Université de Lille
Année Universitaire 2019/2020

Faculté de Pharmacie de Lille

THESE
POUR LE DIPLOME D'ETAT
DE DOCTEUR EN PHARMACIE

Soutenue publiquement le 10 septembre 2020
Par M. BEDART Corentin

CONTRIBUTION DES METHODES *IN SILICO* DANS LE PROCESSUS DE CONCEPTION DE MEDICAMENTS

Membres du jury :

Président : Pr. CHAVATTE Philippe, Professeur des Universités, Faculté de Pharmacie de Lille

Directeur, Conseiller de thèse : Pr. CHAVATTE Philippe, Docteur en Pharmacie, Professeur des Universités, Faculté de Pharmacie de Lille

Assesseur : Dr. STANDAERT Annie, Docteur en Pharmacie, Maître de Conférences des Universités, Faculté de Pharmacie de Lille

Membre extérieur : Dr. PLAETEVOET Marina, Docteur en Pharmacie, Pharmacien d'officine, Pharmacie Plaetevoet à Flines-lez-Râches

Faculté de Pharmacie
de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE
☎ 03.20.96.40.40 - 🖷 : 03.20.96.43.64
http://pharmacie.univ-lille.fr

Université de Lille

Président :	Jean-Christophe CAMART
Premier Vice-président :	Nicolas POSTEL
Vice-présidente formation :	Lynne FRANJIÉ
Vice-président recherche :	Lionel MONTAGNE
Vice-président relations internationales :	François-Olivier SEYS
Vice-président stratégie et prospective	Régis BORDET
Vice-présidente ressources	Georgette DAL
Directeur Général des Services :	Pierre-Marie ROBERT
Directrice Générale des Services Adjointe :	Marie-Dominique SAVINA

Faculté de Pharmacie

Doyen :	Bertrand DÉCAUDIN
Vice-doyen et Assesseur à la recherche :	Patricia MELNYK
Assesseur aux relations internationales :	Philippe CHAVATTE
Assesseur aux relations avec le monde professionnel :	Thomas MORGENROTH
Assesseur à la vie de la Faculté :	Claire PINÇON
Assesseur à la pédagogie :	Benjamin BERTIN
Responsable des Services :	Cyrille PORTA
Représentant étudiant :	Victoire LONG

Liste des Professeurs des Universités - Praticiens Hospitaliers

Civ.	Nom	Prénom	Laboratoire
Mme	ALLORGE	Delphine	Toxicologie et Santé publique
M.	BROUSSEAU	Thierry	Biochimie
M.	DÉCAUDIN	Bertrand	Biopharmacie, Pharmacie Galénique et Hospitalière
M.	DEPREUX	Patrick	Institut de Chimie Pharmaceutique Albert LESPAGNOL
M.	DINE	Thierry	Pharmacologie, Pharmacocinétique et Pharmacie clinique
Mme	DUPONT-PRADO	Annabelle	Hématologie
Mme	GOFFARD	Anne	Bactériologie - Virologie

M.	GRESSIER	Bernard	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	ODOU	Pascal	Biopharmacie, Pharmacie Galénique et Hospitalière
Mme	POULAIN	Stéphanie	Hématologie
M.	SIMON	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	STAELS	Bart	Biologie cellulaire

Liste des Professeurs des Universités

Civ.	Nom	Prénom	Laboratoire
M.	ALIOUAT	El Moukhtar	Parasitologie - Biologie animale
Mme	AZAROUAL	Nathalie	Biophysique et Laboratoire d'application de RMN
M.	CAZIN	Jean-Louis	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	CHAVATTE	Philippe	Institut de Chimie Pharmaceutique Albert LESPAGNOL
M.	COURTECUISSE	Régis	Sciences Végétales et Fongiques
M.	CUNY	Damien	Sciences Végétales et Fongiques
Mme	DELBAERE	Stéphanie	Biophysique et application de RMN
Mme	DEPREZ	Rebecca	Médicaments et molécules pour agir sur les systèmes vivants
M.	DEPREZ	Benoît	Médicaments et molécules pour agir sur les systèmes vivants
M.	DUPONT	Frédéric	Sciences Végétales et Fongiques
M.	DURIEZ	Patrick	Physiologie
M.	FOLIGNÉ	Benoît	Bactériologie - Virologie
M.	GARÇON	Guillaume	Toxicologie et Santé publique
Mme	GAYOT	Anne	Pharmacotechnie industrielle
M.	GOOSSENS	Jean-François	Chimie analytique
M.	HENNEBELLE	Thierry	Pharmacognosie
M.	LEBEGUE	Nicolas	Chimie thérapeutique
M.	LEMDANI	Mohamed	Biomathématiques

Mme	LESTAVEL	Sophie	Biologie cellulaire
Mme	LESTRELIN	Réjane	Biologie cellulaire
Mme	MELNYK	Patricia	Chimie thérapeutique
M.	MILLET	Régis	Institut de Chimie Pharmaceutique Albert LESPAGNOL
Mme	MUHR-TAILLEUX	Anne	Biochimie
Mme	PERROY	Anne-Catherine	Législation et Déontologie pharmaceutique
Mme	ROMOND	Marie-Bénédicte	Bactériologie - Virologie
Mme	SAHPAZ	Sevser	Pharmacognosie
M.	SERGHERAERT	Éric	Législation et Déontologie pharmaceutique
M.	SIEPMANN	Juergen	Pharmacotechnie industrielle
Mme	SIEPMANN	Florence	Pharmacotechnie industrielle
M.	WILLAND	Nicolas	Médicaments et molécules pour agir sur les systèmes vivants

Liste des Maîtres de Conférences - Praticiens Hospitaliers

Civ.	Nom	Prénom	Laboratoire
Mme	BALDUYCK	Malika	Biochimie
Mme	GARAT	Anne	Toxicologie et Santé publique
Mme	GENAY	Stéphanie	Biopharmacie, Pharmacie Galénique et Hospitalière
M.	LANNOY	Damien	Biopharmacie, Pharmacie Galénique et Hospitalière
Mme	ODOU	Marie-Françoise	Bactériologie - Virologie

Liste des Maîtres de Conférences

Civ.	Nom	Prénom	Laboratoire
M.	AGOURIDAS	Laurence	Chimie thérapeutique
Mme	ALIOUAT	Cécile-Marie	Parasitologie - Biologie animale
M.	ANTHÉRIEU	Sébastien	Toxicologie et Santé publique
Mme	AUMERCIER	Pierrette	Biochimie

M.	BANTUBUNGI-BLUM	Kadiombo	Biologie cellulaire
Mme	BARTHELEMY	Christine	Biopharmacie, Pharmacie Galénique et Hospitalière
Mme	BEHRA	Josette	Bactériologie - Virologie
M.	BELARBI	Karim-Ali	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	BERTHET	Jérôme	Biophysique et Laboratoire d'application de RMN
M.	BERTIN	Benjamin	Immunologie
M.	BLANCHEMAIN	Nicolas	Pharmacotechnie industrielle
M.	BORDAGE	Simon	Pharmacognosie
M.	BOSC	Damien	Médicaments et molécules pour agir sur les systèmes vivants
M.	BRIAND	Olivier	Biochimie
M.	CARNOY	Christophe	Immunologie
Mme	CARON-HOUDE	Sandrine	Biologie cellulaire
Mme	CARRIÉ	Hélène	Pharmacologie, Pharmacocinétique et Pharmacie clinique
Mme	CHABÉ	Magali	Parasitologie - Biologie animale
Mme	CHARTON	Julie	Médicaments et molécules pour agir sur les systèmes vivants
M.	CHEVALIER	Dany	Toxicologie et Santé publique
Mme	DANEL	Cécile	Chimie analytique
Mme	DEMANCHE	Christine	Parasitologie - Biologie animale
Mme	DEMARQUILLY	Catherine	Biomathématiques
M.	DHIFLI	Wajdi	Biomathématiques
Mme	DUMONT	Julie	Biologie cellulaire
M.	EL BAKALI	Jamal	Chimie thérapeutique
M.	FARCE	Amaury	Institut de Chimie Pharmaceutique Albert LESPAGNOL
M.	FLIPO	Marion	Médicaments et molécules pour agir sur les systèmes vivants
Mme	FOULON	Catherine	Chimie analytique

M.	FURMAN	Christophe	Institut de Chimie Pharmaceutique Albert LESPAGNOL
M.	GERVOIS	Philippe	Biochimie
Mme	GOOSSENS	Laurence	Institut de Chimie Pharmaceutique Albert LESPAGNOL
Mme	GRAVE	Béatrice	Toxicologie et Santé publique
Mme	GROSS	Barbara	Biochimie
M.	HAMONIER	Julien	Biomathématiques
Mme	HAMOUDI-BEN YELLES	Chérifa-Mounira	Pharmacotechnie industrielle
Mme	HANNOTHIAUX	Marie-Hélène	Toxicologie et Santé publique
Mme	HELLEBOID	Audrey	Physiologie
M.	HERMANN	Emmanuel	Immunologie
M.	KAMBIA KPAKPAGA	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	KARROUT	Younes	Pharmacotechnie industrielle
Mme	LALLOYER	Fanny	Biochimie
Mme	LECOEUR	Marie	Chimie analytique
Mme	LEHMANN	Hélène	Législation et Déontologie pharmaceutique
Mme	LELEU	Natascha	Institut de Chimie Pharmaceutique Albert LESPAGNOL
Mme	LIPKA	Emmanuelle	Chimie analytique
Mme	LOINGEVILLE	Florence	Biomathématiques
Mme	MARTIN	Françoise	Physiologie
M.	MOREAU	Pierre-Arthur	Sciences Végétales et Fongiques
M.	MORGENROTH	Thomas	Législation et Déontologie pharmaceutique
Mme	MUSCHERT	Susanne	Pharmacotechnie industrielle
Mme	NIKASINOVIC	Lydia	Toxicologie et Santé publique
Mme	PINÇON	Claire	Biomathématiques
M.	PIVA	Frank	Biochimie

Mme	PLATEL	Anne	Toxicologie et Santé publique
M.	POURCET	Benoît	Biochimie
M.	RAVAUX	Pierre	Biomathématiques / service innovation pédagogique
Mme	RAVEZ	Séverine	Chimie thérapeutique
Mme	RIVIÈRE	Céline	Pharmacognosie
M.	ROUMY	Vincent	Pharmacognosie
Mme	SEBTI	Yasmine	Biochimie
Mme	SINGER	Elisabeth	Bactériologie - Virologie
Mme	STANDAERT	Annie	Parasitologie - Biologie animale
M.	TAGZIRT	Madjid	Hématologie
M.	VILLEMAGNE	Baptiste	Médicaments et molécules pour agir sur les systèmes vivants
M.	WELTI	Stéphane	Sciences Végétales et Fongiques
M.	YOUS	Saïd	Chimie thérapeutique
M.	ZITOUNI	Djamel	Biomathématiques

Professeurs Certifiés

Civ.	Nom	Prénom	Laboratoire
Mme	FAUQUANT	Soline	Anglais
M.	HUGES	Dominique	Anglais
M.	OSTYN	Gaël	Anglais

Professeur Associé - mi-temps

Civ.	Nom	Prénom	Laboratoire
M.	DAO PHAN	Haï Pascal	Médicaments et molécules pour agir sur les systèmes vivants
M.	DHANANI	Alban	Législation et Déontologie pharmaceutique

Maîtres de Conférences ASSOCIES - mi-temps

Civ.	Nom	Prénom	Laboratoire
Mme	CUCCHI	Malgorzata	Biomathématiques
M.	DUFOSSEZ	François	Biomathématiques
M.	FRIMAT	Bruno	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	GILLOT	François	Législation et Déontologie pharmaceutique
M.	MASCAUT	Daniel	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	ZANETTI	Sébastien	Biomathématiques

AHU

Civ.	Nom	Prénom	Laboratoire
Mme	CUVELIER	Élodie	Pharmacologie, Pharmacocinétique et Pharmacie clinique
Mme	DEMARET	Julie	Immunologie
M.	GRZYCH	Guillaume	Biochimie
Mme	HENRY	Héloïse	Biopharmacie, Pharmacie Galénique et Hospitalière
Mme	MASSE	Morgane	Biopharmacie, Pharmacie Galénique et Hospitalière

ATER

Civ.	Nom	Prénom	Laboratoire
M.	GHARBI	Zied	Biomathématiques
Mme	FLÉAU	Charlotte	Médicaments et molécules pour agir sur les systèmes vivants
Mme	N'GUESSAN	Cécilia	Parasitologie - Biologie animale
M.	RUEZ	Richard	Hématologie
M.	SAIED	Tarak	Biophysique et Laboratoire d'application de RMN
Mme	VAN MAELE	Laurye	Immunologie

Enseignant contractuel

Civ.	Nom	Prénom	Laboratoire
M.	MARTIN MENA	Anthony	Biopharmacie, Pharmacie Galénique et Hospitalière

Université de Lille

Faculté de Pharmacie de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX
Tel. : 03.20.96.40.40 - Télécopie : 03.20.96.43.64
http://pharmacie.univ-lille.fr

Remerciements

Au Pr. Philippe CHAVATTE, pour m'avoir fait l'honneur d'encadrer ma thèse d'exercice et de présider le jury. Merci énormément pour l'intérêt porté pour mes travaux, aussi bien pour cette thèse d'exercice que pour ma thèse de doctorat.

Au Dr. Annie STANDAERT, pour avoir accepté de faire partie de mon jury de thèse malgré un domaine de recherche tout à fait à l'opposé des domaines habituellement traités, et pour son aide précieuse dans le déroulement de ma thèse de doctorat.

Au Dr. Marina PLAETEVOET, pour avoir suivi en entier mes études de Pharmacie, du tout premier stage officinal jusqu'à la soutenance de ma thèse d'exercice, et pour sa confiance qui m'a été accordée.

A mes parents, ma sœur et mon frère, pour m'avoir toujours soutenu, directement ou indirectement, dans mes études, mes envies et mes choix.

A ma famille et belle-famille pour leur soutien moral et leurs encouragements.

A mes amis de ma faculté, pour les bons moments passés ensemble, les belotes, les séances de décoreux, les bières, …

A toutes les équipes de recherche de m'avoir accepté le temps de quelques mois de stage au cours de mes études.

A Par'Immune, l'Unité Inserm U1286 INFINITE et l'Institut de Chimie Pharmaceutique Albert Lespagnol de m'avoir donné l'opportunité de continuer mes études au-delà de mon cursus de Pharmacie dans le domaine de la chémoinformatique, qui m'est particulièrement cher.

A Adeline, Amélie, Xavier, Amaury, Nicolas et Antoine pour m'avoir précieusement aidé, à toute heure, lors de la rédaction de ce mémoire.

Enfin, à Pauline, qui me soutient tous les jours, m'encourage, me pousse à être meilleur, et qui m'a poussé à prendre des risques qui ont été payants et qui me permettent de faire ce que j'aime. Sans toi je n'en serais pas là aujourd'hui. Merci. Je t'aime.

Table des matières

1. Introduction

1.1. Conception de médicaments

1.1.1. Généralités

En 2013, plus de 2800 substances actives différentes correspondant à plus de 11000 spécialités pharmaceutiques étaient disponibles sur le marché français (1). Les avancées de la science et de la médecine ces dernières décennies ont apporté des bénéfices non négligeables pour le patient, avec la diminution du nombre d'effets secondaires induits, l'augmentation de la qualité de vie générale et l'augmentation de l'espérance de vie (2). Toutefois, dans une optique d'amélioration continue de la qualité des produits de santé et l'objectif de traiter des pathologies aux processus complexes, la conception de nouveaux médicaments est devenue un processus bien plus compliqué, long et coûteux (3).

Le développement d'une nouvelle substance active, de la toute première recherche fondamentale à son autorisation de mise sur le marché et sa commercialisation, est un long processus demandant en moyenne plus de 10 ans. Ceci est expliqué entre autres par les nombreux défis techniques, scientifiques et réglementaires. Le coût moyen du développement entier d'un médicament atteignant avec succès le marché a également augmenté exponentiellement, passant de 400 millions de dollars au début des années 1990, à 1 milliard de dollars au début des années 2000, puis à plus de 2,6 milliards de dollars aujourd'hui, soit une augmentation moyenne de 250% par décennie (3) (Figure 1).

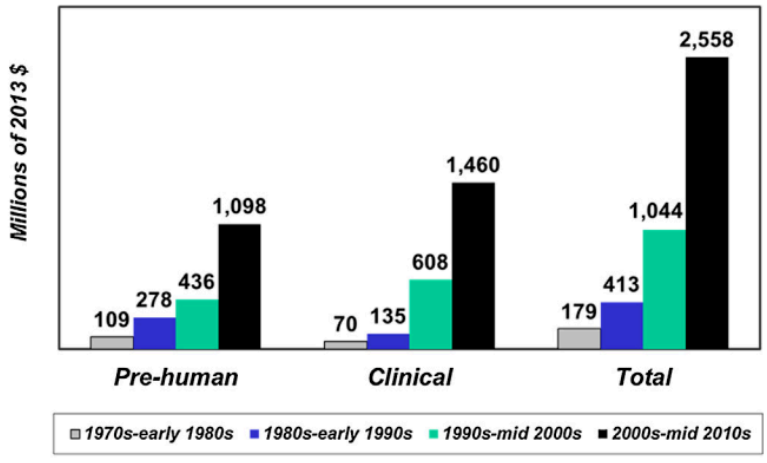


Figure 1 : Evolution du coût des étapes de recherche et de développement préclinique (« *pre-human* ») , du développement clinique et du coût total pour la conception d'un nouveau médicament en millions de dollars de 2013 (3–6)

1.1.2. Processus global

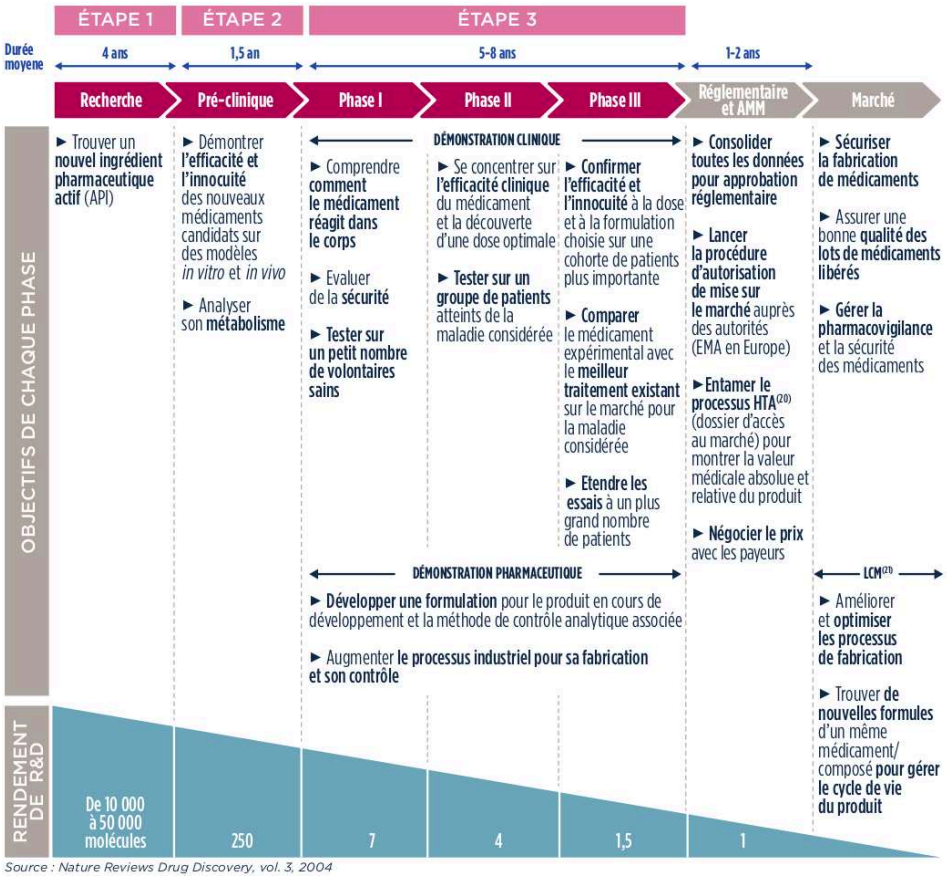


Figure 2 : Grandes étapes de recherche et développement d'un médicament (2)

Le processus moderne de conception d'un médicament s'inscrit au sein de quatre grandes étapes successives (Figure 2).

Recherche

La première étape du processus de conception de médicaments, la phase de recherche, vise à la détermination d'une cible biologique, la recherche de composés chimiques ayant une activité sur celle-ci, et leur protection intellectuelle.

Elle vise dans un premier temps à comprendre les mécanismes inhérents à une pathologie, afin de définir une cible sur laquelle le futur médicament aura un effet : c'est la recherche fondamentale (2). Cette cible peut être dans la grande majorité :

- Une enzyme spécifique possédant une activité chimique précise au sein de l'organisme,
- Un récepteur membranaire ou nucléaire entraînant une cascade d'effets biologiques,
- Un canal ionique ou un transporteur de molécule permettant le passage d'ions ou de molécules entre le milieu extracellulaire et une cellule.

Il est à noter que le temps moyen du développement d'un médicament ne prend pas en compte la durée de la recherche fondamentale réalisée en amont.

Dans un second temps, des milliers de molécules, généralement de 10.000 à 50.000 molécules, sont testées afin d'étudier leurs propriétés chimiques et pharmacologiques sur la cible déterminée au préalable. Les molécules ayant un potentiel effet thérapeutique sont sélectionnées, et appelées « touches » (ou « *hits* ») (7). Ces touches sont optimisées dans le but d'améliorer leur efficacité, leur sélectivité, leur métabolisme agissant sur le temps d'action du médicament, et leurs paramètres physico-chimiques en prévoyant leur future forme galénique, voie d'administration et d'élimination. Les meilleures touches suivent le processus de conception, et sont appelées « chefs de file » (ou « *leads* ») (7). Cette phase demande en moyenne 4 ans de développement (2).

Enfin, dans un troisième temps, il est nécessaire de protéger intellectuellement ces molécules au travers d'un ou de plusieurs dépôts de brevet. Chaque brevet permet une protection des innovations pendant 20 ans, donnant une exclusivité commerciale temporaire. Toutefois, le dépôt de brevet est limité à certains critères, que sont (8):

- Le caractère réellement novateur de la découverte (aucune innovation similaire ne doit être accessible au public)
- L'implication d'une activité inventive (qui ne découle pas de manière évidente de la technique connue par un professionnel du domaine)
- L'existence d'une application industrielle.

Des extensions de durée de protection existent spécifiquement pour les produits pharmaceutiques, avec une possible extension d'un maximum de 5 ans grâce au Certificat Complémentaire de Protection (CCP), ainsi que 3 ans supplémentaires au travers de l'extension d'indication en cas de développement du médicament pour une autre indication thérapeutique (9).

Développement préclinique

Les chefs de file sélectionnés subissent par la suite de nombreux tests *in vitro* puis *in vivo* afin de comprendre plus en détail leurs propriétés intrinsèques, leur efficacité et leur activité. Ces études sont primordiales pour envisager son administration chez l'Homme. Son déroulement prend en moyenne 2 ans (2).

Les premiers tests s'intéressent aux propriétés pharmacologiques des chefs de file, notamment leur pharmacocinétique et pharmacodynamie.

La pharmacocinétique, désignée également sous le terme "ADME", s'intéresse au devenir d'une molécule au sein de l'organisme, dès son administration. Celle-ci comprend quatre étapes (10) :

- L'absorption (A), qui étudie le passage du produit à travers les différents obstacles pour arriver dans le sang, ainsi que l'effet du premier passage qui modifie la quantité disponible de médicament dans le sang.

- La distribution (D), qui implique l'étude de la liaison d'une molécule aux protéines plasmatiques, tissulaires, graisses, … donnant ainsi la notion de volume de distribution.
- La métabolisation (M), qui a lieu principalement au niveau du foie, réalise des biotransformations d'une molécule en métabolites pour soit l'éliminer plus facilement en la rendant hydrophile, soit l'activer.
- L'élimination (E), qui s'intéresse aux mécanismes d'élimination d'une molécule, soit sous forme inchangée, soit sous forme d'un ou plusieurs métabolite(s). Ces composés seront principalement éliminés via l'urine ou les fèces, mais également par la sueur, la salive, le lait maternel ou les larmes.

La pharmacodynamie décrit quant à elle les effets de molécules sur l'organisme, dont l'étude détaillée des interactions entre une molécule et ses différentes cibles. Ces études permettent de distinguer les effets propres à la molécule, en mettant en évidence une relation dose-effet et effet-temps, la Dose Efficace 50 (DE50) permettant l'obtention de 50% de l'effet maximal, et la recherche du mécanisme d'action précis de la molécule (11). Elles permettent également l'étude des impacts d'une molécule sur d'autres cibles, notamment avec la détection d'éventuelles interactions pouvant causer des effets secondaires graves. Des tests sont réalisés systématiquement sur certaines cibles reconnues comme étant critiques, comme par exemple le canal potassique hERG (12) et le récepteur sérotoninergique 2B entraînant de graves problèmes cardiaques pouvant conduire à l'arrêt cardiaque en cas de blocage (13), ou le récepteur sérotoninergique 2A entraînant des hallucinations en cas de stimulation (14).

La seconde batterie de tests à réaliser au sein du développement préclinique est l'étude de l'innocuité des chefs de file. Ces tests s'intéressent aux interactions nocives entre une molécule et l'organisme, en déterminant principalement *in vivo* des doses critiques toxiques servant de référence, et en étudiant les impacts à court, moyen et long terme sur l'organisme, dans le but de sécuriser un maximum les premiers tests chez l'Homme lors du développement clinique. De très nombreux tests sont réalisés, mais les tests les plus critiques sont (15):

- La toxicité aiguë : permet de mesurer la Dose Létale 50 (DL50), la dose causant la mort de la moitié d'une population animale précise, dans des conditions spécifiques et pour une durée minimum de 14 jours.
- La toxicité subaiguë : permet d'obtenir des renseignements sur la capacité du produit à s'accumuler au sein des tissus, en réalisant une surveillance clinique et biologique pendant 28 à 90 jours, et par autopsie précise de chaque organe.
- La toxicité chronique : s'intéresse à l'exposition répétée de la molécule sur une longue période, de 6 mois à 1 an. Elle permet d'étudier les modifications morpho-pathologiques et fonctionnelles qui peuvent être induites par la prise à

long terme du médicament, en établissant les conditions d'apparition de ces altérations.

- Le pouvoir tératogène et l'embryotoxicité : permettent de mesurer la toxicité induite par une molécule sur le développement d'un embryon et d'un nouveau-né, sur une ou plusieurs générations.
- Les pouvoirs cancérigènes et mutagènes : permettent de détecter une éventuelle cancérogénicité, en étudiant une population animale pendant plus de 2 ans et en réalisant des tests *in vitro* sur des cellules et tissus mutés.

Il est à noter que l'absence d'effet toxique ne garantit pas la sécurité totale chez l'Homme à cause de la difficulté d'extrapolation des résultats entre les différentes espèces animales et l'Homme, mais permettent toutefois de sécuriser les premiers essais des études cliniques.

A l'issue de cette étape préclinique, si les molécules étudiées semblent avoir un potentiel en thérapeutique tout en étant non toxiques, elles peuvent être sélectionnées afin de passer à l'étape du développement clinique. Les molécules ainsi sélectionnées sont appelées les « candidats médicaments » (2).

Développement clinique

Les candidats médicaments sélectionnés peuvent passer à l'étape du développement clinique, qui se déroulera sur une période de 5 à 8 ans en moyenne. Cette étape vise à la réalisation des démonstrations pharmaceutiques et cliniques (2).

La démonstration pharmaceutique permet le développement de formulations galéniques administrables à l'Homme, déterminant la voie et le mode d'administration, la forme galénique, la dose en principe actif, …Cette démonstration pharmaceutique vise également à fournir des méthodes d'analyse des performances et des spécifications du principe actif et de la formulation, et le développement d'un processus industriel pour la production et le contrôle à grande échelle (2).

La démonstration clinique évalue quant à elle l'efficacité et la tolérance des candidats médicaments au travers de trois phases d'essais cliniques (2).

- La phase I permet d'évaluer la tolérance d'un candidat médicament chez l'Homme, de comprendre son parcours entier au sein de l'organisme au travers de la pharmacocinétique et pharmacodynamie chez l'Homme, d'évaluer ses potentiels effets indésirables, et de déterminer une dose et une fréquence d'administration qui seront recommandés par la suite. Cette phase est réalisée au travers d'essais menés sur un petit nombre de volontaires sains.

- La phase II se concentre sur la démonstration de l'efficacité du candidat médicament ainsi que sur la détermination de la dose optimale. L'objectif est de confirmer l'activité clinique du médicament à la dose recommandée lors de la phase I, au travers d'essais menés sur un petit nombre de patients atteints de la pathologie cible.
- La phase III a pour but la comparaison du candidat médicament au placebo et/ou à un médicament de référence de la pathologie cible (s'il existe), afin de confirmer son efficacité et son innocuité à la dose et à la formulation choisie.

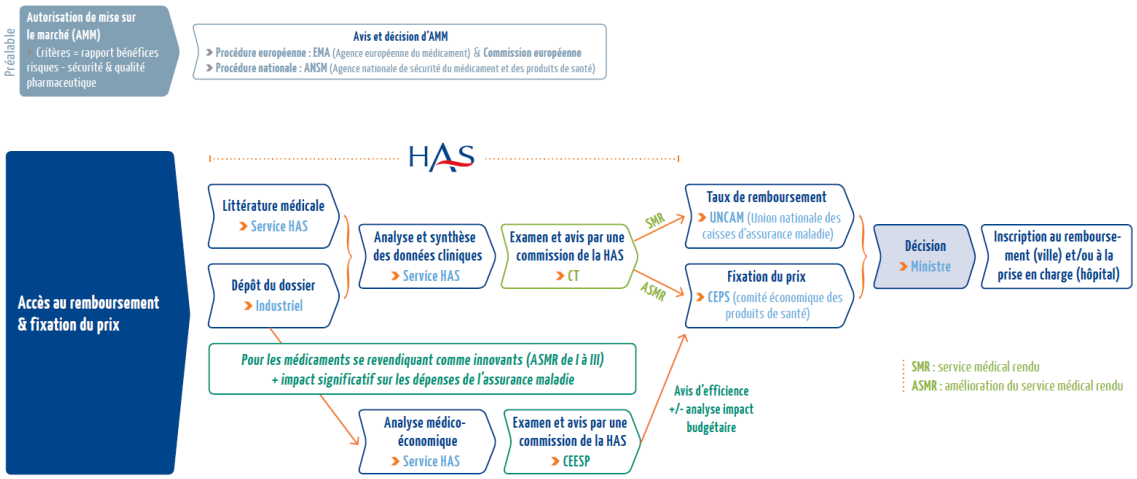


Figure 3 : Circuit d'accès au marché des médicaments en France (16)

Autorisation de Mise sur le Marché

Selon les résultats de la phase III de la démonstration clinique, et après consolidation de toutes données incomplètes, le candidat médicament peut faire l'objet d'une demande d'Autorisation de Mise sur le Marché (AMM) permettant sa commercialisation (17). Cette demande s'effectue auprès d'autorités compétentes, spécifiques pour chaque pays ou organisation de pays. En France, il s'agit de l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM), en Europe de l'Agence Européenne du Médicament (EMA), ou encore aux Etats-Unis de la *Food & Drug Administration* (FDA). L'AMM permet de garantir la qualité, la sécurité et l'efficacité du médicament dans des conditions d'utilisations précises, pour une pathologie précise, au vu uniquement des données scientifiques qui auront été accumulées lors des essais précliniques et cliniques. Cela passe notamment par l'évaluation du rapport bénéfice/risque, où les bénéfices apportés par le candidat médicament surpassent les éventuels risques liés aux effets secondaires.

Prix et remboursement

➔ ***Haute Autorité de Santé***

Après attribution d'une AMM par les autorités compétentes, le candidat médicament doit faire l'objet d'évaluations scientifiques et médico-économiques par la Haute Autorité de Santé (HAS) en France, une autorité administrative indépendante (Figure 3) (16).

La Commission de Transparence de la HAS évalue l'utilité du futur médicament via la détermination de deux critères pour l'attribution du prix et du taux de remboursement du futur médicament (16,18):

- Le Service Médical Rendu (SMR) : permet de prendre en compte la gravité de la pathologie traitée par le candidat médicament, son efficacité et ses effets indésirables, sa place dans l'arsenal thérapeutique en place, et son intérêt pour la santé publique, afin de définir le taux de remboursement. En fonction des résultats, il est catégorisé en SMR insuffisant, en SMR modéré ou faible, ou en SMR majeur ou important.
- L'Amélioration du Service Médical Rendu (ASMR) permet de déterminer le prix d'un médicament remboursable en fonction du progrès thérapeutique amené en comparaison aux alternatives thérapeutiques en place. En fonction du résultat, différents niveaux peuvent être attribués, allant d'ASMR I (amélioration majeure) à ASMR V (amélioration inexistante).

Pour les médicaments possédant une ASMR I à III, ou étant susceptibles d'avoir un impact significatif sur les dépenses de l'Assurance Maladie de part sa haute technicité, la Commission d'Évaluation Économique et de Santé Publique (CEESP) de la HAS réalise une évaluation médico-économique afin de rendre un avis d'efficience. Cet avis sert d'outil d'aide à la décision pour définir un taux de remboursement (16).

Pour les médicaments possédant une AMM mais non éligibles à un remboursement par l'Assurance Maladie de part un SMR insuffisant ou une ASMR V, le laboratoire pharmaceutique peut toutefois commercialiser son médicament à un prix fixé librement.

➔ ***Union Nationale des Caisses d'Assurance Maladie***

Après évaluation d'un SMR et d'une ASMR par la Commission de Transparence de la HAS, et de l'avis d'efficience de la CEESP si nécessaire, l'Union Nationale des Caisses d'Assurance Maladie (Uncam) définit un taux de remboursement allant de 0% à 65% en fonction du SMR qui lui aura été attribué. Sur décision du ministre de la Santé, un médicament pourra être pris en charge à 100% par l'Assurance Maladie si celui-ci est un produit irremplaçable, indispensable ou à un prix élevé (19).

➔ ***Comité Économique des Produits de Santé***

En parallèle de la détermination du taux de remboursement, le Comité Économique des Produits de Santé (CEPS) fixe le prix du futur médicament remboursé (2,18). Ce prix, négocié entre le laboratoire pharmaceutique à l'origine du candidat médicament et le CESP, prend en compte l'ASMR qui aura été attribué, le prix fixé des traitements déjà existants, et les prix fixés à l'étranger pour le même candidat médicament.

