophenyl)-N-methylthiazol-2-amine	2.951	2,706	2.650	2.435
H	H	H	H	Me	4,5-bis(4-chlorophenyl)-N-methylthiazol-2-amine	2.496	2,753	2.817	2.706
H	H	H	H	Me	N-methyl-4,5-di-p-tolylthiazol-2-amine	2.963	2,823	2.894	2.914
H	H	H	H	F	N-ethyl-4,5-di-p-tolylthiazol-2-amine	3.432	3,121	3.590	3.627

Tableau III.8 (suite)

R1	R2	R3	R4	R5	R6	pCl_{50} (µmol/l) Exp.	Prédite MLR	Prédite GFA	Prédite MLP
H	H	H	H	F	N-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	2.936	3,212	3.093	3.054
H	H	H	H	F	N-(4,5-bis(4-chlorophenyl)thiazol-2-yl)acetamide	3.187	3,109	3.017	3.009
H	H	H	H	NO2	N-(4,5-di-p-tolylthiazol-2-yl) acetamide	2.635	2,942	2.917	2.830
H	H	H	H	CN	N-(4,5-di-p-tolylthiazol-2-yl) acetamide	2.893	2,843	2.815	2.778
R1	**R3**	**R5**	**R2**	**R4**	**R6**				
H	H	H	F	F	4,5-bis(4-chlorophenyl)-N-ethylthiazol-2-amine	3.201	3,186	3.190	3.124
H	H	H	Me	Me	N-ethyl-4,5-di-p-tolylthiazol-2-amine	3.523	3,047	3.105	3.038
H	H	H	F	F	N-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	2.611	3,166	3.049	3.099
H	H	H	F	H	N-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	3.244	3,218	3.096	3.046
H	H	H	F	F	N-(4,5-bis(4-chlorophenyl)thiazol-2-yl)acetamide	3.194	2,926	2.857	2.988
H	H	H	F	F	N-(4,5-di-p-tolylthiazol-2-yl) acetamide	2.602	2,77	2.750	2.791
R2	**R4**	**R1**	**R3**	**R5**	**R6**				
H	H	Cl	F	H	4,5-bis(4-chlorophenyl)-N-methylthiazol-2-amine	3.432	3,579	3.095	3.077
H	H	Cl	H	F	N-ethyl-4,5-bis(4-methoxyphenyl)thiazol-2-amine	3.444	3,72	3.623	3.555
H	H	CF3	H	H	N-ethyl-4,5-bis(4-methoxyphenyl)thiazol-2-amine	3.161	2,789	3.230	3.364
H	H	Me	H	H	N-ethyl-4,5-bis(4-methoxyphenyl)thiazol-2-amine	3.229	3,527	3.515	3.449
H	H	CF3	F	H	4,5-bis(4-chlorophenyl)-N-ethylthiazol-2-amine	2.907	3,233	3.313	3.134
H	H	Cl	F	H	4,5-bis(4-chlorophenyl)-N-ethylthiazol-2-amine	3.284	3,581	3.526	3.418
H	H	Me	H	H	4,5-bis(4-chlorophenyl)-N-ethylthiazol-2-amine	3.060	3,561	3.539	3.362

Tableau III.8 (suite)

R1	R2	R3	R4	R5	R6	pCI_{50} (µmol/l)			
						Exp.	Prédite MLR	Prédite GFA	Prédite MLP
H	H	CF3	H	H	N-ethyl-4,5-di-p-tolylthiazol-2-amine	3.398	2,721	2.971	3.192
H	H	Me	H	H	N-ethyl-4,5-di-p-tolylthiazol-2-amine	3.301	3,714	3.754	3.463
H	H	CF3	F	H	N-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	2.910	3,028	3.001	2.927
H	H	CN	H	H	N-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	2.553	3,111	3.002	2.850
H	H	Cl	H	F	N-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	3.387	3,276	3.129	3.109
H	H	CF3	H	H	N-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	3.167	2,788	2.823	2.765
H	H	Me	H	H	N-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	2.873	3,026	2.982	2.850
H	H	CN	H	H	N-(4,5-bis(4-chlorophenyl)thiazol-2-yl) acetamide	2.752	3,087	2.997	2.849
H	H	CF3	H	H	N-(4,5-bis(4-chlorophenyl)thiazol-2-yl) acetamide	2.558	2,626	2.705	2.496
H	H	Me	H	H	N-(4,5-bis(4-chlorophenyl)thiazol-2-yl) acetamide	2.959	2,91	2.900	2.772
H	H	Cl	H	F	N-(4,5-di-p-tolylthiazol-2-yl) acetamide	2.541	2,995	2.942	2.894
H	H	CF3	H	H	N-(4,5-di-p-tolylthiazol-2-yl) acetamide	2.121	2,468	2.614	2.592

III.4.3 Développement des modéles QSAR

L'ensemble de données a été classé à l'aide de la technique de regroupement K-means [48]. Les 99 dérivés de la N-benzylpiperidine ont été divisés en un ensemble d'apprentissage (soit 75% du nombre total de composés) et un ensemble de test (soit 25% du nombre total de composés).

Pour le développement de modèles QSAR, trois approches différentes ont été utilisées: la régression linéaire multiple (MLR) [49], la MLR avec approximation de la fonction génétique (GFA en anglais) [50] et le réseau perceptron multicouche (MLP). La MLR a été réalisée à l'aide du logiciel MLRplusValidation1.3, tandis que la GFA et la MLP ont été réalisées à l'aide du logiciel Accelrys Material Studio 8.0.

III.4.4 Validation des modéles QSAR

La validation des modéles QSAR de cette seconde partie a été menée de la même manière que celle explicitée au paragraphe III.1.4 de la première partie de notre étude.

III.4.5 Domaine d'applicabilité des modéles QSAR

Dans le cadre de cette partie, on a utilisé l'approche de standardisation initiée par Roy et al. [29]. Cette approche est une méthode simple pour définir les valeurs aberrantes des composés (dans le cas de l'ensemble d'apprentissage) et les composés résidant en dehors du domaine d'applicabilité (dans le cas de l'ensemble de test). Le logiciel est accessible à partir du lien suivant: http://dtclab.webs.com/softwaretools ou http://teqip.jdvu.ac.in/QSAR_Tools/. La théorie de fond, l'algorithme et la méthodologie et les avantages de l'approche proposée sont disponibles dans la littérature [29].

III.5 RESULTATS ET DISCUSSION

III.5.1 Calcul et sélection des descripteurs pertinents

Le nombre de descripteurs pertinents obtenus après la sélection effectuée est de 10 : X1Av, SIC2, MATS4v, MATS2e, GATS1m, GATS3e, VEA1, GCUT_SMR_3, SlogP_VSA2 et std_dim3. Ces descripteurs ont été utilisés pour le développement des modèles prédictifs. La signification de ces descripteurs, ainsi que les logiciels qui ont permis leur calcul sont donnés sur le Tableau III.9 ci-contre.

Tableau III.9 Descripteurs utilisés pour le développement des modèles QSAR

Catégorie	Descripteur	Logiciel	Description	Réf
Eigenvalue-based indices	VEA1	Dragon	somme des coefficients du vecteur propre de la matrice d'adjacence	[51]
Connectivity indices	X1Av	Dragon	connectivité de valence moyenne indice d'ordre 1	[52]
Information indices	SIC2	Dragon	Indice de contenu d'informations structurelles (symétrie de voisinage de 2 ordres)	[53]
2D autocorrelations	MATS4v	Dragon	Autocorrélation du décalage 4 du Moran pondéré par le volume de van der Waals	[54]
2D autocorrelations	MATS2e	Dragon	Autocorrélation du décalage 2 du Moran pondéré par l'électronégativité de Sanderson	[54]

Tableau III.9 (suite)

Catégorie	Descripteur	Logiciel	Description	Réf
2D autocorrelations	GATS1m	Dragon	Autocorrélation du décalage 1 du Geary pondéré par la masse	[55]
2D autocorrelations	GATS3e	Dragon	Autocorrélation du décalage 3 du Geary pondéré par l'électronégativité de Sanderson	[55]
Adjacency and distance matrix descriptors	GCUT_SMR_3	MOE	Les descripteurs GCUT utilisant la contribution atomique à la réfractivité molaire (en utilisant la méthode SMR de Wildman et Crippen) au lieu de la charge partielle.	[56]
Subdivided surface areas	SlogP_VSA2	MOE	Somme des surfaces de Van der Waals (Å2) des atomes i (vi) telle que *Li* est dans (-0.20, 0.00]. *Li* désigne la contribution à logP (o / w) pour l'atome i tel que calculé dans le descripteur SlogP	[56]
Surface Area, Volume and Shape Descriptors	std_dim3	MOE	Dimension standard 3: la racine carrée de la troisième plus grande valeur propre de la matrice de covariance des coordonnées atomiques. Une dimension standard est équivalente à l'écart type le long d'un axe de composant principal	[57]

Afin d'étudier la corrélation entre les 10 descripteurs sélectionnés, la matrice de corrélation a été établie. La valeur du coefficient de corrélation de chaque paire de descripteurs sélectionnés a été examinée (Tableau III.10).La plus grande valeur du coefficient de corrélation appartenant à la paire de descripteurs SlogP_VSA2/SIC2 étant égale à 0,455 (donc < 0.7), cela signifie que tous les descripteurs sélectionnés sont indépendants les uns des autres.

Pour évaluer la multi-colinéarité entre les dix descripteurs pertinents, leurs facteurs de variation d'inflation (*VIF*) ont été calculés. Si *VIF* est égal à 1, aucune inter-corrélation n'existe pour chaque variable; si *VIF* est situé dans l'intervalle de 1 à 5, la selection des descripteurs est acceptable et si *VIF* est supérieur à 10, le modèle associé aux descripteurs sélectionnés est instable et une nouvelle sélection est nécessaire [58]. D'après le Tableau III.11, tous les descripteurs ont des valeurs VIF inférieures à 2, ce qui indique que la selection effectuée est bonne et que les descripteurs obtenus ne sont pas corrélés entre eux. La valeur du facteur d'inflation de la variance est calculée par le logiciel XLSTAT.

Tableau III.10 : Matrice des coefficients de corrélation des descripteurs sélectionnés

	X1Av	SIC2	MATS4v	MATS2e	GATS1m	GATS3e	VEA1	GCUT_SMR_3	SlogP_VSA2	std_dim3
X1Av	**1.000**	-0.240	-0.277	-0.247	-0.270	-0.036	-0.145	-0.319	-0.260	-0.211
SIC2		**1.000**	0.276	-0.021	-0.135	-0.120	-0.374	-0.120	**0.455**	-0.207
MATS4v			**1.000**	0.015	0.093	0.336	-0.017	0.173	0.017	-0.118
MATS2e				**1.000**	0.264	0.122	0.117	0.051	-0.112	0.047
GATS1m					**1.000**	0.149	-0.092	-0.117	-0.169	0.207
GATS3e						**1.000**	-0.160	0.147	-0.299	0.201
VEA1							**1.000**	0.111	0.038	0.119
GCUT_SMR_3								**1.000**	-0.156	0.219
SlogP_VSA2									**1.000**	-0.124
std_dim3										**1.000**

Tableau III.11 Valeurs du coefficient VIF des descripteurs sélectionnés.

Descripteur	X1Av	SIC2	MATS4v	MATS2e	GATS1m
VIF	1.972	1.917	1.477	1.228	1.532

Descripteur	GATS3e	VEA1	GCUt_SMR_3	SlogP_VSA2	std_dim3
VIF	1.461	1.553	1.451	1.485	1.248

III.5.2 Modèle MLR

Le modèle MLR pour la prédiction de la concentration inhibitrice des dérivés de la N-benzylpipéridine, obtenu par utilisation de 99 composés, est un modèle linéaire dont l'équation est la suivante :

pCI_{50} = -2.189 (±4. 377) - 31.874 (±5.330) X1Av - 7.499 (±1.861) SIC2 - 2.975 (±0.329) MATS4v - 2.557 (±0.336) MATS2e - 1.439 (±0.302) GATS1m + 1.570 (±0.378) GATS3e - 0.969 (±0.143) VEA1 + 8.537 (±0.705) GCUT_SMR_3 + 0.007 (±0.004) SlogP_VSA2 - 0.891 (±0.274) std_dim3

(III.14)

$N_{training}$ = 74; R^2 = 0.882, R^2_{adj}= 0.864, F = 47.213, p < 0.0001, Q^2 = 0.831, PRESS = 7.297
N_{test} = 25, R^2 = 0.746,$r_m^2 = 0.694$,r^2_{pred} = 0.561 ${}^{c}R_p^2$= 0.807

Les erreurs-types (*standard error*) des coefficients de régression sont données entre parenthèses. Les dix descripteurs pertinents (variables) dans l'équation (III.14) pourraient

expliquer 86,4% de la variance (coefficient de variation ajusté) de la concentration inhibitrice.

Les valeurs prédites de pCI_{50} pour les dérivés de la N-benzylpipéridine dans les ensembles d'apprentissage et de test, reportées sur le Tableau III.8, ont été tracées en fonction de leurs valeurs expérimentales sur la figure III.8. Une corrélation étroite entre les valeurs prédites par le modèle MLP-ANN et les valeurs de toxicité observées ont été trouvées.

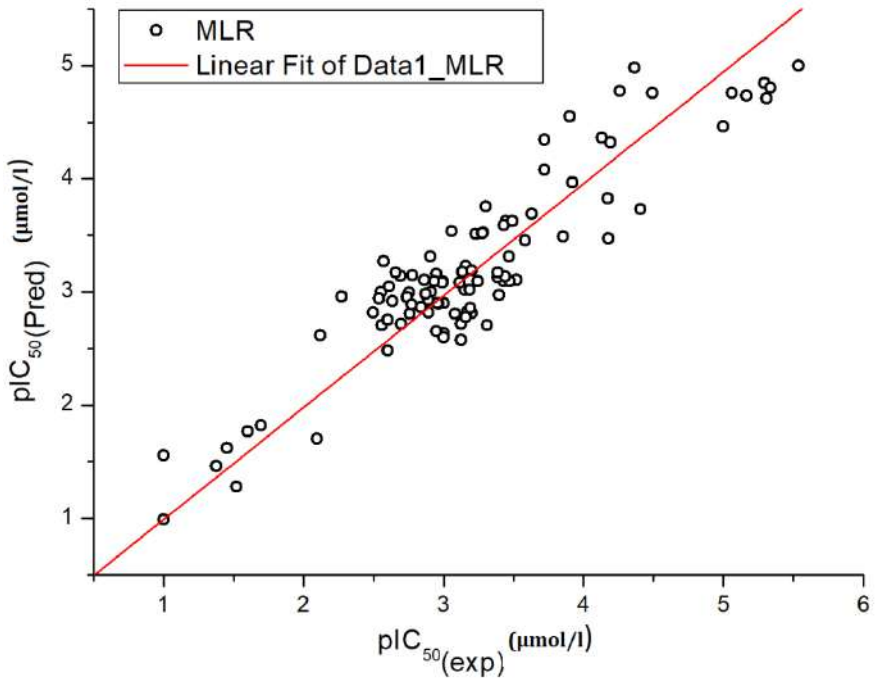


Figure III.8 Courbe de corrélation entre les valeurs prédites de pCI_{50} en fonction des valeurs expérimentales pour les ensembles d'apprentissage et de test (Modèle MLR)

Comme on peut le voir sur le Tableau III.12 ci-dessous, le modèle linéaire MLR a donné de bons résultats pour les coefficients R^2 et R^2_{adj} et F. De plus, la valeur encourageante du coefficient Q^2_{appr} (0.831) suggère une prédictivité interne appréciable du modèle. En plus de cela, la robustesse du modèle a également été vérifiée par un test de randomisation et une valeur élevée du paramètre ${}^{c}R_p^2$ (0.807) suggère que le modèle est dépourvu de toute corrélation de hasard.Le sur-apprentissage a été également vérifié et ce en calculant la différence entre R^2 et Q^2_{appr} qui doit être <0,3. Dans le cadre de ce modèle, la différence de 0,051 implique l'absence du sur-apprentissage. Pour l'ensemble du test, toutes les valeurs des paramètres statistiques (à l'exception du R^2_{pred}) et les critères d'acceptabilité sont satisfaisants, ce qui prouve que ce modèle a un bon pouvoir prédictif.

Tableau III.12 Paramètres de validation interne et externe du modèle MLR

	Paramètres statistiques	**Valeurs trouvées**	**Critères de validité**
Validation interne			
	R^2	0.882	
	R^2_{adj}	0.864	
	F	47.213	
	Q^2_{appr}	0.831	>0.5
	$\overline{r^2_{m(LOO)}}$	0.762	>0.5
	$\Delta r^2_{m(LOO)}$	0.104	<0.2
Validation externe			
	R^2	0.746	>0.6
	r_0^2	0.694	
	k	0.922	0.85<k<1.15
	k'	1.067	0.85<k'<1.15
	$\lvert r_0^2 - r_0'^2 \rvert$	0.050	<0.3
	$\frac{r^2 - r_0^2}{r^2}$	0.070	<0.1
	$\frac{r^2 - r_0'^2}{r^2}$	0.003	<0.1
	R^2_{pred}	0.561	>0.5

Pour distinguer l'importance de chaque descripteur sur le pouvoir inhibiteur des dérivés de la N-benzylpipéridine, les valeurs du t-Test des descripteurs impliqués sont également énumérés dans le Tableau III.13. Comme le montre ce tableau, la valeur du t-Test du descripteur GcUt_SmR_3 est supérieure à celle des autres descripteurs. Cela suggére que l'influence de ce descripteur sur le pouvoir inhibiteur des composés étudiés est la plus forte.

Tableau III.13 Valeurs du paramètre t-Test des descripteurs sélectionnés.

Descripteur	X1Av	SIC2	MATS4v	MATS2e	GATS1m	GATS3e
t-Test	−5.980	−4.030	−9.039	−7.610	−4.770	4.151

Descripteur	VEA1	GCUt_SMR_3	SlogP_VSA2	std_dim3
t-Test	−6.770	12.117	1.740	−3.248

Les coefficients de régression des signes suggèrent la direction de l'influence des descripteurs dans un modèle donné. D'après les coefficients de régression de l'équation III.14, il convient de noter que les descripteurs X1Av, SIC2 et GCUt_SMR_3 sont les principaux descripteurs qui contribuent à l'inhibition. Les coefficients de régression des descripteurs GATS3e, GCUt_SMR_3 et SlogP_VSA2 ont des signes positifs. Ces descripteurs ont donc une influence positive et les valeurs les plus élevées contribuent à améliorer l'inhibition des dérivés de la N-benzylpipéridine. De plus, les coefficients de régression des descripteurs X1Av, SIC2, MATS4v, MATS2e, GATS1m, VEA1 et std_dim3 présentent des signes négatifs, ce qui implique qu'ils ont un impact négatif sur l'inhibition des dérivés de la N-benzylpipéridine. Pour un dérivé donné, de petites valeurs de ces descripteurs aideraient à améliorer son pouvoir inhibiteur.

En interprétant les descripteurs du modèle MLR, il est possible d'avoir une idée des facteurs liés à l'activité inhibitrice. Dans les dix descripteurs sélectionnés, on dénombre quatre descripteurs d'autocorrélation 2D, un descripteur d'indices de connectivité, un descripteur d'indices d'information, un descripteur d'indices à valeur propre, un descripteur de matrice d'adjacence et de distance et enfin un descripteur de surfaces subdivisées. On peut donc en conclure que l'activité inhibitrice est principalement influencée par trois facteurs importants: la réfractivité molaire (GCUT_SMR_3), la structure de la molécule (X1Av, SIC2, MATS4v) et les propriétés électroniques (MATS2e).

III.5.3 Modèle MLP

La modélisation non linéaire au moyen d'un perceptron multicouche (réseau neuronal artificiel) a été également établie. La fonction sigmoïde a été utilisée comme une fonction de transfert de la couche cachée et une fonction d'identité pour la couche de sortie. Un neurone de sortie a été utilisé pour représenter la pCI_{50} expérimentale. L'algorithme d'apprentissage BFGS a été utilisé pour le développement du réseau.

Pour optimiser le nombre de neurones dans la couche cachée, plusieurs calculs ont été effectués avec différents nombres de nœuds cachés (de 1 à 10). Les 99 dérivés de la N-benzylpipéridine ont été répartis en deux ensembles: 75 pour l'apprentissage et 24 pour le

test. Ce réseau se composait de 10 entrées (descripteurs pertinents sélectionnés) et d'une sortie pour pCI_{50}.

Les critères RMSE et Q^2 sont utilisés pour évaluer la précision des ensembles d'apprentissage et de test. Le meilleur modèle a été choisi pour une RMSE minimum et un coefficient Q^2 maximum. Sur le Tableau III.14 sont reportés les résultats de l'optimisation du nombre de neurones.

Tableau III.14 Optimisation du nombre de neurones

Nombre de neurones	R^2 (train)	Q^2 (train)	R^2 (test)	Q^2 (test)	RMSE
1	0.842	0.756	0.785	0.795	0.395
2	0.841	0.725	0.698	0.802	0.397
3	0.844	0.757	0.778	0.800	0.393
4	0.815	0.693	0.829	0.809	0.428
5	0.835	0.768	0.830	0.796	0.404
6	**0.867**	**0.737**	**0.899**	**0.805**	**0.363**
7	0.854	0.756	0.796	0.810	0.380
8	0.863	0.694	0.890	0.794	0.368
9	0.848	0.747	0.784	0.806	0.388
10	0.813	0.742	0.844	0.817	0.430

Les résultats obtenus montrent que le réseau avec six neurones est celui ayant donné les plus grandes valeurs de R^2 et Q^2, ainsi que la plus petite valeur pour la RMSE.

Le modèle MLP à 6 neurones a donné de bons résultats avec des coefficients de corrélation élevés R^2 (0,867 et 0,899), ainsi qu'une meilleure robustesse (Q^2 = 0,737 et 0,805) respectivement pour l'ensemble d'apprentissage et de test. Par conséquent, un MLP avec l'architecture 10-6-1 a été choisi comme modèle final dans ce travail.

Les concentrations inhibitrices prédites par le modèle MLP pour l'ensemble des données (99 composés) ont été obtenus; elles sont rassemblées dans le tableau III.8. Sur la figure III.9 sont représentées les concentrations inhibitrices prédites en fonction des concentrations inhibitrices expérimentales pour l'ensemble d'apprentissage et de test. D'après la figure III.9, une satisfaisante corrélation entre les valeurs prédites par le modèle MLP et les valeurs expérimentale a été constatée.

