الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
ⵓⵖⵏⵄⵎ ⵏ ⵜⵙⵏⴰⵡⵉⵏ ⵜⵉⵙⵄⵔ ⴰ ⵜⵍⴰⵄ ⵜⵙⴰⵡⴰⵍ
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DE MEDEA

Département de Génie des Procédés et Environnement

Laboratoire de Biomatériaux et Phénomènes de Transport

THESE DE DOCTORAT

Présentée par

KHAOUANE Affaf

A L'Université de Médéa

Pour l'obtention du titre de Docteur en Génie Pharmaceutique

Filière : Génie des Procédés Pharmaceutiques

Prédiction des Paramètres Pharmacocinétiques des Médicaments par les RNA-QSAR

Membres du Jury

Pr. Maamar LAIDI	Professeur, Université de Médéa	Président
Pr. Yacine KERCHICH	Professeur, ENP. Alger	Examinateur
Dr. Redha ALOUAOUI	Maître de Conférences (A), Université de Médéa	Examinateur
Dr. Hajira KERMET SAID	Maître de Conférences (A), Université de Médéa	Examinatrice
Pr. Salah HANINI	Professeur, Université de Médéa	Invité
Pr. Samira FERHAT	Professeur, Université de Médéa	Directrice de thèse

Année 2022/2023

Laboratoire de Biomatériaux
et Phénomènes de Transport

THÈSE

Dédicaces

A *mes chers parents,*

A *ma sœur Latifa,*

A *mon frère Djaafer*

A *tous ceux qui m'ont soutenu de près ou de loin.*

Affaf.

Remerciements et gratitude

Cette thèse a été réalisée sous la direction du professeur FERHAT Samira. Je tiens à lui exprimé mes plus sincères remerciements pour son accueil chaleureux et pour la confiance qu'elle m'a accordée. Je lui suis aussi reconnaissante pour les nombreux conseils avisés qu'elle m'a prodiguée au cours de longues discussions.

Je voudrais ensuite remercier, parce que la tradition l'exige certes, mais surtout parce qu'il le mérite, Monsieur **Salah HANINI**, *Professeur à l'UYFM, qui m'a guidé avec grande patience tout au long de l'élaboration de ce travail, pour ses aides précieuses qui ont judicieusement éclairé mon chemin vers l'aboutissement et la réussite de la concrétisation de cette thèse de doctorat.*

Je tiens à exprimer mes plus vifs remerciements à madame **Latifa KHAOUANE**, *professeur à l'UYFM pour ses aides, son accompagnement et son éclairage techniques. On lui doit les remerciements les plus sincères.*

Je tiens également à présenter ma profonde gratitude au membre de jury d'avoir acceptés de juger ce modeste travail.

Merci à tous ceux qui m'ont aidé de près ou de loin durant ces années de recherche et d'études. Je ne peux conclure sans un mot de remerciement à tous mes enseignants et à tous mes collègues.

ملخص النمذجة بالمعلمات الدوائية للأدوية باستعمال QSAR-RNA

تلعب العلاقة الكمية للنشاط بالهيكل QSAR دورا هاما في العلوم الصيدلية. وكان الهدف الرئيسي من هذا العمل هو تطوير اربعة نماذج QSAR باستخدام الشبكة العصبية (RNA) للتنبؤ بارتباط سلسلة متكونة من 277 دواء بالبروتينات البلازمية (*fb.*) انطلاقا من البنية التركيبية للدواء تم إنشاء نماذج مختلفة من واصفات الأدوية (مجموعها 1666). ثم تم استخدام ثلاث طرق اختيار: طريقتان أوليان لاختيار أكثر الأوصاف صلة تسمى «Filter1» و «Wrapper» حيث تكون الأوصاف في شكل متجهات. قدمت تقنية الاختيار هذه 55 و 15 وصفًا على التوالي وطريقة ثالثة تعتبر فيها واصفاتها مصفوفات يتم إدخالها في شبكة عصبية من الالتفاف. يعتبر الأخير الإدخال بمثابة بكسلات للصورة ويسمح لنا باستخراج الخصائص الأساسية لهذه الصورة المزعومة تلقائيًا. لقد استخرجنا متجهين كاراكتيريين بنفس حجم تلك التي تم الحصول عليها لطرق Filter وWrapper من أجل إجراء مقارنة. واستخدمت استراتيجيات ومعايير التصديق المتاحة للتحقق من صحة طريقة الاختيار الجديدة.. غير أن أداء النموذج كالCN-FC15-FFNN كان أفضل بكثير من أداء النموذج CN-FC55-FFNN. من الواضح أن هذا التحقيق أظهر أن استخراج الخصائص باستخدام CNN يستغرق وقتًا أقل بالإضافة إلى الاحتفاظ بجميع المعلومات الواردة في الخصائص الأصلية. على عكس تقنيات اختيار الميزات، لا يتعين علينا فهم الميزات التي تعد أفضل تمثيلات لنموذجنا، بل يتم تعلمها لنا.

كلمات المفاتيح: تصميم -الشبكة العصبية (RNA) -العلاقة الكمية للنشاط بالهيكل QSAR-واصفات-المعلمات الدوائية.

Résumé : Prédiction des paramètres Pharmacocinétique des médicaments par les RNA-QSAR

Les Relations Quantitative Structure Activité (QSAR) sont appelés à jouer un rôle important dans le domaine des sciences pharmaceutiques. L'objectif principal de ce travail a été le développement de quatre modèles QSAR utilisant un réseau de neurone feed-forward pour la prédiction de la fraction du médicament liée aux protéines plasmatiques (*fb*) d'une série de 277 médicaments. Différents types de descripteurs (1666 au total) ont été générées à partir des structures des médicaments. Puis, trois méthodes de sélection ont été utilisées : Deux premières méthodes de sélection des descripteurs les plus pertinents appelées 'Filter' et 'Wrapper' dans lesquelles les descripteurs sont sous forme de vecteurs. Cette technique de sélection a fourni 55 et 15 descripteurs respectivement et une troisième méthode dans laquelle on considère ses descripteurs comme étant des matrices qu'on fait entrer dans un réseau de neurones à convolution. Ce dernier considère l'entrée comme des pixels d'une image et nous permet d'extraire automatiquement les caractéristiques les plus essentiels de cette soi-disant image. Nous avons extrait deux vecteurs de caractéristiques de la même dimension que ceux obtenus pour les méthodes Filter et Wrapper afin de faire une comparaison. Les stratégies et critères de validation disponibles ont été utilisés pour valider la nouvelle méthode de sélection. Cette validation à révéler sa robustesse et fiabilité. Néanmoins, la performance du modèle CN-FC15-FFNN est sensiblement meilleure que celle du modèle CN-FC55-FFNN. IL est évident que cette enquête a montré que l'extraction de caractéristiques à l'aide de CNN prend moins de temps en plus de garder toutes les informations contenues dans les caractéristiques d'origine. Contrairement aux techniques de sélection de caractéristiques, nous n'avons pas à comprendre quelles caractéristiques sont les meilleures représentations pour notre modèle, elles sont apprises pour nous.

Mots clés : Modélisation, RNA, QSAR, $f_{b,}$ Descripteur, pharmacocinétique.

Summary: Prediction of pharmacokinetic parameters of drugs by RNA-QSAR

Quantitative Structure Activity Relationship (QSAR) models are expected to play an important role in the pharmaceutical sciences. The main aim of the present work was the development of four QSAR model using a feed-forward neural network (FFNN) for the prediction of fraction bound to plasma proteins (f_b) of a series of 277 drugs. Different kinds of descriptors (1666 in total) were generated from the drug structures. Then three selection methods were used: Two first methods for selecting the most relevant descriptors are called 'Filter' and 'Rapper' in which the descriptors are in the form of vectors. These selection techniques provided 55 and 15 descriptors respectively and a third method in which the descriptors are considered to be matrices that are entered as input into a convolutional neural network. The latter considers the entry as pixels of an image and allows us to automatically extract the most essential characteristics of this so-called image. We extracted two characteristics vectors of the same size as those obtained for filter and wrapper methods in order to make a comparison. The available validation strategies and criteria were used to validate the new selection method. This validation has revived its robustness and reliability. However, the performance of the CN-FC15-FFNN model is significantly better than that of the CN-FC55-FFNN model. It is clear that this investigation has shown that the extraction of characteristics using CNN takes less time in addition to keeping all the information contained in the original characteristics. Unlike feature selection techniques, we do not have to understand which features are the best representations for our model, they are learned for us.

Keywords: Modélisation, RNA, QSAR, $f_{b,}$ Descriptors, pharmacokinetics.

Table des matières

Table des matières i
Abréviations iv
Liste des tableaux v
Liste des figures vii

Introduction générale 1

Chapitre I. Construction de modèles QSAR dans les méthodes de criblage in silico pour la découverte de nouveaux médicaments

I.1. Découverte de nouveaux médicaments 4
I.1.1. Historique 4
I.1.2. Processus général de R&D 6
I.1.2.1. Identification et validation des cibles 7
I.1.2.2. Identification de « hits » 8
I.1.2.3. Optimisation de « leads » 9
I.1.2.4. Essais précliniques 10
I.1.2.5. Essais cliniques et commercialisation 11
I.1.3. Processus de criblage virtuel 11
I.1.3.1. Intérêts thérapeutiques 11
I.1.3.2. Méthodes de criblages in silico 12
I.1.3.2.1. Criblage virtuel basé sur la structure 13
I.1.3.2.2. Criblage virtuel « ligand-based » 14
I.1.4. Limitations et échecs 14
I.2. Conception de médicament et la théorie QSAR 14
I.2.1. QSAR et Descripteurs 15
I.2.2. Réduction de la dimensionnalité 16
I.3. Les algorithmes de sélection des descripteurs les plus pertinents dans le Modèles de relations quantitatives structure-activité (QSAR) 17
I.3.1. Filter 18
I.3.2. Wrapper 19
I.3.3. Embedded 20
I.4. Construction de modèles QSAR et RNA 21
I.4.1. Le réseau de neurone artificiel (RNA) 21
I.4.2. Le Perceptron Multicouche (PMC ou MLP) 22
I.4.3. Les réseaux de neurones convolutifs 23
I.4.3.1. Architecture 23
I.4.3.1.1. Couches convolutives 24
I.4.3.1.2. Couche de mise en commun (pooling) 25
I.4.3.1.3. La fonction d'activation 25
I.4.3.1.4. Couches entièrement connectées 26
I.4.3.2. Alex Net et transfert d'apprentissage 26

Chapitre II. *Liaisons des médicaments aux protéines circulantes (PROTEIN BINDING)*

II.1. Introduction 33
II.2. Notion de la fixation protéique des médicaments 33

II.3. Protéines fixatrices 37
II.3.1. Albumine 38
II.3.2. Acide Alpha-1-glycoprotéine 39
II.3.3. Autres protéines 40
II.4. Paramètres de la fixation protéique des médicaments 40
II.4.1. Affinité 40
II.4.2. Sites de fixation 41
II.4.3. Pourcentage de fixation 41
II.5. Conséquences des modifications de la fixation protéique 42
II.5.1. Conséquences pharmacocinétiques 42
II.5.2. Conséquences cliniques 43
II.6. Méthodes analytiques pour l'étude de la fixation protéique des médicaments 43
II.6.1. La microdialyse 44
II.6.2. Précipitation 45
II.6.3. Ultrafiltration 46
II.6.4. Dialyse à l'équilibre 46
II.7. Applications des approches *in silico* dans l'étude da la fixation protéique des médicaments 48
II.8. Conclusion 53

Chapitre III. Développement des modèles QSAR pour la prédiction de la liaison des médicaments aux protéines plasmatiques

III.1. Introduction 58
III.2. Procédure de développement des modèles prédictives 59
III.2.1. Collecte des données de la liaison des médicaments aux protéines plasmatique 60
III.2.2. Calcul des descripteurs 61
III.2.3. Sélection des descripteurs les plus influents pour la méthode 'Filter' 62
III.2.4. Sélection des descripteurs les plus influents pour la méthode 'Wrapper' 63
III.2.5. Méthode CNN 63
III.2.6. Développement des modèles prédictives 63
III.2.7. Évaluation et amélioration des performances prédictives des modèles 64
III.2.7.1. Validation interne 64
III.2.7.2. Validation externe 66
III.2.8. Domaine d'applicabilité 67
III.3. Résultat et discussion 68
III.3.1**.** Analyse des descripteurs retenus 68
III.3.1.1. Méthode 'filter' 68
III.3.1.2. Méthode 'wrapper' 74
III.3.2. Conception architectural 76
III.3.2.1. Méthodes 'Filter' et 'Wrapper' 76
III.3.2.2. Méthode CNN 77
III.3.3. Performance du modèle neuronal 79
III.3.3.1. Méthode 'filter' et méthode 'Wrapper' 79
III.3.3.2. Méthode CNN 83
III.4. Validation et comparaison entre les quatre modèles 92
III.5. Domaine d'applicabilité des modèles 93
IV.6. Comparaison avec d'autres modèles parus dans la littérature 97
IV.4. Conclusion 99

Conclusion générale et perspectives 103

Abréviations

CNN-FC15-FFNN	**C**onvolutinal **N**eural **N**etwork-**F**ully **C**onnectad **L**ayer-**F**eed **F**orward **N**eural **N**etwork.
CNN-FC55-FFNN	**C**onvolutinal **N**eural **N**etwork-**F**ully **C**onnectad **L**ayer-**F**eed **F**orward **N**eural **N**etwork.
CPU	**C**entral **P**rocessing **U**nit.
DA	**D**omaine d'**A**pplicabilité.
f_b	la liaison aux protéines plasmatiques.
FFNN	**R**éseau de **N**eurones de type **F**eed **F**orward.
MAE	**M**ean **A**bsolute **E**rror.
MSE	**M**ean **S**quared **E**rror.
OCDE	**O**rganisation de **C**oopération et de **D**éveloppement **E**conomiques.
PMC	**P**erceptron **M**ulti-**C**ouche.
QSAR	**Q**uantitative **S**tructure-**A**ctivity **R**elationship.
QSPR	**Q**uantitative **S**tructure-**P**roperty **R**elationship.
R	Coefficient de correlation.
$\mathbf{R^2}$	Coefficient de determination.
RMSE	**R**oot **M**ean **S**quared **E**rror.
RNA	**R**éseaux de **N**eurones **A**rtificielle.
RSD	**D**éviation **S**tandard **R**elative.
SMILE	**S**implified **M**olecular Input **L**ine **E**ntry **S**pecification.
VIF	**F**acteur d'**I**nflation de la **V**ariance.

Liste des tableaux

Chapitre III. Développement des modèles QSAR pour la prédiction de la liaison des médicaments aux protéines plasmatiques

Tableau III.1	Nombre de descripteurs calculés.	62
Tableau III.2	Notation, définition, type et VIF des descripteurs sélectionnés.	70
Tableau III.3	Notation, définition, type et VIF des descripteurs sélectionnés.	75
Tableau III.4	Critères de sélection de l'architecture du FFNN pour la méthode 'filter'**.**	79
Tableau III.5	Critères de sélection de l'architecture du FFNN pour la méthode 'Wrapper'**.**	82
Tableau III.6	Critères de sélection de l'architecture du FFNN pour le modèle CNN-FC15-FFNN.	84
Tableau III.7	Critères de sélection de l'architecture du FFNN pour le modèle CNN- FC 55 – FFNN.	86
Tableau III.8	*VIF* des entrée sélectionnés pour le modele'CNN-FC55-FFNN'.	91
Tableau III.9	*VIF* des entrée sélectionnés pour le modele'CNN-FC15-FFNN'.	91
Tableau III.10	Validation interne et externe pour les quatre modèles.	93
Tableau III.11	Domaine d'applicabilité du modèle CNN –FC15-FFNN.	94
Tableau III.12	Domaine d'applicabilité du modèle CNN–FC55-FFNN.	96
Tableau III.13	Comparaison avec la littérature.	98

Liste des figures

Chapitre I. Construction de modèles QSAR dans les méthodes de criblage in silico pour la découverte de nouveaux médicaments

Figure I.1	Quelques grandes étapes de la découverte de médicaments.	5
Figure I.2	Etapes du processus de découverte de médicaments.	6
Figure I.3	Méthodes utilisées pour la l'identification et la validation des cibles.	8
Figure I.4	Méthodes utilisées pour la l'identification et la validation HITS.	9
Figure I.5	Méthodes utilisées pour la l'optimisation des LEADS.	10
Figure I.6	Classification des méthodes de criblage virtuel « ligand-based » et « structurebased ».	13
Figure I.7	Illustration du processus de conception de modèles QSAR aboutissant à un modèle statistique.	16
Figure I.8	Technique de sélection des descripteurs.	18
Figure I.9	La procédure du modèle "Filter".	19
Figure I.10	La procédure du modèle "wrapper".	19
Figure I.11	Différentes catégories de RNA.	22
Figure I.12	Architecture du perceptron multicouche.	22
Figure I.13	Les principaux blocs de construction d'une architecture utilisant un réseau de neurones convolutif simple.	24
Figure I.14	Architecture d'Alex Net.	27

Chapitre II. Liaisons des médicaments aux protéines circulantes (PROTEIN BINDING)

Figure II.1	Les composants du sang chez un adulte normal.	35
Figure II.2	Distribution du médicament et liaison aux protéines plasmatiques.	36
Figure II.3	Structure tridimensionnelle de l'albumine et sites de liaison aux médicaments. Elle est composée de trois domaines (I-III) mis en évidence avec des couleurs différentes (bleu clair-domaine I ; vert- domaine II ; domaine rose III). Les autres sites de liaison interagissant avec les ions, les petites molécules et les peptides ne sont pas représentés.	38
Figure II.4	Méthodes séparatives et non séparatives utilisées pour évaluer la liaison des médicaments aux protéines plasmatiques.	44

Chapitre III. Développement des modèles QSAR pour la prédiction de la liaison des médicaments aux protéines plasmatiques

Figure III.1	Organigramme de la procédure suivie.	60
Figure III.2	Boite à moustache pour la répartition des valeurs de (f_b).	61
Figure III.3	Les étapes de la validation croisée Holdout..	65
Figure III.4	Heatmap qui représente les coefficients de corrélation entre les entrée du réseau de neurone pour le modèle 'Filter'.	69
Figure III.5	Heatmap qui représente les coefficients de corrélation entre les entrée du réseau de neurone pour le modèle 'Wrapper'.	74
Figure III.6	L'architecture du RNA optimisée.	77
Figure III.7	Architecture de la procedure 'Méthode CNN'.	78

Figure III.8	Représentation graphique de la valeur de la liaison des médicaments aux protéines plasmatiques expérimentales et celles prédites par le modèle 'Filter'.	80
Figure III.9	Histogramme d'erreur pour la méthode 'Filter'.	80
Figure III.10	Représentation graphique de la valeur de la liaison des médicaments aux protéines plasmatiques expérimentales et celles prédites par le modèle 'Wrapper'.	82
Figure III.11	Histogramme d'erreur pour la méthode 'Wrapper'.	83
Figure III.12	Représentation graphique de la valeur de la liaison des médicaments aux protéines plasmatiques expérimentales et celles prédites par le modèle 'CNN-FC15-FFNN'.	84
Figure III.13	Histogramme d'erreur pour le modèle 'CNN-FC15-FFNN'.	85
Figure III.14	Représentation graphique de la valeur de la liaison des médicaments aux protéines plasmatiques expérimentales et celles prédites par le modèle 'CNN-FC55-FFNN'.	86
Figure III.15	Histogramme d'erreur pour Le modèle 'CNN-FC55-FFNN'.	87
Figure III.16	Heatmap qui représente les coefficients de corrélation entre les entrée du réseau de neurone pour le modèle 'CNN-FC15-FFNN'.	89
Figure III.17	Heatmap qui représente les coefficients de corrélation entre les entrée du réseau de neurone pour le modèle 'CNN-FC55-FFNN'.	90
Figure III.18	Visualisation d'un échantillon d'image pour l'acide 3-carboxy-4-méthyl-5-propyl-2-furanpropionique obtenu.	92
Figure III.19	Diagramme de Williams pour les ensembles d'apprentissage et de test du modèle CNN –FC15-FFNN.	95
Figure III.20	Diagramme de Williams pour les ensembles d'apprentissage et de test du modèle CNN –FC55-FFNN.	96

Introduction générale



La liaison d'un médicament aux protéines plasmatiques, principalement à l'albumine sérique et α-acide glycoprotéine, est l'un des nombreux facteurs qui influencent l'ADME (Adsorption, distribution, métabolisme et excrétion) et la toxicité [1]. La liaison d'un médicament aux protéines plasmatiques réduit la fraction libre du médicament dans la circulation sanguine qui serait autrement disponibles pour la pénétration dans les tissus pour atteindre la cible thérapeutique ou le rein pour l'élimination. De nombreuses méthodes expérimentales sont disponibles pour mesurer la liaison aux protéines plasmatiques (ultrafiltration, ultracentrifugation, dialyse à l'équilibre, l'analyse frontale haute performance, micro extraction en phase solide (SPME), adsorption sur charbon, et micro dialyse in vivo) [2-4].

Cependant, ces approches expérimentales pour le développement de nouveaux médicaments candidats sont assez chères et prennent du temps. Cela a entraîné un besoin urgent pour la modélisation computationnelle pour prédire la liaison protéique d'un médicament. Les approches prédictives comprennent des techniques de modélisation relation quantitatives structure-activité (QSAR), qui attirent de plus en plus les chimistes et les scientifiques dans l'industrie pharmaceutique. Les modèles QSAR sont expectés à jouer un rôle important dans les sciences pharmaceutiques [5].

L'utilisation de ces méthodes alternatives à l'expérimentation sont devenue d'un grand intérêt et sont même recommandées dans les nouvelles règlementations afin d'obtenir les données nécessaires à l'enregistrement des substances. Les méthodes QSAR prennent de plus en plus d'importance dans les études de la conception des précurseurs (lead compound). Ce sont des techniques alternatives et complémentaires aux techniques expérimentales *in vitro* et *in vivo* dans les phases très précoces du développement des médicaments et sont un outil qui permet une prédiction rapide d'une ou plusieurs propriétés biologique. L'objectif d'une modélisation QSAR est de trouver

des modèles précis, applicables et robustes afin de trouver une relation entre la structure et l'activité dans un but de prédiction mais également d'interprétation. La puissance de la prévision de l'activité biologique utilisant l'outil QSAR l'a rendue indispensable dans les recherches pharmaceutiques. Il est intéressant de noter que le QSAR a contribuée principalement au succès de plusieurs molécules mises sur le marché en tant que médicaments. Parmi lesquelles nous citerons [6]: la quinolone carboxylic acid (Norfloxacine antibactérien); l'acridines (Anticancéreux); la Cimitidine (antiulcère); le 3-amino-1-benzylpyrazolin-5-one (Muzolimine: diurétique); la Lomerizine (Antimigraine).

Les relations entre les structures des molécules et leurs propriétés ou activités sont généralement établies à l'aide de méthodes de modélisation par apprentissage statistique. Les techniques usuelles reposent sur la caractérisation des molécules par un ensemble de descripteurs, nombres réels mesurés ou calculés à partir des structures moléculaires [7, 8]. Il est alors possible d'établir une relation entre ces descripteurs et la grandeur modélisée. Ces méthodes présentent cependant plusieurs inconvénients. Elles nécessitent en effet la sélection des descripteurs pertinents ainsi que leur calcul. De plus, leur représentation par un vecteur de données induit donc une perte d'information, qui peut avoir une influence sur la qualité des modèles [9].

Dans le cadre de cette thèse nous avons mis au point une nouvelle méthode comportant la représentation des descripteurs par une matrice au lieu du vecteur conventionnel afin d'éviter la perte d'information. Car la sélection de quelques descripteurs pertinents est une tâche difficile et toutes les idées de sélections sont basées sur le traitement des descripteurs comme étant des vecteurs. En ce sens, l'objectif de ce travail de doctorat est d'étudier la possibilité de trouver un autre processus de sélection des descripteurs les plus pertinents qui soit moins long et moins complexe.

La présente thèse s'organise en trois chapitres :

Le premier chapitre est dédié à la construction de modèles QSAR dans les méthodes de criblage *in silico* pour la découverte de nouveaux médicaments.

Le deuxième chapitre est une présentation de liaisons des médicaments aux protéines circulantes.

Le troisième chapitre est consacré, quant à lui, à l'élaboration du modèle *QSAR*, on discutera du volet matériels et méthodes, de la procédure de développement des modèles et des outils statistiques permettant d'analyser leur performance.

La thèse se termine par une conclusion générale qui synthétise l'ensemble des résultats obtenus et trace les perspectives de recherche pour la continuité de ce travail.

Références

[1] Bohnert T, Gan L-S. Plasma protein binding: from discovery to development. Journal of pharmaceutical sciences 2013;102:2953-94.

[2] Vuignier K, Schappler J, Veuthey J-L, Carrupt P-A, Martel S. Drug–protein binding: a critical review of analytical tools. Analytical and bioanalytical chemistry 2010;398:53-66.

[3] Oravcova J, Bo B, Lindner W. Drug-protein binding studies new trends in analytical and experimental methodology. Journal of Chromatography B: Biomedical Sciences and Applications 1996;677:1-28.

[4] Gleeson MP. Plasma protein binding affinity and its relationship to molecular structure: an in-silico analysis. Journal of medicinal chemistry 2007;50:101-12.

[5] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: where have you been? Where are you going to? Journal of medicinal chemistry 2014;57:4977-5010.

[6] Kraim K. Etudes de QSAR sur des activites biologiques utilisant des produits d'origines naturels 2009.

[7] Katritzky AR, Lobanov VS, Karelson M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chemical Society Reviews 1995;24:279-87.

[8] Mihalić Z, Trinajstić N. A graph-theoretical approach to structure-property relationships. ACS Publications; 1992.

[9] Goodarzi M, Dejaegher B, Heyden YV. Feature selection methods in QSAR studies. Journal of AOAC International 2012;95:636-51.

Chapitre I

Construction de modèles QSAR dans les méthodes de criblage in silico pour la découverte de nouveaux médicaments

I.1. Découverte de nouveaux médicaments

I.1.1. Historique

Basée sur la capacité à partager les connaissances acquises, la diffusion des savoirs médicaux ancestraux a été un facteur déterminant dans la découverte et l'utilisation des médicaments. Longtemps limitée à une diffusion orale, que ce soit par de simples conversations ou par des histoires et des poèmes récités par des bardes, la publication mondiale rapide des résultats scientifiques a contribué en grande partie à de nombreux succès de la recherche pharmaceutique moderne [1]. On peut distinguer cinq phases dans l'évolution de la découverte des médicaments. La première phase a débuté il y a longtemps par l'utilisation des substances naturelles, principalement d'origine végétale mais aussi minérale et animale. Ces substances étaient sélectionnées par l'observation empirique de leurs effets sur le cours des maladies. Quand a la deuxième phase qui a débuté au début du XIXe siècle on peut dire qu'elle a marqué un tournant historique dans la recherche de nouveaux médicaments grâce à l'isolement de principes actifs de substances naturelles précédemment utilisées dans la médecine traditionnelle. Ainsi, la morphine est isolée du pavot en 1803 et ses effets sont décrits par Friedrich Wilhelm Adam Satürner quelques temps plus tard [2] (Figure I.1). L'essor des médicaments de synthèse débute dans le milieu du XIXe siècle avec la synthèse en 1853 de l'acide acétylsalicylique par Charles Gerhardt [3] puis sa

commercialisation vingt ans plus tard sous le nom de marque « Aspirin » par les Laboratoires Bayer [4]. Dans la seconde partie de la décennie passé (phase 3), le développement de chimiothéques a permis la découverte de milliers de nouveaux médicaments qui ont révolutionnés la médecine, spécialement la thérapie de cancer. Dans cette décennie (phase 4), on est équipé par des banques génomiques (génothèques) regroupant les fragments d'ADN représentants le génome entier d'une cellule, cela a permis la découverte de médicaments intelligents qui agissent sur des cibles spécifiques donc ils réduisent la toxicité (phase 5). En d'autre termes cela nous initialise vers la médecine 4P : préventive, personnalisée, prédictive et participative [5].

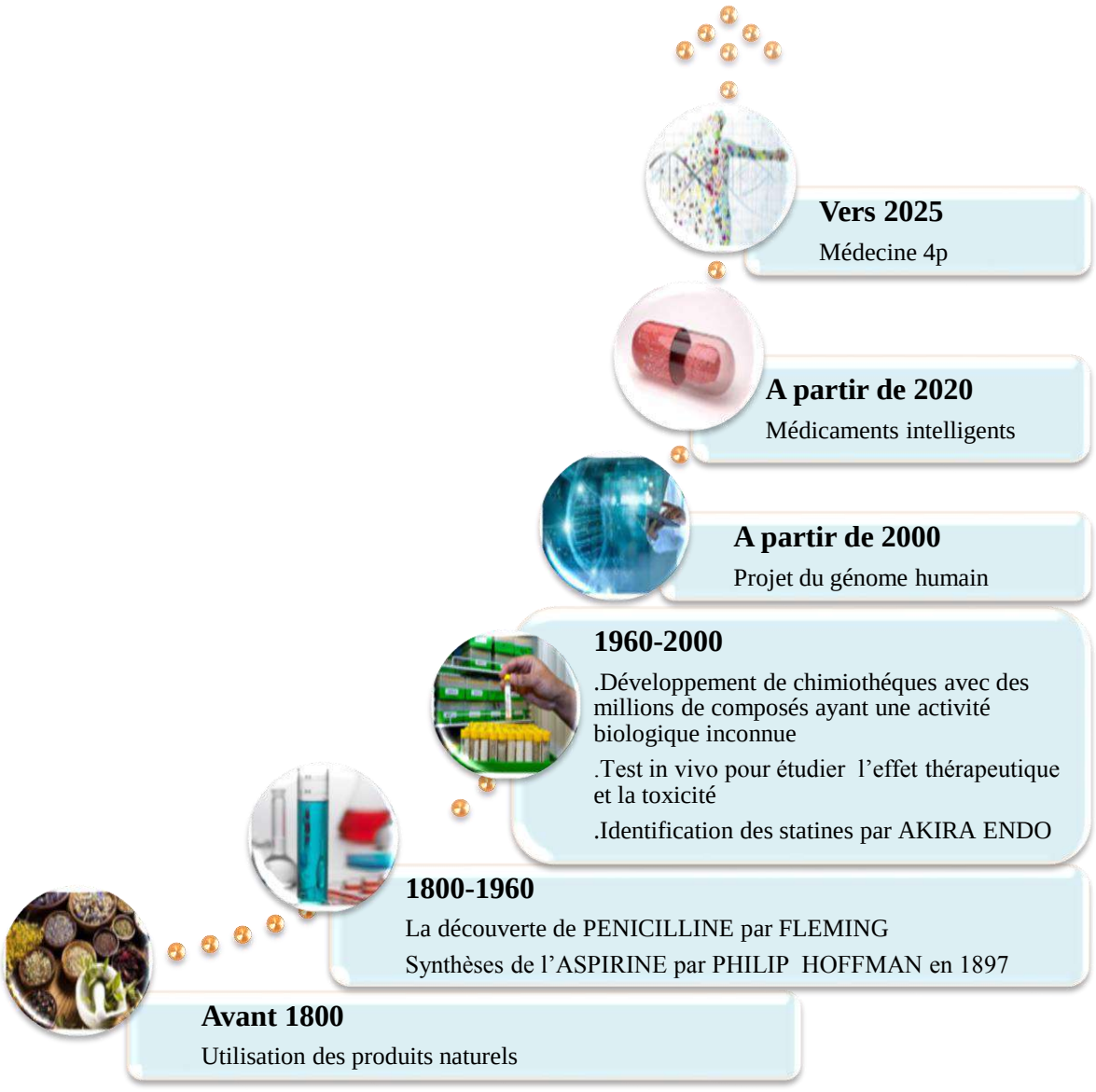


Figure I.1. Quelques grandes étapes de la découverte de médicaments.

I.1.2. Processus général de R&D

La conception de nouveaux médicaments nécessite un processus de recherche long et coûteux, largement entrepris par l'industrie pharmaceutique, avec un soutien académique important. Le délai de découverte d'un nouveau médicament est estimé en moyenne à 12 à 15 ans, de la première phase d'une étude à l'obtention d'une autorisation de mise sur le marché (AMM), pour un coût total de 800 millions à 1 milliard € [5, 6]. Afin d'accélérer le développement de nouveaux médicaments et de réduire les coûts associés à chaque étape, les différentes phases du processus (Figure I.2) doivent donc être optimisées. L'usage des méthodes bioinformatiques et chémoinformatiques est ainsi devenu un standard, celles-ci étant peu coûteuses et devenant de plus en plus efficaces [7].

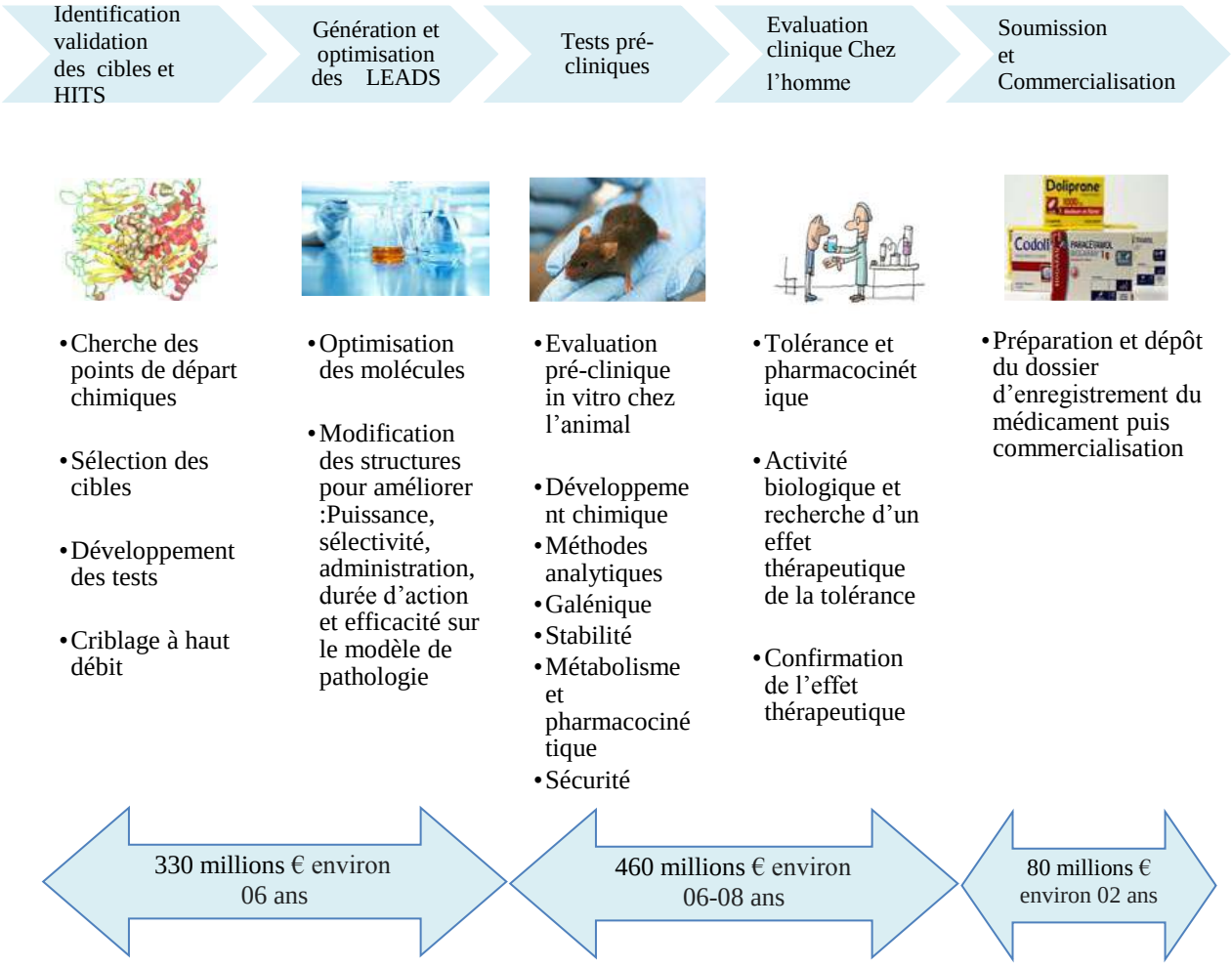


Figure I.2. Etapes du processus de découverte de médicaments.

I.1.2.1. Identification et validation des cibles

Le processus de R&D s'amorce suite à l'émergence d'une maladie dans les populations. Les pathologies les plus répandues, touchant une large partie des populations, sont largement étudiées puisqu'elles représentent un besoin important en termes de réponse thérapeutique, associé également à un fort potentiel commercial. Les maladies neurodégénératives, le Syndrome d'Immunodéficience Acquise (SIDA) et les différents types de cancers font l'objet d'un effort de recherche important, largement financé par les gouvernements et l'industrie pharmaceutique. Inversement, les maladies dites « orphelines » touchent une faible partie des populations et représentent un besoin thérapeutique moindre, associé à un faible potentiel commercial, et tendent à être délaissées. Cependant, une pathologie pour laquelle il n'existe aucune réponse thérapeutique représente, économiquement, l'accès à un monopole de marché lorsque le processus de recherche aboutit à un traitement [8]. Une fois le besoin thérapeutique défini, l'objectif est d'identifier une cible biologique pertinente dont on pourra moduler l'activité afin d'enrayer le processus pathologique. Les cibles thérapeutiques incluent principalement les protéines et plus rarement les gènes (ADN) et les ARNs. Différentes approches permettent d'identifier une cible thérapeutique. Le processus pathologique comporte généralement plusieurs cascades d'évènements, impliquant chacune plusieurs protéines et autres agents biologiques [9, 10]. L'objectif est donc de comprendre ces cascades de réactions, de la manière la plus complète possible, afin d'identifier des cibles thérapeutiques potentielles.

La recherche bibliographique et la compilation des résultats scientifiques peuvent aboutir à l'identification de processus pathologiques et de cibles potentielles. L'exploration de la littérature, grâce aux approches bioinformatiques, permet le tri et l'exploitation de différentes sources de données par « text-mining » (publications, brevets) ou « data-mining » (données d'expression génétiques et protéomiques) [10]. L'étude de l'expression des ARNm ou des protéines permet également l'identification de cibles. Dans ces cas, des cohortes de patients peuvent être constituées. Il est possible, par exemple, de constituer un premier groupe de patients atteints par la pathologie étudiée afin de le comparer à un second groupe de sujets sains [11]. Il est également possible d'établir des échelles de gravité, vitesse de progression, ou résistance à une pathologie, conduisant ainsi à la définition d'un plus grand nombre de groupes de patients et permettant la mise en œuvre de différentes méthodes statistiques [12, 13]. L'étude quantitative de l'expression des entités biologiques permettra de les corréler, ou non, à l'évolution de la pathologie étudiée. Si une protéine est identifiée comme surexprimée dans un groupe de patients atteints, par rapport au groupe contrôle, l'objectif pourra être de rétablir un niveau d'expression normal de cette protéine en agissant sur son processus de synthèse. Plus récemment, les études d'association entre les

polymorphismes génétiques et le développement de pathologies de type GWAS (Genome Wide Association Studies) [14] et eQTL (expression Quantitative Trait Locus) [15] ont permis d'explorer de nouveaux processus pathologiques et d'identifier de nouvelles cibles thérapeutiques.

Les différentes méthodes utilisées pour la l'identification et la validation des cibles sont schématisées dans la figure I.3.