➔ ***Publication au Journal Officiel***

La décision finale de l'inscription au remboursement par l'Assurance Maladie se fait par les ministres de la Santé et de la Sécurité sociale, et est publiée au Journal officiel (2).

Procédures alternatives pour les pathologies rares ou innovations majeures

Pour les innovations majeures ou pour certaines pathologies particulièrement rares ou graves, des procédures existent permettant l'accès à des solutions thérapeutiques pouvant apporter des bénéfices majeurs dans la prise en charge des patients atteints.

- L'Autorisation Temporaire d'Utilisation (ATU) peut être attribuée exceptionnellement en France à un candidat médicament avant son AMM destiné à traiter, prévenir ou guérir des maladies graves ou rares, ou des pathologies pour lesquelles il n'existe pas de traitement sur le marché, en présumant leur efficacité et leur sécurité d'emploi en l'état des connaissances scientifiques (2).
- Le statut de médicament orphelin peut être attribué à un candidat médicament permettant de traiter une pathologie classée comme rare, soit touchant moins d'une personne sur 2000. Ces pathologies sont souvent chroniques, et rapidement fatales sans traitement disponible. Le développement d'un candidat médicament classé comme orphelin permet l'obtention au niveau européen d'aides financières et un gain de 2 ans en moyenne sur des études cliniques via des procédures dérogatoires (2).

1.1.3. Limitations

Il est à noter qu'en parallèle, le nombre d'autorisations de mise sur le marché accordées par la *Food & Drug Administration* reste stable dans le temps (Figure 4), et le taux de succès en phase clinique a diminué en deux décennies, passant de 25% de succès à 10%. Les principales raisons à ces échecs sont le manque d'efficacité dans 50% des cas, ainsi que le manque de sécurité dans 25% des cas

(Figure 5) (20). Ces chiffres s'expliquent notamment par l'augmentation des attentes des autorités de santé, l'amélioration des standards de soins modernes, et les stratégies de recherche et développement axées autour de pathologies complexes comme certaines pathologies infectieuses ou tumorales.

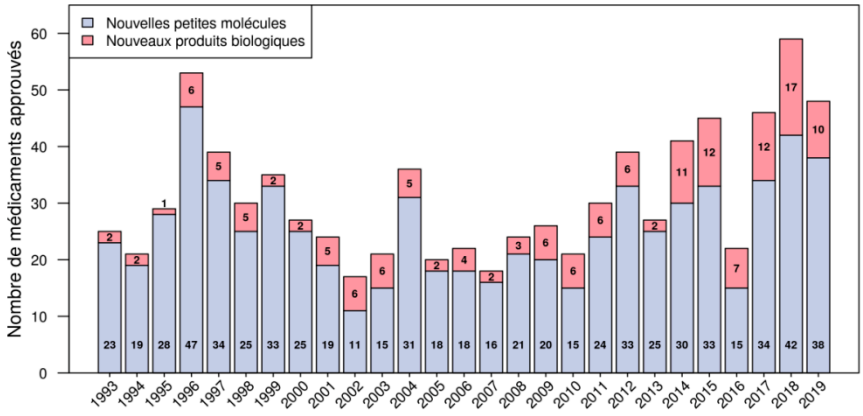


Figure 4 : Nombre d'AMM accordées à de nouvelles petites molécules ou à de nouveaux produits biologiques par l'autorité compétente américaine *Food & Drug Administration* - Adapté de (21).

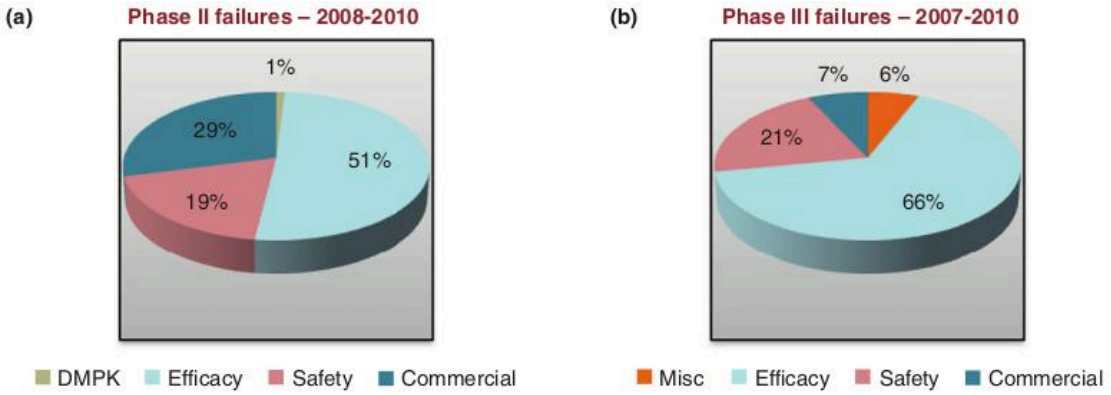


Figure 5 : Raison des échecs en phase II et phase III du développement clinique de 2007 à 2010 (20). DMPK = Métabolisme des médicaments et pharmacocinétique

Tous ces chiffres mettent en évidence l'importance grandissante de l'étape de recherche et développement dans le processus global de conception de médicament, afin de trouver les meilleurs candidats médicaments sûrs et efficaces, d'éviter les échecs en phases précliniques et cliniques, pouvant entraîner la perte sèche de plusieurs millions à milliards de dollars.

1.2. Historique de la conception de médicaments

Bien que très peu de traces permettent d'établir le début de l'utilisation de traitements médicamenteux à des fins médicinales, la plus ancienne preuve écrite d'une préparation médicinale remonte à 2100 av. J.-C. sur une tablette d'argile sumérienne, décrivant de nombreuses recettes sans indication d'utilisation (22). Pendant plusieurs milliers d'années, jusqu'à l'apparition des prémices de techniques et méthodes scientifiques, les remèdes étaient découverts par l'observation empirique des effets produits par certaines substances sur certaines maladies. Le début du XIXème siècle accélère la découverte de nouveaux médicaments avec l'apparition de la médecine moderne et de la chimie. La recherche était basée sur une approche fonctionnelle, où les molécules identifiées devaient démontrer leur effet sur une grande fonction biologique de l'organisme. Enfin, avec l'explosion des connaissances scientifiques et médicales dans les années 1970, le développement des médicaments s'est concentré sur une approche moléculaire, où la démonstration de l'activité d'un médicament se réalise à l'échelle de la molécule (22). Aujourd'hui, de nombreuses approches modernes se sont développées en parallèle, répondant chacune à des problématiques spécifiques.

1.2.1. Approche empirique des médecines antiques

Dans les temps anciens, les médicaments n'avaient pas pour utilité la guérison physique, mais étaient associés à des rites religieux, et administrés par des chefs religieux, à des fins de guérison spirituelle. Les médicaments et traitements populaires prenaient leur source principalement dans les produits végétaux, mélangés à des matières animales ou minérales, ou plus rarement des matières humaines. Ces traitements étaient très probablement découverts par une combinaison d'essais, d'erreurs, et d'observations des réactions humaines et animales causées par l'administration de ceux-ci (22).

La médecine traditionnelle indienne, appelée l'Ayurveda, remonterait à plus de 5000 ans (22,23). Les remèdes utilisés, transmis oralement pendant des milliers d'années, ont été répertoriés dans des textes sacrés, et principalement basés sur l'utilisation de formulations de plantes. La médecine traditionnelle chinoise, prenant son origine aux alentours de 3500 av. J.-C., axe sa pratique autour d'une pharmacopée dense, de massages et d'acupuncture, décrits dans de nombreux livres sacrés, commentés pendant plusieurs millénaires afin de compléter les connaissances. La médecine égyptienne a fourni des écrits au travers de papyrus antiques, notamment via le papyrus Ebers, daté entre le XVIème et le XVème siècle av. J.-C. et rendu public par G.M. Ebers en 1875. Ce papyrus est le plus ancien traité médical connu. Il recense 877 prescriptions classées en 33 groupes d'affections, en précisant le nom des substances utilisées, leurs quantités et leur mode d'administration (22,24).

Les médecines grecques et romaines se basent sur les connaissances des civilisations précédentes. Ce sont les grecs qui vont définir une maladie non pas comme étant lié à une cause surnaturelle ou magique, mais le résultat de causes naturelles. C'est également à cette époque qu'Hippocrate écrira des textes définissant l'éthique médicale, encore utilisés de nos jours comme fondation de la déontologie médicale (22).

Bien que les traitements populaires soient originaires de civilisations différentes, il existe un grand nombre de similarités, notamment dans l'utilisation des mêmes plantes pour soigner des pathologies similaires. Il existe probablement une contribution des anciens marchands, qui ont pu aider à la propagation des connaissances médicales avec leurs voyages. Les traitements traditionnels, découverts sur une période de plusieurs millénaires, ont été pendant très longtemps les seuls traitements disponibles jusqu'à l'apparition des prémices d'approches modernes de découverte de médicaments au XIXème siècle (22).

Toutefois, ces traitements ont encore une place importante, aussi bien par la perpétuation de la pratique de certaines médecines traditionnelles, qu'en tant que source d'inspiration et points de départ pour des processus modernes de découverte de médicaments. La médecine chinoise a notamment permis la découverte de la réserpine et de l'éphédrine, et la médecine égyptienne celle des dérivés du pavot à opium et du saule (22).

1.2.2. Débuts de la chimie médicinale et de la médecine moderne

Le début du XIXème siècle marque un tournant dans les processus de découverte de médicaments, avec l'apparition des premières notions de chimie organique moderne et l'émergence de la chimie analytique, à la base de la chimie médicinale (25). A partir des millénaires de connaissances sur l'utilisation de produits naturels, des composés ont pu être identifiés et isolés, notamment les alcaloïdes. Le premier alcaloïde découvert et isolé a été la morphine en 1805 par F. W. Sertüner, à partir du pavot à opium, puis ont été isolés la quinine en 1823 et l'atropine en 1834 (26,27).

De grandes avancées en chimie ont pu également être réalisées au cours du XIXème siècle, notamment sur la notion d'aromaticité par A. Kekulé en 1865, la classification périodique des éléments par D. Mendeleïev en 1869 confirmant la théorie atomique d'A. Avogadro postulée en 1811, et la découverte de l'électron par J. J. Thomson en 1897 (25,27). Ces nouvelles connaissances ont permis l'élaboration de la théorie des récepteurs en 1905 par P. Ehrlich et J. N. Langley, où

des substances réceptrices vont se lier par liaison chimique avec d'autres substances ou molécules biologiquement actives afin de générer des réponses, entraînant par conséquent des effets biologiques et/ou pharmacologiques (27,28). Cette découverte a conduit au développement de nouveaux concepts, comme la présence d'interactions entre une molécule et son récepteur, ainsi que l'existence de pharmacophore, décrivant les caractéristiques nécessaires pour qu'un médicament puisse se lier à son récepteur et réaliser son action biologique (28).

En 1928, la découverte de la pénicilline comme premier antibiotique par A. Fleming révolutionne le traitement des maladies infectieuses. Celle-ci sera notamment produite en série et utilisée lors de la Seconde Guerre Mondiale pour contrôler les septicémies (22,25). L'élucidation de sa structure en 1942 conduira à l'ère des antibiotiques, permettant de soigner de nombreuses maladies infectieuses auparavant incurables, sauvant des millions de vies et contribuant au bond de l'espérance de vie moyenne au XXème siècle (25). Toutefois, le phénomène de résistance bactérienne a été mis en évidence en parallèle, avec l'apparition de souches bactériennes très résistantes à la pénicilline, produisant une protéine dégradant le médicament (25). Depuis, de très nombreuses familles d'antibiotiques ont été découvertes, ainsi que l'apparition de résistances multiples.

A partir du milieu du XXème, l'étude de la structure tridimensionnelle de la double hélice d'ADN par J. Watson, M. Wilkins et F. Crick en 1953 (29), ainsi que le développement de la technologie de l'ADN recombinant ont révolutionné les connaissances en biologie cellulaire et moléculaire (25). Cela a permis à terme le développement de protéines recombinantes, notamment le développement de l'insuline humaine, qui deviendra le premier médicament biologique autorisé par la FDA en 1982 (25).

1.2.3. Approches modernes de découverte de médicaments

A partir du début des années 1990, les technologies modernes ont vu le jour, avec l'apparition de la chimie computationnelle et de la modélisation moléculaire, ainsi que la chimie combinatoire et le criblage à haut débit. Ces technologies ont permis de révolutionner la découverte et le développement de médicaments.

Le criblage à haut débit, basé sur l'utilisation des connaissances en biochimie, en pharmacologie et en robotique, permet d'identifier de nouvelles molécules bioactives à partir d'une grande collection de molécules. Cette technique se base sur une approche empirique, où toutes les molécules testées seront mises en présence d'une cible thérapeutique afin d'observer la présence ou non d'interactions et d'activité biologique (30).

La chimie combinatoire, basée sur les avancées en protéomique, permet de synthétiser un très grand nombre de composés chimiques par combinaison de plusieurs autres composés sélectionnés. Si un effet sur l'activité biologique est mesuré en utilisant ce mélange de composés synthétisés sur une cible, des processus de déconvolution seront mis en place afin d'identifier la ou les molécules responsables de l'effet observé (31).

La chémoinformatique, également appelé chimie computationnelle, basée sur les évolutions de l'informatique et de la biologie structurale, permet le développement de méthodes *in silico* dans le but de modéliser et de simuler le comportement de molécules biologiques, et d'accélérer la découverte de nouveaux médicaments (32). La chémoinformatique est à la base de l'approche rationnelle, visant à l'obtention de données détaillées sur les étapes biochimiques impliquées, comprenant notamment l'étude des interactions entre un composé chimique et sa cible entraînant un effet thérapeutique. L'expression « *in silico* » signifie « réalisé sur un ordinateur ou via une simulation par ordinateur ». Celle-ci fait référence à l'expression latine *in silicon*, comparable aux expressions latines *in vivo*, *in vitro* et *in situ* communément utilisées en biologie, pouvant être traduite par « dans le silicium », faisant allusion à l'utilisation massive du silicium dans la production des puces informatiques (33).

1.2.4. Histoire de la chémoinformatique

L'histoire de la chémoinformatique et des méthodes *in silico* en général prennent leur source dans l'une des plus grandes inventions du 20ème siècle : l'ordinateur. En 1936, Alan Turing publia dans un article intitulé « *On Computable Numbers with an Application to the Entscheidungsproblem* » les éléments fondateurs de l'informatique permettant la construction de « machines de Turing », fusionnant les concepts d'ordinateur, de langage de programmation et de programme informatique (34). La première application de techniques informatiques pour des problématiques liées à la chimie remonte à 1946, afin d'analyser des spectres en chromatographie (35). Ces premiers travaux de chimie assistés par l'informatique ont donné naissance aux prémices de la chémoinformatique (36).

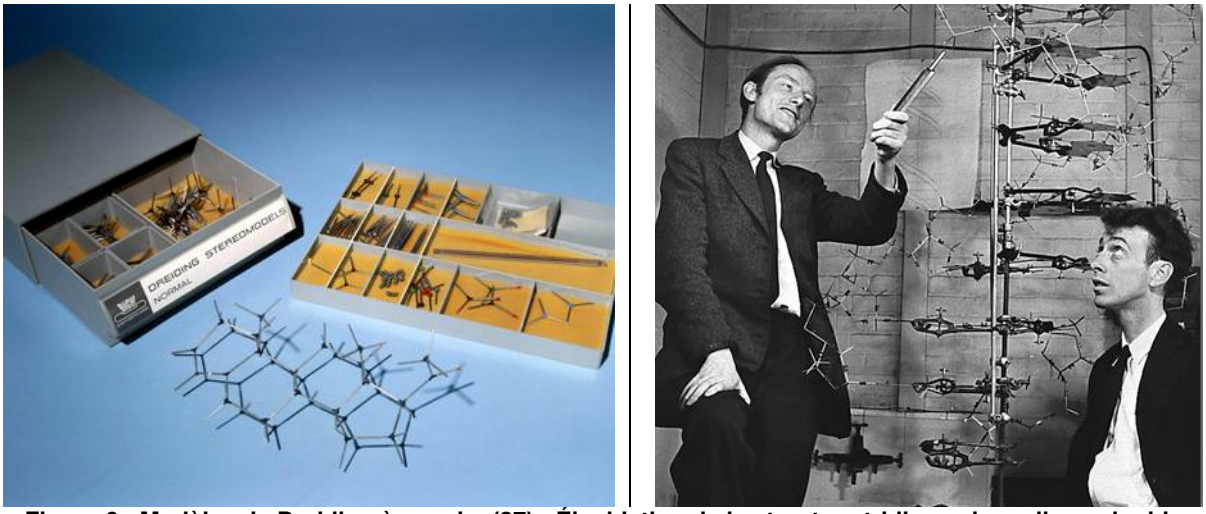


Figure 6 : Modèles de Dreiding à gauche (37) ; Élucidation de la structure tridimensionnelle en double hélice de l'ADN par Watson & Crick en 1962 à droite (38)

En parallèle, la biologie structurale s'est également développée très rapidement, avec l'application de la cristallographie à l'analyse de structures en 1913 par Henry et Lawrence Bragg, récompensés par un prix Nobel en physique en 1915 (39). L'invention de modèles moléculaires squelettiques par André Dreiding en 1958 a permis la représentation de composés chimiques à l'aide de tubes en acier inoxydables correspondant aux liaisons et de petites sphères correspondant aux atomes (Figure 6 - gauche) (40). L'application de ces modèles moléculaires a contribué à la découverte de la structure tridimensionnelle de l'ADN par Francis Crick et James Watson en 1953, récompensés en 1962 par un prix Nobel en médecine (Figure 6 - droite) (29).

Avec le développement des premiers ordinateurs personnels, de la miniaturisation et l'évolution des puissances de calculs, l'informatique est devenue à la portée de nombreux domaines de recherche scientifique. Cette évolution de la puissance de calcul est décrite par la loi de Moore en 1965, qui énonce que « le nombre de transistors par circuit de même taille double, à prix constant, tous les deux ans », entraînant par conséquence une évolution exponentielle de la puissance de calcul (Figure 7) (41).

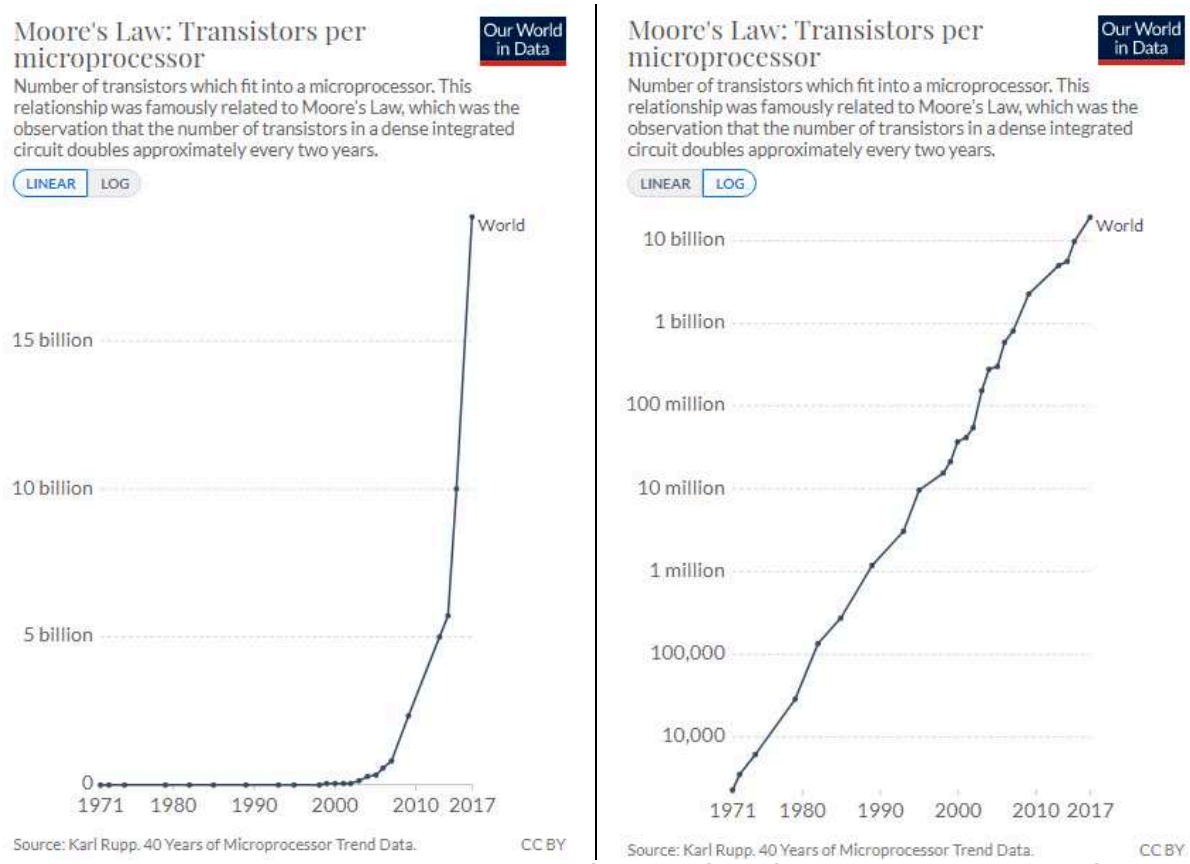


Figure 7 : Illustration de la Loi de Moore dans une échelle linéaire (à gauche) et logarithmique (à droite) (42,43)

En 1976 apparaît le terme « bioinformatique » dans le nom d'un groupe de recherche de l'Université d'Utrecht, et le premier article accessible faisant référence à la bioinformatique a été publié en 1978 par ce même groupe (44). La bioinformatique fut définie comme « l'étude des processus d'information dans les systèmes biotiques » (44,45). Avec l'explosion des données génomiques, notamment avec le projet génome humain lancé en 1988 pour séquencer entièrement l'ADN du génome humain (46), la bioinformatique est devenue une discipline populaire.

Grâce à ces avancées, la « biologie *in silico* » a permis le développement de son homologue, la « chimie *in silico* » (36). Le terme « chémoinformatique » a été pour la première fois utilisé en 1998 par Frank Brown, le définissant comme « l'utilisation de la technologie et de la gestion de l'information comme élément essentiel du processus de découverte de médicaments, en transformant les données chimiques en informations permettant de prendre des décisions plus efficaces et plus rapides dans le domaine de l'identification et du développement de composés chefs de file » (47). Cette discipline a ensuite été démocratisée par le livre de Johann Gasteiger publié en 2003, « Handbook of Chemoinformatics » (48), qui a permis à la chémoinformatique d'être la norme pour l'utilisation de l'informatique en chimie (36).

1.3. Chémoinformatique

1.3.1. Définition

La chémoinformatique est définie comme « toutes les ressources en matière d'information dont un scientifique a besoin pour optimiser les propriétés d'un ligand destiné à devenir un médicament » (47). Située à l'interface de la chimie, la biologie, la biochimie, la physique, des mathématiques et des statistiques en utilisant des outils informatiques, le but principal de la chémoinformatique est d'analyser, simuler, modéliser, visualiser et manipuler des données chimiques (49).

Depuis ses débuts, la chémoinformatique a permis le développement d'outils sophistiqués, utilisés en routine aussi bien dans le milieu académique qu'en industrie pharmaceutique, dans le but notamment de concevoir les médicaments de demain (50). Ces méthodes possèdent trois avantages clés : leur facilité de prise en main et de mise en place, leurs faibles coûts, et leur rapidité d'exécution avec des capacités de calculs suffisantes (51). L'utilisation de la chémoinformatique permet de gagner de 2 à 3 ans sur le temps de conception global et d'économiser près de 200 millions de dollars, équivalent à un retour sur investissement de 3 à 9 dollars pour 1 dollar investi dans ce domaine (52).

Les prédictions réalisées seront confirmées à l'aide de tests expérimentaux ciblés. La chémoinformatique et les méthodes *in silico* s'imposent comme le 3ème pilier de la recherche, en renforçant et accélérant les méthodes traditionnelles *in vivo* et *in vitro*. Son importance grandissante est reconnue par les grandes autorités de santé, notamment par l'EMA dans son rapport stratégique à horizon 2025, promouvant l'utilisation de méthodes *in silico* pour la réduction de l'expérimentation animale et permettant de proposer à terme une alternative aux essais cliniques traditionnels (53).

L'application à la conception de médicaments du vaste spectre de techniques utilisables en chémoinformatique permet (48):

- La conception, la gestion et l'analyse de bases de données chimiques massives (*big data)*, ainsi que leur exploration (*data mining*).
- L'identification de composés pouvant se lier de manière sélective à une cible pharmacologique, connaissant sa structure tridimensionnelle, au travers de techniques comme l'amarrage moléculaire ou le criblage virtuel à haut débit.
- La modélisation moléculaire, afin de modéliser et de mimer le comportement de molécules et de systèmes moléculaires, allant de la petite molécule aux grands complexes protéiques, notamment via l'utilisation de techniques dérivées de la dynamique moléculaire (54).
- La création de modèles de prédiction de propriétés moléculaires, d'activités biologiques ou de toxicité de nouvelles molécules, ainsi que l'identification de

composés similaires à des petites molécules biologiquement actives, notamment au travers de relations quantitatives structure-activité (QSAR), de relations quantitatives structure-propriété (QSPR), ou de la création et l'utilisation de pharmacophores.

En fonction des connaissances disponibles, deux groupes d'approches sont possibles. Le premier groupe d'approches est celui nécessitant la connaissance de la structure de la cible pharmacologique, regroupant notamment l'amarrage moléculaire, le criblage virtuel à haut débit et la dynamique moléculaire.

Le second groupe d'approches est celui nécessitant la connaissance de structures chimiques de petites molécules testées, utilisées dans l'analyse quantitative structure-activité ou la création de pharmacophores.

1.3.2. Différences entre chémoinformatique et bioinformatique

La chémoinformatique est à différencier de la bioinformatique, autre domaine appliquant des méthodes *in silico* pour répondre à des problèmes posés par la biologie, notamment liées à l'analyse et l'interprétation de données biologiques, comme l'étude du génome, la biologie structurale, la protéomique ou la métabolomique. Ce domaine est utilisable en conception de médicaments en permettant de rechercher et de valider une cible thérapeutique dans une pathologie d'intérêt.

La bioinformatique se focalise sur le traitement informatique de bases de données de séquences biologiques, alors que la chémoinformatique va se focaliser sur le traitement informatique de bases de données de structures chimiques, en 2D ou en 3D (55).

1.3.3. Application de la chémoinformatique dans le processus de conception de médicaments

La chémoinformatique permet deux types d'approches, en fonction des connaissances disponibles sur le sujet : basées sur les ligands ou basées sur les structures (Figure 8) (56).

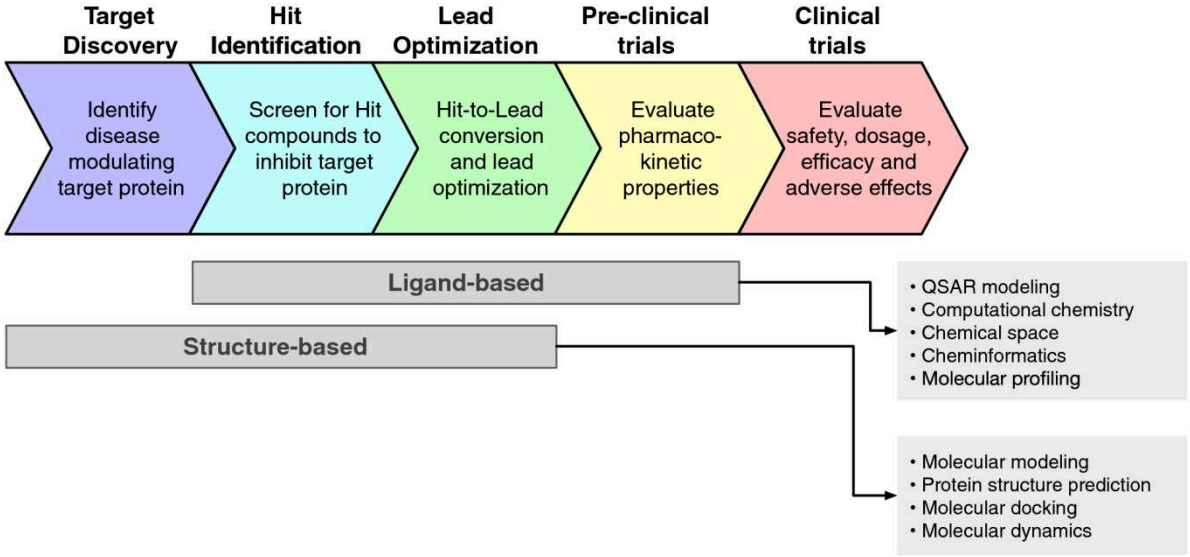


Figure 8 : Résumé schématique du processus de conception de médicaments avec en parallèle les approches *in silico* correspondantes – Adapté de (57)

L'approche basée sur les ligands (*ligand-based*) utilise les connaissances des ligands naturels connus de la cible, comme ses agonistes, antagonistes, … En corrélant leurs activités biologiques avec leur structure, il est possible de prédire l'activité, la toxicité ou des paramètres physico-chimiques ou pharmacocinétiques de nouveaux composés, ou d'établir des règles permettant l'amélioration des molécules testées pour les rendre plus actives, moins toxiques, et/ou ayant des paramètres optimisés.

L'approche basée sur les structures (*structure-based*) nécessite la connaissance de la structure tridimensionnelle de la cible obtenue via des techniques de biologie structurale, notamment par cristallographie ou par spectroscopie RMN. Si cette structure n'est pas disponible, il est possible de créer un modèle de structure à partir d'une structure expérimentale d'une protéine de la même famille. A partir de la structure de la cible, il est imaginable de concevoir des composés chimiques capables de se fixer sélectivement à la cible en étudiant les interactions réalisées au sein d'un site de liaison.

La chémoinformatique nécessite une approche rationnelle, basée sur la connaissance de données détaillées du mécanisme pathologique et des processus associés. Il est donc important de connaître les molécules biologiques impliquées, permettant de faire le lien avec une pathologie, ainsi que la cible pharmacologique impliquée pour obtenir un effet thérapeutique. Cette cible est identifiée et validée par des approches bioinformatiques, ou par des techniques de chémoinformatique basées sur les structures, permettant l'utilisation d'une petite molécule biologique ayant un impact dans une pathologie comme sonde pour un criblage de nombreuses cibles pharmacologiques.

Ensuite, si la structure cristallographique de cette cible est connue, des touches pharmacologiques peuvent être mises en évidence via la réalisation de campagnes de criblage virtuel à haut débit par amarrage moléculaire, en criblant une ou plusieurs chimiothèques généralistes. Si des molécules de référence sont connues, notamment au travers de structures cristallographies, des approches s'appuyant sur le ligand sont également possibles, notamment avec l'application de méthodes pharmacophoriques 2D ou 3D pour le criblage de chimiothèques.
A la différence d'une étape de recherche de touches pharmacologiques classiques, où entre 10.000 et 50.000 molécules sont testées *in vitro*, les méthodes chémoinformatiques permettent le test de plusieurs millions de molécules pour une même cible, entraînant une augmentation des taux de succès et réduisant le besoin d'expériences *in vitro* coûteuses (58).

Le passage des touches pharmacologiques aux composés chefs de file met également en jeu de nouvelles campagnes de criblage virtuel par amarrage moléculaire en se concentrant sur les modulations des touches pharmacologiques, ainsi que l'application de méthodes basées sur la dynamique moléculaire afin d'observer la dynamique des interactions réalisées entre la cible et les composés les plus prometteurs pour le choix de candidats médicaments. Les approches basées sur les ligands permettent la prédiction d'activités biologiques ainsi que des paramètres physico-chimiques au travers de l'étude des relations structure-activité (QSAR) et structure-propriété (QSPR).

Enfin, lors des tests précliniques, les candidats médicaments sélectionnés vont principalement faire l'objet de création de modèles de prédiction de propriétés pharmacocinétiques et toxicologiques. L'utilisation de l'amarrage moléculaire peut également être envisagée pour observer si les candidats médicaments sélectionnés ne se lient pas à des récepteurs pouvant causer de graves effets secondaires, notamment avec le canal potassique hERG pouvant causer des problèmes cardiaques.

2. Notions indispensables

2.1. Représentations des structures chimiques

La représentation des structures chimiques est un des concepts clés de la chémoinformatique afin d'obtenir leur bonne reconnaissance par tous les outils informatiques.

Les structures chimiques sont plus communément représentées à l'aide de formules topologiques, de formules brutes, de noms très différents, ou de différents numéros de registres selon les banques de données. Par exemple, l'ibuprofène possède énormément de noms différents ou d'identifiants, décrivant exactement la même molécule (Figure 9). Afin d'apporter une standardisation, l'Union Internationale de Chimie Pure et Appliquée (IUPAC - *International Union of Pure and Applied Chemistry*) a développé une dénomination standardisée pour les structures chimiques, pour fournir un nom systématique univoque pour chaque molécule.

A partir de ces noms et identifiants, ou de la formule topologique, il est possible d'extraire informatiquement toutes les informations d'une molécule en utilisant des bases de données en libre accès. Toutefois, il existe des limites en fonction du nom utilisé, ou de la structure dessinée, principalement au niveau de la stéréochimie.

En raison du manque de standardisation et de précision dans les représentations facilement écrites et comprises par un humain, des représentations spécifiques ont été développées, comme les notations linéaires ou les tables de connexions.

alpha-Methyl-4-(2-methylpropyl)benzeneacetic Acid	Ibuprofen, Copper (2+) Salt
Brufen	Ibuprofen, Magnesium Salt
Ibumetin	Ibuprofen, Potassium Salt
Ibuprofen	Ibuprofen, Sodium Salt
Ibuprofen Zinc	Ibuprofen, Zinc Salt
Ibuprofen, (+-)-Isomer	Ibuprofen-Zinc
Ibuprofen, (R)-Isomer	Motrin
Ibuprofen, (S)-Isomer	Nuprin
Ibuprofen, Aluminum Salt	Rufen
Ibuprofen, Calcium Salt	Salprofen

Figure 9 : Exemple de dénominations inscrites dans le thésaurus médical MeSH (*Medical Subject Headings*) pour l'Ibuprofène (59)

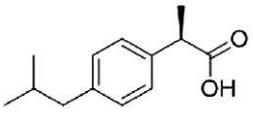
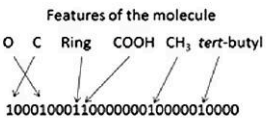
1D-structure	2D-structure	3D-structure
Ibuprofen $C_{13}H_{18}O_2$ SMILES code: c1c(ccc(c1)[C@@H](C(=O)O)C)CC(C)C InChI code: 1S/C13H18O2/c1-9(2)8-11-4-6-12(7-5-11)10(3)13(14)15/h4-7,9-10H,8H2,1-3H3,(H,14,15)	 Molecular fingerprint: 	 Shape: 

Figure 10 : Représentation de l'ibuprofène sous différentes notations (60)

2.1.1. Notations linéaires

Les notations linéaires sont basées sur des méthodes de représentation compactes, pouvant être écrites très simplement sur une seule ligne, mais ont tendance à manquer de précision sur l'état de protonation et la stéréochimie.