Dans le cadre de cette étude, nous avons utilisé l'approche de Roy et al. [23] qui consiste à utiliser les paramètres statistiques de l'ensemble global (overall set) au lieu des paramètres statistiques pour chacun des deux ensembles (apprentissage et test) au cas où la base de données n'est pas assez grande. Les avantages d'une telle approche sont de faciliter l'obtention du meilleur modèle sur la base d'une contribution globale des mesures de validation internes et externes.

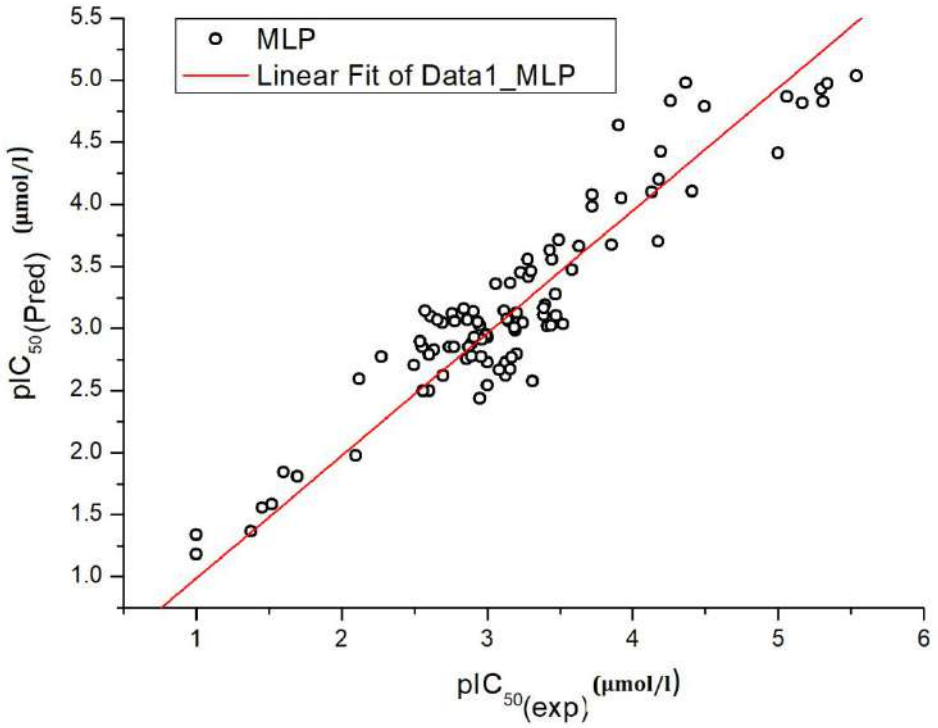


Figure III.9 Courbe de corrélation entre les valeurs prédites de pCI_{50} en fonction des valeurs expérimentales pour les ensembles d'apprentissage et de test (Modèle MLP)

Comme on peut le voir dans le tableau III.15 ci-dessous, le modèle non linéaire des réseaux de neurones (modèle MLP) a donné de bons résultats avec les coefficients $Q^2_{(overall)}$ et $r^2_{m(overall)}$, d'où une bonne robustesse du modèle. De plus, la valeur de $\overline{r^2_{m(overall)}}$ est supérieure à 0.5. Par ailleurs, la valeur de $\Delta r^2_{m(overall)}$ est inférieure à 0,2 [23].

Selon la recommandation de Tropsha et Golbraikh [24], la différence entre R^2 et Q^2 qui est de 0.132 ne dépasse pas 0.3, aussi il y'a lieu de conclure qu'il n'y'a pas eu sur-apprentissage lors de l'élaboration du modèle MLP.

Une analyse de sensibilité utilisant la méthode proposée par Goh [40], a été également réalisée pour déterminer l'importance de chaque variable dans la prédiction de l'activité inhibitrice des dérivés de la N-benzylpipéridine. La contribution de chacun des descripteurs dans le modèle MLP obtenu est reproduite sur la figure III.10 ci-après.

En conclusion, le modèle MLP non linéaire a donné de bons résultats. Les paramètres statistiques obtenus pour l'ensemble global satisfont évidemment aux critères d'acceptabilité et démontrent ainsi le pouvoir prédictif du modèle développé. Ces résultats indiquent que le modèle MLP a non seulement bien fonctionné dans le développement du modèle, mais possède également un excellent pouvoir de prédiction. Ceci laisse supposer

l'existence d'une corrélation non linéaire entre la concentration inhibitrice et les descripteurs pertinents sélectionnés.

Tableau III.15 Paramètres statistiques du modèle MLP

Paramètres statistiques	Modèle MLP
$Q^2_{(overall)}$	0.737
$r^2_{m(overall)}$	0.869
$\overline{r^2_{m(overall)}}$	0.819
$\Delta r^2_{m(overall)}$	0.100

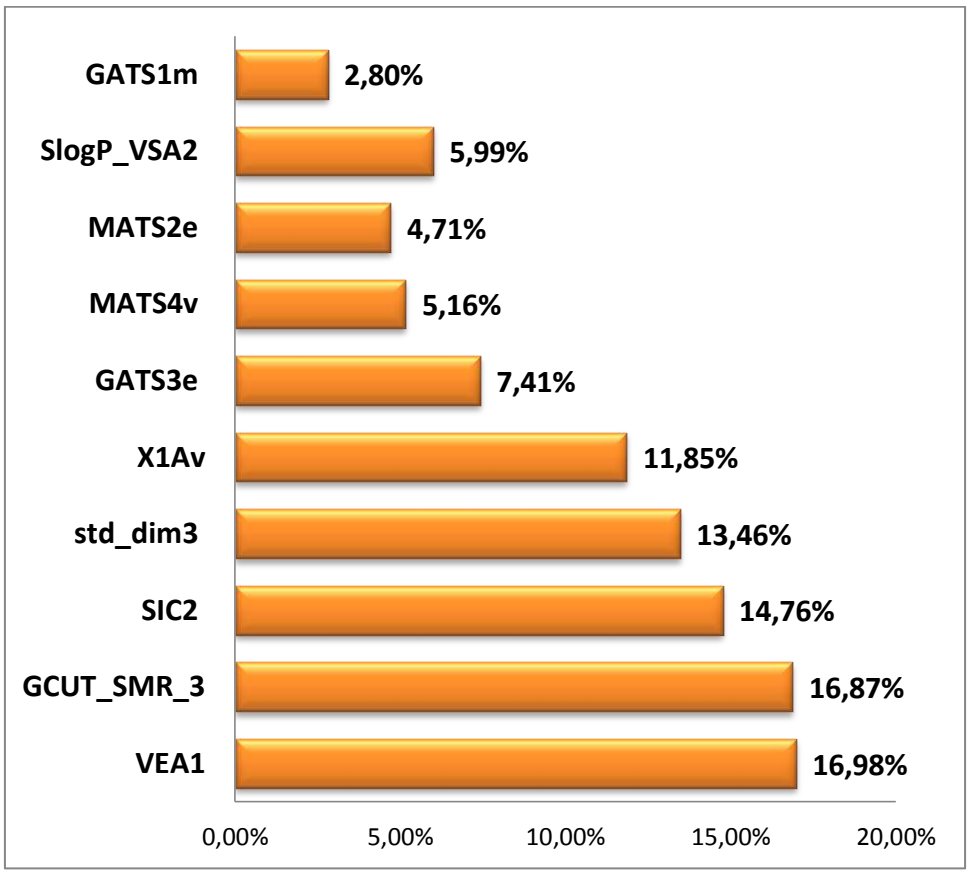


Figure III.10 Contribution à l'effet inhibiteur par type de descripteurs

III.5.4 Modèle GFA

Le modèle GFA pour la prédiction de la concentration inhibitrice des dérivés de la N-benzylpipéridine, obtenu par utilisation de 99 composés, est un modèle dont l'équation qui a donné les meilleurs résultats est la suivante :

pCI_{50} = - 26,034 X1Av - 7,013SIC2 - 2,588 MATS4v - 1,968 MATS2 - 1,289 GATS1m + 1,258 GATS3e - 0,900 VEA1 + 8,784 GCUT_SMR_3 + 0,007 SlogP_VSA2 - 1,026 std_dim3 - 5,279. **(III.15)**

Friedman LOF = 0,579, R^2 = 0,875, F = 61,442, Q^2 = 0,842

Selon l'équation (III.15), les dix descripteurs pertinents pourraient expliquer 87,5% de la variance (coefficient de variation ajusté) de la concentration inhibitrice. La différence entre R^2 et Q^2 est égale à 0,033. Cette différence étant inférieure à 0,3, elle signifie l'absence de sur-apprentissage. La significativité de la régression est donnée par le test F, plus sa valeur est élevée, meilleur est le modèle. Dans le cas de ce modèle, la valeur de F = 61.442, ce qui signifie que la régression est significative. Le score de Friedman (LOF) évalue le modèle QSAR. Plus le LOF est faible, moins il est probable que le modèle GFA corresponde aux données. D'après les coefficients de régression de l'équation III.15, il est important de noter qu'ils sont très proches de ceux obtenus pour le modèle MLR. On retrouve la même influence positive ou négative de ces descripteurs sur l'activité inhibitrice des dérivés de la N-benzylpipéridine.

Les résultats prédictifs du modèle GFA pour l'ensemble de données complet (99 composés) sont reportés sur le Tableau III.8. Sur la figure III.11 ci-dessous, sont représentées les concentrations inhibitrices prédites en fonction des concentrations inhibitrices expérimentales pour l'ensemble d'apprentissage et de test. D'après la figure III.11, une bonne corrélation entre les valeurs prédites par le modèle MLP et les valeurs expérimentale a été constatée.

Comme ce fut le cas pour le modèle MLP, la validation du modèle GFA a été réalisée par étude des coefficients statistiques de l'ensemble de donnée dans sa globalité (apprentissage + test). Les valeurs de ces coefficients sont reportées sur le Tableau III.16 ci-dessous.

D'après le tableau III.16, on note que des résultats statistiquement significatifs pour tous les paramètres métriques ($Q^2{}_{(overall)}$, $r^2_{m(overall)}$ et $\overline{r^2_{m(overall)}}$ indiquent que les valeurs d'activité prédites sont proches des données expérimentales correspondantes. Ainsi, le modèle présente une capacité prédictive élevée. Par ailleurs, la valeur de $\Delta r^2_{m(overall)}$ est inférieure à 0,2 [23]. Par conséquent, ces résultats indiquaient que le modèle GFA a non seulement bien fonctionné lors de son développement, mais possède

également une excellente prédiction, ce qui laisserait supposer l'existence d'une corrélation non linéaire entre la concentration inhibitrice et les descripteurs pertinents.

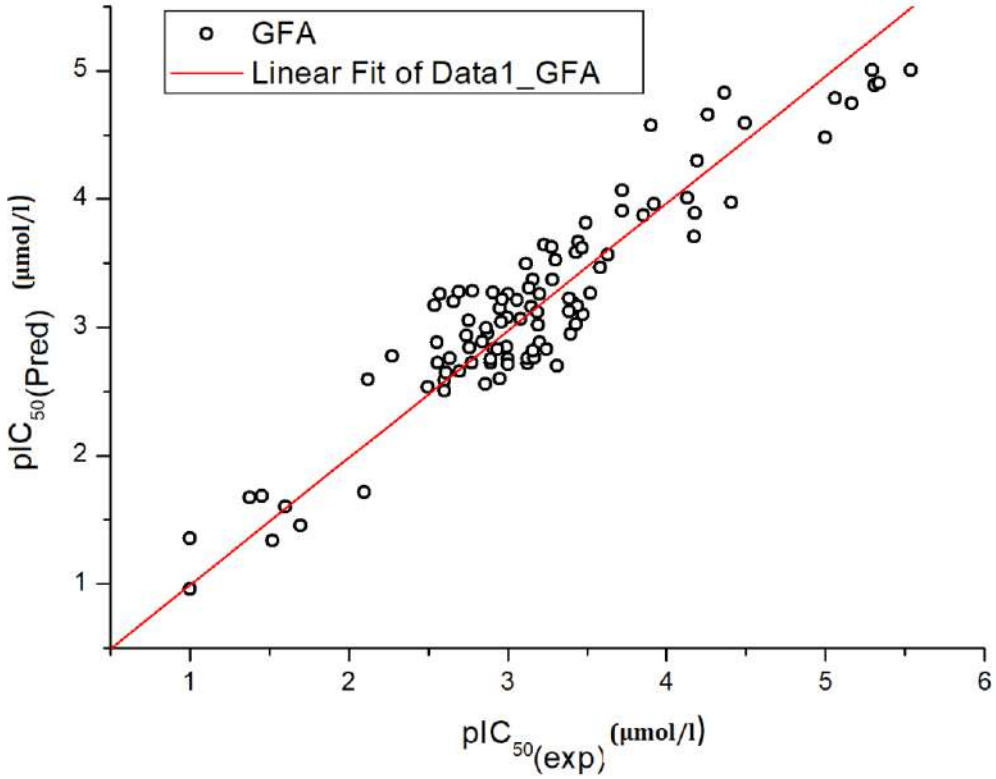


Figure III.11 Courbe de corrélation entre les valeurs prédites de pCI_{50} en fonction des valeurs expérimentales pour les ensembles d'apprentissage et de test (Modèle GFA)

Tableau III.16 Paramètres statistiques du modèle GFA

Paramètres statistiques	Modèle GFA
$Q^2_{(overall)}$	0.842
$r^2_{m(overall)}$	0.875
$\overline{r^2_{m(overall)}}$	0.818
$\Delta r^2_{m(overall)}$	0.113

III.5.4 Comparaison des trois modèles

Pour comparer statistiquement les trois modèles, nous avons utilisé l'approche de Roy et al. [23] inhérente à l'utilisation de la base globale**.** Les résultats sont présentés dans le Tableau III.17 ci-dessous. Sur la base du paramètre statistique $r^2_{m(overall)}$, les modèles GFA et MLP sont légèrement meilleurs que le modèle MLR. La différence entre les modèles GFA et MLP est insignifiante. De plus, les valeurs de $\Delta r^2_{m(overall)}$ et $\overline{r^2_{m(overall)}}$ indiquent que le modèle GFA est légèrement meilleur que le modèle MLP. Ainsi, on peut conclure que les deux modèles MLP et GFA ont un meilleur pouvoir prédictif que le modèle MLR.

Tableau III.17 : Paramètres statistiques des modèles développés

Paramètres statistiques	Modèle MLR	Modèle GFA	Modèle MLP
$Q^2_{(overall)}$	0.831	**0.842**	0.737
$r^2_{m(overall)}$	0.843	**0.875**	0.869
$\overline{r^2_{m(overall)}}$	0.773	**0.818**	**0.819**
$\Delta r^2_{m(overall)}$	0.139	0.113	**0.100**

III.5.5 Domaine d'applicabilité

Le domaine d'applicabilité des trois modèles a été analysé en utilisant une approche de standardisation. Le logiciel est accessible depuis http://teqip.jdvu.ac.in/QSAR_Tools/. Deux composés (un pour l'apprentissage et un pour le test) ont été identifiés comme étant en dehors du domaine d'applicabilité. Il convient de noter que 98% du domaine ont été couverts par les trois modèles lorsqu'ils ont été appliqués pour prédire l'activité inhibitrice des99 dérivés de N-benzylpipéridine. Ainsi, ces résultats montrent que les modèles sont conformes au troisième principe de l'OCDE. Ils peuvent être utilisés pour prédire l'activité inhibitrice des dérivés de la N-benzylpipéridine, en particulier pour ceux qui n'ont pas été testés ainsi que de nouveaux composés.

III.6 COMPARAISON AVEC LES MODELES DE LA LITTERATURE

Comme ce fut le cas pour la première partie de notre travail, nous avons procédé à une comparaison entre les quelques modèles rapportés dans la littérature avec celui qui a été développé dans le cas de l'inhibition de l'enzyme AChE (tableau III.18). Ces modéles sont consacrés à l'inhibition de l'enzyme AChE au moyen de divers composés (dérivés du 1-

benzyl-4-[2-(N-benzoylamino) ethyl] piperidine et N-benzylpiperidine benzisoxazoles, dérivés de la penténone, analogues de la tacrine, organophosphorés, etc.). Comme rapporté au paragraphe III.3, l'évaluation de leurs avantages et inconvénients est assez difficile, car chaque étude publiée utilisait différents ensembles de données et une approche de modélisation différente (descripteurs chimiques, méthodes d'analyse de données, algorithmes, etc.).

Néanmoins, il convient de noter que la plupart des modèles QSAR rapportés dans la littérature ont été obtenus avec des bases de données assez petites à l'exception de celle utilisée par Fernandez et al. [61]. Sur le tableau III.18, il est possible d'observer que, contrairement à notre modèle, aucune approche de la validation externe n'a été réalisée dans les autres modèles, à l'exception du modèle de Jang et al. [66] qui ont utilisé un seul paramètre (Q^2_{ext}). Ainsi, la comparaison était limitée aux résultats obtenus pour les statistiques de la validation interne. Là encore, le nombre de paramètres statistiques utilisés pour la validation interne de ces modèles QSAR est limité en comparaison de ceux utilisés pour notre modèle. Il est possible d'observer que tous ces modèles ont donné des valeurs élevées du coefficient de corrélation R^2 et des valeurs satisfaisantes pour Q^2_{LOO}. De plus, le domaine d'applicabilité des modèles de la littérature n'ont pas été établis. Selon ces résultats, le modèle établi par non soins est plus performant étant donné qu'il satisfait aux directives de l'OCDE. De ce fait, il peut contribuer ainsi à la conception de médicaments anti-Alzheimer, tout en réalisant des économies substantielles d'argent et de temps.

III.7 CONCLUSION

Dans cette seconde étude, des modèles QSAR linéaires et non linéaires ont été utilisés pour prédire l'activité inhibitrice des dérivés de 99 N-benzylpipéridine en fonction de leur structure moléculaire, représentée par 10 descripteurs pertinents. Les données expérimentales ont été sélectionnées à partir de la littérature. Les trois modèles développés ont montré une bonne capacité de prédiction. Pour sélectionner le meilleur modèle, le paramètre r^2_m de la base globale a été utilisé. Selon certains auteurs [60], ce paramètre semble être avantageux par rapport à d'autres paramètres de validation internes et externes dans le cas où la base de travail ne contient pas un grand nombre de composés. Les modèles développés dans ce travail montrent l'importance de la réfractivité molaire, de la structure des molécules et des propriétés électroniques dans le cas de l'activité inhibitrice des dérivés de la N-benzylpipéridine. Les modèles développés peuvent être utiles pour concevoir de puissants inhibiteurs de l'AChE et prédire le potentiel d'inhibition de l'AChE pour de nouveaux candidats médicaments.

Tableau III.18 Comparaison entre les modéles de la littérature et celui développé dans ce travail (enzyme AChE)

Modèles	**Validation interne**						**Validation externe**					
	n	**R^2**	**RMS/SEE**	Q^2_{Loo}	$\overline{r^2_m}$	Δr^2_m	**n**	**R^2_{ext}**	**RMS/SEE**	Q^2_{ext}	$\overline{r^2_m}$	Δr^2_m
Tong et al. [59]	57	0.94	0.32	0.67			20	0.89				
Bernard et al. [60]	82	0.98	0.19	0.75								
Fernandez et al. [61]	108	0.92					28	0.85				
Jung et al. [62]	68	0.96	0.39	0.95			12	0.84				
Liu et al. [63]	30	0.97	0.33	0.63								
Zhou et al. [64]	48	0.99	0.07	0.58			12					
Zhang et al. [65]	34	0.98	0.22	0.76			6	0.98				
Pang et al. [46]	35	0.92					10	0.85				
Jang et al. [66]	56	0.73		0.67			26		0.69	0.67		
Présent travail	99(app + test)	R^2 = **0,875** ; $Q^2_{(overall)}$ = **0.842** ; $r^2_{m(overall)}$ = **0.875** ; $\Delta r^2_{m(overall)}$ = **0.113** ; $\overline{r^2_{m(overall)}}$ = **0.818** ;										

Références

1. Boulebd, H., et al., *New (benz) imidazolopyridino tacrines as nonhepatotoxic, cholinesterase inhibitors for Alzheimer disease.* Future, 2017. 9(8): p. 723-729.
2. Arendt, T., et al., *Changes in acetylcholinesterase and butyrylcholinesterase in Alzheimer's disease resemble embryonic development—a study of molecular forms.* Neurochemistry international, 1992. 21(3): p. 381-396.
3. Estrada, M., et al., *New cinnamic–N-benzylpiperidine and cinnamic–N, N-dibenzyl (N-methyl) amine hybrids as Alzheimer-directed multitarget drugs with antioxidant, cholinergic, neuroprotective and neurogenic properties.* European journal of medicinal chemistry, 2016. 121: p. 376-386.
4. Kostochka, M.L., et al., *Novel Tandem Aldol Intramolecular Cyclization of Substituted N-Benzylpiperidine-4-one: Synthesis of Novel-Type Nitrogen 2, 8-Phenanthroline Heterocycles.* Journal of Heterocyclic Chemistry, 2015. 52(6): p. 1723-1730.
5. Martinez, A., et al., *N-Benzylpiperidine derivatives of 1, 2, 4-thiadiazolidinone as new acetylcholinesterase inhibitors.* European journal of medicinal chemistry, 2000. 35(10): p. 913-922.
6. Rodríguez-Franco, M.I., et al., *Design and synthesis of N-benzylpiperidine–purine derivatives as new dual inhibitors of acetyl-and butyrylcholinesterase.* Bioorganic & medicinal chemistry, 2005. 13(24): p. 6795-6802.
7. Sukumarapillai, D.K., et al., *Design, synthesis and cholinesterase inhibitory evaluation study of fluorescent N-benzylpiperidine-4-one derivatives.* Medicinal Chemistry Research, 2016. 25(8): p. 1705-1715.
8. Więckowska, A., et al., *Synthesis of new N-benzylpiperidine derivatives as cholinesterase inhibitors with β-amyloid anti-aggregation properties and beneficial effects on memory in vivo.* Bioorganic & medicinal chemistry, 2015. 23(10): p. 2445-2457.
9. Contreras, J.-M., et al., *Design, synthesis, and structure− activity relationships of a series of 3-[2-(1-benzylpiperidin-4-yl) ethylamino] pyridazine derivatives as acetylcholinesterase inhibitors.* Journal of medicinal chemistry, 2001. 44(17): p. 2707-2718.
10. Ishichi, Y., et al., *Novel acetylcholinesterase inhibitor as increasing agent on rhythmic bladder contractions: SAR of 8-{3-[1-(3-fluorobenzyl) piperidin-4-yl] propanoyl}-1, 2, 5, 6-tetrahydro-4H-pyrrolo [3, 2, 1-ij] quinolin-4-one (TAK-802) and related compounds.* Bioorganic & medicinal chemistry, 2005. 13(6): p. 1901-1911.
11. Shidore, M., et al., *Benzylpiperidine-linked diarylthiazoles as potential anti-alzheimer's agents: Synthesis and biological evaluation.* Journal of medicinal chemistry, 2016. 59(12): p. 5823-5846.
12. Xie, S.-S., et al., *Design, synthesis and biological evaluation of novel donepezil–coumarin hybrids as multi-target agents for the treatment of Alzheimer's disease.* Bioorganic & medicinal chemistry, 2016. 24(7): p. 1528-1539.
13. Xu, W., et al., *Synthesis and evaluation of donepezil–ferulic acid hybrids as multi-target-directed ligands against Alzheimer's disease.* MedChemComm, 2016. 7(5): p. 990-998.