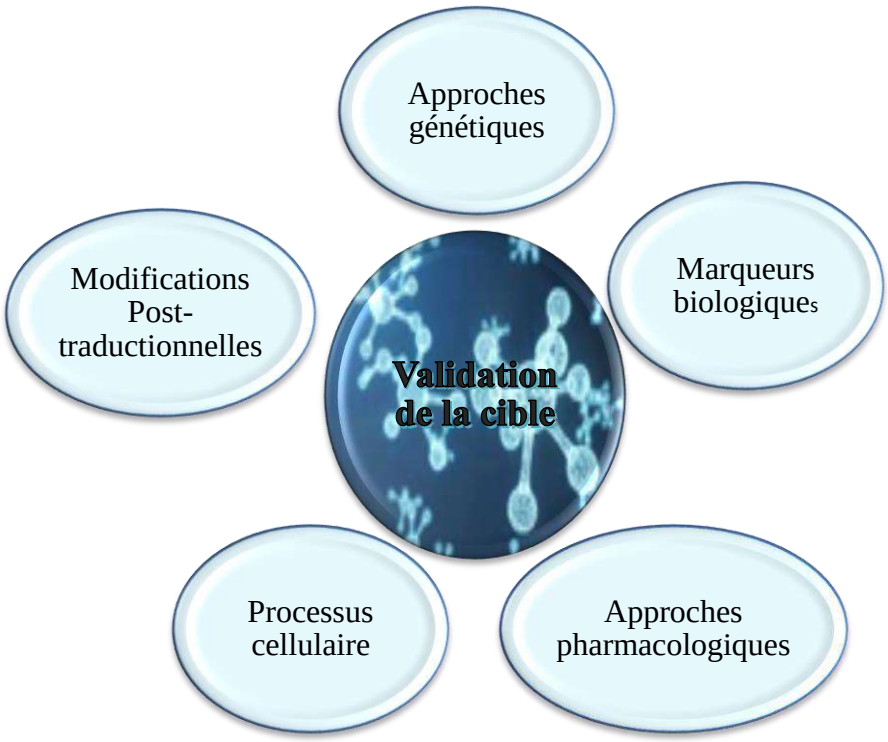


Figure I.3. Méthodes utilisées pour la l'identification et la validation des cibles.

I.1.2.2. Identification de « hits »

Après avoir identifié et choisi la cible thérapeutique, l'objectif est de moduler son activité afin d'altérer le processus pathologique. On appelle touches ou « hits » les composés capables de tels effets. L'approche la plus couramment utilisée dans la recherche de « hits » est le criblage (ou « screening ») de banques de petites molécules (ou chimiothèques). Un criblage peut être réalisé *in vitro* grâce aux techniques HTS permettant d'estimer l'affinité et activité de milliers de composés sur la cible d'intérêt. Cependant, la mise en œuvre de HTS n'est pas toujours possible du fait de leurs coûts importants et du temps nécessaire à leur exécution. Les criblages de type HTS et la validation préclinique des caractéristiques pharmacologiques des composés représenteraient 14% du budget de R&D de l'industrie pharmaceutique. Les techniques de criblage in silico (VHTS), analogues virtuels du HTS, tendent à réduire ces problèmes par trois avantages principaux : une

relative facilité dans leur mise en œuvre, un coût très réduit et l'obtention assez rapide des résultats, en fonction des moyens de calcul disponibles [8, 16]. Généralement utilisé en amont du HTS, le criblage virtuel permet un premier filtrage des chimiothèques en réduisant le nombre de composés à tester expérimentalement et, supposément, en optimisant le potentiel des programmes HTS à découvrir des « hits » réels [17]. Cependant, les résultats de VHTS restent des prédictions dont la fiabilité peut être très variable en fonction de la cible thérapeutique choisie, des caractéristiques des chimiothèques, des méthodes VHTS choisies et de nombreux autres facteurs [18]. Les différentes méthodes d'identification et de validation sont schématisées dans la figure I.4.

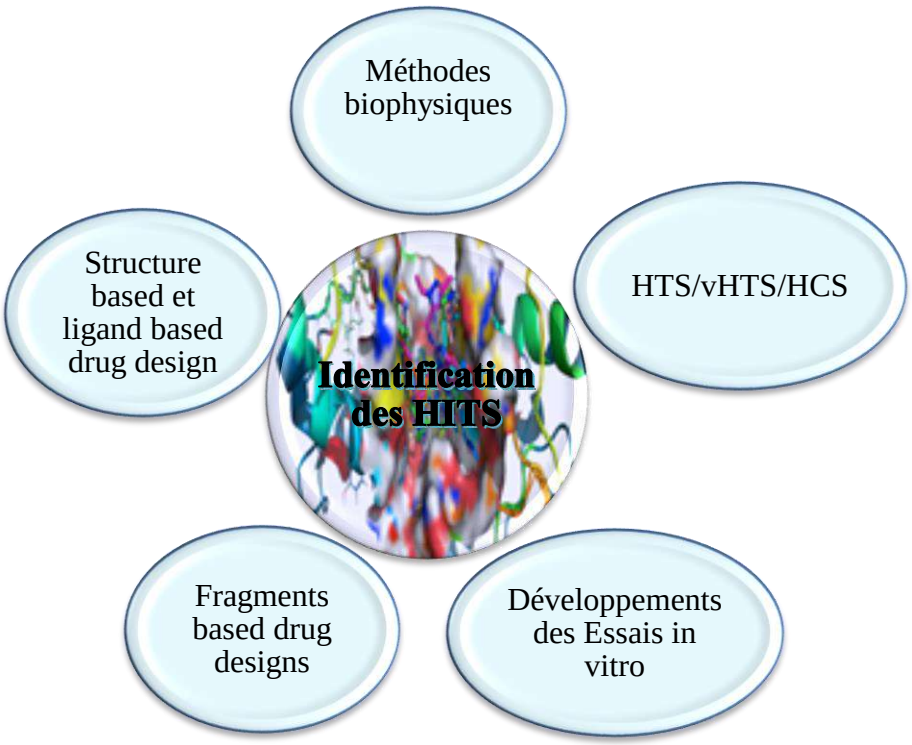


Figure I.4. Méthodes utilisées pour la l'identification et la validation HITS.

I.1.2.3. Optimisation de « leads »

Apres avoir identifiés les leads par criblage virtuel, les chimistes médicinaux procèderont à un processus de modifications chimiques de ces molécules, afin d'optimiser les propriétés et l'activité des futurs médicaments. Cette étape a pour objectif d'obtenir un nombre limité de molécules avec une activité importante, des propriétés physico-chimiques, toxicologiques et biologiques optimales [19]. C'est une étape d'optimisation multi-objective considérée comme la plus critique et la plus difficile. En effet, elle permet d'optimiser en parallèle plusieurs propriétés

qui feront de ces molécules un médicament fiable et efficace et peu ou non toxique. Plusieurs axes de recherche ont été orientés dans ce sens, particulièrement en chémoinformatique [20].

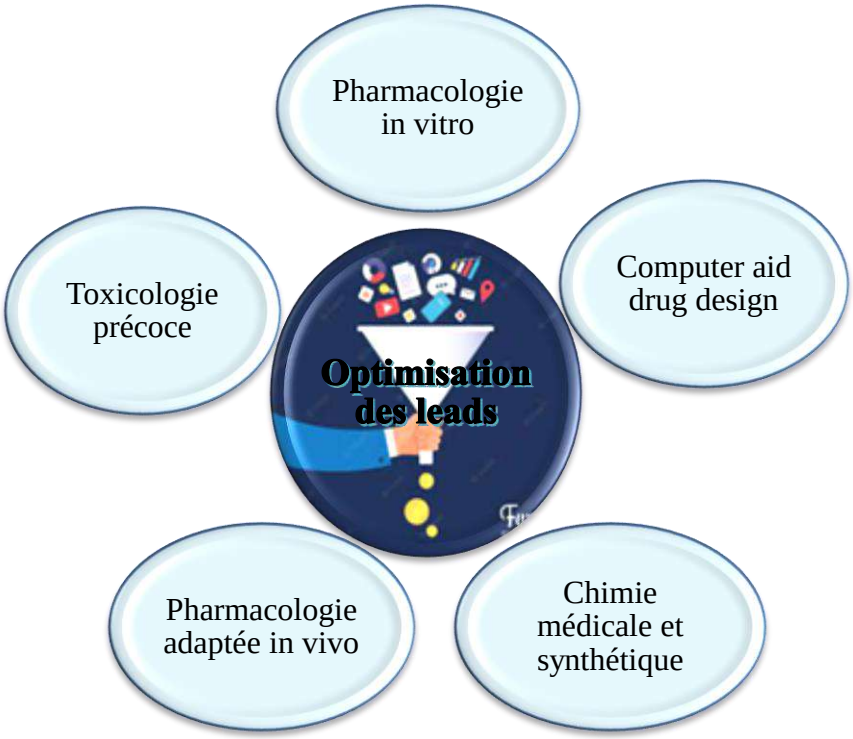


Figure I.5. Méthodes utilisées pour la l'optimisation des LEADS.

I.1.2.4. Essais précliniques

Après avoir obtenu des « leads » optimisés, les essais précliniques consistent en de nombreuse études dont l'objectif est de qualifier le candidat médicament sur les plans pharmacologique, pharmacocinétique et toxicologique. L'usage d'expérimentations animales rationnelles permettra d'envisager l'administration du candidat médicament chez l'homme au cours des essais cliniques suivants. Les études pharmacologiques ont pour objectif de valider le mécanisme d'action du candidat médicament et de mesurer son activité dans des modèles expérimentaux de la pathologie, *in vitro* et *in vivo* chez l'animal. Les études pharmacocinétiques permettent de décrire le devenir d'un composé, sa distribution, son absorption et son métabolisme, puis son élimination par l'organisme. Les études toxicologiques servent à déterminer les doses toxiques du candidat médicament sur l'organisme étudié (principalement des souris ou rats, plus rarement des chats, chiens, porcs, ou primates). Ces données permettront de déterminer les doses à administrer à l'homme lors des essais cliniques et constituent une première approche dans l'étude

de potentiels effets indésirables du candidat médicament, permettant de suivre ces effets de manière proactive. Une partie annexe du développement préclinique consiste également en l'évaluation du risque environnemental lié à la mise sur le marché du candidat médicament. Toutes les informations recueillies lors des essais précliniques seront compilées dans un dossier de demande d'autorisation de mise sur le marché du candidat médicament. Celui-ci sera étroitement étudié par les autorités de santé compétentes (en France, il s'agit de l'Agence Nationale de Sécurité du Médicament (ANSM) et du Comité de Protection des Personnes (CPP)), avant d'autoriser, ou non, l'entrée du candidat médicament en phase d'essais cliniques [21].

I.1.2.5. Essais cliniques et commercialisation

C'est une phase de test sur l'homme. C'est l'étape la plus longue et la plus couteuse du processus. Elle est marquée par un taux d'échec plus élevé. Elle dure entre 6 et 8 ans, et a comme but de tester l'efficacité des médicaments candidats issus de l'étape d'optimisation sur l'homme. Trois grandes étapes caractérisent un essai clinique [22] :

- ***La première phase*** est caractérisée par une étape d'évaluation de la tolérance au médicament sur un nombre restreint de volontaires sains et malades avec l'absence d'effets secondaire frappants.
- ***La deuxième phase*** est réalisée pour faire une bonne estimation des effets secondaire, de prédire l'efficacité thérapeutique et de déterminer les doses nécessaires pour avoir l'effet souhaité. Cette étude est effectuée sur un nombre croissant de patients malades.
- Une étude sera réalisée à ***la troisième phase*** de cette étude dans le but de faire une comparaison entre l'effet du médicament étudié et le traitement de référence. Elle s'effectue sur plusieurs groupes voir des milliers de patients malades.

Une fois les résultats des essais cliniques du médicament sont positifs. Ils seront soumis aux autorités de régulation pour officialiser sa commercialisation. Le processus de pharmacovigilance débute par une surveillance du médicament pour déterminer ses éventuels effets indésirables à court et/ou à long termes.

I.1.3. Processus de criblage virtuel

I.1.3.1. Intérêts thérapeutiques

Au fil des années, le processus et les critères d'acceptation d'un nouveau médicament sont devenus de plus en plus rigoureux. En effet l'augmentation du coût et de la durée de sa mise au point. En vue de cela, le rapport de Price Waterhouse Coopers Pharma en 2005 a déclaré que les industriels pharmaceutiques doivent développer et améliorer l'efficacité des découvertes de médicaments. Le même rapport propose des outils qui permettent de conduire rapidement à des chimiothèques avec une haute valeur ajoutée. Ce sont les méthodes *in silico* qui deviendront un outil dominant pour cette découverte. L'intégration des données expérimentales et informatiques, permet d'initier une pharmacologie *in silico* assemblant tous les types de données [23].

C'est une utilisation systématique d'un grand nombre d'outils de calcul différents visant à améliorer les connaissances sur les interactions entre la cible d'intérêt thérapeutique et le ligand (amarrage moléculaire). Cette approche s'appuie sur l'augmentation de la productivité du criblage (criblage virtuel à haut débit [vHTS]) ou la suggestion de nouveaux candidats principaux potentiels (fragment-fragment-based ligand design [FBD]). Récemment, des méthodes *in silico* ont permis d'activer l'identification et l'optimisation des molécules capables de se fixer sur des cibles d'intérêt thérapeutique, Ils ont donc contribué à la conception et la production d'environ 50 composés qui se sont plié à des essais cliniques, dont certains d'entre eux sont approuvés par la FDA [24].

I.1.3.2. Méthodes de criblages in silico

Le criblage virtuel, analogue *in silico* de l'HTS, peut schématiquement être assimilé à un entonnoir dans lequel on verse un grand nombre (généralement entre cent mille et dix millions) de composés, constituant la chimiothèque à cribler, pour obtenir, à l'aide d'un algorithme de criblage, un plus faible nombre (quelques dizaines à quelques milliers) de composés qui seront ensuite testés expérimentalement. Le rôle des méthodes de criblage est donc d'éliminer les composés supposés inactifs ou les molécules indésirables tout en priorisant les composés les plus susceptibles d'être actifs. Cependant, très souvent, la liste des composés proposés n'est pas utilisée en l'état et les composés à tester expérimentalement sont sélectionnés manuellement par des experts, c'est ce qu'on appelle le « cherry picking » [25]. Deux grandes familles de méthodes de criblages sont distinguées et le choix de leur utilisation est basé sur la disponibilité des données au démarrage du projet [18]. Ainsi, lorsque la structure tridimensionnelle de la cible biologique a été résolue (cristallographie aux rayons X, résonance magnétique nucléaire ou modèle de structure), les méthodes basées sur la structure (ou « structure-based »), évaluent la capacité des ligands à établir des interactions avec le site de liaison étudié pour sélectionner les molécules capables de se lier à la cible. Les méthodes basées sur les ligands (ou « ligand-based ») peuvent pour leur part être mises en œuvre lorsque les valeurs d'activité pour la cible étudiée d'un ensemble de ligands sont disponibles. Les relations

structure-activité de ces molécules sont alors analysées pour découvrir de nouveaux composés susceptibles d'être actifs [26]. Lorsque ces deux types de données sont disponibles simultanément, les méthodes « ligand-based » et « structure-based » peuvent toutes deux être utilisées l'une à la suite de l'autre [27]. Au sein de ces deux grandes familles, différentes méthodes ont été développées (Figure I.6).

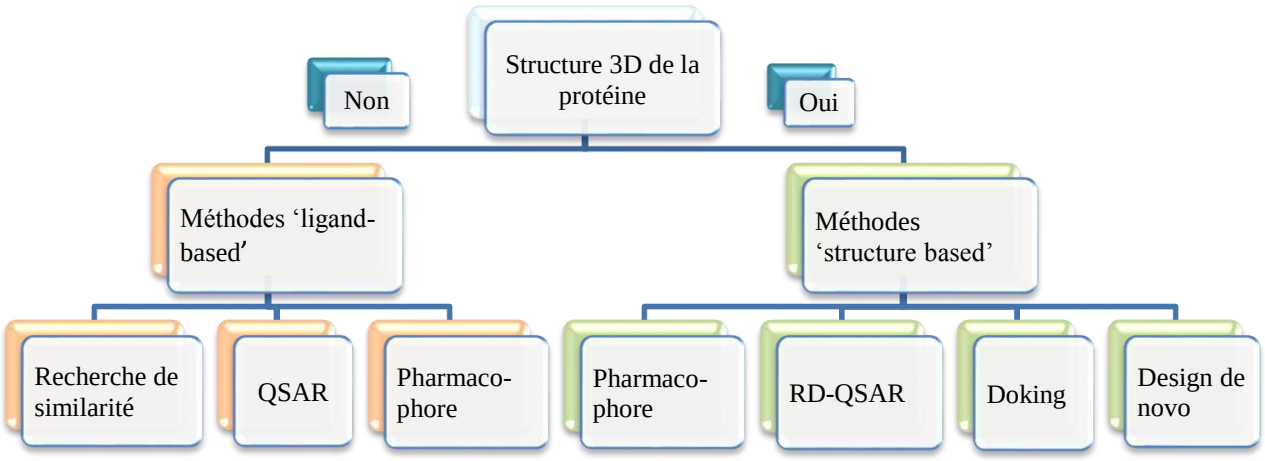


Figure I.6. Classification des méthodes de criblage virtuel « ligand-based »et« structurebased »[28].

Ces différentes méthodes de criblage virtuel peuvent ainsi être utilisées dans les premières phases de développement de nouveaux médicaments, pour guider la sélection des composés les plus prometteurs, que ce soit lors des phases d'identification des hits ou d'optimisation des leads.

I.1.3.2.1. Criblage virtuel basé sur la structure

Le VHTS basé sur la structure est considéré comme l'application la plus simple des algorithmes d'amarrage. Elle consiste à déterminer le mode de liaison de la cible protéique d'intérêt thérapeutique avec une base de données complète de composés existants ou virtuels. Les composés les plus prometteurs sont ensuite utilisés pour des tests expérimentaux.

La découverte de médicaments basée sur la structure (SBDD) est actuellement au cœur des processus de développement des agents thérapeutiques et de leurs processus métaboliques. L'utilisation d'ordinateurs pour des méthodes computationnelles dans cette découverte présente l'avantage de livrer de nouveaux candidats thérapeutiques de manière plus rapide et plus économique que la méthode traditionnelle utilisée auparavant. La méthode SBDD utilise

l'information et la connaissance de la structure tridimensionnelle (3D) de la cible thérapeutique dans le processus, afin d'étudier les interactions moléculaires impliquées dans la liaison ligand-cible [29].

I.1.3.2.2. Criblage virtuel « ligand-based »

Dans cette méthode, le criblage virtuel basé sur le ligand «ligand-based». Elle est utilisée lorsqu'au moins un ligand de la molécule cible est connu. Le principe de cette méthode est que les molécules similaires vont présenter des profils d'activité similaires. Cette similarité est mesurée par des propriétés communes utilisées par la suite comme descripteurs de similarité. Différentes méthodes peuvent être utilisées. Elles varient en fonction du nombre de ligands de la molécule cible et du type de descripteurs. A savoir le criblage par des pharmaco-phores, la recherche de similarité et les méthodes QSAR [30].

I.1.4. Limitations et échecs

La recherche sur les médicaments est donc confrontée à des défis importants, principalement scientifiques, mais aussi financiers et politiques. Le taux de succès des phases d'essai clinique est d'environ 11%, tous types de pathologies confondues. Aujourd'hui, les premières causes d'échec lors des phases d'essai clinique sont le manque d'efficacité des composés (30%) et leur toxicité (30%). Les échecs liés au manque d'efficacité des composés sont plus fréquents lorsque les modèles animaux utilisés sont peu prédictifs, notamment pour les pathologies du système nerveux central et les cancers, induisant de forts taux d'échec en phases II et III [31].

I.2. Conception de médicament et la théorie QSAR

Les modèles quantitatifs de la relation structure-activité (QSAR) sont des méthodes de régression quantitative qui permettent de relier la structure chimique à l'activité biologique [32]. Elles ont été appliquées de manière extensive dans plusieurs disciplines scientifiques : la chimie, la biologie et la toxicologie. Les modèles QSAR sont considérés comme un outil scientifiquement crédible pour prédire et classer les activités biologiques des produits chimiques non testés. Il est donc devenu très essentiel dans l'industrie pharmaceutique comme méthode de découverte et d'optimisation de pointe pour diriger le développement de nouveaux produits [33].

Généralement le processus de conception d'un médicament commence par la découverte d'un composé principal. Cette découverte sera suivie par une série de tests et de synthèse de milliers de dérivés de la structure de ce composé. Ce qui permet éventuellement de trouver enfin un médicament utile et efficace [34].

Le processus de conception est basé sur l'étude des relations quantitatives structure-activité (QSAR) pour prédire l'activité de nouvelles molécules. Principalement, l'étude QSAR qui vise à identifier les relations entre certains aspects de la structure moléculaire et les propriétés telles que la toxicologie, la pharmacocinétique et la pharmacodynamique [35].

La nécessité de dériver de nouveaux aspects structurels ou descripteurs moléculaires est une étape cruciale. Ceci n'est devenu possible que grâce à l'avènement des récents progrès technologiques. En effet ce sont ces progrès qui ont permis l'accélération du processus d'optimisation et la découverte de médicaments.

I.2.1. QSAR et Descripteurs

L'analyse quantitative des relations existantes entre les structures d'un ensemble de composés et leurs activités permet d'identifier et d'évaluer l'impact des propriétés influençant l'activité biologique. Cette relation peut être décrite par une équation qui corrèle mathématiquement les influences réciproques des paramètres concernés. L'extrapolation de ces résultats peut servir de base à la prédiction de l'activité de nouveaux composés. Lors d'une étude QSAR (Figure I.7), il faut calculer à partir de la structure moléculaire de plusieurs actifs tous les descripteurs moléculaires possibles. Après élimination des descripteurs dont la valeur ne varie pas ou peu sur l'ensemble des molécules, une analyse multi variée suivie d'une évaluation statistique sont menées pour obtenir le modèle QSAR. Une dernière étape de validation du modèle est nécessaire pour s'assurer de sa fiabilité [36].

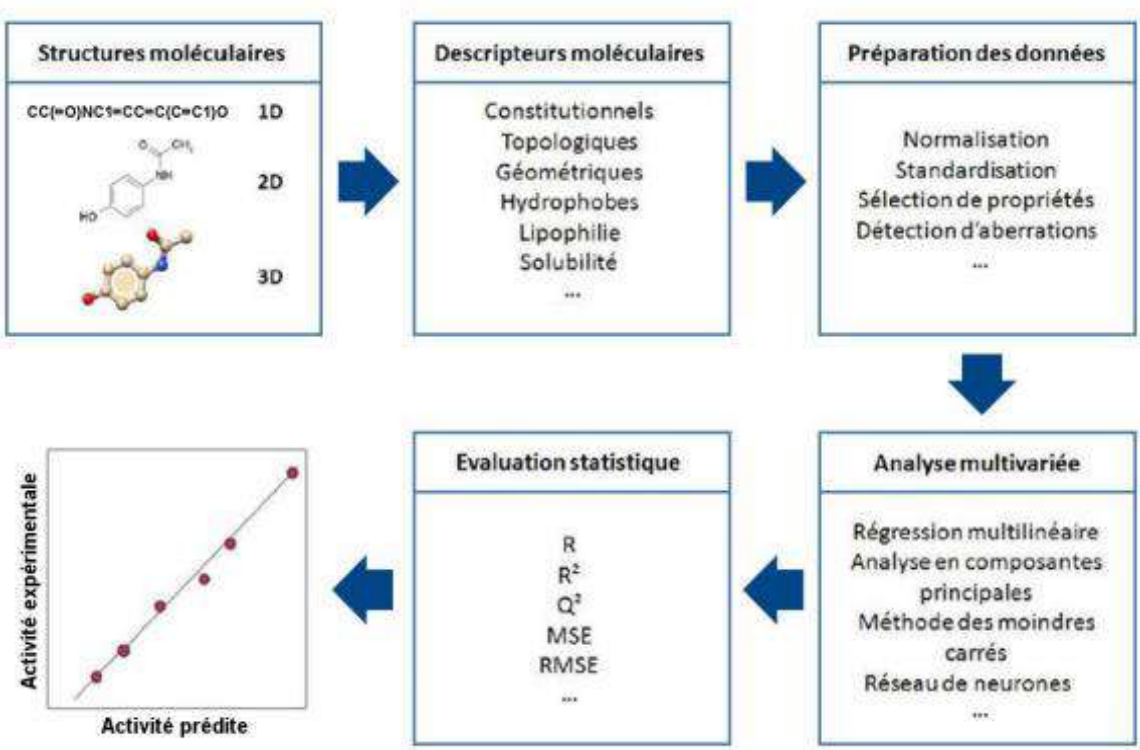


Figure I.7. Illustration du processus de conception de modèles QSAR aboutissant à un modèle statistique [36].

Il existe plus de 3000 descripteurs moléculaires pouvant être utilisés lors d'études QSAR [37], provenant de différentes théories et approches, et qui peuvent être obtenus soit directement à partir de la structure chimique, soit en utilisant des logiciels appropriés. Les descripteurs les plus utilisés lors des études QSAR peuvent être des descripteurs topologiques (2D-QSAR), des descripteurs géométriques ou basés sur des grilles (3D-QSAR). De nouvelles approches plus récentes, QSAR-4D, -5D et -6D ont été développées pour améliorer les approches QSAR-3D, en prenant en compte différentes conformations des ligands (4D), et l'adaptation structurale de la cible au ligand (5D) et les effets de solvants (6D) [38-40].

I.2.2. Réduction de la dimensionnalité

Chercher à réduire la dimensionnalité d'un ensemble de données devient de plus en plus Indispensable en raison de la multiplication des données. Dans de nombreux domaines, le système de résolution d'un problème est fondé sur un ensemble des variables (caractéristiques). L'augmentation du nombre de ces variables (caractéristiques) qui modélisent le problème introduit

des difficultés à plusieurs niveaux comme la complexité, le temps de calcul ainsi que la détérioration du système de résolution en présence de données bruitées. Une méthode de réduction de la dimensionnalité consiste à trouver une représentation des données initiales dans un espace plus réduit. On peut dire que c'est une réduction basée sur une sélection de caractéristiques qui consiste à sélectionner les caractéristiques les plus pertinentes à partir de l'ensemble de données des variables décrivant le phénomène étudié [41].

I.3. Les algorithmes de sélection des descripteurs les plus pertinents dans le Modèles de relations quantitatives structure-activité (QSAR)

La sélection des descripteurs les plus pertinents consiste à extraire les meilleures informations d'une base de données et à évaluer le sous-ensemble de caractéristiques dans les algorithmes de sélection. Ils peuvent être classées en trois catégories principales : "Filter", "Wrapper" et "Embedded" (Figure I.8).

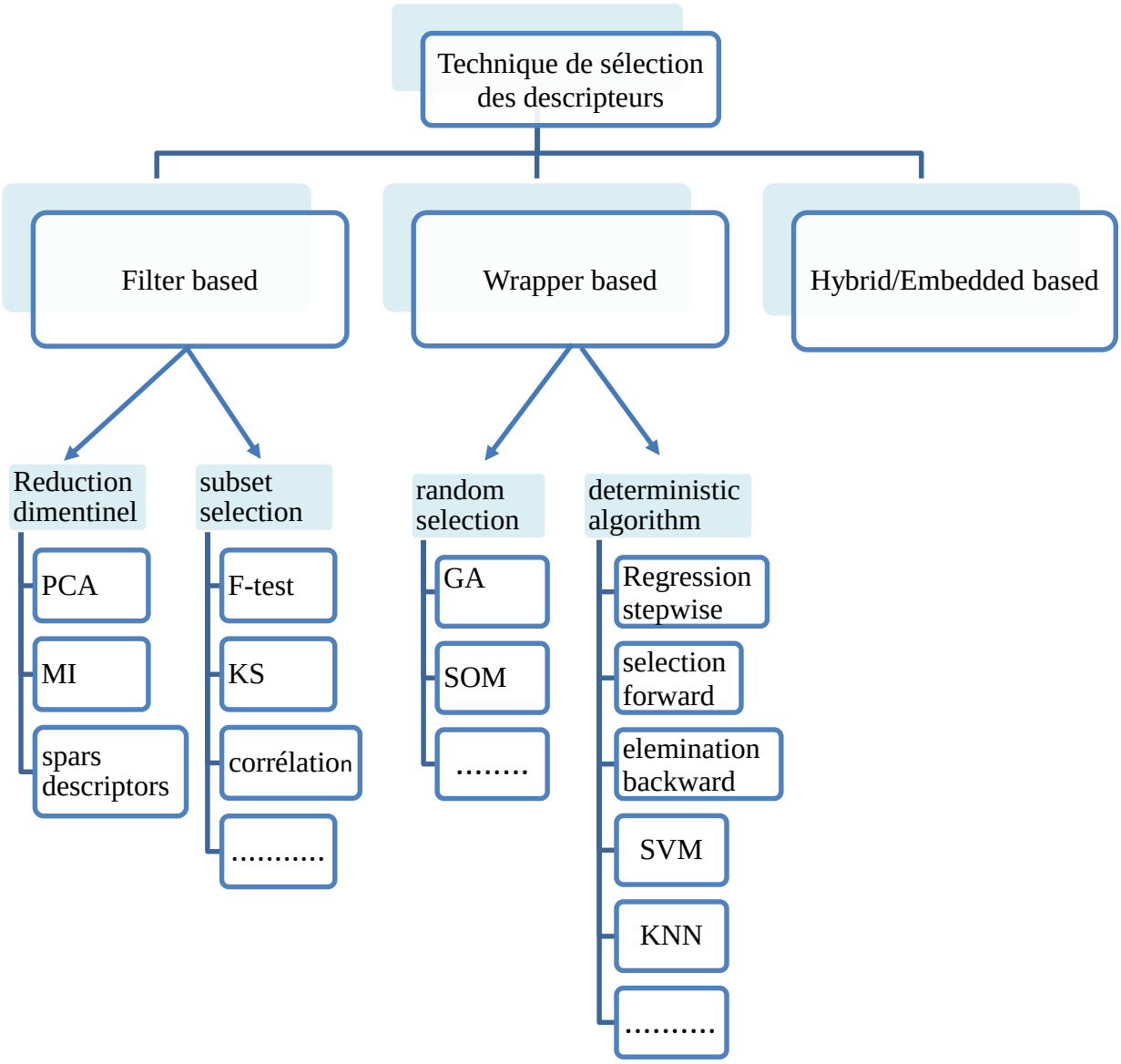


Figure I.8. Technique de sélection des descripteurs [42].

I.3.1. Filter

Le modèle ”Filter” a ´été le premier utilisé pour la s´élection de caractéristiques. Dans celui-ci, le critère d’´évaluation utilisé évalue la pertinence d’une caractéristique selon des mesures qui reposent sur les propriétés des données d’apprentissage. Cette méthode est considérée, davantage comme une étape de prétraitement (filtrage) avant la phase d’apprentissage. En d’autres termes, l’évaluation se fait généralement indépendamment d’un classificateur. Les méthodes qui se basent sur ce modèle pour l’évaluation des caractéristiques, utilisent souvent une approche heuristique comme stratégie de recherche. La procédure du modèle ”Filter” est illustrée par la figure I.9 [43].

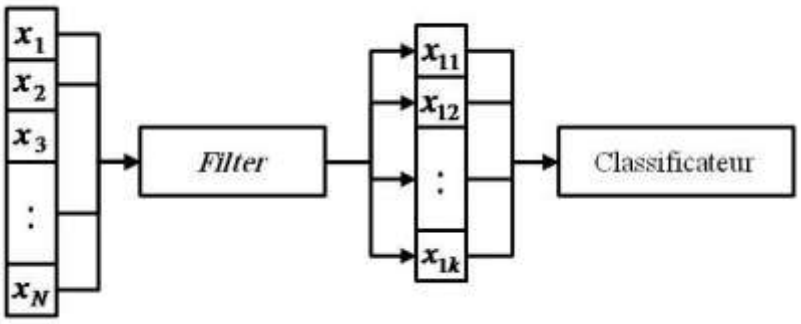


Figure I.9. La procédure du modèle ”Filter” [44].

I.3.2. Wrapper

Le principal inconvénient des approches ”Filter” sont le fait qu'elles ignorent l'influence des caractéristiques s´sélectionnées sur la performance du classificateur à utiliser par la suite. Les méthodes ”Wrapper”, appelées aussi méthodes enveloppantes, évaluent un sous-ensemble de caractéristiques par sa performance de classification en utilisant un algorithme d'apprentissage. La procédure du modèle ”Wrapper” est illustrée par la figure I.10. L'´évaluation se fait à l'aide d'un classificateur qui estime la pertinence d'un sous-ensemble [45].

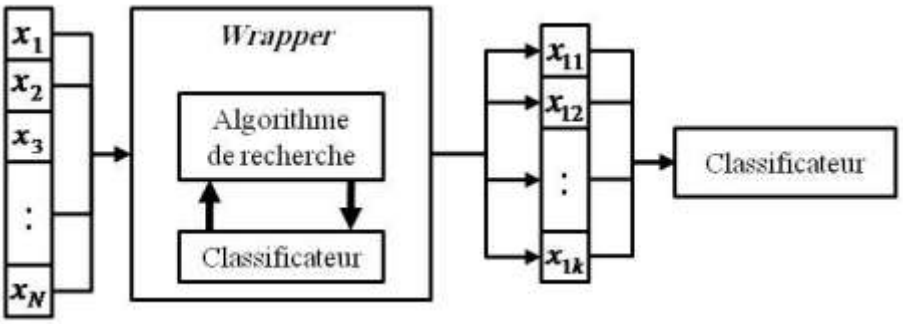


Figure I.10. La procédure du modèle ”wrapper” [44].

Les sous-ensembles de caractéristiques s´sélectionnés par cette méthode sont bien adaptés à l'algorithme de classification utilisé, mais ils ne sont pas forcément valides si on change le classificateur. La complexité de l'algorithme d'apprentissage rend les méthodes ”Wrapper” très

coûteuses en temps de calcul. En général, pour diminuer le temps de calcul et pour éviter les problèmes de sur-apprentissage, le mécanisme de validation croisée est fréquemment utilisé. Une meilleure performance des méthodes ”Wrapper” par rapport à certaines méthodes de filtrage a été démontré par Kohavi et John [45]. Le problème de la complexité de cette technique rend impossible l'utilisation d'une stratégie de recherche exhaustive. Par conséquent, des méthodes de recherche heuristiques ou aléatoires peuvent être utilisées. La recherche devient néanmoins, de plus en plus irréalisable avec l'augmentation de la taille de l'ensemble initial de caractéristiques. Il existe deux inconvénients principaux qui limitent ces méthodes [43] :

- La complexité et le temps de calcul nécessaire pour la sélection : Le principal inconvénient de l'approche ”Wrapper” est le temps nécessaire pour la sélection des caractéristiques, il est nettement plus long que celui des approches de filtrage et d'autres approches de sélection de caractéristiques. L'utilisation d'un classificateur pour évaluer les sous-ensembles ainsi que les techniques d'´évaluation (validation croisée par exemple) rendent les méthodes ”Wrapper” très coûteuses en termes de temps de calcul.
- La dépendance des caractéristiques pertinentes s´sélectionnées par rapport au classificateur utilisé : La deuxième limitation de l'approche ”Wrapper” est que l'évaluation des caractéristiques se fait par un seul classificateur lors de la sélection. Chaque classificateur a ses spécificités et ses hypothèses. Ainsi le sous-ensemble sélectionné dépend toujours du classificateur utilisé.

I.3.3. Embedded

A la différence des méthodes ”Wrapper” et ”Filter”, les méthodes ”Embedded” (appelées aussi méthodes intégrées) incorporent la sélection de variables lors du processus d'apprentissage. Un tel mécanisme intégré pour la sélection des caractéristiques peut être trouvé, par exemple, dans les algorithmes de type SVM, AdaBoost, ou dans les arbres de décisions. Dans les méthodes de sélection de type ”Wrapper”, la base d'apprentissage est divisée en deux parties : une base d'apprentissage et une base de validation pour valider le sous-ensemble de caractéristiques s´sélectionné. En revanche, les méthodes intégrées peuvent se servir de tous les exemples d'apprentissage pour établir le système. Cela constitue un avantage qui peut améliorer les résultats. Un autre avantage de ces méthodes est leur plus grande rapidité par rapport aux approches ”Wrapper” parce qu'elles évitent que le classificateur recommence de zéro pour chaque sous-ensemble de caractéristiques [46].

I.4. Construction de modèles QSAR et RNA

De nombreuses méthodes statistiques peuvent être mises en œuvre pour la construction de modèles QSAR, parmi les plus accessibles : les régressions linéaires multiples (Multiple Linear Regression ou MLR) ou de moindres carrés (Partial Least Squares ou PLS), les arbres de décision, l'analyse en composantes principales (Principal Component Analysis ou PCA) et la méthode des plus proches voisins [47]. Les méthodes des forêts aléatoires [48], des machines à vecteurs de support [49] ou des réseaux de neurones peuvent également être appliquées à ces fins [47]. Plus récemment, des techniques de « deep learning » ont également été appliquées à la prédiction de la solubilité de molécules « druglike » [50].

I.4.1. Le réseau de neurone artificiel (RNA)

Un RNA est un ensemble de neurones formels associés en couches et fonctionnant en parallèle. Ils ont la capacité de stocker de la connaissance empirique et de la rendre disponible à l'usage. Les habiletés de traitement du réseau sont stockées dans des poids synaptiques, obtenus par des processus d'adaptation ou d'apprentissage. En ce sens, les RNA ressemblent au cerveau car, non seulement la connaissance est acquise à travers un apprentissage, mais de plus, cette connaissance est stockée soit dans les connexions entre entités, soit dans les poids synaptiques. La structure des connexions détermine alors la topologie du réseau. La Figure I.11 présente une classification des différents types de RNA regroupés en quatre principales catégories [51].

Dans les réseaux de type Feed Forward, l'information se propage de couche en couche sans que le retour en arrière ne soit possible. Nous nous intéressons dans cette thèse à deux RNA : PMC et CNN.

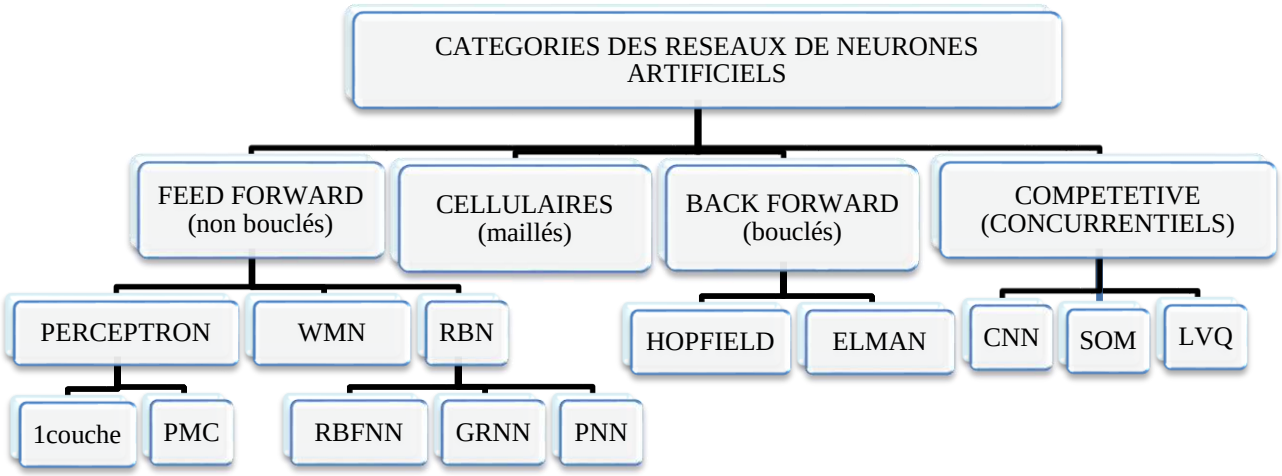


Figure I.11. Différentes catégories de RNA [52] .