Parmi ces notations linéaires, on retrouve l'InChI, ou Identifiant Chimique International (*International Chemical Identifier*) développé par l'IUPAC, permettant un encodage des informations moléculaires (61). Cet identifiant est composé d'une succession informations, comprenant au minimum la formule chimique, la charge et la stéréochimie (Figure 10).
L'avantage principal de cette notation est la possibilité de réaliser une recherche générique, afin de trouver des identifiants qui ne correspondent qu'à certaines couches d'informations, comme les différents états possibles d'oxydation et de protonation d'une molécule.

On retrouve également parmi ces notations linéaires la notation SMILES (*Simplified Molecular Input Line Entry Specification*), créée en 1988 permettant de décrire une structure chimique sous la forme d'une chaine de caractères (Figure 10) (62). Les chaînes SMILES ont pour avantages d'être plus compréhensible par une personne physique que l'InChI, et sont prises en charge par la très grande majorité des logiciels informatiques.

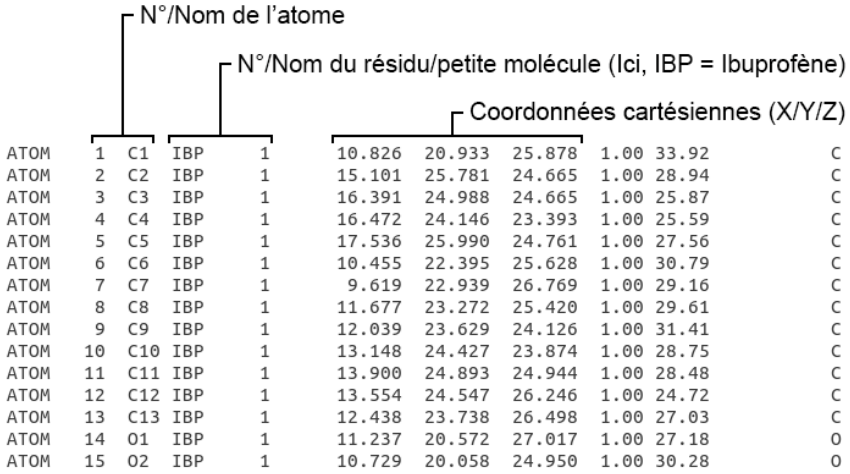
2.1.2. Tables de connexions

Les tables de connexions sont des formats spécifiques de représentations de structures chimiques permettant la reconnaissance de structures bidimensionnelles et tridimensionnelles par un système informatique.

La table de connexion la plus simple possible est composée de la liste de tous les atomes de la molécule, et de la liste de toutes les liaisons présentes sous la

forme de paires d'atomes liés. Des informations plus précises peuvent être présentes, comme l'état d'hybridation de chaque atome, leur charge ou leurs coordonnées. Le renseignement de ces coordonnées sous la forme cartésienne permet d'obtenir une représentation bidimensionnelle (coordonnées X/Y) ou tridimensionnelle (coordonnées X/Y/Z), qui sera interprété informatiquement et visualisable à l'aide de logiciels.
Il est à noter que les hydrogènes ne sont pas nécessairement renseignés explicitement, leurs positions pouvant être recalculées en fonction de l'état d'hybridation et de la charge des atomes.

Parmi les formats composés de tables de connexions, l'un des plus utilisés est le format *Protein Data Bank*, possédant l'extension « .pdb », aussi bien pour des structures chimiques que pour des structures plus complexes. Il inclut entre autres le numéro, le nom et le type de l'atome codé en fonction de son état d'hybridation, le numéro et le nom de l'acide aminé et les coordonnées cartésiennes (Figure 11). D'autres formats de tables de connexion sont couramment utilisés en chémoinformatique, comme les formats *Standard Data File* (.sdf) ou Sybyl Mol2 (.mol2).




```
ATOM    1  C1  IBP     1      10.826  20.933  25.878  1.00 33.92           C
ATOM    2  C2  IBP     1      15.101  25.781  24.665  1.00 28.94           C
ATOM    3  C3  IBP     1      16.391  24.988  24.665  1.00 25.87           C
ATOM    4  C4  IBP     1      16.472  24.146  23.393  1.00 25.59           C
ATOM    5  C5  IBP     1      17.536  25.990  24.761  1.00 27.56           C
ATOM    6  C6  IBP     1      10.455  22.395  25.628  1.00 30.79           C
ATOM    7  C7  IBP     1       9.619  22.939  26.769  1.00 29.16           C
ATOM    8  C8  IBP     1      11.677  23.272  25.420  1.00 29.61           C
ATOM    9  C9  IBP     1      12.039  23.629  24.126  1.00 31.41           C
ATOM   10  C10 IBP     1      13.148  24.427  23.874  1.00 28.75           C
ATOM   11  C11 IBP     1      13.900  24.893  24.944  1.00 28.48           C
ATOM   12  C12 IBP     1      13.554  24.547  26.246  1.00 24.72           C
ATOM   13  C13 IBP     1      12.438  23.738  26.498  1.00 27.03           C
ATOM   14  O1  IBP     1      11.237  20.572  27.017  1.00 27.18           O
ATOM   15  O2  IBP     1      10.729  20.058  24.950  1.00 30.28           O
```

Figure 11 : Exemple de table de connexion au format PDB pour l'Ibuprofène (Identifiant PDB : 4PH9) (63)

2.2. Descripteurs moléculaires

Un descripteur moléculaire est une information déterminée à partir d'un composé chimique (Figure 12 – gauche) (64,65). Ces informations sont obtenues à partir de mesures expérimentales, ou calculées à partir de la structure du composé. Les descripteurs moléculaires ont un rôle clé dans les études des relations quantitatives structure-activité (QSAR) et structure-propriété (QSPR), où ils sont utilisés pour établir une relation mathématique entre la structure et une variable à prédire.

Un très grand nombre de descripteurs moléculaires différents existent, se différenciant par leur complexité ou leurs règles de conception : plus de 9000 descripteurs moléculaires différents ont été répertoriés en 2009 (66). Tous ces descripteurs sont calculables à partir de nombreux logiciels ou de librairies de chémoinformatique, notamment Dragon (67), RDKIT (68) ou CDK (69).
Les descripteurs moléculaires sont classés en plusieurs catégories, en fonction de la représentation du composé utilisé.

2.2.1. Descripteurs unidimensionnels (1D)

Les descripteurs 1D sont les descripteurs moléculaires les plus simples, déterminés à partir de la formule brute de la molécule. Cela inclut le nombre et le type des atomes de la molécule, le poids moléculaire, la masse molaire, ...

2.2.2. Descripteurs bidimensionnels (2D)

Les descripteurs 2D sont déterminés à partir de la structure plane de la molécule. Ils sont plus complexes que les descripteurs 1D car ils représentent des informations sur la taille, la forme et la distribution électronique de la molécule. Il existe deux catégories de descripteurs 2D :

- Les descripteurs constitutionnels comprenant le nombre total et le type de liaisons, le nombre et le type de cycles aromatiques ou non aromatiques, etc…
- Les descripteurs topologiques obtenus à partir des connexions interatomiques de la molécule afin de calculer une série de chiffres binaires qui représentent la présence (codée 1) ou l'absence (codée 0) de sous-structures particulières au sein de la molécule, appelées les empreintes moléculaires (Figure 12 – droite).

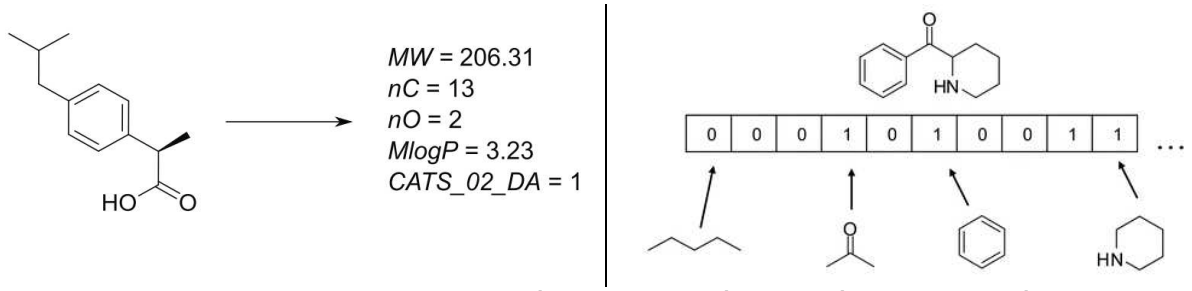


Figure 12 : Exemple de descripteurs moléculaires (Adapté de (70)) à gauche ; Représentation schématique d'empreintes moléculaires (71) à droite.
MW = Masse moléculaire ; nC = Nombre d'atomes de carbone ;
nO = Nombre d'atomes d'oxygène ; MlogP = Coefficient de partage prédit ;
CATS_02_DA = Nombre de paires d'accepteur-donneur de liaison hydrogènes séparées par deux liaisons

2.2.3. Descripteurs tridimensionnels (3D)

Les descripteurs 3D sont déterminés à partir des coordonnées dans l'espace de tous les atomes composant la molécule. Ce sont des descripteurs très complexes, nécessitant une puissance de calcul élevée afin de les obtenir, mais ils apportent des informations nécessaires pour la modélisation d'activités ou de propriétés qui dépendent directement de la structure 3D de la molécule. Il existe quatre catégories de descripteurs 3D :

- Les descripteurs électroniques, ou quantiques, comme l'électronégativité, le moment dipolaire ou certaines énergies.
- Les descripteurs géométriques comme le nombre de liaisons rotatives ou le volume moléculaire.
- Les descripteurs thermodynamiques se basant sur les lois de la thermodynamique, comme le point de fusion, la température thermodynamique ou la pression critique.
- Les descripteurs physico-chimiques comme le nombre de donneurs ou accepteurs de liaisons hydrogène ou la réfractivité.

2.2.4. Espace chimique et similarité moléculaire

En se basant sur les propriétés physico-chimiques qu'un composé chimique devrait posséder pour être un médicament idéal administré par voie orale, il est possible d'énumérer plus de 10^{60} composés. L'ensemble de ces petites molécules est appelé un espace chimique (72,73). Actuellement, seul une fraction de cet espace chimique n'a été étudiée. Pour exemple, la plus grande base de données virtuelle de composés testés expérimentalement, PubChem, recense en 2020 environ 100 millions de composés chimiques, soit 10^8 composés (74).

Etant donné la taille gigantesque de l'espace chimique, il est nécessaire de se focaliser sur une toute petite partie de cet espace. Un des moyens d'y parvenir est la recherche de composés par similarité moléculaire. La similarité entre deux petites molécules peut être obtenue à partir de descripteurs moléculaires via le calcul d'un indice de similarité, exprimant un pourcentage situé entre 0 (totalement différent) et 1 (identique) (32).

A partir d'une base de composés testés expérimentalement et/ou générés virtuellement, une recherche basée sur la similarité permet d'identifier des molécules similaires en comparant leurs empreintes moléculaires à celle d'un composé de référence. Cette recherche permet l'obtention d'un nombre prédéterminé de composés, ou de tous les composés possédant un indice de similarité supérieur à un seul défini au préalable.

La méthode de similarité la plus couramment utilisée est celle basée sur les empreintes moléculaires, un descripteur 2D topologique vu précédemment. A partir de deux séries de chiffres binaires correspondant à l'absence (codé 0) ou à la présence (codé 1) de sous-structures spécifiques pour deux composés sélectionnés, il est possible de calculer un indice de similarité, exprimant le pourcentage de similarité entre ces deux molécules, via une fonction mathématique de distance (32).

Dans le cas de la similarité moléculaire, la formule la plus utilisée est le coefficient de Tanimoto. Celui-ci estime la similarité par rapport au nombre de sous-structures spécifiques présentes chez la première molécule (Figure 13 – variable a), chez la seconde molécule (Figure 13 – variable b), ainsi que le nombre de sous-structures en commun (Figure 13 – variable c) (32).

D'autres manières d'estimer une similarité moléculaire existent, avec l'utilisation d'autres descripteurs moléculaires (propriétés physico-chimiques, descripteurs 3D, etc…) ou avec l'utilisation d'autres formules mathématiques (distance euclidienne, distance de Manhattan, etc…) (75).

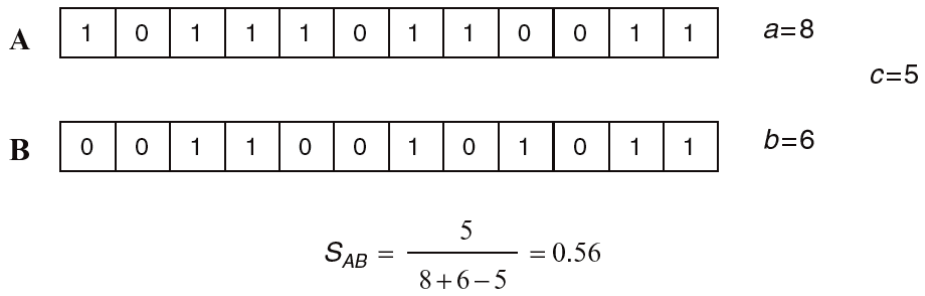


Figure 13 : Exemple de calcul d'une similarité moléculaire par le coefficient de Tanimoto S_{AB}, à partir de deux empreintes moléculaires A et B (32)

2.3. Représentations des protéines

2.3.1. Protéines, peptides et acides aminés

Les protéines sont des macromolécules biologiques constituées par l'association d'acides aminés unis entre eux par une liaison peptidique, assurant des fonctions très diverses au sein des cellules et des tissus (76). Par convention, une protéine composée de moins de 50 acides aminés est appelée peptide.

Un « acide aminé », aussi appelé « résidu », est une molécule organique comprenant un groupement amine, un groupement acide, et une chaine latérale variable en fonction de l'acide aminé en question, tous portés par un carbone asymétrique appelé le « carbone alpha » (77). L'association du groupement acide d'un premier acide aminé avec le groupement amine d'un second acide aminé

permet la formation d'une liaison peptidique, composant la chaine carbonée principale de la protéine. Chez l'Homme, toutes les protéines sont composées de 20 acides aminés dits « standards », tous différenciés par leur chaîne latérale, qui donnent leur nom à l'acide aminé. Il existe des codes à une lettre et à trois lettres pour chaque acide aminé, permettant de les identifier.

2.3.2. Structure primaire

La succession linéaire d'acides aminés dans une protéine est appelée une séquence protéique, ou structure primaire. Par convention, cette séquence est notée en partant de la partie N-terminale (-NH2), où l'extrémité amine est libre, jusqu'à la partie C-terminale (-COOH), où l'extrémité acide est libre (78). Cette structure primaire est déterminée par la séquence codante du gène correspondant à la protéine en question.

La séquence protéique est représentée par une succession spécifique de lettres correspondant aux 20 acides aminés existants (Figure 14).

```
3LII_hAChE   1 MRPPQCLLHTPSLASPLLLLLLWLLGGGVGAEGREDAELLVTVRGGRLRGIRLKTPGGPV
3LII_hAChE  61 SAFLGIPFAEPPMGPRRFLPPEPKQPWSGVVDATTFQSVCYQYVDTLYPGFEGTEMWNPN
3LII_hAChE 121 RELSEDCLYLNVWTPYPRPTSPTPVLVWIYGGGFYSGASSLDVYDGRFLVQAERTVLVSM
3LII_hAChE 181 NYRVGAFGFLALPGSREAPGNVGLLDQRLALQWVQENVAAFGGDPTSVTLFGESAGAASV
3LII_hAChE 241 GMHLLSPPSRGLFHRAVLQSGAPNGPWATVGMGEARRRATQLAHLVGCPPGGTGGNDTEL
3LII_hAChE 301 VACLRTRPAQVLVNHEWHVLPQESVFRFSFVPVVDGDFLSDTPEALINAGDFHGLQVLVG
3LII_hAChE 361 VVKDEGSYFLVYGAPGFSKDNESLISRAEFLAGVRVGVPQVSDLAAEAVVLHYTDWLHPE
3LII_hAChE 421 DPARLREALSDVVGDHNVVCPVAQLAGRLAAQGARVYAYVFEHRASTLSWPLWMGVPHGY
3LII_hAChE 481 EIEFIFGIPLDPSRNYTAEEKIFAQRLMRYWANFARTGDPNEPRDPKAPQWPPYTAGAQQ
3LII_hAChE 541 YVSLDLRPLEVRRGLRAQACAFWNRFLPKLLSATDTLDEAERQWKAEFHRWSSYMVHWKN
3LII_hAChE 601 QFDHYSKQDRCSDL
```

Figure 14 : Structure primaire/Séquence protéique de l'acétylcholinestérase humaine (Identifiant PDB : 3LII) (79)

2.3.3. Structure secondaire

La succession de certains acides aminés dans une séquence entraîne la formation de structures secondaires, en réalisant des interactions locales entre eux par l'intermédiaire de liaisons non covalentes, majoritairement des liaisons hydrogènes (78). Les structures secondaires retrouvées le plus couramment sont :

- Les hélices α, formées par l'enroulement sur elle-même d'une partie de la structure primaire afin de créer une hélice où chaque acide aminé forme une liaison hydrogène avec l'acide aminé situé à quatre positions avant et/ou après lui (Figure 15 – gauche).
- Les feuillets β, formés par l'association de 2 à 15 brins β reliés latéralement par des liaisons hydrogènes, formant un plan plissé généralement torsadé. Un brin β est une partie d'une structure primaire de 3 à 10 acides aminés en moyenne (Figure 15 – droite).

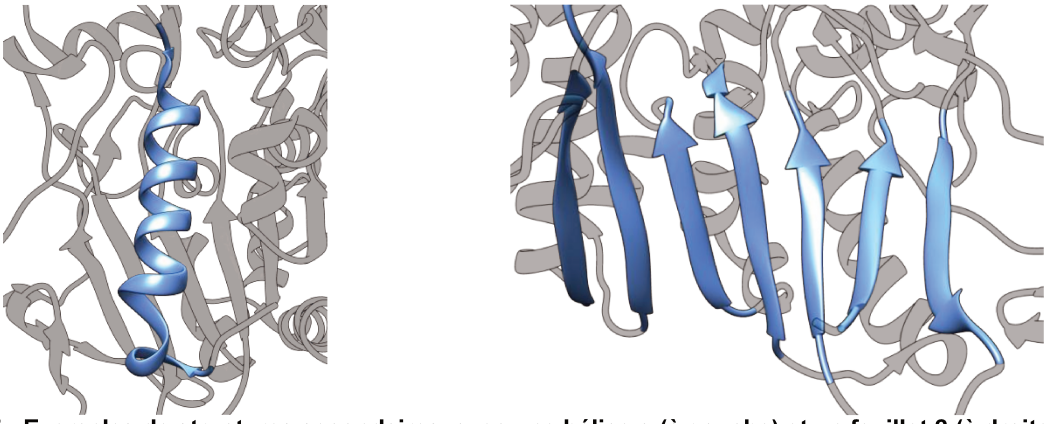


Figure 15 : Exemples de structures secondaires, avec une hélice α (à gauche) et un feuillet β (à droite) obtenus à partir de la structure cristallographique de l'acétylcholinestérase humaine (Identifiant PDB : 3LII) (79)

2.3.4. Structure tertiaire

Le repliement tridimensionnel d'une structure primaire d'une protéine, et l'association de plusieurs structures secondaires sous la forme de domaines protéines, forment la structure tertiaire de la protéine (Figure 16) (78). Ce repliement permet l'obtention de fonctions très diverses, comprenant entre autres :

- Des fonctions de catalyse de réactions chimiques, les protéines sont alors appelées des enzymes.
- Des fonctions de signalisation cellulaire et de transduction de signaux, comprenant les hormones peptidiques ou les anticorps.
- Des fonctions de structuration, conférant de la rigidité aux cellules et tissus avec les protéines structurelles, ou générant des forces mécaniques essentielles à la motilité avec les protéines motrices.

La structure tertiaire est stabilisée par la présence de plusieurs types de liaisons faibles comprenant des liaisons hydrogènes, des interactions hydrophobes, des liaisons ioniques, etc… entre les chaînes latérales des acides aminés. La structure tertiaire est définie par les coordonnées de tous les atomes de la protéine dans l'espace.

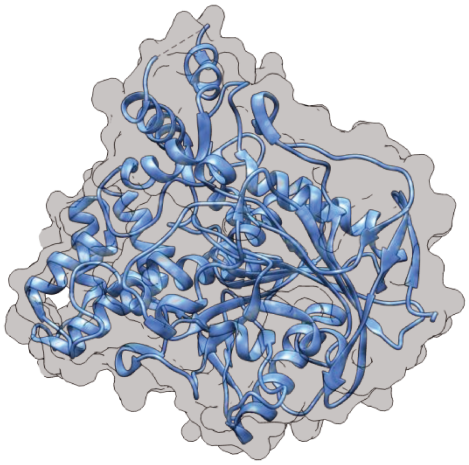


Figure 16 : Structure tertiaire de l'acétylcholinestérase humaine (Identifiant PDB : 3LII) (79)

2.3.5. Structure quaternaire

Certaines structures tertiaires possèdent la capacité se combiner entre elles afin de former des structures tridimensionnelles plus complexes, formant une structure quaternaire (78). Chaque structure tertiaire comprise dans une structure quaternaire est appelée une sous-unité. L'association de deux sous-unités est appelée un dimère, de trois sous-unités un trimère, de quatre sous-unités un tétramère, etc…

La structure quaternaire permet également d'apporter des fonctions supplémentaires aux protéines, et est représentée de la même manière que les structures tertiaires avec les coordonnées de tous les atomes des protéines dans l'espace.

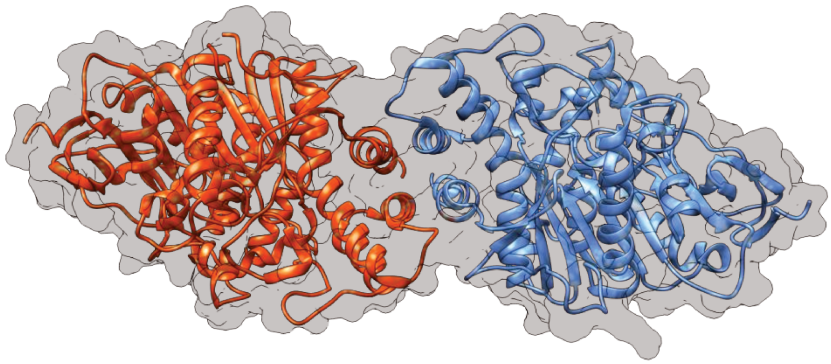


Figure 17 : Structure quaternaire de l'acétylcholinestérase humaine (Identifiant PDB : 3LII) (79)

2.3.6. Détermination de la structure tertiaire et quaternaire

Des méthodes expérimentales biophysiques permettent la détermination de la structure tertiaire ou quaternaire de protéine, via l'utilisation de la cristallographie à rayons X, de la spectroscopie par résonance magnétique nucléaire ou plus récemment de la cryomicroscopie (80). Ces méthodes mesurent la distribution des électrons de la protéine sous la forme de densités, pour en déduire les coordonnées tridimensionnelles de chaque atome avec une certaine précision, appelée résolution.

Il est également possible de prédire la structure tertiaire de protéines par des méthodes automatiques, en se basant sur la structure de protéines proches connues (méthodes par homologie) ou sur les propriétés physico-chimiques des atomes (méthodes *ab initio* et *de novo*).

2.3.7. Bases de données de structures protéiques

La grande majorité des structures tridimensionnelles élucidées sont répertoriées dans une banque de données sur les protéines, appelée *Protein Data Bank* ou encore PDB. Les structures expérimentales obtenues sont déposées dans la PDB afin d'entrer dans le domaine public et d'être accessibles gratuitement (32).

2.4. Champs de forces classiques

En chimie, un champ de forces est un jeu de paramètres déterminés de manière empirique décrivant toutes les interactions entre atomes dans un système moléculaire (54). Ce champ de forces permet de calculer l'énergie potentielle d'un système moléculaire en connaissant les coordonnées des atomes et leurs liaisons. L'énergie potentielle peut être calculée en additionnant toutes les interactions liées et non liées réalisées au sein du système.

Parmi les interactions liées, on peut prendre en compte, entre autres :

- L'élongation des liaisons chimiques, qui sont considérées comme des potentiels harmoniques, comparable à des ressorts,
- La déformation des angles entre atomes, également considérés comme des potentiels harmoniques, comparable à des charnières,
- La déformation des angles dièdres, qui concernent la rotation autour de la liaison centrale de trois liaisons consécutives, comparable à des torsions.

Les interactions non liées prennent en compte les interactions réalisées à distance, avec des atomes situés soit à plus de trois liaisons, soit situés sur une autre molécule. On y trouve :

- Les interactions électrostatiques, qui sont estimées via la loi de Coulomb qui exprime la force des interactions électriques entre deux atomes chargés.
- Les forces de van der Walls, qui sont dues aux interactions entre deux dipôles, permanents et/ou induits, estimées par le potentiel de Lennard-Jones.

L'addition de tous ces potentiels donne l'énergie potentielle totale du système, où les termes surlignés en rouge concernent les paramètres empiriques du champ de forces déterminés pour les états d'équilibre, et les termes surlignés en bleu concernent les paramètres variant avec la géométrie du système moléculaire étudié (Figure 18).

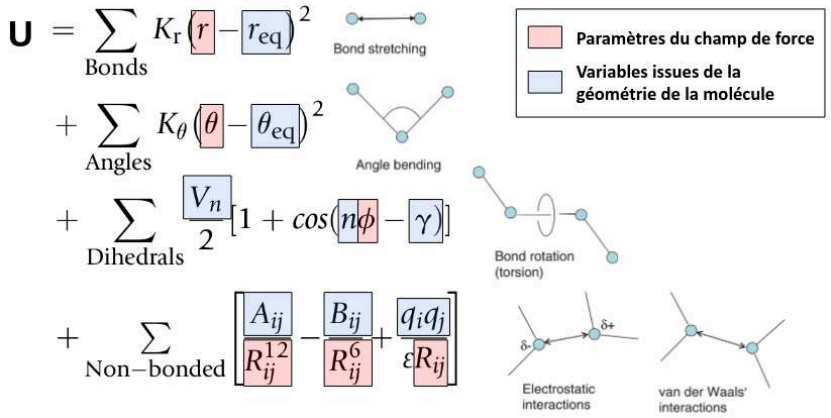


Figure 18 : Equation de l'énergie potentielle d'un système et de la signification des différents termes, avec la mise en évidence de la contribution du champ de force (en rouge) et de la géométrie de la structure observée (en bleu). Adapté de (54,81)

Les champs de forces les plus couramment utilisés pour des simulations de biomolécules et décrits dans la littérature sont AMBER (*Assisted Model Building and Energy Refinement*) (82), CHARMM (*Chemistry at HARvard Molecular Mechanics*) (83), GROMOS (*GROningen MOlecular Simulation package*) (84) ou encore OPLS (*Optimized Potential for Liquid Simulations*) (85).

Certains champs de forces peuvent prendre en compte des interactions supplémentaires, comme les torsions impropres qui sont des mouvements de courbure hors du plan permettant de maintenir la planéité des cycles aromatiques, ou d'autres méthodes de calcul des interactions électrostatiques considérant les variations de charge causées par l'environnement chimique proche (54).

3. Approches basées sur la structure des cibles

3.1. Amarrage moléculaire & criblage virtuel à haut-débit

3.1.1. Définition & introduction

L'amarrage moléculaire, également appelé « *docking* », est une technique permettant la prédiction précise des interactions entre deux molécules afin de former un complexe stable (Figure 19) (50,86,87). Cette technique, utilisée dans le cadre de l'approche basée sur la structure des cibles, permet de prédire les interactions réalisées par la liaison d'une petite molécule (appelé le « ligand ») au sein d'un site de liaison d'une protéine cible (appelée le « récepteur »). Chaque conformation du ligand obtenue est appelée une pose, et est classée en utilisant une fonction de score.

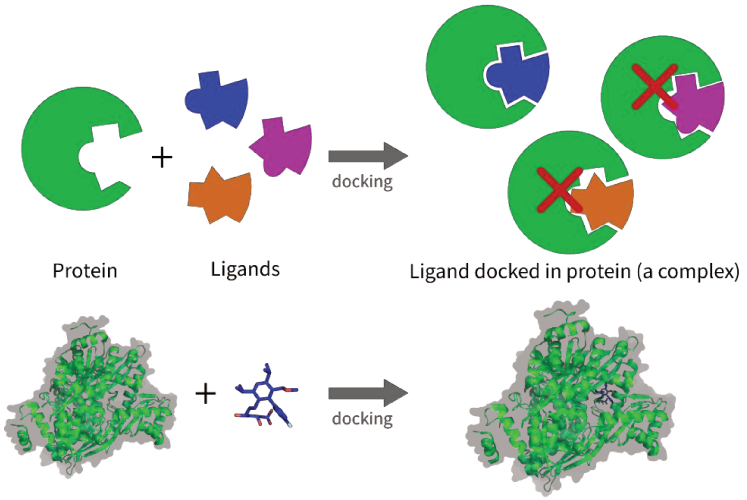


Figure 19 : Représentation schématique du principe d'amarrage moléculaire (88)

Afin de réaliser une recherche rapide de petites molécules pouvant se lier à une protéine d'intérêt, il est possible d'appliquer l'amarrage moléculaire à une plus grande échelle et de manière plus grossière, dans le cadre d'un criblage virtuel à haut débit (« *High-Throughput Virtual Screening* ») (Figure 20) (32,50). A partir de grandes bases de données de petites molécules comprenant des milliers à des millions de composés différents, un sous-ensemble de ligands pertinents peut être mis en évidence pour une activité biologique souhaitée. L'idée est donc de tester en amont les composés par ordinateur afin de réduire drastiquement le nombre de composés qui devront être testés expérimentalement, permettant de réduire le temps et le coût des expérimentations *in vitro*.

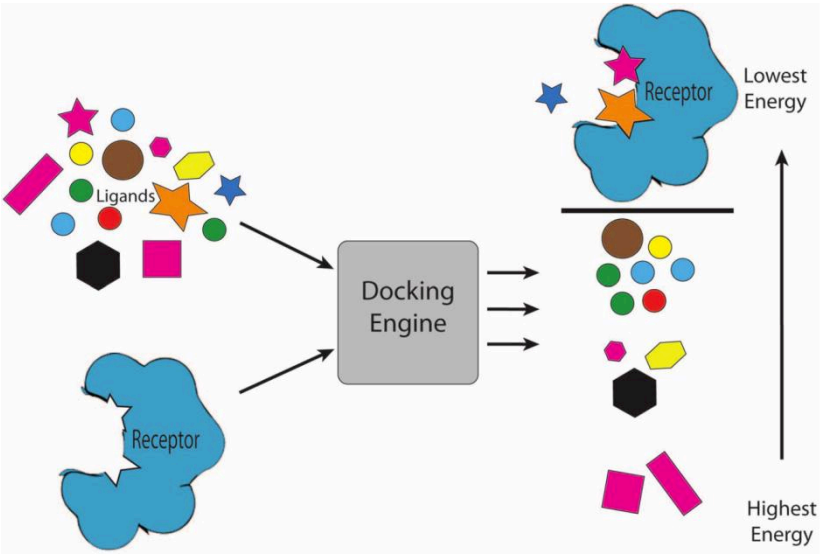


Figure 20 : Représentation schématique du principe du criblage virtuel à haut débit (89)

L'amarrage moléculaire et le criblage virtuel à haut débit permettent entre autre (50,87,90):

- De déterminer précisément les interactions intermoléculaires réalisées entre une petite molécule et sa protéine cible afin de comprendre son mode de liaison et d'en déduire son impact,
- D'identifier de nouvelles touches lors de la phase de recherche exploratoire, en criblant *in silico* des bases de données de molécules pour mettre en évidence des molécules susceptibles de se lier à la protéine d'intérêt,
- D'optimiser les chefs de file confirmés, en proposant des modifications permettant l'obtention d'analogues plus sélectifs et plus puissants en se basant sur leur orientation au sein de la protéine d'intérêt.

3.1.2. Mise au point

Préparation du récepteur

L'amarrage moléculaire et le criblage virtuel à haut débit nécessitant la structure tridimensionnelle d'un récepteur, celui-ci est choisi avec attention.

Cette structure peut être déterminée par des techniques expérimentales, comprenant la Résonance Magnétique Nucléaire (RMN) ou la cristallographie par rayons X (80). La qualité de la structure du récepteur est évaluée selon :

- Sa résolution (qui doit être préférentiellement inférieure à 2 Å),
- Le facteur de Debye-Waller également appelé le facteur B exprimant la précision de la position de chaque atome,
- La comparaison de paramètres observés dans la structure par rapport aux paramètres moyens de toutes les cristallographies.

La présence d'un ligand dans la structure est une information importante, permettant de déterminer précisément le site de liaison en fonction de ses interactions avec certains acides aminés clés de la protéine.

S'il n'existe pas de structure déterminée expérimentalement, des techniques de modélisation *in silico* peuvent être utilisées, notamment la modélisation par homologie afin de générer une structure 3D à partir d'une protéine de la même famille possédant une structure 3D déterminée expérimentalement. Cette reconstruction se base sur l'alignement des séquences des protéines utilisées, et des modèles corrects peuvent être créés avec un minimum de 50% d'acides aminés identiques entre les deux séquences (91).

Afin de les préparer aux différentes techniques, ces structures devront être modifiées en fonction des besoins, notamment avec la suppression de toutes les molécules d'eau et des ligands si ceux-ci se trouvent au niveau du site de liaison (certaines molécules d'eau étant primordiales pour la liaison entre un ligand et son récepteur, il est possible de les garder en conséquence), l'ajout de tous les hydrogènes des acides aminés, et du calcul des charges (50).

Préparation des ligands

L'amarrage moléculaire et le criblage virtuel à haut débit nécessitent également une sélection de ligands, sous des conformations 3D. Ces ligands sont choisis en fonction des besoins liés au sujet d'étude. Après les avoir soit récupérées dans des bases de données, soit dessinées dans des logiciels spécifiques, des conformations 3D sont générées en prenant en compte l'état d'ionisation des molécules, les tautomères ou les isomères si nécessaire.