14. Ellman, G.L., et al., *A new and rapid colorimetric determination of acetylcholinesterase activity.* Biochemical pharmacology, 1961. 7(2): p. 88IN191-9095.
15. Liu, T., et al., *BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities.* Nucleic acids research, 2006. 35(suppl_1): p. D198-D201.
16. Mayers, D.L., et al., *Antimicrobial Drug Resistance: Clinical and Epidemiological Aspects*. 2017: Springer International Publishing. 1627.
17. Environment, M.O., *Chemical Computing Group*. 2007, Montreal Canada.
18. Todeschini, R. and V. Consonni, *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*. Vol. 41. 2009: John Wiley & Sons.
19. Hamadache, M., et al., *Prediction of acute herbicide toxicity in rats from quantitative structure–activity relationship modeling.* Environmental Engineering Science, 2014. 31(5): p. 243-252.
20. Confais, J. and M. Le Guen, *Premiers pas en régression linéaire avec SAS®.* 2007.
21. Wang, R., et al., *Prediction of impact sensitivity of nitro energetic compounds by neural network based on electrotopological-state indices.* Journal of hazardous materials, 2009. 166(1): p. 155-186.
22. Gramatica, P., *Principles of QSAR models validation: internal and external.* Molecular Informatics, 2007. 26(5): p. 694-701.
23. Roy, K., et al., *Comparative studies on some metrics for external validation of QSPR models.* Journal of chemical information and modeling, 2012. 52(2): p. 396-408.
24. Golbraikh, A. and A. Tropsha, *Beware of q2!* Journal of molecular graphics and modelling, 2002. 20(4): p. 269-276.
25. Chirico, N. and P. Gramatica, *Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient.* Journal of chemical information and modeling, 2011. 51(9): p. 2320-2335.
26. Lawrence, I. and K. Lin, *A concordance correlation coefficient to evaluate reproducibility.* Biometrics, 1989: p. 255-268.
27. Ghamali, M., et al., *Méthodologie générale d'une étude RQSA/RQSP.*
28. Tetko, I.V., et al., *Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection.* Journal of chemical information and modeling, 2008. 48(9): p. 1733-1746.
29. Roy, K., S. Kar, and P. Ambure, *On a simple approach for determining applicability domain of QSAR models.* Chemometrics and Intelligent Laboratory Systems, 2015. 145: p. 22-29.
30. Sahigara, F., et al., *Comparison of different approaches to define the applicability domain of QSAR models.* Molecules, 2012. 17(5): p. 4791-4810.
31. Sahigara, F., et al., *Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions.* Journal of cheminformatics, 2013. 5(1): p. 27.

32. Tropsha, A., P. Gramatica, and V.K. Gombar, *The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models.* Molecular Informatics, 2003. 22(1): p. 69-77.
33. de Oliveira, D.B. and A.C. Gaudio, *BuildQSAR: a new computer program for QSAR analysis.* Molecular Informatics, 2000. 19(6): p. 599-601.
34. Labute, P., *A widely applicable set of descriptors.* Journal of Molecular Graphics and Modelling, 2000. 18(4-5): p. 464-477.
35. Eriksson, L., et al., *Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs.* Environmental health perspectives, 2003. 111(10): p. 1361.
36. Mitra, I., A. Saha, and K. Roy, *Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants.* Molecular Simulation, 2010. 36(13): p. 1067-1079.
37. Othman, F. and M. Naseri, *Reservoir inflow forecasting using artificial neural network.* International journal of physical sciences, 2011. 6(3): p. 434-440.
38. Lee, T.-L., *Back-propagation neural network for the prediction of the short-term storm surge in Taichung harbor, Taiwan.* Engineering Applications of Artificial Intelligence, 2008. 21(1): p. 63-72.
39. Sedki, A., D. Ouazar, and E. El Mazoudi, *Evolving neural network using real coded genetic algorithm for daily rainfall–runoff forecasting.* Expert Systems with Applications, 2009. 36(3): p. 4523-4527.
40. Goh, A., *Back-propagation neural networks for modeling complex systems.* Artificial Intelligence in Engineering, 1995. 9(3): p. 143-151.
41. Wang, W., et al., *Determination of the spread parameter in the Gaussian kernel for classification and regression.* Neurocomputing, 2003. 55(3): p. 643-663.
42. Cooperation, O.f.E. and Development, *Guidance Document on the Validation of (Quantitative) Structure–Activity Relationships [(Q) SAR] Models.* ENV/JM/MONO, 2007. 2: p. 1-154.
43. Castilho, M., R. C Guido, and A. Andricopulo, *Classical and hologram QSAR studies on a series of tacrine derivatives as butyrylcholinesterase inhibitors.* Letters in Drug Design & Discovery, 2007. **4**(2): p. 106-113.
44. Uddin, R., et al., *Receptor-based modeling and 3D-QSAR for a quantitative production of the butyrylcholinesterase inhibitors based on genetic algorithm.* Journal of chemical information and modeling, 2008. **48**(5): p. 1092-1103.
45. Fang, J., et al., *Molecular Modeling on Berberine Derivatives toward BuChE: An Integrated Study with Quantitative Structure–Activity Relationships Models, Molecular Docking, and Molecular Dynamics Simulations.* Chemical biology & drug design, 2016. **87**(5): p. 649-663.
46. Pang, X., et al., *Evaluation of novel dual acetyl-and butyrylcholinesterase inhibitors as potential anti-Alzheimer's disease agents using pharmacophore, 3D-QSAR, and molecular docking approaches.* Molecules, 2017. **22**(8): p. 1254.
47. Toropov, A.A. and A.P. Toropova, *The index of ideality of correlation: A criterion of predictive potential of QSPR/QSAR models?* Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 2017. 819: p. 31-37.
48. Roy, K. and P.P. Roy, *Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR,*

PLS, GFA, G/PLS and ANN techniques. European journal of medicinal chemistry, 2009. 44(7): p. 2913-2922.

49. Snedecor, G. and W. Cochran, *Statistical methods 6th edition Oxford and IBH Publishing Co.* New Delhi, 1967.
50. Rogers, D. and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships.* Journal of Chemical Information and Computer Sciences, 1994. 34(4): p. 854-866.
51. Balaban, A.T., D. Ciubotariu, and M. Medeleanu, *Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors.* Journal of chemical information and computer sciences, 1991. 31(4): p. 517-523.
52. Kier, L.B. and L.H. Hall, *Molecular connectivity in structure-activity analysis*. 1986: Research Studies.
53. Bonchev, D., *Information theoretic indices for characterization of chemical structures*. 1983: Research Studies Press.
54. Moran, P.A., *Notes on continuous stochastic phenomena.* Biometrika, 1950. 37(1/2): p. 17-23.
55. Geary, R.C., *The contiguity ratio and statistical mapping.* The incorporated statistician, 1954. 5(3): p. 115-146.
56. Wildman, S.A. and G.M. Crippen, *Prediction of physicochemical parameters by atomic contributions.* Journal of chemical information and computer sciences, 1999. 39(5): p. 868-873.
57. Sauer, W.H. and M.K. Schwarz, *Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity.* Journal of chemical information and computer sciences, 2003. 43(3): p. 987-1003.
58. Jaiswal, M., et al., *Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides.* Bioorganic & medicinal chemistry letters, 2004. **14**(12): p. 3283-3290.
59. Tong, W., et al., *A comparative molecular field analysis study of N-benzylpiperidines as acetylcholinesterase inhibitors.* Journal of medicinal chemistry, 1996. **39**(2): p. 380-387.
60. Bernard, P., et al., *Automated docking of 82 N-benzylpiperidine derivatives to mouse acetylcholinesterase and comparative molecular field analysis with'natural'alignment.* Journal of computer-aided molecular design, 1999. **13**(4): p. 355-371.
61. Fernandez, M., et al., *Modeling of acetylcholinesterase inhibition by tacrine analogues using Bayesian-regularized Genetic Neural Networks and ensemble averaging.* Journal of enzyme inhibition and medicinal chemistry, 2006. **21**(6): p. 647-661.
62. Jung, M., et al., *Quantitative structure–activity relationship (QSAR) of tacrine derivatives against acetylcholinesterase (AChE) activity using variable selections.* Bioorganic & medicinal chemistry letters, 2007. **17**(4): p. 1082-1090.
63. Liu, A., et al., *3D-QSAR analysis of a new type of acetylcholinesterase inhibitors.* Science in China Series C: Life Sciences, 2007. **50**(6): p. 726-730.

64. Zhou, A., et al., *Combined 3D-QSAR, molecular docking, and molecular dynamics study of tacrine derivatives as potential acetylcholinesterase (AChE) inhibitors of Alzheimer's disease.* Journal of molecular modeling, 2015. **21**(10): p. 277.
65. Zhang, S., et al., *Design and prediction of new acetylcholinesterase inhibitor via quantitative structure activity relationship of huprines derivatives.* Archives of pharmacal research, 2016. **39**(5): p. 591-602.
66. Jang, C., et al., *Identification of novel acetylcholinesterase inhibitors designed by pharmacophore-based virtual screening, molecular docking and Bioassay.* Scientific reports, 2018. **8**(1): p. 14921.

CONCLUSION GENERALE ET PERSPECTIVES

La modélisation des activités biologiques de molécules étant un champ de recherche important dans le domaine de la conception de médicaments, l'objectif de ce travail s'inscrit dans le cadre d'une contribution à la synthèse de médicaments anti-Alzheimer. Il est question de développer des modèles QSAR fiables et prédictifs pour la prédiction de la concentration inhibitrice de deux séries de composés (dérivés de la N-benzylpipéridine et dérivés de la Tacrine) vis-à-vis des enzymes acétylcholinestérase (AChE) et butyrylcholinestérase (BuChE).

Le choix de la maladie d'Alzheimer n'est pas fortuit. En effet, d'après la recherche bibliographique inhérente à différentes maladies, on a constaté que la maladie d'Alzheimer représente un véritable enjeu de santé publique de par le nombre de personnes atteintes au niveau mondial mais aussi de par son incidence extrêmement élevée et de sa fatalité. A titre d'exemple, en 2016, il a été rapporté que cette maladie est la 3éme cause de mortalité en Algérie. C'est une maladie qui demeure incurable et sans approches thérapeutiques efficaces. Par ailleurs, la revue bibliographique nous a édifiés sur l'importance de l'inhibition de l'activité enzymatique dans la conception de médicaments. A ce titre, deux enzymes (AChE et BuChE) et deux séries de composés ont retenu notre attention du fait de leur incidence sur la maladie d'Alzheimer. De plus, eu égard aux processus traditionnels de conception et de développement de médicaments qui s'avèrent très compliqués, couteux et qui durent dans le temps, la tendance actuelle est à la recherche de méthodes alternatives. Parmi ces dernières, figurent en bonne place celles dites "*in silico*". L'approche QSAR utilisée dans ce présent travail se veut justement une contribution à l'utilisation de ces méthodes alternatives.

La qualité des modèles QSAR est sensible à de nombreux facteurs. La sélection des données expérimentales de référence (nature et précision), tout comme le calcul et la sélection de descripteurs pertinents conditionnent largement la réussite de ce type d'étude. C'est ainsi qu'un soin particulier a été apporté au recueil des données expérimentales de concentration inhibitrice (CI_{50}) et ce en vérifiant que ces valeurs ont été obtenues dans les mêmes conditions expérimentales. Au total, on a recueilli la CI_{50} de 151 dérivés de la Tacrine et de 99 dérivés de la N-benzylpipéridine. Par ailleurs, la sélection des descripteurs les plus pertinents, c'est-à-dire ceux qui sont susceptibles de décrire au mieux le phénomène de l'inhibition, a été entamée. Au final, sur plusieurs centaines de descripteurs calculés, 8 et 10 d'entre eux ont été sélectionnés respectivement dans le cas des dérivés de la Tacrine et de la N-benzylpipéridine. La méthode d'analyse de données est également importante pour le développement de modèles QSAR. A ce titre, quatre méthodes ont été employées. Il s'agit de la régression linéaire multiple (MLR), des réseaux de neurones artificiels (MLP), de la régression de vecteur support (SVR) et de la régression linéaire avec approximation de la fonction génétique (GFA). Une fois développés, Il est impératif de valider les modèles QSAR afin de s'assurer de leur robustesse et de leur pouvoir prédictif. Tous les outils de validation interne et externe ont été appliqués à ces modèles QSAR dédiés à la prédiction de la concentration inhibitrice des dérivés de la Tacrine et des dérivés de la N-benzylpipéridine vis-à-vis des enzymes butyrylcholinestérase (BuChE) et acétylcholinestérase (AChE) respectivement.

Dans une première partie de ce travail, trois modèles QSAR (MLR, MLP et SVR) ont été développés pour prédire la concentration inhibitrice des dérivés de la Tacrine vis-à-vis de l'enzyme butyrylcholinestérase (BuChE). Les modèles conçus, par utilisation de 151 dérivés et de 8 descripteurs, ont montré une très bonne corrélation des données expérimentales avec celles prédites. Les coefficients statistiques de la validation interne pour les 3 modèles sont tous acceptables et satisfaisants. En conséquence, ces modèles sont robustes. La qualité des modèles a également été jugée en termes de validation externe. Là aussi, les 3 modèles se distinguent par un excellent pouvoir prédictif. De plus, la précision et l'exactitude des modèles sont vérifiées avec les valeurs du coefficient "CCC". Par ailleurs, on a noté l'absence de sur-apprentissage pour les trois modèles et ce par vérification des recommandations recommandation de Tropsha et Golbraikh. Néanmoins, la comparaison entre ces trois modèles montre une amélioration substantielle des paramètres statistiques pour le modèle SVR. Par conséquent, ce dernier semble légèrement meilleur que les deux autres. L'établissement du domaine d'applicabilité du modèle SVR selon six approches différentes a montré que 93 à 100% de composés du test appartiennent à ce domaine. Enfin, une analyse de sensibilité a été également réalisée pour déterminer l'importance de chaque descripteur dans la prédiction de l'activité inhibitrice des dérivés de la Tacrine. Aussi, il a été constaté cette activité est influencée par la forme et les masses atomiques des molécules, la charge négative relative, la polarisabilité atomique et l'électronégativité atomique des molécules. En conséquence,

l'identification de ces paramètres qui agissent sur l'activité inhibitrice nous orientera pour concevoir et synthétiser de nouveaux composés actifs sur l'enzyme considérée.

Dans une seconde partie de notre travail, il a été question de développer trois modèles QSAR (MLR, MLP et GFA) pour la prédiction de la concentration inhibitrice des dérivés de la N-benzylpipéridine vis-à-vis de l'enzyme acétylcholinestérase (AChE) moyennant l'utilisation de 99 dérivés et de 10 descripteurs. Les trois modèles conçus ont montré une très bonne corrélation des données expérimentales avec celles prédites. Les coefficients statistiques de la validation interne et externe appliquée à l'ensemble des données sont tous acceptables et satisfaisants. En conséquence, ces modèles sont robustes et se distinguent par un bon pouvoir prédictif. Une étude comparative des 3 modèles a montré que les meilleures performances de prédiction ont été trouvées dans le cas des modèles GFA et MLP, contrairement au modèle MLR. Le domaine d'applicabilité a montré que 98% des composés ont été couverts par les trois modèles lorsqu'ils ont été appliqués pour prédire l'activité inhibitrice des 99 dérivés de N-benzylpipéridine. Ainsi, ces résultats montrent que les modèles sont conformes au troisième principe de l'OCDE. Les modèles développés dans cette seconde partie de ce travail montrent l'importance de la réfractivité molaire, de la structure des molécules et des propriétés électroniques dans le cas de l'activité inhibitrice des dérivés de la N-benzylpipéridine. Comme ce fut le cas pour les dérivés de la Tacrine, la connaissance des facteurs structuraux qui agissent sur l'activité inhibitrice des dérivés de la N-benzylpipéridine nous aidera à concevoir et à synthétiser de nouveaux composés actifs sur l'enzyme acétylcholinestérase (AChE).

En conclusion, nous estimons que l'objectif assigné à cette étude a été atteint en ce sens que les résultats satisfaisants obtenus confirment que les approches QSAR peuvent effectivement servir d'alternative aux méthodes traditionnelles de conception de médicaments (phase de synthèse et de détermination expérimentales de la concentration inhibitrice de composés).

A l'issue de ce travail et au vu des résultats obtenus, de nombreuses perspectives pouvant faire l'objet de recherches sont apparues. Celles-ci se résument à ce qui suit :

- Application des formules mathématiques extraites des modèles pour la prédiction de la concentration inhibitrice de composés qui n'ont pas été utilisés durant ce travail. Par comparaison aux valeurs expérimentales, on s'assurera de la fiabilité ou non des modèles élaborés.
- Calcul et utilisation d'une autre série de descripteurs pour développer d'autres modèles de prédiction.
- Les modéles développés dans ce travail l'ont été sur des séries analogues de composés. Il serait intéressant de voir la possibilité de développer des modéles pour des composés ayant des structures chimiques différentes.

PUBLICATIONS SCIENTIFIQUES

1) **Said Bitam**, Mabrouk Hamadache and Salah Hanini (2017). " ***QSAR model for prediction of the therapeutic potency of N-benzylpiperidine derivatives as AChE inhibitors***." *SAR and QSAR in Environmental Research 28(6): 471-489.*

2) **Said Bitam**, Mabrouk Hamadache and Salah Hanini (2018). " ***Prediction of therapeutic potency of tacrine derivatives as BuChE inhibitors from quantitative structure–activity relationship modelling.***" *SAR and QSAR in Environmental Research 29(3): 213–230*

SAR AND QSAR IN ENVIRONMENTAL RESEARCH, 2017
https://doi.org/10.1080/1062936X.2017.1331467

Taylor & Francis
Taylor & Francis Group

Check for updates

QSAR model for prediction of the therapeutic potency of *N*-benzylpiperidine derivatives as AChE inhibitors

S. Bitam, M. Hamadache and S. Hanini

Department of Process Engineering and Environment, Université Dr Yahia Fares de Médéa, Médéa, Algeria

ABSTRACT

A new family of AChE inhibitors, *N*-benzylpiperidines, showed exceptional efficacy *in vitro* and *in vivo*, minimal side effects and high selectivity for acetylcholinesterase (AChE). Three regression methods were chosen in this work to develop robust predictive models, namely multiple linear regression (MLR), genetic function approximation (GFA) and multilayer perceptron network (MLP). Ten descriptors were selected for a dataset of 99 molecules, using a genetic algorithm. The best results were obtained for MLP with a 10-6-1 artificial neural network model trained with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Statistical prediction for MLR and GFA were r^2 = 0.882 and r^2 = 0.875, respectively. Because internal and external validation strategies play an important role, we adopted all available validation strategies to check the robustness of the models. All criteria used to validate these models revealed the superiority of the GFA model. Therefore, the models developed in this study provide an excellent prediction of the inhibitory concentration of a new family of AChE inhibitors.

ARTICLE HISTORY
Received 15 February 2017
Accepted 14 May 2017

KEYWORDS
QSAR; MLP-ANN; GFA; MLR; AChE inhibitors

Introduction

Tens of millions of people around the world are believed to suffer from Alzheimer's disease (AD). This disease is characterized by progressive memory loss, and a decline in language skills and other cognitive abilities [1]. One of the factors that cause this disease is a decrease in acetylcholine (ACh). This decrease is due to the action of acetylcholinesterase (AChE), which hydrolyses ACh into acetic acid and choline [2]. Acetylcholinesterase inhibitors have been shown to be effective in the treatment of Alzheimer's disease. Inhibition of cholinesterases leads to an increase in the synaptic concentration of acetylcholine. The beneficial effects reported in the treatment of Alzheimer's disease are related to the anticholinesterase action of inhibitors at the cerebral level (especially for AChE). On the other hand, inhibition at the peripheral level (in particular butyrylcholinesterase (BChE)) is the cause of undesirable effects [3]. Also, the few drugs used for the treatment of Alzheimer's disease (tacrine, galanteamine, dopepezil and rivastigmine) are AChE inhibitors [4]. However these AChE inhibitors are known to have side effects such as hepatotoxicity and cardiovascular disease.

CONTACT S. Bitam bitam.said@univ-medea.dz

© 2017 Informa UK Limited, trading as Taylor & Francis Group

Because of these side effects, along with other shortcomings such as the short life of the inhibitors, the search for better AChE inhibitors is still going on [5]. *N*-benzylpiperidines are an example of AChE inhibitors (AChEIs). It has been experimentally demonstrated that this family displays enhanced inhibitory activity (exceptional efficacy *in vitro* and *in vivo*, minimal side effects and high selectivity) towards AChE compared to others inhibitors [6–12].

Quantitative analysis of structure–activity relationships has gained great popularity recently for several reasons, including the reduction of costs, time and labour requirements [13]. Quantitative structure–activity relationships (QSAR) enable the prediction of biological activities for compounds of interest as a function of their descriptors through the use of statistical or machine-learning methods. Aside from the ability to predict activity, QSAR models have been instrumental in enabling the understanding of the origin of these biological activities through the interpretation of the descriptors used in building such models [14]. The success of any QSAR model depends on the accuracy of the input data, the selection of appropriate descriptors and statistical tools and, most importantly, the validation of the developed model [15].