I.4.2. Le Perceptron Multicouche (PMC ou MLP)

Le PMC est l'extension du perceptron monocouche à une ou plusieurs couches cachées (Figure I.12). Le perceptron monocouche, inventée en 1957 par Franck Rosenblatt, puis redéveloppé par Bernard Widrow et Ted Ho dans les années1960 sous une version nommée ADALINE, ne contient qu'un unique neurone et ne peut ainsi résoudre que des problèmes de classification linéaire. Ceci constitue sa principale limite. Le PMC par contre, permet de résoudre des problèmes non linéairement séparables et des problèmes logiques plus compliqués, notamment le fameux problème du XOR ; c'est pourquoi il est dit « Approximateur universel » [52].

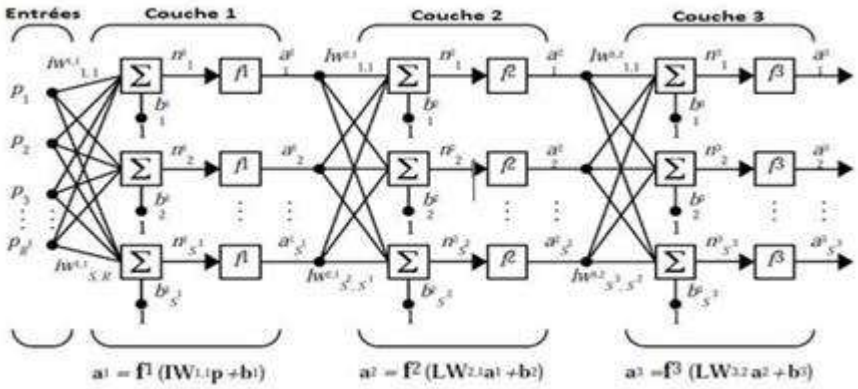


Figure I.12. Architecture du perceptron multicouche [52].

I.4.3. Les réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNN ou ConvNet - Convolutional Neural Network) sont un type particulier de réseaux de neurones multi-couches feedforward souvent utilisés en traitement d'images. Introduits initialement par Fukushima [53] pour une tâche de reconnaissance de forme dont l'architecture du réseau est inspirée du cortex visuel des animaux [54], et popularisés par LeCun en 1990 pour la tâche de la reconnaissance de caractères, Bien que cette approche donne des résultats, ses progrès et son évolution ont été limités par les progrès technologiques en matière de micro-processeurs, de puissance de calculs, et du manque d'accessibilités à des données afin de pouvoir entraîner les neurones. Cependant certains chercheurs ont continué à travailler sur ce modèle pendant environ deux décennies. Et, avec l'aide des évolutions en matière de technologies mais surtout avec la disponibilité toujours plus grande de données, ont pu améliorer cette technique. Ce n'est qu'en 2012 que le Deep Learning est remis au goût du jour en remportant avec succès le concours de reconnaissance d'image fondé par l'université de Stamford (Large Squale Visual Recognition Challenge : ILSVRC) grâce à l'évolution en matière de technologies et à la disponibilité toujours plus grande de données. Un nouvel algorithme de Deep Learning explose les records Il s'agit d'un réseau de neurones convolutif appelé AlexNet, largement inspiré du réseau LeNet. ImageNet regroupant 15 000 000 d'images naturelles comportant différents objets et diverses scènes (véhicules, animaux,...). Aujourd'hui, les réseaux de neurones convolutifs, sont toujours les modèles les plus performants pour la classification d'images. Google, Microsoft, Facebook, Baidu (le moteur de recherche chinois), Alibaba (site marchand chinois), Nvidia (géant du processeur graphique)…utilisent les CNN dans leurs applications [55].

I.4.3.1. Architecture

En traitement d'images, l'extraction des caractéristiques s'effectue sur des petites régions de l'image, dont chaque filtre se déplace dans deux directions (horizontalement et verticalement) sur l'entrée du réseau. Contrairement aux approches d'apprentissage standards qui sont fondées sur des caractéristiques pré-définies, les réseaux de neurones convolutifs sont capables de détecter, d'extraire et d'apprendre des traits spécifiques adaptés à la tâche visée au moment de l'apprentissage sans avoir besoin de ressources, d'implémentation des algorithmes ou d'outils pour extraire des « traits pré-définis » (engineered features). Un réseau de neurone convolutif se compose de plusieurs couches comme le montre la figure I.13. Une architecture CNN est formée par un empilement de couches de traitement indépendantes [56] :

- La couche de convolution (CONV) qui traite les données d'un champ récepteur.

- La couche de pooling (POOL), qui permet de compresser l'information en réduisant la taille de l'image intermédiaire (souvent par sous-échantillonnage).
- la fonction d'activation ou La couche de correction (ReLU), souvent appelée par abus 'ReLU' en référence à la fonction d'activation (Unité de rectification linéaire).
- La couche "entièrement connectée" (FC), qui est une couche de type perceptron.

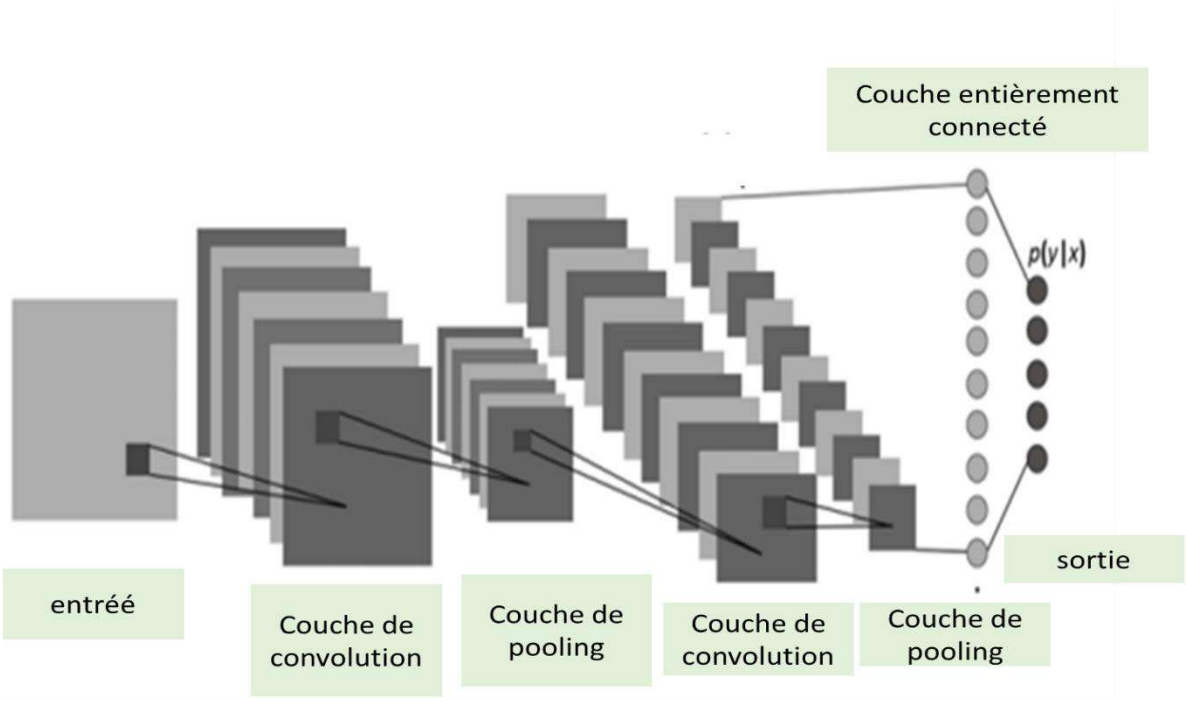


Figure I.13. Les principaux blocs de construction d'une architecture utilisant un réseau de neurones convolutif simple [57].

I.4.3.1.1. Couches convolutives

Les couches convolutives constituent le cœur des réseaux convolutifs. Ces couches se composent d'une grille rectangulaire de neurones ayant un petit champ réceptif étendu à travers toute la profondeur du volume d'entrée. Le terme ‘’convolution’’ vient du fait que les réseaux CNN utilisent des opérations de convolution comme dans le traitement du signal ou d'images. Dans la couche de convolution, on applique un produit de convolution entre l'image originale ou celle obtenue dans la couche précédente avec un filtre dont des coefficients sont assimilés à des poids synaptiques. Le produit de convolution d'une image I avec un filtre F de taille (s*s) est :

$$\boldsymbol{I_F(x,y) = \sum_{i=-s/2}^{s/2} \sum_{j=-s/2}^{s/2} I(x+i, y+j) F(i+\frac{s}{2}, j+\frac{s}{2})} \qquad \textbf{(I.1)}$$

La convolution agit comme un filtrage. Pour effectuer cette opération sur toute l'image, on définit une fenêtre de voisinage de taille (s*s) qui va se déplacer à travers toute l'image. Au tout

début de la convolution, la fenêtre sera positionnée tout en haut à gauche de l'image puis elle va se décaler d'un certain nombre de cases (appelé le pas) vers la droite et lorsque le Réseaux de neurones convolutifs arrivera au bout de l'image, elle se décalera d'un pas vers le bas ainsi de-suite jusqu'à ce que le filtre est parcourue la totalité de l'image. L'opération de convolution produit, en sortie une image appelée ''carte de caractéristiques'' ou ''featuremaps''. De plus, elle peut être appliquée sur des images multi composantes (plusieurs canaux) [58] .

I.4.3.1.2. Couche de mise en commun (pooling)

Ce type de couche est souvent placé entre deux couches de convolution. Elle reçoit en entrée plusieurs featuremaps, pour ensuite réduire la taille de chacune d'entre elles, Autrement dit ; l'opération de pooling consiste à réduire la taille des images caractéristiques, tout en préservant leurs caractéristiques importantes. Elle effectue une sortie de sous échantillonnage. Il existe plusieurs manières de réaliser le pooling. La méthode la plus utilisée est le ''max-pooling'. Celle-ci consiste à découper l'image en cellules régulière, puis on garde au sein de chaque cellule la valeur maximale. En pratique, on utilise souvent des cellules carrées de petite taille pour ne pas perdre trop d'informations. Les choix les plus communs sont des cellules adjacentes de taille (2*2) ou (3*3) qui ne se chevauchent pas, on obtient en sortie le même nombre de feature maps qu'en entrée, mais de tailles plus petites. Finalement la couche de pooling permet de réduire le nombre de paramètres par conséquent d'accélérer les calculs dans le réseau. Elle permet aussi d'améliore l'efficacité du réseau et éviter le problème sur-apprentissage [55] .

I.4.3.1.3. La fonction d'activation

Une fois la convolution effectuée, une fonction d'activation est appliquée sur toutes les valeurs de l'image filtrée. Il existe beaucoup de fonctions d'activation, par exemple on peut utiliser le ReLU [59, 60] qui se définit comme f(σ) = max(0, σ), la fonction tanh [61] ou bien la fonction sigmoïde [62]. La valeur de sortie s (l) k d'un neurone k de la couche l dépend donc de sa fonction d'activation et est définie comme :

$$\boldsymbol{S_k^{(1)} = f(\sigma_k^{(1)})} \qquad \textbf{(I.2)}$$

Avec $S_k^{(1)}$la valeur de sortie du neurone k de la couche l, f la fonction d'activation et $\sigma_k^{(1)}$ la valeur du neurone k.

I.4.3.1.4. Couches entièrement connectées

Enfin, après plusieurs couches de convolution et pooling, le raisonnement de haut niveau dans le réseau neuronal se fait via des couches totalement connectées. Dans les réseaux de neurones convolutifs, chaque couche agit comme un filtre de détection pour la présence de caractéristiques spécifiques ou de motifs présents dans les données d'origine. Les premières couches d'un convolutif détectent des caractéristiques qui peuvent être reconnues et interprétées relativement facilement. Les couches ultérieures détectent de plus en plus des caractéristiques plus abstraites. La dernière couche du réseau convolutif est capable de faire une classification ultra-spécifique en combinant toutes les caractéristiques spécifiques détectées par les couches précédentes dans les données d'entrée [63] .

I.4.3.2. Alex Net et transfert d'apprentissage

Il est plus usuel d'utiliser des modèles de CNNs déjà entrainer et de les réadapter pour le problème de classification ou autre, c'est ce que l'on appelle le transfert d'apprentissage. Il s'agit de transférer l'apprentissage d'un modèle traitant un problème vers un autre type de problème. Il existe un grand nombre de réseaux CNN pré entrainés, les plus connus sont : Google Net, VGG, LeNet, Alex Net…etc [64].

Alex Net : est le nom d'un réseau neuronal convolutif, initialement écrit avec CUDA pour fonctionner avec le support du GPU, qui a participé au défi de reconnaissance visuelle à grande échelle Image Net en 2012. Alex Net a largement surperformé tous ses concurrents précédents et a remporté le défi en réduisant l'erreur de 26% à 15,3% il a été conçu par le groupe Supervision, composé d'Alex krizhevsky. Alex Net a eu un grand impact dans le domaine de l'apprentissage automatique, en particulier dans l'application de l'apprentissage en profondeur à la vision artificielle. Alex Net contenait huit couches ; les cinq première étaient des couches convolutionnelles et les trois dernières étaient des couches entièrement connectées il a parallèlement deux lignes CNN formées sur deux GPU avec des connexions croisées et il a utilisé la fonction d'activation non saturante de Relu, qui a montré une amélioration des performances d'entrainement par rapport au Tanh et au sigmoïde. Il a été formé pendant 6 jours simultanément sur deux GPU NvidiaGeforce GTX [65] (figure I.14).

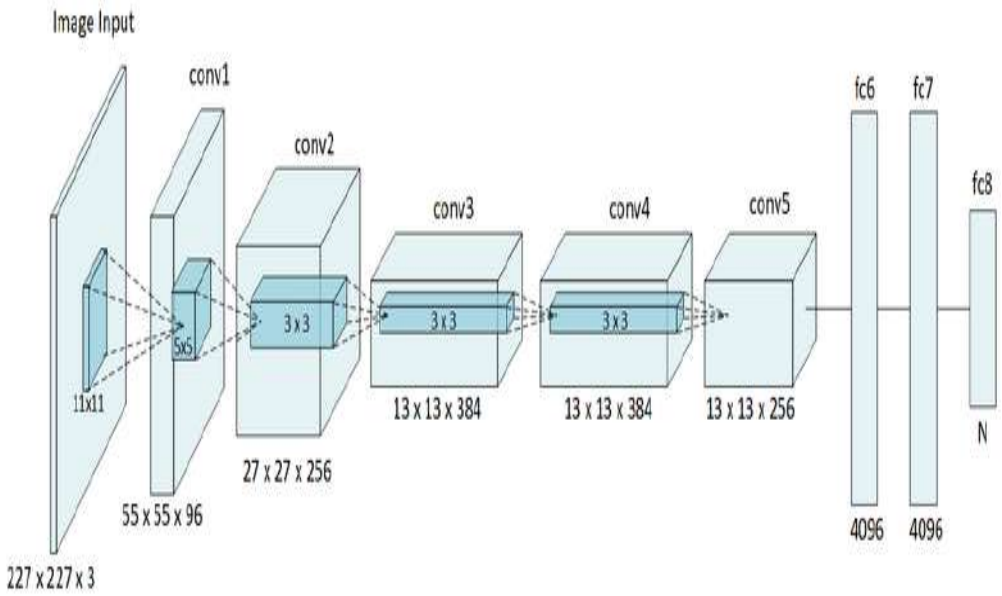


Figure I.14. Architecture d'Alex Net [66].

Références

[1] Sneader W. Drug discovery: a history: John Wiley & Sons; 2005.

[2] Krishnamurti C, Rao SC. The isolation of morphine by Serturner. Indian journal of anaesthesia 2016;60:861.

[3] Ankri J. Système du médicament. Actualité et dossier en santé publique (AdSP) 1999;27.

[4] Lafont O. Du saule à l'aspirine. Revue d'histoire de la pharmacie 2007;94:209-16.

[5] Amelio I, Bertolo R, Bove P, Candi E, Chiocchi M, Cipriani C, et al. Cancer predictive studies. Biology direct 2020;15:1-7.

[6] Boschetti P. Le prix des médicaments: l'industrie pharmaceutique suisse: Editions d'en bas; 2006.

[7] Chanda SK, Caldwell JS. Fulfilling the promise: drug discovery in the post-genomic era. Drug discovery today 2003;8:168-74.

[8] Hubbard RE. 3D structure and the drug-discovery process. Molecular BioSystems 2005;1:391-406.

[9] Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. Health policy 2011;100:4-17.

[10] Aronson J. Rare diseases and orphan drugs. British journal of clinical pharmacology 2006;61:243.

[11] Owens J. Determining druggability. Nature Reviews Drug Discovery 2007;6:187-.

[12] Hopkins AL, Groom CR. The druggable genome. Nature reviews Drug discovery 2002;1:727-30.

[13] Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. British journal of pharmacology 2011;162:1239-49.

[14] Yang Y, Adelstein SJ, Kassis AI. Target discovery from data mining approaches. Drug discovery today 2012;17:S16-S23.

[15] Limou S, Le Clerc S, Coulonges C, Carpentier W, Dina C, Delameau O, et al. Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). The Journal of infectious diseases 2009;199:419-26.

[16] Peltason L, Bajorath J. Systematic computational analysis of structure–activity relationships: concepts, challenges and recent advances. Future Medicinal Chemistry 2009;1:451-66.

[17] Wang D, Bakhai A. Clinical trials: a practical guide to design, analysis, and reporting: Remedica; 2006.

[18] Oprea TI, Matter H. Integrating virtual screening in lead discovery. Current opinion in chemical biology 2004;8:349-58.

[19] Hann MM, Leach AR, Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. Journal of chemical information and computer sciences 2001;41:856-64.

[20] Lusher SJ, McGuire R, Azevedo R, Boiten J-W, van Schaik RC, de Vlieg J. A molecular informatics view on best practice in multi-parameter compound optimization. Drug discovery today 2011;16:555-68.

[21] Hann MM, Oprea TI. Pursuing the leadlikeness concept in pharmaceutical research. Current opinion in chemical biology 2004;8:255-63.

[22] Arrowsmith J. Trial watch: phase III and submission failures: 2007-2010. Nat Rev Drug Discov 2011;10:87.

[23] Aradi I, Érdi P. Computational neuropharmacology: dynamical approaches in drug discovery. Trends in pharmacological sciences 2006;27:240-3.

[24] Jorgensen WL. The many roles of computation in drug discovery. Science 2004;303:1813-8.

[25] Tanrikulu Y, Krüger B, Proschak E. The holistic integration of virtual screening in drug discovery. Drug discovery today 2013;18:358-64.

[26] Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: a critical review. Current medicinal chemistry 2013;20:2839-60.

[27] Perez-Pineiro R, Burgos A, Jones DC, Andrew LC, Rodriguez H, Suarez M, et al. Development of a novel virtual screening cascade protocol to identify potential trypanothione reductase inhibitors. Journal of medicinal chemistry 2009;52:1670-80.

[28] Lagarde N. Méthodes de criblage virtuel in silico: importance de l'évaluation et application à la recherche de nouveaux inhibiteurs de l'interleukine 6: Paris, CNAM; 2014.

[29] Klebe G. Virtual ligand screening: strategies, perspectives and limitations. Drug discovery today 2006;11:580-94.

[30] Johnson MA, Maggiora GM. Concepts and applications of molecular similarity: Wiley; 1990.

[31] Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. Annual reports in computational chemistry: Elsevier; 2008. p. 217-41.

[32] Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature 1962;194:178-80.

[33] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: where have you been? Where are you going to? Journal of medicinal chemistry 2014;57:4977-5010.

[34] Todeschini R, Consonni V. Handbook of molecular descriptors: John Wiley & Sons; 2008.

[35] Greener M. Drug safety on trial: Last year's withdrawal of the anti-arthritis drug Vioxx triggered a debate about how to better monitor drug safety even after approval. EMBO reports 2005;6:202-4.

[36] G Damale M, N Harke S, A Kalam Khan F, B Shinde D, N Sangshetti J. Recent advances in multidimensional QSAR (4D-6D): a critical review. Mini reviews in medicinal chemistry 2014;14:35-55.

[37] Dunn WJ. Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry Series. Volume 11 By Roberto Todeschini and Viviana Consonni (Universita degli Studi di Milano-Bicocca). Edited by R. Mannold, H. Kubinyi, and H. Timmerman. Wiley-VCH: Weinheim and New York. 2000. xxi+ 668 pp. 498 DM. ISBN 3-527-29913-O. ACS Publications; 2001.

[38] Andrade CH, Pasqualoto KF, Ferreira EI, Hopfinger AJ. 4D-QSAR: perspectives in drug design. Molecules 2010;15:3281-94.

[39] Vedani A, Dobler M. 5D-QSAR: the key for simulating induced fit? Journal of medicinal chemistry 2002;45:2139-49.

[40] Vedani A, Descloux A-V, Spreafico M, Ernst B. Predicting the toxic potential of drugs and chemicals in silico: A model for the peroxisome proliferator-activated receptor γ (PPAR γ). Toxicology letters 2007;173:17-23.

[41] Liu H, Motoda H. Feature transformation and subset selection. IEEE Intell Syst Their Appl 1998;13:26-8.

[42] Ghasemi F, Mehridehnavi A, Perez-Garrido A, Perez-Sanchez H. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. Drug Discov Today 2018;23:1784-90.

[43] John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. Machine learning proceedings 1994: Elsevier; 1994. p. 121-9.

[44] Chouaib H. Sélection de caractéristiques: méthodes et applications. Paris Descartes University: Paris, France 2011.

[45] Kohavi R, John GH. Wrappers for feature subset selection. Artificial intelligence 1997;97:273-324.

[46] Wang S, Tang J, Liu H. Embedded unsupervised feature selection. Proceedings of the AAAI Conference on Artificial Intelligence2015.

[47] Lewis RA, Wood D. Modern 2D QSAR for drug discovery. Wiley Interdisciplinary Reviews: computational molecular science 2014;4:505-22.

[48] Breiman L. Random forests. Machine learning 2001;45:5-32.

[49] Cortes C, Vapnik V. Support-vector networks. Machine learning 1995;20:273-97.

[50] Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. Journal of chemical information and modeling 2013;53:1563-75.

[51] Güreken M. Neural network based beamforming for linear and cylindrical array applications: Middle East Technical University; 2009.

[52] KEPCHABE S, TONYE E. Optimisation des performances des antennes intelligentes par synthèse au moyen de réseaux de neurones.

[53] Fukushima K, Miyake S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. Competition and cooperation in neural nets: Springer; 1982. p. 267-85.

[54] Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology 1962;160:106.

[55] Noura H, Nadia K. Classification des textures par les réseaux de neurones convolutifs: Université Mouloud Mammeri; 2018.

[56] Malki N, Guerram T. Classification automatique des textes par Les réseaux de neurones à convolution. 2019.

[57] Blanc-Durand P. Réseaux de neurones convolutifs en médecine nucléaire : Applications à la segmentation automatique des tumeurs gliales et à la correction d'atténuation en TEP/IRM 2018.

[58] Hakim B, Brahim S. Classification des images avec les réseaux de neurones convolutionnels: Université Mouloud Mammeri; 2018.

[59] Jarrett K, Kavukcuoglu K, Ranzato MA, LeCun Y. What is the best multi-stage architecture for object recognition? 2009 IEEE 12th international conference on computer vision: IEEE; 2009. p. 2146-53.

[60] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Icml2010.

[61] Nguyen D, Widrow B. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. 1990 IJCNN International Joint Conference on Neural Networks: IEEE; 1990. p. 21-6.

[62] Norouzi M, Ranjbar M, Mori G. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. 2009 IEEE Conference on Computer Vision and Pattern Recognition: IEEE; 2009. p. 2735-42.

[63] Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. 2017 international conference on engineering and technology (ICET): Ieee; 2017. p. 1-6.

[64] Hussain M, Bird JJ, Faria DR. A study on cnn transfer learning for image classification. UK Workshop on computational Intelligence: Springer; 2018. p. 191-202.

[65] Abd Almisreb A, Jamil N, Din NM. Utilizing AlexNet deep transfer learning for ear recognition. 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP): IEEE; 2018. p. 1-5.

[66] Hemmer M, Khang HV, Robbersmyr K, Waag T, Meyer T. Fault Classification of Axial and Radial Roller Bearings Using Transfer Learning through a Pretrained Convolutional Neural Network. Designs 2018;2:56.

Chapitre II

Liaisons des médicaments aux protéines circulantes (PROTEIN BINDING)

II.1. Introduction

Les protéines plasmatiques assurent le transport et la distribution des médicaments dans tout l'organisme. Si quelques-uns utilisent des protéines spécifiques comme la transcortine pour les glucocorticoïdes, la plupart d'entre eux sont liés à l'albumine (HSA), à l'alpha-1 glycoprotéine acide (AGP) et aux lipoprotéines (LP). L'éventail très large des structures chimiques des médicaments fixés montre qu'il n'existe pas de spécificité étroite de reconnaissance des substances liées (ligands). Cependant, les liaisons les plus affines se forment sur des sites de liaisons parfaitement individualisés sur l'HSA et sur l'AGP. Pratiquement toutes ces interactions ont en commun une liaison élémentaire hydrophobe qui implique une liposolubilité minimale du médicament et une zone apolaire (le site) de la protéine. Les interactions avec les LP sont le plus souvent de simples liposolubilisations dans la fraction lipidique de la protéine sans valeur rétentionnelle [1].

II.2. Notion de la fixation protéique des médicaments

Après leur absorption, les molécules médicamenteuses vont être distribuées aux différents tissus par la circulation générale. Le sang est un tissu c'est-à-dire un ensemble de cellules ayant la

particularité d'être liquide contrairement aux autres tissus que l'on trouve dans le corps humain. Le sang permet le transport de nutriments (sucres, acides aminés, vitamines, …) et de dioxygène (O2) à travers l'organisme tout en assurant aussi l'élimination des déchets (par exemple le dioxyde de carbone ; CO2). Le sang permet également le transfert d'informations entre des cellules éloignées via la diffusion d'hormones. Il régule aussi le PH, la température corporelle et la teneur en eau des cellules. Il protège aussi contre les pertes de sang par le biais de la coagulation, et contre les maladies grâce aux leucocytes et a des protéines comme les anticorps, les interférons et le complément. Le sang est composé d'une partie liquide appelée plasma et de cellules qui flottent dans ce milieu. On peut isoler le plasma, composante liquide du sang, en faisant tourner un tube de sang complet à haute vitesse dans une centrifugeuse. Les cellules et plaquettes plus denses se déplacent vers le bas du tube, en formant des couches rouges et blanches, tandis que le plasma reste au sommet, formant une couche jaune. Le plasma est composé d'environ 90 % d'eau, les 10 % restant sont composés d'ions, de protéines, de nutriments, de déchets et de gaz dissous. Les trois fractions significatives de protéines plasmatiques sont l'albumine, la globuline et le fibrinogène (Figure II.1). Dans le sang, le médicament existe sous deux formes : une forme libre et une forme liée, cette liaison, se fait par fixation principalement sur ces protéines plasmatiques. On aboutit en général à un équilibre entre une forme liée, non diffusable, et donc sans action et une forme libre, capable de diffuser vers les tissus et donc active. L'interaction avec les protéines plasmatiques assure des fonctions temporaires de fixation (stockage et de transport) et protège la molécule de la dégradation [2].

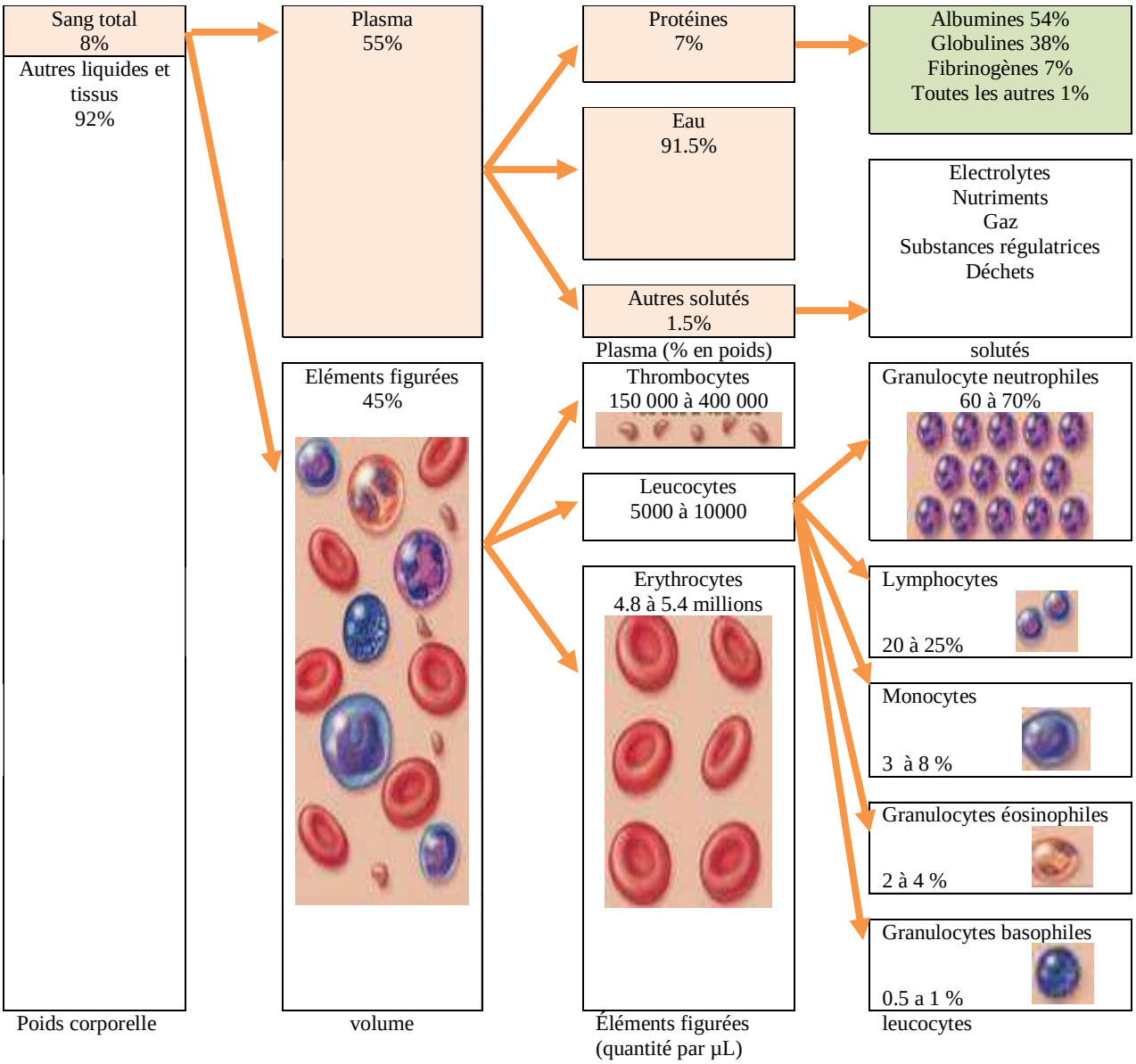


Figure II.1. Les composants du sang chez un adulte normal [2].

La liaison d'un médicament aux protéines plasmatiques joue un rôle important dans ses propriétés pharmacocinétiques et sa pharmacodynamique. La liaison au plasma peut également être considérée comme un stock en circulation, libérant le médicament dans les tissus ayant une plus grande affinité. La fraction non liée (fu : fraction unbound) du médicament est considérée comme la seule fraction capable de diffuser à travers les membranes et d'interagir avec un site effecteur pour produire une réponse thérapeutique [3] (Figure II.2).

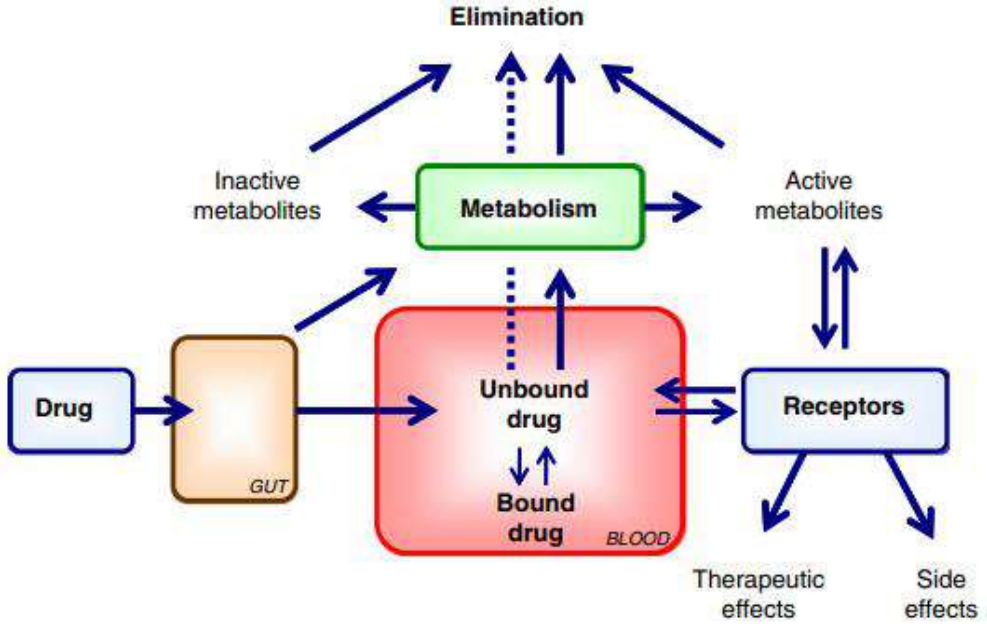


Figure II.2. Distribution du médicament et liaison aux protéines plasmatiques [3].

Les changements des taux de (*fu*) se traduisent directement par des changements du volume de distribution (*V*). L'étendue de la liaison aux protéines plasmatiques ou tissulaires contrôle le volume de distribution et affecte la clairance hépatique et rénale[4]. Ainsi, le volume de distribution (*V*) dépend de la fraction non liée dans le plasma (*fu*), de la fraction non liée dans le tissu *(fuT*), du volume de tissu (*VT*) et du volume de plasma (*VP*) au moyen de l'équation :

$$\boldsymbol{V = \left[\frac{(fu)}{fuT}\right] VT + VP} \qquad \textbf{(II.1)}$$

Les effets de la liaison au plasma sur l'élimination du médicament peuvent être mieux compris en prenant en compte la clairance. Tous les modèles de calcul de la clairance des organes prennent en considération la liaison aux protéines, en particulier la *(fu)* : fraction non liée du médicament. Exemple du calcul de la clairance hépatique [5] :

$$\boldsymbol{CL\ hep = Q * (1 - e^{\frac{-cl\ int * fu}{Q}})} \qquad \textbf{(II.2)}$$

Avec *Q* : le flux sanguin hépatique ; *cl int* : la clairance intrinsèque.

Seule la fraction plasmatique non liée (*fu*) est filtrée par les reins. Par conséquent, le débit de la filtration est le produit du débit de filtration glomérulaire (*DFG*) et de la concentration plasmatique non liée du médicament [53]. Pour un médicament qui est uniquement filtré par les reins et totalement excrété dans l'urine, sa clairance rénale (*CLR*) est le taux de la filtration du produit :

$$\boldsymbol{CLR = fu * DFG} \qquad \textbf{(II.3)}$$

II.3. Protéines fixatrices

Certaines protéines plasmatiques possèdent la propriété de fixer des substances endogènes mais également les produits exogènes comme les médicaments. Il en résulte la formation d'un complexe [protéine – médicament] menant à la distinction du médicament sous une forme libre et sous forme liée. La liaison médicament-protéine est réversible. Elle répond à la loi d'action de masse :

Médicament libre + protéine plasmatique ↔ médicament lié à la protéine

Seule la forme libre du médicament est active pharmacologiquement. La forme liée est inactive pharmacologiquement et ne peut diffuser pour atteindre son lieu d'action. Cette inactivité n'est que temporaire car les formes liée et libre sont en équilibre réversible. Au fur et à mesure de la disparition de la forme libre (par diffusion vers les tissus ou élimination), il y a passage de la forme liée vers la forme libre. Différentes protéines plasmatiques sont impliquées dans la fixation des médicaments : • l'albumine • l'α1 glycoprotéine acide • les lipoprotéines • la γ globuline. L'albumine et l'α1 glycoprotéine acide sont les deux principales protéines impliquées dans la fixation des médicaments. La fixation aux protéines plasmatiques dépend des caractéristiques acido-basiques des médicaments [3].

Un sujet de 70 kg a environ 5 litres de sang, ce qui équivaut à 3 litres de plasma. Le plasma est défini comme la phase liquide du sang et consiste en un liquide résiduel obtenu après coagulation du sang (environ 55% du volume sanguin). Le plasma est composé d'eau (91%) et de protéines (7%), mais contient également des sels, des glucides, des lipides, des vitamines et des hormones. Dans un plasma adulte sain, la concentration totale en protéines est habituellement évaluée entre 60 et 85 g / L. Au premier rang figure le Sérum Albumine Humaine (SAH), chez un sujet normal la concentration d'Albumine peut aller de 35 à 50 g/L avec une moyenne de 41.8g/L,

la protéine la plus abondante dans le plasma sanguin humain, au plan qualitatif, elle représente 60% [6], suivit des α1glycoprotéines (0.4 à 1.0 g/L) et autres protéine.

II.3.1. Albumine

La molécule d'albumine contient plusieurs sites de liaison avec une affinité différente pour différents espèces (Figure II.3).

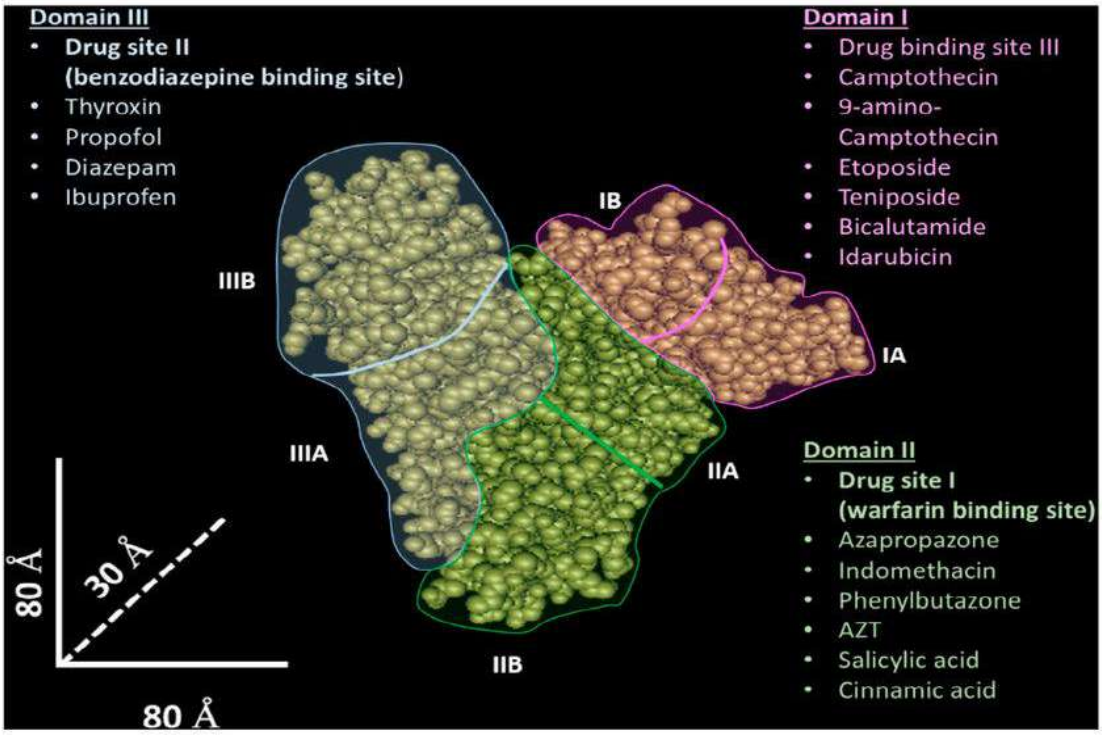


Figure II.3. Structure tridimensionnelle de l'albumine et sites de liaison aux médicaments. Elle est composée de trois domaines (I-III) mis en évidence avec des couleurs différentes (bleu clair-domaine I ; vert- domaine II ; domaine rose III). Les autres sites de liaison interagissant avec les ions, les petites molécules et les peptides ne sont pas représentés [7].