Pour procéder au criblage virtuel à haut débit, il faut créer une grande sélection de composés chimiques à partir de bases de données, appelé des chimiothèques virtuelles, comprenant des milliers à des millions de composés différant par leur taille et leurs caractéristiques. Ces chimiothèques peuvent être publiques, regroupant les molécules décrites dans la littérature ou testées puis répertoriées publiquement, commerciales, où les molécules sont également proposées à l'achat physique auprès de fournisseurs, ou privées, créées et utilisées pour un usage interne uniquement. Plusieurs types de chimiothèques existent :

- Les chimiothèques dites classiques, regroupant plusieurs millions à dizaine de millions de composés chimiques différents, avec une diversité chimique très importante.
- Les chimiothèques focalisées, dédiées pour une famille de cibles spécifique, avec l'utilisation de composés chimiques aux caractéristiques physico-chimiques proches de composés de référence.

- Les chimiothèques de produits naturels, qui proposent des composés extraits ou dérivés de sources naturelles.
- Les chimiothèques de fragments, regroupant des composés chimiques de petite taille, permettant d'obtenir des points de départ qui seront transformés en molécules plus grandes et plus affines en prenant en compte leur positionnement dans le site de liaison.

Des molécules testées *in vitro* et/ou *in vivo*, décrites comme étant soit actives soit inactives, sont ajoutées aux composés sélectionnés afin d'introduire des témoins positifs et négatifs.

Après la sélection d'une ou de plusieurs chimiothèques virtuelles à utiliser pour le criblage, il est nécessaire de préparer l'ensemble des composés en les transformant dans un format adapté et en réalisant des étapes de filtrage pour réduire la taille de la chimiothèque qui sera effectivement criblée (50,75).

Dans un premier temps, les composés possédant des structures incorrectes, au minimum un centre asymétrique non défini, ou étant en doublon sont retirés du jeu de données. Les ions ou solvants présents dans les structures seront retirés. De la même manière que pour l'amarrage moléculaire, s'il est nécessaire de prendre en compte l'état d'ionisation, les tautomères ou les isomères, les molécules sont générées en conséquence. Tous les composés sont ensuite transformés en structure 3D.

Dans un second temps, la base de données de molécules est filtrée en fonction de différents critères, afin de diminuer sa taille et d'optimiser les résultats qui seront obtenus. A partir des propriétés physico-chimiques et en appliquant les règles de Lipinski, il est possible de limiter la base de données aux composés les plus susceptibles d'être biodisponibles. Ces règles de Lipinski prennent en compte le poids moléculaire, la polarité et le nombre d'accepteurs ou de donneurs de liaisons hydrogènes des composés. Il est possible de les compléter en appliquant les règles de Veber, prenant en compte le nombre de liaisons en libre rotation et l'aire de la surface polaire des composés.

En complément, il est également possible de filtrer la base de données en prenant en compte la toxicité potentielle des composés, au travers de modèles de prédiction QSAR (voir la partie X.X) ou de la présence de sous-structures classées comme toxiques. Il est également possible de filtrer la base de composés en fonction des composés qui pourraient poser problème lors des tests *in vitro*, via notamment la présence de structures d'interférences, les PAINS (*Pan-assay interference compounds*), donnant souvent des résultats faussement positifs ou étant agrégateurs (92–94).

3.1.3. Algorithmes de recherche

Les méthodes d'amarrage moléculaire mettent en jeu des algorithmes de recherche afin de produire des poses pour chaque composé. Une « pose » est un mode de liaison proposé par l'algorithme. La flexibilité du ligand est traitée différemment au sein de trois groupes de méthodes (50,90).

Les méthodes systématiques ne sont utilisables que pour l'amarrage moléculaire rigide, où tous les ligands sont considérés comme des objets rigides. Aucune flexibilité ne leur est accordée, avec la fixation des liaisons et des angles, mais plusieurs conformations des ligands sont générées en amont afin d'outrepasser ce manque de flexibilité. L'algorithme de recherche est principalement basé sur une combinaison de translations et de rotations du ligand au sein du site de liaison. Ces méthodes comprennent également les algorithmes *de novo*, permettant de construire un ligand progressivement ou d'assembler des fragments dans le site de liaison.

Afin d'appréhender la flexibilité des ligands, (i) les méthodes stochastiques se basent sur des modifications aléatoires de la conformation des ligands dans le but de réaliser de l'échantillonnage conformationnel. Elles mettent en jeu l'algorithme de Monte-Carlo Metropolis, visant à introduire un biais dans l'échantillonnage pour générer préférentiellement des configurations stables de basse énergie, ainsi que les algorithmes génétiques, qui intègrent les mécanismes de la génétique naturelle et de la théorie de l'évolution.

(ii) Les méthodes par simulation emploient la minimisation d'énergie et la dynamique moléculaire. Ces techniques, demandant des ressources de calcul importantes et n'explorant pas aussi bien que les méthodes systématiques, sont principalement utilisées pour affiner les résultats obtenus. La minimisation d'énergie être également utilisée en complément des autres méthodes afin d'obtenir des conformations stables énergétiquement.

3.1.4. Fonctions de score

Suite à l'échantillonnage et la détermination de poses, il est nécessaire de pouvoir classer ces poses en fonction de leur pertinence. Pour cela, une fonction de score est utilisée, permettant l'obtention d'un score qui est une approximation de l'affinité entre le ligand et la protéine étudiée, à ne pas confondre avec l'activité du ligand (50,87,90). Les fonctions de score les plus communément utilisées se classent en trois catégories :

- Les fonctions de score empiriques, basées sur le nombre et le type d'interactions réalisées entre le ligand et son site de liaison, le nombre de liaisons en libre rotation et la surface des atomes ne contribuant pas au contact avec le solvant, pondérés en utilisant des transformations statistiques.

- Les fonctions de score basées sur l'utilisation d'un champ de forces, où l'affinité est estimée en se basant sur les interactions de van der Waals et les interactions électrostatiques entre tous les atomes du ligand et du site de liaison, ainsi que l'énergie des liaisons et des angles du ligand.
- Les fonctions de score basées sur l'utilisation de potentiels énergétiques moyens obtenus par l'analyse de grandes bases de données d'observations expérimentales de complexes ligand-récepteur.

Pour éviter les faux positifs, où le ligand est prédit comme se liant à la cible mais non confirmé par des tests expérimentaux, ainsi que les faux négatifs, le cas inverse, il est fortement suggéré d'utiliser deux ou plusieurs fonctions de score pour évaluer l'affinité de la liaison. Des protocoles d'affinage, appelés également protocoles de *rescoring*, permet de calculer de nouveaux scores en utilisant d'autres fonctions (87). D'autres méthodes peuvent également être employées, comme la mesure précise des énergies conformationnelles et de solvatation en utilisant d'importantes ressources de calcul (87).

3.1.5. Validation

Comme toute approche, il est nécessaire de valider le système obtenu pour s'assurer que l'on peut se fier aux résultats. L'approche la plus communément utilisée est de regarder les résultats pour des molécules de référence testées *in vitro* et/ou *in vivo* pour la cible étudiée, décrits comme étant actifs (témoins positifs) ou inactifs (témoins négatifs), et qui sont ajoutées lors de la préparation du jeu de composés (50).

Dans le cas idéal où des structures tridimensionnelles de la cible co-cristallisées avec un ou plusieurs ligands de références existent, la position de ceux-ci dans leur structure cristallographique est comparée aux poses obtenues lors de l'amarrage moléculaire. Dans le cas où il n'existe pas de structure tridimensionnelle en complexe, il est nécessaire d'identifier le site de liaison le plus probable des ligands de référence avec la cible en identifiant toutes les interactions réalisables par les ligands ainsi que les acides aminés pouvant y contribuer, à partir de données biologiques ou d'approches basées sur les ligands.

3.1.6. Sélection des meilleurs composés

Les résultats d'un amarrage moléculaire ou d'un criblage virtuel doivent être observés et filtrés avant de sélectionner les meilleurs composés qui seront testés biologiquement (50,95).

Les scores ne doivent pas être les seuls critères de sélection, car ils sont principalement basés sur une notion quantitative des interactions réalisées entre un

ligand et sa cible. Il est donc possible d'obtenir des composés possédant des scores très hauts, mais présentant des conformations non stables ou ne possédant pas d'interactions clés décrites dans la littérature. Les scores permettent donc de créer une présélection de composés, dont les poses seront par la suite inspectées visuellement pour apprécier leur mode de liaison, les interactions clés réalisées, leur forme, etc…

Il est également possible d'utiliser des techniques d'exploration de données (*data mining*) pour d'analyser de très grands jeux de poses afin de sélectionner des composés très différents, aussi bien dans leurs structures chimiques que leurs modes de liaison (96). De plus, en cas d'utilisation des scores, ceux-ci peuvent être normalisés en calculant l'efficacité du ligand (*Ligand Efficiency*) en les divisant par les masses moléculaires (97).

Enfin, des modèles issus d'approches basées sur les ligands, notamment des modèles pharmacophoriques, peuvent être appliqués sur toutes les poses de tous les composés dans le but de sélectionner ceux respectant un mode de liaison déterminé (98).

3.1.7. Exemples d'application

Une application possible de l'amarrage moléculaire et du criblage virtuel à haut débit est le repositionnement de molécules existantes (possédant une AMM ou testées dans des essais cliniques) vers de nouvelles cibles afin de traiter des pathologies différentes. Dans un article publié en 2020, cette application est utilisée dans le but de trouver des traitements potentiels pour traiter la COVID-19, causée par le virus SARS-CoV-2 (*Severe acute respiratory syndrome coronavirus 2*) (99). Dans cet article, une étude par amarrage moléculaire a testé 61 molécules déjà utilisées en tant qu'agents antiviraux ou au stade de l'essai clinique pour cibler la protéase principale du SARS-CoV-2 (Figure 21).

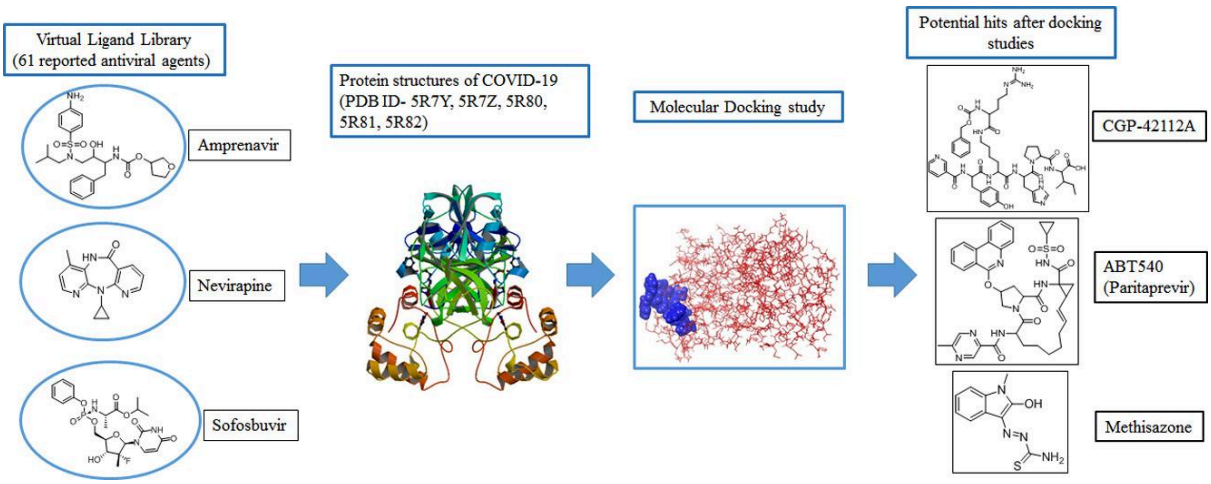


Figure 21 : Schéma du protocole d'amarrage virtuel et de criblage virtuel à haut débit utilisé (99)

Ce jeu de molécules a été extrait d'une base de données, et des conformations tridimensionnelles ont été générées et minimisées. En ce qui concerne le choix de la cible, 5 structures cristallographiques ont été extraites d'une base de données, et ont été préparées. L'amarrage moléculaire a employé une méthode stochastique et une fonction de score empirique, calculé à partir des interactions réalisées. Les meilleures poses proposées par le logiciel présentent les scores les plus négatifs.

La figure 22 présente les résultats comparatifs de 37 composés interagissant avec au minimum 2 structures cristallographiques sur 5, avec au minimum un score inférieur à -6.5. Ici, on peut voir que des molécules déjà connues contre la COVID-19 ressortent du lot, avec 4 inhibiteurs de la protéase du VIH (le Lopinavir, l'Asunaprevir, l'Indinavir, et le Ritonavir) ainsi que le Remdesivir. Ces molécules possèdent de très bons scores, et interagissent avec des acides aminés clés de la protéase. De plus, de nouvelles molécules non-observées jusqu'ici ont pu être mises en évidence, comme le Methisazone, le Paritaprevir (ABT450) et le CGP42112A, qui pourront faire l'objet dans le futur de tests *in vitro* et *in vivo* afin d'évaluer leur pertinence.

Comparative docking study results on COVID-19 enzymes.

Sr. No	Drug	Dock score[a]				
		5R7Y	5R7Z	5R80	5R81	5R82
1.	NSC306711 (Ferristatin II)	−7.331	–	–	−9.147	–
2.	Lopinavir	−6.834	−6.968	−7.331	−8.44	−7.58
3.	Elbasvir	–	–	–	−9.027	–
4.	Asunaprevir	–	−7.725	–	−8.257	−6.597
5.	Simeprevir	–	−7.778	–	−7.784	–
6.	Remdesivir	–	−7.001	−7.674	–	−7.911
7.	CGP42112A	−7.108	–	–	−7.521	−7.243
8.	Indinavir	–	−7.058	−7.157	−6.834	−6.796
9.	Ritonavir	−7.621	–	−6.736	−6.764	−7.316
10.	ABT450	–	−6.330	–	−7.327	−6.617
11.	Marboran/ Methisazone	–	−7.542	−6.829	−6.928	–
12.	Galidesivir	−6.857	–	–	–	−7.239
13.	Saquinavir	–	–	−7.227	−7.632	–
14.	Baricitinib	–	–	−7.075	−7.43	–
15.	Raltegravir	–	–	−7.057	−7.257	–
16.	Delavirdine	–	−7.634	–	–	–
17.	Elvitegravir	–	−7.189	–	–	–
18.	Danoprevir	–	−6.956	–	−6.83	–
19.	Galidesivir	–	−6.737	−6.873	−6.601	–
20.	Entecavir	−6.715	−6.687	–	–	–
21.	Famciclovir	–	−6.521	−6.965	–	–
22.	Uprifosbuvir	–	–	−7.558	–	−6.805
23.	Oseltamivir	–	–	−6.891	–	–
24.	Azidothimidine	–	–	−6.873	–	–
25.	Sofosbuvir	–	–	−6.756	–	−7.037
26.	Tenofovir	–	–	−6.644	−6.739	–
27.	Mericitabine	–	–	−6.572	–	–
28.	Zanamivir	–	–	−6.548	–	–
29.	Didanosine	–	–	−6.526	–	–
30.	faldaprevir	–	–	–	−7.652	−6.703
31.	Grazoprevir	–	–	–	−7.237	–
32.	Vedroprevir	–	–	–	−7.172	–
33.	Ravidasvir	–	–	–	−7.031	–
34.	Amprenavir	–	–	–	−6.583	−6.744
35.	Efavirenz	–	–	–	−6.509	−6.657
36.	Telaprevir	–	–	–	–	−7.083
37.	Daclatasvir	–	–	–	–	−6.881

– Indicates that dock score value is higher than −6.5.
[a] Dock score value of −6.5 or lower is mentioned in table.

Figure 22 : Etude comparative des résultats d'amarrage moléculaire sur cinq structures cristallographiques de protéases du SARS-CoV-2 (99)

3.2. Dynamique moléculaire

3.2.1. Introduction

L'amarrage moléculaire et le criblage virtuel à haut débit sont des techniques populaires et puissantes pour la découverte de nouvelles touches, mais celles-ci ne prennent que très peu en considération la flexibilité de la cible étudiée. En effet, ces techniques se basent sur le modèle "clé-serrure" proposé par E. Fischer, posant l'hypothèse de la complémentarité de forme entre le ligand et son site de liaison, où il n'existe aucune flexibilité lors de l'interaction entre les deux partenaires (28). Certaines techniques permettent l'échantillonnage conformationnel du ligand afin d'amener une notion de flexibilité au système, mais le site de liaison protéique est très généralement considéré comme statique dans le temps.

Or, les protéines sont des entités dynamiques, où leur fonction est liée de très près à leurs mouvements. La prise en compte de la flexibilité des protéines est donc fondamentale. En conséquence, D. Koshland a proposé une adaptation du modèle "clé-serrure" afin de tenir compte de la flexibilité des protéines, appelé le modèle de l'ajustement induit (100). Ce modèle considère le changement dynamique du site de liaison et du ligand, afin d'obtenir l'interaction la plus optimale possible. Le site de liaison protéique est toujours en mouvement, lors de la liaison, de l'activité de la protéine, et de la libération des produits finaux obtenus pour revenir à un état de base.

La méthode de la dynamique moléculaire, appliquée dans le domaine de la biologie, est capable de rendre compte de la flexibilité et de la dynamique des protéines en augmentant la précision du calcul des interactions liées et non-liées entre chaque atome du système (54,101–103). Cette précision accrue est réalisée au détriment des ressources de calculs nécessaires à l'application de cette méthode. Toutefois, grâce aux progrès matériels et logiciels considérables, l'utilisation de ces simulations assistées par ordinateur est maintenant compatible avec les délais classiques des projets de découverte et de développement de médicaments.

3.2.2. Principes de la dynamique moléculaire

La dynamique moléculaire classique est une technique de simulation assistée par ordinateur de l'évolution d'un système à l'échelle de l'atome en fonction du temps, en décrivant les mouvements des atomes et des molécules qui le compose, en suivant les lois du mouvement de Newton (appelé également mécanique Newtonienne) (54,101–103). Cette technique permet entre autres :

- L'étude du repliement des protéines,
- L'étude des mouvements du squelette et des structures secondaires d'une protéine dans le temps afin d'identifier des sites de liaisons transitoires,

- L'étude dynamique des mécanismes physico-chimiques mis en jeu lors de la liaison d'une molécule à un site de liaison protéique,
- L'obtention d'informations au niveau atomique sur la mécanistique des fonctions de la protéine étudiée, utiles lors de la conception et l'amélioration de touches ou de chefs de file.

La simulation d'une dynamique moléculaire classique consiste à l'intégration des lois du mouvement de Newton.

Equations du mouvement

Première loi de Newton

La première loi de Newton, ou le principe d'inertie, est énoncée selon ces termes: "Tout corps persévère dans l'état de repos ou de mouvement uniforme en ligne droite dans lequel il se trouve, à moins que quelque force n'agisse sur lui, et ne le contraigne à changer d'état" (104).

Cette loi indique que si la somme de toutes les forces qui s'appliquent sur un objet est nulle, alors la vitesse de cet objet est constante. Un objet en mouvement ne changera donc pas de vitesse sauf si une nouvelle force vient s'appliquer. Si ce principe d'inertie est vérifié, le référentiel utilisé, classiquement cartésien et orthonormé où un point est repéré par trois coordonnées notées x, y et z, sera appelé le référentiel galiléen.

Seconde loi de Newton

La seconde loi de Newton est le principe fondamental de la dynamique. "Les changements qui arrivent dans le mouvement sont proportionnels à la force motrice ; et se font dans la ligne droite dans laquelle cette force a été imprimée" (104). Dans un référentiel galiléen, la seconde loi de Newton dit qu'une force qui s'applique à un objet est égale à l'accélération de l'objet multipliée par sa masse. Plus un objet sera lourd, et plus la force requise pour l'accélérer à une vitesse définie sera importante.

$$\boldsymbol{F = m \cdot a}$$

Avec :
- F : Force s'appliquant à un objet
- m : Masse de l'objet
- a : Accélération de l'objet

De plus, toujours dans un référentiel galiléen, l'énergie potentielle est la force qui est exercée par un champ de forces, ici un champ de forces moléculaire (54,103). Cette énergie potentielle, notée U, est égale au travail à fournir contre cette force pour déplacer un objet de sa position d'origine à une nouvelle position. La force qui doit être exercée doit être égale à l'énergie potentielle, mais dirigée de façon

opposée. La force exercée par le champ de forces tend toujours vers une énergie totale plus faible, soit l'état le plus stable, et donc réduit l'énergie potentielle.

En résumé, si l'on applique la seconde loi de Newton à un système moléculaire composé de i atomes, on peut en déduire cette équation différentielle (54,103):

$$-\frac{\boldsymbol{dU}}{\boldsymbol{dx_i}} = \boldsymbol{m_i} \cdot \boldsymbol{a_i}$$

Avec :
- U : Energie potentielle
- x_i : Coordonnées d'un atome i
- a_i : Accélération d'un atome i
- m_i : Masse d'un atome i

Cette équation permet de faire le lien entre une structure moléculaire aux coordonnées connues et l'énergie potentiel du système, précisant le coût énergétique d'un déplacement atomique, ici via l'accélération.

Algorithmes d'intégration

En se basant sur un développement en série de Taylor, il est également possible d'exprimer la nouvelle position d'un objet aux coordonnées $x_i(t+\Delta t)$, en connaissant ses coordonnées à l'instant d'avant $x_i(t)$, ainsi que sa vitesse, son accélération, etc… (54,102,103)

$$\boldsymbol{x_i(t + \Delta t) = x_i(t) + v_i(t)\Delta t + \frac{1}{2}a_i(t)\Delta t^2 + \ldots}$$

Avec :
- x_i : Coordonnées d'un atome i
- v_i : Vitesse d'un atome i
- a_i : Accélération d'un atome i
- t : Temps d'origine
- Δt : Intervalle de temps

Ce développement étant infini, sa résolution complète est impossible. Toutefois, il est possible d'obtenir une estimation réaliste en s'arrêtant au second ordre. Certains algorithmes d'intégration proposent des résolutions estimées de cette équation, notamment l'algorithme de Verlet, permettant d'exprimer les nouvelles coordonnées d'un objet en connaissant ses deux précédentes positions ainsi que son accélération, pour de très petits intervalles de temps, de l'ordre de la femtoseconde (10^{-15} seconde) (54,102,103).

$$\boldsymbol{x_i(t + \Delta t) = 2x_i(t) - x_i(t - \Delta t) + a(t)\Delta t^2}$$

Avec :

- x_i : Les coordonnées d'un atome i
- a_i : L'accélération d'un atome i
- t : Temps d'origine
- Δt : Intervalle de temps

Il est donc possible, en utilisant la seconde loi de Newton et un algorithme d'intégration, ici Verlet, de déterminer la position d'un atome en connaissant :

- Ses deux positions précédentes, avec la position de départ obtenue via des structures cristallographiques ou des modélisations tridimensionnelles *in silico*, et la seconde position en estimant le développement en série de Taylor de la manière la plus précise possible.
- Son accélération, qui est directement calculée à partir de ses coordonnées, de sa masse et de son énergie potentielle, elle-même obtenue via l'utilisation d'un champ de forces.

Thermodynamique et ensembles statistiques

Les simulations de dynamique moléculaire sont réalisées à une échelle microscopique : il est donc nécessaire de convertir les données obtenues à une échelle macroscopique, afin de pouvoir rendre comparable les simulations réalisées avec des résultats expérimentaux.

Au niveau macroscopique, en thermodynamique, l'état d'équilibre d'un système peut être précisément défini par des variables d'état comprenant la température (T), le volume (V), la pression (P), le nombre de particules (n) ou l'énergie du système (E), liés notamment au travers de la loi des gaz parfaits (Équation ci-dessous) (103). Les autres propriétés thermodynamiques peuvent être déduites directement avec l'application des grands principes de la thermodynamique.

$$\boldsymbol{E = \frac{P \cdot V}{n \cdot R \cdot T}}$$ avec R = constante des gaz parfaits = 8.3 $J.K^{-1}.mol^{-1}$

Afin de réaliser le lien entre ces variables macroscopiques et les données obtenues lors de dynamiques moléculaires, comme les coordonnées des atomes ou leur vitesse, des concepts de physique statistique sont appliqués, notamment les concepts d'ensemble statistique (101,103,105,106).

Un ensemble statistique vise à représenter les états possibles d'un système mécanique en équilibre thermique avec un bain de chaleur à température fixe (Figure 23). Trois ensembles statistiques sont principalement utilisés en dynamique moléculaire :

- L'ensemble microcanonique NVE, où le système est totalement isolé thermodynamiquement : il ne peut échanger ni particules (N), ni énergie (E) avec l'extérieur, et ne peut pas changer de volume (V).
- L'ensemble canonique NVT, où le système est considéré être en contact avec un « thermostat » appliquant une température (T) fixée, en autorisant uniquement les échanges d'énergie : le nombre de particules (N) et le volume (V) sont également fixes. Cet ensemble permet de mimer des réactions biologiques se produisant à des températures constantes.
- L'ensemble isotherme-isobare NPT, où le système est considéré être au contact d'un « barostat » appliquant une pression (P) fixée, en autorisant uniquement les échanges d'énergie : le nombre de particules (N) et la température (T) sont également fixes. Cet ensemble permet de simuler des réactions chimiques réalisées à l'air libre, sous la pression atmosphérique considérée comme constante.

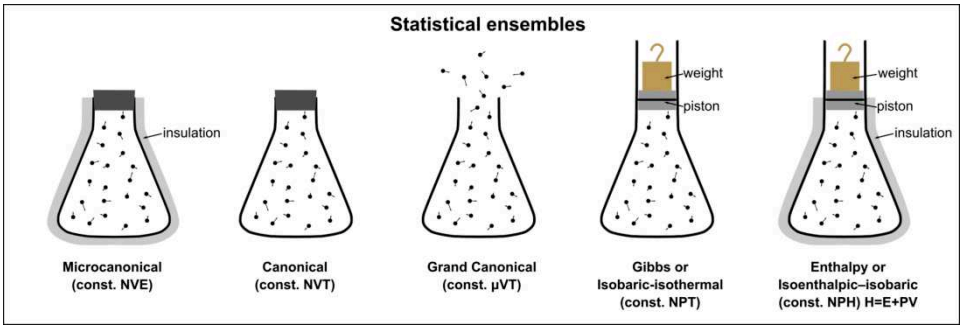


Figure 23 : Représentation visuelle d'ensembles statistiques (107)

La dynamique moléculaire se basant sur les lois de Newton, l'ensemble se rapprochant le plus des conditions de simulation est l'ensemble microcanonique NVE, mais celui-ci ne correspond que très rarement aux conditions expérimentales, où la température, la pression, ou les deux paramètres sont fixés. Dans le but de se rapprocher de ces conditions expérimentales, des méthodes d'incorporation de thermostat ou de barostat sont utilisées lors des simulations pour travailler dans un ensemble canonique NVT ou un ensemble isotherme-isobare NPT.

Un thermostat ou un barostat en dynamique moléculaire vise à modifier les vitesses du système lors de la simulation afin que l'énergie cinétique moyenne soit égale à l'énergie cinétique correspondant à une température cible, pour se rapprocher des conditions expérimentales. Les thermostats les plus utilisés sont ceux

de Langevin, d'Andersen ou de Nosé-Hoover, et les barostats les plus utilisés sont ceux de Berendsen ou de Monte Carlo.

3.2.3. Aspects pratiques de la dynamique moléculaire

Chaque dynamique moléculaire se décompose en 5 étapes (54,102,108):

- L'initialisation du système, pour préparer la structure souhaitée
- La minimisation énergétique, pour libérer les tensions causées entre autre par de mauvais contacts entre atomes
- Le chauffage, pour augmenter la température du système jusqu'à celle qui sera utilisée lors de la simulation, et générer des vitesses pour chaque atome
- L'équilibrage, pour équilibrer le système en utilisant le protocole de simulation normal
- La production, qui sera l'étape utile où les trajectoires seront sauvegardées

Initialisation du système

Une dynamique moléculaire se base sur une conformation de départ, obtenue via une structure cristallographique ou une modélisation *in silico*. La structure tridimensionnelle de départ est préparée en (108):

- Supprimant tout ce qui n'est pas nécessaire à la simulation : les solvants de co-cristallisation, les molécules d'eau, les ions, les ligands non nécessaires, …
- Reconstruisant si nécessaire les chaînes latérales incomplètes et les acides aminés non résolus dans la structure cristallographique,
- Ajoutant tous les hydrogènes,
- Calculant les charges de chaque atome en fonction du champ de force utilisé.

Traitement du solvant biologique

L'eau joue un rôle fondamental dans la structuration des protéines et dans leurs fonctions. Afin de se rapprocher des conditions biologiques, il est nécessaire d'ajouter du solvant sous la forme d'une boîte, où l'eau est représentée de manière explicite par des modèles de molécules d'eau et où des contre-ions sont ajoutés si nécessaire pour neutraliser le système (Figure 24) (102,106). La géométrie de la boîte utilisée (cubique, prismatique, octaédrique, …) ainsi que sa taille ont un impact direct sur les ressources de calcul nécessaires, en lien avec le nombre de molécules d'eau présentes dans le système, et donc en lien avec le nombre de calculs des interactions augmentant de manière exponentielle pour chaque itération.

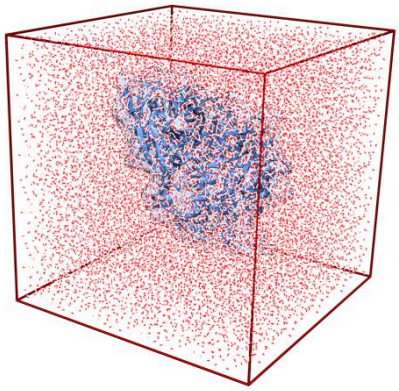


Figure 24 : Structure cristallographique de l'acétylcholinestérase humaine après définition d'une boîte de simulation et ajout de molécules d'eau (Identifiant PDB : 3LII) (79)

Condition périodique aux limites

Le système étudié étant microscopique et isolé dans un espace, le calcul de l'énergie du système serait biaisé avec la non-présence de partenaires aux limites de la boîte, et il existerait des phénomènes de tension superficielle très important entraînant l'éclatement de la boîte de simulation (101–103).

Pour éviter ces problèmes, le système simulé est représenté comme un système périodique afin d'obtenir une approximation d'un grand système (considéré comme infini) via la répétition des coordonnées de la boîte principale, où se déroule la simulation, dans tous les axes de l'espace : c'est la condition périodique aux limites (*Periodic boundary conditions*) (Figure 25).

En pratique, si une molécule quitte la boîte de la simulation d'un côté, elle sera remplacée par une molécule correspondante située dans l'image de la boîte centrale de l'autre côté, permettant de garder un nombre de particules fixe tout au long de la simulation (Figure 26).

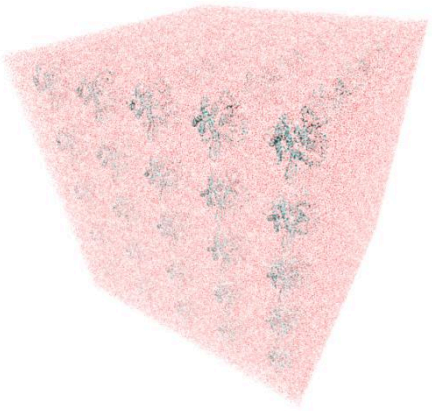


Figure 25 : Représentation du système obtenu à partir de la cristallographique de l'acétylcholinestérase humaine en appliquant la condition périodique aux limites (Identifiant PDB : 3LII) (79)

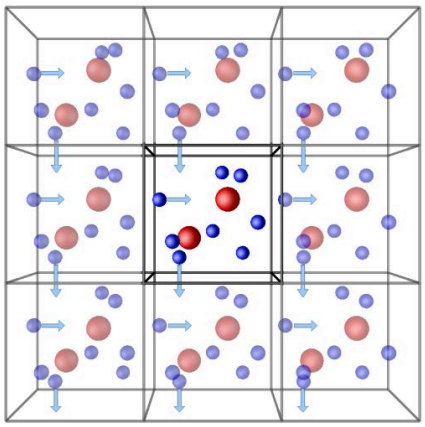


Figure 26 : Illustration du principe de la condition périodique aux limites. Un atome qui sort de la boîte de simulation par un côté sera réintroduite dans la boîte par le côté opposé, dans les trois dimensions de l'espace (109)

Minimisation d'énergie

Une minimisation de l'énergie potentielle est réalisée dans un premier temps, afin d'optimiser la structure d'entrée en éliminant les contacts défavorables potentiels entre atomes, et en cherchant une géométrie permettant un état stable du système, possédant une basse énergie (54,102,108).

Les algorithmes de minimisation qui sont utilisés cherchent tous à minimiser la fonction d'énergie potentielle, décrivant un paysage énergétique conformationnel de la protéine, où chaque minimum local représente un état stable de la protéine, et où le minimum global représente l'état le plus stable possible de la protéine (Figure 27). Les algorithmes permettent de naviguer au sein de ce paysage énergétique afin de converger dans un des minima locaux le plus proche de la structure de départ.

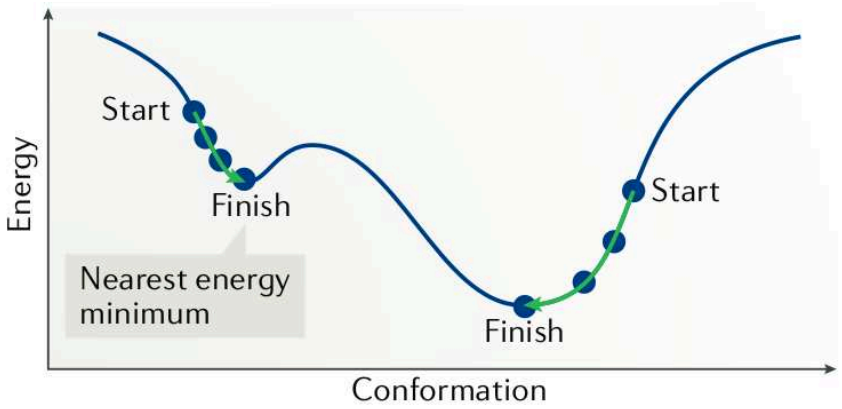


Figure 27 : Représentation d'un paysage énergétique conformationnel à une dimension, avec deux minima locaux dont un minimum global (110)

Les algorithmes de minimisation les plus utilisés se basent sur le calcul de manière itérative des dérivées de la fonction d'énergie (54):

- La méthode de la plus grande pente (*Steepest descent*) se base sur le calcul de la pente de la courbe. C'est une méthode efficace quand le minimum local est proche de la structure de base, mais relativement lente.
- La méthode du gradient conjugué (*Conjugate gradient*) se base également sur le calcul de la pente de la courbe, mais en prenant compte également des valeurs calculées aux étapes précédentes. Cette méthode permet donc d'accélérer la convergence, mais elle n'est pas adaptée aux structures présentant de nombreux mauvais contacts.
- Les méthodes de Newton-Raphson utilisent la dérivée seconde, permettant une minimisation rapide mais demandant plus de ressources de calcul.