In recent years, a lot of research has been directed towards molecular modelling studies of AChE inhibitors. For example, some authors [1,16–19] have published QSAR studies on tacrine derivatives, among the QSAR research carried out on other AChE inhibitors. Pasquale et al. [20] published in 2012 a QSAR study on chalcone derivatives. In 2014, Fang et al. [21] studied two genistein derivatives, which were reported to be inhibitors of AChE and butyrylcholinesterase (BChE), and differences in the inhibition of AChE were described. In the same year, Ambure et al. [22] utilized molecular docking to build a pharmacophore virtual screening model that provided new suggestions for designing novel AChE inhibitors. A set of 84 known *N*-aryl-monosubstituted derivatives that display inhibitory activity toward both AChE and BuChE was considered by Correa-Basurto et al. [23] for their QSAR study. These QSAR studies led the authors to propose helpful advice regarding the design of AChE and BuChE inhibitors. Recently, Vats et al. [24] utilized group-based QSAR molecular docking and molecular dynamic simulations for 27 novel flavonoid derivatives to build a 3D-QSAR model that provided new useful guides for the design of AChE inhibitors. A QSAR study was carried out by Jiang et al. [25] on AChE inhibitors with comparative molecular field analysis (CoMFA), comparative molecular similarity indices analysis (CoMSIA) and hologram quantitative structure–activity relationship (HQSAR). Statistical results revealed a considerable predictive ability. The authors concluded that the established models may contribute to the development of novel and potential AChE inhibitors. More recently, two different sets of AChE inhibitors (30 compounds and 20 compounds) were investigated through multiple linear regressions (MLR) aided linear and support vector machines (SVM) aided non-linear QSAR models by Pulikkal et al. [26]. The QSAR models obtained were found to be statistically accurate, stable and predictive on validation scales. The authors concluded that the common SAR identified in these linear and non-linear QSAR models could be considered for the design of novel inhibitors of AChE with improved biological activity.

However, among this abundant literature devoted towards molecular modelling studies of AChE inhibitors, QSAR studies specifically dedicated to *N*-benzylpiperidines appear to be rather limited. With this in aim, Cardozo et al. (1992) investigated a series of 18 substituted indanones and benzylpiperidines (18 compounds). A set of QSARs was constructed and evaluated for substituents on the aromatic ring of the benzylpiperidine substructure. The most significant QSAR involved a representation of molecular shape, the largest principal

moment of inertia and the HOMO of the substituted aromatic ring. According to the QSAR analysis, Cardozo et al. [27] suggested that the inhibition potency can be better enhanced by substitution on the indanone ring, as compared to the aromatic sites of the benzylpiperidine ring. In another study, Tong et al. [28] proposed a QSAR model for a series of 1-benzyl-4-[2-(*N*-benzoylamino) ethyl] piperidine derivatives and of *N*-benzylpiperidine benzisoxazoles using the CoMFA approach. The resulting 3D-QSAR indicated a strong correlation between the inhibitory activity of these *N*-benzylpiperidines and the steric and electronic factors that modulate their biochemical activity. Three years later, 3D QSAR studies using the CoMFA method were performed by Bernard et al. [29]. The established CoMFA model was applied to a series of 29 *N*-benzylpiperidine derivatives whose AChE inhibitory activity data were measured under different experimental conditions. A good correlation between predicted and experimental activity data showed that the model could be extended to AChE inhibitory activity data measured on another AChE and/or at different incubation times and pH level. The electronic-topological method (ETM) was applied to three series of *N*-benzylpiperidine derivatives to examine the structures' AChE inhibitor activity relationships [30].

Herein, we propose a QSAR models for a set of 99 compounds derived from *N*-benzylpiperidine. The analyses were performed with relevant descriptors calculated using the Dragon 1.0 software (http://www.vcclab.org) and the results were discussed with regards to the relevance and the predictive potential of the model related to the inhibitory activity. The statistical tools used were multiple linear regression (MLR), genetic function algorithm (GFA), and the non-linear methods artificial neural network (ANN). QSAR models were built in accordance with the guidelines of the Organisation for Economic Cooperation and Development (OECD) [31].

Materials and methods

Datasets

It is well known that high-quality experimental data are essential for the development of high-quality QSAR models. If they are unreliable, the model will be unreliable. The 50% inhibitory concentration (IC_{50}) values of 99 *N*-benzylpiperidine derivatives (Table 1) that were retrieved from the literature [6,9,32–36] were used as the model data set for the present study. The common structural skeletons of AChE inhibitors are given in Figure 1. IC_{50} data correspond to the concentration (mol/l) of a drug that is required for 50% inhibition *in vitro*, and only those measured by Ellman et al.'s method [37] were selected. The IC_{50} values were converted into negative logarithmic scale ($pIC_{50} = -\log IC_{50}$ (M)**)**. All compounds were optimized using the molecular mechanic (MM+) method for generating the initial structures. In order to obtain the minimum energy, a second geometric optimization was performed using a semi-empirical method at the AM1 level implemented in MOPAC 2012 software [38].

Molecular descriptors

One important step in obtaining a QSAR model is the numerical representation of the structural features of molecules by means of molecular descriptors. Molecular descriptors play a fundamental role in developing QSAR models. The descriptors (2489) were obtained for each

Table 1. Dataset corresponding to the observed and predicted values of pIC_{50} of *N*-benzylpiperidine derivatives.

*R1**	*R2*	*R3*	*R4*	*R5*	*R6*	*Reference*	*pIC_{50}*	*GFA*	*MLP*
H	H	H	H	H	Propan-1-ol	[6]	1.000	1.557	1.341
H	H	H	H	H	Ethylpropionate	[6]	1.523	1.281	1.587
H	H	H	H	H	Ethyl acrylate	[6]	1.000	0.991	1.183
H	H	H	H	H	Ethanol	[6]	2.097	1.703	1.977
H	H	H	H	H	Methylacetate	[6]	1.602	1.770	1.841
H	H	H	H	H	8-(Ethoxymethyl)-1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione	[6]	3.000	2.901	2.928
H	H	H	H	H	8-Ethyl-1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione	[6]	2.602	2.480	2.498
H	H	H	H	H	1,3-Dimethyl-7-propyl-3,7-dihydro-1H-purine-2,6-dione	[6]	2.699	2.720	2.620
H	H	H	H	H	9-Ethyl-6-(methylthio)-9H-purine	[6]	1.456	1.619	1.555
H	H	H	H	H	9-Ethyl-6-methoxy-9H-purine	[6]	1.377	1.464	1.367
H	H	H	H	H	7-Ethyl-1-methyl-3,7-dihydro-1H-purine-2,6-dione	[6]	1.699	1.822	1.810
H	H	H	H	H	3-Phenyl-6-propoxypyridazine	[9]	1.523	1.281	1.587
H	H	H	H	H	*N*-propyl-6,7-dihydro-5H-benzo[6,7]cyclohepta[1,2-c]pyridazin-3-amine	[9]	3.921	3.967	4.051
H	H	H	H	H	*N*-propylbenzo[h]cinnolin-3-amine	[9]	3.721	4.343	3.984
H	H	H	H	H	*N*-propyl-5,6-dihydrobenzo[h]cinnolin-3-amine	[9]	3.721	4.080	4.076
H	H	H	H	H	N-propyl-5H-indeno[1,2-c]pyridazin-3-amine	[9]	4.131	4.362	4.097
H	H	H	H	H	5-Ethyl-6-phenyl-N-propylpyridazin-3-amine	[9]	4.180	3.471	4.202
H	H	H	H	H	5-Methyl-6-phenyl-N-propylpyridazin-3-amine	[9]	4.409	3.729	4.104
H	H	H	H	H	6-Phenyl-N-propylpyridazin-3-amine	[9]	3.854	3.490	3.675
H	H	H	H	H	(*E*)-3-(3,4-dihydroxyphenyl)-*N*-propylacrylamide	[32]	1.000	0.991	1.183
H	H	H	H	H	(*E*)-3-(2,4-dihydroxyphenyl)-*N*-propylacrylamide	[32]	3.004	2.600	2.727
H	H	H	H	H	(*E*)-3-(4-hydroxyphenyl)-*N*-propylacrylamide	[32]	2.097	1.703	1.977
H	H	H	H	H	(*E*)-3-(3-hydroxyphenyl)-*N*-propylacrylamide	[32]	1.602	1.770	1.841
H	H	H	H	H	(*E*)-3-(4-hydroxy-3-methoxyphenyl)-*N*-propylacrylamide	[32]	3.409	3.125	3.018
H	H	H	H	H	(*E*)-3-(2,4-dimethoxyphenyl)-*N*-propylacrylamide	[32]	2.602	2.480	2.498
H	H	H	H	H	(*E*)-3-(4-methoxyphenyl)-*N*-propylacrylamide	[32]	2.699	2.720	2.620
H	H	H	H	H	(*E*)-3-(3-methoxyphenyl)-*N*-propylacrylamide	[32]	3.125	2.720	2.619
H	H	H	H	H	(*E*)-3-(3,4-dimethoxyphenyl)-*N*-propylacrylamide	[32]	3.854	3.490	3.675
H	H	H	H	H	2-Ethyl-5,6-dimethoxy-2,3-dihydro-1H-inden-1-one	[32]	5.000	4.459	4.415
H	H	H	H	H	4,5-bis(4-Chlorophenyl)-*N*-methylthiazol-2-amine	[34]	2.860	2.853	2.756
H	H	H	H	H	*N*-methyl-4,5-di-p-tolylthiazol-2-amine	[34]	2.893	2.927	2.878
H	H	H	H	H	4-Methyl-7-((propylamino)methyl)-2H-chromen-2-one	[35]	4.174	3.826	3.702
H	H	H	H	H	3-(2-(Methylamino)ethoxy)-7,8,9,10-tetrahydro-6H-benzo[c]chromen-6-one	[35]	2.863	3.105	3.067
H	H	H	H	H	3-Chloro-4-methyl-7-(2-(methylamino)ethoxy)-2H-chromen-2-one	[35]	2.740	2.955	2.851
H	H	H	H	H	3,4-Dimethyl-7-(2-(methylamino)ethoxy)-2H-chromen-2-one	[35]	2.772	2.886	2.853
H	H	H	H	H	4-Chloro-7-(2-(methylamino)ethoxy)-2H-chromen-2-one	[35]	2.690	3.139	3.044
H	H	H	H	H	3-(4-Hydroxy-3-methoxyphenyl)-N-propylpropanamide	[36]	2.097	1.703	1.977

H	H	H	H	H	(*E*)-3-(4-nitrophenyl)-*N*-propylacrylamide	[36]	3.472	3.097	3.102
H	H	H	H	H	(*E*)-3-(4-chlorophenyl)-*N*-propylacrylamide	[36]	2.863	3.105	3.067
H	H	H	H	H	(*E*)-N-propyl-3-(3,4,5-trimethoxyphenyl)acrylamide	[36]	2.772	2.886	2.853
H	H	H	H	H	(*E*)-3-(4-methoxyphenyl)-*N*-propylacrylamide	[36]	3.161	2.778	2.672
H	H	H	H	H	(*E*)-3-(3,4-dimethoxyphenyl)-*N*-propylacrylamide	[36]	3.279	3.517	3.556
H	H	H	H	H	(*E*)-2-methoxy-4-(3-oxo-3-(propylamino)prop-1-en-1-yl)phenylacetate	[36]	3.470	3.315	3.276
H	H	H	H	H	*N*-propylcinnamamide	[36]	3.386	3.169	3.167
H	H	H	H	H	(*E*)-3-(2-hydroxyphenyl)-*N*-propylacrylamide	[36]	2.572	3.270	3.140
H	H	H	H	H	(*E*)-3-(3,4-dihydroxyphenyl)-*N*-propylacrylamide	[36]	2.690	3.139	3.044
H	H	H	H	H	(*E*)-3-(4-hydroxyphenyl)-*N*-propylacrylamide	[36]	2.658	3.172	3.070
H	H	H	H	H	(*E*)-3-(4-hydroxy-3,5-dimethoxyphenyl)-*N*-propylacrylamide	[36]	3.493	3.625	3.711
H	H	H	H	H	(*E*)-3-(3-hydroxy-4-methoxyphenyl)-*N*-propylacrylamide	[36]	3.000	2.634	2.542
H	H	H	H	H	(*E*)-3-(4-hydroxy-3-methoxyphenyl)-*N*-propylacrylamide	[36]	3.000	2.901	2.928
H	H	H	H	CN	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	4.260	4.773	4.835
H	H	H	H	Ome	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	4.194	4.319	4.424
H	H	H	H	Cl	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	5.292	4.844	4.928
H	H	H	H	Me	8-(Ethoxymethyl)-1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione	[6]	3.000	2.634	2.542
H	H	H	H	CN	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	5.167	4.733	4.817
H	H	H	H	NO2	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	5.538	4.999	5.036
H	H	H	H	OH	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	5.060	4.754	4.868
H	H	H	H	Ome	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	3.903	4.548	4.640
H	H	H	H	Cl	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	5.310	4.708	4.826
H	H	H	H	CN	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	4.495	4.758	4.787
H	H	H	H	NO2	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	4.367	4.982	4.978
H	H	H	H	F	8-Butyryl-1,2,5,6-tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	[33]	5.337	4.804	4.972
H	H	H	H	CF3	4,5-bis(4-Methoxyphenyl)-*N*-methylthiazol-2-amine	[34]	3.125	2.575	2.726
H	H	H	H	Ome	4,5-bis(4-Methoxyphenyl)-*N*-methylthiazol-2-amine	[34]	3.114	3.082	3.142
H	H	H	H	Me	4,5-bis(4-Methoxyphenyl)-*N*-methylthiazol-2-amine	[34]	3.081	2.804	2.667
H	H	H	H	CN	4,5-bis(4-Chlorophenyl)-*N*-methylthiazol-2-amine	[34]	2.991	3.084	2.948
H	H	H	H	Ome	4,5-bis(4-Chlorophenyl)-*N*-methylthiazol-2-amine	[34]	2.951	2.650	2.435
H	H	H	H	Me	4,5-bis(4-Chlorophenyl)-*N*-methylthiazol-2-amine	[34]	2.496	2.817	2.706
H	H	H	H	Me	*N*-methyl-4,5-di-*p*-tolylthiazol-2-amine	[34]	2.963	2.894	2.914
H	H	H	H	F	*N*-ethyl-4,5-di-*p*-tolylthiazol-2-amine	[34]	3.432	3.590	3.627
H	H	H	H	F	*N*-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	[34]	2.936	3.093	3.054
H	H	H	H	F	*N*-(4,5-bis(4-chlorophenyl)thiazol-2-yl)acetamide	[34]	3.187	3.017	3.009
H	H	H	H	NO2	*N*-(4,5-di-*p*-tolylthiazol-2-yl)acetamide	[34]	2.635	2.917	2.830
H	H	H	H	CN	*N*-(4,5-di-*p*-tolylthiazol-2-yl)acetamide	[34]	2.893	2.815	2.778

(*Continued*)

Table 1. (*Continued*).

*R1**	*R2*	*R3*	*R4*	*R5*	*R6*	*Reference*	*pIC_{50}*	*GFA*	*MLP*
H	H	H	F	F	4,5-bis(4-Chlorophenyl)-*N*-ethylthiazol-2-amine	[34]	3.201	3.190	3.124
H	H	H	Me	Me	*N*-ethyl-4,5-di-*p*-tolylthiazol-2-amine	[34]	3.523	3.105	3.038
H	H	H	F	F	*N*-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	[34]	2.611	3.049	3.099
H	H	H	F	H	*N*-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	[34]	3.244	3.096	3.046
H	H	H	F	F	*N*-(4,5-bis(4-chlorophenyl)thiazol-2-yl)acetamide	[34]	3.194	2.857	2.988
H	H	H	F	F	*N*-(4,5-di-*p*-tolylthiazol-2-yl)acetamide	[34]	2.602	2.750	2.791
H	H	Cl	F	H	4,5-bis(4-Chlorophenyl)-*N*-methylthiazol-2-amine	[34]	3.432	3.095	3.077
H	H	Cl	H	F	*N*-ethyl-4,5-bis(4-methoxyphenyl)thiazol-2-amine	[34]	3.444	3.623	3.555
H	H	CF3	H	H	*N*-ethyl-4,5-bis(4-methoxyphenyl)thiazol-2-amine	[34]	3.161	3.230	3.364
H	H	Me	H	H	*N*-ethyl-4,5-bis(4-methoxyphenyl)thiazol-2-amine	[34]	3.229	3.515	3.449
H	H	CF3	F	H	4,5-bis(4-Chlorophenyl)-*N*-ethylthiazol-2-amine	[34]	2.907	3.313	3.134
H	H	Cl	F	H	4,5-bis(4-Chlorophenyl)-*N*-ethylthiazol-2-amine	[34]	3.284	3.526	3.418
H	H	Me	H	H	4,5-bis(4-Chlorophenyl)-*N*-ethylthiazol-2-amine	[34]	3.060	3.539	3.362
H	H	CF3	H	H	*N*-ethyl-4,5-di-*p*-tolylthiazol-2-amine	[34]	3.398	2.971	3.192
H	H	Me	H	H	*N*-ethyl-4,5-di-*p*-tolylthiazol-2-amine	[34]	3.301	3.754	3.463
H	H	CF3	F	H	*N*-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	[34]	2.910	3.001	2.927
H	H	CN	H	H	*N*-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	[34]	2.553	3.002	2.850
H	H	Cl	H	F	*N*-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	[34]	3.387	3.129	3.109
H	H	CF3	H	H	*N*-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	[34]	3.167	2.823	2.765
H	H	Me	H	H	*N*-(4,5-bis(4-methoxyphenyl)thiazol-2-yl)acetamide	[34]	2.873	2.982	2.850
H	H	CN	H	H	*N*-(4,5-bis(4-chlorophenyl)thiazol-2-yl)acetamide	[34]	2.752	2.997	2.849
H	H	CF3	H	H	*N*-(4,5-bis(4-chlorophenyl)thiazol-2-yl)acetamide	[34]	2.558	2.705	2.496
H	H	Me	H	H	*N*-(4,5-bis(4-chlorophenyl)thiazol-2-yl)acetamide	[34]	2.959	2.900	2.772
H	H	Cl	H	F	*N*-(4,5-di-p-tolylthiazol-2-yl)acetamide	[34]	2.541	2.942	2.894
H	H	CF3	H	H	*N*-(4,5-di-p-tolylthiazol-2-yl)acetamide	[34]	2.121	2.614	2.592

*Substituents in Figure 1.

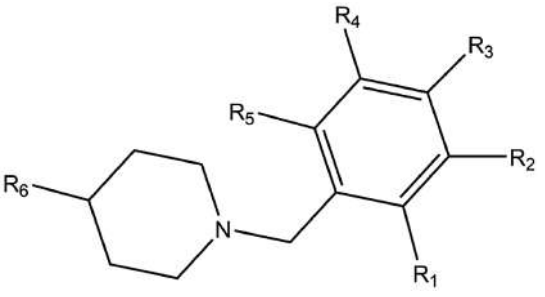


Figure 1. Structure of *N*-benzylpiperidine derivatives.

compound from the online program E-Dragon 1.0 (www.vcclab.org) and Molecular Operating Environment (MOE) software [39]. All of the necessary information about these descriptors is given in the literature [39,40]

Selection of relevant descriptors

An important step in QSAR modeling is to select robust and informative descriptors from a variety of descriptors. As the number of descriptors generated is huge, it must be subjected to an efficient and fast selection of relevant descriptors to avoid the phenomenon of overfitting. Several methods to simplify a database are used, for example: principal component analysis (PCA) [41], partial least squares regression (PLS) [42,43], the selection of relevant variables (forward selections, stepwise selection), and more effective methods such as probe descriptors [44].The method used to select the most significant descriptors has been previously described [45]. In a first step, invariant descriptors, namely those with absent values (represented by the code 999), were manually removed. Next, any descriptor that had identical values for >75% of the samples and any descriptors with a relative standard deviation <0.05 were removed. Finally, half of the descriptors showing an absolute value of the Pearson correlation coefficient >0.75 were also removed. The number of descriptors obtained after selection was 83. For relevant descriptors' selection, a genetic algorithm variable selection was then made using BuildQsar software [46]. Finally, the 10 descriptors obtained after feature pre-screening were used to develop the predictive models.

Model development

The predictive potential of a model on the new data set is influenced by the similarity of the chemical nature between the training set and the test set [15]. Thus, to build robust and predictive QSAR models, the selection of the training set is very important. In this paper, we have classified the data set into clusters using the K-means clustering method [15], implemented in XLSTAT software (http://www.xlstat.com). The 99 compounds were divided into training (75% of the total number of compounds) and test (25% of the total number of compounds) sets. QSAR models were developed using the training set compounds and the developed models were then validated internally by the leave-one-out method and then externally by predicting the inhibitory concentration (IC_{50}) values of the corresponding test set.

For the development of models, three different approaches were used: multiple linear regression (MLR) [47], MLR with genetic function approximation (GFA) [48] and multilayer perceptron network (MLP). MLR was performed using MLRplusValidation1.3 software (dtclab.webs.com/software-tools). GFA and MLP were performed using Accelrys Material Studio 8.0 software (accelrys.com/products/materials-studio).

Statistical methods

MLR

The linear model was developed by applying MLR. The MLR technique has proved by different research groups to be a multidisciplinary technique of valuable applicability for establishing predictive QSAR models. These are the most widely used and known modelling methods, and are used as the basis for a number of multivariate methods [49]. MLR consists of a quantitative relationship between a group of predictor variables (X) and a response Y (activity) as shown by

$$Y = A_k + \sum_{k=1}^{N} A_k X_k \tag{1}$$

where Y is the dependent variable; X_k represents the molecular descriptors; A_k represents the coefficients of those descriptors and A_0 is the intercept of the equation. MLR calculations were performed using MLRplusValidation 1.3.

GFA

The genetic function approximation (GFA) method developed by Rogers and Hopfinger [48] was used to derive the QSAR models. This method combines Holland's genetic algorithm (GA) with Friedman's multivariate adaptive regression splines (MARS) [50]. A distinctive feature of GFA is that it produces a population of models, instead of generating a single model, as do most other statistical methods [51]. The genetic function approximation was used with a mutation probability of 0.1, including full cubic terms. The maximum number of terms allowed in a model is taken to be equal to the number of descriptors used, i.e. 10. The best model was selected according to the lack-of-fit (LOF) values, so that the best model receives the best fitness score. The error measurement term *LOF* is determined by

$$LOF = \frac{LSE}{\left(1 - \frac{c - d*p}{M}\right)^2} \tag{2}$$

In Equation (2), *LSE* is the least squares error, c is the number of basis functions, d is a smoothing parameter (adjustable by the user), p is the total number of features contained in all basis functions and M is the number of samples in the training set.