L'albumine est la protéine la plus importante dans le sang humain, de masse moléculaire d'environ 66 kDa, elle est très abondante dans le plasma (~40 g/L), et porte une charge négative au pH physiologique. Capable de fixer des ligands endogènes (acides gras et ions métalliques) ainsi que des ligans exogènes (médicaments) [8]. Deux sites de liaison incluant différents domaines sensibles au pH ont été identifiés sur la molécule d'albumine, les médicaments pouvant se fixer sur l'un ou l'autre des deux sites ou sur les deux [9]. Les médicaments de nature basique, peu ionisés au

pH plasmatique, semblent se lier sur l'albumine mais aussi sur d'autres fractions protéiques circulantes, de façon non spécifique, avec une faible affinité, sur un grand nombre de sites. A l'inverse, les médicaments de nature acide, fortement ionisés à pH 7,4, se fixent en quasi-totalité à l'albumine avec une forte affinité et sur un petit nombre de sites. Compte tenu de la fixation exclusive sur l'albumine et du petit nombre de sites offert, la fixation protéique de ces médicaments dépend de l'albuminémie et il peut exister des saturations dans diverses circonstances physiologiques ou pathologiques accompagnées d'hypo albuminémie, telles que l'âge, l'insuffisance hépatique ou l'insuffisance rénale. En effet, la concentration d'albumine dans le plasma à la naissance, qui est en moyenne de 35 à 37 g/L, augmente lentement durant la première année de vie pour atteindre une concentration d'environ 50 g/L chez l'adulte [10]. Cette diminution de l'albuminémie chez les nouveau-nés et les jeunes enfants risque de provoquer une augmentation de la fraction libre des médicaments, d'autant plus que l'albumine fœtale présente une affinité moindre pour les médicaments que l'albumine de l'adulte [11]. Au plan pathologique, l'insuffisance rénale entraîne une hypo albuminémie due à une perte urinaire de protéines, à une diminution de la synthèse protéique et à une moindre résorption des acides aminés. De plus, des changements de conformation de l'albumine ont été mis en évidence dans l'insuffisance rénale, ce qui pourrait être à l'origine d'une diminution d'affinité de la protéine [12]. De même, l'insuffisance hépatique s'accompagne très souvent d'une hypo albuminémie, conséquence de la diminution de la synthèse hépatique des protéines et d'un nombre de sites de fixation réduit.

II.3.2. Acide Alpha-1-glycoprotéine

L'acide alpha-1-glycoprotéine est une molécule de poids moléculaire de (~38-48 kDa) et dont la concentration physiologique est faible (~0,4 à 1,0 g/L). Les médicaments basiques et neutres s'y lient sur un site de fixation avec une forte affinité. Cette liaison peut être rapidement saturable. La concentration en alpha-1-glycoprotéine acide varie en fonction de l'âge. Faible à la naissance, augmente au cours de la première année de vie pour atteindre des valeurs similaires à celles de l'adulte (0.55-14mg/ml chez un adulte sain). Une diminution de la concentration sanguine d'alpha-1- glycoprotéine acide peut s'observer au cours d'une insuffisance hépatique, notamment en cas de dénutrition, ou d'une fuite excessive par le rein en cas de syndrome néphrotique. Une augmentation de concentration traduit un état inflammatoire, une infection ou une insuffisance rénale, le rein étant alors incapable d'éliminer correctement cette protéine [13].

II.3.3. *Autres protéines*

Les lipoprotéines fixent les médicaments les plus lipophiles. On constate parfois une fixation sur les gammaglobulines, celles-ci étant le plus souvent très faibles et non saturables. Des fixations additionnelles peuvent également être mises en évidence sur les globules rouges, les leucocytes et les plaquettes [14].

II.4. Paramètres de la fixation protéique des médicaments

La liaison des médicaments aux protéines plasmatiques est un phénomène dynamique qui dépend du nombre de sites de fixation sur la protéine, de la constante d'affinité et de la concentration molaire en protéine [15]. Les modes de fixation diffèrent selon la structure chimique et les caractéristiques physicochimiques du médicament. Il existe plusieurs types de fixation aux protéines plasmatiques, qui expliquent les différentes affinités. Il est possible de classer la fixation des médicaments aux protéines en deux catégories selon leur force, c'est-à-dire leur affinité : Liaisons faibles au labiles, de type hydrogènes ou hydrophobes. Liaisons fortes de type ioniques plus rarement de type covalentes. Les premières caractérisent les médicaments à faible affinité, elles sont facilement révisables et dissociables. Les secondes sont propres aux médicaments à forte affinité, elles sont stables et difficiles « à briser ». Des données suffisantes de liaisons doivent être obtenues pour que les paramètres de liaison aux protéines puissent être calculés. Spécifiquement, les paramètres caractérisant la force de l'association médicament-protéine, le nombre de classes de sites de liaison et le nombre de sites de liaison par classe/protéine doivent être déterminés. La connaissance de ces paramètres de liaison aux protéines est essentielle pour l'évaluation des implications pharmacocinétiques de la liaison aux médicaments [16].

II.4.1. Affinité

La fixation d'un médicament M sur une protéine P résulte d'un équilibre réversible entre la forme libre [M] et la forme liée [MP] qui obéit à la loi d'action de masse :

(II.1)

$$[M] + [P] \rightleftharpoons [MP]$$

Seule la forme libre [*M*] est active et peut diffuser à travers les membranes. La forme liée [*MP*] peut être considérée comme une forme de stockage qui ne traverse pas les membranes. L'affinité définit la solidité de la fixation du médicament à la protéine. Selon la loi d'action de

masse, *Ka* étant la constante d'affinité du médicament pour la protéine exprimée en litre/mol, on a à l'équilibre [17] :

$$\boldsymbol{Ka} = [\boldsymbol{MP}]/([\boldsymbol{M}] * [\boldsymbol{P}]) \quad \textbf{(II.4)}$$

II.4.2. Sites de fixation

La liaison des médicaments aux protéines s'effectue sur des sites localisés à la surface des protéines. On appelle généralement n le nombre de sites de fixation. Il peut exister plusieurs familles de sites sur une protéine et, pour un médicament donné, une protéine offre un ou plusieurs sites de fixation appartenant à une ou plusieurs classes, chacune caractérisée par des affinités différentes. Des phénomènes de compétition de fixation peuvent se développer entre plusieurs médicaments lorsqu'une protéine ne possède qu'un nombre limité de sites communs aux substances présentes. Si deux médicaments, de même nature, se fixant sur les mêmes sites, sont administrés simultanément, celui qui a la plus forte affinité se fixant préférentiellement, déplace ou empêche la fixation de l'autre. Ce phénomène de compétition peut aussi survenir avec des substances endogènes, telles que la bilirubine par exemple, qui est aussi fixée à l'albumine et peut entrer en compétition avec les médicaments acides, entraînant une défixation protéique de ces médicaments [17].

II.4.3. Pourcentage de fixation

Le paramètre de liaison le plus utilisé est le pourcentage de liaison du médicament aux protéines plasmatiques. La fraction liée aux protéines (*fb*) et la fraction non liée ou libre (*fu*) auront des valeurs comprises entre 0 et 100%. Si la valeur de *fb* est supérieure à 90%, on considère que la liaison aux protéines est importante, alors que des valeurs de *fb* inférieures à 20% indiquent que la liaison est faible ou négligeable. Les valeurs de *fb* et *fu* seront déterminées par les relations suivantes :

$$\boldsymbol{fb} = ([\boldsymbol{Concentration\ de\ la\ forme\ liée}] \,/\, [\boldsymbol{Concentration\ des\ formes\ libre + liée}]) \quad \textbf{(II.5)}$$

$$\boldsymbol{fb} = \textbf{\textit{Concentration plasmatique totale * 100}} \quad \textbf{(II.6)}$$

$$\boldsymbol{fu} = \mathbf{100} - \boldsymbol{fb} \quad \textbf{(II.7)}$$

Ce pourcentage de fixation est directement fonction du produit n*[P]*Ka.

Avec n : le nombre de sites de fixation ; Ka : la constante d'affinité du médicament pour la protéine et [P] : la concentration molaire en protéine [17].

II.5. Conséquences des modifications de la fixation protéique

II.5.1. Conséquences pharmacocinétiques

Le volume de distribution d'un médicament dépend de la fixation plasmatique et de la fixation tissulaire :

$$\boldsymbol{V = VP + VT * fu/fuT} \qquad \textbf{(II.8)}$$

Avec *fu* : la fraction libre plasmatique ; *fuT* : la fraction libre tissulaire ; *VT* : le volume tissulaire et *VP* le volume plasmatique.

Pour les médicaments ayant un volume de distribution supérieur ou égal à 30 L (environ 0,5 L/kg), toute modification de la fraction libre plasmatique entraînera une modification du volume de distribution [18]. Cependant, les conséquences de cette modification de la fraction libre plasmatique seront faibles car la fixation tissulaire du médicament est très supérieure à la fixation aux protéines plasmatiques. En résumé, la fixation aux protéines plasmatiques n'est un facteur limitant la distribution d'un médicament que si elle s'effectue avec une affinité plus forte que celle de la fixation tissulaire. Une modification de la fraction libre peut également influencer la vitesse d'élimination d'un médicament éliminé principalement par biotransformations hépatiques et donc sa clairance hépatique définie par l'équation suivante :

$$\boldsymbol{CLH = (QH * fu * CLi)/(QH + fu * CLi)} \qquad \textbf{(II.9)}$$

Avec *ClH* : la clairance hépatique ; *QH* : le débit sanguin hépatique ; *Cli* : la clairance intrinsèque et *fu* : la fraction libre [18] . Pour les médicaments à coefficient d'extraction hépatique élevé ($E > 0,7$), $QH << fu \times Cli$, la clairance hépatique est indépendante de la fraction libre et ne dépend que du débit hépatique *QH*. Pour les médicaments à coefficient d'extraction hépatique faible ($E < 0,3$), $QH >> fu \times Cli$, la clairance dépend de la fraction libre du médicament. Toute élévation de la fraction libre augmentera la clairance [18]. Enfin, la filtration glomérulaire d'un médicament est limitée par sa fixation protéique. La clairance rénale des médicaments excrétés uniquement par

filtration glomérulaire est directement liée à la fraction libre du médicament et au débit de filtration glomérulaire. La filtration glomérulaire est d'autant plus faible que la fixation protéique est élevée. Ainsi, des variations dans la concentration d'albumine peuvent entraîner des modifications des paramètres pharmacocinétiques des médicaments qui y sont liés [19].

II.5.2. Conséquences cliniques

Les conséquences pharmacocinétiques d'une modification de la fixation protéique n'auront pas toujours de répercussions au niveau clinique. En effet, une modification de la fixation aux protéines plasmatiques entraîne instantanément un changement de la concentration de médicament libre, de façon transitoire, jusqu'à ce que le complexe médicament-protéine se dissocie et libère le médicament libre, non fixé. Lors de cette dissociation du médicament des protéines plasmatiques, l'augmentation de la fraction libre s'accompagne d'un accroissement de la concentration libre, disponible pour l'élimination et la distribution tissulaire. Les deux formes libres et liées du médicament étant en équilibre réversible, au fur et à mesure de la disparition de la forme libre diffusant vers les tissus ou se trouvant éliminée, il y a passage d'une certaine proportion de forme liée sous forme libre. La concentration libre active du médicament reste inchangée, il est donc peu probable que l'activité du médicament soit modifiée [18]. Les conséquences pharmacocinétiques de la défixation protéique seront plus importantes pour un médicament fixé à plus de 90-95%, c'est-à-dire pour lequel une faible défixation protéique entraînera une grande augmentation de la fraction libre active.

II.6. Méthodes analytiques pour l'étude de la fixation protéique des médicaments

L'importance de la fraction libre du médicament (non liée aux protéines plasmatiques) pour la distribution dans les tissus pathologiques et les organes de clairance a incité de nombreuses sociétés de recherche biomédicale à mettre en œuvre des méthodes analytiques dans le but d'étudier la liaison des médicaments aux protéines plasmatiques. Ces techniques de séparation et d'étude sont très nombreuses, continuellement développées et bien décrites [20, 21]. Les approches utilisées pour étudier les interactions de la liaison des médicaments aux protéines plasmatiques sont divisées en approches séparatives et non séparatives (Figure II.4). Le premier groupe implique la séparation du ligand libre de l'espèce liée et est utilisé pour déterminer directement soit la concentration du médicament non lié, soit la concentration du médicament lié [22]. Le deuxième groupe s'appuie sur

détection d'une modification d'une propriété physico-chimique de soit le ligand soit la protéine en raison de la liaison [23].

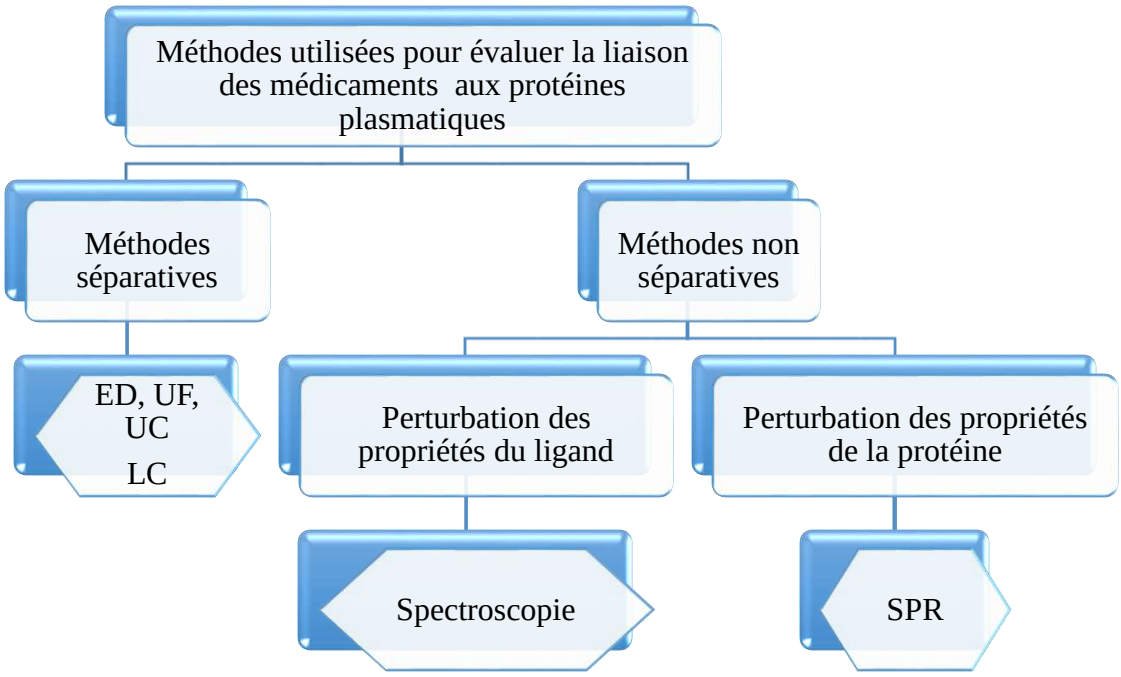


Figure II.4. Méthodes séparatives et non séparatives utilisées pour évaluer la liaison des médicaments aux protéines plasmatiques [24].

Plusieurs méthodes sont utilisées pour la détermination de la liaison des médicaments aux protéines plasmatiques : elles comprennent la microdialyse, la dialyse à l'équilibre ED, l'ultrafiltration UF, la dialyse dynamique, l'ultracentrifugation UC, chromatographie en phase liquide LC, la filtration sur gel, l'électrophorèse CE, la spectroscopie des plasmons de surface SPR, la spectrophotométrie et les méthodes cinétiques enzymatiques. Dans ce travail, nous avons détaillé trois techniques analytiques in vitro (Technique de précipitation, d'ultrafiltration et de dialyse à l'équilibre).

II.6.1. La microdialyse

La microdialyse est une méthode développée pour la mesure des concentrations de médicament non liées. C'est la seule méthode permettant de déterminer in vivo les concentrations de médicament non liées dans le plasma, les tissus et d'autres liquides biologiques tels que le

liquide céphalorachidien. Une sonde de microdialyse contenant une membrane de dialyse est implantée chirurgicalement dans un vaisseau sanguin, un espace contenant un fluide ou un espace tissulaire. Le tampon de dialysat est pompé à faible débit à travers la sonde et le médicament non lié dans le sang ou le liquide tissulaire diffuse à travers la membrane dans la sonde via un gradient de concentration. Le dialysat est collecté au fil du temps et analysé pour déterminer les concentrations de médicament non liées (*Cu*). Comme cette méthode ne mesure que les concentrations de médicament non liées, aucune information sur la concentration totale de médicament (*Ct*) ou la concentration de médicament liée (*Cb*) n'est obtenue à partir d'expériences de microdialyse. Pour déterminer *(Ct*), les échantillons de plasma doivent être prélevés et analysés séparément. *Cb* peut alors être calculé à partir de *Ct - Cu*. Etant donné que l'équilibre n'est pas atteint, il est essentiel que la récupération du médicament au cours de la microdialyse soit quantifiée. Un procédé *in vitro* y parvient en plaçant la sonde de microdialyse dans un tampon contenant une concentration connue de médicament (*C* tampon). La concentration du médicament du dialysat après perfusion de la solution (*C* dialysat) peut être déterminée et utilisée pour calculer la récupération. La récupération peut ensuite être calculée comme indiqué dans l'équation [25, 26] :

$$\boldsymbol{concentration\ de\ medicament\ non\ liées} = \frac{(\boldsymbol{C\ dialysat})}{(\boldsymbol{C\ tompan})} \qquad \textbf{(II.10)}$$

II.6.2. Précipitation

La précipitation des protéines à partir de fluides biologiques a été observée depuis des milliers d'années (par exemple, la précipitation de la caséine à partir du lait par un acide dilué). Cependant, le fractionnement du plasma humain par les méthodes de précipitations a été reporté au cours de la seconde guerre mondiale et a rendu possible la préparation de nombreux composants protéiques du plasma humain [27]. La précipitation des protéines est déterminée par quatre variables : le pH, la force ionique, la température et la concentration en protéines. De nombreuses techniques ont été développées pour la précipitation des protéines en modifiant un ou plusieurs de ces paramètres (par exemple en utilisant l'éthanol ou l'acide trichloracétique ATC) [28]. L'ATC conduit à une forte diminution du pH, entraînant une dénaturation et par conséquent une précipitation de la protéine. Une étude réalisée par Sivaraman et collaborateurs [29] a montré que les trois groupes chloro de la molécule jouent également un rôle important dans la précipitation des protéines.

II.6.3. Ultrafiltration

L'ultrafiltration est une technique rapide, efficace et simple pour la détermination de la fraction libre du médicament. L'avantage fondamental de cette méthode est qu'elle ne nécessite pas l'utilisation d'un tampon pour les mesures de liaison aux protéines et peut également être appliquée à différents types de matrices biologiques, y compris les homogénats tissulaires. L'ultrafiltration (UF) avec des membranes d'exclusion de tailles semi-perméables produit une séparation du médicament libre des macromolécules en utilisant un gradient de pression, qui force les petites molécules à traverser la membrane. Cette méthode de séparation du ligand lié du ligand libre est extrêmement populaire en raison de la manipulation facile d'un grand nombre d'échantillons, ainsi que de la disponibilité commerciale d'une variété de dispositifs de filtration [30]. Durant l'ultrafiltration d'une solution protéique contenant un ligand, le volume total de solutions d'ultrafiltrat plus de rétentat doit être égal au volume de la solution initiale. De plus, la concentration en ligand libre sera constante dans l'ultrafiltrat et le rétentat. Toute diminution de la concentration de ligand libre entraînerait une dissociation du ligand lié selon la loi de l'action de masse. Le rapport ligand lié / concentration de protéine libre dans le rétentat est indépendant du volume d'ultrafiltrat. Néanmoins, il a été montré que la concentration en ligand libre est constante au cours du procédé, à condition que le volume d'ultrafiltrat n'excède pas 40% du volume initial introduit [31].

II.6.4. Dialyse à l'équilibre

Dans la recherche fondamentale, la dialyse à l'équilibre (DE) est la méthode la plus fréquemment utilisée pour la détermination de la liaison des médicaments aux protéines plasmatiques. Bien qu'il n'y ait pas de «méthode standard» pour mesurer la liaison, la DE est souvent considérée comme la «méthode de référence» pour la détermination du profil de liaison médicament-protéine. Davis [32] et Klotz [33] ont été les premiers à utiliser la dialyse à l'équilibre pour étudier la liaison des ligands de masse moléculaire faible (PM inferieur a 1000) aux protéines. Cette méthode est encore largement utilisée pour déterminer les paramètres de liaison caractérisant l'interaction de ligands ou de médicaments avec des biopolymères tels que les protéines plasmatiques [15]. Cette technique utilise une cellule de dialyse qui contient 2 réservoirs séparés par une membrane de dialyse semi-perméable (Généralement en cellulose), cette membrane est disponible dans divers seuils de poids moléculaire et doit être choisie en fonction du poids moléculaire des composés et des protéines d'intérêt. Le plasma et le tampon sont placés dans leurs réservoirs respectifs séparés par la membrane de dialyse. Le système doit être incubé à la température physiologique (37°C) et laissé atteindre l'équilibre. Après avoir atteint l'équilibre, le

fluide de chaque réservoir est retiré et analysé pour la concentration du médicament. Puisque la concentration du médicament après la dialyse dans le réservoir tampon est en équilibre avec la concentration du médicament dans le réservoir de protéines, la concentration du médicament dans le tampon est dite libre (concentration unbound: Cu). Le temps d'équilibre est une variable importante dans la dialyse à l'équilibre. Pour que cette méthode soit valide, le médicament non lié dans les compartiments tampon et plasma doit être à l'équilibre [30]. Par conséquent, des expériences préliminaires devraient être effectuées pour déterminer le temps nécessaire pour atteindre l'équilibre. Au moment où aucun changement n'est détecté dans les concentrations de protéines et de tampon entre plusieurs points temporels, à ce moment-là, on peut dire qu‘un équilibre a été atteint. Il existe plusieurs variables qui doivent être contrôlées dans des expériences de dialyse à l'équilibre afin d'obtenir des résultats précis. Le pH plasmatique doit être ajusté au pH physiologique (7.4), car la liaison aux protéines plasmatiques peut dépendre du pH. La dialyse à l'équilibre étant une technique *in vitro*, la solution tampon utilisée n'est pas physiologiquement identique au sérum. Les ions dans le tampon peuvent affecter la liaison du médicament aux protéines plasmatiques. Des déplacements de fluide du compartiment tampon vers le compartiment plasmatique peuvent se produire en raison de la pression osmotique exercée par les protéines plasmatiques. Ceci entraîne une dilution des protéines plasmatiques. Des facteurs de correction mathématiques ont également été développés pour corriger les changements de fluides. Dans une expérience typique de dialyse à l'équilibre, s‘il n'y a pas d'altération du ligand ou de la (des) protéine (s), le système atteint un état d'équilibre connu sous le nom « d'état stationnaire», la concentration de ligand lié est égale à la différence entre les concentrations totales de ligand (liées et libres) des solutions présentes dans chaque compartiment. Selon la loi de Fick, la vitesse de diffusion du ligand est fonction de la surface et de l'épaisseur de la membrane, du gradient de concentration, des volumes des compartiments et du coefficient de diffusion. Cette dernière quantité, qui caractérise une molécule de ligand particulière, est fonction de la masse moléculaire du ligand et de la température (le taux de dialyse augmente avec la température). Evidemment, la membrane doit être choisie de manière à obtenir la plus grande vitesse de diffusion du ligand avec la capacité de rétention la plus élevée pour toutes les protéines.

II.7. Applications des approches *in silico* dans l'étude da la fixation protéique des médicaments

La liaison aux protéines plasmatiques joue un rôle clé dans la pharmacothérapie qui affecte la pharmacocinétique et la pharmacodynamique du médicament, car elles sont souvent directement liées à la concentration de médicament libre dans le plasma [34].

Ces dernières années, plusieurs modèles d'intelligence artificielle QSAR ont été développés pour prédire la liaison aux protéines plasmatiques. Dans ce paragraphe, il sera question d'une revue bibliographique sur les différentes méthodes *in sillico* utilisée pour prédire la valeur de la liaison des médicaments aux protéines plasmatiques. Cette section résume ces travaux, lesquels sont présentés par ordre chronologique de publication.

Dans une étude datant de 2000, Roustem Saiakhov et ses collaborateurs ont construits Une base de données comportant le taux de liaison aux protéines plasmatique de 154 médicaments et ils l'ont fait soumettre au programme M-CASE (The Multiple Computer-Automated Structure Evaluation program) qui a séparé l'ensemble de données total en groupes de médicaments ayant des caractéristiques structurelles communes. Ils ont supposé que les molécules d'un même sous-ensemble agissent par le même mécanisme. Ainsi Pour chacun de ces groupes, une relation quantitative structure-activité (QSAR) a été obtenue. Le processus est totalement automatisé et se déroule sans intervention humaine .ainsi, Le modèle, composé de plusieurs modèles QSAR locaux, était capable de prédire avec précision le pourcentage de liaison aux protéines pour 80% des médicaments inclus dans la base de données avec une erreur moyenne de seulement ~14 % [35].

Durant la même année, S. Agatonovic-Kustrin avec ces Co- auteurs ont mené une étude dont Le but était de développer un modèle de réseau neuronal génétique (GNN) pour prédire la quantité des médicaments transférée dans le lait maternelle. Des architectures de perceptron multicouche (MLP) ont été utilisées pour corréler l'activité avec les descripteurs qui ont été présélectionnés par un modèle d'algorithme génétique. Une Forte corrélation entre les valeurs prédites et expérimental a été observé (R^2>0.96) pour le meilleur modèle ANN (26-5-5-1) [36].

En 2002, Gilles Klopman et ses collègues ont développé un modèle basé sur une méthode appelé méthode de groupe de contribution modifiée dans laquelle les paramètres de base sont des descripteurs structuraux identifiés par le programme ' CAS'. Le modèle comprend 37 descripteurs structurels dérivés des structures chimiques d'un ensemble de données contenant 417 médicaments. Le modèle a pu prédire le pourcentage de médicament absorbé par le tractus gastro-intestinal avec un R^2 de 0,79 des composés de l'ensemble d'apprentissage [37].

Deux ans plus tard, Joseph VTurner avec ses collaborateurs ont réalisé une étude dans laquelle ils ont utilisé des réseaux de neurones artificiels (ANN) pour la prédiction des clairances, de la fraction liée aux protéines plasmatiques et du volume de distribution d'une série de composés. Un certain nombre de descripteurs théoriques ont été générés à partir des structures de médicaments. La corrélations pour les composés de test était de 0,992 Pour la liaison au protéines plasmatiques [38].

En 2006, Dans leur travail, Junmei Wang et ses amis chercheurs ont appliqué un algorithme génétique (GA) pour construire un ensemble de modèles QSPR (relation quantitative structure-propriété) pour la biodisponibilité orale absolue humaine, la liaison aux protéines plasmatiques et l'excrétion urinaire en utilisant le nombre de fragments moléculaires comme descripteurs [39].

Durant la même période Joseph R. Votano et ses confrères ont employé Quatre techniques de modélisation, utilisant des descripteurs topologiques pour représenter la structure moléculaire, pour produire des modèles de liaison aux protéines sériques humaines sur un ensemble de données de 1008 valeurs expérimentales, soigneusement sélectionnées à partir de sources accessibles au public. Les données ont été réparties en un ensemble d'apprentissage de 808 composés et un ensemble de test de validation externe de 200 composés. Les quatre techniques de modélisation comprennent la régression linéaire multiple (MLR), les réseaux de neurones artificiels (ANN), les k-plus proches voisins (kNN) et les machines à vecteurs de support (SVM). Les coefficients de corrélation des ensembles d'apprentissage et l'erreur absolue moyenne variaient de R^2 = 0,90 et MAE = 7,6 pour ANN à R^2 = 0,61 et MAE = 16,2 pour MLR. Les résultats de prédiction de l'ensemble de validation ont produit des coefficients de corrélation et des erreurs absolues moyennes allant de R^2 = 0,70 et MAE = 14,1 pour ANN à un minimum de R^2 = 0,59 et MAE = 18,3 pour le modèle SVM [40].

Un an plus tard, Tiago L. Moda et ses camarades ont utilisé une méthode appelée Hologram quantitative structure-activity relationships (HQSAR). HQSAR ne nécessite que des structures 2D et la valeur de la propriété d'intérêt comme entrées. Les analyses, les calculs et les visualisations de la modélisation HQSAR ont été effectués à l'aide du package SYBYL 7.3 (TriposInc., Saint-Louis, États-Unis). Le meilleur modèle statistique (Q^2 = 0,72, R^2 = 0,91) a été utilisé pour prédire le PPB de 62 composés de test [41].

Dans une étude, datant de 2008, Chang-Ying Ma et ses partenaires ont utilisé une méthode appelé (GA – CG – SVM) : les machine à vecteurs de support (SVM) combinée à l'algorithme

génétique (GA) pour la sélection des caractéristiques et la méthode du gradient conjugué (CG) pour l'optimisation des paramètres afin de faire la prédiction de la protéine plasmatique humaine (PPBR).ils ont utilisé Five-fold cross-validation pour la validation interne qui a donné une précision de 86%. Tandis que le validation externe a donné une précision de 81% [42].

En 2011, Une nouvelle méthode adopté par Haiyan Li et ses collaborateurs, appelée analyse QSAR issu de l'interaction des protéines plasmatiques (PPI-QSAR) a été utilisée pour construire les modèles QSAR afin de prédire la liaison aux protéines plasmatiques humaines. Le modèle in silico pour un ensemble de données de 65 antibiotiques a été construit par la méthode de régression linéaire multiple et validé par l'analyse résiduelle. Les valeurs de R^2 et Q^2 pour l'ensemble de données complet étaient de 0,87 et 0,77, respectivement, et pour l'ensemble d'apprentissage ils étaient de 0,86 et 0,72, respectivement [43].

En 2012, ZVETANKA ZHIVKOVA et ses Co-auteurs ont focalisé leurs études sur un ensemble de 132 médicaments acides liés aux protéines plasmatiques. Ils ont développé une relations quantitatives structure-protéine plasmatique (QSPPBR) pour prédire leur fraction non liée dans le plasma (*fu*) à l'aide de 178 descripteurs moléculaires à partir des logiciels ACD/ LogD version 9.08 (Advanced Chemistry Development Inc., Ontario, Canada) et MDL QSAR version 2.2 (MDL Information Systems, Inc., San Leandro, California). La sélection des variables s'est faites par algorithme génétique et le modèle a été établie par la régression pas à pas. Le modèle final a une valeur R^2 de 0,771, une valeur Q^2 de 0,737 [44].

En 2013, Taravat Ghafourian et ses collègues ont sélectionnés Les valeurs de liaison aux protéines pour 794 composés à partir de la littérature. Les données ont été réparties en un ensemble d'apprentissage de 662 composés et un ensemble de validation externe de 132 composés. Les descripteurs ont été calculés pour chaque composé à l'aide des logiciels ACD labs/logD, MOE (Chemical Computing Group) et Symyx QSAR. Ils onts construits leurs modèles utilisant une analyse de régression pas à pas, des arbres de classification et de régression (CART), des arbres boostés et les forêts aléatoire. Le meilleur modèle était celui de modèle d'arbre de régression boosté qui avait une erreur absolue moyenne pour l'ensemble d'apprentissage de 13,25 et pour l'ensemble de validation de 14,96 [45].

En 2014, abhishek singh chauhan et ses adjoints ont séparé leurs base de donne de 715 médicaments en deux séries contenant les médicaments a haut valeur de ppb et ceux ayant une faible valeur de ppb. Ils ont utilisés les logiciels TSAR, DRAGON et SCHRODINGER pour extraire 295 descripteurs puis ils ont sélectionné 95 descripteurs à travers différentes méthodes

statistiques qu'ils ont utilisés comme entrée pour un algorithme SVM et un modèle ANN. Dans la méthode ANN, ils ont obtenu une précision de 95 % pour les données d'apprentissage et 86% pour les données de test. Tandis que pour la méthode SVM, ils ont obtenus une précision de 90 % pour les données d'apprentissage et 80 % pour les donnes données de test [46].

En 2016, Anu Grover et ses camarades ont sélectionné une base de donne de 39 céphalosporines à partir de la littérature. Seuls les descripteurs significatifs calculés par QikProp et CODESSA ont été prises dans l'étude de RLM (régression linéaire multiple). Ils ont obtenue d'Excellentes corrélations pour les liaisons aux protéines pour la série des céphalosporines qui a utilisé plus de 4 descripteurs. Avec un R^2 élevé (~0,82) et un Q^2 (~0,54) [47].

En 2017, Ning-NingWang et ses compagnons ont collecté une base de donnes de 1830 médicaments. Ils ont utilisé la combinaison d'un algorithme génétique (NSGA-II) avec la régression des moindres carrés ((PLS) pour la sélection des descripteurs les plus pertinents qu'ils ont utilisés comme entrée pour cinq modèles prédictives (random forest (RF), support vector machine (SVM), Cubist, Gaussian process (GP), et Boosting). Ils ont obtenu un modèle de consensus (pour une validation croisée quintuple : Q^2 = 0,750 ; RMSE = 16,151). En outre, un ensemble de tests et deux ensembles de données de validation externes ont été appliqués pour valider sa robustesse et sa praticité. Pour l'ensemble de test, R^2 = 0,787 et RMSE = 14,154 ; lorsque deux ensembles de données externes ont été appliqués, ils ont obtenus R^2 = 0,704 et 0,703, RMSE = 18,194 et 17,233 respectivement [48].

En 2018, Lixia Sun et ses co-travailleur, dans leurs études, 967 valeurs de fractions liées aux protéines plasmatiques (fb) ont été utilisés pour construire des modèles de relation quantitative structure-activité (QSAR) par six algorithmes d'apprentissage avec 26 descripteurs moléculaires : random forest (RF), boost tree (BT), multiple linear regression (MLR), *k*-nearest neighbor (kNN), support vector regression (SVR), et multi-layer neural network (MNN). Les performances du modèle ont été estimées par une validation croisée décuplée et trois ensembles de validation externes comprenant respectivement 242 produits pharmaceutiques, 397 produits industriels et 231 produits chimiques nouvellement conçus. Les modèles ont montré d'excellentes performances pour l'ensemble du test, avec une erreur absolue moyenne (MAE) allant de 0,126 à 0,178 [49].

En 2018 aussi, Rajnish Kumar ses amis. Ont utilisé un ensemble de données de 736 médicaments pour faire des prédictions de la liaison des médicaments aux protéines plasmatiques en utilisant Support vector machine, Artificial neural network, k- nearest neighbor, Probabilistic neural network, Partial least square et Linear discriminant analysis. Le meilleur modèle était celui

utilisant Support vector machine. la précision de l'ensemble d'apprentissage, de l'ensemble de validation, l'exactitude, la précision, la sensibilité, la spécificité et le score F1 pour Support vector machine se sont révélés être 89,73 %, 89,97 %, 92,56 %, 87,26 %, 91,97 % et 0,898, respectivement [50].

En 2019, Cosimo Toma et all. Ont tiré 670 valeurs de plasma protéine binding de la littérature. Puis ils utilisé deux approches pour obtenir des modèles QSAR pour PPB. Le premier applique des algorithmes d'apprentissage automatique sur des descripteurs moléculaires basés sur les caractéristiques chimiques des composés. La deuxième approche a utilisé le logiciel CORAL qui implémente un algorithme d'extraction de descripteurs à partir d'une chaîne SMILES.les descripteurs obtenues ont été utilisés dans l'algorithme Random Forest pour faire la prédiction. La validation interne des modèles sélectionnés a donné des valeurs Q^2 proches de 0,60. La validation externe a également donné des valeurs de R^2 toujours supérieures à 0,60. Le modèle basé sur le descripteur CORAL pour était le meilleur, avec R^2 =0,74 en validation externe [51].

YawenYuan ses partenaires ont construit en 2020 un modèle global de relation quantitative structure-activité (QSAR) de PPB sur la base d'un vaste ensemble d'entraînement comprenant plus de 5000 composés pour représenter une grande diversité structurelle.ils ont proposé une nouvelle stratégie pour construire des modèles pour différents niveaux de liaison. Le meilleur modèle a donné une erreur absolue moyenne (MAE) de 0.076 sur l'ensemble de test qui est une valeur beaucoup plus faible que les modèles publiés et a donné une MAE égale à 0.041 et cela pour les niveaux élevé de liaison plasmatiques (0,8–1) [52].

En 2021, Jianan Li ses collègues ont développé une méthode de prédiction qu'ils ont appelé CycPeptPPB (cyclique peptife for plasma protéine binding) en utilisant les réseaux de neurone convolutionaire 1D-CNN. L'anneau macrocyclique des peptides cycliques a été décomposé. Les descripteurs basés sur les résidus de décomposition ont été organisés en fonction des informations obtenue du peptide cyclique. De plus, une méthode d'augmentation des données a été utilisée. CycPeptPPB a présenté d'excellentes performances, avec une erreur absolue moyenne (MAE) de 4,79 % et un coefficient de corrélation (R) de 0,92 [53].

En 2022, Chaofeng Lou et ses confrères ont introduit IDL-PPBopt, une stratégie de calcul pour prédire et optimiser la valeur de la liaison aux protéines plasmatiques (PPB) basée sur une méthode d'apprentissage en profondeur interprétable. Au début, un ensemble de données PPB organisé a été utilisé pour construire un modèle d'apprentissage en profondeur interprétable, qui a montré d'excellentes performances prédictives avec une erreur quadratique moyenne de 0,112 pour

l'ensemble de test. Ensuite, ils ont conçu un protocole de détection basé sur le modèle et sur le test de Wilcoxon pour identifier les sous-structures liées au PPB (nommées sous-structures privilégiées, PSubs) pour chaque molécule. Au total, 22 sous-structures privilégiées générales (GPSubs) ont été identifiées, qui partagent certaines caractéristiques.ils ont démontré que ces paires de sous-structures étaient également applicables en dehors de l'ensemble d'apprentissage et ont personnalisé en conséquence les schémas de modification structurelle pour chaque GPSub, ce qui a fourni des alternatives pour l'optimisation de la propriété PPB. Par conséquent, IDL-PPBopt fournit un schéma prometteur pour la prédiction et l'optimisation de la propriété PPB et serait utile pour l'optimisation principale d'autres propriétés pharmacocinétiques [54].

II.8. Conclusion

Les liaisons plasmatiques de forte affinité ont une incidence significative sur les autres paramètres cinétiques des médicaments. Actuellement, seules les liaisons à l'HSA et à l'AGP peuvent remplir ce rôle. Toute modification de leurs capacités de liaison peut entraîner alors des variations de distribution, d'élimination et d'effets (intensité et durée). Le plus souvent, les concentrations plasmatiques des protéines sont modifiées entraînant alors des variations de concentrations de forme libre des médicaments. On peut observer aussi une diminution des pourcentages de liaison alors que les concentrations de protéines ne sont pas ou peu modifiées : la présence de ligands endogènes ou exogènes explique la perte de capacité de fixation des médicaments. Il est aujourd'hui possible de programmer les liaisons plasmatiques d'une substance nouvelle. Le plus souvent, on cherche soit à prolonger la durée d'un effet en diminuant la vitesse de dissociation de la liaison, soit à limiter sa distribution par une liaison restrictive. On peut aussi choisir le site de liaison de la protéine pour rendre la liaison du médicament indépendante de celles des autres ligands présents dans le plasma des patients.

Références

[1] Tillement J-P, Duché J-C, Barré J. Liaisons des médicaments aux protéines circulantes: caractéristiques, rôles et modifications physio-pathologiques. Bulletin de l'Académie nationale de médecine 2006;190:935-47.

[2] Mohammad S. PHYSIOLOGY OF BLOOD2021.

[3] Loichot C, Grima M. Faculté de Médecine de Strasbourg, Module de Pharmacologie Générale DCEM1. 2005.