En pratique, une combinaison de méthodes est utilisée dans un processus de minimisation de structure protéique, avec notamment la combinaison des méthodes de la plus grande pente et du gradient conjugué, utilisées l'une après l'autre. L'étape de minimisation permet l'obtention de structures optimisées, libérées de potentiels

excès de tensions causés notamment par des mauvais contacts entre atomes trop proches.

Chauffage

Afin d'amener le système à la température de simulation souhaitée, il est nécessaire de réaliser une courte étape de chauffage (102,108). En se basant sur un ensemble canonique NVT, un thermostat est utilisé afin d'augmenter progressivement la température de la simulation jusqu'à la température qui sera utilisée par la suite, sur un ordre de temps allant des picosecondes à la nanoseconde. Pour chaque pas, des vitesses sont assignées à chaque atome, variables en fonction de la température atteinte.
En règle générale, le chauffage est réalisé par paliers ou par augmentation linéaire de la température, de 0 K jusqu'à 300 K, entre 10 et 100ps.

L'étape de chauffage permet l'obtention d'un système ayant atteint la température de simulation souhaitée, avec des vitesses générées pour chaque atome du système.

Équilibrage

L'équilibrage des molécules d'eau autour de la protéine est nécessaire. Pour cela, la protéine doit être maintenue en place via l'utilisation de contraintes de position sur son squelette carboné, tout en permettant aux molécules de solvant de se déplacer librement dans le système. Cette phase est réalisée dans un ensemble NPT, où le nombre de particules, la pression et la température du système sont constants, permettant la stabilisation du volume du système, via l'utilisation d'un thermostat et d'un barostat (102,108).

Production

Quand l'étape d'équilibrage est terminée, les contraintes de position qui ont été imposées sont relâchées, permettant la réalisation de l'étape de production (54,108). Cette étape étant la suite directe de l'équilibrage, elle est également réalisée dans un ensemble NPT, avec l'utilisation d'un thermostat et d'un barostat. Les énergies calculées et les coordonnées de chaque atome sont enregistrées de manière périodique dans un fichier, appelé fichier de la « trajectoire » de la dynamique moléculaire. Chaque jeu de coordonnées enregistré à un temps défini est appelé une « image ». Ce processus est réalisé sur des temps de l'ordre de la nanoseconde et de la microseconde (Figure 28).

La trajectoire obtenue représente le mouvement de la protéine au cours de la simulation, et peut être affichée sous forme de film à l'aide de logiciels de visualisation.

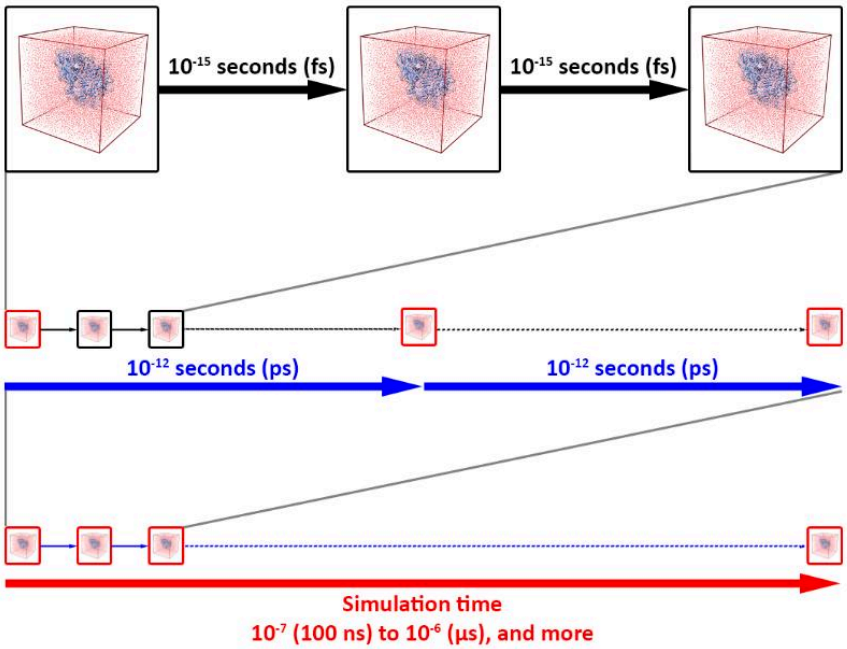


Figure 28 : Etape de production d'une dynamique moléculaire. Les jeux de coordonnés enregistrés pour obtenir la trajectoire de la dynamique moléculaire sont encadrés en rouge.

3.2.4. Analyse des résultats

L'analyse des résultats intervient dès l'obtention d'une trajectoire. Les types d'analyse réalisables sont illimités, une simulation de dynamique moléculaire peut donc être réutilisée de nombreuses fois (111).

Les premières analyses qui sont réalisées visent à simplifier la visualisation de la trajectoire ainsi qu'à la préparer pour l'obtention de données. Les structures composant celle-ci sont alignées avec la structure de départ, et les molécules d'eau sont éliminées si elles ne sont pas indispensables pour le sujet traité. L'élimination de ces molécules d'eau permet d'optimiser l'espace disque et la quantité de mémoire vive nécessaire aux analyses, car la grande majorité des données de la trajectoire recense les coordonnées de ces molécules. Enfin, l'observation de trajectoires est réalisée avec l'aide de logiciels de visualisation, comme UCSF Chimera (112) ou VMD (113) .

Séries temporelles

Pour toutes les conformations que comportent une trajectoire de dynamique moléculaire, la variation d'angles ou de distances entre des atomes choisis au préalable sont observables dans le temps. Il est également possible d'observer l'évolution de structures tridimensionnelles, comme le changement des structures secondaires au cours du temps (Figure 29), la variation de surfaces et/ou de volumes de poches contenant de potentiels sites de liaison, ou encore l'évolution d'interactions moléculaires comme les liaisons hydrogènes et les liaisons ioniques.

Pour toutes ces informations, des statistiques sont réalisées au travers du calcul de moyennes et d'écart types, de distributions, etc…

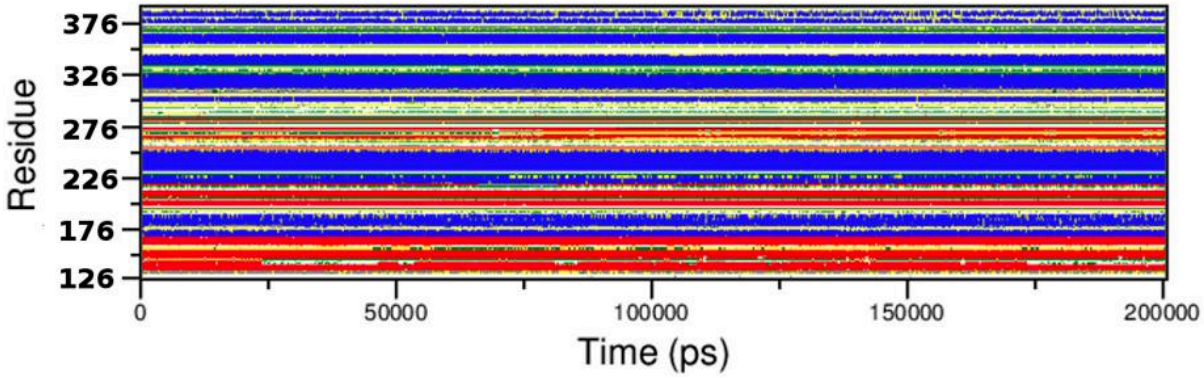


Figure 29 : Graphique de l'évolution des structures secondaires dans le temps d'une simulation de dynamique moléculaire réalisée sur la protéine Aurora-A (114)

Grandeurs calculées

Des grandeurs peuvent également être calculées, comme par exemple le RMSD et le RMSF permettant d'obtenir des informations sur l'ensemble de la structure au cours de la simulation.

Le RMSD (*Root-Mean-Square Deviation*), ou écart quadratique moyen, est une mesure de la distance moyenne entre les coordonnées de deux groupes d'atomes (54,115). Cette mesure peut être appliquée aux conformations observées au cours de la dynamique moléculaire, en comparant les images de la trajectoire avec une structure de référence ou avec la première image de la trajectoire. Le graphique obtenu, traçant l'évolution du RMSD dans le temps, permet d'évaluer les changements de conformation de la biomolécule étudiée (Figure 30). L'augmentation du graphique RMSD en fonction du temps montre que la protéine s'écarte progressivement de sa conformation d'origine.

Le RMSF (*Root-Mean-Square Fluctuation*), ou fluctuation quadratique moyenne, est une mesure proche du RMSD, visant à mesurer la variation de position moyenne de chaque objet étudié (atomes, groupes d'atomes, résidus, …) au cours du temps par rapport à une position de référence (115). Le RMSF permet donc d'analyser les portions de structure qui fluctuent plus ou moins par rapport à une structure de référence ou la première image de la trajectoire au cours de la simulation de la dynamique moléculaire (Figure 30). Les valeurs élevées de RMSF sont très souvent corrélées à des régions possédant plus de flexibilité conformationnelle, ayant donc tendance à avoir des amplitudes de mouvement plus importantes.

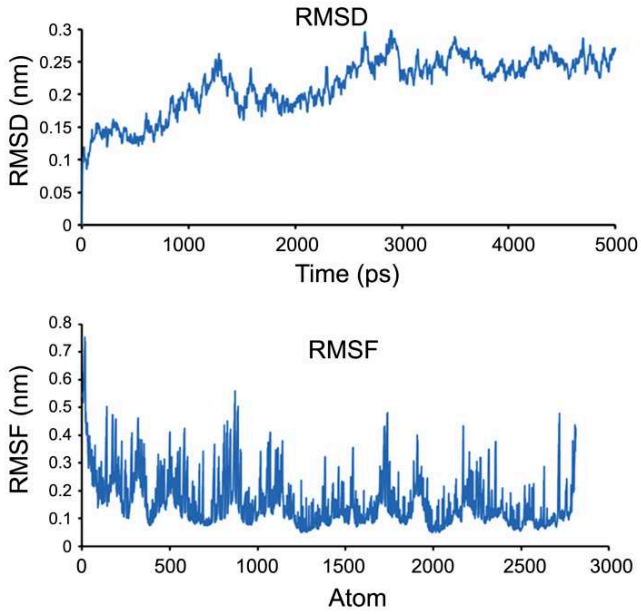


Figure 30 : Graphiques d'un RMSD (en haut) et d'un RMSF (en bas) d'une simulation de dynamique moléculaire réalisée sur la protéine Aurora-B (116)

MM/GBSA et MM/PBSA

A partir de trajectoires de dynamiques moléculaires mettant en jeu un récepteur et un ligand, il est possible d'estimer l'énergie libre de liaison et d'identifier la contribution des acides aminés participant à cette liaison en utilisant les méthodes MM/GBSA (*Molecular Mechanics Generalized Born Surface Area*) et MM/PBSA (*Molecular Mechanics Poisson-Boltzmann Surface Area*). A partir de configurations extraites d'une trajectoire, ces méthodes vont estimer des énergies libres observées lors du complexe récepteur-ligand, et en prenant à part le ligand puis le récepteur. L'énergie libre de liaison est calculée en soustrayant l'énergie libre de la protéine seule et du ligand seul à l'énergie libre du complexe récepteur-ligand (117).

Ces informations permettent par exemple de rationaliser des résultats expérimentaux, ou d'améliorer des résultats obtenus lors de protocoles d'amarrage moléculaire ou de criblage virtuel à haut débit. Toutefois, ces méthodes manquent de précision car elles ne prennent pas en compte le nombre et l'énergie libre des molécules d'eau situées au niveau du site de liaison (117).

3.2.5. Techniques dérivées

Une des techniques dérivées est la dynamique moléculaire dirigée (*Steered Molecular Dynamics*), permettant de « tirer » virtuellement sur certains atomes de la protéine en appliquant des forces externes. Cette technique permet simuler la

déliaison d'un ligand, ou d'entrainer des changements conformationnels souhaités (118).

Une seconde méthode dérivée de la dynamique moléculaire est la métadynamique. La structure tridimensionnelle utilisée lors de l'initialisation d'une dynamique moléculaire ne présente d'une seule conformation stable d'une protéine. La métadynamique permet de faire de l'échantillonnage conformationnel en ajoutant petit à petit de l'énergie au système, permettant de ne pas revoir les conformations précédentes (119). L'énergie est amenée via la modification de variables du système, allant de simples distances entre atomes à des fonctions mathématiques complexes (120). Appliquée au domaine de la conception de médicaments, cette méthode permet d'étudier le processus de liaison ou de déliaison entre un ligand et son récepteur (121).

3.2.6. Avantages et limites de la dynamique moléculaire

La dynamique moléculaire permet d'étudier avec précision le comportement de protéines et d'autres biomolécules à l'échelle atomique, dans une très petite échelle de temps. Appliquée à la conception de médicaments, elle permet d'élucider le mécanisme moléculaire d'une protéine, de déterminer les acides aminés clés participant à son activité, et d'étudier le comportement d'un ligand au sein de son site de liaison (106,122).

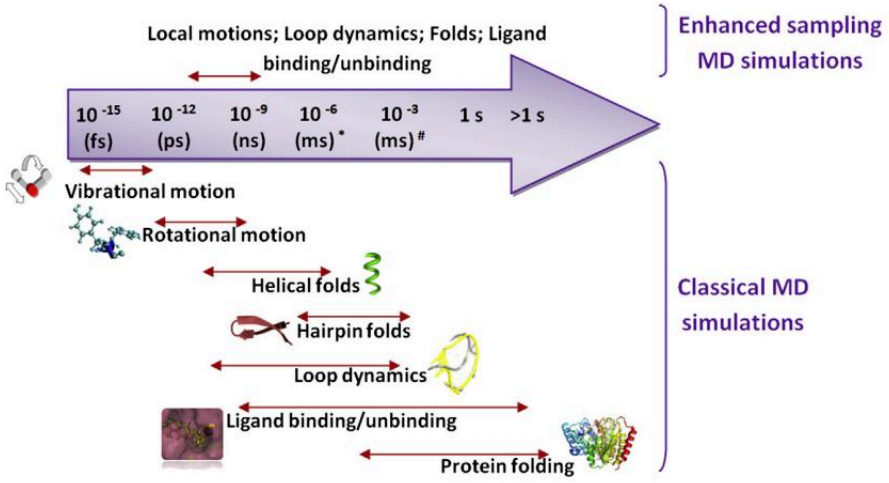


Figure 31 : Échelle de temps montrant les différents mouvements des protéines lors de simulations de dynamique moléculaire (123)

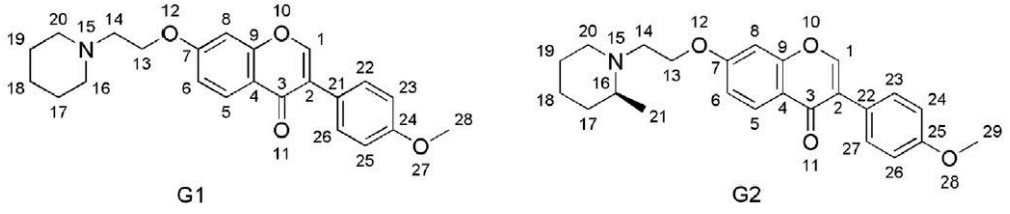
Toutefois, les échelles de temps utilisées lors de la dynamique moléculaire sont également problématiques pour étudier des mouvements protéiques complexes (106,122). En effet, l'étude du repliement des protéines intervient dans des temps de l'ordre de la milliseconde à quelques secondes, très loin des nanosecondes et microsecondes simulables. L'étude de la liaison et de la déliaison d'un ligand se

déroule également pendant quelques nanosecondes à plusieurs secondes, mais ces phénomènes sont étudiables en utilisant des techniques dérivées mettant en jeu l'ajout d'énergies comme la métadynamique (Figure 31).

De plus, la taille du système qui peut être étudié est limitée par la puissance de calcul et le temps accordé à une simulation. Enfin, la dynamique moléculaire se base sur des données expérimentales avec la formulation d'hypothèses qui conditionnent le comportement du système. Des approximations et simplifications sont également utilisées, avec l'utilisation de champs de forces non adaptés, l'utilisation d'algorithmes moins précis ou le calcul des interactions entre chaque atome. Ceci amène des erreurs dans les calculs réalisés. Une validation expérimentale est donc nécessaire, soit de manière directe avec le calcul de grandeurs expérimentales, soit de manière indirecte via l'évaluation de conséquences vérifiables expérimentalement. Cette évaluation indirecte peut être réalisée via des mesures de biophysique ou l'utilisation de la mutagenèse, en mutant des acides aminés clés dans l'activité de la protéine étudiée (111).

3.2.7. Exemples d'applications

Un exemple d'utilisation de dynamique moléculaire peut être réalisé pour l'étude du mode de liaison de ligands. Dans un article publié en 2014, une équipe de recherche chinoise présente une étude réalisée sur la différence d'inhibition de l'acétylcholinestérase observée chez deux ligands (124).



G1

3-(4-methoxyphenyl)-7-(2-(piperidin-1-yl)ethoxy)-4*H*-chromen-4-one

G2

(*S*)-3-(4-methoxyphenyl)-7-(2-(2-methylpiperidin-1-yl)ethoxy)-4*H*-chromen-4-one

Figure 32 : Structures 2D des ligands G1 (à gauche) et G2 (à droite) (124)

Ces ligands, appelés G1 et G2, ne diffèrent que d'un groupement méthyl, mais l'effet inhibiteur observé pour le ligand G1 est 80 fois supérieur à celui du ligand G2 (Figure 32). Pour comprendre cette différence, des dynamiques moléculaires classiques ont été réalisées à partir de résultats d'amarrage moléculaire, positionnant les ligands G1 et G2 au sein de leur site de liaison, et donnant les complexes AChE/G1 et AChE/G2.

Le protocole de dynamique moléculaire a été réalisé en utilisant le champ de force AMBER ff12SB. Les complexes AChE/G1 et AChE/G2 ont été neutralisés en

ajoutant 10 ions sodium, ont été placés dans une boîte de simulation rectangulaire avec des molécules d'eau, et des conditions périodiques aux limites ont été appliquées. Après une courte étape de minimisation d'énergie combinant la méthode de la plus grande pente puis la méthode du gradient conjugué, les complexes ont été chauffés de 0K à 300K en 50 ps, puis équilibrés pendant 1 ns. Enfin, une production a été réalisée pendant 10 ns, en enregistrant les configurations toutes les 1 ps.

De plus, des énergies libres de liaison ont été calculées en utilisant un algorithme MM/GBSA.

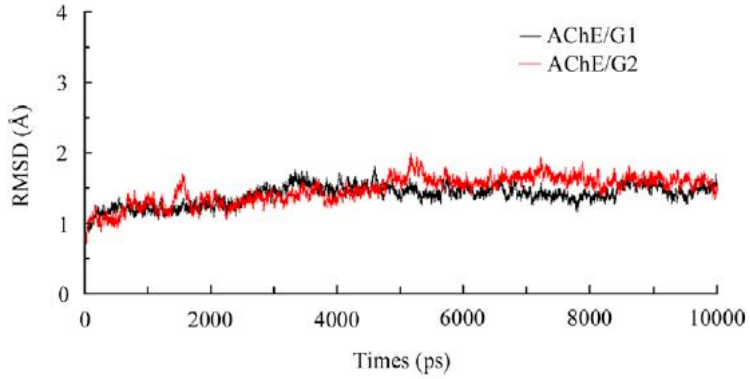


Figure 33 : Evolution du RMSD des atomes du squelette carboné en fonction du temps (124)

Dans un premier temps, la stabilité des complexes est étudiée via le calcul du RMSD des atomes du squelette carboné en fonction du temps (Figure 33). Le RMSD indique ici que les deux trajectoires ont atteint un état d'équilibre après 5000 ps (5 ns). Le calcul des énergies libres de liaison est donc réalisé à partir des conformations situées de 5 ns à 10 ns.

Inhibitor	ΔE_{vdw}	ΔE_{ele}	ΔG_{GB}	ΔG_{SA}	ΔG_{pred}	ΔG_{exp}
G1	−54.8±2.8	−11.6±2.2	29.1±2.2	−6.6±0.2	−43.9±2.9	−9.0
G2	−54.8±2.4	−3.9±3.2	26.0±2.8	−6.2±0.2	−38.9±2.6	−6.4

Figure 34 : Composantes du calcul de l'énergie libre de liaison (en kcal/mol) pour l'acétylcholinestérase en complexe avec les ligands G1 ou G2, à partir d'une méthode MM/GBSA (124)

A partir des composantes du calcul de l'énergie libre de liaison (Figure 34), il peut être observé une différence d'énergie libre de liaison calculée (ΔG_{pred}) du même ordre que l'énergie libre de liaison expérimentale (ΔG_{exp}) calculée à partir de la concentration inhibitrice médiane IC50. De plus, les contributions des termes électrostatiques (ΔE_{ele} et ΔG_{GB}) jouent un rôle important dans la différenciation des activités de G1 et G2.

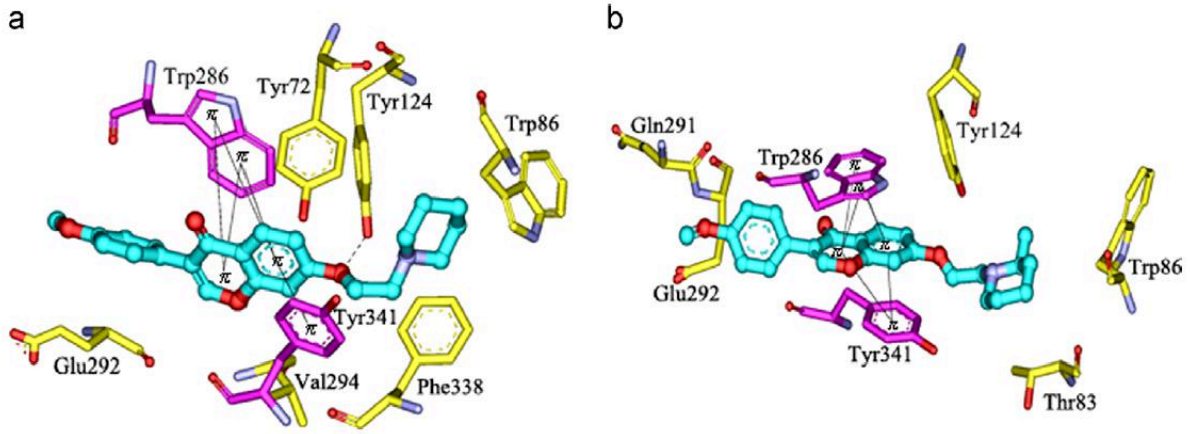


Figure 35 : Modes de liaison des complexes AChE/G1 et AChE/G2 (124)

Les résultats obtenus précédemment proposent une explication à la différence d'activités entre les ligands G1 et G2 au niveau des énergies libres. Les énergies étant définies par les structures et interactions, il est intéressant de regarder les différences au niveau du mode de liaison de chaque ligand à l'acétylcholinestérase (Figure 35). En superposant les deux complexes AChE/G1 et AChE/G2, il

Les ligands G1 (Figure 35 – a) et G2 (Figure 35 – b) vont réaliser des interactions similaires, notamment avec les acides aminés Trp286, Tyr341 et Trp86. Cependant, du fait de la présence d'un groupement méthyle en plus sur le ligand G2, ce dernier se place différemment, et ne réalise plus d'interactions hydrogènes avec l'acide aminé Tyr124, ainsi que quelques interactions hydrophobes avec notamment Tyr72, Val294 et Phe338.

En conclusion, à partir de différentes approches comme l'amarrage moléculaire, la dynamique moléculaire et les méthodes MM/GBSA, il a pu être démontré que la présence d'un groupement méthyle supplémentaire implique une différence au niveau de l'activité inhibitrice de l'acétylcholinestérase. Les informations obtenues permettent de contribuer à la meilleure connaissance du potentiel thérapeutique de cette famille chimique de ligands, et ouvrent la voie à une optimisation rationnelle de ligands afin d'obtenir des inhibiteurs d'acétylcholinestérase plus efficaces.

4. Approches basées sur la structure des ligands

4.1. QSAR/QSP

4.1.1. Introduction

Les relations quantitatives structure-activité, également appelées QSAR (*Quantitative Structure-Activity Relationship*), et structure-propriété, également appelées QSPR (*Quantitative Structure-Property Relationship*) sont des modèles statistiques de prédiction qui cherchent à relier des descripteurs moléculaires à des réponses quantitatives ou qualitatives observées dans un système expérimental comme l'activité, la toxicité ou les propriétés physico-chimiques de composés, au travers de relations mathématiques (Figure 36) (125). Ces modèles de prédiction empiriques se basent sur les tendances observées et les corrélations entre les descripteurs moléculaires et les réponses biologiques, afin d'identifier et de comprendre l'impact de changements structurels d'une série chimique sur leur activité, leurs propriétés ou toxicité.

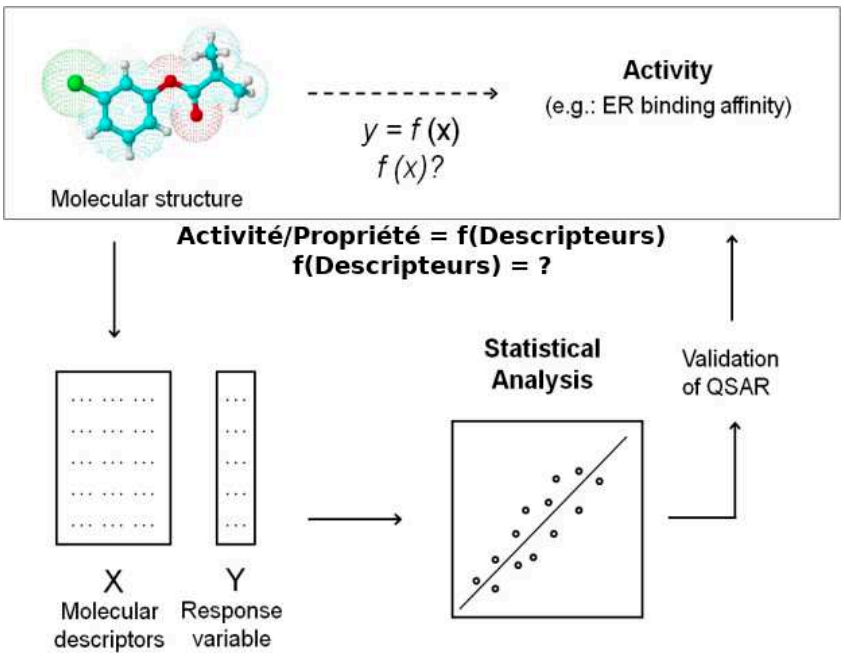


Figure 36 : Principe de la modélisation des relations quantitatives structure-activité/propriété (126)

Les méthodes QSAR/QSPR sont utilisées historiquement en chimie organique physique, et appliquées dans un large champ de domaines, dont la chimie médicinale pour l'optimisation de têtes de file, ainsi que la toxicologie prédictive pour évaluer les risques de petites molécules (127). Ces techniques permettent l'augmentation des capacités de recherche de nouveaux candidats médicaments via la diminution des besoins de synthèse chimique et de criblage biologique, entraînant un gain de temps et d'argent significatif. Elles permettent également de filtrer un jeu de molécules pour ne garder que celles possédant les caractéristiques souhaitées ou

pour éliminer celles potentiellement toxiques, permettant d'aider à la hiérarchisation des tests expérimentaux.

Les prémices de l'utilisation des méthodes QSAR/QSPR remontent aux travaux d'Overton et Meyer en 1899, qui ont observé que la toxicité des produits chimiques organiques simples pour les espèces aquatiques augmente en fonction de leur coefficient de partage, donc en fonction de leur lipophilie (128). L'utilisation des méthodes QSAR/QSPR s'est ensuite démocratisée à partir des années soixante avec les travaux de Hansch et Fujita, qui ont démontré que des activités biologiques peuvent être modélisées mathématiquement à partir des propriétés physico-chimiques de petites molécules, en proposant une première application sur l'évaluation des effets herbicides de composés chimiques avec l'utilisation de modèles de régression linéaire simple (129).

Avec le développement de l'informatique et l'augmentation exponentielle des puissances de calcul, la modélisation QSAR/QSPR s'est diversifiée avec l'apparition de très nombreuses approches statistiques utilisables, augmentant son importance dans les protocoles de conception de médicaments et les travaux de réglementation. L'utilisation de modèles QSAR/QSPR est en effet suggérée par le règlement REACH (Registration, Evaluation, Authorisation and restriction of CHemicals) et par le Conseil International d'Harmonisation des exigences techniques pour l'enregistrement des médicaments à usage humain pour notamment l'évaluation in silico de la génotoxicité (127,130).

4.1.2. Préparation

Le développement classique d'un modèle QSAR/QSPR consiste en (Figure 37) (32,127):

- La recherche de données expérimentales fiables et comparables pour un jeu de composés chimiques,
- La détermination de leurs descripteurs chimiques,
- Le développement d'un modèle statistique en utilisant une ou plusieurs approches différentes,
- La validation et la détermination du domaine d'applicabilité de ce modèle statistique.

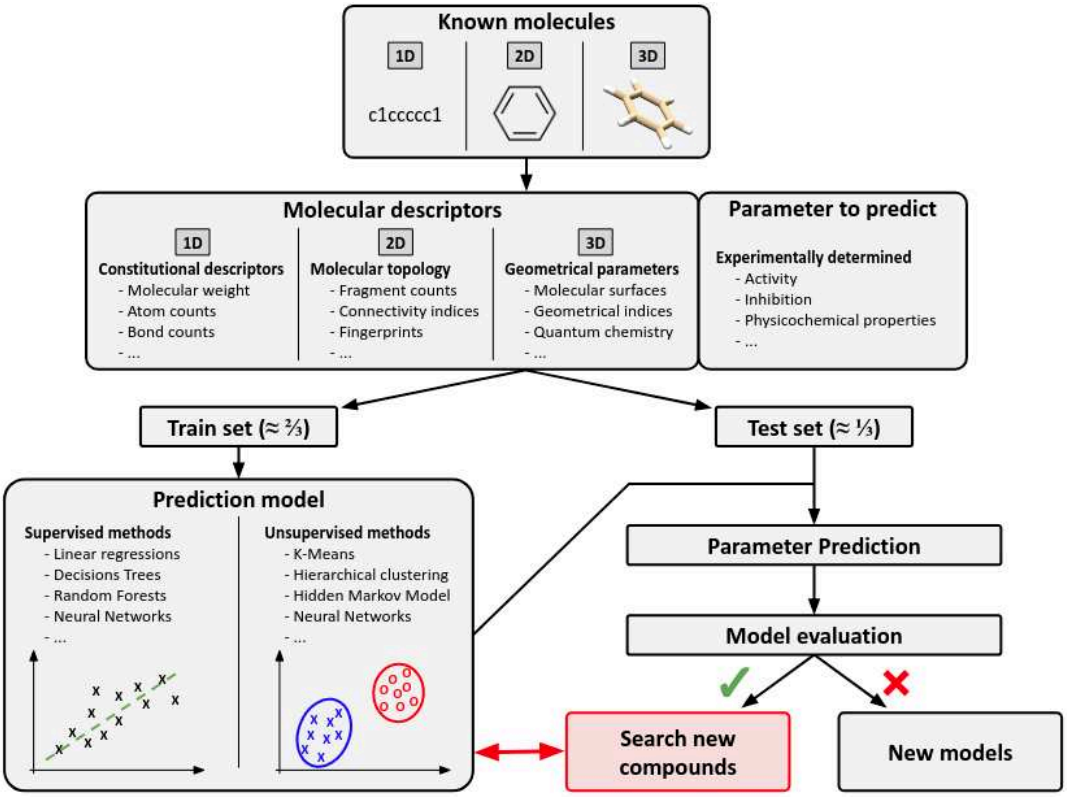


Figure 37 : Développement classique d'un modèle QSAR/QSPR

Choix du jeu de molécules

La première étape de la modélisation QSAR/QSPR est le choix du jeu de données qui permet la création du modèle prédictif (127). Celui-ci doit inclure un jeu de molécules issues de plusieurs séries chimiques avec une diversité structurale, ainsi que des données expérimentales pour la variable qui est prédite, comparables pour chaque molécule. Cette variable peut être entre autres soit une mesure de la bioactivité (concentration inhibitrice médiane IC50, constante d'inhibition Ki, activité, …), soit une mesure de la toxicité (dose létale médiane DL50, dose efficace médiane DE50, dose toxique minimale DTM, …), soit un paramètre physicochimique.

Des bases de données existent pour récupérer des données expérimentales de molécules ayant été testées sur des cibles d'intérêt, en répertoriant les informations publiées dans la littérature ou brevetées. La base de données en accès public ChEMBL, version 27, répertorie 2 millions de molécules bioactives différentes, manuellement organisées et vérifiées, avec 16 millions d'informations d'activités pour 13000 cibles distinctes. La base de données privée Reaxys Medicinal Chemistry répertorie quant à elle plus de 6.5 millions de molécules différentes pour 33.5 millions de données de bioactivité sur 20000 cibles. Les molécules extraites se présentent sous la forme d'une liste de notations linéaires, généralement sous forme de SMILES, soit sous la forme de fichiers avec les coordonnées 2D ou 3D.

Préparation des données

Une des étapes critiques dans la mise en place d'un modèle QSAR/QSPR est la préparation des données, la qualité des données en entrée influant sur la qualité du modèle obtenu. Il est généralement nécessaire de vérifier et de corriger les données, en faisant attention :

- Aux erreurs dans les structures,
- Au manque de standardisation,
- A la présence de molécules en doublons,
- A la présence de composés inorganiques ou sous forme de sels,
- A la tautomérie et à la stéréo-isomérie si les molécules utilisées sont en 3D.

Calcul des descripteurs

Le calcul et la préparation des descripteurs est la seconde étape critique de la mise en place de modèles QSAR/QSPR (127). Il est possible de générer plusieurs milliers de descripteurs pour chaque molécule, mais certaines méthodes sont sensibles au nombre de descripteurs proposés en entrée. Un nombre élevé de descripteurs augmente la probabilité de trouver des corrélations par hasard, surtout dans le cas où il y a plus de descripteurs que de molécules bioactives. Il est donc primordial de choisir avec parcimonie les descripteurs à utiliser, en prenant en compte les spécificités du jeu de données utilisé, et la sensibilité des méthodes qui sont appliquées.