MLP/ANN

The artificial neural network (ANN) is an information processing pattern that is inspired by the way biological nervous systems, such as the brain, process information. There are plenty of different models of neural networks to choose from, each one having its specific properties and advantages for its particular application. One of the most successful and most popular

is the feed-forward multi-layered perceptron (MLP) [45]. The structure of an MLP consists of one input layer (corresponding to the independent variables: descriptors), one intermediate or hidden layer, and one output layer corresponding to the dependent variable (inhibitory concentration). Each layer can have a number of neurons, which are connected linearly by weights to the neurons in the neighbouring layers. ANN calculations were performed by Accelrys Material Studio 8.0 to study the structure–activity relationship of various *N*-benzylpiperidine derivatives. A set of 10 descriptors were used as input parameters of the network.

Model validation

Validation is a crucial and important aspect for the determination of the reliability of models. There are several approaches for validation, including internal validation and external validation. Recent studies [52] indicated that internal validation is considered to be necessary for model validation. Recently, Ojha et al. [53] proposed r^2_m metrics as additional validation parameters. The r^2_m metrics can be computed using the RmSquare Calculator available at http://aptsoftware.co.in/rmsquare/.

The most important statistical parameters used in our study to check the performance of the model are the root mean square error (RMSE), the determination coefficient (r^2), the determination coefficient adjusted (r^2_{adj}), the cross-validated correlation coefficient (Q^2) and the $\overline{r^2_m}$, Δr^2_m values for the training and test set. The equations for these statistical parameters are available in the literature [44]. A value of $Q^2_{loo} > 0.5$ is considered satisfactory, and a Q^2_{loo} value >0.9 is excellent [54]. The values of both $r^2_{m(loo)}$ for the training set and $r^2_{m(test)}$ for the test set should be higher than 0.5. Thus, Δr^2_m values should be checked for both training and test sets. The values of $\Delta r^2_{m(loo)}$ and $\Delta r^2_{m(test)}$ should be lower than 0.2. Furthermore, the external validation is a significant and necessary validation method used to determine both the generalizability and the true predictive ability of the QSAR models. The external predictive ability of the developed models on the external prediction set was judged based on the predictive r^2 (r^2_{pred}) values calculated according to the equation [15]

$$r^2_{pred} = 1 - \frac{\sum \left(Y_{pred\ (test)} - Y_{obs\ (test)}\right)^2}{\sum \left(Y_{obs\ (test)} - \bar{Y}_{obs\ (training)}\right)^2} \tag{3}$$

In Equation (3), $Y_{pred\ (test)}$ and $Y_{obs\ (test)}$ indicate predicted and observed activity values, respectively, of the test set compounds, and $\bar{Y}_{obs\ (training)}$ indicates the mean activity value of the training set.

Applicability domain using standardization approach

This approach is a simple method for defining outliers (in the case of the training set) and the compounds residing outside the AD (in the case of the test set). An open access standalone application has also been developed for the calculation of the AD for QSAR models. The software can be accessed from the following links: dtclab.webs.com or http://teqip.jdvu.ac.in/QSAR_Tools/. The background theory, the algorithm and methodology and the advantages of the proposed approach are available in the literature [55].

Results and discussion

Selection of relevant descriptors

As the number of *N*-benzylpiperidine derivatives in the training set was 74, it was important to reduce the number of descriptors until the ratio of *N*-benzylpiperidine derivatives/predictors was ≥5 [56]. After using the genetic algorithm, 10 descriptors were selected: VEA1, GCUT_SMR_3, SIC2, std_dim3, X1Av, GATS3e, MATS4v, MATS2e, SlogP_VSA2 and GATS1m.

In order to study the correlation between the selected descriptors, a correlation matrix was established. The value of the correlation coefficient for each pair of selected descriptors was examined. The greatest value of the correlation coefficient (0.455) is that belonging to the pair of descriptors SlogP_VSA2 and SIC2, which means that the selected descriptors were independent. In order to assess the multi-collinearity between the 10 relevant descriptors, their variation inflation factors (VIF) were calculated. If VIF is equal to 1, then no inter-correlation exists for each variable; if VIF is located in the range 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary [57]. The corresponding VIF and tolerance values of the 10 descriptors are presented in Table 2. As can be seen from this table, all of the variables have VIF values of less than five, indicating that the obtained model has statistical significance, and the descriptors were found to be reasonably orthogonal.

MLR predictive model

The model obtained for the prediction of inhibitory concentration of *N*-benzylpiperidine derivatives, using 99 compounds, is the following linear model (Equation (4)) with the reported statistical parameters

$$\begin{aligned} pIC_{50} = &-2.189(\pm4.377) - 31.874(\pm5.330)X1Av - 7.499(\pm1.861)SIC2 \\ &- 2.975(\pm0.329)MATS4v - 2.557(\pm0.336)MATS2e - 1.439(\pm0.302)GATS1m \\ &+ 1.570(\pm0.378)GATS3e - 0.969(\pm0.143)VEA1 + 8.537(\pm0.705)GCUT_{S}MR_{3} \\ &+ 0.007(\pm0.004)S\log P_{V}SA2 - 0.891(\pm0.274)std_{dim3} \end{aligned} \quad (4)$$

where $n_{\text{training}} = 74$, $r^2 = 0.882$, $r^2_{\text{adj}} = 0.864$, $F = 47.213$, $p < 0.0001$, $Q^2 = 0.831$, PRESS = 7.297, $n_{\text{test}} = 25$, $r^2 = 0.746$, $r_0^2 = 0.694$, $r_m^2 = 0.694$, $r^2_{\text{pred}} = 0.561$

The standard errors of the regression coefficients are given within parentheses. The 10 relevant descriptors (variables) in Equation (4) could explain 86.4% of the variance (adjusted coefficient of variation) of the inhibitory concentration. The difference between r^2 and Q^2 is

Table 2. List of descriptors used in the development of different models.

Descriptors	*Category*	*Tolerance*	*VIF*	*t-Test*
X1Av	Connectivity indices	0.507	1.972	−5.980
SIC2	Information indices	0.522	1.917	−4.030
MATS4v	2D autocorrelations	0.677	1.477	−9.039
MATS2e	2D autocorrelations	0.814	1.228	−7.610
GATS1m	2D autocorrelations	0.653	1.532	−4.770
GATS3e	2D autocorrelations	0.684	1.461	4.151
VEA1	Eigenvalue-based indices	0.644	1.553	−6.770
GCUT_SMR_3	Adjacency and distance matrix descriptors	0.689	1.451	12.117
SlogP_VSA2	Subdivided surface areas	0.673	1.485	1.740
std_dim3	Surface area, volume and shape descriptors	0.801	1.248	−3.248

equal to 0.051. This difference is less than 0.3, signifying the robustness of the model [53]. While Equation (4) was applied for prediction of test set compounds, the predictive r^2_{pred} value for the test set was found to be 0.561. An acceptable value of r^2_m (0.694) was obtained. A list of values of different parameters for MLR model is given in Table 3. As can be seen from Table 3, the MLR model gave good results. The predictive results from the MLR model for the entire dataset (99 compounds) are obtained and presented in Table 1. Figure 2 shows the regression line of the model equation, i.e. predicted vs experimental results. A close correlation between the values predicted by the MLR model and the observed values of IC_{50} was found.

Principle 5 of the OECD [31] guidelines requires that a QSAR model should be mechanistically interpretable. By interpreting the descriptors contained in the QSAR model, it is possible to gain some insights into factors, which are related to the inhibitory activity. In the 10 selected descriptors, there are four 2D autocorrelations descriptors, one connectivity index descriptor, one information index descriptor, one eigenvalue-based index descriptor, one adjacency and distance matrix descriptor, one subdivided surface area descriptors, and one surface area, volume and shape descriptor.

To distinguish the importance of each descriptor on inhibitory activity, test values of the involved descriptors are also listed in Table 2. As shown in Table 2, the *t*-test value of adjacency and distance matrix descriptors (GCUT_SMR_3) is 12.117, larger than that of the other descriptors, which indicates that, in this MLR model, the influence of GCUT_SMR_3 on inhibitory activity is stronger. The signs of regression coefficients suggest the direction of the influence of descriptors in a given model. According to the regression coefficients in Equation (4), it is noticeable that the connectivity index (X1Av), the information index (SIC2), the 2D autocorrelations (Mats4v and Mats2e) and the adjacency and distance matrix (GCUT_SMR_3) descriptor are the main contributors to the inhibitory activity. The regression coefficients of the descriptors Gats3e, GCUT_SMR_3 and SlogP_VSA2 have positive signs. These descriptors provide a positive influence, and the highest values were conducive to the improvement of the inhibitory activity of *N*-benzylpiperidine derivatives. In addition, the regression coefficients of the descriptors X1Av, SIC2, MATS4v, MATS2e, GATS1m, VEA1 and std_dim3 have

Table 3. Internal and external criteria of the MLR model.

Validation parameters	*Value*	*Threshold value*
Internal		
r^2	0.882	
r^2_{adj}	0.864	
F	47.213	
Q^2	0.831	>0.5
$\overline{r^2_{m(loo)}}$	0.762	>0.5
$\Delta r^2_{m(LOO)}$	0.104	<0.2
External		
r^2	0.746	>0.6
r^2_0	0.694	
k	0.922	$0.85 < k < 1.15$
k'	1.067	$0.85 < k' < 1.15$
$\left\|r^2_0 - r'^2_0\right\|$	0.050	
$\frac{r^2-r^2_0}{r^2}$	0.070	<0.1
$\frac{r^2-r'^2_0}{r^2}$	0.003	<0.1

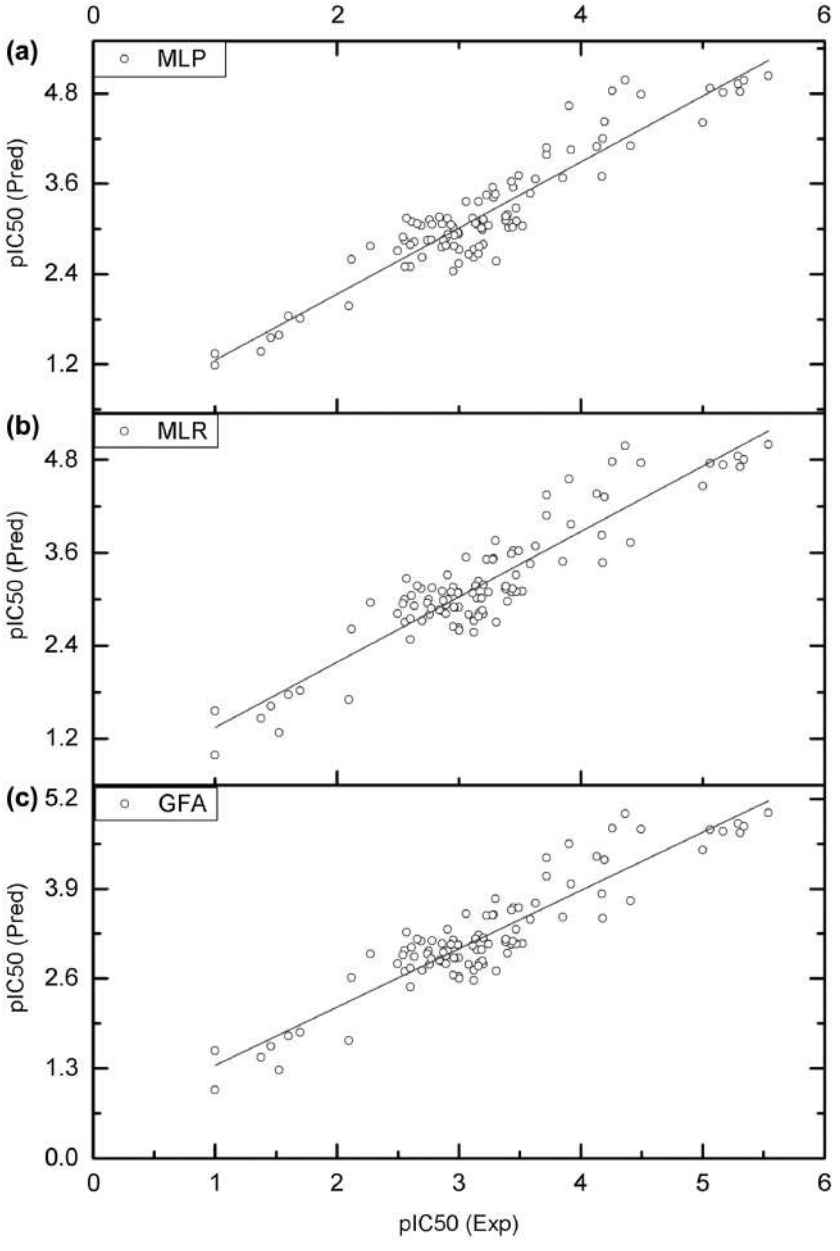


Figure 2. Plot of observed versus predicted pIC_{50} values from the MLR, MLP-ANN and GFA models. (a) MLP; (b) MLR; (c) GFA.

negative signs, thus giving a negative impact on inhibitory activity. For a given *N*-benzylpiperidine derivative, small values for these descriptors would help in improving its inhibitory activity. Thus, a conclusion can be drawn that inhibitory activity is mainly influenced by three important factors: the molar refractivity (GCUT_SMR_3), molecule's structure (X1Av, SIC2, MATS4v) and the electronic properties (MATS2e).

GFA predictive model

In the case of GFA, the following equation with 10 variables appeared to be the best equation

$$\begin{aligned} pIC_{50} = &- 26.034X1Av - 7.013SIC2 - 2.588MATS4v - 1.968MATS2 \\ &- 1.289GATS1m + 1.258GATS3e - 0.900VEA1 + 8.784GCUT_{S}MR_{3} \\ &+ 0.007SlogP_{V}SA2 - 1.026std_{d}im3 - 5,279. \end{aligned} \quad (5)$$

where Friedman $LOF = 0.579$, $r^2 = 0.875$, $F = 61.442$, $Q^2 = 0.842$.

In Equation (5), the 10 relevant descriptors could explain 87.5% of the variance (adjusted coefficient of variation) of the inhibitory concentration. The difference between r^2 and Q^2 is equal to 0.033. This difference is less than 0.3, signifying the robustness of the model. The predictive results from the GFA model for the entire dataset (99 compounds) were obtained and are presented in Table 1. Figure 2 shows the regression line of the model equation, i.e. predicted vs. experimental results. A close correlation between the values predicted by the GFA model and the observed values of IC_{50} was found.

According to the regression coefficients of Equation (5), it is important to note that they are very close to those obtained for the MLR model. We thus find the same positive or negative influence of these descriptors on the inhibitory activity of *N*-benzylpiperidine derivatives.

MLP predictive model

Non-linear modelling with an artificial neural network was also tried. In this investigation, the sigmoid function was used as a transfer function of the hidden layer and an identity function for the output layer. The number of hidden neurons was optimized by a trial and error procedure in the training process. One output neuron was used to represent the observed IC_{50} values. The network was trained using the BFGS quasi-Newton methods algorithm. To optimize the number of nodes in the hidden layer, several calculations were performed with different numbers of hidden nodes (1–10). The 99 *N*-benzylpiperidine derivatives were divided into two groups: training and test sets comprising 75 and 24 compounds, respectively. This network consisted of 10 input (relevant) descriptors used in the other statistical methods (MLR and GFA) and one output for IC_{50}.

The RMSE and Q^2 criteria are employed for the evaluation of the accuracy of both the training and testing sets. The best model was chosen according to the minimum RMSE and the maximum Q^2. Table 4 shows 10 network models that were developed by taking all descriptors. The results obtained show that network 6 with six neurons is the best model. As can be seen from Table 4, the non-linear MLP model gave good results with higher correlation coefficients r^2 (0.867 and 0.899), as well as better robustness Q^2 (0.737 and 0.805) in the training and test sets. Therefore, an MLP with the 10-6-1 architecture was chosen as the final model in this work. The predictive results from the MLP model for the entire dataset (99 compounds) are presented in Table 1. Figure 2 shows the regression line of the model equation, i.e. predicted vs. experimental results. A close correlation between the values predicted by the ANN model and the observed values of IC_{50} was found.

Table 4. Selected criteria of the different multi-layer perceptron.

Number of hidden neurons	r^2 *(train)*	Q^2 *(train)*	r^2 *(test)*	Q^2 *(test)*	*RMSE*
1	0.842	0.756	0.785	0.795	0.395
2	0.841	0.725	0.698	0.802	0.397
3	0.844	0.757	0.778	0.800	0.393
4	0.815	0.693	0.829	0.809	0.428
5	0.835	0.768	0.830	0.796	0.404
6	**0.867**	**0.737**	**0.899**	**0.805**	**0.363**
7	0.854	0.756	0.796	0.810	0.380
8	0.863	0.694	0.890	0.794	0.368
9	0.848	0.747	0.784	0.806	0.388
10	0.813	0.742	0.844	0.817	0.430

Table 5. Selected criteria of the different models.

Statistical parameter	*MLR*	*GFA*	*MLP*
$Q^2_{(\text{overall})}$	0.831	0.842	0.737
$\underline{r^2_{m(\text{overall})}}$	0.843	0.875	0.869
$r^2_{m(\text{overall})}$	0.773	0.818	0.819
$\Delta r^2_{m(\text{overall})}$	0.139	0.113	0.100
$\underline{r^2_{m(\text{overall})}}$ (scaled)	0.843	0.875	0.868
$r^2_{m(\text{overall})}$ (scaled)	0.776	0.820	0.821
$\Delta r^2_{m(\text{overall})}$ (scaled)	0.134	0.109	0.094

Each entry for the submitted dataset is scaled as follows. Scaled Z(observed or predicted) = (Z − minimum of observed)/(maximum of observed − minimum of observed).

To see the importance of each variable for the prediction of inhibitory activity of *N*-benzylpiperidine derivatives, a sensitivity analysis was conducted This method, proposed by Garson [58] then taken by Goh [59], provides a quantification of the relative importance of different inputs (variables) on the output of the neural network. The contribution of each of the descriptors in the MLP model is as follows: VEA1 (16.98%), GCUT_SMR_3 (16.87%), SIC2 (14.76%), std_dim3 (13.46%), X1Av (11.85%), GATS3e (7.41%), MATS4v (5.16%), MATS2e (4.71%), SlogP_VSA2 (5.99%) and GATS1m (2.80%).

Comparison of three statistical models

To compare statistically the three models, we used the approach of Roy and his co-workers. Roy and colleagues [60] have shown that the parameters $r^2_{m(\text{loo})}$, $r^2_{m(\text{test})}$, $\Delta r^2_{m(\text{loo})}$ and $\Delta r^2_{m(\text{test})}$ can be extended to the whole data set. The new parameters have been designated by $r^2_{m(\text{overall})}$ and $\Delta r^2_{m(\text{overall})}$. The advantages of such an approach are to facilitate the selection of the best model on the basis of an overall contribution of internal and external validation measures. For the present QSAR study, we have determined r^2_m values for the whole set for the three models and the results are shown in Table 5. Based on the r^2_m (whole) statistic, the GFA and MLP models are better than the MLR model. The difference between the two models (GFA and MLP) is insignificant. Moreover, the values of $\Delta r^2_{m(\text{overall})}$ and $\underline{r^2_{m(\text{overall})}}$ imply that the MLP model is slightly better than the GFA model. Thus, it can be concluded that both the MLP and GFA models have better predictive power than the MLR model. This means that the models obtained allows the establishment, to a large extent, of a nonlinear relationship

between the therapeutic activity of *N*-benzylpiperidine derivatives and the structural information (selected molecular descriptors).

Applicability domain

The applicability domain of the models was analysed using a standardization approach. The software can be accessed from dtclab.webs.com or http://teqip.jdvu.ac.in/QSAR_Tools/. In the training set, one compound was identified as an outlier and in the test set one compound was identified as being outside the AD by the standardization approach. It should be noted that over 97.5% of the domain was covered by the models when they are applied to predict the inhibitory activity of 99 *N*-benzylpiperidine derivatives in the training and test sets. Thus, these results show that the models comply with the third principle of the OECD. They can be used to predict the inhibitory activity of 99 *N*-benzylpiperidine derivatives, particularly for those that have not been tested as well as new compounds.

Overview

Three statistical methods, MLR, GFA and MLP have been applied for the linear and non-linear modelling of *N*-benzylpiperidine derivative inhibitors using 10 descriptors. In all cases, the variable inflation factor was found to be less than 1.972 and the tolerance value was more than 0.507, which suggests an absence of multicollinearity. The whole data set was divided into training set (74 compounds) and test set (25 compounds) based on *K*-means clustering. Models were developed from the training set. In this present QSAR study, we have determined $r^2_{m(\text{overall})}$ and $r^2_{m(\text{overall})}$ for the whole set for the three models and the results are shown in Table 5. A comparison of statistical quality of the different models is given in Table 5. The MLR model obtained from the whole data set are of statistical significance but its performance is inferior to the GFA and MLP-derived models in terms of $\Delta r^2_{m(\text{overall})}$ and $r^2_{m(\text{overall})}$

Conclusions

In this study, linear and non-linear QSAR models were used to predict the inhibitory activity of 99 *N*-benzylpiperidine derivatives based on their molecular structure, represented by 10 relevant descriptors. Experimental data have been selected from literature. The three developed models showed good predictive capacity. To select the best model, the parameter $r^2_{m(\text{overall})}$ was used. According to some authors (for example Roy and co-workers[60]) this parameter appears to be advantageous over other internal and external validation parameters in that it is based on the prediction of both training and test set compounds and thus involving more compounds in the prediction process. The best model is that established with GFA. The developed models in this work show the importance of the molar refractivity, the molecules' structure and electronic properties in describing the *N*-benzylpiperidine derivatives' inhibitory activity. The developed models may be helpful in designing potent AChE inhibitors and predicting the AChE inhibition potential of novel drug candidates.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] K.Y. Wong, A.G. Mercader, L.M. Saavedra, B. Honarparvar, G.P. Romanelli, and P.R. Duchowicz, *QSAR analysis on tacrine-related acetylcholinesterase inhibitors*, J. Biomed. Sci. 21 (2014): 84. doi: https://doi.org/10.1186/s12929-014-0084-0.

[2] A. Zhou, J. Hu, L. Wang, G. Zhong, J. Pan, Z. Wu, and A. Hui, *Combined 3D-QSAR, molecular docking, and molecular dynamics study of tacrine derivatives as potential acetylcholinesterase (AChE) inhibitors of Alzheimer's disease*, J. Mol. Model. 21(2015): 277. doi: https://doi.org/10.1007/s00894-015-2797-8.

[3] B.J. Vellas and P. Robert, *Fédération nationale des centres Mémoire de Ressources et de Recherche, Traité sur la maladie d'Alzheimer*, Springer-Verlag France, Paris, 2013.