[4] Rowland M, Tozer TN. Clinical pharmacokinetics/pharmacodynamics: Lippincott Williams and Wilkins Philadelphia; 2005.

[5] Mahmood I. Interspecies scaling: role of protein binding in the prediction of clearance from animals to humans. The Journal of Clinical Pharmacology 2000;40:1439-46.

[6] Yamasaki K, Chuang VTG, Maruyama T, Otagiri M. Albumin–drug interaction and its clinical implication. Biochimica et Biophysica Acta (BBA)-General Subjects 2013;1830:5435-43.

[7] Parodi A, Miao J, Soond SM, Rudzińska M, Zamyatnin Jr AA. Albumin nanovectors in cancer therapy and imaging. Biomolecules 2019;9:218.

[8] Matsuda R, Anguizola J, Joseph K, Hage DS. Analysis of drug interactions with modified proteins by high-performance affinity chromatography: binding of glibenclamide to normal and glycated human serum albumin. Journal of Chromatography A 2012;1265:114-22.

[9] Hinderling PH, Hartmann D. The pH dependency of the binding of drugs to plasma proteins in man. Therapeutic drug monitoring 2005;27:71-85.

[10] Strolin Benedetti M, Baltes E. Drug metabolism and disposition in children. Fundamental & clinical pharmacology 2003;17:281-99.

[11] Kearns GL, Abdel-Rahman SM, Alander SW, Blowey DL, Leeder JS, Kauffman RE. Developmental pharmacology—drug disposition, action, and therapy in infants and children. New England Journal of Medicine 2003;349:1157-67.

[12] Soobis SW. Alteration of plasma albumin in relation to decreased drug binding in uremia. Clinical Pharmacology & Therapeutics 1977;22:147-53.

[13] Kopecký Jr Vr, Ettrich R, Hofbauerová K, Baumruk V. Structure of human α1-acid glycoprotein and its high-affinity binding site. Biochemical and biophysical research communications 2003;300:41-6.

[14] Kremer JM, Wilting J, Janssen L. Drug binding to human alpha-1-acid glycoprotein in health and disease. Pharmacological reviews 1988;40:1-47.

[15] Wright JD, Boudinot FD, Ujhelyi MR. Measurement and analysis of unbound drug concentrations. Clinical pharmacokinetics 1996;30:445-62.
[16] Rolan P. Plasma protein binding displacement interactions--why are they still regarded as clinically important? British journal of clinical pharmacology 1994;37:125.
[17] CHEMLAL L. ETUDE DE LA FIXATION AUX PROTEINES PLASMATIQUES DU 99MTC-DTPA: PRODUIT RADIOPHARMACEUTIQUE UTILISE EN SCINTIGRAPHIE RENALE. 2019.
[18] Benet LZ, Hoener B-A. Changes in plasma protein binding have little clinical relevance. MOSBY, INC 11830 WESTLINE INDUSTRIAL DR, ST LOUIS, MO 63146-3318 USA; 2002. p. 115-21.
[19] Benichou AS, Blanchet B, Conti F, Hornecker M, Bernard D, Taieb F, et al. Variability in free mycophenolic acid exposure in adult liver transplant recipients during the early posttransplantation period. The Journal of Clinical Pharmacology 2010;50:1202-10.
[20] Oravcova J, Bo B, Lindner W. Drug-protein binding studies new trends in analytical and experimental methodology. Journal of Chromatography B: Biomedical Sciences and Applications 1996;677:1-28.
[21] Gleeson MP. Plasma protein binding affinity and its relationship to molecular structure: an in-silico analysis. Journal of medicinal chemistry 2007;50:101-12.
[22] Sebille B, Zini R, Madjar C-V, Thuaud N, Tillement J-P. Separation procedures used to reveal and follow drug—protein binding. Journal of Chromatography B: Biomedical Sciences and Applications 1990;531:51-77.
[23] Klotz IM. Physicochemical aspects of drug-protein interactions: a general perspective. Annals of the New York Academy of Sciences 1973;226:18-35.
[24] Vuignier K, Schappler J, Veuthey J-L, Carrupt P-A, Martel S. Drug–protein binding: a critical review of analytical tools. Analytical and bioanalytical chemistry 2010;398:53-66.
[25] Sarre S, Van Belle K, Smolders I, Krieken G, Michotte Y. The use of microdialysis for the determination of plasma protein binding of drugs. Journal of pharmaceutical and biomedical analysis 1992;10:735-9.
[26] Quellec AL, Dupin S, Tufenkji AE, Genissel P, Houin G. Microdialysis: an alternative for in vitro and in vivo protein binding studies. Pharmaceutical research 1994;11:835-8.
[27] Burgess RR. Protein precipitation techniques. Methods in enzymology 2009;463:331-42.
[28] Zellner M, Winkler W, Hayden H, Diestinger M, Eliasen M, Gesslbauer B, et al. Quantitative validation of different protein precipitation methods in proteome analysis of blood platelets. Electrophoresis 2005;26:2481-9.

[29] Sivaraman T, Kumar T, Jayaraman G, Yu C. The mechanism of 2, 2, 2-trichloroacetic acid-induced protein precipitation. Journal of protein chemistry 1997;16:291-7.
[30] Vogel HG, Maas J, Hock FJ, Mayer D. Drug discovery and evaluation: safety and pharmacokinetic assays: Springer; 2013.
[31] Basken NE, Mathias CJ, Lipka AE, Green MA. Species dependence of [64Cu] Cu-Bis (thiosemicarbazone) radiopharmaceutical binding to serum albumins. Nuclear medicine and biology 2008;35:281-6.
[32] Davis BD. The binding of sulfonamide drugs by plasma proteins. A factor in determining the distribution of drugs in the body. The Journal of Clinical Investigation 1943;22:753-62.
[33] Klotz IM. Spectrophotometric investigations of the interactions of proteins with organic anions. Journal of the American Chemical Society 1946;68:2299-304.
[34] Bohnert T, Gan L-S. Plasma protein binding: from discovery to development. Journal of pharmaceutical sciences 2013;102:2953-94.
[35] Saiakhov RD, Stefan LR, Klopman G. Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs. Perspectives in Drug Discovery and Design 2000;19:133-55.
[36] Agatonovic-Kustrin S, Tucker I, Zecevic M, Zivanovic L. Prediction of drug transfer into human milk from theoretically derived descriptors. Analytica chimica acta 2000;418:181-95.
[37] Klopman G, Stefan LR, Saiakhov RD. ADME evaluation: 2. A computer model for the prediction of intestinal absorption in humans. European journal of pharmaceutical sciences 2002;17:253-63.
[38] Turner JV, Maddalena DJ, Cutler DJ. Pharmacokinetic parameter prediction from drug structure using artificial neural networks. International journal of pharmaceutics 2004;270:209-19.
[39] Wang J, Krudy G, Xie X-Q, Wu C, Holland G. Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. Journal of chemical information and modeling 2006;46:2674-83.
[40] Votano JR, Parham M, Hall LM, Hall LH, Kier LB, Oloff S, et al. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure− information representation. Journal of medicinal chemistry 2006;49:7169-81.
[41] Moda TL, Montanari CA, Andricopulo AD. In silico prediction of human plasma protein binding using hologram QSAR. Letters in Drug Design & Discovery 2007;4:502-9.
[42] Ma C-Y, Yang S-Y, Zhang H, Xiang M-L, Huang Q, Wei Y-Q. Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA–CG–SVM method. Journal of pharmaceutical and biomedical analysis 2008;47:677-82.

[43] Li H, Chen Z, Xu X, Sui X, Guo T, Liu W, et al. Predicting human plasma protein binding of drugs using plasma protein interaction QSAR analysis (PPI-QSAR). Biopharmaceutics & Drug Disposition 2011;32:333-42.
[44] Zhivkova Z, Doytchinova I. Quantitative structure—plasma protein binding relationships of acidic drugs. Journal of pharmaceutical sciences 2012;101:4627-41.
[45] Ghafourian T, Amin Z. QSAR models for the prediction of plasma protein binding. BioImpacts: BI 2013;3:21.
[46] Chauhan AS, Raj U, Varadwaj PK. PREDICTION OF PLASMA PROTEIN BINDING AFFINITY BY SUPPORT VECTOR MACHINE AND ARTIFICIAL NEURAL NETWORK. 2014.
[47] Grover A, Grover M, Sharma K. Quantitative structure property relationship studies for predicting plasma protein binding properties of Cephalosporins.
[48] Wang N-N, Deng Z-K, Huang C, Dong J, Zhu M-F, Yao Z-J, et al. ADME properties evaluation in drug discovery: Prediction of plasma protein binding using NSGA-II combining PLS and consensus modeling. Chemometrics and Intelligent Laboratory Systems 2017;170:84-95.
[49] Sun L, Yang H, Li J, Wang T, Li W, Liu G, et al. In silico prediction of compounds binding to human plasma proteins by QSAR models. ChemMedChem 2018;13:572-81.
[50] Kumar R, Sharma A, Siddiqui MH, Tiwari RK. Prediction of drug-plasma protein binding using artificial intelligence based algorithms. Combinatorial chemistry & high throughput screening 2018;21:57-64.
[51] Toma C, Gadaleta D, Roncaglioni A, Toropov A, Toropova A, Marzo M, et al. QSAR development for plasma protein binding: influence of the ionization state. Pharmaceutical Research 2019;36:1-9.
[52] Yuan Y, Chang S, Zhang Z, Li Z, Li S, Xie P, et al. A novel strategy for prediction of human plasma protein binding using machine learning techniques. Chemometrics and Intelligent Laboratory Systems 2020;199:103962.
[53] Li J, Yanagisawa K, Yoshikawa Y, Ohue M, Akiyama Y. Plasma protein binding prediction focusing on residue-level features and circularity of cyclic peptides by deep learning. Bioinformatics 2022;38:1110-7.
[54] Lou C, Yang H, Wang J, Huang M, Li W, Liu G, et al. IDL-PPBopt: A Strategy for Prediction and Optimization of Human Plasma Protein Binding of Compounds via an Interpretable Deep Learning Method. Journal of Chemical Information and Modeling 2022.

Chapitre III

Développement des modèles QSAR pour la prédiction de la liaison des médicaments aux protéines plasmatiques

III.1. Introduction

La liaison aux protéines plasmatiques joue un rôle clé dans la thérapie médicamenteuse qui affecte la pharmacocinétique et la pharmacodynamique du médicament, car elles sont souvent directement liées à la concentration du médicament libre dans le plasma [1].

La construction de modèles *in silico* établissant une relation mathématique entre la structure des médicaments et les propriétés considérées est une étape importante dans la découverte de médicaments, car elle nous permet d'éviter la synthèse chimique et de réduire les tests de laboratoire couteux et fastidieux [2]. Cependant, comme constaté dans la revue bibliographique (voir chapitre I), la sélection de quelques descripteurs pertinents est une tâche difficile et nous avons remarqué aussi que toutes les idées de sélections étaient basées sur le traitement des descripteurs comme étant des vecteurs.

En ce sens, l'objectif de ce travail de doctorat est d'étudier la possibilité de trouver un autre processus de sélection des descripteurs les plus pertinents qui soit moins long et moins complexe.

Tout d'abord, nous avons construits une méthode 'Filtrer' en utilisant une corrélation inter-descripteurs pour la sélection des descripteurs les plus pertinents. Ensuite, nous avons effectués une méthode 'Wrapper' utilisant une régression pas à pas. Et en fin, nous avons élaboré une nouvelle méthode utilisant un réseau de neurone convolutif pré-entrainé pour extraire les caractéristiques les plus essentiels de notre entrés qui n'est rien d'autre que nos descripteurs redimensionner en matrices.

III.2. Procédure de développement des modèles prédictives

La figure III.1 présente un schéma simplifié de la procédure adopté pour la modélisation prédictive de f_b. Cette procédure comprend l'utilisation de trois méthodes : Deux premières méthodes dans lesquelles les descripteurs sont sous forme de vecteurs et utilisant deux modèles de sélection des descripteurs les plus pertinents et une troisième méthode dans laquelle on considère ses descripteurs comme étant des matrices qu'on fait entrer dans un réseau de neurone a convolution. Ce dernier considère l'entrée comme des pixels d'une image et nous permet d'extraire automatiquement les caractéristiques les plus essentiels de cette soi-disant image. Par la suite, on procède à la génération du modèle, qui permet l'établissement de relations statistiquement significatives entre la propriété ciblée (taux de fixation aux protéines plasmatique) et les valeurs obtenue par les trois méthodes.

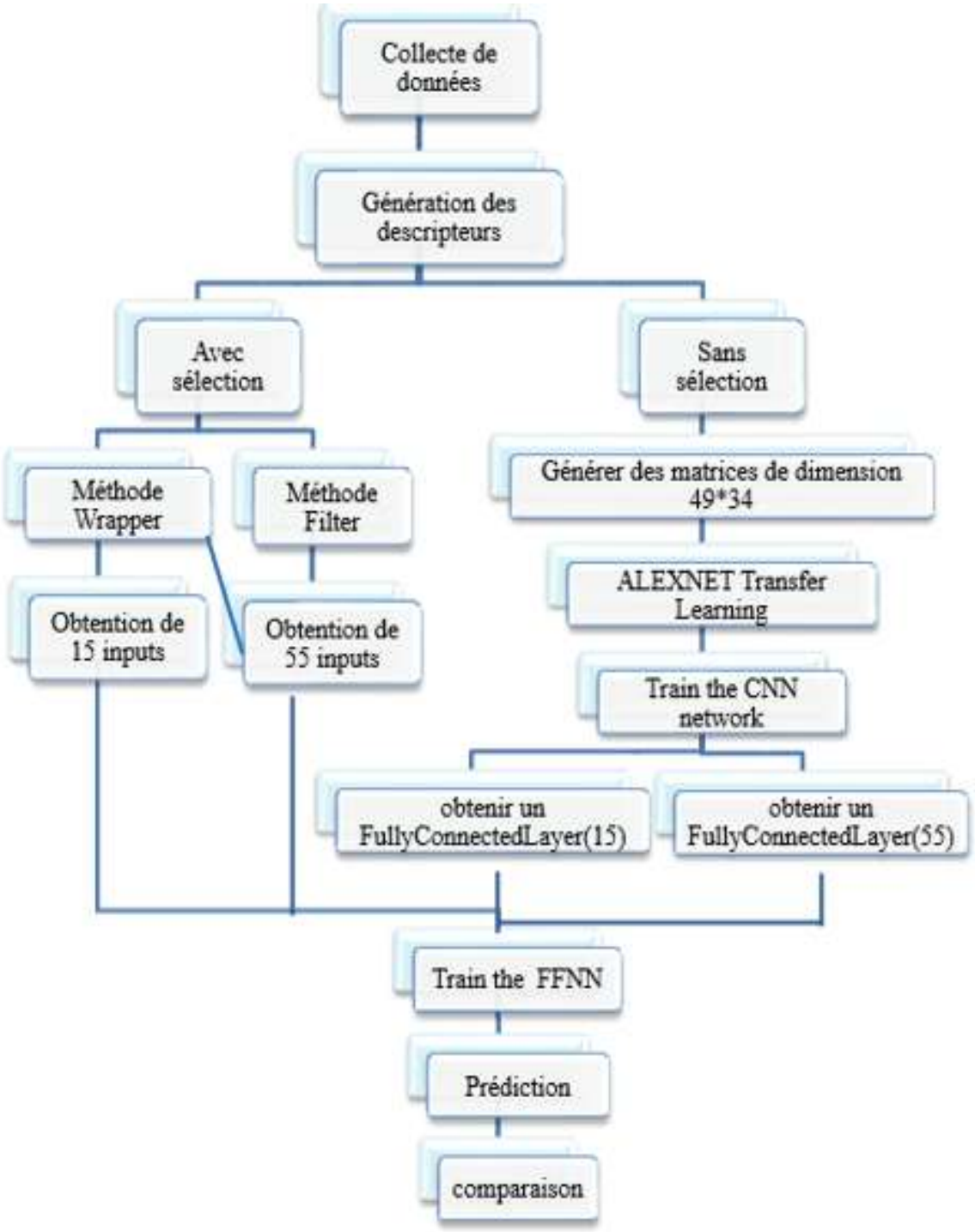


Figure III.1. Organigramme de la procédure suivie.

III.2.1. Collecte des données de la liaison des médicaments aux protéines plasmatique

La fraction expérimentale du médicament lié aux protéines plasmatiques (f_b) d'une série de 277 médicaments a été tiré de la littérature [3-10]. Pour résumer et comprendre la répartition des

valeurs de notre variable, nous avons utilisés un Box Plot que l'on appelle aussi une boite à moustache (voir figure III.2).

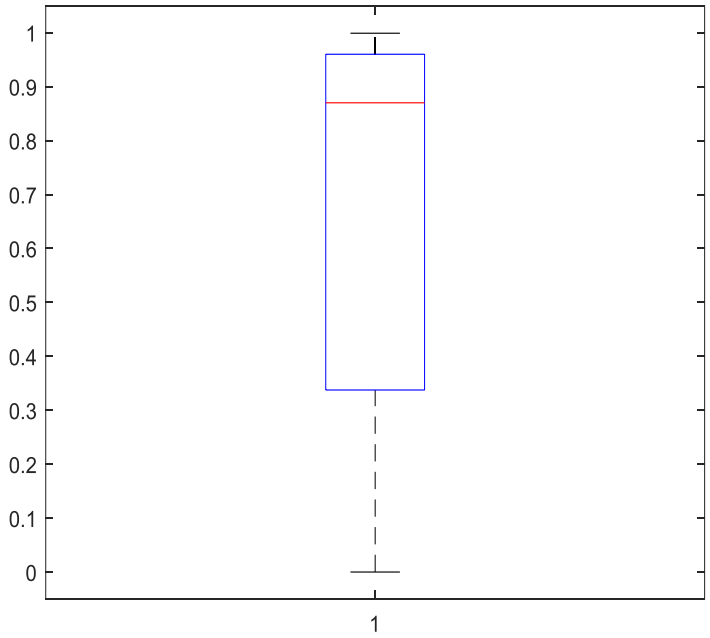


Figure III.2. Boite à moustache pour la répartition des valeurs de (f_b).

Les lignes qui s'étendent de la boite sont appelées moustaches. Ils représentent la variation des données qui est de 0 à 0.99. Ces données sont dissymétriques puisque la médiane de 0.87 est plus proche du 75 éme centile de 0.96 que du 25 éme centile de 0.33.

III.2.2. Calcul des descripteurs

La représentation numérique de la structure moléculaire a été évaluée en termes de descripteurs moléculaires. La notation SMILES nécessaire au calcul des descripteurs a été tirée à partir du site DrugBank database [11]. Ces notations ont été utilisées pour générer 1666 descripteurs répartis en vingt catégories: constitutional (0D), topological (2D), walk and path counts (2D), connectivity indices (2D), information indices (2D), two-dimensional autocorrelations (2D), edge adjacency indices (2D), Burden eigenvalues (2D), topological charge indices (2D), eigenvalue-based indices (2D), Randic molecular profiles (3D), geometrical descriptors (3D), RDF descriptors (3D), 3D-MoRSE descriptors (3D), WHIM descriptors (3D), GETAWAY descriptors (3D), functional group counts (1D), atom-centered fragments (1D), charge descriptors (1D), et molecular

properties (1D). Tous les descripteurs ont été obtenus par le biais du programme en ligne E-Dragon 1,0 [12]. Dans le tableau III.1, est répertorié le nombre de descripteurs obtenus par catégorie.

Tableau III.1. Nombre de descripteurs calculés [12].

Catégorie de descripteurs	Nombre par catégorie
Constitutional descriptors	48
Topological descriptors	119
Walk and pathcounts	47
Connectivity indices	33
Information indices	47
2D autocorrelations	96
Edgeadjacency indices	107
Burdeneigen value descriptors	64
Topological charge indices	21
Eigen value based indices	44
Randic molecular profiles	41
Geometrical descriptors	74
RDF descriptors	150
3D-morse descriptors	160
WHIM descriptors	99
GETAWAY descriptors	197
Functional group counts	154
Atomcentred fragments	120
Charge descriptors	14
Molecular properties	31
Total	**1666**

III.2.3. Sélection des descripteurs les plus influents pour la méthode 'Filter'

Le prétraitement de la base de données a pour finalité d'éliminer les descripteurs qui ne sont pas pertinents afin d'éviter le phénomène de sur-apprentissage. Il s'agit alors de la réduction de la dimension de l'espace des variables (descripteurs) du modèle avec l'élimination des entrées qui n'ont pas ou possèdent peu d'influence sur la sortie (taux de fixation aux protéines plasmatiques). Une procédure basée sur un prétraitement de la base de données a été utilisé pour la sélection des descripteurs les plus pertinents suivants les étapes suivantes [13] :

1. On calcule le minimum et le maximum pour chaque descripteur à l'aide d'un logicielle STATISTICA, puis on enlève les descripteurs qui ont le maximum et le minimum égaux.

2. On élimine le descripteur quasi-constants (1^{er} quartile 25%=$2^{éme}$ quartile 75%).

3. On calcule la déviation standard relative (*RSD*) pour chaque descripteur et on élimine ceux ayant des valeurs de *RSD* inférieures à 0.05.

4. On obtient alors une matrice qui représente la corrélation entre f_b et les descripteurs retenus. On classe les descripteurs selon la valeur décroissante du coefficient de corrélation. On prend le descripteur ayant la plus grande corrélation et on le compare aux autres descripteurs de la matrice. Ceux dont la valeur du coefficient de corrélation est supérieure à 0.75 sont éliminés à leur tour. On répète la même chose avec le descripteur classé juste après le premier et ainsi de suite. Pour un modèle plus robuste, nous avons calculé le facteur d'inflation de la variance *VIF* et tous les descripteurs possédants un *VIF* >5 ont été éliminé de la base de données.

III.2.4. Sélection des descripteurs les plus influents pour la méthode 'Wrapper'

Comme la méthode **'**Filter' est une méthode de pré-sélection, On utilise un programme basé sur la méthode régression pas à pas (forward stepwise regression) pour la sélection des descripteurs les plus pertinents parmi ceux obtenus à l'étape 4 à l'aide d'un logicielle STATISTICA.

III.2.5. Méthode CNN

Comme les réseaux de neurones convolutifs requièrent une matrice en entrée, nous avons redimensionnés tous les vecteurs des descripteurs pour chaque médicament (1666 descripteurs) en matrices de dimensions 49*34 en utilisant MATLAB.

III.2.6. Développement des modèles prédictives

La base de donnes contenant 277 médicaments a été divisée en trois sous-ensembles en utilisant la méthode holdout cross validation : un ensemble d'apprentissage contenant 70 % de l'échantillon, un ensemble de validation et de test avec 15% d'échantillon chacun. La méthode utilisée pour l'établissement des modèles QSAR est les réseaux de neurones artificiels (RNA). Les Réseaux de Neurones Artificiels (RNA) sont particulièrement bien adaptés pour les modèles QSAR / QSPR en raison de leur capacité à extraire des informations non-linéaires présentes dans la matrice de données [14]. Pour trouver un modèle non linéaire entre les descripteurs moléculaires sélectionnés et le taux de fixation aux protéines plasmatiques, l'utilisation de la méthode RNA est

nécessaire et est devenue une technique de modélisation importante. Elle est largement utilisée pour les études de QSAR non linéaires et est un outil puissant pour construire des modèles prédictifs. Le type de réseaux de neurones retenu pour ce travail est le perceptron multicouche (PMC). Le perceptron multicouche est un des réseaux de neurones les plus utilisés pour des problèmes d'approximation, de classification et de prédiction (voir chapitre I).

III.2.7. Évaluation et amélioration des performances prédictives des modèles

Une fois développé, le modèle doit être validé afin de vérifier sa stabilité, son pouvoir explicatif et son pouvoir prédictif.

III.2.7. 1. Validation interne

La validation croisée est une technique d'évaluation de modèle utilisé pour évaluer les performances d'un algorithme d'apprentissage en faisant des prédictions sur de nouveaux ensembles de données sur lesquels il n'a pas été formé. Cela se fait en partitionnant l'ensemble de données connu, en utilisant un sous-ensemble pour former l'algorithme et les données restantes pour les tests. Lors de l'entraînement d'un modèle, il est important de ne pas le sur-adapter ou le sous-adapter avec des algorithmes trop complexes ou trop simples. Le choix d'ensemble d'apprentissage et d'ensemble de test est essentiel pour réduire ce risque. Cependant, il est difficile de diviser l'ensemble de données pour maximiser à la fois l'apprentissage et la validité des résultats des tests. C'est là que la validation croisée entre en pratique. La validation croisée propose plusieurs techniques qui divisent les données différemment, pour trouver le meilleur algorithme pour le modèle. La validation croisée aide également à choisir le modèle le plus performant en calculant l'erreur à l'aide de l'ensemble de données de test, qui n'a pas été utilisé pour l'apprentissage. L'ensemble de données de test permet de calculer la précision du modèle et la manière dont il se généralisera avec les données futures. De nombreuses techniques sont disponibles pour la validation croisée. Parmi les plus courantes figurent :

k-fold : divise les données en k sous-ensembles choisis au hasard de taille à peu près égale. Un sous-ensemble est utilisé pour valider le modèle formé à l'aide des sous-ensembles restants. Ce processus est répété k fois de sorte que chaque sous-ensemble est utilisé exactement une fois pour la validation. Il s'agit de l'une des techniques de validation croisée les plus populaires, mais son exécution peut prendre beaucoup de temps car le modèle doit être entraîné à plusieurs reprises [15].

Holdout : partitionne les données de manière aléatoire en deux sous-ensembles de ratio spécifié pour l'apprentissage et le test (figure III.3). Cette méthode n'effectue l'apprentissage et le

test qu'une seule fois, ce qui réduit le temps d'exécution sur les grands ensembles, mais interprète l'erreur signalée avec prudence sur les petits ensembles de données [16].

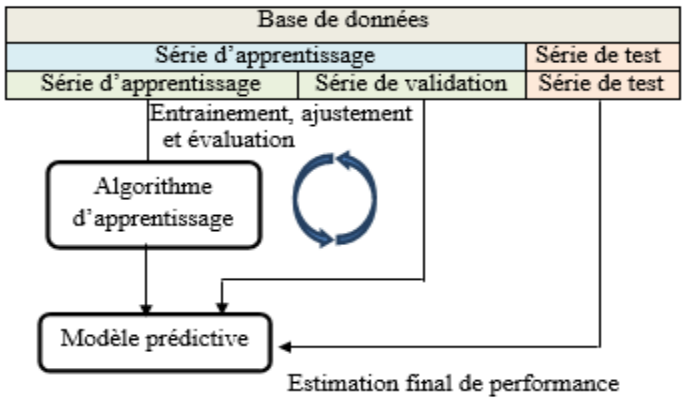


Figure III.3. Les étapes de la validation croisée Holdout.

Leaveout : partitionnez les données en utilisant l'approche k-fold où k est égal au nombre total d'observations dans les données et toutes les données seront utilisés une fois comme ensemble de test. Également connue sous le nom de validation croisée Leave-one-out (LOOCV) [17].

Sous-échantillonnage aléatoire répété : crée plusieurs partitions aléatoires de données à utiliser comme ensemble d'apprentissage et ensemble de test à l'aide de la méthodologie Monte Carlo et agrège les résultats sur toutes les exécutions. Cette technique a une idée similaire à celle du k-fold, mais chaque ensemble de tests est choisi indépendamment, ce qui signifie que certains points de données peuvent être utilisés pour tester plus d'une fois[18].

Stratifier : partitionner les données de manière à ce que les ensembles d'apprentissage et de test aient à peu près les mêmes proportions de classe dans la réponse ou la cible[19].

Afin de déterminer la fiabilité des modèles, des critères de validation internes ont été utilisés pour la validation des modèles et des critères de validation externes qui sont nécessaires pour déterminer à la fois la généralisabilité et la capacité prédictive réelles des modèles QSAR. Les principaux paramètres statistiques les plus importants utilisés dans ce travail pour vérifier les performances des modèles sont les suivants :

- L'erreur quadratique moyenne (en anglais **R**oot **M**ean **S**quares **E**rror *RMSE*).
- Le coefficient de correlation *R*.
- Le coefficient de détermination R^2_{train}.
- L'erreur absolue moyenne (en anglais Mean Absolute Error *MAE*).

- Erreur quadratique moyenne (en anglais Mean Squared Error *MSE*).
- Le coefficient de validation croisée Q^2_{cv}.

III.2.7.2. Validation externe

Il n'est pas suffisant de développer un modèle QSAR avec une excellente qualité d'ajustement et de prévision, mais il est également nécessaire de généraliser ces prévisions en l'appliquant sur un échantillon externe. Les prévisions résultantes pour le test set déterminent alors la validité externe du modèle. Les mesures quantitatives de la qualité de la prévision par validation externe peuvent être identiques à celles utilisées pour la validation interne. Les paramètres et critères de validations externes utilisées dans ce travail sont les suivants :

- L'erreur quadratique moyenne *RMSE*.
- Le coefficient de correlation *R*.
- Le coefficient Q^2_{test} .
- L'erreur absolue moyenne *MAE*.
- Erreur quadratique moyenne *MSE*.

Les équations de ses différents paramètres de validation interne et externe ainsi que la signification des différents termes sont données ci-dessous

$$\boldsymbol{RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}}} \quad \textbf{(III.1)}$$

$$\boldsymbol{R = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}} \quad \textbf{(III.2)}$$

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \textbf{(III.3)}$$

$$\boldsymbol{Q^2_{cv} = 1 - \frac{\sum_{i=1}^{n}(y_{exp(app)} - \hat{y}_{app})^2}{\sum_{i=1}^{n}(y_{exp(app)} - \bar{y}_{app})^2}} \quad \textbf{(III.4)}$$

$$\boldsymbol{Q^2_{test} = 1 - \frac{\sum_{i=1}^{n}(y_{exp(test)} - \hat{y}_{test})^2}{\sum_{i=1}^{n}(y_{exp(test)} - \bar{y}_{test})^2}} \quad \textbf{(III.5)}$$

$$\boldsymbol{MSE = \sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}} \quad \textbf{(III.6)}$$

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|} \quad \textbf{(III.7)}$$

$\widehat{y_i}$: sonts les valeurs prédites; y_i : les valeurs observées ; n: le nombre d'observation; $\bar{y}$: la moyenne des valeurs observées; $\overline{\widehat{y_i}}$: la moyenne des valeurs prédites.

De plus, d'autres paramètres doivent être vérifiés. Ces paramètres sont connus sous le nom «critères de validation externe » ou souvent appelés « critères de Trophsa » [20] .

$$(r^2 - r_0^2) < 0.3 \quad \textbf{(III.8)}$$

$$(r^2 - r_0^2)/r^2 < 0.1 \quad \textbf{(III.9)}$$

0.85<k< 1.15

$$(r^2 - r'^2_0)/r^2 < 0.1 \quad \textbf{(III.10)}$$

0.85<k'< 1.15

Avec

r^2 : Le carré du Coefficient de corrélation pour les molécules de la série de test ; r_0^2 : Le carré du coefficient de corrélation entre les valeurs prédites et expérimentales pour la série de test à zéro intercepte ; r'^2_0 : Coefficient de corrélation entre les valeurs expérimentales et prédites pour la série de test à zéro intercepte ; k : est la constante de la droite de corrélation (valeurs prédites en fonction des valeurs expérimentales) à zéro intercepte ; k′ : est la constante de la droite de corrélation (valeurs expérimentales en Fonction des valeurs prédites) à zéro intercepte.

IV.2.8. Domaine d'applicabilité

Un modèle idéale est celui qui est capable de prédire l'activité de n'importe quelle molécule imaginable. Malheureusement, cela est loin d'être toujours possible : la taille limitée du jeu d'entrainement rend l'espace chimique des modèles construits limitées. Pour prévenir ce type de problème on utilise le 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 [21-23].

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 ses collaborateurs 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.

III.3. Résultat et discussion

III.3.1. Analyse des descripteurs retenus

III.3.1.1. Méthode 'Filter'

La procédure de sélection des variables a généré 55 descripteurs (voir tableau III.2). Pour obtenir un aperçu de la structure de corrélation, nous avons utilisé une carte thermique (heat map) pour mettre en évidence ce qui est important (figure III.4). La valeur du coefficient de corrélation entre chaque paire de descripteurs est inférieur à 0,75 ce qui signifie que les descripteurs sélectionnés sont indépendants. En outre, toutes les variables ont des valeurs du facteur d'inflation de la variance (*VIF*) dans la gamme 1-5 (voir le tableau III.2), ce qui indique que le modèle obtenu à une signification statistique évidente [24] et les descripteurs se sont révélés être relativement orthogonaux.

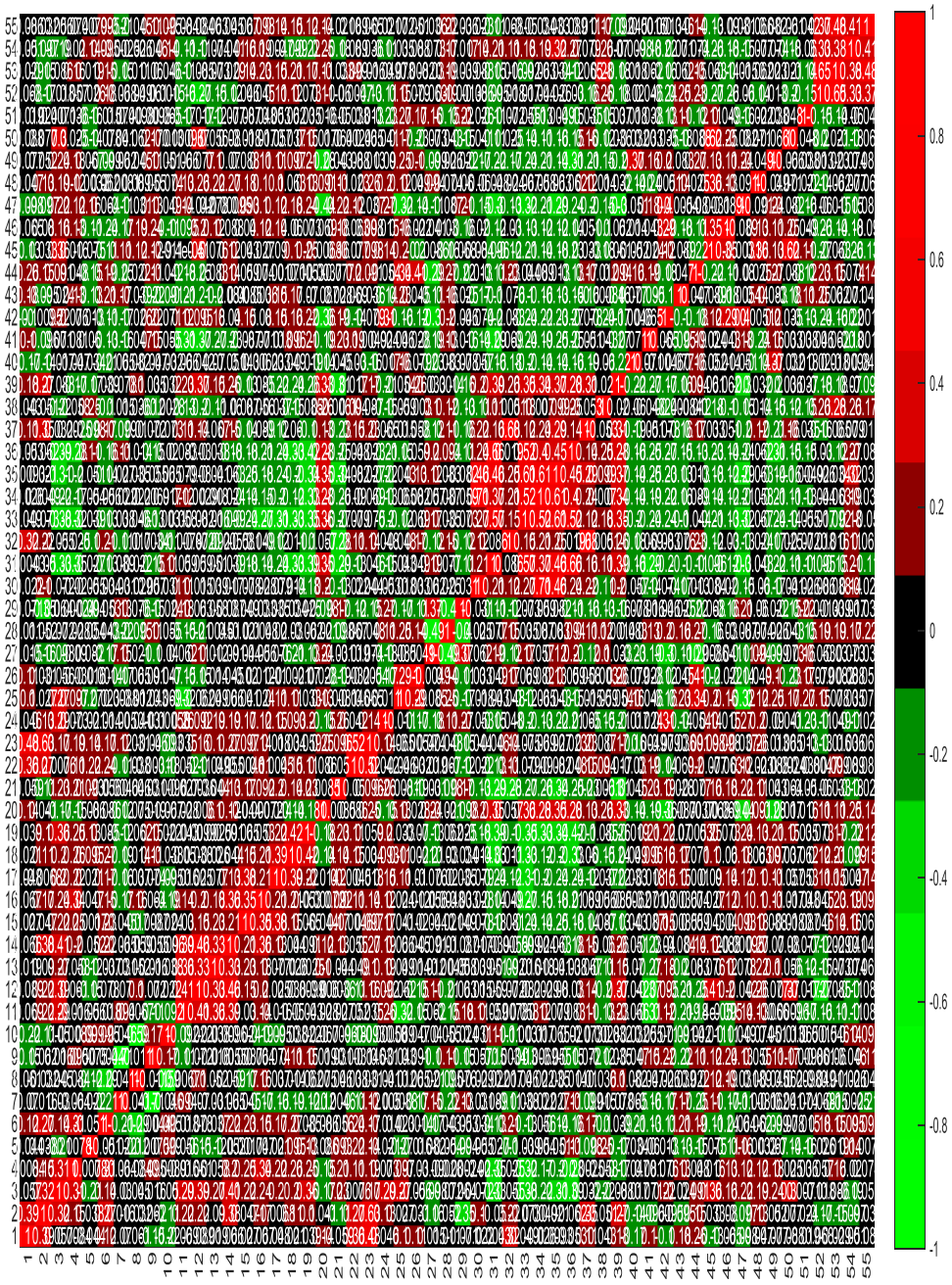


Figure III.4. Heatmap qui représente les coefficients de corrélation entre les entrée du réseau de neurone pour le modèle 'Filter'.

Tableau III.2. Notation, définition, type et *VIF* des descripteurs sélectionnés

N0	descripteur	type	description	*VIF*
1	nX	Constitutional descriptors	number of halogen atoms maximal	2.6615
2	MAXDN	Topological descriptors	electrotopological negative variation	4.4004
3	MAXDP	Topological descriptors	maximal electrotopological positive variation	3.5133
4	PJI2	Topological descriptors	2D Petitjean shape index	1.7543
5	Lop	Topological descriptors	Lopping centric index	1.9696
6	MATS1m	2D autocorrelations	Moran autocorrelation - lag 1 / weighted by atomic masses	2.9735
7	MATS2m	2D autocorrelations	Moran autocorrelation - lag 2 / weighted by atomic masses	4.0324
8	MATS4m	2D autocorrelations	Moran autocorrelation - lag 4 / weighted by atomic masses	2.6631
9	GATS2m	2D autocorrelations	Geary autocorrelation - lag 2 / weighted by atomic masses	3.3624
10	GATS4m	2D autocorrelations	Geary autocorrelation - lag 4 / weighted by atomic masses	2.7846
11	JGI2	Topological charge indices	mean topological charge index of order2	2.4942
12	JGI3	Topological charge indices	mean topological charge index of order3	3.5132
13	JGI4	Topological charge indices	mean topological charge index of order4	1.8913
14	JGI5	Topological charge indices	mean topological charge index of order5	2.3317
15	JGI6	Topological charge indices	mean topological charge index of order6	1.9848

16	JGI7	Topological charge indices	mean topological charge index of order7	2.0805
17	JGI8	Topological charge indices	mean topological charge index of order8	1.6600
18	JGI9	Topological charge indices	mean topological charge index of order9	2.0077
19	JGI10	Topological charge indices	mean topological charge index of order10	1.8634
20	FDI	Geometrical descriptors	folding degree index	2.7337
21	PJI3	Geometrical descriptors	3D Petitjean shape index	1.8725
22	DISPm	Geometrical descriptors	d COMMA2 value / weighted by atomic masses	2.3418
23	DISPe	Geometrical descriptors	d COMMA2 value / weighted by atomic Sanderson electronegativities	4.0247
24	Mor04m	3D-MoRSE descriptors	3D-MoRSE - signal 04 / weighted by atomic masses	2.4145
25	Mor12m	3D-MoRSE descriptors	3D-MoRSE - signal 12 / weighted by atomic masses	3.0770
26	Mor17m	3D-MoRSE descriptors	3D-MoRSE - signal 17 / weighted by atomic masses	1.8941
27	Mor26m	3D-MoRSE descriptors	3D-MoRSE - signal 26 / weighted by atomic masses	2.3693
28	Mor28m	3D-MoRSE descriptors	3D-MoRSE - signal 28 / weighted by atomic masses	2.5970

29	Mor31m	3D-MoRSE descriptors	3D-MoRSE - signal 31 / weighted by atomic masses	2.7935
30	G2u	WHIM descriptors	2st component symmetry directional WHIM index / unweighted	2.5765
31	G2m	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic masses	2.7232
32	E2m	WHIM descriptors	2nd component accessibility directional WHIM index / weighted by atomic masses	2.9830
33	G2v	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic van der Waals volumes	2.6210
34	G2e	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic Sanderson electronegativities	3.5147
35	G2p	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic polarizabilities	2.8308
36	G2s	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic electrotopological states	2.9183
37	E2s	WHIM descriptors	2nd component accessibility directional WHIM index / weighted by atomic electrotopological states	3.0706
38	ISH	GETAWAY descriptors	standardized information content on the leverage equality	1.7718
39	HATS4m	GETAWAY descriptors	leverage-weighted autocorrelation of lag 4 / weighted by atomic masses	3.2173
40	C-005	Atom-centred fragments	atom-centred fragments	2.0372
41	C-006	Atom-centred fragments	atom-centred fragments	2.3826
42	C-008	Atom-centred fragments	atom-centred fragments	2.7744
43	C-025	Atom-centred fragments	atom-centred fragments	2.4806

44	C-026	Atom-centred fragments	atom-centred fragments	3.2520
45	C-040	Atom-centred fragments	atom-centred fragments	3.5563
46	H-048	Atom-centred fragments	atom-centred fragments	1.9890
47	H-052	Atom-centred fragments	atom-centred fragments	2.3870
48	O-057	Atom-centred fragments	atom-centred fragments	2.2844
49	O-060	Atom-centred fragments	atom-centred fragments	2.7103
50	N-072	Atom-centred fragments	atom-centred fragments	2.5844
51	N-075	Atom-centred fragments	atom-centred fragments	2.0612
52	Inflammat-80	Molecular properties	Ghose-Viswanadhan-Wendoloski antiinflammatory at 80% (drug-like index)	2.6589
53	Hypertens-80	Molecular properties	Ghose-Viswanadhan-Wendoloski antihypertensive at 80% (drug-like index)	2.8495
54	Hypnotic-80	Molecular properties	Ghose-Viswanadhan-Wendoloski hypnotic at 80% (drug-like index)	2.4199
55	Neoplastic-50	Molecular properties	Ghose-Viswanadhan-Wendoloski antineoplastic at 50% (drug-like index)	1.8707

III.3.1.2. Méthode 'Wrapper'

Le nombre de descripteurs obtenus après la sélection effectuée était de 15. Afin de vérifier l'inter-corrélation des quinze descripteurs, l'analyse du facteur d'inflation de la variance *(VIF)* a été effectuée. D'après le tableau III.2, tous les descripteurs ont des valeurs *VIF* inférieures à 5, ce qui indique que la sélection effectuée est bonne et que les descripteurs obtenus ne sont pas corrélés entre eux. Pour obtenir un aperçu de la structure de corrélation, nous avons utilisé une carte thermique (heat map) pour mettre en évidence ce qui est important. La valeur du coefficient de corrélation de chaque paire de descripteurs est inférieur à 0,75 ce qui signifie que les descripteurs sélectionnés sont indépendants (figure III.5).