Ensuite, une étape de préparation des données est réalisée, afin de vérifier la présence ou non de descripteurs possédant une variance nulle ou des descripteurs fortement corrélés entre eux, ne permettant pas une discrimination entre les molécules ou de repérer de potentielles erreurs obtenues lors de la génération des descripteurs.

Certaines méthodes sont sensibles aux écarts de grandeurs entre chaque descripteur, où certains descripteurs possédant des ordres de grandeurs importants prennent le pas sur les descripteurs aux valeurs infinitésimales. Pour éviter ceci, l'ensemble des données est souvent normalisé en soustrayant la valeur moyenne de toutes les valeurs d'un descripteur, puis divisé par l'écart-type, permettant d'obtenir des valeurs ayant pour moyenne 0 et un écart-type de 1/4. Cette étape permet de donner un poids égal à chaque descripteur dans la contribution au modèle.

Génération de jeux de composés de test et d'apprentissage

L'ensemble des données préparées est divisé en deux sous-ensembles, un « jeu de composés d'apprentissage » et un « jeu de composés de test » (127). Le jeu d'apprentissage est utilisé pour créer le modèle à partir de la méthode choisie, lors de la phase d'apprentissage. Le jeu de test est utilisé pour évaluer la qualité du

modèle QSAR, notamment sa capacité à prédire la variable étudiée d'un ensemble de molécules qui ne faisait pas parti de la phase d'apprentissage. Il est nécessaire de réaliser cette séparation car les méthodes de modélisation QSAR/QSPR utilisées sont très efficaces pour adapter les descripteurs aux variables à prédire, entraînant parfois un sur-apprentissage, où le modèle est trop bien adapté aux données en entrée. Toutefois, l'inconvénient principal de cette séparation est la limitation du nombre de molécules qui sont utilisées afin d'entraîner le modèle QSAR/QSPR.
Il existe plusieurs méthodes de détermination des jeux d'apprentissage et de test, notamment :

- Par choix aléatoire en utilisant des algorithmes de randomisation,
- Par regroupement afin d'obtenir une diversité chimique similaire entre les deux jeux,
- Par utilisation de la procédure de validation croisée à k blocs, où le jeu de données est divisé en k parties, avec une partie utilisée pour le jeu de test et les parties restantes pour le jeu d'apprentissage, permettant l'obtention de k validations de k modèles créés.

La taille des jeux d'apprentissage et de test dépend du nombre de molécules dans l'ensemble des données préparées. L'approche courante consiste à sélectionner 1/3 des molécules dans le jeu de test, et les 2/3 restants dans le jeu d'apprentissage. Si l'ensemble des données possède un très grand nombre de molécules, il est possible de diminuer encore la répartition pour atteindre 10% pour le jeu de test et 90% pour le jeu d'apprentissage.

4.1.3. Méthodes statistiques

Les modèles QSAR/QSPR sont obtenus à partir de méthodes statistiques basées sur la régression, la classification, ou à partir de techniques d'apprentissage automatique se basant sur l'intelligence artificielle.

Analyse en composantes principales (PCA)

L'analyse en composantes principales, appelé également PCA (*Principal Component Analysis*) est une méthode permettant de transformer un grand nombre de descripteurs étudiés en de nouvelles variables factorielles dans le but d'obtenir un résumé le plus pertinent des données initiales, facilitant son exploration et son étude (127,131). Ces nouvelles variables sont appelées les « composantes principales », ou encore les « axes principaux ».

Tous les descripteurs peuvent contenir de l'information potentiellement pertinente. La PCA permet de créer un nouveau jeu de données avec un nombre de composantes principales égal au nombre de descripteurs en entrée. La PCA permet donc de passer de la représentation des données sous la forme de descripteurs à

une représentation sous la forme de facteurs, les composantes principales, définie par les vecteurs propres d'une matrice de corrélation entre tous les descripteurs.

Après l'obtention des composantes principales, il est nécessaire de choisir le nombre de ces descripteurs à garder pour permettre un résumé suffisamment précis de l'information contenue dans les données initiales. Chaque composante principale explique une partie de l'information globale, où la première composante principale explique la plus grande partie de l'information, la seconde une plus petite partie, ainsi de suite jusqu'à obtenir 100% de l'information expliquée par toutes les composantes principales. Afin de simplifier le nombre de composantes principales, il est nécessaire de choisir à partir de quel stade l'information expliquée par les dernières composantes principales peut être considérée comme négligeable. Plusieurs techniques peuvent être choisies, comprenant le choix d'un pourcentage de l'information expliquée arbitraire, un ratio entre le nombre de données et le nombre de composantes principales, le degré de corrélation entre les variables, l'utilisation des règles de Kaiser, ...

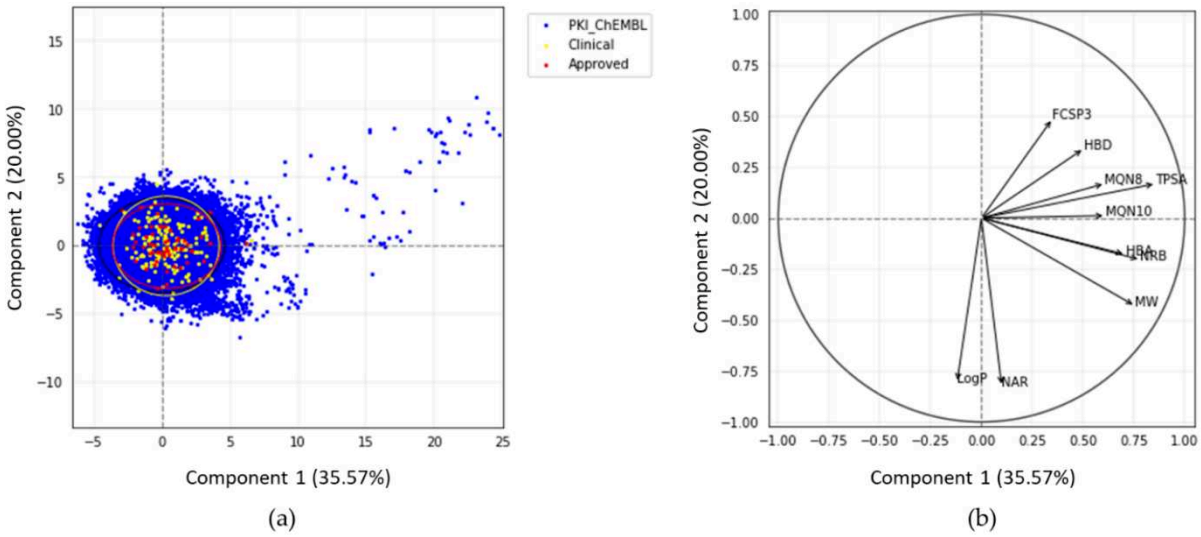


Figure 38 : Exemple de graphique des observations (à gauche) et de cercle des corrélations (à droite), prenant en compte les deux premières composantes de l'Analyse en Composantes Principales expliquant 55.57% de l'information globale (132)

Ensuite, la PCA permet d'obtenir deux graphiques présentant des informations sur la composition du jeu de données utilisé, à partir desquels il est possible d'établir des groupes de données, et de pouvoir expliquer la différence entre ces groupes via les descripteurs utilisés en entrée.

Le premier graphique est le cercle des corrélations, qui est la projection des descripteurs moléculaires utilisés sur le plan des composantes principales (Figure 38 – droite). Les axes des abscisses et des ordonnées représentent des composantes principales, et les flèches situés à l'intérieur, partant du centre, les descripteurs utilisés. L'objectif de ce graphique est de pouvoir expliquer la liaison entre les

descripteurs, en cherchant à comprendre s'il existe des groupes de descripteurs qui sont fortement corrélés entre eux :

- Plus deux descripteurs sont proches, plus ils sont corrélés.
- De la même manière, plus deux descripteurs sont éloignés, plus ils sont anti-corrélés.
- Enfin, un angle de 90° entre deux descripteurs exprime une non-corrélation.

Le second graphique est le graphique des observations, permettant d'afficher toutes les molécules dans les mêmes dimensions de la PCA utilisées précédemment pour le cercle des corrélations (Figure 38 – gauche). La position des molécules sur le graphique est directement liée aux positions des descripteurs moléculaires du cercle des corrélations. Un groupe de molécules situé d'un côté peut être différencié d'un autre en fonction de la position des flèches des descripteurs moléculaires.

La PCA est utilisée afin d'analyser de grands jeux de données pour comprendre les liens entre les molécules entre elles, les descripteurs moléculaires entre eux ainsi qu'entre les molécules et descripteurs. Elle permet également d'être le point de départ de méthodes permettant l'obtention de modèles prédictifs pour ne conserver que le nombre minimal de descripteurs non corrélés contenant le maximum d'information.

Approches basées sur la régression

Régression linéaire multiple

La régression linéaire multiple est une méthode QSAR/QSPR populaire par sa simplicité, sa reproductibilité et sa facilité d'interprétation. Cette méthode est basée directement sur le principe de la régression linéaire afin de prédire une variable à partir de plusieurs autres variables, en supposant qu'il existe une relation linéaire entre eux (127). Dans le cas de la QSAR/QSPR, elle permet de décrire les variations d'une réponse souhaitée en fonction de plusieurs descripteurs moléculaires, qui sont pondérés par des coefficients. Un descripteur possédant un coefficient positif signifie que celui-ci participe positivement à l'activité ou propriété modélisée, un coefficient négatif signifie un impact négatif, et un coefficient proche de 0 indique une non-participation.

Trois types de régressions linéaires multiples sont possibles :

- La méthode progressive ascendante, qui ajoute les descripteurs au modèle un à un, en sélectionnant à chaque étape le descripteur possédant la plus grande participation (positive ou négative) au modèle jusqu'à ce que l'amélioration du modèle ne soit plus significative par rapport aux paramètres de validation.
- La méthode progressive descendante, à l'inverse, utilise au début tous les descripteurs, puis élimine un à un le descripteur possédant la plus faible

contribution au modèle jusqu'à observer une modification significative sur les paramètres de validation.

- La méthode pas à pas, qui combine les deux méthodes précédentes, où les descripteurs sont ajoutés un à un à progressivement, comme lors de la méthode progressive ascendante, mais qui réévalue la contribution du descripteur possédant la plus faible contribution au modèle et peut l'éliminer en conséquence, comme lors de la méthode progressive descendante.

Le bilan de la régression linéaire multiple peut être observé au travers d'un graphique montrant l'évolution des valeurs des paramètres prédits en fonction des valeurs des paramètres observés, présentant les composés et la droite de régression (Figure 39). Le modèle prédictif est de qualité si les composés sont proches de la droite de régression. Au contraire, le modèle prédictif est mauvais si les composés sont éloignés de la droite de régression, pouvant exprimer une corrélation non linéaire entre le paramètre à prédire et les descripteurs moléculaires utilisés.

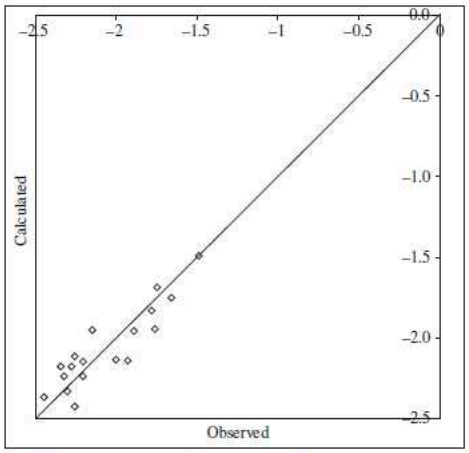


Figure 6.1 A scatter plot of the observed and calculated activity (pIC_{50}) for the MLR equation described in Section 6.2.1.3.

Figure 39 : Exemple de graphique via la régression linéaire multiple et la régression PLS (127)

En pratique, les méthodes progressives sont plutôt utilisées pour savoir si l'ajout d'un nouveau descripteur ou le retrait d'un descripteur existant dans le modèle augmente significativement la qualité de la prédiction. La méthode pas à pas est plutôt utilisée pour obtenir la meilleure combinaison linéaire de descripteurs parmi ceux existants pour obtenir la meilleure prédiction possible (133).

Régression des moindres carrés partiels (PLS)

La méthode de la régression par les moindres carrés partiels (*Partial Least Square* - PLS), est une généralisation de la régression linéaire multiple, cherchant également à prédire une variable à partir d'autres variables, donc dans le cas d'une QSAR/QSPR à prédire une activité biologique ou une propriété à partir de

descripteurs moléculaires (127,134). Avec l'utilisation d'algorithmes spécifiques, il est possible de prédire un groupe de variables à partir des descripteurs.

La régression PLS cherche à réduire le nombre de descripteurs en réalisant une transformation du même type que celle réalisée dans l'analyse en composantes principales, puis une régression linéaire en appliquant la méthode des moindres carrés en minimisant l'erreur entre les paramètres observés et les paramètres prédits.

De la même manière que pour la régression linéaire multiple, le bilan de la régression PLS peut être observé à l'aide d'un graphique montrant l'évolution des valeurs des paramètres prédits en fonction des valeurs des paramètres observés, présentant les composés et la droite de régression (Figure 39).

La régression PLS est une méthode rapide et efficace, utilisée entre autres si la régression linéaire multiple ne peut pas être employée, notamment dans le cas d'un plus grand nombre de descripteurs que de composés ou s'il existe de nombreux descripteurs très corrélés avec la variable à prédire. Toutefois, comme cette méthode se base également sur des modèles linéaires, les corrélations non linéaires ne peuvent être étudiées.

Approches basées sur la classification

Analyse discriminante linéaire (LDA)

L'analyse discriminante linéaire (*Linear Discriminant Analysis* - LDA) est une technique de classification permettant la séparation d'un jeu de molécules en deux ou plusieurs groupes (127,135). Le paramètre à prédire est forcément qualitatif, par exemple actif/non actif, ou encore toxique/peu toxique/non toxique.

La méthode LDA crée des groupes en cherchant une ou plusieurs droites permettant de séparer au mieux ces groupes. Pour cela, un algorithme permet de chercher les axes ou les ensembles d'axes qui dispersent au maximum les composés s'ils appartiennent à deux groupes différents en maximisant la variance intergroupes, et qui en même temps rassemblent tous les composés du même groupe en minimisant la variance intra-groupes (Figure 40). Enfin, la prédiction d'un nouveau composé se fait par rapport aux équations des droites obtenues.

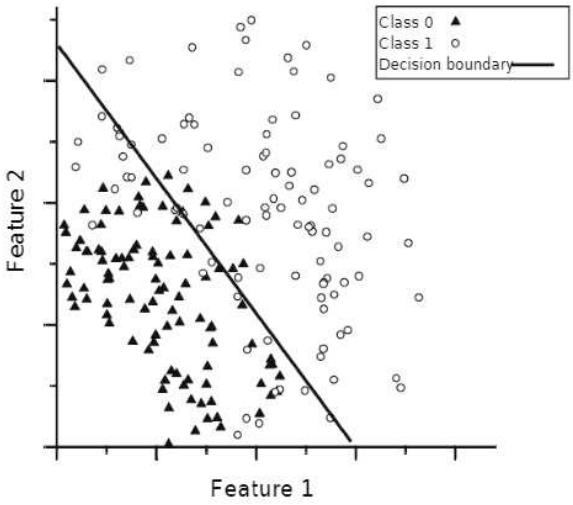


Figure 40 : Exemple d'analyse discriminante linéaire séparant deux groupes au moyen d'une frontière de décision (131)

L'analyse discriminante linéaire est une méthode simple à utiliser permettant une classification rapide de composés en deux ou plusieurs groupes en utilisant des droites, sans passer par des modèles d'apprentissage automatique. Toutefois, la compréhension de ce modèle est assez complexe du fait de son utilisation comme une boîte noire, sans considérer son fonctionnement interne. Il est également difficile de l'utiliser dans le cas où le jeu de données est de petite taille, et si la séparation ne peut pas se faire de manière linéaire demandant une complexité supérieure. Des modèles quadratiques dérivés de la LDA peuvent toutefois être utilisés pour résoudre ce dernier cas.

Régression logistique

La régression logistique est une méthode basée également sur le principe de la régression linéaire classique, permettant de déterminer les relations entre une variable qualitative binaire (par exemple actif/non actif, ou encore toxique/non toxique) et des descripteurs moléculaires (127,136). Le LDA permet de modéliser la probabilité d'être dans une des deux modalités. L'algorithme utilisé s'appuie sur une régression linéaire, à laquelle une transformation logarithmique est appliquée afin d'obtenir une courbe sigmoïde.

La régression logistique estime la probabilité que le composé soit dans un des deux groupes si la probabilité calculée est supérieure à 0.5. De la même manière que pour la régression linéaire multiple, les coefficients de régression peuvent décrire l'influence des descripteurs moléculaires. Si un descripteur possède un coefficient élevé, il affecte fortement la probabilité du résultat. Au contraire, si le coefficient est proche de zéro, le descripteur en question n'a aucune influence sur la probabilité du résultat.

La technique de la régression logistique est une technique efficace qui ne demande pas de grandes ressources de calcul et qui permet l'obtention de résultats

interprétables, notamment en connaissant l'impact des différents descripteurs moléculaires utilisés sur la prédiction. De plus, les variables à prédire n'ont pas besoin d'être d'égale variance ou normalement distribuées, ce qui en fait une méthode très souple. Toutefois, comme elle se base sur une régression linéaire, il n'est pas possible de résoudre des problèmes non linéaires. Enfin, cette technique est facilement surpassée par des algorithmes plus complexes comme les techniques d'apprentissage automatique.

Partitionnement des données – K-moyennes (k-means)

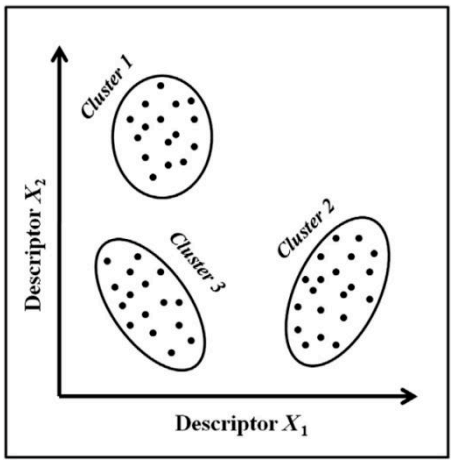


Figure 41 : Exemple de k-moyennes (127)

Le partitionnement k-moyennes (*k-means*) est une méthode de classification utilisable quand le nombre de groupes k à obtenir est connu en amont. Cet algorithme permet d'analyser un jeu de composés afin de regrouper les composés possédant des descripteurs moléculaires similaires (Figure 41) (127,137).
L'algorithme utilisé dans cette méthode commence sa recherche de groupes en choisissant au hasard des composés qui sont considérés comme les centres des différents groupes. Puis, de manière itérative, l'algorithme :

- Classe chaque composé du jeu de données en fonction du centre le plus proche.
- Recalcule un nouveau centre en fonction de la moyenne des descripteurs du groupe.

Après un certain nombre d'itérations, si l'algorithme détermine un découpage stable du jeu de données, celui-ci a convergé : les groupes déterminés sont définitifs.

L'algorithme des k-moyennes est une méthode simple et facile à comprendre et à mettre en place, mais présente des difficultés si les groupes à déterminer sont trop complexes, et ne se limite qu'à la classification en un nombre de groupes prédéterminé en amont.

Partitionnement des données – Regroupement hiérarchique (*Hierarchical clustering*)

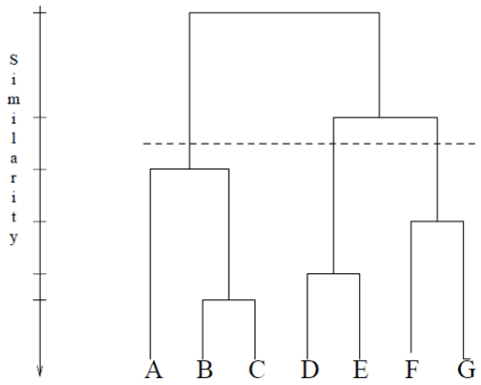


Figure 42 : Exemple de regroupement hiérarchique. La ligne en pointillée représente la division choisie. Ici, trois groupes sont obtenus, composés de A/B/C, D/E et F/G (138)

Le regroupement hiérarchique est une méthode de classification qui ne nécessite pas de connaître au préalable le nombre de groupes à former (127,139). Il existe deux approches au regroupement hiérarchique :

- L'approche ascendante, ou classification agglomérative, où chaque composé est considéré comme un groupe. Les deux groupes les plus proches sont ensuite fusionnés en un seul. Enfin, cette étape est répétée jusqu'à ce qu'il n'existe plus qu'un grand groupe unique regroupant tous les composés.
- L'approche descendante, ou classification divisive, qui fonctionne de manière opposée, où tous les composés sont regroupés dans un grand groupe unique, puis divisés successivement jusqu'à ce que chaque composé possède son propre groupe.

Le graphique obtenu à la suite de ces approches s'appelle le dendrogramme. Celui-ci débute en haut du graphique, puis se divise progressivement jusqu'à retrouver autant de groupes que de composés. En fonction du choix du nombre de groupes, ce dendrogramme sera coupé plus ou moins haut, permettant l'attribution facile de chaque composé dans des groupes retenus (Figure 42).

Afin de réaliser ces approches, il est toutefois nécessaire de connaître la distance entre deux groupes. Pour cela, des méthodes de liens sont utilisées, permettant d'associer un seul point pour chaque composé qui est considéré pour le calcul de la distance : soit les composés les plus au centre du groupe sont utilisés, soit les composés les plus éloignés, soit la moyenne, soit le centre de masse.

L'avantage de cette méthode de classification est qu'il n'est pas nécessaire de déterminer un nombre de groupes en amont. Avec le dendrogramme, il est facile de tester plusieurs possibilités pour trouver le nombre de groupes idéal. Toutefois, comme il est nécessaire de calculer pour chaque itération la distance entre toutes les

paires de groupes possibles afin de déterminer les deux groupes les plus proches ou éloignés, cette méthode nécessite beaucoup de ressources de calcul, d'espace mémoire et de temps pour les grands jeux de données. Une solution utilisable pour parer à ce défaut est de combiner l'approche des k-moyennes avec celle du regroupement hiérarchique, où la première permet de diminuer très rapidement le nombre de groupes, et d'optimiser le temps et les ressources requises pour la seconde.

Techniques d'apprentissage automatique

Machines à Vecteurs de Support (SVM)

Les machines à vecteurs de support, ou SVM (*Support Vector Machine*), sont des méthodes d'apprentissage automatique permettant de résoudre des problèmes de classification ou de régression (127,140). Le but de ces méthodes est de séparer l'ensemble des données en entrée sous la forme de classes en utilisant des frontières linéaires, appelées hyperplans, qui séparent au mieux toutes les classes. Si cette séparation linéaire est impossible, l'espace décrit par les descripteurs est transformé en un nouvel espace de plus grande dimension dans lequel il est possible de réaliser une séparation linéaire à l'aide d'une fonction mathématique, appelée fonction noyau.
Ces méthodes SVM ont une grande précision de prédiction et peuvent fonctionner sur des petits jeux de données, mais elles ne conviennent pas aux grands jeux de données de part les importantes ressources de calcul nécessaires, et posent des problèmes en cas de présence d'*outliers*.

Arbre de décision et forêts aléatoires

L'apprentissage par arbre de décision, également appelé CART (*Classification And Regression Tree*), est une méthode d'apprentissage supervisé se basant sur la création d'un arbre de décision comme modèle prédictif, permettant la prédiction de variables quantitatives (arbre de régression) ou qualitatives (arbre de classification) (127,141). L'arbre de décision possède à sa racine l'ensemble des observations, puis chaque division sépare cet ensemble en deux parties, appelés nœuds, en fonction d'un critère dépendant de la valeur d'un descripteur. A la fin de l'arbre, les feuilles terminales permettent d'attribuer une valeur prédictive, qualitative ou quantitative (Figure 43 – gauche).

Enfin, cet arbre est élagué, permettant de simplifier le modèle et d'éviter le sur-apprentissage des données. Pour cela, un algorithme d'élagage est utilisé afin de créer une suite de sous-arbres dérivés du grand arbre, permettant d'éliminer des

parties de l'arbre de décision ne fournissant que très peu de séparation des données, en vérifiant à chaque fois que les sous-arbres ne perdent pas en pouvoir prédictif.

La méthode CART est également utilisée à la base d'une autre méthode, celle des forêts aléatoires (127,142). Cette méthode permet de diminuer la variance globale d'un arbre de décision seul en s'appuyant sur plusieurs arbres de décisions pour améliorer les performances. Une forêt aléatoire est composée d'un ensemble d'arbres de décision, générés en parallèle, pouvant chacun prédire une réponse. La réponse globale de la forêt aléatoire est établie sur cet ensemble de réponses, soit en votant pour la classe la plus populaire dans le cas d'une classification, soit en combinant les réponses sous la forme d'une moyenne dans le but d'obtenir une estimation quantitative dans le cas d'une régression (Figure 43 – droite).

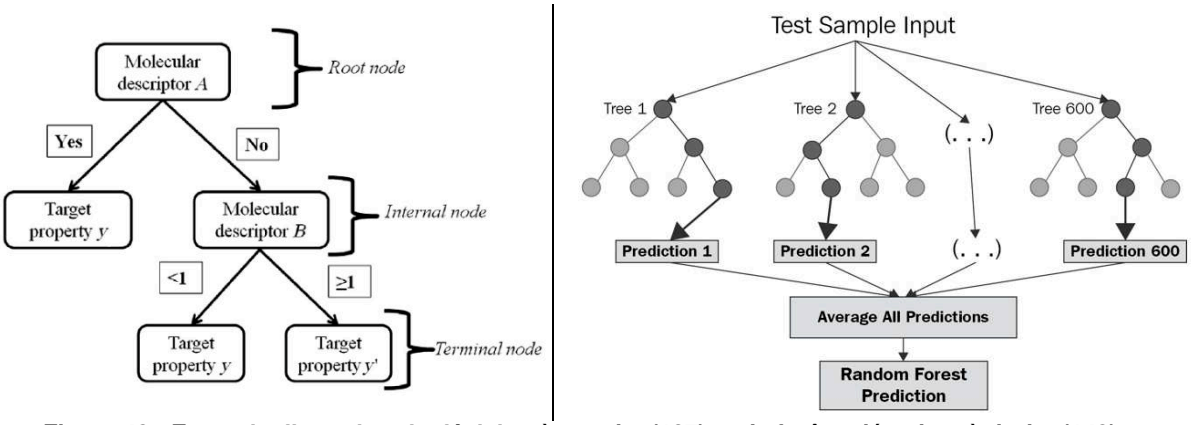


Figure 43 : Exemple d'un arbre de décision à gauche (127) et de forêts aléatoires à droite (143)

Réseaux de neurones artificiels

La méthode des réseaux de neurones artificiels s'inspire du fonctionnement du cerveau humain et de la mémoire associative, où une somme d'informations en entrée est traitée ensemble afin de fournir une réponse (127,144).

Un réseau de neurones artificiels se présente sous la forme de connexions entre plusieurs couches de neurones représentant une ou plusieurs fonctions algébriques : la première couche est la couche d'entrée, la dernière couche celle de la réponse, et les couches intermédiaires les couches cachées. L'information circule dans un seul sens, allant des neurones d'entrée, ici les descripteurs moléculaires, jusqu'aux neurones de sortie, ici la réponse à prédire (Figure 44).

La spécificité des réseaux de neurones artificiels est qu'ils ne peuvent pas être programmés directement pour effectuer leur action : le réseau se modifie de lui-même en fonction du résultat de ses actions, permettant son apprentissage et la résolution de problèmes sans algorithme. Cet apprentissage peut être de trois types différents :

- L'apprentissage supervisé, où le réseau de neurones artificiels optimise le modèle à partir d'une partie des données de manière itérative, en se modifiant pour être capable de prédire la réponse observée.
- L'apprentissage non-supervisé, où le réseau de neurones artificiels analyse l'ensemble des données en entrée et estime son éloignement du résultat souhaité, de manière itérative. Le réseau s'adapte ensuite pour augmenter la précision du modèle.
- L'apprentissage par renforcement, où le réseau de neurones artificiels est récompensé pour les résultats positifs, mais sanctionné pour les résultats négatifs, pour permettre au modèle d'apprendre de ses propres erreurs.

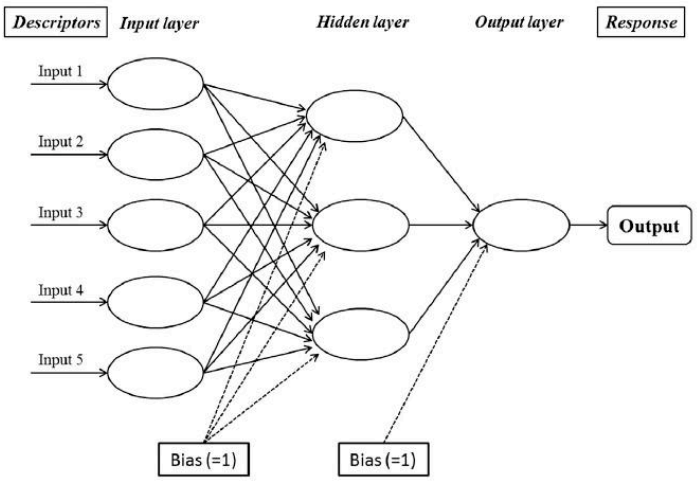


Figure 44 : Exemple d'un réseau de neurones artificiels de trois couches (127)

4.1.4. 3D-QSAR

Le principe de base des méthodes QSAR/QSPR est que la différence d'activité biologique ou des valeurs d'une propriété est causée par les différences structurales des composés (127). Dans les études QSAR/QSPR vues précédemment, les paramètres prédits sont corrélés avec les descripteurs moléculaires choisis, mais les propriétés structurales tridimensionnelles ne sont que très peu prises en compte au travers de l'utilisation de descripteurs 3D difficiles à obtenir et à interpréter. Les méthodes 3D-QSAR/QSPR ont été développées afin de mieux exploiter les propriétés 3D des composés, et d'établir une relation entre une activité biologique/propriété à prédire et les conformations 3D des composés. Elles permettent également l'obtention de modèles interprétables visuellement, et d'indications sur les modifications structurales qui peuvent potentiellement améliorer la puissance des composés.

Les méthodes 3D-QSAR/QSPR reposent sur le calcul de propriétés stériques selon les conformations des composés et de propriétés électrostatiques en

appliquant un champ de forces pour calculer des potentiels d'énergie. Ces propriétés sont obtenues en utilisant une grille tridimensionnelle dans laquelle les composés sont placés afin de calculer leurs champs d'interactions. Par la suite, des méthodes QSAR/QSPR vues précédemment sont appliquées pour aboutir à une prédiction.

La méthode CoMFA (*Comparative Molecular Field Analysis*), développée à la fin des années 80, est l'ancêtre de toutes les méthodes 3D-QSAR/QSPR développées jusqu'à maintenant (10,127). A partir d'un jeu de composés partageant le même mode de liaison, ceux-ci sont alignés dans leurs conformations bioactives sur un ou plusieurs composés de référence, idéalement possédant une conformation 3D déterminée par cristallographie. L'ensemble des composés est ensuite séparé en un jeu d'apprentissage et un jeu de test. Tous les composés du jeu d'apprentissage sont ensuite placés dans une grille tridimensionnelle suffisamment grande pour englober tous les composés dans toutes les directions de l'espace, avec une taille de grille de l'ordre de 1 ou 2 Å. Des sondes, généralement des atomes de carbone hybridé sp3 portant une charge positive, sont placées à chaque nœud de grille. Les énergies d'interactions stériques et électrostatiques entre les composés et ces sondes sont ensuite calculées et enregistrées.

Après avoir éliminé celles possédant un écart-type faible, ces données sont utilisées en tant que descripteurs pour des méthodes QSAR/QSPR. Le jeu de données à analyser étant très grand, la méthode PLS est favorisée, ou une réduction des descripteurs au préalable par une méthode PCA est utilisée avant l'application d'autres modèles. Dans le cas de l'application d'une méthode PLS, il est possible de projeter les coefficients obtenus lors de cette méthode au niveau de chaque sonde dans l'espace 3D des composés, afin d'obtenir des informations visuellement interprétables sur l'impact des propriétés stériques et électrostatiques sur le paramètre à prédire (Figure 45).

De nombreuses méthodes ont été dérivées de la CoMFA, comme la CoMSIA (*Comparative Molecular Similarity Indices Analysis*) utilisant des notions de similarité entre composés.

Les méthodes 3D-QSAR/QSPR permettent l'interprétation visuelle des informations obtenues à partir du modèle prédictif, et l'obtention d'indices sur les possibilités d'amélioration des composés. Toutefois, il est nécessaire d'avoir au préalable un grand jeu de composés possédant un mode de liaison identique pour permettre des prédictions significatives. Elles nécessitent également l'alignement de composés entre eux, donc la connaissance de conformations 3D de composés de référence obtenues au travers de structures cristallographiques.