[4] K.K. Roy, A. Dixit, and A.K. Saxena, *An investigation of structurally diverse carbamates for acetylcholinesterase (AChE) inhibition using 3D-QSAR analysis*, J. Mol. Graph. Model. 27 (2008), pp. 197–208.

[5] M.R. Islam, A. Zaman, I. Jahan, R. Chakravorty, and S. Chakraborty, *In silico QSAR analysis of quercetin reveals its potential as therapeutic drug for Alzheimer's disease*, J. Young Pharm. 5 (2013), pp. 173–179.

[6] M. Estrada, C. Herrera-Arozamena, C. Pérez, D. Viña, A. Romero, J.A. Morales-García, A. Pérez-Castillo, and M.I. Rodríguez-Franco, *New cinnamic–N-benzylpiperidine and cinnamic–N, N-dibenzyl (N-methyl) amine hybrids as Alzheimer-directed multitarget drugs with antioxidant, cholinergic, neuroprotective and neurogenic properties*, Eur. J. Med. Chem. 121 (2016), pp. 376–386.

[7] M.L. Kostochka, J. Zajicek, J.A. Fuselier, M.A. Etienne, L. Sun, and D.H. Coy, *Novel tandem aldol intramolecular cyclization of substituted n-benzylpiperidine-4-one: Synthesis of novel-type nitrogen 2, 8-phenanthroline heterocycles*, J. Heterocyclic Chem. 52 (2015), pp. 1723–1730.

[8] A. Martinez, E. Fernandez, A. Castro, S. Conde, I. Rodriguez-Franco, J.-E. Baños, and A. Badia, *N-Benzylpiperidine derivatives of 1, 2, 4-thiadiazolidinone as new acetylcholinesterase inhibitors*, Eur. J. Med. Chem. 35 (2000), pp. 913–922.

[9] M.I. Rodríguez-Franco, M.I. Fernández-Bachiller, C. Pérez, A. Castro, and A. Martínez, *Design and synthesis of N-benzylpiperidine–purine derivatives as new dual inhibitors of acetyl-and butyrylcholinesterase*, Bioorgan. Med. Chem. 13 (2005), pp. 6795–6802.

[10] D.K. Sukumarapillai, K. Kooi-Yeong, Y. Kia, V. Murugaiyah, and S.K. Iyer, *Design, synthesis and cholinesterase inhibitory evaluation study of fluorescent N-benzylpiperidine-4-one derivatives*, Med. Chem. Res. 25 (2016), pp. 1705–1715.

[11] A. Więckowska, K. Więckowski, M. Bajda, B. Brus, K. Sałat, P. Czerwińska, S. Gobec, B. Filipek, and B. Malawska, *Synthesis of new N-benzylpiperidine derivatives as cholinesterase inhibitors with β-amyloid anti-aggregation properties and beneficial effects on memory in vivo*, Bioorgan. Med. Chem. 23 (2015), pp. 2445–2457.

[12] M. Shidore, J. Machhi, K. Shingala, P. Murumkar, M.K. Sharma, N. Agrawal, A. Tripathi, Z. Parikh, P. Pillai, and M.R. Yadav, *Benzylpiperidine-linked diarylthiazoles as potential anti-Alzheimer's agents: Synthesis and biological evaluation*, J. Med. Chem. 59 (2016), pp. 5823–5846.

[13] K. Roy, S. Kar, and R.N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, 2015: 1–46.

[14] S. Simeon, N. Anuwongcharoen, W. Shoombuatong, A.A. Malik, V. Prachayasittikul, J.E. Wikberg, and C. Nantasenamat, *Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking*, PeerJ 4 (2016), p. e2322.

[15] K. Roy and P.P. Roy, *Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques*, Eur. J. Med. Chem. 44 (2009), pp. 2913–2922.

[16] N. Akula, L. Lecanu, J. Greeson, and V. Papadopoulos, *3D QSAR studies of AChE inhibitors based on molecular docking scores and CoMFA*, Bioorgan. Med. Chem. Lett. 16 (2006), pp. 6277–6280.

[17] M. Jung, J. Tak, Y. Lee, and Y. Jung, *Quantitative structure–activity relationship (QSAR) of tacrine derivatives against acetylcholinesterase (AChE) activity using variable selections*, Bioorgan. Med. Chem. Lett. 17 (2007), pp. 1082–1090.
[18] M. Saracoglu and F. Kandemirli, *The investigation of structure–activity relationships of tacrine analogues: Electronic-topological method*, Open Med. J. 2 (2008), pp. 75–80.
[19] N. Chen, C. Liu, L. Zhao, and H. Zhang, *3D-QSAR study of multi-target-directed AchE inhibitors based on autodocking*, Med. Chem. Res. 21 (2012), pp. 245–256.
[20] G. Pasquale, G.P. Romanelli, J.C. Autino, J. García, E.V. Ortiz, and P.R. Duchowicz, *Quantitative structure–activity relationships of mosquito larvicidalchalcone derivatives*, J. Agric. Food Chem. 60 (2012), pp. 692–697.
[21] J. Fang, P. Wu, R. Yang, L. Gao, C. Li, D. Wang, S. Wu, A.-L. Liu, and G.-H. Du, *Inhibition of acetylcholinesterase by two genistein derivatives: Kinetic analysis, molecular docking and molecular dynamics simulation*, Acta Pharm. Sin. B 4 (2014), pp. 430–437.
[22] P. Ambure, S. Kar, and K. Roy, *Pharmacophore mapping-based virtual screening followed by molecular docking studies in search of potential acetylcholinesterase inhibitors as anti-Alzheimer's agents*, Biosystems 116 (2014), pp. 10–20.
[23] J. Correa-Basurto, M. Bello, M.C. Rosales-Hernandez, M. Hernández-Rodríguez, I. Nicolás-Vázquez, A. Rojo-Domínguez, J.G. Trujillo-Ferrara, R. Miranda, and C. Flores-Sandoval, *QSAR, docking, dynamic simulation and quantum mechanics studies to explore the recognition properties of cholinesterase binding sites*, Chem. Biol. Interact. 209 (2014), pp. 1–13.
[24] C. Vats, J.K. Dhanjal, S. Goyal, N. Bharadvaja, and A. Grover, *Computational design of novel flavonoid analogues as potential AChE inhibitors: Analysis using group-based QSAR, molecular docking and molecular dynamics simulations*, Struct. Chem. 26 (2015), pp. 467–476.
[25] Y.-R. Jiang, Y.-Y. Yang, Y.-L. Chen, and Z.-J. Liang, *CoMFA, CoMSIA and HQSAR studies of acetylcholinesterase inhibitors*, Curr. Comput. Aided Drug Des. 9 (2013), pp. 385–395.
[26] B.P. Pulikkal, *Common SAR derived from linear and non-linear QSAR studies on AChE inhibitors used in the treatment of Alzheimer's disease*, Curr. Neuropharmacol. 14, pp. 1–7.
[27] M.G. Cardozo, Y. Imura, H. Sugimoto, Y. Yamanishi, and A.J. Hopfinger, *QSAR analyses of the substituted indanone and benzylpiperidine rings of a series of indanone-benzylpiperidine inhibitors of acetylcholinesterase*, J. Med. Chem. 35 (1992), pp. 584–589.
[28] W. Tong, E.R. Collantes, Y. Chen, and W.J. Welsh, *A comparative molecular field analysis study of N-benzylpiperidines as acetylcholinesterase inhibitors*, J. Med. Chem. 39 (1996), pp. 380–387.
[29] P. Bernard, D.B. Kireev, J.R. Chrétien, P.-L. Fortier, and L. Coppet, *Automated docking of 82 N-benzylpiperidine derivatives to mouse acetylcholinesterase and comparative molecular field analysis with 'natural' alignment*, J. Comput. Aided Mol. Des. 13 (1999), pp. 355–371.
[30] A.S. Dimoglo, N.M. Shvets, I.V. Tetko, and D.J. Livingstone, *Electronic-topological investigation of the structure – acetylcholinesterase inhibitor activity relationship in the series of n-benzylpiperidine derivatives*, Quant. Struct.-Act. Rel. 20 (2001), pp. 31–45.
[31] OECD, *Guidance document on the validation of (quantitative) structure–activity relationships [(Q) SAR] models*, OECD Series on Testing and Assessment No. 69. ENV/JM/MONO (2007) 2 154, OECD, Paris. 2007.
[32] J.-M. Contreras, I. Parrot, W. Sippl, Y.M. Rival, and C.G. Wermuth, *Design, synthesis, and structure–activity relationships of a series of 3-[2-(1-benzylpiperidin-4-yl) ethylamino] pyridazine derivatives as acetylcholinesterase inhibitors*, J. Med. Chem. 44 (2001), pp. 2707–2718.
[33] Y. Ishichi, M. Sasaki, M. Setoh, T. Tsukamoto, S. Miwatashi, H. Nagabukuro, S. Okanishi, S. Imai, R. Saikawa, and T. Doi, *Novel acetylcholinesterase inhibitor as increasing agent on rhythmic bladder contractions: SAR of 8-{3-[1-(3-fluorobenzyl) piperidin-4-yl] propanoyl}-1, 2, 5, 6-tetrahydro-4H-pyrrolo [3, 2, 1-ij] quinolin-4-one (TAK-802) and related compounds*, Bioorgan. Med. Chem. 13 (2005), pp. 1901–1911.
[34] M. Shidore, J. Machhi, K. Shingala, P. Murumkar, M.K. Sharma, N. Agrawal, A. Tripathi, Z. Parikh, P. Pillai, and M.R. Yadav, *Benzylpiperidine-linked diarylthiazoles as potential anti-Alzheimer's agents-synthesis and biological evaluation*, J. Med. Chem. 59 (2016), pp. 5823–46.

[35] S.-S. Xie, J.-S. Lan, X. Wang, Z.-M. Wang, N. Jiang, F. Li, J.-J. Wu, J. Wang, and L.-Y. Kong, *Design, synthesis and biological evaluation of novel donepezil–coumarin hybrids as multi-target agents for the treatment of Alzheimer's disease*, Bioorgan. Med. Chem. 24 (2016), pp. 1528–1539.
[36] W. Xu, X.-B. Wang, Z.-M. Wang, J.-J. Wu, F. Li, J. Wang, and L.-Y. Kong, *Synthesis and evaluation of donepezil–ferulic acid hybrids as multi-target-directed ligands against Alzheimer's disease*, Med. Chem. Comm. 7 (2016), pp. 990–998.
[37] G.L. Ellman, K.D. Courtney, V. Andres, and R.M. Featherstone, *A new and rapid colorimetric determination of acetylcholinesterase activity*, Biochem. Pharmacol. 7 (1961), pp. 88IN191–9095.
[38] *MOPAC 2012, Version 15.038W*. Stewart Computational Chemistry, 2012; software available at http://OpenMOPAC.net.
[39] *Molecular Operating Environment (MOE)*, Chemical Computing Group Inc., Montreal, Canada, 2012.
[40] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Vol. I and II, (2009), pp. 26–29, 173, 176, 757.
[41] I.T. Jolliffe and J. Cadima, *Principal component analysis: A review and recent developments*, Phil. Trans. R. Soc. A 374 (2016), p. 20150202.
[42] A.-L. Boulesteix and K. Strimmer, *Partial least squares: A versatile tool for the analysis of high-dimensional genomic data*, Brief. Bioinformatics 8 (2007), pp. 32–44.
[43] N. Urbach and F. Ahlemann, *Structural equation modeling in information systems research using partial least squares*, J. Inform. Technol. Theory Appl. 11 (2010), pp. 5–40.
[44] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, *Ranking a random feature for variable and feature selection*, J. Mach. Learn. Res. 3 (2003), pp. 1399–1414.
[45] M. Hamadache, S. Hanini, O. Benkortbi, A. Amrane, L. Khaouane, and C.S. Moussa, *Artificial neural network-based equation to predict the toxicity of herbicides on rats*, Chemometr. Intell. Lab. Syst. 154 (2016), pp. 7–15.
[46] D.B. de Oliveira and A.C. Gaudio, *BuildQSAR: A new computer program for QSAR analysis*, Quant.-Struct.-Act. Rel. 19 (2000), pp. 599–601.
[47] G. Snedecor and W. Cochran, *Statistical Methods*, Oxford and IBH Publishing Co., New Delhi, 1967.
[48] D. Rogers and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships*, J. Chem. Inform. Comput. Sci. 34 (1994), pp. 854–866.
[49] M. Hamadache, L. Khaouane, O. Benkortbi, C. Si Moussa, S. Hanini, and A. Amrane, *Prediction of acute herbicide toxicity in rats from quantitative structure–activity relationship modeling*, Environ. Eng. Sci. 31 (2014), pp. 243–252.
[50] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, and J.N. Weinstein, *Quantitative structure-antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies*, J. Med. Chem. 44 (2001), pp. 3254–3263.
[51] P. Pratim Roy, S. Paul, I. Mitra, and K. Roy, *On two novel parameters for validation of predictive QSAR models*, Molecules 14 (2009), pp. 1660–1701.
[52] R. Wang, J. Jiang, Y. Pan, H. Cao, and Yi. Cui, *Prediction of impact sensitivity of nitro energetic compounds by neural network based on electrotopological-state indices*, J. Hazard. Mater. 166 (2009), pp. 155–186.
[53] P.K. Ojha, I. Mitra, R.N. Das, and K. Roy, *Further exploring metrics for validation of QSPR models*, Chemometr. Intell. Lab. Syst. 107 (2011), pp. 194–205.
[54] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs*, Environ. Health Persp. 111 (2003), pp. 1361–1375.
[55] K. Roy, S. Kar, and P. Ambure, *On a simple approach for determining applicability domain of QSAR models*, Chemometr. Intell. Lab. Syst. 145 (2015), pp. 22–29.
[56] A. Tropsha, P. Gramatica, and V.K. Gombar, *The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models*, QSAR Comb. Sci. 22 (2003), pp. 69–77.
[57] M. Jaiswal, P.V. Khadikar, A. Scozzafava, and C.T. Supuran, *Carbonic anhydrase inhibitors: The first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides*, Bioorgan. Med. Chem. Lett. 14 (2004), pp. 3283–3290.

[58] G.D. Garson, *Interpreting neural network connection weights*, AI Expert 6 (1991), pp. 47–51.
[59] A.T. Goh, *Back-propagation neural networks for modeling complex systems*, Artif. Intell. Eng. 9 (1995), pp. 143–151.
[60] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, and H. Kabir, *Comparative studies on some metrics for external validation of QSPR models*, J. Chem. Inform. Model. 52 (2012), pp. 396–408.

SAR AND QSAR IN ENVIRONMENTAL RESEARCH, 2018
VOL. 29, NO. 3, 213–230
https://doi.org/10.1080/1062936X.2018.1423640

Taylor & Francis
Taylor & Francis Group

Check for updates

Prediction of therapeutic potency of tacrine derivatives as BuChE inhibitors from quantitative structure–activity relationship modelling

S. Bitam, M. Hamadache and S. Hanini

Département du Génie des Procédés et Environnement, Université de Médéa, Quartier Ain D'heb, Médéa, Algeria

ABSTRACT

Numerous studies show that tacrine derivatives exhibit increased inhibitory activity against butyrylcholinesterase (BuChE) and acetylcholinesterase (AChE). However, the screening assays for currently available BuChE inhibitors are expensive, time consuming and dependent on the inhibitory compound. It is therefore desirable to develop alternative methods to facilitate the screening of these derivatives in the early phase of drug discovery. In order to develop robust predictive models, three regression methods were chosen in this study: multiple linear regression (MLR), support vector regression (SVR) and multilayer perceptron network (MLP). Eight relevant descriptors were selected on a dataset of 151 molecules using a method based on genetic algorithms. Internal and external validation strategies play an important role. Also, to check the robustness of the selected models, all available validation strategies were used, and all criteria used to validate these models revealed the superiority of the SVR model. The statistical parameters obtained with the SVR model were RMSE = 0.197, r^2 = 0.969 and Q^2 = 0.964 for the training set, and r^2 = 0.906 and Q^2 = 0.891 for the test set. Therefore, the model developed in this study provides an excellent prediction of the inhibitory concentration of tacrine derivatives.

ARTICLE HISTORY
Received 24 October 2017
Accepted 1 January 2018

KEYWORDS
QSAR; multilayer perceptron network; support vector regression; BuChE inhibitors

Introduction

It is estimated that more than 35 million elderly people in the world suffer from Alzheimer's disease (AD). This disease has emerged as the leading cause of dementia, but also the most common cause of memory loss and cognitive impairment [1]. The incidence of AD increases sharply with age. As a result, the increasing proportion of people who survive to age 80 and older will increase the incidence of this disease in the coming years [2].

The only symptomatic treatment of this disease proven to date is the use of cholinesterase (ChE) inhibitors to increase cholinergic activity. There are two types of ChE that are ubiquitous throughout the body. In the central nervous system, acetylcholinesterase (AChE) is mainly localized in neurons, while butyrylcholinesterase (BuChE) is mainly associated with glial cells

CONTACT S. Bitam bitam.said@univ-medea.dz
Supplemental data for this article can be accessed here at https://doi.org/10.1080/1062936X.2018.1423640

© 2018 Informa UK Limited, trading as Taylor & Francis Group

and is secreted by these [3]. Some authors [3–5] have shown that up to 45% of AChE can be lost in some areas of the brain during AD progression, while activity levels of BuChE increase up to 90%. These studies focus on BuChE and its role in the pathology of AD and highlight the utility of BuChE inhibition in slowing the progression of the disease [6]. In other words, the design of new and more potent and selective BuChE inhibitors is of great importance in drug discovery.

Tacrine is one of the most potent inhibitors of BuChE known to date [1]. It is considered a reference in the development of new BuChE inhibitors as potential drugs for AD. For example, numerous studies show that tacrine derivatives exhibit increased inhibitory activity against BuChE and AChE [7–10]. However, the currently available BuChE inhibitor screening assays are expensive, labour-intensive and compound-dependent. It would therefore be desirable to develop computational, reliable and rapid methods, which will facilitate the screening of these derivatives in the early phase of drug discovery.

Quantitative analysis of the structure–activity relationship has recently gained wide popularity for a number of reasons, including reduced costs, time and labour requirements [11]. The quantitative structure–activity relationship (QSAR) allows the prediction of biological activities for compounds of interest according to their descriptors through the use of statistical methods or machine learning. In addition to the ability to predict activity, QSAR models have made it possible to understand the origin of these biological activities through the interpretation of the descriptors used in the construction of such models [12]. The success of any QSAR model depends on the accuracy of the input data, the selection of appropriate descriptors, statistical tools, and especially the validation of the developed model [13].

Several QSAR studies dedicated to the inhibitory power of tacrine derivatives on AChE have been performed. Three-dimensional (3D) QSAR studies on AChE inhibitors, based on molecular docking scores and comparative molecular field analysis (CoMFA), were performed by Akula et al. [14]. The high correlation coefficient reveals that the model is a useful tool for the prediction of test set as well as newly designed structures against AChE activity. In the same year, AChE inhibition was modelled by Fernández et al. [15] for a set of 136 tacrine analogues using Bayesian-regularized genetic neural networks (BRGNNs). The predictive capacity of selected models was evaluated by averaging multiple validation sets generated as members of diverse-training set neural network ensembles (NNEs). A reliable measure of training and test set R values were obtained, i.e. 0.921 and 0.851, respectively. Saracoglu and Kandemirli [16] investigated the structure–activity relationships by using the electron-topological method (ETM) for a class of AChE inhibitors related to tacrine (9-amino-1,2,3,4-tetrahydroacridine) and 11*H*-indeno-[1,2-*b*]-quinolin-10-ylamine, a tetracyclic tacrine analogue, a drug currently in use for the treatment of AD. The result of testing showed the high ability of ETM in predicting the activity and inactivity in investigated series. Various QSAR models for a set of 80 tacrine derivatives have been developed by Jung et al. [17] using variable selections of stepwise multiple linear regression (MLR), genetic algorithm (GA)-MLR and simulated annealing (SA)-MLR. The best equation was obtained from SA-MLR with greater explanatory capability and better prediction, with a smaller standard error than other methods. Furthermore, the hydrophilic and topological features of molecules were shown to decrease AChE activity. Chekmarev et al. [18] developed and tested support vector machine (SVM) classification and regression models with molecular descriptors derived from shape signatures and the molecular operating environment (MOE) application software. According to the results obtained with a set of piperidine AChE inhibitors (N = 110), they concluded

that these new machine learning models with combined descriptor schemes may find utility for predicting novel AChE inhibitors. Recently, Wong et al. [19] have used 10 groups of new developed tacrine-related inhibitors and established a QSAR on their AChE inhibitory activity. The QSAR models developed in this work have a satisfactory predictive ability.

Unfortunately, it should be noted that there are only few QSAR studies on the inhibitory activities of tacrine derivatives against BuChE. Castilho et al. [20] conducted two-dimensional (2D) QSAR studies on a set of 40 potent inhibitors of human BuChE using classical and hologram QSAR (HQSAR) approaches. The best HQSAR model was used to predict the potency of 10 test set compounds, and the predicted values were in good agreement with the experimental results, showing the potential of this model for new untested compounds. Recently, in an investigation conducted by Fang et al. [21], SVM models and naive Bayesian models were built to discriminate BuChE inhibitors from the non-inhibitors. The models generated from SVM and naive Bayesian approaches successfully predicted BuChE inhibitors.

The goal of this work was therefore to develop QSAR models that are predictive of the inhibitory activity of 151 compounds derived from tacrine; their inhibitory activity regarding not only AChE, but also and especially BuChE was examined. For this, three techniques were used: MLR, support vector regression (SVR) and artificial neural network (ANN). Analyses were performed with relevant descriptors calculated using the Dragon 1.0 software and the MOE application software. The QSAR models have been developed in accordance with the guidelines of the Organization for Economic Co-operation and Development (OECD) [22].

Materials and methods

Datasets

Inhibitory concentration (IC_{50}) values of the tacrine derivatives against BuChE (Table 1) were recovered from the BindingDB database [23]. The techniques used to compile the datasets were as follows:

- Select inhibitors with tacrine central scaffolding.
- Eliminate duplicate compounds.
- Selection of all structures for which cholinesterase tests were performed using human serum with the Elman-reported colour method [24].