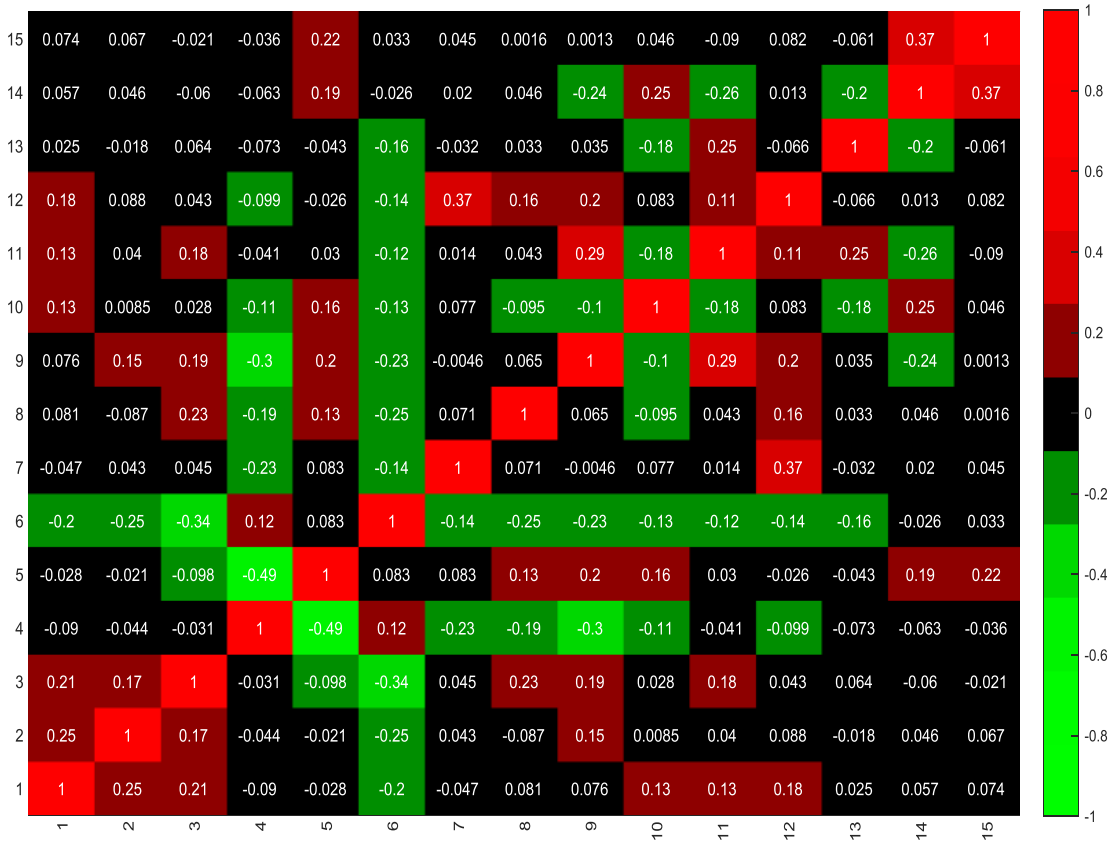


Figure III.5. Heatmap qui représente les coefficients de corrélation entre les entrée du réseau de neurone pour le modèle 'Wrapper'.

Tableau III.3. Notation, définition, type et *VIF* des descripteurs sélectionnés

N0	descripteur	type	description	*VIF*
1	PJI2	Topological descriptors	2D Petitjean shape index	1.2179
2	JGI6	Topological charge indices	mean topological charge index of order6	1.1909
3	PJI3	Geometrical descriptors	3D Petitjean shape index	1.2777
4	Mor26m	3D-MoRSE descriptors	3D-MoRSE - signal 26 / weighted by atomic masses	1.5615
5	Mor28m	3D-MoRSE descriptors	3D-MoRSE - signal 28 / weighted by atomic masses	1.6117
6	G2p	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic polarizabilities	1.4653
7	C-005	Atom-centred fragments	Atom-centred fragments	1.3155
8	C-006	Atom-centred fragments	Atom-centred fragments	1.2867
9	C-008	Atom-centred fragments	Atom-centred fragments	1.5295
10	C-025	Atom-centred fragments	Atom-centred fragments	1.2771
11	H-048	Atom-centred fragments	Atom-centred fragments	1.2713
12	O-060	Atom-centred fragments	Atom-centred fragments	1.3680
13	N-072	Atom-centred fragments	Atom-centred fragments	1.1847
14	Inflammat-80	Molecular properties	Ghose-Viswanadhan-Wendoloski hypnotic at 80% (drug-like index)	1.4384
15	Neoplastic-50	Molecular properties	Ghose-Viswanadhan-Wendoloski antineoplastic at 50% (drug-like index)	1.23866

III.3.2. Conception architectural

III.3.2.1. Méthodes 'Filter' et 'Wrapper'

Jusqu'à présent, il n'y a pas de réponse claire pour préciser le nombre de neurones cachés requis pour une tâche de modélisation. Cela reste encore une question ouverte. Il existe différentes approches pour connaître ce nombre, expliquées en détail dans les revues, parmi lesquelles une revue des méthodes de sélection du nombre de nœuds cachés dans le réseau de neurones artificiels [25]. Pour déterminer le nombre de neurones dans la couche cachée, nous avons utilisé la technique résumée ci-dessous [26].

1. Au début, nous n'utilisions que quelques neurones cachés, cinq.

2. Nous avons entraîné le réseau jusqu'à ce que l'erreur quadratique moyenne ne semble plus s'améliorer.

3. À ce moment, nous avons ajouté quelques (cinq) neurones à la couche cachée, chacun avec des poids initialisés de manière aléatoire, et reprenons l'entraînement.

4. Nous avons répété les deux étapes précédentes jusqu'à ce qu'un critère de terminaison soit satisfait.

Pour la méthode 'Filter', la meilleure performance du modèle avait une topologie de (55-40-1) : 55 nœud d'entrée, 1 couche cachée avec 40 nœud ayants la tangent hyperbolique comme fonction de transfert et purelin comme fonction de transfert pour la couche de sortie.

Pour l'approche 'Wrapper', la meilleure performance du modèle avait une topologie de (15-45-1) : 15 nœud d'entrée, 1 couche cachée avec 45 nœud ayants la tangent hyperbolique comme fonction de transfert ,1 couche de sortie avec purelin comme fonction de transfert.

Le réseau de neurones a été implémenté à l'aide de Neural Network Toolbox dans MATLAB. Le type de réseau était un réseau feed-forward avec la fonction d'entraînement de rétro-propagation de Levenberg-Marquardt et une descente de gradient avec une fonction d'apprentissage du poids et du biais. L'architecture du RNA optimisée est représentée dans la figure III.6.

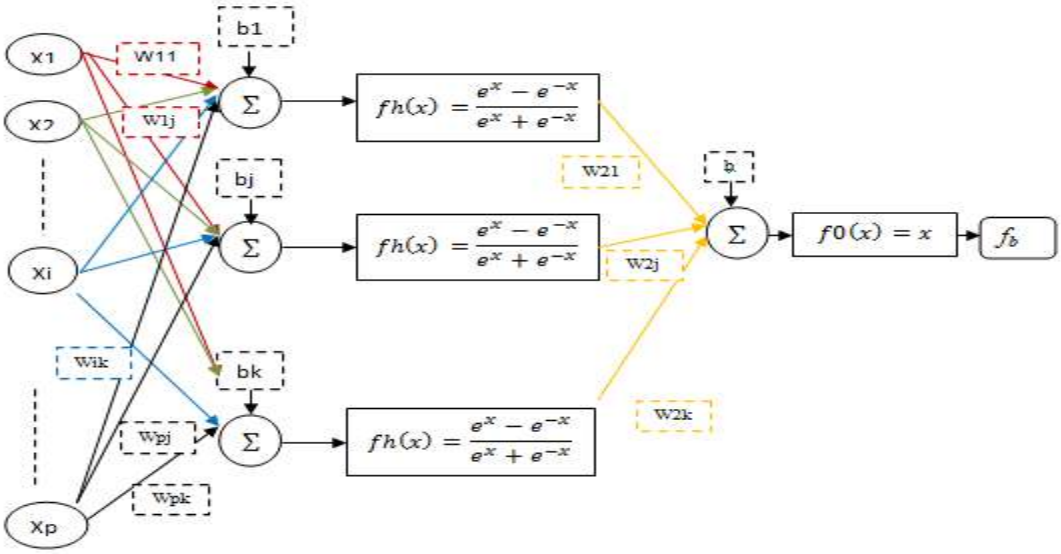


Figure III.6. L'architecture du RNA optimisée.

X_i ($i=1...P$) : sont les entrées qui correspondent au nombre de données incluses dans l'apprentissage du RNA ; i : allants de 1 à 55 pour la méthode 'Filter' et de 1 to 15 pour la méthode 'Wrapper' ; W_{ij} ($i=1...P$, $j=1...K$) : sont des poids de l'entrée à la couche cachée ; b_j ($j=1...K$) : sont des biais des neurones dans la couche cachée ; k=40 pour la méthode 'Filter' et 45 pour la méthode 'Wrapper' , w_{2j} (j $1...K$) : sont des fonctions de pondération de la couche cachée à la couche de sortie ; b : est le biais du neurone de sortie et $\boldsymbol{f_b}$ est la sortie. L'équation mathématique du modèle, qui peut être utilisée pour la prédiction de la liaison aux protéines, est présentée ci-dessous :

$$\boldsymbol{fb = \sum_{j=1}^{k} w2k\left(\frac{exp(\sum_{i=1}^{p} xi+wij+bj)-exp(-\sum_{i=1}^{p} xi+wij+bj)}{exp(\sum_{i=1}^{p} xi+wij+bj)+exp(-\sum_{i=1}^{p} xi+wij+bj)}\right) + b} \qquad \textbf{(III.11)}$$

III.3.2. 2. Méthode CNN

Le réseau de neurones convolutif est une puissante technique d'apprentissage automatique. L'apprentissage d'un CNN nécessite une grande collection de données diverses, ce qui n'est pas une tâche facile. Cependant, il existe un moyen simple, nous pouvons utiliser un réseau neuronal convolutif pré-entraîné, cela permet d'économiser énormément de temps et d'efforts, car l'apprentissage d'un réseau par transfert Learning est généralement beaucoup plus rapide et plus facile que l'apprentissage d'un réseau à partir de zéro. Et c'est ce que nous avons fait dans ce

travail. Nous avons appliqué l'apprentissage par transfert à l'aide du réseau de neurones convolutif ALEXNET. Le détail de la procédure est schématisé dans la figure III.7.

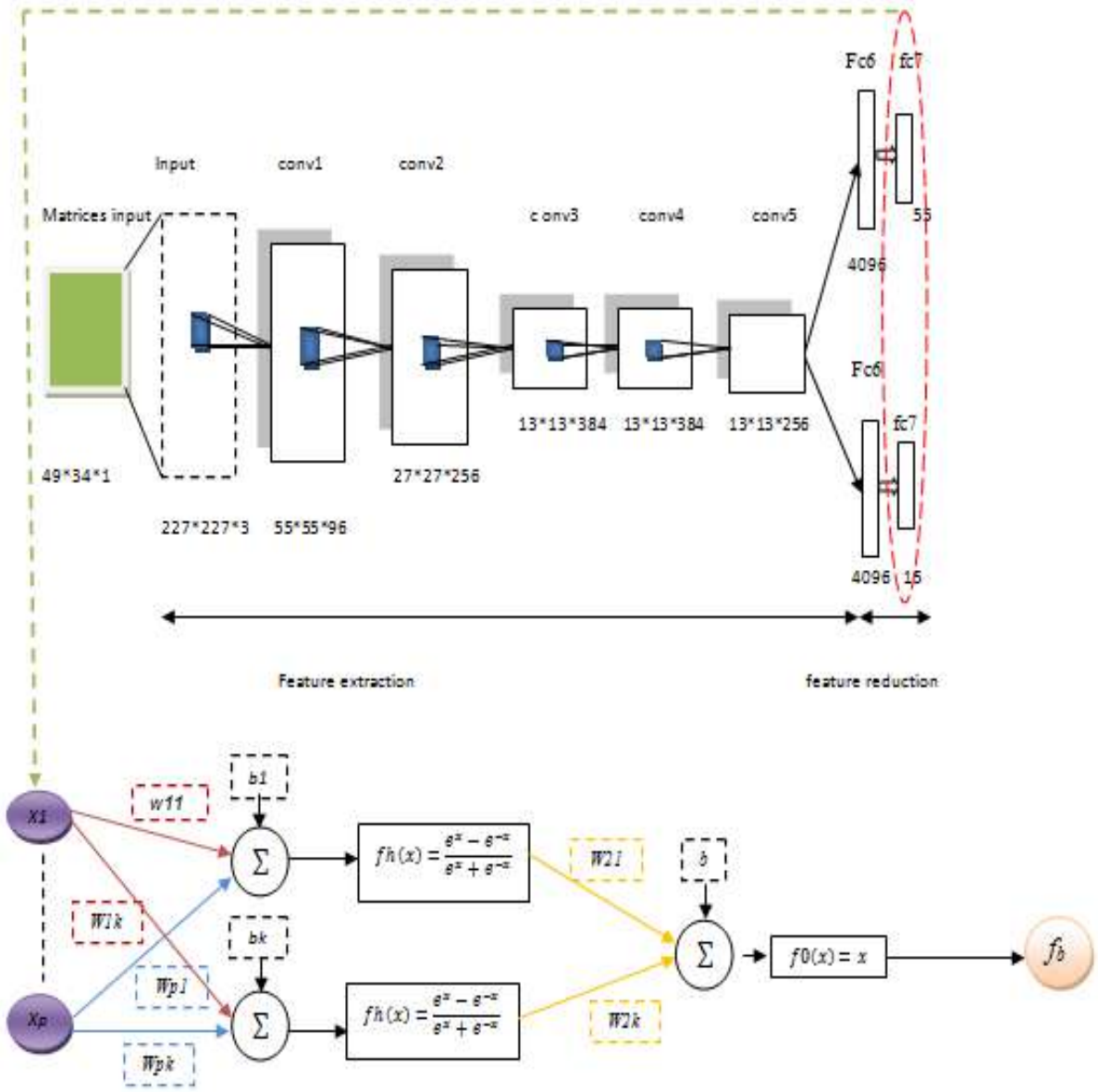


Figure III.7. Architecture de la procedure ‘Méthode CNN’.

III.3.3. Performance du modèle neuronal

III.3.3.1. Méthode 'filter' et méthode 'Wrapper'

Pour spécifier le nombre de neurones cachés requis, nous avons suivi la procédure détaillée ci-dessus. Les critères *R (*all*),* $\boldsymbol{R^2_{app}}$*, MSE* (validation) et $\boldsymbol{Q^2_{cv}}$ ont été utilisés pour l'évaluation de la performance optimale du modèle. Le meilleur modèle a été choisi en fonction du *MSE* (validation) minimum, du *R* (all) maximum et $\boldsymbol{Q^2_{cv}}$ maximum. Le tableau III.4 Montre 10 modèles de réseau qui ont été développés pour la méthode 'Filter'. Les résultats obtenus montrent que le réseau 8 avec 40 neurones est le meilleur modèle avec le plus grand coefficient de corrélation *R* (all)=0.99, $\boldsymbol{R^2_{app}}$ =0.981, $\boldsymbol{Q^2_{cv}}$ =0.989 et *MSE* (validation)=0.002. Par conséquent, la meilleure performance du modèle avait une topologie de (55-40-1) : 55 nœuds d'entrée, 1 couche cachée avec 40 nœuds ayant la tangente hyperbolique comme fonction de transfert, et une fonction d'identité dans la couche de sortie.

Tableau III.4. Critères de sélection de l'architecture du FFNN pour la méthode 'filter'**.**

Nombre de neurones dans la couche cachée	***R*(all)**	$\boldsymbol{R^2_{app}}$	$\boldsymbol{Q^2_{cv}}$	***MSE*(validation)**
5	0.849	0.743	0.707	0.039
10	0.857	0.729	0.664	0.032
15	0.870	0.774	0.743	0.031
20	0.872	0.780	0.714	0.038
25	0.917	0.839	0.780	0.026
30	0.957	0.918	0.882	0.020
35	0.955	0.953	0.818	0.024
40	**0.990**	**0.981**	**0.989**	**0.002**
45	0.944	0.901	0.875	0.014
50	0.832	0.694	0.714	0.027

Les valeurs d'apprentissage, de la validation et du test utilisant le modèle RNA ont été tracés par rapport aux valeurs expérimentales de la liaison aux protéines dans la figure III.8. Après avoir trouvé le modèle de régression et pour visualiser la plage d'erreur dans l'ensemble de données, nous avons tracé l'histogramme d'erreur (figure III.9).

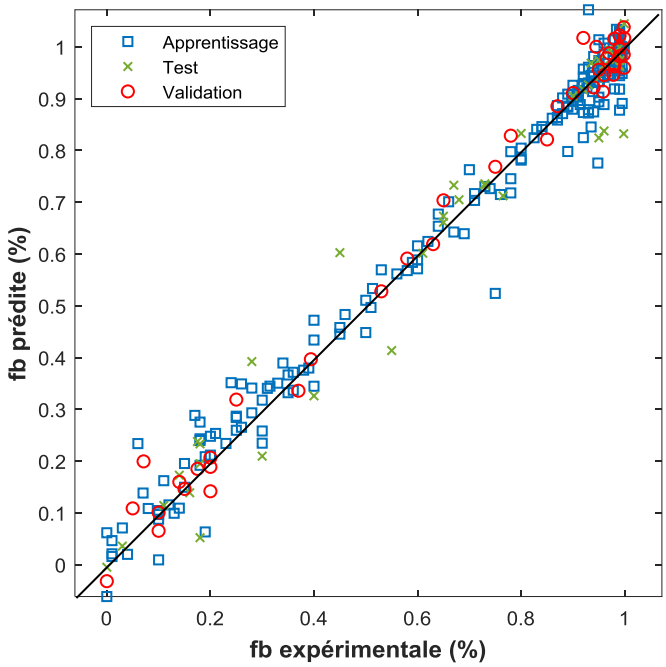


Figure III.8. Représentation graphique de la valeur de la liaison des médicaments aux protéines plasmatiques expérimentales et celles prédites par le modèle 'Filter'.

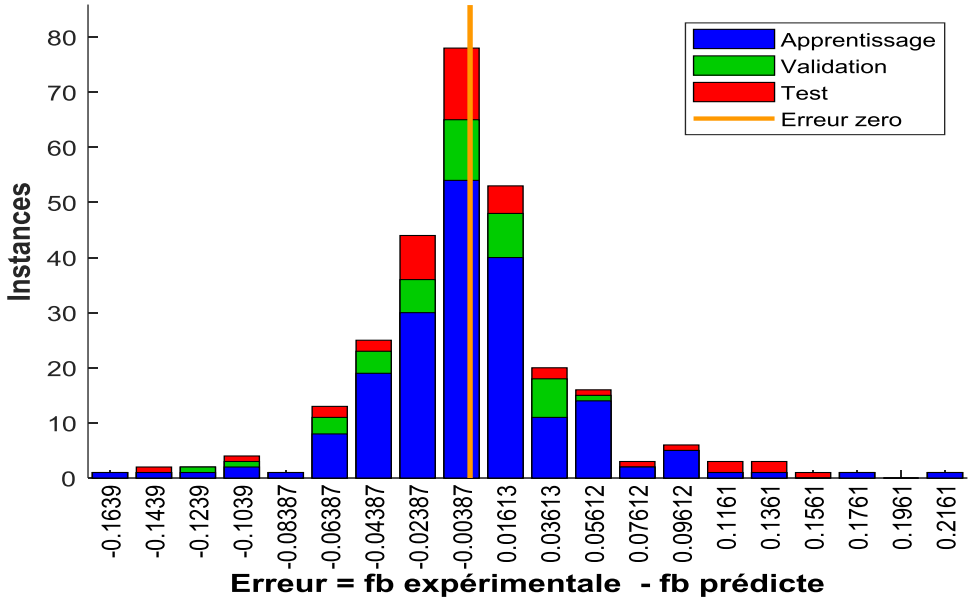


Figure III.9. Histogramme d'erreur pour la méthode 'Filter'.

D'après la figure III.8, nous avons observées une étroite corrélation entre les valeurs de liaison aux protéines plasmatiques prédites par le réseau de neurones artificiels et les valeurs observées. Selon les recommandations de Tropsha et Golbrzaikh, si la différence entre $\boldsymbol{R^2_{app}}$ et $\boldsymbol{Q^2_{cv}}$ ne dépasse pas 0.3, il n'y a pas de sur apprentissage dans le modèle [27]. Dans ce cas, cette différence est égale à 0.008. En conséquent, il y a absence de sur-apprentissage pour le modèle 'Filter '.

L'histogramme de distribution des résidus montre une distribution normale ayant une moyenne se situant à -0.00387. Cette moyenne négative indique que les valeurs calculées sont généralement sous-estimées. Nous avons un bin correspondant à l'erreur de -0.00387 et la hauteur de ce bin pour l'ensemble des données d'apprentissage et de 54, cela signifie que 54 échantillons de notre ensemble d'apprentissage, soit environ 28%, ont une erreur dans cette plage. Pour la validation et le test, les valeurs d'environ 26% et 30% respectivement. Les résidus se répartissent entre -0.163 et 0.2161. Il faut noter que 95% des résidus doivent se trouver dans l'intervalle (-2, +2) et toutes valeurs en dehors de cet intervalle est révélatrice d'une donnée suspecte. Dans notre cas pour l'ensemble des données d'apprentissage nous avons 0.5% des échantillons se trouvant hors intervalle et c'est une valeur inferieur à la valeur de 5% demandée. Pour l'ensemble de validation et de test, nous n'avons aucune valeur hors intervalle. Ces résultats démontrent l'excellente qualité du modèle et son potentiel prédictif performant.

Le tableau III.5 montre 10 modèles de réseau qui ont été développés pour l'approche 'Wrapper'. Les résultats obtenus montrent que le réseau 9 ayant 45 neurones est le meilleur modèle avec *R* (all)=0.935**,** R^2_{app} =0.875, Q^2_{cv} =0.871 et *MSE* (validation)=0.015. La meilleure performance du modèle avait une topologie de (15-45-1) : 15 nœuds d'entrée, 1 couche cachée avec 45 nœuds ayant la tangente hyperbolique comme fonction de transfert, 1 couche de sortie avec une fonction d'identité. Les ensembles d'entraînement, de validation et de test utilisant le modèle RNA ont été tracés par rapport aux valeurs expérimentales de liaison aux protéines dans la figure III.10. Les résultats montrent une étroite corrélation entre la liaison aux protéines plasmatiques prédite et observée. Les réseaux de neurones ont été implémentés à l'aide de Neural Network Toolbox pour MATLAB. Le type de réseau était un réseau Feed-Forward avec la fonction d'entraînement de rétropropagation de Levenberg-Marquardt et une descente de gradient avec une fonction d'apprentissage du poids et du biais.

Tableau III.5. Critères de sélection de l'architecture du FFNN pour la méthode 'Wrapper'**.**

Nombre de neurones dans la couche cachée	R(all)	R^2_{app}	Q^2_{cv}	MSE(validation)
5	0.754	0.539	0.632	0.048
10	0.830	0.679	0.728	0.033
15	0.838	0.694	0.756	0.025
20	0.840	0.724	0.651	0.038
25	0.849	0.712	0.712	0.035
30	0.894	0.810	0.780	0.025
35	0.855	0.707	0.764	0.029
40	0.907	0.814	0.871	0.015
45	**0.935**	**0.875**	**0.871**	**0.015**
50	0.903	0.806	0.821	0.023

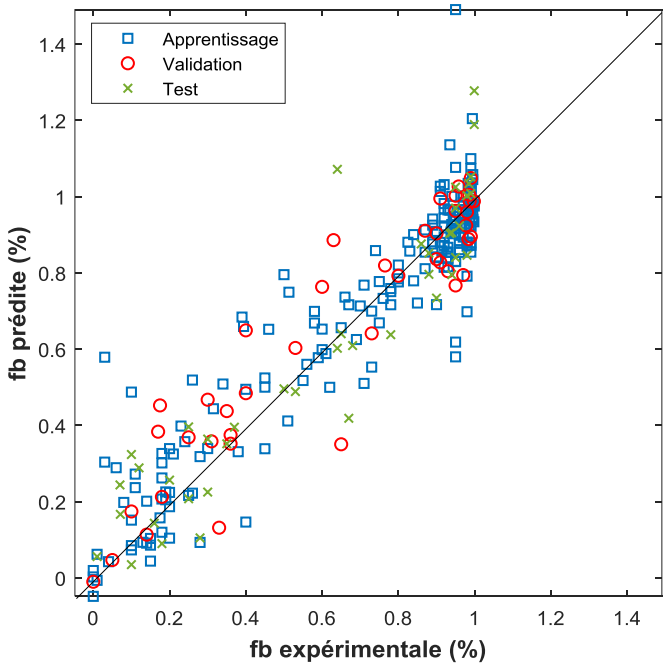


Figure III.10. Représentation graphique de la valeur de la liaison des médicaments aux protéines plasmatiques expérimentales et celles prédites par le modèle 'Wrapper'.

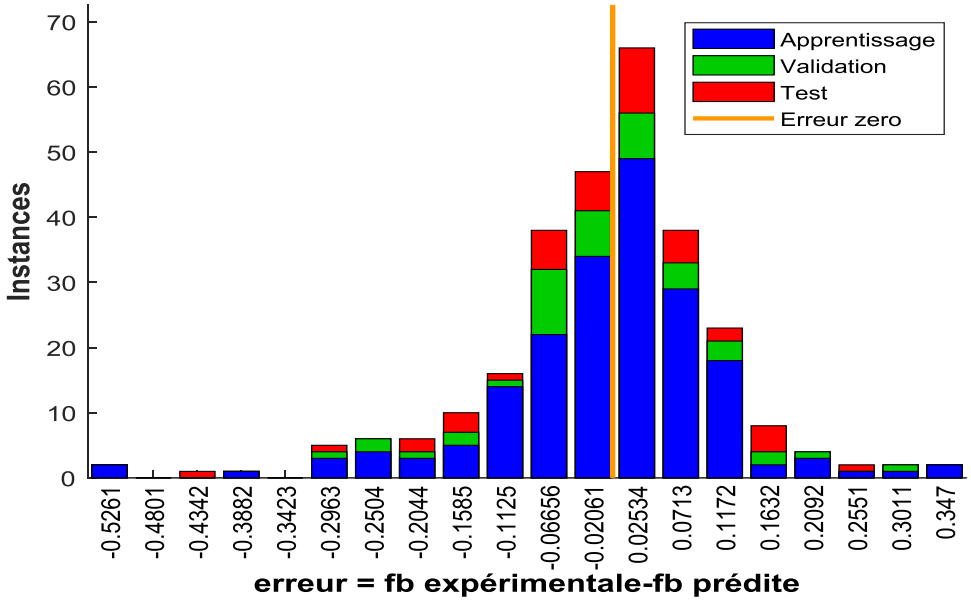


Figure III.11. Histogramme d'erreur pour la méthode 'Wrapper'.

D'aprés la figure III.10, nous avons obtenue une etroite correlation entre les valeurs prédite de la liaison des medicaments aux proteines plasmatique et les valeurs experimentale, ce modèle a egalement donné de bons resultats avec $\boldsymbol{R^2_{app}}$=0.875 et une bonne robustesse avec $\boldsymbol{Q^2_{cv}}$=0.871. La différence entre $\boldsymbol{R^2_{app}}$ et $\boldsymbol{Q^2_{cv}}$ est égale à 0.004. En conséquent, il y a absence de sur-apprentissage pour le modèle 'Wrapper '.

L'histogramme de distribution des residus (figure III.11**)** nous indique une distribution reguliére ayant une moyenne de 0.02534. la courbe présente un décallage positif sur le graphique. Cela signifie que le modèle predit globalement des valeurs plus elevées que les valeurs observées. Les veleurs des residus se repartissent entre -0.521 et 0.347. pour l'enssemble des données d'apprentissage, nous avons 8.8% des echantillons hors intervalle .pour les donnes de validation ,on a 14% et pour les donnes du test on a 7.14%. ces valeurs sonts superieur a la limite de 5% recherches donc ces resultats demontrents que le le modele n'est pas un bon modele predictif.

III.3.3.2. Méthode CNN

Pour spécifier le nombre de neurones cachés requis, nous avons adopté la même procédure pour les méthodes 'Filter' et 'Wrapper'. Les résultats sont rapportés dans les tableaux III.6 et III.7. Pour le modèle CNN-FC15-FFNN, les résultats obtenus montrent que le réseau 7 est le meilleur

modèle avec R (all)=0.960**,** R^2_{app} =0.905, Q^2_{cv} =0.964 et *MSE* (validation)=0.005. La meilleure performance du modèle avait une topologie de (15-35-1) : 15 nœuds d'entrée, 1 couche cachée avec 35 nœuds ayant la tangente hyperbolique comme fonction de transfert et une fonction de transfert linéaire dans la couche de sortie. La formation, la validation et les tests utilisant le modèle RNA ont été tracés par rapport aux valeurs expérimentales de liaison aux protéines dans la figure III.12.

Tableau III.6. Critères de sélection de l'architecture du FFNN pour le modèle CNN-FC15-FFNN.

Nombre de neurones dans la couche cachée	***R*(all)**	R^2_{app}	Q^2_{cv}	***MSE*(validation)**
5	0.882	0.812	0.701	0.040
10	0.908	0.832	0.798	0.009
15	0.913	0.856	0.760	0.028
20	0.933	0.857	0.948	0.009
25	0.910	0.834	0.830	0.022
30	0.905	0.830	0.775	0.022
35	**0.960**	**0.905**	**0.964**	**0.005**
40	0.922	0.832	0.961	0.007
45	0.940	0.874	0.899	0.012
50	0.922	0.830	0.841	0.009

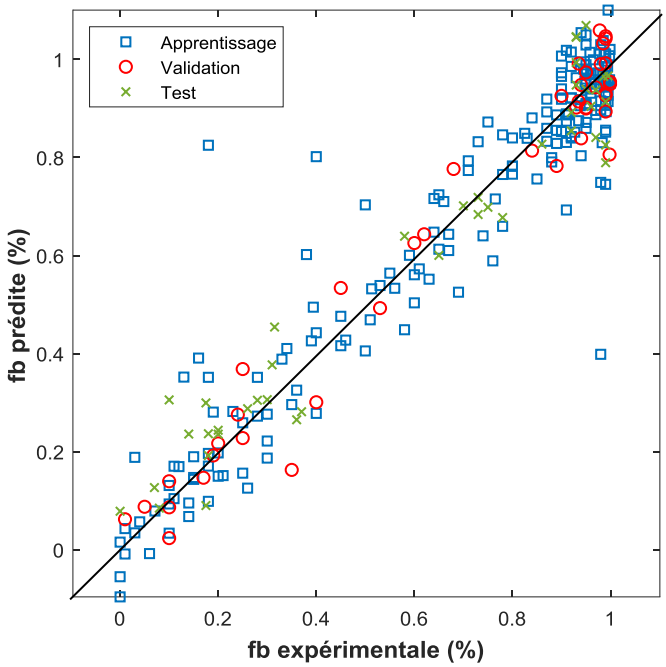


Figure III.12. Représentation graphique de la valeur de la liaison des médicaments aux protéines plasmatiques expérimentales et celles prédites par le modèle 'CNN-FC15-FFNN'.

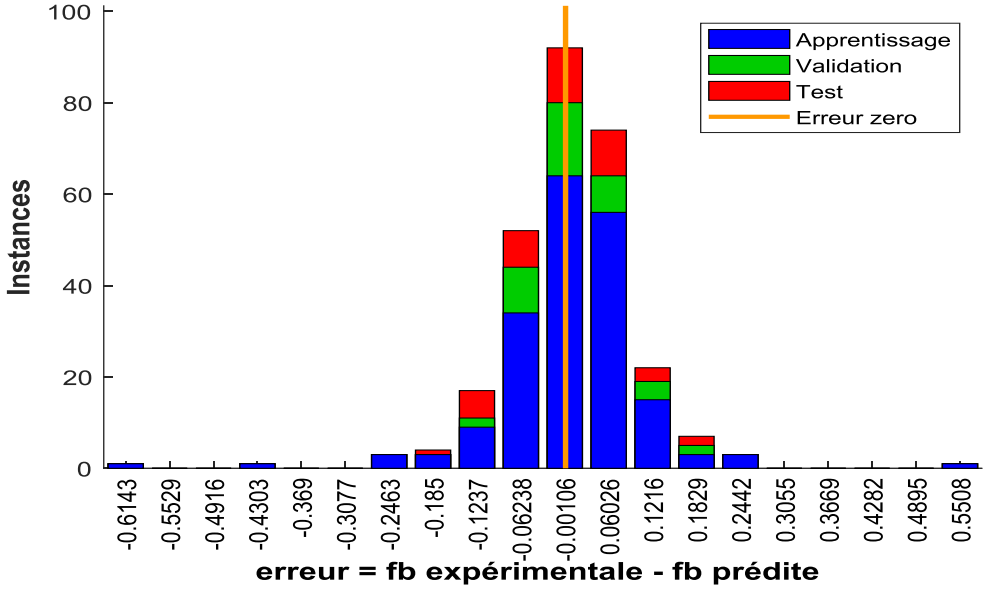


Figure III.13. Histogramme d'erreur pour le modèle 'CNN-FC15-FFNN'.

D'apres la figure III.12, nous avons obtenue une etroite correlation entre les valeurs prédite de la liaison des medicaments aux proteines plasmatique et les valeurs experimentale, ce modèle a egalement donné de bons resultats avec $\boldsymbol{R^2_{app}}$=0.905 et une bonne robustesse avec $\boldsymbol{Q^2_{cv}}$=0.964. La différence entre $\boldsymbol{R^2_{app}}$ et $\boldsymbol{Q^2_{cv}}$ égale à 0.059. En conséquent, il y a absence de sur-apprentissage pour le modèle 'CNN-FC15-FFNN'.

L'histogramme de distribution des residus (figure III.13) présente une moyenne de -0.00106. les veleurs des residus se repartissent entre -0.6143 et 0.5508. pour l'enssemble des données d'apprentissage, nous avons 4.6% des echantillons hors intervalle .pour les donnes de validation et les donnes du test on a aucune valeurs hors intervalle donc ces resulatts sonts inferieurs a la limite recherchée de 5% et ils demontrents que le le modele est un bon modele predictif.

Pour l'approches CNN-FC55-FFNN, les résultats obtenus montrent que le réseau 6 est le meilleur modèle avec R (all)=0.957, R^2_{app} =0.903, Q^2_{cv} =0.961 et *MSE* (validation)=0.005. La meilleure performance du modèle avait une topologie de (55-30-1) : 55 nœuds d'entrée, 1 couche cachée avec 30 nœuds ayant la tangente hyperbolique comme fonction de transfert, 1 couche de sortie avec une fonction de transfert linéaire. Les réseaux de neurones ont été implémentés à l'aide de Neural Network Toolbox dans MATLAB. Le type de réseau était un réseau feed-forward avec la

fonction d'entraînement de rétropropagation de Levenberg-Marquardt et une descente de gradient avec une fonction d'apprentissage du poids et du biais. La formation, la validation et les tests utilisant le modèle RNA ont été tracés par rapport aux valeurs expérimentales de liaison aux protéines dans la figure III.14.

Tableau III.7. Critères de sélection de l'architecture du FFNN pour le modèle CNN- FC 55 –FFNN

Nombre de neurones dans la couche cachée	R**(all)**	R^2_{app}	Q^2_{cv}	MSE**(validation)**
5	0.869	0.769	0.774	0.034
10	0.900	0.805	0.882	0.014
15	0.933	0.858	0.914	0.011
20	0.894	0.796	0.843	0.022
25	0.928	0.854	0.840	0.019
30	**0.957**	**0.903**	**0.961**	**0.005**
35	0.923	0.854	0.840	0.019
40	0.955	0.939	0.832	0.019
45	0.930	0.880	0.854	0.018
50	0.912	0.828	0.840	0.022

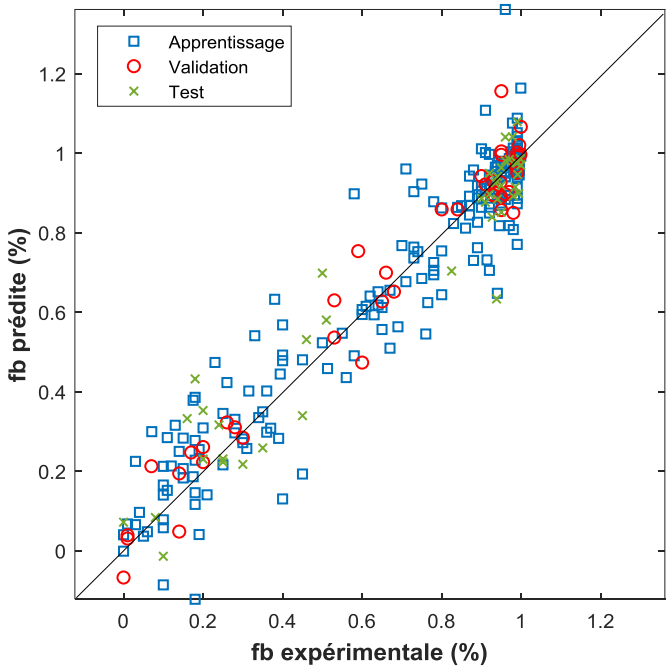


Figure III.14. Représentation graphique de la valeur de la liaison des médicaments aux protéines plasmatiques expérimentales et celles prédites par le modèle 'CNN-FC55-FFNN'.