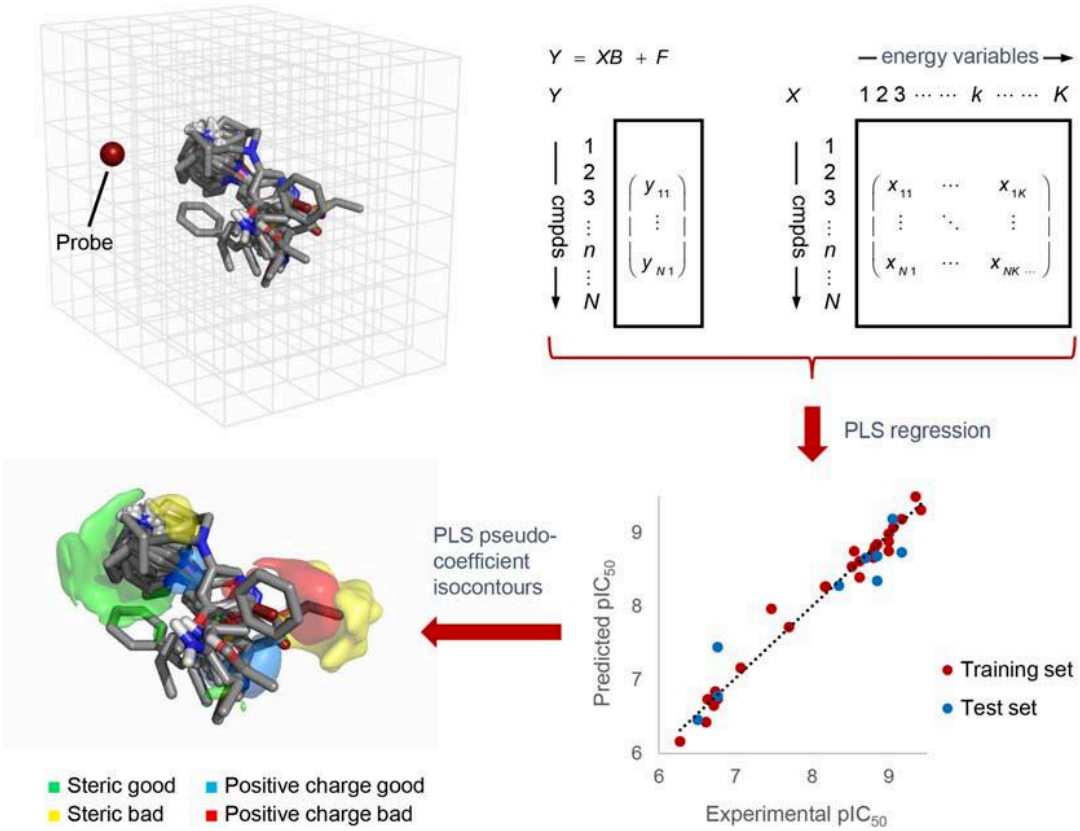


Figure 45 : Procédure d'une méthode CoMFA (145)

4.1.5. Validation des modèles

La validation d'un modèle QSAR/QSPR nécessite d'évaluer la qualité du modèle avec le jeu de données de test, en réalisant la prédiction du paramètre d'intérêt et en le comparant à ce même paramètre observé expérimentalement, ou en mesurant l'impact de l'influence des composés ou de l'aléatoire sur le pouvoir prédictif du modèle.

Coefficients et tests statistiques

Pour valider la qualité d'un modèle prédictif QSAR/QSPR, il est possible de calculer certains coefficients permettant d'évaluer la qualité du modèle (127):

- Coefficient de corrélation R : Ce coefficient permet de déterminer la part de la variance de l'activité ou de la propriété expliquée par le modèle QSAR/QSPR, en comparant le paramètre observé expérimentalement au paramètre prédit par le modèle. Une bonne corrélation entre les deux activités est observée quand R tend vers 1. Au contraire, une non-corrélation est observée quand R tend vers 0.
- Coefficient de détermination R^2 : Le coefficient de détermination permet d'obtenir une information sur l'adéquation entre le modèle et les données observées. Pour un R^2 égal à 0.9, 90% des activités ou propriétés prédites peuvent être expliquées par le modèle QSAR/QSPR, et 10% sont expliquées

par le hasard. Plus R^2 tend vers 1, plus les paramètres observés et prédits sont corrélées, et plus R^2 tend vers 0, plus elles ne sont pas corrélées. Enfin, un modèle prédictif est qualifié d'acceptable quand le R^2 est supérieur à 0.5, c'est à dire quand les prédictions obtenues sont plus expliquées par le modèle QSAR/QSPR que par l'aléatoire.

- Biais : Le biais permet d'évaluer si les prédictions réalisées par le modèle sont précises, ou si le modèle a tendance à surestimer ou sous-estimer les valeurs de la variable prédite. Plus le biais est proche de 0, plus la prédiction est de qualité.
- Erreur quadratique moyenne (RMSE - *Root-Mean-Square Error*) : Cet indice permet d'obtenir une information sur la dispersion et la variabilité de la qualité de la prédiction, donc de la variance du modèle. Afin de rendre cet indice plus facilement interprétable, celui-ci est souvent normalisé pour obtenir un pourcentage de la variance du modèle par rapport à la moyenne des observations.

Des tests statistiques peuvent également être utilisés, comme le test de Student pour évaluer la pertinence d'un ou de plusieurs descripteurs dans le modèle en supposant que ces descripteurs ne soient pas significatifs.

Validation interne

La stabilité de la prédiction obtenue par un modèle prédictif est évaluée en prenant en compte l'influence de chaque composé sur celui-ci. Le principe global d'une validation interne consiste à mettre de côté un nombre déterminé de composés, puis à créer un nouveau modèle avec les composés restants. Réalisé de manière répété, et en évaluant pour chaque nouveau modèle son coefficient de détermination R^2, il est possible d'évaluer le pouvoir de prédiction interne du modèle prédictif exprimé par le coefficient de détermination de la validation interne, noté Q^2, calculé de manière similaire au R^2. Plus le Q^2 se rapproche de 1, plus le pouvoir de prédiction du modèle est optimal. Pour qu'un modèle soit qualifié d'acceptable, le Q^2 doit être supérieur à 0.5 (146).

La procédure de validation croisée (*cross-validation* - CV) est une des méthodes les plus employées pour déterminer la stabilité d'un modèle prédictif et tester l'influence de chaque composé sur le modèle final (127,147). Deux approches sont utilisées :

- La validation croisée à k-blocs (*k-cross cross-validation*), où le jeu de données global est divisé en k parties. Une de ces parties est mise de côté pour faire parti du jeu de test, et les autres parties sont utilisées comme jeu d'apprentissage pour créer le modèle prédictif. Répété k fois en faisant varier les jeux utilisés, un coefficient de détermination de la validation interne R^2_{CV} également appelé Q^2, est calculé après k itérations.

- La validation croisée d'un contre tous (*leave-one-out cross-validation*), qui est un cas particulier de la validation croisée à k-blocs, où il existe autant de parties que de composés dans le jeu de données global. Pour chaque itération, le jeu de test n'est formé que d'un seul composé, et le jeu d'apprentissage inclut tous les autres composés. De la même manière, un coefficient de détermination de la validation interne Q^2 sera calculé après k itérations.

Les valeurs de Q^2 obtenues à l'aide de procédures de validation croisée pouvant être surestimée dans le cas de corrélations dues au hasard, une validation par test de randomisation peut être utilisé en complément (127,147). Cette approche, répétée plusieurs fois, mélange aléatoirement tous les paramètres observés expérimentalement des composés présents dans le jeu d'apprentissage, un nouveau modèle prédictif est créé, et un nouveau coefficient de détermination Q^2_{rand} déterminé. Si ce Q^2_{rand} est petit, le modèle prédictif peut être validé, car il garanti que le modèle développé n'a pas été obtenu par hasard.

Validation externe

Pour évaluer la puissance de la prédiction du modèle obtenu, il est nécessaire de tester sa capacité à prédire l'activité biologique ou les propriétés de composés présents dans un jeu indépendant n'ayant pas été utilisé pour créer ce modèle (127). Le jeu de test, créé au début de la procédure QSAR/QSPR, a été formé dans ce but. Cette puissance de prédiction est basée sur le calcul d'un coefficient de détermination R^2_{test} à partir des paramètres observés expérimentalement et des paramètres prédits par le modèle. Comme pour le jeu d'apprentissage, le R^2_{test} doit tendre vers 1, en étant supérieur à 0.5, pour avoir un modèle prédictif de qualité.

Domaine d'applicabilité

Le modèle statistique QSAR/QSPR étant créé à partir d'un jeu de composés limité, celui-ci ne peut pas être considéré comme un modèle de prédiction universel permettant de prédire une variable pour tous les composés imaginables. Pour délimiter le champ d'action utilisable du modèle, il est nécessaire de définir son domaine d'applicabilité, c'est à dire l'espace chimique dans lequel un composé pourra être prédit avec fiabilité. Il existe différentes méthodes de détermination d'un domaine d'applicabilité, dont (147,148):

- L'étude de la variation des descripteurs d'un nouveau composé parmi la moyenne de ces descripteurs du jeu d'apprentissage permettant de calculer un index de similarité ainsi qu'une valeur levier, estimant leur ressemblance au jeu d'apprentissage.
- L'utilisation d'autres techniques statistiques en complément, comme l'Analyse en Composantes Principales (PCA) pour étudier visuellement l'emplacement

du nouveau composé par rapport au jeu d'apprentissage, ou l'utilisation d'une approche basée sur la classification.

- L'utilisation de techniques mathématiques permettant de mesurer la distance par rapport au modèle, comme le calcul de la déviation standard de l'ensemble des prédictions (STD), la mesure de l'incertitude de la prédiction pour des modèles de classification (CLASS-LAG).

4.1.6. Exemples d'applications

Une première application des méthodes QSAR/QSPR peut être réalisée pour découvrir et développer de nouvelles molécules bioactives. Dans un article publié en 2020, une équipe a utilisé entre autres des méthodes QSAR et QSPR afin de concevoir de nouveaux analogues de l'Énalapril, un médicament utilisé dans le traitement de l'hypertension artérielle en ciblant l'enzyme de conversion. Ils ont pu obtenir des composés ayant une meilleure affinité pour son enzyme, une plus faible concentration inhibitrice médiane (IC50) et une plus faible toxicité (149).

Les principales méthodes utilisées ont été la régression linéaire multiple (MLR) ainsi que la régression des moindres carrés partiels (PLS), permettant la prédiction d'une IC50 pour l'activité anti-hypertensive à partir de descripteurs moléculaires. Après création des modèles, seuls trois descripteurs ont été choisis, corrélés avec l'IC50 des composés (Figure 46).

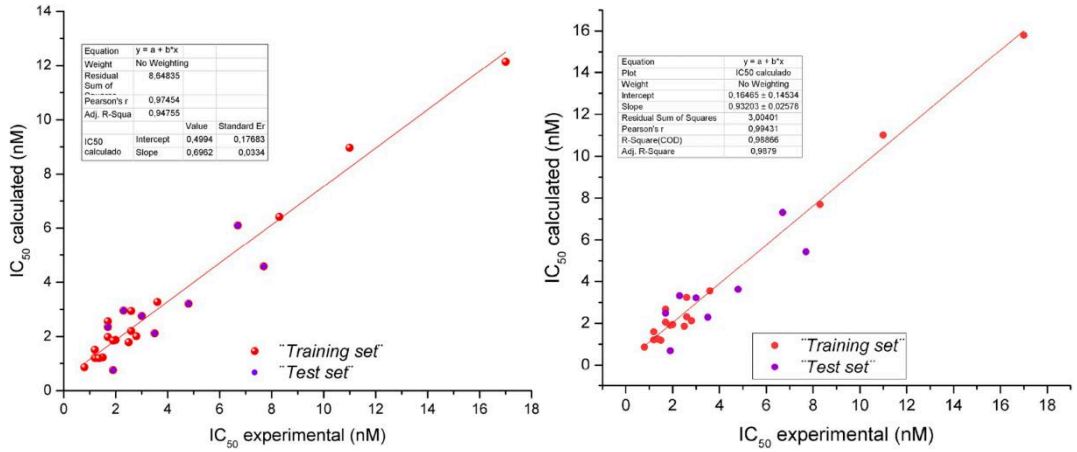


Figure 46 : Relations entre l'IC50 observée expérimentalement et prédite de composés commerciaux et non-commerciaux d'inhibiteurs de l'enzyme de conversion à partir de (A) du modèle MLR (B) du modèle PLS. En rouge sont représentés les composés du jeu d'apprentissage et en violet les composés du jeu de test (149)

Les modèles MLR et PLS obtenus montrent de hauts coefficients de corrélations R^2 supérieurs à 0.95 (indiquant que le modèle prédit 95% de la variation de l'IC50, les 5% restants étant dû au hasard), des coefficients de corrélation en validation croisée aussi importants et des écarts-types faibles, ce qui indique que les modèles prédictifs possèdent une capacité prédictive élevée.

Une fois les modèles réalisés, plus de 200 analogues de l'Énalapril ont été testés via les modèles QSAR créés, via des amarrages moléculaires pour calculer l'affinité de liaison, et via l'utilisation de modèles QSPR propriétaires utilisant le logiciel PreADMET (150) pour prédire leur toxicité (Figure 47). Deux composés ont été mis en évidence, les composés 20 et 21, possédant des concentrations inhibitrices médianes IC50 plus de 100 fois inférieures à celle de l'Énalapril, signifiant qu'il faut 100 fois moins de composés 20 ou 21 pour obtenir le même niveau d'inhibition de l'enzyme de conversion que l'Énalapril. Ces composés ont également montré une toxicité pour l'environnement 100 fois inférieure (*Algae_at* et *Daphnia_at*). Enfin, le composé 20 ne présente pas de cancérogénicité pour des lignées cellulaires de rat et de souris, mais le composé 21 présente une activité cancérigène pour des lignées cellulaires de rat sans être cancérigène pour des lignées cellulaires de souris.

		Enalapril	6	14	20	21
MLR QSAR Model	Predicted IC_{50} nM (MLR)	1.2 (1.2)[a]	1.3	1.1	0.009	0.009
Molecular Docking	Intermolecular Energy (kcal/mol)	−7.50	−10.11	−9.48	−8.90	−9.30
Toxicity PreADME	Algae_at	0.0263535	0.0547759	0.0443204	0.0002601	0.0005878
	Carcino_Rat	Positive	Positive	Negative	Positive	Negative
	Carcino_Mouse	Negative	Negative	Negative	Negative	Negative
	Daphnia_at	0.132963	0.521971	0.411032	0.000911	0.0023671
	Ames_test	Non-mutagenic	Non-mutagenic	Non-mutagenic	Non-mutagenic	Non-mutagenic

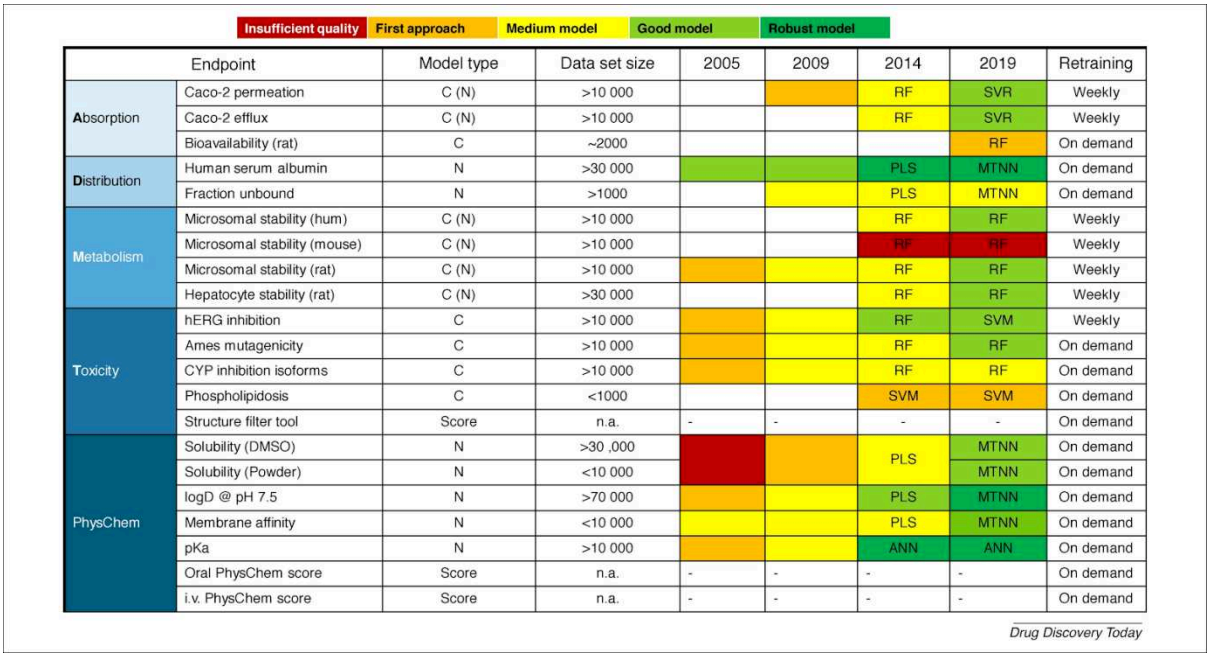
Tableau 47 : Résumé de l'activité antihypertensive prédite par le modèle QSAR de régression linéaire multiple, de l'affinité de liaison et de toxicités évaluées (149)

Les approches in silico réalisées dans cette étude mettent donc en évidence la possibilité de découvrir et de développer de nouvelles molécules anti-hypertensives possédant des activités biologiques supérieures aux molécules actuellement utilisées en thérapeutique, tout en étant moins toxiques pour l'environnement.

Une seconde application des méthodes QSAR/QSPR est fournie par la prédiction de nombreux paramètres pharmacocinétiques et toxicologiques. Au cours des 20 dernières années, une plateforme de prédiction de propriétés pharmacocinétiques et toxicologiques a été créée au sein de l'entreprise pharmaceutique Bayer afin de développer des modèles de prédiction QSPR pour les appliquer lors des processus de conception de médicaments (151). Ces modèles sont très souvent utilisés pour aider à la sélection et à la conception de nouvelles touches pharmacologiques, ainsi que leur optimisation.

Parmi ces modèles, les méthodes les plus utilisées sont les machines à vecteurs de support (SVM), les forêts aléatoires (RF) et les réseaux de neurones (ANN/MTNN), des techniques d'apprentissage automatique principalement

appliquées dans le cadre de régressions (Figure 48). Tous ces modèles sont utilisés au sein de l'entreprise via une plateforme de données développée en interne regroupant toutes les molécules en cours de test ou déjà testées, les molécules présentes dans des chimiothèques à disposition, ainsi que des outils de visualisation des données prédites (Figure 49).



	Endpoint	Model type	Data set size	2005	2009	2014	2019	Retraining
Absorption	Caco-2 permeation	C (N)	>10 000			RF	SVR	Weekly
	Caco-2 efflux	C (N)	>10 000			RF	SVR	Weekly
	Bioavailability (rat)	C	~2000				RF	On demand
Distribution	Human serum albumin	N	>30 000			PLS	MTNN	On demand
	Fraction unbound	N	>1000			PLS	MTNN	On demand
Metabolism	Microsomal stability (hum)	C (N)	>10 000			RF	RF	Weekly
	Microsomal stability (mouse)	C (N)	>10 000			RF	RF	Weekly
	Microsomal stability (rat)	C (N)	>10 000			RF	RF	Weekly
	Hepatocyte stability (rat)	C (N)	>30 000			RF	RF	Weekly
Toxicity	hERG inhibition	C	>10 000			RF	SVM	Weekly
	Ames mutagenicity	C	>10 000			RF	RF	On demand
	CYP inhibition isoforms	C	>10 000			RF	RF	On demand
	Phospholipidosis	C	<1000			SVM	SVM	On demand
	Structure filter tool	Score	n.a.	-	-	-	-	On demand
PhysChem	Solubility (DMSO)	N	>30 ,000			PLS	MTNN	On demand
	Solubility (Powder)	N	<10 000				MTNN	On demand
	logD @ pH 7.5	N	>70 000			PLS	MTNN	On demand
	Membrane affinity	N	<10 000			PLS	MTNN	On demand
	pKa	N	>10 000			ANN	ANN	On demand
	Oral PhysChem score	Score	n.a.	-	-	-	-	On demand
	i.v. PhysChem score	Score	n.a.	-	-	-	-	On demand

Figure 48 : Portefeuille de modèles de prédiction in silico développés par Bayer afin de prédire des propriétés pharmacocinétiques, toxicologiques et physico-chimiques. La qualité des modèles est indiquée par un code couleur allant du rouge pour les modèles possédant une qualité de prédiction insuffisante, jusqu'au vert foncé pour les modèles robustes possédant un pouvoir de prédiction de haute qualité (151)

Figure 49 : Outil interne à Bayer permettant notamment l'accès aux modèles prédictifs, aidant tous les employés à la conception de médicaments (151)

4.2. Pharmacophore

Un pharmacophore est défini comme l'ensemble des propriétés moléculaires nécessaires à la réalisation des interactions optimales avec une cible biologique afin de réaliser ou bloquer une réponse biologique (152). Un pharmacophore ne représente donc pas des motifs structuraux spécifiques ou l'association de groupements fonctionnels, mais utilise des descriptions abstraites des propriétés stériques et électroniques essentielles du ligand pour réaliser les interactions optimales avec son récepteur. De plus, des composés possédant un même pharmacophore interagissent de la même manière avec un même site de liaison, faisant d'un pharmacophore une référence afin de trouver de nouvelles touches pharmacologiques.

Le première modèle de pharmacophore a été décrit par Beckett en 1963, et permettait de définir les distances entre les acides aminés du site de liaison de récepteurs muscariniques (153). Celui-ci a été vérifié par Kier en 1971, qui lui donne le nom de « pharmacophore » (154).

L'approche pharmacophorique *in silico* se base sur la définition de points pharmacophoriques considérés comme des groupes fonctionnels, comme des donneurs ou des accepteurs de liaisons hydrogène, des zones d'interactions électrostatiques, des cycles aromatiques ou des zones hydrophobes. Dans le cas de l'approche s'appuyant sur les ligands, et ne connaissant pas la structure de la cible, des composés actifs (idéalement rigides) sont pris comme référence (32,60). Cette technique peut également être utilisée dans une approche basée sur les structures, en se fondant sur une conformation tridimensionnelle de la cible étudiée.

Les modèles pharmacophoriques obtenus permettent, entre autres, l'identification de nouvelles touches pharmacologiques de familles chimiques différentes respectant un mode de liaison spécifique, ou l'optimisation des chefs de file en rajoutant des groupements fonctionnels de manière rationnelle (155). Les pharmacophores peuvent également être utilisés afin de prédire les interactions de composés d'intérêt avec des cibles non souhaitées, entraînant de potentiels effets indésirables (60).

4.2.1. Pharmacophores 2D

Les méthodes de création de pharmacophores 2D se reposent sur des descripteurs 2D, majoritairement basées sur la connectivité des atomes, pour définir et comparer des points pharmacophoriques. Elles n'utilisent donc pas les coordonnées atomiques des ligands dans l'espace, permettant l'obtention de méthodes rapides et ne demandant que peu de ressources de calculs.

L'une des méthodes pionnières des pharmacophores 2D est la méthode CATS 2D (*Chemically Advanced Templace Search*) (156). Celle-ci, créée en 1999, part d'un graphe moléculaire, avec la définition de points pharmacophoriques : lipophile, aromatique, accepteur ou donneur de liaison hydrogène, ou encore ionisable négatif ou positif (Figure 50 – A/B). Ensuite, une énumération de toutes les distances entre chaque point pharmacophorique est réalisée, en mesurant le nombre de liaisons séparant un point de tous les autres points pharmacophoriques. Cette énumération permet l'obtention d'un tableau regroupant les 21 combinaisons possibles de type de points pharmacophoriques, et le nombre de paires observées pour chaque distance de liaisons (Figure 50 – C). Enfin, ce tableau est normalisé en divisant les colonnes par la somme des éléments la composant pour obtenir des proportions entre 0 et 1. Ceci permet l'obtention d'un vecteur pharmacophorique pour une molécule, qui est utilisé afin de le comparer à d'autres molécules en calculant notamment une distance euclidienne.

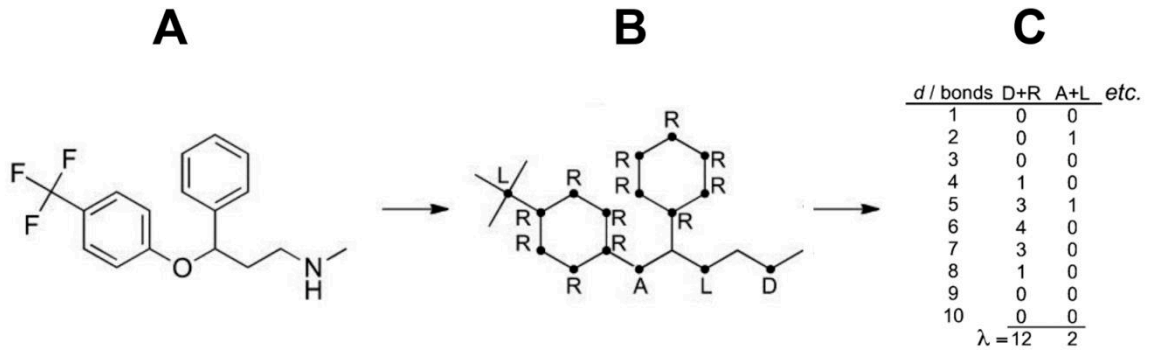


Figure 50 : Représentation schématique d'une méthode CATS – Adapté de (156)

Une autre approche pionnière est la méthode des *feature trees*, créée en 1998 (157). Cette approche permet de représenter la structure d'un ligand sous la forme d'un arbre, où chaque nœud représente un point pharmacophorique, pour décrire les propriétés chimiques des différents groupes d'atomes (Figure 51). La similarité entre deux ligands est évaluée via la superposition de deux *feature trees* en essayant de superposer un maximum de nœuds identiques, avec l'attribution d'un score en fonction du nombre et de la qualité de cet alignement.

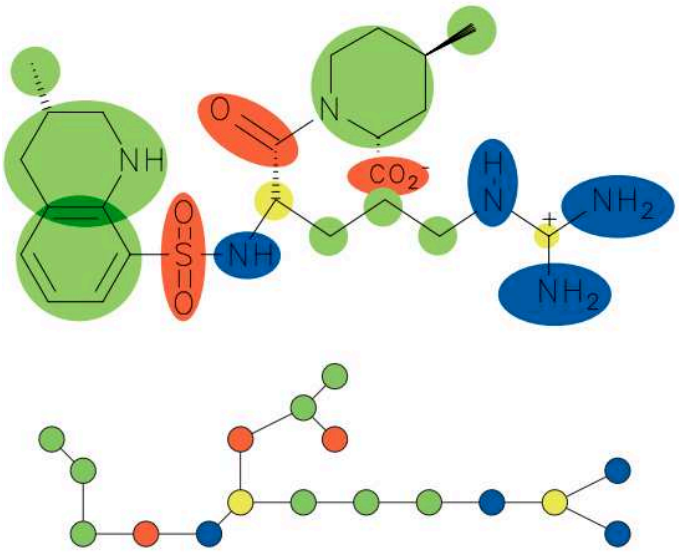


Figure 51 : Représentation d'une méthode *Feature Trees* (157)

Des méthodes dérivées de ces deux approches ont été développées. La méthode des *Smilog keys*, à mi-chemin entre l'approche pharmacophorique CATS 2D et la recherche de similarité à partir de descripteurs 2D, permet l'obtention de vecteurs comparant plusieurs ligands. La méthode des graphes réduits (*reduced graphs*), et graphes réduits étendus (*extended reduced graphs*) représentent un ligand sous la forme d'un arbre simplifié, pour obtenir une représentation plus globale et simplifiée.

Par leur simplicité d'exécution et par la faible puissance de calcul nécessaire, les méthodes pharmacophoriques 2D sont utiles afin de trouver des ligands similaires dans un grand jeu de données. Cependant, elles identifient des ligands similaires à la référence utilisée, et ne prennent pas en compte la flexibilité conformationnelle. Des méthodes pharmacophoriques 3D ont donc été développées en conséquence.

4.2.1. Pharmacophores 3D

Les méthodes permettant de créer des modèles pharmacophoriques 3D prennent en compte l'arrangement tridimensionnel des caractéristiques chimiques essentielles d'une molécule afin de déterminer des points pharmacophoriques (X). Les relations spatiales entre ces points sont spécifiées sous la forme de distances (Figure 52 – droite), d'angles, de plans, de centroïdes ou de zones de tolérance dans l'espace représentées sous la forme de sphères (Figure 52 – gauche) (An introduction to chemoinformatics). Afin de générer un modèle pharmacophorique 3D, plusieurs étapes sont nécessaires.

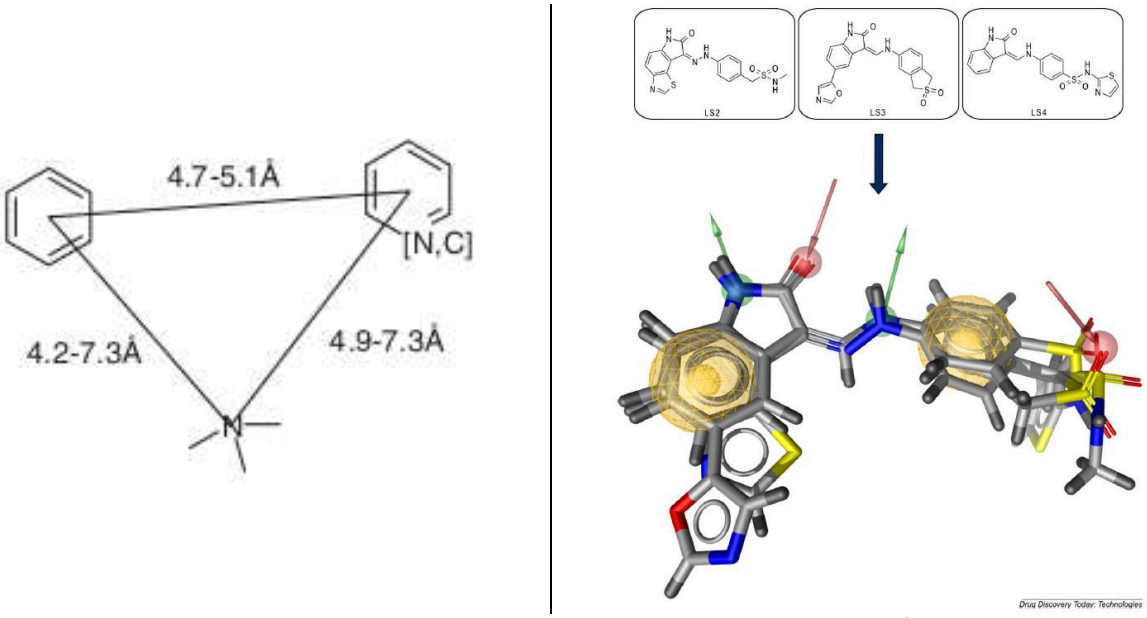


Figure 52 : Représentation d'un pharmacophore 3D sous la forme de distances à gauche (32,158) ; Représentation d'un pharmacophore 3D sous la forme de sphères de tolérance à droite (60)

Identification de ligands de référence

A partir de bases de données de composés chimiques, des ligands de référence présentant une activité biologique pour la cible biologique étudiée sont sélectionnés. Ces ligands sont à la base du modèle pharmacophorique qui sera obtenu. Leur choix a un impact direct sur la qualité du modèle. Si des ligands ne se fixent pas à un même site de liaison, ceux-ci peuvent fausser la construction du modèle ou la rendre impossible (32). Il est également nécessaire d'utiliser un jeu de ligands le plus diversifié possible, afin de pouvoir mettre en évidence les points pharmacophoriques les plus importants, critiques pour la liaison (155).

De la même manière que pour les méthodes QSAR/QSPR, le jeu des ligands de référence est séparé en deux, une sous-partie permettant la création du modèle pharmacophorique, et une seconde permettant sa validation (50).

Recherche conformationnelle

Une fois les ligands de référence sélectionnés, leurs conformations tridimensionnelles doivent être générées. En effet, les conformations adoptées par les ligands de référence pour réaliser les interactions optimales étant le plus souvent inconnues, il est nécessaire de générer de manière exhaustive l'espace conformationnel de chaque ligand. Deux approches peuvent être adoptées :

- Soit il est nécessaire de générer les conformations au préalable. Certains modèles peuvent nécessiter de nombreuses conformations différentes par ligand afin de garantir la précision des résultats. Une étape de minimisation puis une étape d'élimination des conformations redondantes sont réalisées pour l'obtention d'un nombre limité de conformations, représentatives de l'espace conformationnel global.

- Soit les conformations sont générées en parallèle de la création du modèle pharmacophorique, à l'aide de conformations aléatoires obtenues par exemple via l'utilisation d'un algorithme génétique.

Identification des points pharmacophoriques

Ensuite, à partir des ligands de référence, des points pharmacophoriques pertinents pour l'obtention d'un modèle pharmacophorique sont identifiés. Trois niveaux de points pharmacophoriques peuvent être utilisés, en fonction de la complexité souhaitée (155):

- Les points pharmacophoriques basés sur les atomes sont la manière la plus simple de définir un point, en utilisant les coordonnées et le type d'un atome.
- Les points basés sur la topologie sont issus de groupes d'atomes créés en fonction de leur topologie, comme par exemple un cycle aromatique spécifique comme un groupe phényle.
- Les points basés sur les fonctions regroupent des atomes en caractéristiques fonctionnelles chimiques, qui décrivent un type d'interaction important pour la liaison entre le ligand et son site de liaison. Les plus courants sont le groupe hydrophobe, le cycle aromatique, l'accepteur ou le donneur de liaison hydrogène, ainsi que les fonctions acides ou basiques chargées à pH physiologique (pH 7). En règle générale, ces points pharmacophoriques sont dotés d'une sphère de tolérance et un centre, ainsi que parfois d'une direction d'interaction sous la forme d'un vecteur. Ils sont également pondérés, permettant de décrire l'importance du point pharmacophorique.

Création de modèles pharmacophoriques

L'étape suivante est la création d'un ou de plusieurs pharmacophores regroupant des points pharmacophoriques communs entre les ligands de référence. La méthode la plus populaire est celle des structures communes maximales (*Maximum Common Features*), permettant de réaliser l'alignement des points pharmacophoriques identifiés de tous les ligands de référence. Cet alignement est soit réalisé en superposant les points pharmacophoriques, soit en utilisant des descripteurs moléculaires.

D'autres méthodes existent, mais sont généralement spécifiques à un logiciel, utilisant ses propres méthodes de recherche et d'alignement des points pharmacophoriques.

Validation et évaluation des modèles pharmacophoriques

Pour permettre l'identification des meilleurs modèles pharmacophoriques parmi ceux générés, un score est attribué à l'aide d'une fonction de score. Cette fonction prend en compte, entre autres, le nombre et la qualité de la superposition des différents points pharmacophoriques, l'énergie conformationnelle, le volume de recouvrement des ligands, ou encore la distance entre les points pharmacophoriques des ligands et du modèle (via notamment le calcul de RMSD) (60). Des critères peuvent être également pris en compte, comme la spécificité de points pharmacophoriques, retrouvée dans une très grande quantité de composés chimiques et donc peu discriminante (159).

Une fois les meilleurs modèles pharmacophoriques sélectionnés en fonction de leur score et par une observation visuelle, ces modèles sont validés via l'utilisation d'un jeu de ligands de validation indépendant. Ce jeu de validation est composé de la sous-partie des ligands actifs de référence mise de côté lors de l'étape de préparation et non utilisée lors de la création des modèles, ainsi que de composés inactifs. Un ensemble de leurres peut également être utilisé, formé à partir de composés similaires aux ligands actifs du jeu de validation, possédant des propriétés physico-chimiques proches (50). Tous ces jeux de composés sont filtrés par le modèle pharmacophorique afin d'obtenir une prédiction sur leur pertinence, et pour évaluer la qualité du modèle au travers du calcul d'indicateurs.