For this purpose, 151 inhibitors with a tacrine scaffold designed with different functional groups were selected (see supplementary information, available online at the journal's website at: https://doi.org/10.1080/1062936X.2018.1423640). The structural skeletons of BuChE inhibitors are given in Figure 1. The IC_{50} data correspond to the concentration (μmol/L) of a drug required for a 50% inhibition in vitro. The values of the IC_{50} were converted to negative

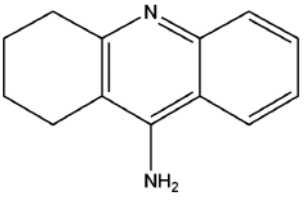


Figure 1. Molecular structure of tacrine.

Table 1. Dataset corresponding to the observed and predicted values of pIC50 of tacrine derivatives.

N°	PubChem CID	pIC_{50} (μm)	pIC_{50} (MLR)	pIC_{50} (MLP)	pIC_{50} (SVR)
Training set					
1	16094876	3.857	3.206	3.526	3.715
2	16094860	3.851	3.648	4.030	3.709
3	16094859	3.646	3.490	3.747	3.548
4	90654210	3.62	3.560	3.650	3.480
5	90654209	3.602	3.110	3.243	3.328
6	54768920	1.420	1.807	1.483	1.420
7	71653762	1.420	0.646	0.628	1.069
8	1935	1.398	1.192	1.383	1.258
9	101886789	1.387	1.070	1.061	1.246
10	14372818	1.387	1.141	1.307	1.246
11	54769157	1.387	0.606	0.601	1.247
12	23644776	1.347	1.267	1.151	1.207
13	11304963	1.323	1.150	0.944	1.464
14	205772	1.284	1.849	1.385	1.214
15	10625415	1.276	1.202	1.340	1.136
16	11753295	1.268	1.875	1.956	1.410
17	71654239	1.260	1.169	1.217	1.173
18	46855186	1.260	1.233	1.019	1.119
19	71653619	1.244	1.264	1.227	1.154
20	11455410	1.229	1.481	1.527	1.219
21	11329824	1.222	1.453	1.523	1.190
22	71653919	1.208	0.864	0.863	1.066
23	11169647	1.196	1.263	1.002	1.152
24	54577394	1.194	0.507	0.484	0.510
25	71653918	1.174	1.122	1.088	1.110
26	71654077	1.114	1.180	1.129	1.034
27	11191030	1.102	1.303	1.331	0.962
28	122199174	1.097	0.474	0.624	0.804
29	22023	1.097	0.655	0.748	0.956
30	118732688	1.097	0.884	0.823	0.980
31	10883396	1.097	0.664	0.591	0.956
32	71653620	1.086	1.031	1.026	1.113
33	1593398	1.081	1.031	1.189	1.137
34	44342760	1.046	0.630	0.534	0.618
35	54768918	1.041	1.378	1.149	1.085
36	71653761	1.036	0.886	0.933	1.136
37	23644778	1.032	1.278	1.268	1.173
38	23644783	1.009	1.355	1.224	1.150
39	54768917	0.996	1.407	1.328	1.137
40	54768922	0.996	0.876	0.672	0.927
41	71653617	0.963	0.834	0.834	1.104
42	71653760	0.921	1.321	1.274	1.061
43	52948017	0.921	0.224	0.164	0.424
44	118734712	0.907	0.281	0.247	0.766
45	44342769	0.886	0.457	0.330	0.263
46	118732691	0.886	1.017	0.805	0.547
47	24949525	0.863	1.146	1.204	0.924
48	118732692	0.796	1.175	0.936	0.654
49	9850331	0.785	1.390	1.174	1.076
50	24949524	0.724	1.059	1.132	0.864
51	118732690	0.721	0.990	0.824	0.677
52	9851127	0.712	0.564	0.469	0.570
53	101886794	0.706	0.699	0.604	0.555
54	16077296	0.699	0.621	0.799	0.839
55	122199187	0.699	0.272	0.412	0.684
56	24949523	0.676	0.856	0.904	0.817
57	118709901	0.642	1.077	0.664	0.783
58	101886788	0.592	0.525	0.472	0.467
59	9828345	0.590	1.232	1.033	0.977
60	101886796	0.588	0.912	0.830	0.580
61	102369422	0.573	0.338	0.211	0.431
62	118721872	0.542	0.554	0.456	0.682

(*Continued*)

Table 1. (*Continued*).

N°	PubChem CID	pIC_{50} (μm)	pIC_{50} (MLR)	pIC_{50} (MLP)	pIC_{50} (SVR)
63	9861757	0.538	0.422	0.372	0.391
64	23644777	0.506	0.350	0.233	0.490
65	102369423	0.484	0.941	0.680	0.526
66	118721629	0.483	0.552	0.493	0.343
67	72549439	0.456	0.399	0.344	0.219
68	52946279	0.444	0.447	0.193	0.584
69	44342592	0.398	0.739	0.573	0.429
70	1933	0.398	0.717	1.030	0.540
71	11067083	0.398	0.278	0.236	0.120
72	101886793	0.300	−0.053	0.100	0.159
73	90655199	0.292	−0.245	0.093	0.432
74	44342795	0.268	0.922	0.696	0.408
75	78350462	0.266	0.709	0.660	0.425
76	118732689	0.252	0.697	0.548	0.493
77	44224105	0.232	−0.189	0.126	0.092
78	56955111	3.097	2.745	3.014	2.726
79	71478328	3.056	2.543	2.443	2.869
80	16094856	3.014	2.323	2.512	2.874
81	11713155	3.000	2.733	3.038	2.860
82	56955109	3.000	2.889	3.184	2.859
83	90654212	2.921	2.980	3.004	2.864
84	71478270	2.883	3.190	3.193	2.868
85	16094874	2.860	2.508	2.725	2.719
86	56955100	2.824	2.359	2.507	2.491
87	90654211	2.785	2.918	2.738	2.926
88	90654203	2.690	2.925	3.016	3.001
89	56955199	2.602	2.418	2.564	2.461
90	9926968	0.208	0.218	0.210	0.314
91	118721626	0.197	0.199	0.229	0.200
92	90655197	0.190	0.738	0.512	0.332
93	11130919	0.161	0.532	0.417	0.281
94	44342503	0.155	0.772	0.575	0.319
95	9926905	0.155	0.580	0.434	0.295
96	10433216	0.148	0.322	0.241	0.139
97	118720935	0.130	−0.349	0.090	0.145
98	118709900	0.116	0.284	0.206	0.256
99	9882480	0.114	0.422	0.333	0.597
100	118709916	0.082	0.422	0.291	0.223
101	118721630	0.013	−0.048	0.086	0.130
102	118720934	0.008	0.021	0.154	0.148
103	52941983	0.000	−0.361	0.092	0.141
104	10049496	−0.004	0.556	0.284	0.137
105	44224109	−0.032	−0.082	0.126	0.109
106	72549683	−0.041	0.630	0.456	0.276
107	118721628	−0.107	0.296	0.322	0.235
108	44357449	−0.114	−0.471	−0.168	0.027
109	122199179	−0.146	−0.176	−0.176	−0.005
110	118709917	−0.152	0.324	−0.137	−0.011
111	70692913	−0.182	0.477	0.339	0.109
112	122199188	−0.204	−0.511	−0.440	−0.345
113	54577395	−0.220	0.260	0.128	0.174
114	44224101	−0.286	−0.556	0.053	−0.145
115	122199176	−0.380	−0.553	−0.436	−0.282
116	11352229	2.489	2.083	1.985	2.349
117	90654202	2.369	2.360	2.348	2.510
118	122199180	−0.623	−0.543	−0.464	−0.682
119	122199177	−0.716	−0.092	−0.042	−0.414
120	118709918	−0.750	−0.999	−0.531	−0.609
121	122199190	−0.833	−0.757	−0.596	−0.692
Validation set					
1	90654201	3.602	2.961	3.111	3.149
2	71478203	3.553	3.347	3.455	3.498
3	90654204	3.420	2.604	2.656	2.552

(*Continued*)

Table 1. (*Continued*).

N°	PubChem CID	pIC_{50} (μm)	pIC_{50} (MLR)	pIC_{50} (MLP)	pIC_{50} (SVR)
4	24800299	1.387	1.488	1.420	1.312
5	54768919	1.367	1.627	1.352	1.275
6	118721875	1.319	1.163	1.069	1.224
7	9549196	1.268	0.978	1.069	1.078
8	71654078	1.237	0.918	1.040	1.227
9	23644779	1.229	1.269	1.255	1.092
10	23644784	1.125	1.502	1.359	1.199
11	10322257	1.046	0.745	0.315	0.178
12	118721874	1.018	0.827	0.737	1.047
13	118732693	0.959	1.421	1.132	0.903
14	118734713	0.947	0.181	0.203	0.638
15	24881944	0.866	0.054	0.183	0.834
16	11341105	0.866	0.997	0.903	0.737
17	70684451	0.676	−0.086	0.079	0.154
18	44224107	0.662	−0.226	0.114	0.023
19	118721873	0.631	0.896	0.755	0.867
20	44224108	0.631	−0.231	0.111	0.059
21	24949376	0.529	0.355	0.284	0.365
22	90654213	3.108	3.131	2.896	2.787
23	56955099	2.824	2.691	2.917	2.414
24	90654208	2.684	2.598	2.794	2.599
25	118721871	0.130	0.336	0.288	0.229
26	122199184	−0.279	0.175	0.134	0.366
27	11504245	2.523	1.945	2.293	2.459
28	56955198	2.301	2.566	2.740	2.600
29	11385572	2.298	2.287	2.640	2.472
30	10711007	−0.602	−0.031	−0.018	−0.260

logarithmic scale [$pIC_{50} = -\log IC_{50}$ (μM)]. All compounds were optimized using the molecular mechanics (MM+) method to generate initial structures. In order to obtain a minimum of energy, a second geometrical optimization was carried out using a semi-empirical method at the AM1 level implemented in the MOPAC software [25].

Selection of relevant descriptors

The numerical representation of the structural characteristics of molecules, referred to as 'molecular descriptors', is an important step in obtaining a QSAR model. These descriptors play a fundamental role in the development of QSAR models. From the E-Dragon 1.0 online program (www.vcclab.org) and the MOE software [26], 2430 descriptors were calculated for each inhibitor. Given this large number of descriptors and to avoid the phenomenon of over-learning, it was necessary to carry out an effective preselection to keep only those which have the capacity to represent the essential determinants of the molecular interaction. The method used to select the most significant descriptors was described previously [27]. The number of descriptors obtained was 21. The selection of relevant descriptors by GAs was then performed using the BuildQsar software [28]. This second selection allowed us to obtain eight relevant descriptors, which were used to develop the predictive models.

Model development

The dataset was split into two sets: learning and test sets, with 80% and 20% of the total number of compounds, respectively. The learning set was dedicated to develop the QSAR model. The test set, which included tacrine derivatives that had not been used for the development of the QSAR models, was left for the external validation.

For the development of QSAR models, three different approaches were used: MLR [29], SVR and multilayer perceptron network (MLP). The MLR was conducted using the MLR plus Validation1.3 software (http://dtclab.webs.com/softwaretools or http://teqip.jdvu.ac.in/QSAR_Tools/). The SVR and the MLP were performed using the STATISTICA software (STATISTICA 8.0, StatSoft, Inc., Tulsa, OK, USA).

Statistical methods

MLR

The linear model was developed by applying MLR. The MLR technique has proved to be a multidisciplinary technique of valuable applicability for establishing predictive QSAR models by different research groups [19]. An MLR model is constructed under the assumption that a linear relationship exists between a set of molecular descriptors of a compound (which is represented by a feature vector (X) with each descriptor as its component) and a specific activity (y) [30], as shown in

$$y = A_0 + \sum_{k=1}^{N} A_k X_k \tag{1}$$

where y is the dependent variable, X_k represents molecular descriptors, A_k represents the coefficients of those descriptors and A_0 is the intercept of the equation. MLR calculations were performed using MLR plus Validation1.3 from the following link: (http://dtclab.webs.com/softwaretools or http://teqip.jdvu.ac.in/QSAR_Tools/)

SVR

SVMs [31, 32] are learning machines implementing the structural risk minimization (SRM) inductive principle to obtain good generalization on a limited number of learning patterns recognition (classification). The SRM principle has been shown to be superior to the traditional empirical risk minimization (ERM) principle employed by conventional neural networks. SRM minimizes an upper bound of the generalization error on Vapnik–Chernoverkis (VC) dimension ('generalization error'), as opposed to ERM that minimizes the training error. So SVM is usually less vulnerable to the overfitting problem.

With the introduction of the ε-insensitive loss function, the SVM has been extended to solve nonlinear regression estimation. Considering the problem of approximating the set of data $G = \left\{ (x_i, d_i) \right\}_{i=1}^{n}$; x_i is the input vector, d_i is the desired value and n is the total number of data patterns). In SVM method, the regression function is approximated, in a feature space F, by the following function:

$$f(x) = w\varphi(x_i) + b \tag{2}$$

where w is a vector in F and $\varphi(x_i)$ maps the input x to a vector in F. The coefficients w and b are estimated by minimizing

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, f(x_i)) + \frac{1}{2}||w^2|| \tag{3}$$

$$L_{\varepsilon}\left(d_i, f(x)\right) = \begin{cases} |d - f(x)| \geq \varepsilon \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

In equation (3), R is the regularized risk function and the first term $C\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon}(d_i, f(x_i))$ is the empirical error (risk). They are measured by the ε-insensitive loss function (L_{ε}) given by equation (4). The second term $\frac{1}{2}||w||^2$ is the regularization term. ε is called the tube size; it is equivalent to the approximation accuracy placed on the training data points, so that if predicted value is within the tube, the loss is zero, while if predicted point is outside the tube, the loss is the magnitude of the difference between the predicted value and the radius ε of the tube. C is the penalty parameter; it determines the trade-off between the empirical risk and the regularization term. Both C and ε are user-prescribed parameters. Introduction of Lagrange multipliers (a_i, a_i^*) and exploitation of the optimality constraints lead to the decision function given by

$$f\left(x, a_i, a_i^*\right) = \sum \left(a_i - a_i^*\right) K\left(x, x_i\right) + b \tag{5}$$

Based on the Karush–Kuhn–Tucker (KKT) conditions of quadratic programming [33], the kernel function K corresponds to $K(x, x_i) = \varphi(x)\varphi(x_i)$. Any function that satisfies the Mercer's condition [34, 35] can be used as the kernel function. The type of kernel function and its parameter are defined by the user. In SVR, the Gaussian radial basis function (RBF) is commonly used, and is defined in

$$K\left(x_i, x_j\right) = \exp\left(\frac{-x_i - x_j^2}{2\gamma^2}\right) \tag{6}$$

where γ represents the kernel width. The overall performance of SVR was evaluated in terms of root mean-square error (RMSE), which was calculated from the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{\text{obs}} - y_{\text{pred}})^2}{n_s}} \tag{7}$$

where y_{obs} is the desired output, y_{pred} is the value predicted by the model and n_s is the number of the molecules in dataset.

MLP/ANN

The ANN is an information-processing model that is inspired by biological nervous systems, such as the brain; it mimics the brain process of information processing. There are different models of neural networks available, each with its specific properties and benefits for particular applications. One of the most successful and most popular is the feed-forward MLP [36]. The structure of an MLP consists of an input layer (it corresponds to the independent variables: eight relevant descriptors), an intermediate or hidden layer and an output layer corresponding to the dependent variable (IC_{50}). Each layer can have a number of neurons, which are linearly connected by weights to the neurons in the neighbouring layers.

Validation of the model

Validation is an important and necessary step to test the performance and robustness of models. There are several validation approaches including internal validation and external validation. Recent studies [37, 38] have indicated that internal validation is essential for the validation of a QSAR model. The most important statistical parameters used in our study to check the performance of the models are the values of the coefficient of determination (r^2), the cross validated correlation coefficient $\left(Q^2\right)$, $\bar{r}_m^2$, Δr_m^2 values for the training and test sets. A value of $Q^2 > 0.5$ is considered satisfactory and a Q^2 value > 0.9 is excellent [39]. The values of both $r^2_{m(\text{test})}$ for the training set and r_m^2 for the test set should be higher than 0.5. Thus, Δr_m^2 values should be checked for both training and test sets. The values of $\Delta r^2_{m(\text{LOO})}$ and $\Delta r^2_{m(\text{test})}$ should be lower than 0.2. The metrics parameters introduced by Roy et al. [40] may be computed by the following equations:

$$\overline{r_m^2} = \frac{(r_m^2 + r_m'^2)}{2}$$

$$\Delta r_m^2 = \left| r_m^2 - r_m'^2 \right|$$

$$r_m^2 = r^2(1 - \sqrt{\left(r^2 - r_0^2\right)}$$

$$r_m'^2 = r^2\left(1 - \sqrt{\left(r^2 - r_0'^2\right)}\right)$$

r^2 and r_0^2 are the squared correlation coefficients between the observed and (leave-one-out) predicted values of the compounds with and without intercept, respectively. The parameter $r_0'^2$ has the same meaning but uses the reversed axes. The r_m^2 is the average value of r_m^2, and Δr_m^2 is the absolute difference between r_m^2 and $r_m'^2$.

Furthermore, the external validation is a significant and necessary validation method used to determine both the generalizability and the true predictive ability of the QSAR models. Furthermore, we used the concordance correlation coefficient (CCC) [41, 42], which measures both precision and accuracy, detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. Any deviation of the regression line from the concordance line (a line passing through the origin) gives a value of CCC smaller than 1.

Results and discussion

Selection of relevant descriptors

As the number of tacrine derivatives in the training set was 121, it was important to reduce the number of descriptors until the ratio 'number of tacrine derivatives/descriptors' was ≥ 5 [43]. After using the GA, eight descriptors were selected: ATSC5e, MATS5s, GATS7c, SpMax3_Bhm, SpMin1_Bhi, RotBFrac, SRW9 and vsa_don.

The fifth principle of the OECD [22] guidelines requires that a QSAR model should be mechanistically interpretable. In the eight selected descriptors, there are three 2D autocorrelation descriptors, two Burden modified eigenvalue descriptors, one PaDEL rotatable bond count descriptor, one walk count descriptor and one pharmacophore atom type descriptor calculated by MOE.

The variance inflation factor (VIF) analysis was performed to check the inter-correlation of the eight selected descriptors by MLR model [44]. The corresponding VIF values of the eight descriptors are shown in Table 2. As can be seen from this table, all variables had VIF values less than 5, indicating that the obtained model has obvious statistical significance.

In addition, for examination of the relative importance and contribution of each descriptor in the QSAR models, the value of the mean effect (MF) was calculated for each descriptor. Its sign displays the variation direction in the values of the activities as a result of the increase or decrease of this descriptor value. The calculated values of the MF for all descriptors are listed in Table 2.

ATSc5e is a 2D Moreau–Broto autocorrelation descriptor weighted by partial charges, i.e. an indicator of spatial partial charge association and its MF had a negative sign, which indicates that an increase in spatial partial charge in a molecule leads to decrease the activity of the molecule. MATS5s is a 2D Moran autocorrelation weighted by the intrinsic state molecular pseudo connectivity indices. Thus increasing the size of the molecule increases its MATS5s value. The MF of MATS5s has a positive sign, which indicates that an increase in ATSc5s leads to increase the activity of the molecule. GATS7c is a 2D Geary autocorrelation weighted by charges and its MF had a positive sign, which indicates that an increase in GATS7c leads to increase the activity of the molecule.

Burden modified eigenvalues are the eigenvalues of a modified connectivity matrix associated with atomic properties such as relative atomic mass, polarizability, Sanderson electronegativity and van der Waals (VDW) volume. The largest absolute eigenvalue of Burden that is weighted by relative mass (SpMax3_Bhm) and smallest absolute eigenvalue of Burden that is weighted by relative first ionization potential (SpMin1_Bhi) showed a positive MF contribution towards inhibitory activities of tacrine derivatives. Hence, the increased optimal level of the descriptor values in a molecule could improve inhibitory potency towards BuChE. RotBtFrac (fraction of rotatable bonds, excluding terminal bonds) belongs to the 2D PaDEL rotatable bond count descriptor. This descriptor indicates the number of rotatable bonds. RotBtFrac is the number of bonds in the molecule having rotations that are considered to be

Table 2. List of descriptors used in the development of different models.

Descriptor	*Description*	*VIF*	*MF*
ATSC5e	Centred Broto–Moreau autocorrelation – lag 5 / weighted by Sanderson electronegativities	1.759	–0.002
MATS5s	Moran autocorrelation – lag 5 / weighted by I-state	2.495	0.010
GATS7c	Geary autocorrelation – lag 7 / weighted by charges	1.836	0.090
SpMax3_Bhm	Largest absolute eigenvalue of Burden modified matrix – n 3 / weighted by relative mass	1.665	0.133
SpMin1_Bhi	Smallest absolute eigenvalue of Burden modified matrix – n 1 / weighted by relative first ionization potential	1.313	0.860
RotBFrac	Fraction of rotatable bonds, excluding terminal bonds	1.679	–0.062
SRW9	Self-returning walk count of order 9	1.368	–0.002
vsa_don	Approximation to the sum of VDW surface areas of pure hydrogen bond donors	1.308	–0.028

meaningful for molecular mechanics. All terminal H atoms are ignored. The RotBtFrac MF has a negative sign. This sign suggests that the activity is indirectly related to the descriptor.

Self-returning walk can be associated with possible electron movements. The larger the number of SRWi for a specific atom, the larger its fractional electronic charge. As ATSc5e, SRW9 had an MF with a negative sign; an increase in SRW9 leads to an increase in the activity of the molecule. vsa_don (approximation to the sum of VDW surface areas of pure hydrogen bond donors) belong to the 2D MOE describing pharmacophoric features and the total molecular surface area of pure hydrogen bond donors. Its MF had a negative sign, which indicates that an increase in total molecular surface area of pure hydrogen bond donors in a molecule leads to decrease the activity of the molecule.

According to the above discussion, it was concluded that the shape and atomic masses of molecules (MATS5s, SpMax3_Bhm and RotBtFrac), as well as the relative negative charge, the atomic polarizability and the atomic electronegativity of molecules (ATSc5e, GATS7c, SpMin1_Bhi, SRW9, and vsa_don), play an important role in the activity of tacrine derivatives.

MLR predictive model

The model obtained for the prediction of IC_{50} of tacrine derivatives, using 151 compounds, is given by the following linear model with the reported statistical parameters:

$$\begin{aligned} pIC_{50} = {} & 30.20351(\pm 4.94621) - 0.10371(\pm 0.05156)ATSC5e \\ & + 3.48641(\pm 0.48069)MATS5s - 3.50392(\pm 0.25099)GATS7c - 1.05831(\pm 0.30975)SpMax3_Bhm12.42265 \\ & (\pm 2.6109)SpMin1_Bhi + 7.56893(\pm 0.5521)RotBFrac0.02208(\pm 0.01261)SRW9 + 0.0691(\pm 0.00657)vsa_don \end{aligned} \tag{8}$$

with $n_{training} = 121$, $r^2 = 0.879$, $r^2_{adj} = 0.870$, $F = 101.462$, $p < 0.0001$, $Q^2 = 0.857$, PRESS = 16.125
$n_{test} = 30$, $r^2 = 0.847$, $r_0^2 = 0.828$, r_m^2 and ${}^c r_p^2 = 0.849$.