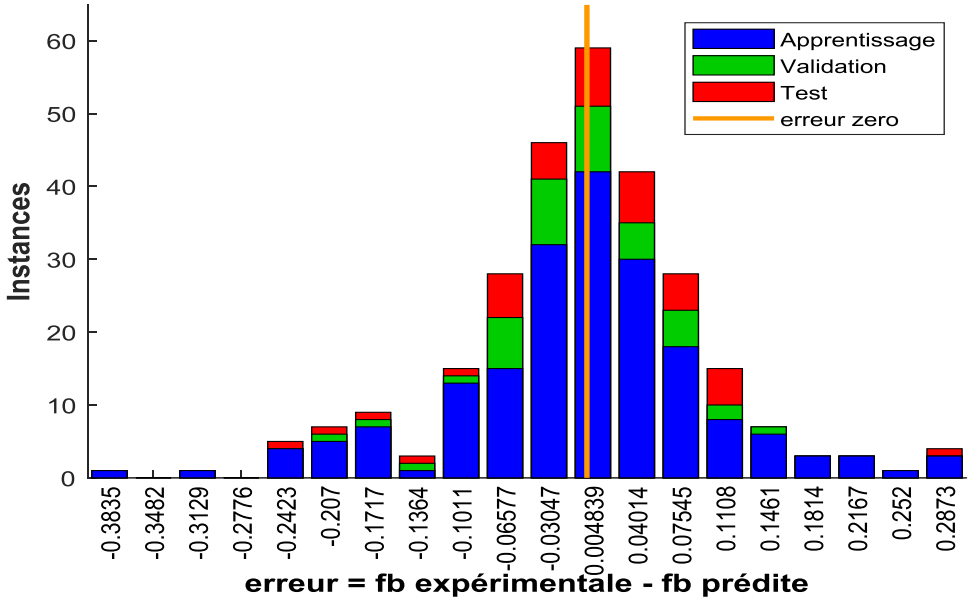


Figure III.15. Histogramme d'erreur pour Le modèle 'CNN-FC55-FFNN'.

La figure III.14, nous renseigne sur la bonne correlation entre les valeurs prédite de la liaison des medicaments aux proteines plasmatique et les valeurs experimentale, ce modèle a egalement donné de bons resultats avec $\boldsymbol{R^2_{app}}$=0.903 et une bonne robustesse avec $\boldsymbol{Q^2_{cv}}$ =0.961. La différence entre $\boldsymbol{R^2_{app}}$ et $\boldsymbol{Q^2_{cv}}$ égale à 0.058. En conséquent, il y a absence de sur-apprentissage pour le modèle 'CNN-FC55-FFNN'.

L'histogramme de distribution des residus (figure III.15) présente une moyenne de 0.004839. les veleurs des residus se repartissent entre -0.3835 et 0.2873. pour l'enssemble des données d'apprentissage, nous avons 9.3% des echantillons hors intervalle(-2,2) . pour les donnes de validation et les donnes du test on a 2.4 %et 7.1 % respectivement donc certeins resulatts sonts superieurs a la limite recherchée de 5% ey ils demontrents que le le modele contient des données suspecte .

Pour les modèles CNN-FC15-FFNN et CNN-FC55-FFNN, nous avons vérifié la corrélation entre chaque paire de nos variables d'entrée au FFNN. Pour obtenir un aperçu de la structure de corrélation, nous avons utilisé une carte thermique (heat map) pour mettre en évidence ce qui est important (figures III.16 et III.17). Pour la méthode 'CNN-FC15-FFNN'. La valeur du coefficient de corrélation entre chaque paire de descripteurs est inférieur à 0,77 ce qui signifie que les descripteurs sélectionnés sont indépendants (figure III.16). En outre, toutes les variables ont des

valeurs du facteur d'inflation de la variance (*VIF*) dans la gamme 1-5 (voir le tableau III.9), ce qui indique que le modèle obtenu à une signification statistique évidente [24]. Pour le modèle 'CNN-FC55-FFNN', La valeur du coefficient de corrélation entre chaque paire de descripteurs est inférieur à 0.8 (figure III.17) et toutes les variables ont des valeurs du facteur d'inflation de la variance (*VIF*) dans la gamme 1-23 (voir le tableau III.8). Il semble que lorsque le vecteur du FullyconnectedLayer de sortie du CNN a une petite dimension, il donne de meilleures valeurs non corrélées entre eux. La question maintenant, c'est qu'elle est le taux de corrélation maximum toléré entre nos entrées de réseau neuronal ? Il semble que la règle générale est que si le coefficient de corrélation simple entre deux régresseurs est supérieur à 0,8 ou 0,9, la multicolinéarité est un problème sérieux. Cependant, avec toutes les règles empiriques que nous obtenons dans les statistiques, il n'y a rien de noir ou blanc à ce sujet [28]. En fait, dans notre cas, il semble qu'une prédiction avec un coefficient de corrélation de 0,8 ne soit pas fatale à notre modèle de régression. Nous pouvons dire que les résultats ci-dessus ont servi notre théorie où plus de descripteurs sont précieux et bienvenus, car nous acceptons qu'en supprimant certaines des descripteurs, nous éliminons également certaines des données que nous avons sur le problème.

Nos résultats décrivent pour la première fois l'utilisation de CNN comme méthode d'extraction des descripteurs combinés à un FFNN pour un ensemble de descripteurs sous forme de matrice au lieu de l'extraction des caractéristiques essentielles d'une image. Les dernières caractéristiques apprises dans les couches du CNN ont été transformées en vecteur et transmises à un FFNN. L'une des énormes circonstances favorables de cette méthode est qu'elle ne prend pas beaucoup de temps par rapport aux méthodes 'Filter' et 'Wrapper' en seulement dix minutes utilisant un CPU, nous obtenons automatiquement quelles caractéristiques sont la meilleure représentation de notre matrice d'entrée, car dans ce travail, les entrées du CNN étaient simplement construit en redimensionnant les 1666 descripteurs en matrices de dimensions 49 × 34. Nous savons qu'il est recommandé d'utiliser les CNN sur des données qui ont des caractéristiques spatiales, et le fait que cela nous a donné de bonnes caractéristiques à utiliser est que le CNN utilisé a découvert que nos données ont une sorte de caractéristiques spatiales. Nous voulions savoir ce que ce réseau de neurones convolutif profond a vraiment vu, et comment il a compris les entrées dont nous l'avons alimentée avec. En visualisant les valeurs numériques de nos matrices d'entrées au CNN, nous avons obtenu des images grayscale donnant 255 différentes nuances de gris possibles du noir au blanc. La figure III.18 montre un échantillon d'image pour l'acide 3-carboxy-4-méthyl-5-propyl-2-furanpropionique obtenu. Les valeurs numériques de zéro et moins vont à zéro et celles de un et plus haut passant à 255. Les valeurs entre zéro et un sont des nuances de gris avec des valeurs comprises entre 0 et 255. En prenant les valeurs de nos descripteurs et en les mappant soudainement

à la couleur correspondante, nous pouvons voir la structure qui se trouve dans nos données. Nous pouvons dire que nous avons fait une interface entre nos données et notre cerveau en visualisant les descripteurs dans lesquels les informations physiques et chimiques sur les molécules deviennent accessibles sous une nouvelle forme d'empreintes digitales. Ils deviennent des storytellers des secrets des molécules en juste une seule image.



Figure III.16. Heatmap qui représente les coefficients de corrélation entre les entrée du réseau de neurone pour le modèle ‘CNN-FC15-FFNN’.



Figure III.17. Heatmap qui représente les coefficients de corrélation entre les entrée du réseau de neurone pour le modèle ‘CNN-FC55-FFNN’.

Tableau III.8. *VIF* des entrée sélectionnés pour le modèle‘CNN-FC55-FFNN’.

Entrée N	1	2	3	4	5	6	7	8	9	10	11	12	23	14	15
VIF	16.8	4.6	15.8	9	4	5	11.1	12.6	12.3	11.6	8.4	12.7	6.8	5	6.4

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
14.9	6.4	17.3	2.9	10	9.7	6.4	13.1	5.7	13.5	6.1	7	4.7	6.2	8.3	12.6	7.5

33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
8.6	13	4.2	5.4	4.1	6.7	3.7	9	23	10	17	18	16	9.6	10	6.2	3.9	8.1	12

52	53	54	55
5.6	7	10.9	3.2

Tableau III.9. *VIF* des entrée sélectionnés pour le modèle‘CNN-FC15-FFNN’.

Entrée N	1	2	3	4	5	6	7	8	9	10	11	12	23	14	15
VIF	3.3	4.9	2.5	2.6	3.4	2	2.4	2	1.3	1.9	3.2	2	2.6	1.8	1.5

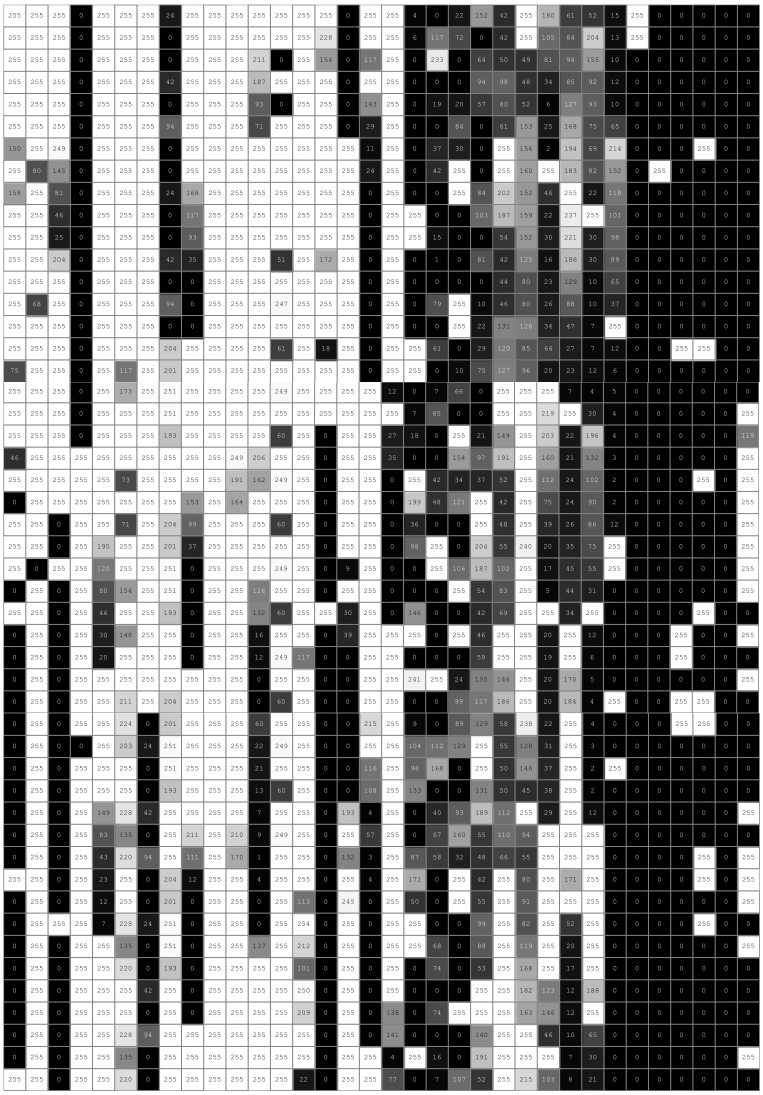


Figure III.18. Visualisation d'un échantillon d'image pour l'acide 3-carboxy-4-méthyl-5-propyl-2-furanpropionique obtenu.

III.4. Validation et comparaison entre les quatre modèles

Une évaluation statistique reportée dans le tableau III.10 est réalisé pour comparer les performances et la qualité de prédiction des quatre modèles développés dans ce travail. Les coefficients statistiques de la validation interne pour les 4 modèles sont tous acceptables et satisfaisants. En conséquence, ces modèles sont robustes. La qualité des modèles a également été jugée en termes de paramètres de validation externe. Là aussi les 4 modèles se distinguent par un excellent pouvoir de prédiction.

Tableau III.10. Validation interne et externe pour les quatre modèles.

	Méthode 'Filter'	**Méthode 'Wrapper'**	**CNN-FC55-FFNN**	**CNN-FC15-FFNN**	**Critère de validité**
Validation interne					
R	0.991	0.935	0.950	0.951	
R^2	0.981	0.875	0.903	0.905	>0.6
Q^2_{cv}	0.989	0.871	0.961	0.964	>0.5
MSE	0.002	0.015	0.005	0.005	
MAE	0.028	0.090	0.054	0.054	
RMSE	0.039	0.122	0.070	0.069	
Validation externe					
R	0.983	0.933	0.961	0.972	
Q^2	0.966	0.870	0.923	0.945	>0.6
MSE	0.004	0.016	0.009	0.007	
MAE	0.042	0.094	0.071	0.068	
RMSE	0.063	0.129	0.093	0.085	
$r^2 - r_0^2$	-0.022	-0.034	-0.02	0.021	<0.3
k	0.99	0.9	0.97	0.88	0.85< k<1.15
k'	1	1	0.97	1	0.85< k'<1.15
$(r^2 - r_0^2)/ r^2$	-0.023	-0.04	-0.022	-0.033	<0.1
$(r^2 - r_0'^2)/ r^2$	-0.023	-0.04	-0.022	-0.033	<0.1

III.5. Domaine d'applicabilité des modèles

Le troisième principe des lignes directrices de l'OCDE exige un domaine d'applicabilité défini. Dans ce travail, le domaine d'applicabilité a été élaboré en utilisant différentes méthodes par le biais de MATLAB : 'bounding box', 'leverage approach','euclidean distance (95percentile)','classical KNN '.

Tableau III.11. Domaine d'applicabilité du modèle CNN –FC15-FFNN.

Approche utilisée	**Nombre de composée à l'intérieur du DA**	**Nombre de composes à l'extérieur du DA**
Bounding box	42	0
leverage	41	1
Euclidan distance (95 percentile)	42	0
Classical KNN (euclidean distance,k=5)	40	2
KNN (euclidean distance (k=25)	42	0

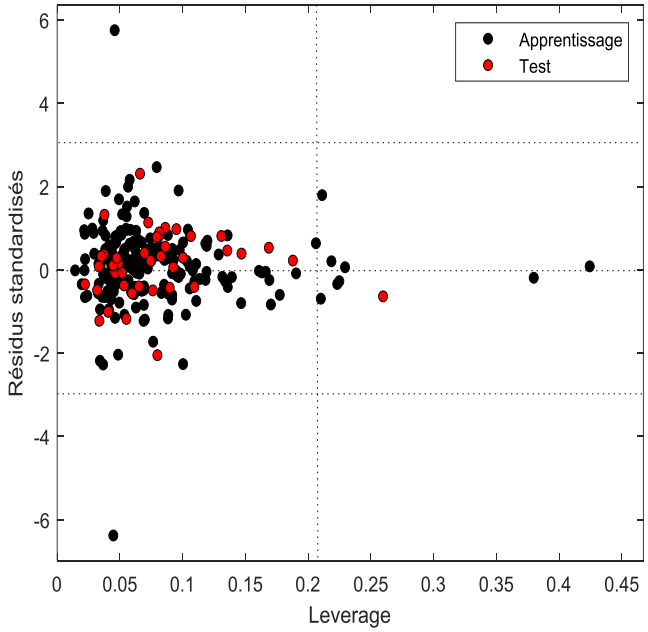
Pour les modèles KNN (euclidean distance (k=25), Euclidan distance (95 percentile) et la méthode du Bounding box aucun des composée de l'ensemble test n'est en dehors du domaine d'applicabilité tandis que le modèle Classical KNN (euclidean distance, k=5) montrent que deux composés sont à l'extérieur du domaine d'applicabilité. Par contre, pour la méthode du leverage un seul composé est à l'extérieur du domaine d'applicabilité. A titre d'exemple, nous reproduisons le diagramme de wiliams (figure III.19) obtenue par l'approche du leverage pour les ensembles d'apprentissage et de test du modèle CNN-FC15-FFNN. Cette approche est basée sur la variation des résidus r_i de prédiction standardisée (équation III.12) en fonction des valeurs des leviers h_i pour chacun des composée .la valeur du levier critique h^* est donne par l'équation III.13.

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

Généralement -3<= r_i<=3

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

Ou *n, k* et *s* représentent respectivement le nombre d'observation de l'ensemble d'apprentissage, le nombre de descripteurs et l'écart type.



$$h^* = 0.248$$

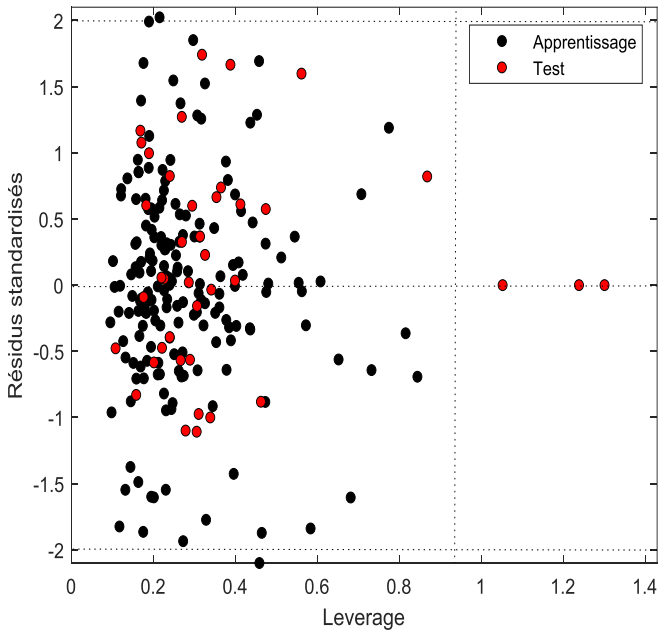
Figure III.19. Diagramme de Williams pour les ensembles d'apprentissage et de test du modèle CNN –FC15-FFNN.

On remarque que un seul composé appartenant à l'ensemble de test est en dehors du domaine d'applicabilité (aire limite par les deux lignes horizontales et ligne vertical). Un point de levier est un type de valeur aberrante. Cependant les points de leviers ne sont pas forcément des valeurs aberrantes de régression. Lorsque les points de levier suivent la tendance linéaire générale de la majorité des observations, on parle de bons points de levier. Dans notre cas, bien que notre point soit éloigné du reste des observations, il est proche de la ligne d'ajustement de régression car il a un petit résidu. Ce résultat montre que le modèle CNN-FC15-FFNN développé est conforme au troisième principe de l'OCDE, il peut être ainsi utilisé pour prédire la liaison des médicaments aux protéines plasmatique pour de nouveaux médicaments.

Tableau III.12. Domaine d'applicabilité du modèle CNN–FC55-FFNN.

Approche utilisée	Nombre de composés à l'intérieur du DA	Nombre de composés à l'extérieur du DA
Bounding box	34	8
leverage	39	3
Euclidan distance (95 percentile)	41	1
Classical KNN (euclidean distance,k=5)	41	1
KNN(euclidean distance(k=25)	42	0

Pour les modèles, Euclidan distance (95 percentile) et la méthode Classical KNN (euclidean distance,k=5) un seule composée de l'ensemble test est en dehors du domaine d'applicabilité tandis que le modèle KNN(euclidean distance(k=25) montre que aucun composés n'est à l'extérieur du domaine d'applicabilité. Par contre, avec le méthode du leverage, 3 composé sont à l'extérieur du domaine d'applicabilité. Pour le modèle Bounding box huit composée sont en dehors du domaine d'applicabilité. La figure III.20 montre le diagramme de wiliams obtenue par l'approche du leverage pour les ensembles d'apprentissage et de test du modèle CNN-FC55-FFNN.



$h^* = 0.87$

Figure III.20. Diagramme de Williams pour les ensembles d'apprentissage et de test du modèle CNN –FC55-FFNN.

On Remarque que trois composé appartenant à l'ensemble de test sont en dehors du domaine d'applicabilité (aire limite par les deux lignes horizontales et ligne vertical). Bien que nos points soient éloignés du reste des observations, ils sont très proches de la ligne d'ajustement de régression car ils ont un très faible résidu. Ils sont donc de bons points de levier. Ce résultat montre que le modèle CNN-FC55-FFNN développé est conforme au troisième principe de l'OCDE, il peut être ainsi utilise pour prédire la liaison des médicaments aux protéines plasmatique pour de nouveaux médicaments.

III.6. Comparaison avec d'autres modèles parus dans la littérature

Nous avons comparés nos résultats à un nombre limité de modèles QSAR développées pour la prédiction du taux de fixation aux protéines plasmatiques disponibles dans la littérature (tableau III.13). Yawen Yuan avec ses collaborateurs ont construits un modèle QSAR pour la prédiction du taux de fixation aux protéines plasmatiques sur la base d'un vaste ensemble d'apprentissage comprenant plus de 5000 composés et ils ont proposés une nouvelle stratégie pour construire des modèles pour différents niveaux de liaisons [29].

Lixia Sun et ses co-auteurs ont développés des modèles QSAR pour six algorithmes d'apprentissage pour la prédiction du taux de fixation aux protéines plasmatiques contenant 967 produits pharmaceutiques. Le modèle utilisant le RNA a abouti à une *MAE* pour l'ensemble de test de 0.126 [30].

Rajnish Kumar et ses camarades ont présentés six algorithmes d'apprentissages pour des modèles QSAR utilisant une base de données contenant 735 médicaments. Le modèle avec le RNA a donné une *MSE* train= 0.869 et une *MSE* test=0.888[31].

Haiyan Li et ses collègues ont développés une nouvelles méthode appelée analyse QSAR de l'interaction des protéines plasmatiques (PPI-QSAR) avec une base de données de 65 antibiotiques qui a donné R^2 train=0.86 Q^2 train=0.72 [32].

Taravat Ghafourian avec son groupe de chercheurs ont collectés une base de donnes de 794 médicaments et ont utilisés 04 algorithmes d'apprentissage le meilleur modèles était Boosted Regression Tree qui a donné une *MAE* train=13.25 et une *MAE* val=14.96 [33].

Moda Tiago et ses collaborateurs ont développé une relation structure activité par hologramme (HQSAR) sur une série de 312 médicaments. Ils ont obtenu un Q^2=0.72 et R^2 =0.91[34].

Nous nous ne sommes pas intéressés à l'évaluation des avantages et inconvénients de ces méthodes car l'objectif de notre travail était de prouver que notre nouvelle méthode était aussi efficace que ces anciennes méthodes. De plus, selon nos résultats, nous avons prouvé que le modèle établi par nos soins peut être utilisé de manière prometteuse pour prédire la concentration des médicaments liés aux protéines plasmatiques tout en réalisant des économies de temps et d'argent.

Tableau III.13. Comparaison avec la littérature.

Méthode	***MAE***	***R^2***	***R***	***MSE***
Méthode suggérée (Filter)	Train 0.0313 Validation 0.0284 Test 0.0423	Train 0.981 Validation 0.989 Test 0.966	Train0.991 Validation 0.995 Test 0.983	Train 0.031 Validation 0.028 Test 0.042
Méthode suggérée (Wrapper)	Train 0.080 Validation 0.091 Test 0.094	Train 0.875 Validation 0.871 Test 0.870	Train 0.935 Validation 0.933 Test 0.933	Train 0.014 Validation 0.015 Test 0.017
Méthode suggéréé (CNN-FC55-FFNN)	Train 0.074 Validation 0.054 Test 0.072	Train 0.903 Validation 0.961 Test 0.923	Train 0.950 Validation 0.980 Test 0.961	Train 0.011 Validation 0.005 Test 0.009
Méthode suggéréé (CNN-FC15-FFNN)	Train 0.066 Validation 0.054 Test 0.068	Train 0.905 Validation 0.964 Test 0.945	Train 0.951 Validation 0.982 Test 0.972	Train 0.011 Validation 0.005 Test 0.007
Yawen Yuan 2020[29]	Test 0.076			
Lixia Sun 2018[30]	Test 0.126			
Rajnish Kumar 2018[31]				Train 0.869 Test 0.888
Haiyan Li 2011[32]		Train 0.86		
Taravat Ghafourian 2013[33]	Train 13.25 Validation 14.96	Train 0.717 Validation 0.646	Train 0.681 Validation 0.641	
Moda Tiago L 2007[34]		Test 0.91		

III.7. Conclusion

Nous avons expérimenté trois approches différentes pour la sélection de caractéristiques dans l'apprentissage automatique. Un ensemble de données a été sélectionné ; les trois stratégies de sélection de caractéristiques ont été suivies pour produire des entrées, qui ont été utilisées dans un réseau neuronal pour faire des prédictions. Les paramètres utilisés pour la validation ont indiqué que la précision de la nouvelle approche de sélection des caractéristiques extraites à partir d'un CNN est similaire à la précision avec des caractéristiques obtenues par certains prédicateurs actuels. Il est évident que cette enquête a montré que l'extraction de caractéristiques à l'aide de CNN prend moins de temps en plus de garder toutes les informations contenues dans les caractéristiques d'origine. Contrairement aux techniques de sélection de caractéristiques, nous n'avons pas à comprendre quelles caractéristiques sont les meilleures représentations pour notre modèle, elles sont apprises pour nous. Au final, on peut dire que cette recherche nous a permis de développer une nouvelle méthodologie pour prédire les propriétés des composés à l'aide de descripteurs.

Références

[1] Bohnert T, Gan L-S. Plasma protein binding: from discovery to development. Journal of pharmaceutical sciences 2013;102:2953-94.

[2] Loubna A. CONCEPTION ET EVALUATION DES MOLECULES ANTICANCEREUSES PAR APPROCHES IN SILICO: Inhibiteurs des Kinases AKT1 et LMTK3. 2020.

[3] Alberts DS, Chang SY, Chen HS, Evans TL, TE. M. Oral melphalan kinetics. Clinical Pharmacology & Therapeutics 1979;26:737-45.

[4] Anderson PO, Knoben JE, Troutman WG. Handbook of Clinical Drug Data: McGraw-Hill Medical Pub. Division; 2002.

[5] Atiba JO, Taylor G, Pershe RA, Blaschke TF. Plasma antipyrine half-life can be determined from urine data. British Journal of Clinical Pharmacology 1987;23:715-9.

[6] Goodman LS, Brunton LL, Blumenthal DK, Murri N, Hilal-Dandan R. Goodman & Gilman's the pharmacological basis of therapeutics, 12th Edition. New York: McGraw-Hill Medical.; 2011.

[7] Goodman LS, Gilman A, Brunton LL, Lazo JS, Parker KL. Goodman & Gilman's the pharmacological basis of therapeutics, 11th Edition. New York: McGraw-Hill; 2006.

[8] Goodman LS, Gilman A, Hardman JG, Gilman AG, Limbird LE. Goodman & Gilman's the pharmacological basis of therapeutics, 9th Edition. New York: McGraw-Hill, Health Professions Division; 1996.

[9] Turner JV, Maddalena DJ, Cutler DJ. Pharmacokinetic parameter prediction from drug structure using artificial neural networks. International Journal of Pharmaceutics 2004;270:209-19.

[10] Weber C, Gasser R, G. H. Absorption, excretion, and metabolism of the endothelin receptor antagonist bosentan in healthy male subjects. Drug Metabolism & Disposition 1999;27:810-5.

[11] DrugBank database (www.drugbank.ca). September 2020.

[12] Virtual Computational Chemistry Laboratory (www.vcclab.org). September 2020.

[13] Hamadache M, Khaouane L, Benkortbi O, Si Moussa C, Hanini S, Amrane A. Prediction of Acute Herbicide Toxicity in Rats from Quantitative Structure–Activity Relationship Modeling. Environmental Engineering Science 2014;31:243-52.

[14] COMPUTATIONNELLE C. TOUHAMI Moufida: Université de Saida; 2019.

[15] Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The 'K'in K-fold cross validation. 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN): i6doc. com publ; 2012. p. 441-6.

[16] Yadav S, Shukla S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. 2016 IEEE 6th International conference on advanced computing (IACC): IEEE; 2016. p. 78-83.

[17] Berrar D. Cross-Validation. 2019.

[18] Van Wilgenburg S, Sólymos P, Kardynal K, Frey M. Paired sampling standardizes point count data from humans and acoustic recorders. Avian Conservation and Ecology 2017;12.

[19] Denis E, Thiboud P, Estrade V. Reconnaissance endoscopique des calculs par intelligence artificielle: résultats préliminaires. Progrès en Urologie 2021;31:827.

[20] Alexander DL, Tropsha A, Winkler DA. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. Journal of chemical information and modeling 2015;55:1316-22.

[21] Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems 2015;145:22-9.

[22] Hanser T, Barber C, Marchaland J, Werner S. Applicability domain: towards a more formal definition. SAR and QSAR in Environmental Research 2016;27:865-81.

[23] Sushko I, Novotarskyi S, Körner R, Pandey AK, Kovalishyn VV, Prokopenko VV, et al. Applicability domain for in silico models to achieve accuracy of experimental measurements. Journal of Chemometrics 2010;24:202-8.

[24] Agrawal VK, Khadikar PV. QSAR prediction of toxicity of nitrobenzenes. Bioorganic & Medicinal Chemistry 2001;9:3035-40.

[25] Panchal FS, Panchal M. Review on methods of selecting number of hidden nodes in artificial neural network. International Journal of Computer Science and Mobile Computing 2014;3:455-64.

[26] Kubat M. An introduction to machine learning: Springer; 2017.

[27] Golbraikh A, Tropsha A. Beware of q2! Journal of molecular graphics and modelling 2002;20:269-76.

[28] Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. Journal of Interdisciplinary Mathematics 2010;13:253-67.

[29] Yuan Y, Chang S, Zhang Z, Li Z, Li S, Xie P, et al. A novel strategy for prediction of human plasma protein binding using machine learning techniques. Chemometrics and Intelligent Laboratory Systems 2020;199:103962.

[30] Sun L, Yang H, Li J, Wang T, Li W, Liu G, et al. In silico prediction of compounds binding to human plasma proteins by QSAR models. ChemMedChem 2018;13:572-81.

[31] Kumar R, Sharma A, Siddiqui MH, Tiwari RK. Prediction of drug-plasma protein binding using artificial intelligence based algorithms. Combinatorial chemistry & high throughput screening 2018;21:57-64.

[32] Li H, Chen Z, Xu X, Sui X, Guo T, Liu W, et al. Predicting human plasma protein binding of drugs using plasma protein interaction QSAR analysis (PPI-QSAR). Biopharmaceutics & Drug Disposition 2011;32:333-42.

[33] Ghafourian T, Amin Z. QSAR models for the prediction of plasma protein binding. BioImpacts: BI 2013;3:21.

[34] Moda TL, Montanari CA, Andricopulo AD. In silico prediction of human plasma protein binding using hologram QSAR. Letters in Drug Design & Discovery 2007;4:502-9.

Conclusion générale et perspectives



La finalité de la recherche en chimie est la mise sur le marché de nouveaux produits ayant par rapport aux produits existants, au moins un avantage sur le plan de l'efficacité, de la tolérance ou du prix. Comment choisir et arriver à ces nouveaux produits ? En chimie thérapeutique par exemple, malgré des travaux considérables, il ne semble pas exister de procédé permettant de prévoir à coup sûr les actions pharmacologiques d'une substance de structure entièrement nouvelle. Seuls quelques processus généraux d'innovation sont à la base de la découverte de médicaments : le hasard, la chimie combinatoire, le criblage et l'inspiration à partir des structures de substances à activité biologique connue.

Les études des QSAR s'assignent comme objectif de relier l'activité biologique d'une série de composés homologues à leurs structures chimiques ou plutôt à leurs propriétés physico-chimiques. Ils permettent de

- ✓ Connaître les facteurs qui gouvernent l'activité,
- ✓ Prédire l'efficacité des composés synthétisés ou non synthétisés,
- ✓ Orienter la recherche vers de nouveaux composés actifs.

Les QSAR permettent ainsi de réduire le nombre de produits synthétisés, et par conséquent les coûts dépensés dans ce domaine. Elles se servent de certaines méthodes statistiques pour la modélisation des relations entre l'activité et la structure chimique des composés.

L'objectif de cette thèse était de mettre au point un modèle statistique de prédiction du taux de fixation des médicaments aux protéines plasmatiques. Cette étude s'intègre dans le cadre de la

quantification de ce taux sans recourir à l'expérimentation animale. Cette dernière, en plus du fait qu'elle est coûteuse et laborieuse, est décriée par une multitude d'organisations de défense des animaux. Les principaux inconvénients des techniques traditionnelles de modélisation établissent une relation entre la grandeur modélisée et un vecteur de variables qui la déterminent résident dans la difficulté du choix des variables pertinentes, et dans leur calcul ou leur mesure préalable. Nous avons développé une technique qui s'affranchit de ces problèmes en structurant les données sous forme matricielle.

Pour atteindre cet objectif, les modèles QSAR ont été établis en utilisant les réseaux de neurones artificiels à 3 couches (une couche d'entrées, une couche cachée et une couche de sortie) et nous avons effectué trois techniques de sélection des descripteurs les plus pertinents. Dans ce contexte, nous avons suivi les démarches suivantes :

Tout d'abord, nous avons élaboré une base de données avec comme paramètre à modéliser le taux de fixation aux protéines plasmatiques. Au total, on a recueilli le taux de fixation aux protéines plasmatiques de 277 médicaments. Le choix des variables nécessaires à la description de f_b a été fait sur la base du concept qui stipule que le phénomène à modéliser (f_b) est intimement lié à la structure moléculaire. A cette fin, nous avons calculé 1666 descripteurs moléculaires pour chaque médicament.

Par la suite, nous avons élaboré un modèle de prédiction du taux de fixation aux protéines pour 277 médicaments. Les descripteurs ont été sélectionnés par trois méthodes puis utilisés pour développer quatre modèles QSAR prédictif, par utilisation des réseaux de neurones (*PMC-RNA*). Une fois que le choix des conditions de travail (répartition de la base de données en ensemble d'apprentissage et de validation, algorithme d'apprentissage, fonctions d'activation de la couche cachée) eut été fait, un modèle neuronal a été élaboré.

Les quatre modèles conçus ont montrées une très bonne corrélation des données expérimentales avec celles prédites. Par contre les modèles 'Wrapper 'et 'CNN-FC55-FFNN' ont montré l'existence de données suspectes. Les coefficients statistiques de la validation interne et externe appliqué à l'ensemble des donnés sont tous acceptables et satisfaisant. En conséquence, ces modèles sont robustes et se distinguent par un bon pouvoir prédictif. Le domaine d'applicabilité pour la nouvelle méthode a monté que les deux modèles sont conforme au troisième principe de l'OCDE.

En conclusion, nous estimons que l'objectif assigné à cette étude a été atteint. En ce sens, on peut confirmer que la nouvelle méthode de sélection est aussi efficace que les méthodes

traditionnelles aves l'avantage d'être moins expansif en temps et en complexité. Car les caractéristiques essentielles sont automatiquement apprises pour nous par le CNN et cela en dix minutes scellements avec un CPU au lieu de 03 mois par les méthodes Filter et Wrapper.

A l'issu de ce travail et en vus des résultats, de nombreuses voies de recherche ont émergé. Ces perspectives concernent plus particulièrement la prédiction d'autres phénomènes biologiques. D'autre part, il faudrait s'intéresser a utilisé la prédiction directe du taux de fixation aux protéines plasmatiques utilisant les CNN et utilisé la représentation de la structure moléculaire au lieu des descripteurs comme entrée dans un CNN et faire la comparaison.

PUBLICATION SCIENTIFIQUE



Affaf Khaouane, Samira Ferhat, Salah Hanini. **« A Novel Methodology for Human Plasma Protein Binding: Prediction, Validation, and Applicability Domain ».** Pharmaceutical and Biomedical Research. 2022; 8(4):311-322.

Original Article

A Novel Methodology for Human Plasma Protein Binding: Prediction, Validation, and Applicability Domain

Affaf Khaouane[1*], Samira Ferhat[1], Salah Hanini[1]

1. Laboratory of Biomaterial and Transport Phenomena, University of Médéa, Médéa, Algeria.

*** Corresponding Author:**
Affaf Khaouane, PhD.
Address: Laboratory of Biomaterial and Transport Phenomena, University of Médéa, Médéa, Algeria.
Phone: +213 (791) 79 32 84
E-mail: affoufa80@gmail.com

Copyright© 2020, The Authors.

Article info:
Received: 11 Apr 2022
Accepted: 01 Jun 2022

Keywords:
Quantitative structure-activity relationship, Convolutional neural network, Feed-forward neural network, Protein binding.

ABSTRACT

Background: Plasma protein binding is a key component in drug therapy as it affects the pharmacokinetics and pharmacodynamics of drugs.

Objectives: This study aimed to predict the fraction of plasma protein binding.

Methods: A quantitative structure-activity relationship, convolutional neural network, and feed-forward neural network (QSAR-CNN-FFNN) methodology was used. CNN was used for feature selection, which is known as a difficult task in QSAR studies. The values of the descriptors acquired without the preprocessing procedures were rearranged into matrices, and features from a deep fully connected layer of a pre-trained CNN (ALEXNET) were extracted. Then, the latest features learned from the CNN layers were flattened out and passed through an FFNN to make predictions.

Results: The external accuracy of the validation set (Q^2=0.945, RMSE=0.085) showed the performance of this methodology. Another extremely favorable circumstance of this method is that it does not take a lot of time (only a few minutes) compared to the QSAR-Wrapper-FFNN method (days of hard work and concentration) and it automatically gives us the characteristics that are the best representations of our input.

Conclusion: We can say that this model can be used to predict the fraction of human plasma protein binding for drugs that have not been tested to avoid chemical synthesis and reduce expansive laboratory tests.

Citation Khaouane A, Ferhat S, Hanini S. A Novel Methodology for Human Plasma Protein Binding: Prediction, Validation, and Applicability Domain. Pharmaceutical and Biomedical Research. 2022; 8(4):311-322

http://dx.doi.org/10

Introduction

Building a quantitative structure-activity relationship (QSAR) model, known as in silico model, is an important step in drug discovery. It allows us to avoid chemical synthesis and reduce expansive and tedious laboratory tests [1, 2]. Different QSAR approaches have been developed over the past few years [3-5]. These approaches can determine the quantitative relations between the variation in the values of calculated descriptors and the biological activity of a series of chemical molecules [6]. Thus, once a correlation has been established, it can be used to predict the property or biological activity of new structures [7, 8]. In QSAR studies, selecting a few relevant descriptors from a large number of features is a difficult task. Feature selection techniques are applied to reduce the number of attributes in the dataset by choosing features that will give us better accuracy with fewer data [9-11]. It also reduces the overfitting and the overtraining risk [12]. Feature selection methods can be divided into three categories: filter, wrapper, and embedded methods. Filters are applied independently of the mapping method used. They are executed before the mapping, to reduce the number of descriptors, following some objective criteria [13]. The wrapper techniques are used to select the optimum subset of features based on the error reduction of classifier algorithms. Wrapper methods perform better than filter methods but are more expensive and time-consuming [14]. Features selected with embedded or hybrid methods are sensitive to the structure of the underlying classifiers. Thus, in most cases, the features selected by one embedded method might not be suitable for others [15].

Plasma protein binding plays a key role in drug therapy that affects the pharmacokinetics and pharmacodynamics of the drug, as they are often directly related to the free drug concentration in plasma [16, 17]. In recent years, several QSAR-artificial intelligence models have been developed to predict plasma protein binding, such as support vector machines and their derivatives [18-20], random forest [21], neural networks [22, 23], and gradient-boosting decision trees [24]. In 2017, Sun et al. constructed QSAR models by six machine-learning algorithms with 26 molecular descriptors [25]. Kumar et al. in 2018 presented a systematic approach using a support vector machine, artificial neural network, K-nearest neighbor, probabilistic neural network, partial least square, and linear discriminant analysis to a diverse dataset of 735 drugs [26]. Yuan et al. in 2020 published a global QSAR model for plasma protein binding and developed a novel strategy to construct a robust QSAR model for plasma protein binding prediction [27]. Recently, deep learning algorithms have attracted the attention of scientists and become an important option for pharmaceutical research. Ramsundar et al. presented deep learning techniques for healthcare that efficiently predict drug activity and structure [28]. Wallach et al. introduced AtomNet, known as the first structure-based deep convolutional neural network (CNN), to predict the bioactivity of small molecules for drug discovery applications [29].

In this work, a novel combined methodology based on QSAR, CNN, and a feed-forward neural network (FFNN) was used to predict plasma protein binding for 277 molecules. The CNN, known to be the most popular algorithm for deep learning, was used for feature selection as an alternative to the wrapper method. The FFNN was then used for the prediction of the plasma protein binding from the extracted features.

Material and Metods

In order to predict the plasma protein binding, a methodology based on five steps was used as summarized in Figure 1: (1) data set collection, (2) molecular descriptors generation, (3) selection of relevant descriptors by a wrapper method and a CNN method, (4) FFNN modeling, and (5) validation of models.

Data set collection

The experimental data of protein binding of the 277 drugs used in this study were selected from the pharmacological basis of the therapeutics handbook [30] and the handbook of clinical drug data [31]. Chemical names and experimental protein binding values are presented in Supplemental Material 1.

Molecular descriptors generation

The numerical representation of molecular structure was assessed in terms of molecular descriptors. The SMILES script (simplified molecular-input line-entry system) required to calculate descriptors was extracted from the open-access database PubChem [32]. SMILES is a specification in the form of a line notation describing the structure of chemical species [33]. The SMILES scripts for the 277 drugs were used to generate 1666 descriptors divided into 20 categories: *(i)* constitutional descriptors, *(ii)* topological descriptors, *(iii)* walk and path counts; *(iv)* connectivity indices; *(v)* index information; *(vi)* 2Dautocorrelations; *(vii)* edge adjacency indices; *(viii)* burden eigenvalue

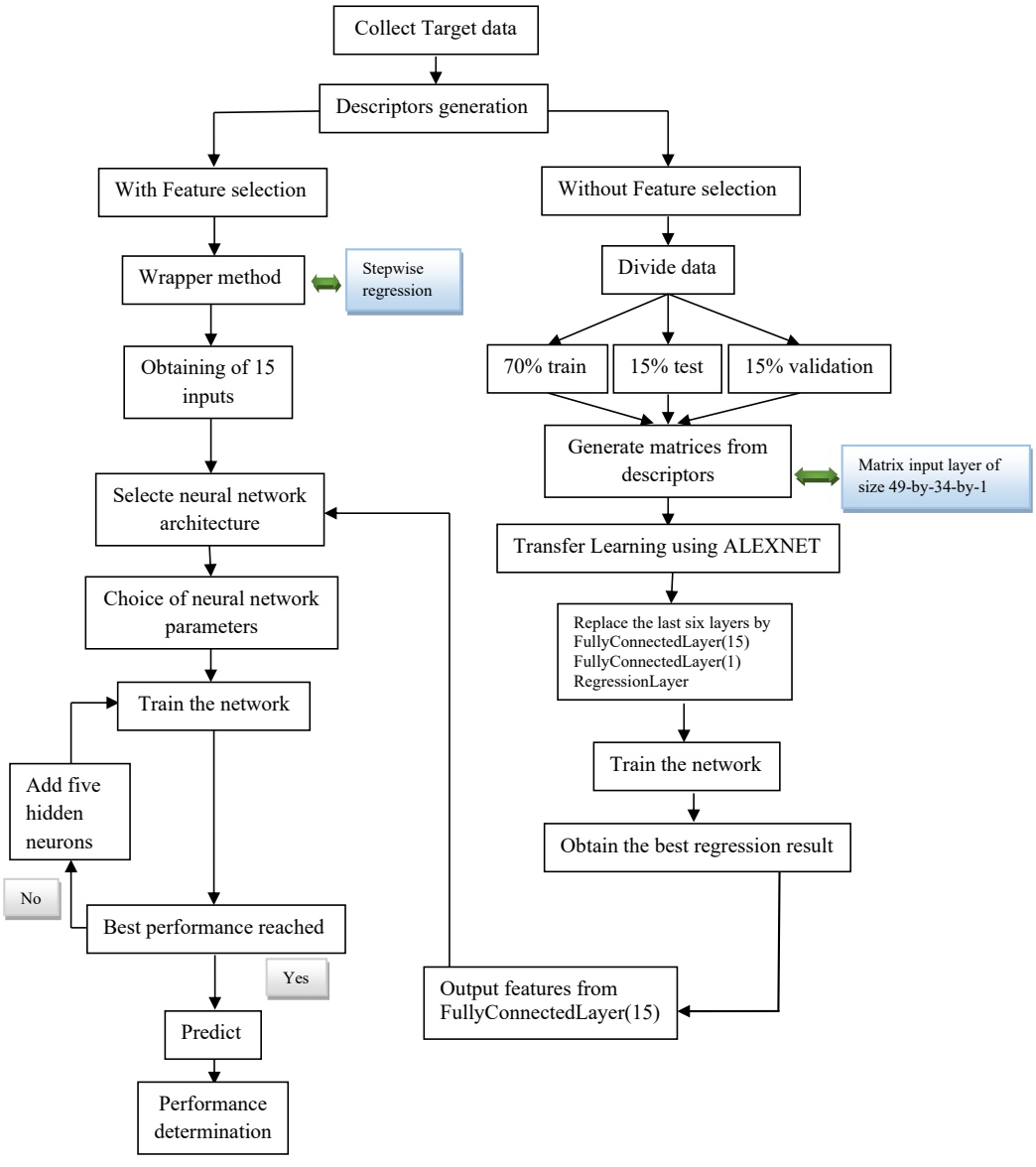


Figure 1. Flow sheet of the procedure followed

PBR

descriptors; *(ix)* topological charge indices; *(x)* eigenvalue-based indices; *(xi)* Randic molecular profiles; *(xii)* geometrical descriptors; *(xiii)* RDF descriptors; *(xiv)* 3D-MoRSE descriptors; *(xv)* WHIM descriptors; *(xvi)* GETAWAY descriptors; *(xvii)* functional group counts; *(xviii)* atom-centered fragments; *(xix)* charge descriptors; and *(xx)* molecular properties.

All descriptors were obtained through the E-Dragon online programs [34], which is known as the electronic remote version of the well-known software DRAGON developed by the Milano Chemometrics and QSAR Research Group by Prof. R. Todeschini. The name and number of calculated descriptors are reported in Supplemental Material 2.

Selection of relevant descriptors

The selection of the most efficient descriptors, which is an important step in QSAR modeling, was made by two techniques in this work.

Wrapper method

The number of molecular descriptors was reduced by the following procedure [35]:

1) Descriptors having constant values (min=max) were eliminated.

2) Quasi-constant descriptors (1st quartile 25%=2nd quartile 75%) were removed.

3) Descriptors with standard relative deviation *RSD*<0.05 were deleted.

The three steps above were performed using STATISTICA software [36].

4) Matrices of the pairwise linear correlation between each pair of columns in the input matrices were calculated via MATLAB [37]. In addition, all variables with a correlation coefficient R>0.75 were removed. For more robustness of the model, the variance inflation factor *VIF* was calculated (Equation 1):

$$1.\ VIF_i = 1/(1-R_i^2)$$

, Where is the squared correlation coefficient of the i^{th} descriptor versus the remaining ones. All the descriptors with VIF>5 were eliminated from the model [38].

5) The remaining descriptors were analyzed by stepwise regression in STATISTICA software [36]. Only 15 descriptors were selected and used as input variables in the FFNN models.

CNN method

CNNs are one of the most popular algorithms for deep learning, and they are mostly used for image classification. Computers do not see an image as an image, but as an array of pixels and the whole process in CNNs depends on it. Therefore, we provided our CNN network with previously digitalized data as input. Because for deep learning, in image classification, we do not have to understand which features are the best representations of the object, they are taught to us.

In our approach, the 1666 descriptors calculated with E-DRAGON were transformed for each drug into matrices of size 49-by-34-by-1 dimension in MATLAB because CNN requires matrices as inputs [39]. The latest features learned from the CNN layers were flattened out and passed through an FFNN to make predictions.

FFNN modeling

QSAR-wrapper-FFNN

The selected descriptors by the wrapper technique were set as input in an FFNN. So far, there is no clear answer to specify the number of hidden neurons required for a modeling task and it remains an open question. Different approaches are discovering this number, explained in detail in reviews, including methods of selecting the number of hidden nodes in the artificial neural networks review [40]. To decide the number of neurons in the hidden layer, the procedure summarized below was used [41]:

1) In the beginning, only a few hidden neurons (five) were used.

2) The network was trained until the mean square error no longer seemed to improve.

3) At this moment, a few (five) neurons were added to the hidden layer, each with randomly initialized weights, and resumed training.

4) The previous two steps were repeated until a termination criterion has been satisfied.

The mathematical equation of the model, for the prediction of protein binding, is shown below (Equation 2):

$$2.\ fb=\sum_{j=1}^{k} w2j \left(\frac{(\exp(\sum_{i=1}^{P} xi+wij+bj)-\exp(-\sum_{i=1} xi+wij+bj)}{(\exp(\sum_{i=1}^{P} xi+wij+bj)+\exp(-\sum_{i=1} xi+wij+bj)}+b\right)$$

x i (i=1…p) is the input that corresponds to the number of data included in the training of the ANN, I from 1 to 15, wij (i=1…p, j=1…k) weight from input to hidden layer, b j (j=1…k) are biases of the neurons in the hidden layer, k=45 for Wrapper method, w2j (j=1…k) are weights from the hidden to the output layer, b is the bias of the output neuron, and fb is the output.

QSAR-CNN-FFNN

CNN is a powerful machine learning technique. Training a CNN requires a large collection of diverse data, which is not an easy task [42]. However, there is an easy way, we can use a pre-trained CNN, which saves a huge amount of time and effort because fine-tuning a network with transfer learning is usually much faster and easier than training a network from scratch [43]. In this work, transfer learning using the ALEXNET CNN was applied to perform feature extraction and then feature reduction with adjustment in the last six layers by new ones. The CNN model was used to extract representative feature vectors from the penultimate fully connected layer, and then these ”deep features” were used for training an FFNN. The choice of dimension 15 for the feature reduction is to perform a comparison with the features selected by wrapper methods. The details of the 22 layers are reported in Supplemental Material 3. To train the network, holdout cross-validation was used to divide the data randomly 70% for training, 15% for validation, and 15% for testing.

Validation of the models

To determine both the generalizability and the actual predictive capacity of the QSAR models, internal and external validation criteria were used for the validation of the models. The statistical parameters used in our study to check the performance of the models were the coefficient of determination (R^2), the correlation coefficient (R), the predictive squared correlation coefficient (Q^2), and the mean squared error values (MSE) (Equations 3-5):

3. $R^2=1-RSS/SS$
4. $MSE=\sum_i^n=1 \frac{(y_i^{pred}-y_i)^2}{2}$
5. $Q^2=1-PRESS/SS$

The residual sum of squares (RSS) is the difference between the fitted values and the observed values. It refers to the difference between the observation and their mean. The predictive RSS is the difference between the predictions and the observations.

Results

QSAR-Wrapper-FFNN method

The results obtained from the selection of the most important descriptors by the wrapper method, using the correlation coefficient R and the variance inflation factor VIF, showed that 15 descriptors seemed to be the most appropriate. The calculated *VIF*s among the values of the selected descriptors were less than five, indicating that multicollinearity between the selected descriptors is acceptable. Supplemental Material 4 shows the *VIF* values for the selected descriptors and their meanings. To specify the number of hidden neurons required, the procedure detailed above was followed. R (all), MSE (validation), and (validation) criteria were employed for the evaluation of the accuracy of the best model. The best model was chosen according to the minimum MSE (validation) and the maximum R (all), and (validation) [35, 44].

Table 1 shows ten network models developed after the wrapper approach. The results obtained showed that network nine with 45 neurons was the best model with R (all)=0.935, R^2_{train}=0.875, Q^2=0.871, and MSE (validation)=0.015. The best performance of the model had a topology of 15-[45]-1:15 input nodes, one hidden layer with 45 nodes having the hyperbolic tangent as a transfer function, and one output layer with an identity transfer function. The neural networks were implemented using the neural network toolbox for MATLAB [37]. Figure 2 shows a comparison between the experiment and predicted plasma protein binding values for training, validation, and testing sets. The results showed a close correlation between predicted and observed plasma protein binding. The FFNN models were trained with the Levenberg-Marquardt backpropagation training function and gradient descent with momentum weight and bias learning function, and the data were partitioned

Table 1. Selected criteria of the different feed-forward neural networks (FFNNs) obtained by the quantitative structure-activity relationship (QSAR)-Wrapper-FFNN methodology

Number of Hidden Neurons	R (all)	R^2_{train}	Q^2	Mean Squared Error (Validation)
5	0.754	0.539	0.632	0.048
10	0.830	0.679	0.728	0.033
15	0.838	0.694	0.756	0.025
20	0.840	0.724	0.651	0.038
25	0.849	0.712	0.712	0.035
30	0.894	0.810	0.780	0.025
35	0.855	0.707	0.764	0.029
40	0.907	0.814	0.871	0.015
45	**0.935**	**0.875**	**0.871**	**0.015**
50	0.903	0.806	0.821	0.023

Table 2. Selected criteria for the different feed-forward neural networks (FFNNs) obtained from the quantitative structure-activity relationship (QSAR)- convolutional neural network (CNN)-FFNN method

Number of Hidden Neurons	R (all)	R^2_{train}	Q^2	Mean squared error (Validation)
5	0.882	0.812	0.701	0.040
10	0.908	0.832	0.798	0.009
15	0.913	0.856	0.760	0.028
20	0.933	0.857	0.948	0.009
25	0.910	0.834	0.830	0.022
30	0.905	0.830	0.775	0.022
35	0.960	0.905	0.964	0.005
40	0.922	0.832	0.961	0.007
45	0.940	0.874	0.899	0.012
50	0.922	0.830	0.841	0.009

PBR

using holdout cross-validation. The difference between R^2_{train} and Q^2 was equal to 0.004. This difference did not exceed 0.3 indicating the robustness of the model [45].

QSAR-CNN-FFNN method

Regarding the QSAR-CNN-FFNN, the results obtained (Table 2) showed that network seven was the best model with R (all)=0.960, R^2_{train}=0.905, Q^2=0.964, and MSE (validation)=0.005. The best performance of the model had a topology of 15-[35]-1:15 input nodes, one hidden layer with 35 nodes having the hyperbolic tangent as a transfer function, and an identity transfer function in the output layer. Figure 3 shows a comparison between experimental and predicted plasma protein binding for training, validation, and testing sets. The difference between R^2_{train} and Q^2 is equal to 0.059. This difference is less than 0.3 indicating the robustness of the model [45]. For QSAR-CNN-FFNN, the correla-

Table 3. External and internal criteria of the two models

	Parameters	QSAR-Wrapper-FFNN Method	QSAR-CNN-FFNN Method
Internal validation	R (all)	0.935	0.960
	R^2_{train}	0.875	0.905
	Q^2	0.871	0.964
	MSE	0.015	0.005
	MAE	0.090	0.054
	RMSE	0.122	0.069
	$R^2_{adjusted}$	0.874	0.963
External validation	R	0.933	0.982
	Q^2	0.870	0.945
	MSE	0.016	0.007
	MAE	0.094	0.068
	RMSE	0.129	0.085

PBR

Table 4. Comparison with the literature

Method/Ref.	MAE	R^2	R	MSE
Suggested method (QSAR-Wrapper-FFNN)	Train 0.080 Validation 0.090 Test 0.094	Train 0.875 Validation 0.871 Test 0.870	Train 0.935 Validation 0.933 Test 0.933	Train 0.014 Validation 0.015 Test 0.017
Suggested method (QSAR-CNN-FFNN)	Train 0.066 Validation 0.054 Test 0.068	Train 0.905 Validation 0.964 Test 0.945	Train 0.951 Validation 0.972 Test 0.982	Train 0.011 Validation 0.005 Test 0.007
Yuan et al. [27]	Test 0.076			
Sun et al. [25]	Test 0.126			
Kumar et al. [26]				Train 0.869 Test 0.8881
Li et al. [49]		Train 0.86		
Ghafourian et al. [21]	Train 13.25 Validation 14.96	Train 0.717 Validation 0.646	Train 0.681 Validation 0.641	
Moda et al. [50]	-	Test 0.91		

PBR

tion among each pair of our variables was verified and the results are presented in Supplemental Materials 5. In order to obtain an overview of the correlation structure, a heatmap was used to highlight what is important (Figure 4). The question now is how much is too much correlation between our neural network inputs. It seems that the general rule of thumb is that if the simple correlation coefficient between two regressors is greater than 0.8 or 0.9, multicollinearity is a serious problem [46]. However, with all the rules of thumb, we get in the statistics, there is nothing black or white about this. In our case, a prediction with a correlation coefficient of 0.8 does not appear fatal for our regression model. We can say that the above results served our theory, where more descriptors are valuable and welcome because we accept that by removing some of the features we are also discarding some of the data we have about the problem.

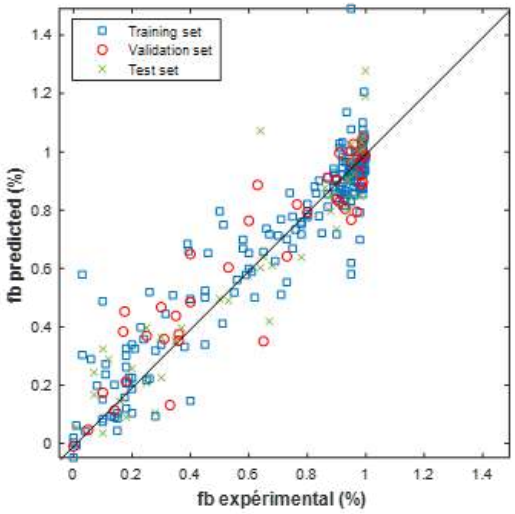
Discussion

Our results described for the first time the use of CNNs as a feature extraction method in QSAR studies for a set of descriptor matrices instead of local pattern extraction from images. One of the huge favorable circumstances of this method is that it does not take much time (only a few minutes) compared to the wrapper method (which took us days and concentration). It also automatically provides us with features that are the best representation of the non-image input, because in this work, we simply constructed the CNN inputs by resizing the 1666 features to 49×34 matrices. We knew that CNNs are used on data that has spatial features, and the fact that it gave us good features to use is that the CNN used discovered that our data has some sort of spatial features. We wanted to know what this deep CNN actually saw, and how it understands the inputs, we feed them. By visualizing the digital values of our matrices, grayscale images were obtained, giving 255 possible different shades of grey

Table 5. Applicability domain of the new method

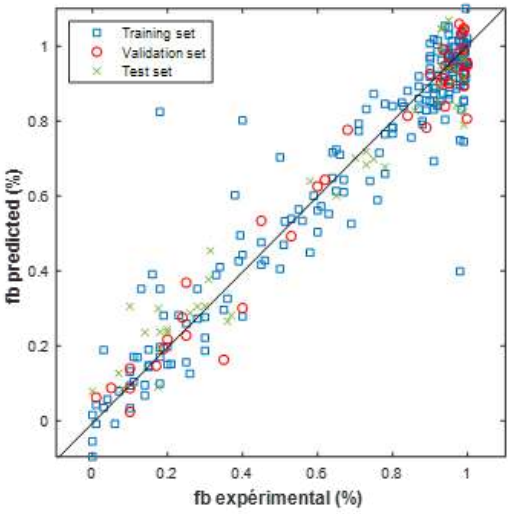
Approaches	Test Inside AD	Test Outside AD
Bounding box	40	2
Leverage	41	1
Euclidean distance (95 percentile)	41	1
Classical KNN (euclidean distance, k=5)	42	0
KNN (euclidean distance (k=25))	42	0

PBR



PBR

Figure 2. Comparison between experimental and calculated plasma protein binding predicted by the QSAR-Wrapper-FFNN technique.



PBR

Figure 3. Comparison between experimental and calculated values for plasma protein binding predicted by the QSAR-CNN-FFNN technique.

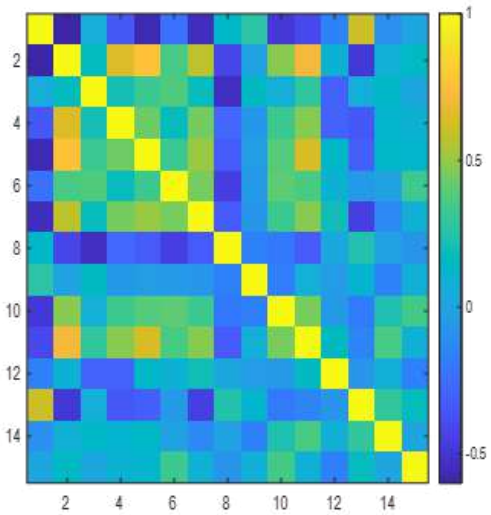
from black to white. Figure 5 shows a sample image of 3-carboxy-4-methyl-5-propyl-2-furanpropionic acid obtained. Numeric values below zero go to zero and those over one go to 255. Values between zero and one are shades of grey, with values between zero and 255. By taking the values of our descriptors and mapping them to the corresponding color suddenly, we can see the structure within our data. We can say that we have made an interface between our data and our brain by visualizing descriptors, in which physical and chemical information about molecules becomes accessible in a new form of fingerprints. They turn into storytellers of molecules secret in just an image.

Comparison of the two models

A statistical evaluation presented in Table 3 was conducted to compare the performance and quality of the predictions by the FFNN of the two models developed for this work. The statistical coefficients of the internal validation for the two models are all acceptable and satisfactory (high R^2_{train}, Q^2 (validation), $R^2_{adjusted}$ and lowest MSE, RMSE, and MAE) and consequently, these models are robust. The quality of the models was also evaluated in terms of external validation criteria. For the wrapper feature extraction method, a value of $Q^2>0.5$ is considered satisfactory and for the QSAR-CNN feature extraction method, the value of $Q^2>0.9$ is considered excellent [45]. We can say that the two models are distinguished by excellent predictive power.

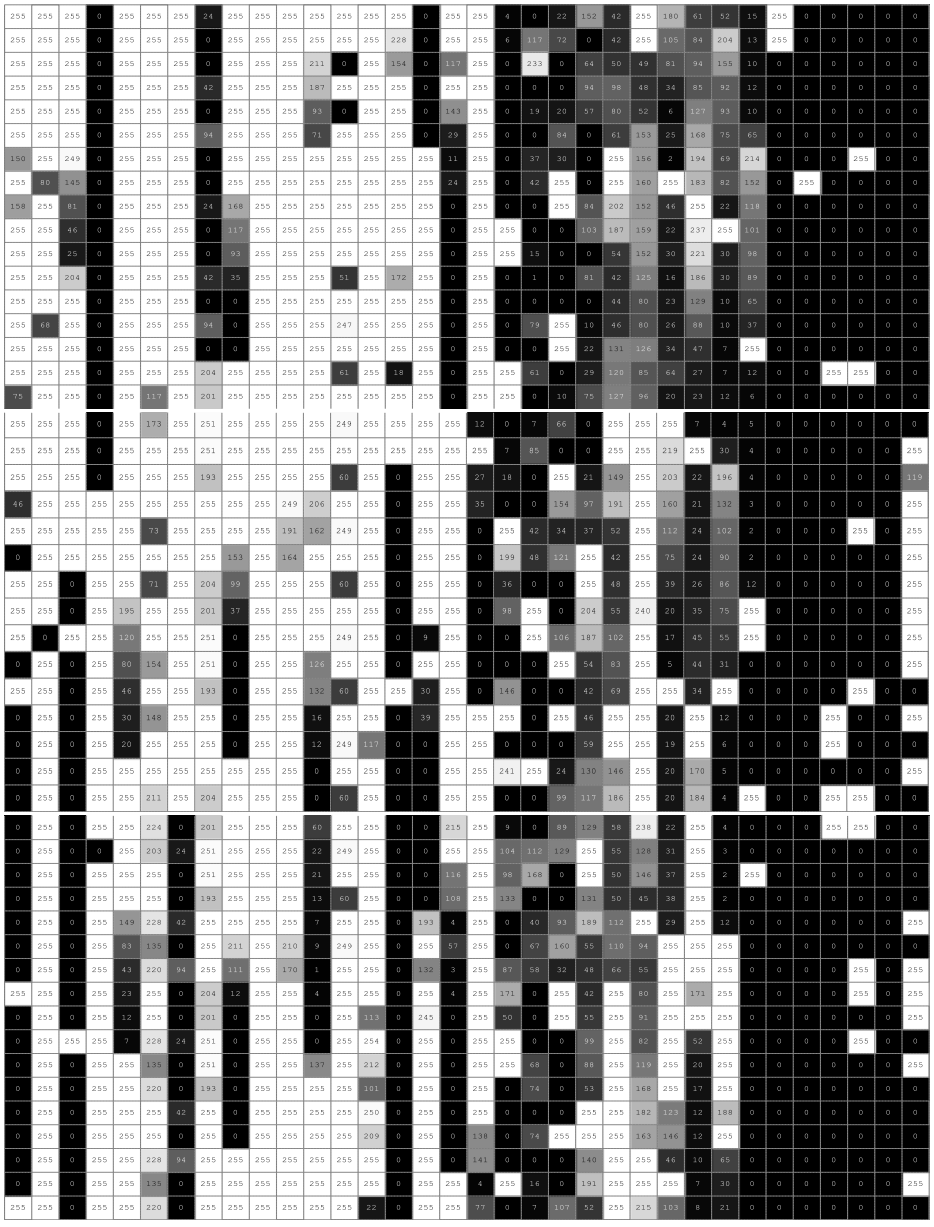
Comparison between models from the literature

A comparison was made between the few models reported in the literature with those developed in this work for the prediction of the binding of drugs to plasma proteins (Table 4). Yawen Yuan et al. constructed a QSAR model for predicting plasma protein binding based on a large training set comprising more than 5000 compounds and proposed a new strategy for constructing models for different binding levels. Lixia Sun et al. de-



PBR

Figure 4. Heatmap of the correlation matrix for the QSAR-CNN-FFNN method.

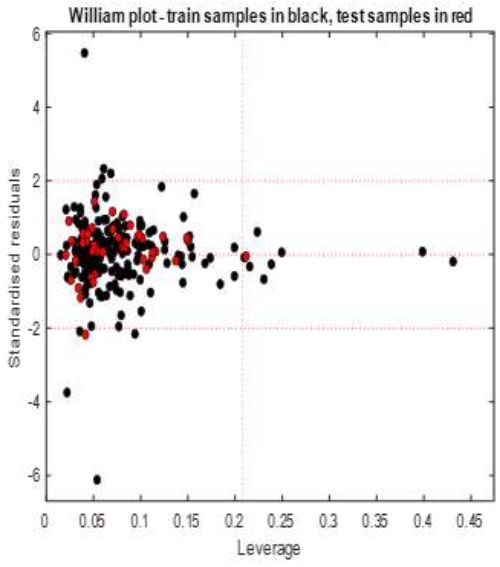


PBR

Figure 5. Image of the input matrices of 3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid with the corresponding numerical values of shade of gray.

veloped QSAR models by six machine learning algorithms with 26 molecular descriptors for predicting the protein-bound fractions of 967 pharmaceuticals. The neural network model yielded an MAE error of 0.126. Rajnish Kumar et al. presented six machine learning algorithms for QSAR models using a database containing 735 drugs. The model with the artificial neural network gave an MSE train=0.869 and MSE val=0.881. Haiyan Li et al. developed a new method called QSAR Plasma Protein Interaction QSAR Analysis (PPI-QSAR) with a database of 65 antibiotics, providing R^2_{train}=0.86 and Q^2=0.72. Taravat Ghafourian et al. collected a database of 794 drugs and used four data mining tools; the best model was boosted trees providing an error MAE train=13.25 and an MAE=14.96. Moda Tiago et al. developed a hologram quantitative structure-activity relationship (HQSAR) on a series of 312 drugs, and they obtained a Q^2=0.72 and R^2=0.91.

We were not interested in evaluating the advantages and disadvantages of these methods because it is quite difficult (each study used different data sets and different modeling approaches) and because the objective of our work was to prove that our new feature selection method



PBR

Figure 6. Plot of the residuals for calculated values of plasma protein binding from the new approach versus their experimental values for training and test sets.

is as effective as these old methods. According to these results, the models developed in this work performed as well as these methods.

Applicability domain

The third principle of the OECD guidelines [47] recommends a defined applicability domain. In this work, the domain of applicability was analyzed with different approaches reported in Table 5. The algorithm and methodology of the proposed approaches are available in the literature [48]. The number of samples inside the applicability domain varied depending on the method used: the bounding box considered two test samples out of the applicability domain, while leverage and Euclidean distance (95 percentile) identified one test sample out of the domain of applicability as shown in Williams plot (Figure 6). Despite its distance from the other observations, our point is close to the regression fitted line since it has a small residual. Classical KNN (Euclidean distance, k=5) and KNN (Euclidean distance (k=25)) showed none of the test samples out of the applicability domain. These results show that we can use the model to predict plasma protein binding for new compounds that have not been tested.

Conclusion

Two different approaches to feature selection in machine learning were tested. These feature selection strategies were followed to produce inputs, which were used in an FFNN to make predictions. Examination of the estimates of external and internal criteria revealed that the QSAR model developed by the new feature selection method is robust, externally predictive, and distinguished by a good applicability domain. The external accuracy of the validation set was calculated by the Q^2 and RMSE, which are equal to 0.945 and 0.085, respectively indicating that the accuracy of NN trained with extracted features from CNNs is slightly better than the accuracy of NN trained with features obtained by the current predictors. This investigation showed that extracting features using CNNs takes less time (only a few minutes) and summarizes much of the information contained in the original features. Contrary to feature selection techniques, we do not have to figure out which features are the best representations of the input, they are learned for us. In the end, according to the OECD principle, we can say that we can use this QSAR-CNN-FFNN model to predict the fraction of plasma protein binding to human plasma for drugs that have not been tested to avoid chemical synthesis and reduce expansive laboratory tests.

Ethical Considerations

Compliance with ethical guidelines

There were no ethical considerations to be considered in this research.

Funding

The authors declare no conflict of interest.

Authors' contributions

Study design: Salah Hanini; Investigation, data collection, and data analysis: Khaouane Affaf; Supervision and Writing–original draft: Samira Ferhat; Writing–review & editing: all authors.

Conflict of interest

The authors declared no conflict of interest.

Acknowledgments

The authors thank the Laboratory of Biomaterial and Transport Phenomena (LBMPT), University of Médéa, for providing the facilities to conduct this study.

References

[1] Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr. Computational methods in drug discovery. Pharmacol Rev. 2013; 66(1):334-95. [DOI:10.1124/pr.112.007336] [PMID] [PMCID]

[2] Lin X, Li X, Lin X. A review on applications of computational methods in drug screening and design. Molecules. 2020; 25(6):1375. [DOI:10.3390/molecules25061375] [PMID] [PMCID]

[3] Kubinyi H. QSAR and 3D QSAR in drug design Part 1: methodology. Drug Discovery Today. 1997; 2(11):457-67. [DOI:10.1016/S1359-6446(97)01084-2]

[4] Khan MT. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. Curr Drug Metab. 2010; 11:285-95. [DOI:10.2174/138920010791514306] [PMID]

[5] Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. Drug Discov Today. 2017; 22(11):1680-5. [DOI:10.1016/j.drudis.2017.08.010] [PMID]

[6] Rasulev BF, Abdullaev ND, Syrov VN, Leszczynski J. A quantitative structure-activity relationship (QSAR) study of the antioxidant activity of flavonoids. QSAR Comb Sci. 2005; 24:1056-65. [DOI:10.1002/qsar.200430013]

[7] Nirmalakhandan N, Speece RE. ES & T critical review: Structure-activity relationships. Quantitative techniques for predicting the behavior of chemicals in the ecosystem. Environ Sci Technol. 1988; 22:606-15. [DOI:10.1021/es00171a002]

[8] Grover M, Singh B, Bakshi M, Singh S. Quantitative structure-property relationships in pharmaceutical research-Part 1. Pharm Sci Technol. 2000; 3(2):50-57. [DOI:10.1016/S1461-5347(99)00214-X]

[9] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinformatics. 2015; 2015:198363. [DOI:10.1155/2015/198363] [PMID] [PMCID]

[10] Langley P. Selection of relevant features in machine learning. Palo Alto: Institute for the Study of Learning and Expertise; 1994. [DOI:10.21236/ADA292575]

[11] Dash M, Liu H. Feature selection for classification. Intell Data Anal. 1997; 1(1-4):131-56. [DOI:10.1016/S1088-467X(97)00008-5]

[12] Tetko IV, Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. J Chem Inf Comput Sci. 1995:35(5):826–33. [DOI:10.1021/ci00027a006]

[13] Dudek AZ, Arodz T, Gálvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. Comb Chem High Throughput Screen. 2006; 9(3):213-28. [DOI:10.2174/138620706776055539] [PMID]

[14] Ghasemi F, Mehridehnavi A, Pérez-Garrido A, Pérez-Sánchez H. Neural network and deep-learning algorithms used in QSAR studies: Merits and drawbacks. Drug Discov Today. 2018; 23(10):1784-90. [DOI:10.1016/j.drudis.2018.06.016] [PMID]

[15] Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: Paradigm for drug design. Drug Discov Today. 2016; 21(8):1291-302. [DOI:10.1016/j.drudis.2016.06.013] [PMID]

[16] Bohnert T, Gan LS. Plasma protein binding: From discovery to development. J Pharm Sci. 2013; 102(9):2953-94. [DOI:10.1002/jps.23614] [PMID]

[17] Trainor GL. The importance of plasma protein binding in drug discovery. Expert Opin Drug Discov. 2007; 2(1):51-64. [DOI:10.1517/17460441.2.1.51] [PMID]

[18] Zsila F, Bikadi Z, Malik D, Hari P, Pechan I, Berces A, et al. Evaluation of drug-human serum albumin binding interactions with support vector machine aided online automated docking. Bioinformatics. 2011; 27(13):1806-13. [DOI:10.1093/bioinformatics/btr284] [PMID]

[19] Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. 2008; 36(9):3025-30. [DOI:10.1093/nar/gkn159] [PMID] [PMCID]

[20] Zhao C, Zhang H, Zhang X, Zhang R, Luan F, Liu M, et al. Prediction of milk/plasma drug concentration (M/P) ratio using support vector machine (SVM) method. Pharmaceutical Res. 2006; 23(1):41-8. [DOI:10.1007/s11095-005-8716-4] [PMID]

[21] Ghafourian T, Amin Z. QSAR models for the prediction of plasma protein binding. Bioimpacts. 2013; 3(1):21-7. [DOI:10.5681/bi.2013.011] [PMID] [PMCID]

[22] Fu X, Wang G, Gao J, Zhan S, Liang W. Prediction of plasma protein binding of cephalosporins using an artificial neural network. Pharmazie. 2007; 62(2):157-8. [PMID]

[23] Turner JV, Maddalena DJ, Cutler DJ. Pharmacokinetic parameter prediction from drug structure using artificial neural networks. Int J Pharm. 2004; 270(1-2):209-19. [DOI:10.1016/j.ijpharm.2003.10.011] [PMID]

[24] Deng L, Sui Y, Zhang J. XGBPRH: Prediction of binding hot spots at protein-RNA interfaces utilizing extreme gradient boosting. Genes. 2019; 10(3):242. [DOI:10.3390/genes10030242] [PMID] [PMCID]

[25] Sun L, Yang H, Li J, Wang T, Li W, Liu G, et al. In silico prediction of compounds binding to human plasma proteins by QSAR models. Chem Med Chem. 2018; 13(6):572-81. [DOI:10.1002/cmdc.201700582] [PMID]

[26] Kumar R, Sharma A, Siddiqui MH, Tiwari RK. Prediction of drug-plasma protein binding using artificial intelligence based algorithms. Comb Chem High Throughput Screen. 2018; 21(1):57-64. [DOI:10.2174/1386207321666171218121557] [PMID]

[27] Yuan Y, Chang S, Zhang Z, Li Z, Li S, Xie P, et al. A novel strategy for prediction of human plasma protein binding using machine learning techniques. Chemometr Intell Lab Syst. 2020; 199:103962. [DOI:10.1016/j.chemolab.2020.103962]

[28] Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. ACS Cent Sci. 2017; 3(4):283-93. [DOI:10.1021/acscentsci.6b00367] [PMID] [PMCID]

[29] Wallach I, Dzamba M, Heifets A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXi v:1510.02855v1. 2015; 1-11. [DOI:10.48550/arXiv.1510.02855]

[30] Brunton L, Lazo J, Parker K. Goodman & Gilman's the pharmacological basis of therapeutics. New York: McGraw Hill Professional; 2005. [Link]

[31] Troutman WG. Handbook of clinical drug data. New York: McGraw-Hill; 2002. [Link]

[32] PubChem. National Institutes of Health (NIH). https://pubchemdocs.ncbi.nlm.nih.gov/

[33] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988; 28(1):31–6. [DOI:10.1021/ci00057a005]

[34] Virtual Computational Chemistry Laboratory (VCCLAB). http://www.vcclab.org/

[35] Hamadache M, Benkortbi O, Hanini S, Amrane A, Khaouane L, Moussa CS. A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction. J Hazard Mater. 2016; 303:28-40. [DOI:10.1016/j.jhazmat.2015.09.021] [PMID]

[36] Stat Soft Inc. STATISTICA, Version 8 [Internet]. 2007 [Updated November 2022]. Available from: **[Link]**

[37] Math Works MATLAB R2019b. MathWorks, Inc. [Internet]. 2022 [Updated November 2022]. Available from: [Link]

[38] Akinwande MO, Dikko HG, Samson A. Variance inflation factor: As a condition for the inclusion of suppressor variable (s) in regression analysis. "Open J Stat. 2015; 5:754-67. [DOI:10.4236/ojs.2015.57075]

[39] Bazgir O, Zhang R, Dhruba SR, Rahman R, Ghosh S, Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. Nat Commun. 2020; 11:4391. [DOI:10.1038/s41467-020-18197-y] [PMID] [PMCID]

[40] Panchal FS, Panchal M. Review on methods of selecting number of hidden nodes in artificial neural network. Int J Comput Sci Mob Computing. 2014; 3(11):455-64. [Link]

[41] Kubat M. An introduction to machine learning. Cham: Springer; 2017. [DOI:10.1007/978-3-319-63913-0]

[42] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging. 2016; 35(5):1299-312 [DOI:10.1109/TMI.2016.2535302] [PMID]

[43] Schwarz M, Schulz H, Behnke S. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. Paper presented at: 2015 IEEE International Conference on Robotics and Automation (ICRA). 26-30 May 2015; Seattle, USA. [DOI:10.1109/ICRA.2015.7139363]

[44] Bitam S, Hamadache M, Hanini S. QSAR model for prediction of the therapeutic potency of N-benzylpiperidine derivatives as AChE inhibitors. SAR QSAR Environ Res. 2017; 28(6):471-49. [DOI:10.1080/1062936X.2017.1331467] [PMID]

[45] Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect. 2003; 111(10):1361-75. [DOI:10.1289/ehp.5758] [PMID] [PMCID]

[46] Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. J Interdiscip Math. 2010; 13:253-67. [DOI:10.1080/09720502.2010.10700699]

[47] Organisation for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models., Paris: OECD Publishing; 2014. [Link]

[48] Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst. 2015; 145:22-9. [DOI:10.1016/j.chemolab.2015.04.013]

[49] Li H, Chen Z, Xu X, Sui X, Guo T, Liu W, et al. Predicting human plasma protein binding of drugs using plasma protein interaction QSAR analysis (PPI-QSAR). Biopharm Drug Dispos. 2011; 32(6):333-42. [DOI:10.1002/bdd.762] [PMID]

[50] Moda TL, Montanari CA, Andricopulo AD. *In silico* prediction of human plasma protein binding using hologram QSAR. Lett Drug Des Discov. 2007; 4(7):502-9. [DOI:10.2174/157018007781788480]