The standard errors of regression coefficients are given in parenthesis. The eight relevant descriptors (variables) in equation (8) could explain 87% of the variance (adjusted coefficient of variation) of the IC_{50}. The difference between r^2 and Q^2 was equal to 0.022. This difference was less than 0.3 indicating the robustness of the model [39]. Y-randomization was performed in order to ensure the robustness of the developed QSAR model. The degree of variation in the values of the squared mean correlation coefficient of the randomized model r_r^2 and squared correlation coefficient of the nonrandom model (r^2) was reflected in the value of the ${}^c r_p^2$ parameter [45]. The value of ${}^c r_p^2$ (0.847) was greater than 0.5 (threshold value). An acceptable value of r_m^2 (0.787) was obtained. The predictive results from the MLR model for the entire dataset (151 compounds) were obtained and presented in Table 1.

MLP predictive model

The sigmoid function and the identity function were used as transfer functions of the input layer to the hidden layer and the hidden layer to the output layer, respectively. The network was formed using the algorithm of quasi-Newton methods of BFGS. The complete dataset (151 compounds) of tacrine derivatives was divided into two groups: a training set and a validation set consisting of 121 and 30 compounds, respectively. To optimize the number of nodes in the hidden layer, several calculations were performed with different numbers of

hidden nodes (1–10). The RMSE and Q^2 criteria were employed for the evaluation of the accuracy of both training and validating sets. The best model was chosen according to the minimum RMSE (0.0023) and the maximum Q^2 (0.911). Therefore, the final model in this work was chosen with an MLP architecture 8-2-1 with input and hidden biases. The predictive results from the MLP model for the entire dataset (151 compounds) were obtained and presented in Table 1.

To see the importance of each variable for the prediction of inhibitory activity of tacrine derivatives, a sensitivity analysis was conducted. This method, proposed by Garson then taken by Goh [46], provides a quantification of the relative importance of different inputs (variables) on the output of the neural network. The contribution of each of the descriptors in the MLP model was as follows: ATSC5e (11.04%), MATS5s (10.55%), GATS7c (19.35%), SpMax3_Bhm (4.83%), SpMin1_Bhi 25.91%), RotBFrac (0.57%), SRW9 (2.477%) and vsa_don (25.27%).

SVR predictive model

In the SVR algorithm, three parameters must be optimized, the regularization constant C, ε in the ε-insensitive function and the Gaussian function parameter γ. In order to adjust these parameters, a grid search was carried out based on leave-one-out cross-validation on the training set. The values of the C parameter where tested from 10 to 100 with incremental steps of 1, γ ranging from 0.01 to 2.5 with incremental steps of 0.01 and ε from 0.01 to 0.5 with incremental steps of 0.01. The best SVR model was obtained with γ equal to 2 (with C = 10 and ε = 0.06). The statistical parameters found for this model were RMSE = 0.197, r^2 = 0.969 and Q^2 = 0.964 for the training set, and r^2 = 0.906 and Q^2 = 0.891 for the test set. The predictive results from the SVR model for the entire dataset (151 compounds) were obtained and presented in Table 1.

Comparison of three statistical models

Validation of QSAR models is performed using internal validation and external validation. The statistical parameters for the three considered models are summarized in Table 3. All of them seemed satisfactory with good data fit and robustness (high r^2, r_m^2 and Q^2 values). It can however be noted that the SVR model appeared slightly better, particularly as regards robustness. The quality of the model was judged in terms of different external validation metrics. According to Roy et al. [40], when the predicted values deviate much from the observed values, the value of r_0^2 should be inferior to r^2. As it is therefore expected, the values of r_0^2 and r^2 in SVR model were quite equal, whereas r_0^2 looked inferior to r^2 in the other models.

Table 3. External criteria of the MLR, MLP and SVR models.

Parameters	*MLR*	*MLP*	*SVR*
r^2	0.847	0.8884	0.9066
Q^2	0.8052	0.895	0.8915
r_m^2	0.7856	0.8206	0.8718
r_0^2	0.8282	0.8731	0.9011
$\overline{r_m^2}$	0.7869	0.8352	0.8683
Δr_m^2	0.0848	0.0921	0.0224
CCC	0.9099	0.9355	0.9422

For an acceptable prediction, the value of $\Delta r^2_{m(\text{test})}$ should be less than 0.2, provided that the value of $r^2_{m(\text{test})}$ is more than 0.5; these thresholds are well satisfied for the three models, as it can be seen in Table 3.

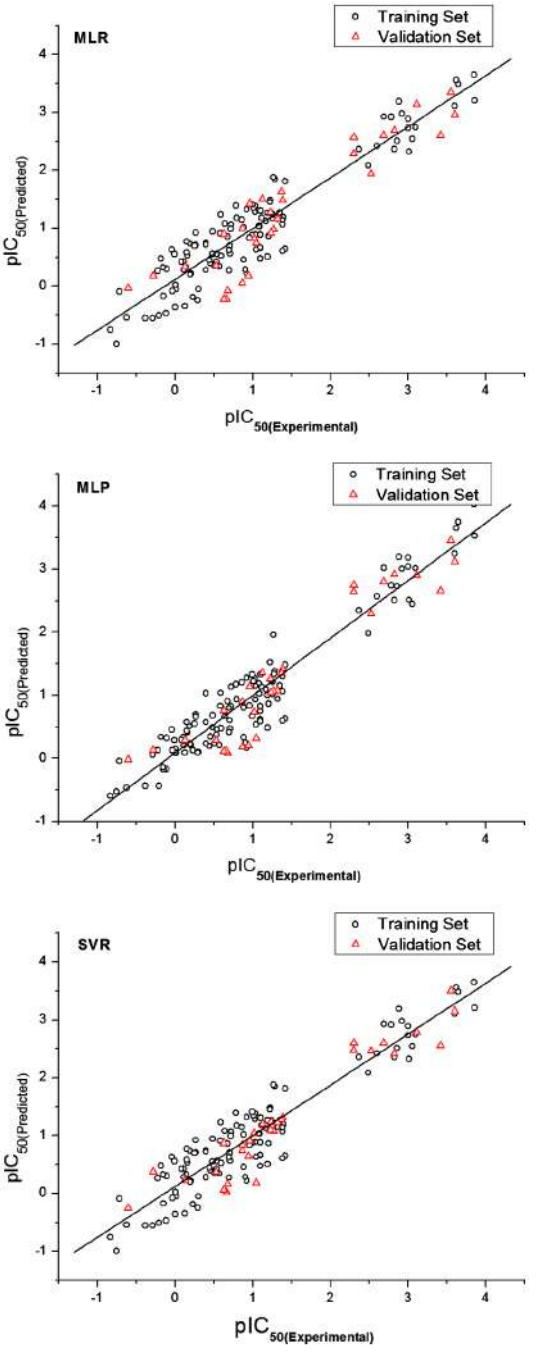


Figure 2. Plot of observed versus predicted pIC_{50} values from the MLR, MLP and SVR models.

Table 4. Comparison of AD methods applied to the test set.

Approach	*Options*	*Test inside AD*	*Test outside AD*	*List of samples outside AD*
Bounding box	—	30	0	None
Leverage	—	28	2	11 and 23
Euclidean distance (95 percentile)	Euclidean	30	0	None
Classical kNN (Euclidean dist., $k = 5$)	Euclidean	29	1	10
kNN (Euclidean dist., $k = 23$)	Euclidean	30	0	None
Standardization approach	—	30	0	None

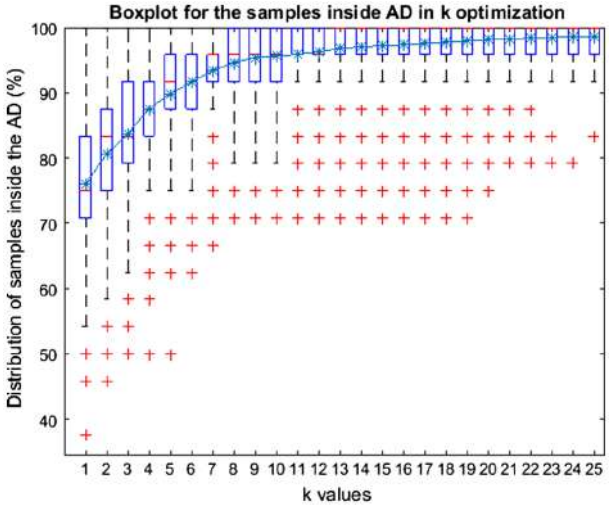


Figure 3. Box-and-whisker plot of test samples (%) retained within the applicability domain for different k values.

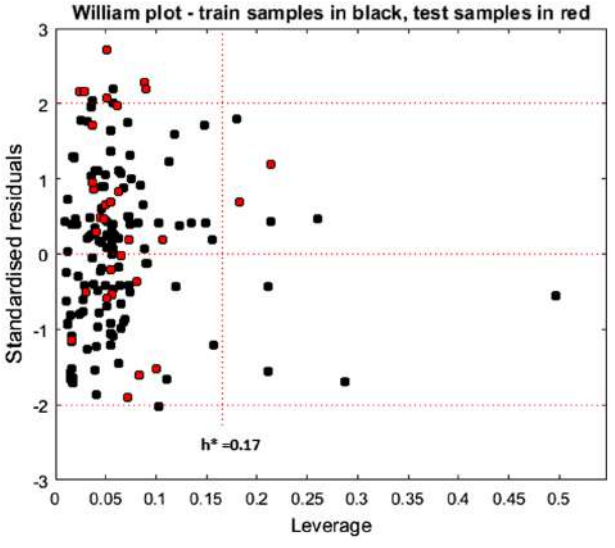


Figure 4. Plot of the residuals for calculated values of pIC_{50} from the SVR model versus their experimental values for the training and validation sets.

Additionally, the performance of the external validation criteria is evaluated by verifying the discrepancy from reference values. Accordingly, Figure 2 shows a close correlation between the values predicted by the MLR, MLP and SVR models and the observed values of IC_{50}. The great precision and accuracy was verified with CCC values; this coefficient measures both precision (how far the observations are from the fitting line) and accuracy (how far the regression line deviates from the slope line passing through the origin, i.e. the concordance). It can be noted that the points in the three models were almost perfectly aligned on the fitting lines while the accuracy increased gradually from MLR to SVR, suggesting that CCC well reflects the data point distribution in the SVR model. It can be concluded that the SVR model has better predictive power than the MLR and MLP models. This means that the models obtained allow the establishment, to a large extent, of a nonlinear relationship between the therapeutic activity of tacrine derivatives and the structural information (selected molecular descriptors).

Applicability domain

The third principle of the OECD [22] guidelines requires a defined domain of applicability. In this work, the applicability domain was analysed using different approaches: bounding box, leverage, Euclidean distance (95 percentile), classical kNN, modified kNN and standardization approach. The background theory, the algorithm and methodology and the advantages of the proposed approaches are available in the literature [47–49]. Table 4 reports these results; the number of samples retained within the applicability domain varied significantly depending on what strategy was used: bounding box, Euclidean distance (95 percentile), modified kNN (Euclidean distance and $k = 23$) and standardization approach, considered none of the test samples outside the applicability domain, while the classical kNN method identified one compound outside the applicability domain, as shown in Figure 3. Williams plot (Figure 4) shows two compounds from the test set that are identified as outside the applicability domain. These results show that the models comply with the third principle of the OECD. They can be used to predict the inhibitory activity of 151 tacrine derivatives, particularly for those that have not been tested as well as new compounds.

Conclusion

In this study, three linear and nonlinear QSAR models were developed to predict the inhibitory activity of 151 tacrine derivatives according to their molecular structure, represented by eight relevant descriptors. The experimental data was rigorously selected from the BindingDB database. The three models developed showed a good predictive ability. To select the best model, the metric parameters were used. According to some authors, this parameter appears to be advantageous over other internal and external validation parameters, since it relies on a prediction of the test and drive compounds and therefore implies more compounds in the process of prediction. The best model was that established with the SVR algorithm. The models developed in this work showed the importance of molar refractivity, the structure of the molecule and the electronic properties in the contribution to the inhibitory activity of the derivatives of tacrine. The models developed can be useful both in the field of designing potent inhibitors of BuChE and in predicting the potential for BuChE inhibition of new drug candidates.

References

[1] H. Boulebd, L. Ismaili, H. Martin, A. Bonet, M. Chioua, J. Marco Contelles, and A. Belfaitah, *New (benz) imidazolopyridino tacrines as nonhepatotoxic, cholinesterase inhibitors for Alzheimer disease*, Future Med. Chem. 9 (2017), pp. 723–729.

[2] I. McDowell, *Alzheimer's disease: Insights from epidemiology*, Aging Clin. Exp. Res. 13 (2001), pp. 143–162.

[3] T. Arendt, M.K. Brückner, M. Lange, and V. Bigl, *Changes in acetylcholinesterase and butyrylcholinesterase in Alzheimer's disease resemble embryonic development: A study of molecular forms*, Neurochem. Int. 21 (1992), pp. 381–396.

[4] E.K. Perry, R.H. Perry, G. Blessed, and B.E. Tomlinson, *Changes in brain cholinesterases in senile dementia of Alzheimer type*, Neuropathol. Appl. Neurobiol. 4 (1978), pp. 273–277.

[5] A. Enz, R. Amstutz, H. Boddeke, G. Gmelin, and J. Malanowski, *Brain selective inhibition of acetylcholinesterase: A novel approach to therapy for Alzheimer's disease*, Prog. Brain Res. 98 (1993), pp. 431–438.

[6] N.H. Greig, D.K. Lahiri, and K. Sambamurti, *Butyrylcholinesterase: An important new target in Alzheimer's disease therapy*, Int. Psychogeriatr. 14 (2002), pp. 77–91.

[7] P.W. Elsinghorst, C.M. Tanarro, and M. Gütschow, *Novel heterobivalent tacrine derivatives as cholinesterase inhibitors with notable selectivity toward butyrylcholinesterase*, J. Med. Chem. 49 (2006), pp. 7540–7544.

[8] S. Hamulakova, L. Janovec, M. Hrabinova, K. Spilovska, J. Korabecny, P. Kristian, K. Kuca, and J. Imrich, *Synthesis and biological evaluation of novel tacrine derivatives and tacrine–coumarin hybrids as cholinesterase inhibitors*, J. Med. Chem. 57 (2014), pp. 7073–7084.

[9] J.L. Marco, C. de los Ríos, M.C. Carreiras, J.E. Baños, A. Badía, and N.M. Vivas, *Synthesis and acetylcholinesterase/butyrylcholinesterase inhibition activity of new tacrine-like analogues*, Bioorg. Med. Chem. 9 (2001), pp. 727–732.

[10] S. Thiratmatrakul, C. Yenjai, P. Waiwut, O. Vajragupta, P. Reubroycharoen, M. Tohda, and C. Boonyarat, *Synthesis, biological evaluation and molecular modeling study of novel tacrine-carbazole hybrids as potential multifunctional agents for the treatment of Alzheimer's disease*, Eur. J. Med. Chem. 75 (2014), pp. 21–30.

[11] K. Roy, S. Kar, and R.N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, San Diego, CA, 2015.

[12] S. Simeon, N. Anuwongcharoen, W. Shoombuatong, A.A. Malik, V. Prachayasittikul, J.E.S. Wikberg, and C. Nantasenamat, *Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking*, PeerJ 4 (2016), e2322.

[13] K. Roy and P.P. Roy, *Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques*, Eur. J. Med. Chem. 44 (2009), pp. 2913–2922.

[14] N. Akula, L. Lecanu, J. Greeson, and V. Papadopoulos, *3D QSAR studies of AChE inhibitors based on molecular docking scores and CoMFA*, Bioorg. Med. Chem. Lett. 16 (2006), pp. 6277–6280.

[15] M. Fernández, M.C. Carreiras, J.L. Marco, and J. Caballero, *Modeling of acetylcholinesterase inhibition by tacrine analogues using Bayesian-regularized genetic neural networks and ensemble averaging*, J. Enzyme Inhib. Med. Chem. 21 (2006), pp. 647–661.

[16] M. Saracoglu and F. Kandemirli, *The investigation of structure–activity relationships of tacrine analogues: Electronic-topological method*, Open Med. Chem. J. 2 (2008), pp. 75–80.

[17] M. Jung, J. Tak, Y. Lee, and Y. Jung, *Quantitative structure–activity relationship (QSAR) of tacrine derivatives against acetylcholinesterase (AChE) activity using variable selections*, Bioorg. Med. Chem. Lett. 17 (2007), pp. 1082–1090.

[18] D. Chekmarev, V. Kholodovych, S. Kortagere, W.J. Welsh, and S. Ekins, *Predicting inhibitors of acetylcholinesterase by regression and classification machine learning approaches with combinations of molecular descriptors*, Pharm. Res. 26 (2009), pp. 2216–2224.

[19] K.Y. Wong, A.G. Mercader, L.M. Saavedra, B. Honarparvar, G.P. Romanelli, and P.R. Duchowicz, *QSAR analysis on tacrine-related acetylcholinesterase inhibitors*, J. Biomed. Sci. 21 (2014), p. 84.

[20] M.S. Castilho, R.V.C. Guido, and A.D. Andricopulo, *Classical and hologram QSAR studies on a series of tacrine derivatives as butyrylcholinesterase inhibitors*, Lett. Drug Des. Discov. 4 (2007), pp. 106–113.
[21] J. Fang, R. Yang, L. Gao, D. Zhou, S. Yang, A.L. Liu, and G.H. Du, *Predictions of BuChE inhibitors using support vector machine and naive bayesian classification techniques in drug discovery*, J. Chem. Inf. Model. 53 (2013), pp. 3009–3020.
[22] OECD, *Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q) SAR] Models*, Series on Testing and Assessment N° 69, OECD, Paris, 2007.
[23] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, and M.K. Gilson, *BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities*, Nucleic Acids Res. 35 (2007), D198–D201.
[24] G.L. Ellman, K.D. Courtney, V.Jr. Andres, and R.M. Feather-stone, *A new and rapid colorimetric determination of acetylcholinesterase activity*, Biochem. Pharmacol. 7 (1961), pp. 88–95.
[25] A.A. Toropov and A.P. Toropova, *The index of ideality of correlation: A criterion of predictive potential of QSPR/QSAR models?*, Mutat. Res. 819 (2017), pp. 31–37.
[26] MOE, *Molecular Operating Environment*, Chemical Computing Group, Montreal, Canada, 2007.
[27] S. Bitam, M. Hamadache, and S. Hanini, *QSAR model for prediction of the therapeutic potency of N-benzylpiperidine derivatives as AChE inhibitors*, SAR QSAR Environ. Res. 28 (2017), pp. 471–489.
[28] D.B. de Oliveira and A.C. Gaudio, *BuildQSAR: A new computer program for QSAR analysis*, Mol. Inform. 19 (2000), pp. 599–601.
[29] G. Snedecor and W. Cochran, *Statistical Methods*, 6th ed., Oxford and IBH Publishing Co, New Delhi, 1967.
[30] C.W. Yap, H. Li, Z.L. Ji, and Y.Z. Chen, *Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties*, Mini Rev. Med. Chem. 7 (2007), 1097–1107.
[31] C.J.C. Burges, *A tutorial on support vector machines for pattern recognition*, Data Min. Knowl. Discov. 2 (1998), pp. 121–167.
[32] V.N. Vapnik and S. Kotz, *Estimation of Dependences Based on Empirical Data*, Springer Series in Statistics, Springer-Verlag, New York, 1982.
[33] J.C.G. Boot, *Quadratic Programming: Algorithms, Anomalies, Applications*, North-Holland, Amsterdam, 1964.
[34] B.E. Boser, I.M. Guyon, and V.N. Vapnik, *A training algorithm for optimal margin classifiers, in Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ACM, Pittsburgh, PA, 1992, pp. 144–152.
[35] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, 2013.
[36] L.V. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*, Prentice-Hall, Upper Saddle River, NJ, 1994.
[37] R. Wang, J. Jiang, Y. Pan, H. Cao, and Y. Cui, *Prediction of impact sensitivity of nitro energetic compounds by neural network based on electrotopological-state indices*, J. Hazard Mater. 166 (2009), pp. 155–186.
[38] P.K. Ojha, I. Mitra, R.N. Das, and K. Roy, *Further exploring rm2 metrics for validation of QSPR models*, Chemom. Intell. Lab. 107 (2011), pp. 194–205.
[39] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs*, Environ. Health Perspect. 111 (2003), pp. 1361–1375.
[40] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, and H. Kabir, *Comparative studies on some metrics for external validation of QSPR models*, J. Chem. Inf. Model. 52 (2012), pp. 396–408.
[41] N. Chirico, and P. Gramatica, *Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient*, J. Chem. Inf. Model. 51 (2011), pp. 2320–2335.
[42] L.I. Lin, *A concordance correlation coefficient to evaluate reproducibility*, Biometrics 45 (1989), pp. 255–268.
[43] A. Tropsha, P. Gramatica, and V.K. Gombar, *The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models*, Mol. Inform. 22 (2003), pp. 69–77.

[44] V.K. Agrawal and P.V. Khadikar, *QSAR prediction of toxicity of nitrobenzenes*, Bioorg. Med. Chem. 9 (2001), pp. 3035–3040.
[45] I. Mitra, A. Saha, and K. Roy, *Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants*, Mol. Simul. 36 (2010), pp. 1067–1079.
[46] M. Hamadache, S. Hanini, O. Benkortbi, A. Amrane, L. Khaouane, and C.S. Moussa, *Artificial neural network-based equation to predict the toxicity of herbicides on rats*, Chemom. Intell. Lab. 154 (2016), pp. 7–15.
[47] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, and R. Todeschini, *Comparison of different approaches to define the applicability domain of QSAR models*, Molecules 17 (2012), pp. 4791–4810.
[48] F. Sahigara, D. Ballabio, R. Todeschini, and V. Consonni, *Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions*, J. Cheminform. 5 (2013), p. 27.
[49] K. Roy, S. Kar, and P. Ambure, *On a simple approach for determining applicability domain of QSAR models*, Chemom. Intell. Lab. 145 (2015), pp. 22–29.