L'indicateur le plus utilisé est le facteur d'enrichissement, qui décrit le nombre de composés actifs retrouvés par le modèle pharmacophorique comparé au nombre d'actifs trouvés par une méthode aléatoire (50,160).

Un autre indicateur très utilisé est le score GH (Güner-Henry) (161), qui correspond à deux facteurs opposés permettant d'évaluer la qualité du modèle pharmacophorique : la précision et la sensibilité. La précision est la fraction des composés du jeu de validation correctement prédits parmi tous les composés du jeu. La sensibilité est la fraction des composés actifs du jeu de validation bien prédits comme actifs, parmi les composés prédits comme étant actifs (160).

Criblage de chimiothèques à partir d'un modèle pharmacophorique

Une fois l'obtention d'un modèle pharmacophorique suffisamment sélectif et discriminant obtenu, celui-ci peut être utilisé pour permettre l'identification de nouvelles touches pharmacophoriques au travers du criblage de chimiothèques (155).

De la même manière que pour les ligands de référence, l'espace conformationnel de chaque composé de la chimiothèque sélectionnée est exploré de manière exhaustive au travers de l'obtention de nombreuses conformations. Il est également possible d'utiliser une approche alternative qui consiste en l'ajustement

dynamique de la conformation d'un composé en utilisant un algorithme génétique lors de l'étape suivante, mais cette approche demande des temps et des ressources de calculs plus importants (155).

Puis, un alignement de ces conformations sur le modèle pharmacophorique déterminé est réalisé afin d'évaluer la superposition des différents points pharmacophoriques des composés testés avec le modèle. L'évaluation du nombre de points pharmacophoriques communs, avec le calcul de la distance entre les points pharmacophoriques à l'aide du calcul d'un RMSD, permet d'identifier des composés respectant ce pharmacophore (60,162).

L'étape d'alignement demandant des ressources de calcul très importantes, il est parfois nécessaire de réaliser une étape de filtrage en amont, notamment pour les grandes chimiothèques. Ce filtrage consiste à éliminer les composés qui ne pourront jamais respecter le pharmacophore au vu du nombre et du type de leurs points pharmacophoriques (162). Seuls les composés présentant au minimum le même nombre de points pharmacophoriques que le modèle pharmacophorique utilisé sont gardés. Il est également possible d'utiliser en amont une méthode pharmacophorique 2D, afin de réaliser une présélection demandant beaucoup moins de temps et de ressources (162).

4.2.2. Avantages et inconvénients de l'approche pharmacophorique

Les approches pharmacophoriques 2D et 3D permettent de mettre en évidence des propriétés physico-chimiques importantes pour le mode de liaison à une cible pharmacologique étudiée, ainsi que l'obtention d'une activité biologique. Toutefois, ces approches ne permettent pas la quantification de cette activité biologique (162). Une solution possible pour passer outre ce défaut est la génération de modèles QSAR à partir des pharmacophores obtenus.

Ces approches permettent également l'identification de nouveaux composés de familles chimiques différentes comme touches pharmacologiques, possédant un même mode de liaison, visant à améliorer la diversité chimique. Les chefs de file peuvent également être optimisés, par la modification rationnelle de ceux-ci à partir des informations obtenues avec les modèles pharmacophoriques 3D (155).

Enfin, la qualité des modèles est en relation directe avec la qualité de l'exploration conformationnelle de chaque ligand de référence. L'utilisation de trop de conformations permet l'obtention de bons modèles mais nécessite l'utilisation de beaucoup de ressources et de temps de calculs, et à l'opposé, l'utilisation de peu de conformations entraîne une exploration non exhaustive et l'obtention de modèles

pharmacophoriques de mauvaise qualité. L'étape de la recherche conformationnelle est donc l'étape critique du processus, qui doit être contrôlée précisément (127).

4.2.3. Exemple d'application

Les méthodes pharmacophoriques peuvent fournir de nouvelles familles chimiques de ligands pour une cible étudiée. Dans un article de 2009, une équipe autrichienne a mis en évidence de nouvelles familles chimiques de ligands du récepteur CB2 (*Cannabinoid receptor type 2*) en utilisant un criblage de bases de données virtuelles à partir de modèles pharmacophoriques (Figure 53) (163). La découverte de nouveaux ligands du récepteur CB2 permettrait l'obtention à terme de candidats médicaments pour la prise en charge de troubles inflammatoires, de l'athérosclérose et de l'ostéoporose (164).

A partir de 5 ligands de référence présentant des constantes d'inhibition de l'ordre du nanomolaire pour le récepteur CB2, jusqu'à 250 conformations différentes ont été générées par ligand. Ensuite, 10 modèles pharmacophoriques 3D ont été créés en utilisant la méthode des structures communes maximales, avec l'algorithme HipHop (165) implémenté dans le logiciel Catalyst. Deux modèles pharmacophoriques ont été validés à partir de 15 nouveaux ligands de référence, différents de ceux ayant permis la création des modèles.

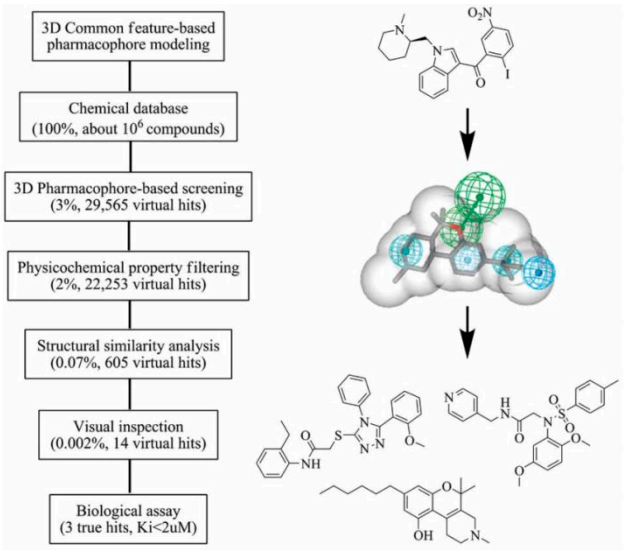


Figure 53 : Processus du criblage utilisant un modèle pharmacophorique dans la découverte de nouveaux ligands du récepteur CB2 (163)

En appliquant ces deux pharmacophores sur 6 bases de données comprenant au total 923.000 molécules, 29.565 composés ont pu être mis en évidence en correspondant aux deux modèles en même temps. Une étape de filtrage à partir de propriétés physico-chimiques, en appliquant des règles de Lipinski adaptées à l'étude, a ensuite été réalisée pour sélectionner 22.253 composés susceptibles d'être biodisponibles. Puis une analyse de similarité structurale a été réalisée, permettant l'obtention de 605 composés les plus diversifiés structuralement.

Enfin, une inspection visuelle de ces composés a été effectuée pour éviter les composés non stables chimiquement, non adaptés ou pouvant être toxiques. Quatorze composés ont été sélectionnés puis achetés, pour réaliser une confirmation expérimentale avec des tests biologiques, et mesurer notamment des constantes d'inhibition chez les récepteurs CB2 et CB1.

Suite aux tests expérimentaux, 7 composés ont présenté des constantes d'inhibition inférieures à 25 µM (Figure 54). Parmi eux, 3 composés prometteurs ont pu être mis en évidence, notamment un agoniste partiel compétitif d'une famille chimique déjà identifiée dans la littérature (Figure 54 – composé 8), ainsi que deux composés de nouvelles familles chimiques non connues dans la littérature, un antagoniste non compétitif (Figure 54 – composé 12) et un agoniste inverse compétitif (Figure 54 – composé 18).

compd	hCB_1 K_i (μM)[a]	hCB_2 K_i (μM)	(hCB_1 K_i)/(hCB_2 K_i)	binding behavior	receptor activity
5	0.797 ± 0.018	0.023 ± 0.004	35		
8	>50	1.78 ± 0.09	>27	competitive	partial agonist
11	>50	<25	>2		
12	>50	1.35 ± 0.17	>37	noncompetitive	antagonist
15	>50	<25	>2		
16	>50	<25	>2		
17	>50	<25	>2		
18	>50	2.1 ± 0.11	>23	competitive	inverse agonist

[a] K_i values are the mean values of three separate experiments performed in triplicate. The standard deviations for all K_i determinations are less than 15%.

Figure 54 : Mesures expérimentales des constantes d'inhibitions (Ki) pour les récepteurs CB1 et CB2 de 7 nouveaux composés identifiés à partir des modèles pharmacophoriques, et d'un agoniste bien connu du récepteur CB2 (composé 5) (163)

Cette étude à donc permis la découverte de nouveaux points de départ pour le développement de nouvelles molécules, agonistes comme antagonistes, ciblant le récepteur CB2.

5. Conclusion

Le processus de conception de médicaments est un parcours long et coûteux, demandant un investissement de plusieurs milliards de dollars sur près d'une décennie. Ce processus comporte quatre étapes successives, comprenant :

- La recherche d'un nouvel ingrédient pharmaceutique actif,
- Son développement préclinique dans le but de démontrer son efficacité et son innocuité sur des modèles *in vitro* et *in vivo*, ainsi que de comprendre son métabolisme,
- Son développement clinique au travers des démonstrations cliniques et pharmaceutiques, pour confirmer son efficacité et son innocuité chez l'Homme, pour le comparer à des traitements existants et développer sa formulation galénique,
- La procédure d'Autorisation de Mise sur le Marché, ainsi que la détermination de son prix et de son taux de remboursement.

Toutefois, très peu de candidats médicaments développés vont réussir à atteindre le marché, en raison de l'inflation des attentes des autorités de santé, de l'amélioration des standards de soins modernes et des pathologies ciblées de plus en plus complexes.
Afin de proposer des alternatives de développement, des méthodes *in silico* sont apparues avec l'évolution de l'informatique, notamment via l'utilisation de techniques de bioinformatique et de chémoinformatique. Ces techniques permettent de gagner 2 à 3 ans sur le temps de conception global d'un nouveau médicament, et d'économiser en moyenne près de 200 millions de dollars.

La chémoinformatique s'intéresse particulièrement à l'analyse, la simulation, la modélisation, la visualisation et la manipulation de données chimiques au travers de l'utilisation de l'informatique. Ce domaine permet entre autres d'identifier des composés chimiques, de modéliser et de mimer le comportement de molécules, ou encore de créer des modèles de prédictions d'activité biologique ou de propriétés moléculaires. En fonction des connaissances disponibles, des approches différentes sont utilisées.

Des approches basées sur la connaissance de structures tridimensionnelles des cibles pharmacologiques modélisent et évaluent les interactions entre un ligand et son site de liaison.

L'amarrage moléculaire et le criblage virtuel à haut débit permettent de déterminer précisément les interactions intermoléculaires réalisées entre une petite molécule et la cible étudiée, et d'identifier et optimiser de nouveaux composés prometteurs. La dynamique moléculaire étudie avec précision le comportement de protéines et d'autres biomolécules à l'échelle atomique, pour élucider leur

mécanisme moléculaire et étudier le comportement de ligands au sein de son site de liaison.

Des approches basées sur la connaissance de ligands naturels prédisent des propriétés ou établissent des règles pour les rendre plus actifs, moins toxiques, et/ou ayant des paramètres physico-chimiques optimisés.

La création de modèles de prédiction à partir de l'étude des relations quantitatives structure-activité (QSAR) et structure-propriété (QSPR) permettent de découvrir et d'optimiser de nouvelles molécules bioactives en prédisant leur activité, leur toxicité ou des paramètres moléculaires. La détermination de pharmacophores met en évidence les interactions importantes dans le mode de liaison à une cible pharmacologique, permet la recherche de nouveaux composés possédant un même mode de liaison, et leur optimisation rationnelle.

Toutes ces techniques sont rarement utilisées seules, mais combinées pour améliorer les hypothèses qui sont formulées. Elles sont utilisées tout le long de la conception de médicaments, de l'identification d'une cible pharmacologique jusqu'aux essais précliniques. La validation expérimentale *in vitro* et *in vivo* reste nécessaire pour valider les hypothèses formulées à partir des résultats *in silico* obtenus. Toutefois, les méthodes *in silico* permettent d'optimiser les tests expérimentaux en proposant une focalisation sur les composés les plus prometteurs, ce qui entraine la réduction du nombre d'expériences à réaliser.

L'amélioration exponentielle de la puissance de calcul, décrit par la loi de Moore, promet un avenir radieux pour les méthodes *in silico*. Cette puissance permettra l'application de méthodes dans de nouveaux contextes, comme le développement de la dynamique moléculaire à haut débit pour fournir des informations sur la cinétique, l'affinité et le mode de liaison de très nombreux ligands. Elle permettra également d'améliorer les techniques d'amarrage moléculaire et de criblage virtuel à haut débit via l'optimisation des fonctions de score, en se basant notamment sur le calcul d'énergies libres et de paramètres thermodynamiques. Enfin, elle conduira à la démocratisation des techniques de dynamique quantique, qui s'intéressent aux mouvements de biomolécules à l'échelle de l'électron à partir de la physique quantique.

Enfin, l'évolution rapide des modèles de prédiction basés sur l'intelligence artificielle et l'apprentissage profond (*deep learning*) entraîne une révolution, en permettant l'apprentissage de données complexes et abstraites pour l'obtention de modèles de prédictions s'approchant du raisonnement humain. Dans le processus de conception de médicaments, l'apprentissage profond permettra, à terme, d'optimiser

les essais cliniques en diminuant le besoin de patients nécessaires pour chaque étape, tout en permettant de traiter des quantités de données astronomiques. Les essais précliniques et cliniques *in silico* sont déjà reconnus par des autorités de santé comme l'Agence Européenne du Médicament, pour la réduction de l'expérimentation animale et la proposition d'alternatives aux études cliniques traditionnelles.

6. Bibliographie

1. Agence nationale de sécurité du médicament et des produits de santé. Analyse des ventes de médicaments en France en 2013 [Internet]. 2014. Disponible sur: ansm.sante.fr
2. LEEM. Médicaments : Rapport sur le progrès thérapeutique [Internet]. 2018. Disponible sur: leem.org
3. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016;47:20-33.
4. Hansen R, Chien R. The pharmaceutical development process: estimates of current development costs and times and the effects of regulatory changes. Issues in Pharmaceutical Economics. Lexington Books, Lexington, MA; 1979.
5. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ. 2003;22(2):151-85.
6. DiMasi JA, Hansen RW, Grabowski HG, Lasagna L. Cost of innovation in the pharmaceutical industry. J Health Econ. 1991;10(2):107-42.
7. Hughes J, Rees S, Kalindjian S, Philpott K. Principles of early drug discovery: Principles of early drug discovery. Br J Pharmacol. 2011;162(6):1239-49.
8. INPI. Comprendre la propriété intellectuelle - Le brevet [Internet]. Disponible sur: https://www.inpi.fr/fr/comprendre-la-propriete-intellectuelle/le-brevet
9. INPI. Bien préparer son dépôt - Cas particulier : les produits pharmaceutiques [Internet]. Disponible sur: https://www.inpi.fr/fr/comprendre-la-propriete-intellectuelle/le-brevet/cas-particulier-les-produits
10. Thomas G. Medicinal chemistry: an introduction. Chichester ; New York: Wiley; 2000. 539 p.
11. Collège National de Pharmacologie Médicale. Relation dose-effet clinique [Internet]. Disponible sur: https://pharmacomedicale.org/pharmacologie/pharmacologie-medicale-vue-d-ensemble/35-relation-dose-effet-clinique
12. Sanguinetti MC, Tristani-Firouzi M. hERG potassium channels and cardiac arrhythmia. Nature. 2006;440(7083):463-9.
13. Roth BL. Drugs and Valvular Heart Disease. N Engl J Med. 2007;356(1):6-9.
14. Nichols DE. Psychedelics. Barker EL, éditeur. Pharmacol Rev. 2016;68(2):264-355.
15. Collège National de Pharmacologie Médicale. Essais pré-cliniques des futurs médicaments [Internet]. Disponible sur: https://pharmacomedicale.org/pharmacologie/developpement-et-suivi-des-medicaments/25-essais-pre-cliniques-des-futurs-medicaments
16. HAS. Evaluation des médicaments en vue de leur remboursement - CT & CEEST [Internet]. 2017. Disponible sur: https://www.has-sante.fr/upload/docs/application/pdf/2017-03/dir4/v13ok-circuit_medicament_ct_ceesp-160317.pdf
17. ANSM. L'AMM et le parcours du médicament [Internet]. Disponible sur: https://www.ansm.sante.fr/Activites/Autorisations-de-Mise-sur-le-Marche-AMM/L-AMM-et-le-parcours-du-medicament/(offset)/3
18. HAS. Le parcours du médicament en France [Internet]. Disponible sur: https://www.has-sante.fr/upload/docs/application/pdf/2019-03/le_parcours_du_medicaments_en_france.pdf

19. Amelie. Remboursement des médicaments et tiers payant [Internet]. Disponible sur: https://www.ameli.fr/assure/remboursements/rembourse/medicaments-vaccins-dispositifs-medicaux/remboursement-medicaments-tiers-payant
20. Khanna I. Drug discovery in pharmaceutical industry: productivity challenges and trends. Drug Discov Today. 2012;17(19-20):1088-102.
21. Mullard A. 2018 FDA drug approvals. Nat Rev Drug Discov. 2019;18(2):85-9.
22. Ng R. Drugs [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2008. Disponible sur: http://doi.wiley.com/10.1002/9780470403587
23. Chopra A, Doiphode VV. Ayurvedic medicine: core concept, therapeutic principles, and current relevance. Med Clin North Am. 2002;86(1):75-89.
24. Bryan CP. The papyrus Ebers. 1930.
25. Pina AS, Hussain A, Roque ACA. An Historical Overview of Drug Discovery. In: Roque ACA, éditeur. Ligand-Macromolecular Interactions in Drug Discovery [Internet]. Totowa, NJ: Humana Press; 2010. p. 3-12. (Methods in Molecular Biology; vol. 572). Disponible sur: http://link.springer.com/10.1007/978-1-60761-244-5_1
26. Kubinyi H. General Aspects of Medicinal Chemistry. :61.
27. Drews J. Drug Discovery: A Historical Perspective. Science. 2000;287(5460):1960-4.
28. Maehle A-H. A binding question: the evolution of the receptor concept. Endeavour. 2009;33(4):135-40.
29. Watson JD, Crick FH, others. A structure for deoxyribose nucleic acid. Nature. 1953;171(4356):737–738.
30. Inglese J, Auld DS. High Throughput Screening (HTS) Techniques: Applications in Chemical Biology. In: Wiley Encyclopedia of Chemical Biology [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2008. p. wecb223. Disponible sur: http://doi.wiley.com/10.1002/9780470048672.wecb223
31. Liu R, Li X, Lam KS. Combinatorial chemistry in drug discovery. Curr Opin Chem Biol. 2017;38:117-26.
32. Leach AR, Gillet VJ. An introduction to chemoinformatics [Internet]. Dordrecht; London: Springer; 2007. Disponible sur: http://www.springerlink.com/openurl.asp?genre=book&isbn=978-1-4020-6290-2
33. Merriam-Webster. « in silico. » Merriam-Webster.com Dictionary [Internet]. Disponible sur: https://www.merriam-webster.com/dictionary/in%20silico
34. Turing AM. On Computable Numbers, with an Application to the Entscheidungsproblem. Proc Lond Math Soc. 1937;s2-42(1):230-65.
35. King GW, Cross PC, Thomas GB. The Asymmetric Rotor III. Punched- Card Methods of Constructing Band Spectra. J Chem Phys. 1946;14(1):35-42.
36. Chen WL. Chemoinformatics: Past, Present, and Future †. J Chem Inf Model. 2006;46(6):2230-55.
37. Science Museum / Science & Society Picture Library.
38. A. BARRINGTON BROWN, © GONVILLE & CAIUS COLLEGE / SCIENCE PHOTO LIBRARY.
39. Brooks-Bartlett JC, Garman EF. The Nobel Science: One Hundred Years of Crystallography. Interdiscip Sci Rev. 2015;40(3):244-64.
40. Drieding models, designed by Drieding at the University of Zurich, as available from Fisher Scientific (Pittsburgh, PA).

41. Moore GE. Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. IEEE Solid-State Circuits Soc Newsl. 2006;11(3):33-5.
42. Our World in Data. Moore's Law: Transistors per microprocessor [Internet]. Disponible sur: https://ourworldindata.org/grapher/transistors-per-microprocessor
43. Rupp K. 40 Years of Microprocessor Trend Data [Internet]. Disponible sur: https://github.com/karlrupp/microprocessor-trend-data
44. Hogeweg P. Simulating the growth of cellular forms. SIMULATION. 1978;31(3):90-6.
45. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology. Searls DB, éditeur. PLoS Comput Biol. 2011;7(3):e1002021.
46. Hood L, Galas D. The digital code of DNA. Nature. 2003;421(6921):444-8.
47. Brown FK, others. Chemoinformatics: what is it and how does it impact drug discovery. Annu Rep Med Chem. 1998;33:375–384.
48. Gasteiger J, éditeur. Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes [Internet]. 1re éd. Wiley; 2003. Disponible sur: https://onlinelibrary.wiley.com/doi/book/10.1002/9783527618279
49. Begam BF, Kumar JS. A Study on Cheminformatics and its Applications on Modern Drug Discovery. Procedia Eng. 2012;38:1264-75.
50. Lill MA, Future Science Ltd. In silico drug discovery and design [Internet]. 2013. Disponible sur: http://www.futuremedicine.com/doi/pdf/10.4155/9781909453012
51. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go: Docking/scoring methods-a review. Br J Pharmacol. 2009;153(S1):S7-26.
52. Goldbeck G. The economic impact of molecular modelling of chemicals and materials [Internet]. 2012. Disponible sur: https://gerhardgoldbeck.files.wordpress.com/2014/01/the-economic-impact-of-modelling.pdf
53. EMA. EMA Regulatory Science to 2025 - Strategic reflection [Internet]. 2020. Disponible sur: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection_en.pdf
54. Leach AR. Molecular modelling: principles and applications. 2nd ed. Harlow, England ; New York: Prentice Hall; 2001. 744 p.
55. Robson B, McBurney R. The role of information, bioinformatics and genomics. In: Drug Discovery and Development [Internet]. Elsevier; 2013. p. 77-94. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/B978070204299700007X
56. Engel ET, Gasteiger J. Applied Chemoinformatics. :646.
57. Schaduangrat N, Lampa S, Simeon S, Gleeson MP, Spjuth O, Nantasenamat C. Towards reproducible computational drug discovery. J Cheminformatics. 2020;12(1):9.
58. Bruno A, Costantino G, Sartori L, Radi M. The In Silico Drug Discovery Toolbox: Applications in Lead Discovery and Optimization. Curr Med Chem. 2019;26(21):3838-73.
59. Pubchem - Ibuprofen [Internet]. Disponible sur: https://pubchem.ncbi.nlm.nih.gov/compound/Ibuprofen

60. Vuorinen A, Schuster D. Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. Methods. 2015;71:113-34.
61. IUPAC. The IUPAC International Chemical Identifier: InChI—A New Standard for Molecular Informatics [Internet]. 2006. Disponible sur: http://publications.iupac.org/ci/2006/2806/2806-pp12-15.pdf
62. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Model. 1988;28(1):31-6.
63. Orlando BJ, Lucido MJ, Malkowski MG. The structure of ibuprofen bound to cyclooxygenase-2. J Struct Biol. 2015;189(1):62-6.
64. Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References [Internet]. 1re éd. Wiley; 2009. (Methods and Principles in Medicinal Chemistry; vol. 41). Disponible sur: https://onlinelibrary.wiley.com/doi/book/10.1002/9783527628766
65. Roy K, Kar S, Das RN. Chemical Information and Descriptors. In: Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment [Internet]. Elsevier; 2015. p. 47-80. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/B9780128015056000028
66. Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. Weinheim: Wiley-VCH; 2009. (Methods and principles in medicinal chemistry).
67. Mauri A, Consonni V, Pavan M, Todeschini R. Dragon software: An easy approach to molecular descriptor calculations. Match. 2006;56(2):237–248.
68. Landrum G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Academic Press; 2013.
69. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminformatics. 2017;9(1):33.
70. Grisoni F, Consonni V, Todeschini R. Impact of Molecular Descriptors on Computational Models. In: Brown JB, éditeur. Computational Chemogenomics [Internet]. New York, NY: Springer New York; 2018. p. 171-209. (Methods in Molecular Biology; vol. 1825). Disponible sur: http://link.springer.com/10.1007/978-1-4939-8639-2_5
71. Fernández-de Gortari E, García-Jacas CR, Martinez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. J Cheminformatics. 2017;9(1):9.
72. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996;16(1):3-50.
73. Dobson CM. Chemical space and biology. Nature. 2004;432(7019):824-8.
74. PubChem - Statistics [Internet]. Disponible sur: https://pubchemdocs.ncbi.nlm.nih.gov/statistics
75. Bunin BA, Siesel B, Morales G, Bajorath J. Chemoinformatics: Theory, Practice, & Products. :302.
76. Dictionnaire Larousse. Définition de « protéine » [Internet]. Disponible sur: https://www.larousse.fr/dictionnaires/francais/prot%C3%A9ine/64527
77. Mader SS, Windelspecht M. Human biology. 12th ed. New York, NY: McGraw-Hill; 2012. 1 p.
78. Cooper GM. The cell: a molecular approach. Eighth edition. Oxford ; New York: Sinauer Associates, an imprint of Oxford University Press; 2019.

79. Dvir H, Silman I, Harel M, Rosenberry TL, Sussman JL. Acetylcholinesterase: From 3D structure to function. Chem Biol Interact. 2010;187(1-3):10-22.
80. Banaszak L. Foundations of structural biology. San Diego, Calif: Academic; 2000. 168 p.
81. Leach AR. Ligand-Based Approaches: Core Molecular Modeling. In: Comprehensive Medicinal Chemistry II [Internet]. Elsevier; 2007. p. 87-118. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/B008045044X002467
82. Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham, III TE, Cruzeiro VWD. AMBER 2018. Univ Calif San Franc. 2018;
83. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. J Phys Chem B. 1998;102(18):3586-616.
84. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1-2:19-25.
85. Jorgensen WL, Tirado-Rives J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc. 1988;110(6):1657-66.
86. Lengauer T, Rarey M. Computational methods for biomolecular docking. Curr Opin Struct Biol. 1996;6(3):402-6.
87. Stouten PFW, Kroemer RT. Docking and Scoring. In: Comprehensive Medicinal Chemistry II [Internet]. Elsevier; 2007. p. 255-81. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/B008045044X002534
88. Kurji N. Computational Drug Discovery: A look at yesterday, today, and tomorrow — Biophysics and Artificial Intelligence [Internet]. 2018. Disponible sur: https://blog.usejournal.com/artificial-intelligence-in-drug-discovery-a-special-perspective-from-dr-5adcd688b0ec
89. Jacob RB, Andersen T, McDougal OM. Accessible High-Throughput Virtual Screening Molecular Docking Software for Students and Educators. Lewitter F, éditeur. PLoS Comput Biol. 2012;8(5):e1002499.
90. Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. Curr Comput Aided-Drug Des. 2011;7(2):146-57.
91. Krieger E, Nabuurs SB, Vriend G. Homology modeling. Methods Biochem Anal. 2003;44:509–524.
92. Baell JB, Holloway GA. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. J Med Chem. 2010;53(7):2719-40.
93. Baell J, Walters MA. Chemistry: Chemical con artists foil drug discovery. Nature. 2014;513(7519):481-3.
94. Baell JB, Nissink JWM. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. ACS Chem Biol. 2018;13(1):36-44.
95. Höltje H-D, éditeur. Molecular modeling: basic principles and applications. 2nd ed. Weinheim: Wiley-VCH; 2003. 228 p. (Methods and principles in medicinal chemistry).
96. Waszkowycz B. Towards improving compound selection in structure-based virtual screening. Drug Discov Today. 2008;13(5-6):219-26.
97. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. Proc Natl Acad Sci. 1999;96(18):9997-10002.

98. Peach ML, Nicklaus MC. Combining docking with pharmacophore filtering for improved virtual screening. J Cheminformatics. 2009;1(1):6.
99. Shah B, Modi P, Sagar SR. In silico studies on therapeutic agents for COVID-19: Drug repurposing approach. Life Sci. 2020;252:117652.
100. Koshland DE. The Key–Lock Theory and the Induced Fit Theory. Angew Chem Int Ed Engl. 1995;33(2324):2375-8.
101. Tuckerman ME, Martyna GJ. Understanding Modern Molecular Dynamics: Techniques and Applications. J Phys Chem B. 2000;104(2):159-78.
102. Schlick T. Molecular Modeling and Simulation: An Interdisciplinary Guide: An Interdisciplinary Guide [Internet]. New York, NY: Springer New York; 2010. (Interdisciplinary Applied Mathematics; vol. 21). Disponible sur: http://link.springer.com/10.1007/978-1-4419-6351-2
103. Vlugt TJH, Eerden JPJM van der, Dijkstra M, Smit B, Frenkel D. Introduction to molecular simulation and statistical thermodynamics [Internet]. 2009. Disponible sur: http://homepage.tudelft.nl/v9k6y/imsst/book-15-6-2009.pdf
104. Newton I. Principes mathématiques de la philosophie naturelle. 1687.
105. Gibbs JW. Elementary principles in statistical mechanics. 1902. N Y Charles Scribner's Sons. 1960;
106. Ganesan A, Coote ML, Barakat K. Molecular dynamics-driven drug discovery: leaping forward with confidence. Drug Discov Today. 2017;22(2):249-69.
107. Nzjacobmartin. Figure by Nzjacobmartin - Own work, CC BY-SA 4.0, [Internet]. Disponible sur: https://commons.wikimedia.org/w/index.php?curid=61475757
108. Madej BD, Walker R. An Introduction to Molecular Dynamics Simulations using AMBER [Internet]. 2015. Disponible sur: https://ambermd.org/tutorials/basic/tutorial0/index.htm
109. Central Michigan University. Model Box Periodic Boundary Conditions - P.B.C. [Internet]. Disponible sur: http://isaacs.sourceforge.net/phys/pbc.html
110. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol. 2019;20(11):681-97.
111. Duneau J-P. Simulation de dynamique moléculaire [Internet]. 2004. Disponible sur: https://lism.cnrs-mrs.fr/JS_files/Page_JP/Biomolec/DM5DynTheo.pdf
112. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera - A visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.
113. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. J Mol Graph. 1996;14(1):33-8.
114. Kumar A, Purohit R. Use of Long Term Molecular Dynamics Simulation in Predicting Cancer Associated SNPs. MacKerell AD, éditeur. PLoS Comput Biol. 2014;10(4):e1003318.
115. Frenkel D, Smit B. Understanding molecular simulation: from algorithms to applications. 2nd ed. San Diego: Academic Press; 2002. 638 p. (Computational science series).
116. Kim, Yongseong, Lee, Geun-U. Pharmacophore Modeling and Molecular Dynamics Simulation to Find the Potent Leads for Aurora Kinase B. Bull Korean Chem Soc. 2012;33(3):869-80.
117. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov. 2015;10(5):449-61.

118. Izrailev S, Stepaniants S, Isralewitz B, Kosztin D, Lu H, Molnar F, et al. Steered Molecular Dynamics. In: Deuflhard P, Hermans J, Leimkuhler B, Mark AE, Reich S, Skeel RD, éditeurs. Computational Molecular Dynamics: Challenges, Methods, Ideas [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999. p. 39-65. Disponible sur: http://link.springer.com/10.1007/978-3-642-58360-5_2

119. Laio A, Parrinello M. Escaping free-energy minima. Proc Natl Acad Sci. 2002;99(20):12562-6.

120. Fiorin G, Klein ML, Hénin J. Using collective variables to drive molecular dynamics simulations. Mol Phys. 2013;111(22-23):3345-62.

121. Incerti M, Russo S, Callegari D, Pala D, Giorgio C, Zanotti I, et al. Metadynamics for Perspective Drug Design: Computationally Driven Synthesis of New Protein–Protein Interaction Inhibitors Targeting the EphA2 Receptor. J Med Chem. 2017;60(2):787-96.

122. Vlachakis D, Bencurova E, Papangelopoulos N, Kossida S. Current State-of-the-Art Molecular Dynamics Methods and Applications. In: Advances in Protein Chemistry and Structural Biology [Internet]. Elsevier; 2014. p. 269-313. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/B978012800168400007X

123. Kalyaanamoorthy S, Chen Y-PP. Modelling and enhanced molecular dynamics to steer structure-based drug discovery. Prog Biophys Mol Biol. 2014;114(3):123-36.

124. Fang J, Wu P, Yang R, Gao L, Li C, Wang D, et al. Inhibition of acetylcholinesterase by two genistein derivatives: kinetic analysis, molecular docking and molecular dynamics simulation. Acta Pharm Sin B. 2014;4(6):430-7.

125. Davis AM. Quantitative Structure–Activity Relationships. In: Comprehensive Medicinal Chemistry III [Internet]. Elsevier; 2017. p. 379-92. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/B9780124095472123480

126. Chtita S, Bouachrine M, Lakhlifi T. Basic approaches and applications of QSAR/QSPR methods. Rev Interdiscip. 2016;1(1).

127. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment [Internet]. Elsevier; 2015. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/C20140002869

128. Meyer H. Zur theorie der alkoholnarkose. Arch Für Exp Pathol Pharmakol. 1899;42(2-4):109–118.

129. Fujita T. In memoriam Professor Corwin Hansch: birth pangs of QSAR before 1961. J Comput Aided Mol Des. 2011;25(6):509-17.

130. Fioravanzo E, Bassan A, Pavan M, Mostrag-Szlichtyng A, Worth AP. Role of in silico genotoxicity tools in the regulatory assessment of pharmaceutical impurities. SAR QSAR Environ Res. 2012;23(3-4):257-77.

131. Principal Component Analysis [Internet]. New York: Springer-Verlag; 2002. (Springer Series in Statistics). Disponible sur: http://link.springer.com/10.1007/b98835

132. Bournez C, Carles F, Peyrat G, Aci-Sèche S, Bourg S, Meyer C, et al. Comparative Assessment of Protein Kinase Inhibitors in Public Databases and in PKIDB. Molecules. 2020;25(14):3226.

133. Tabachnick BG, Fidell LS. Using multivariate statistics. 3rd ed. New York, NY: HarperCollins College Publishers; 1996. 880 p.

134. Tobias RD, others. An introduction to partial least squares regression. In: Proceedings of the twentieth annual SAS users group international conference. SAS Institute Inc Cary; 1995.
135. Balakrishnama S, Ganapathiraju A. Linear discriminant analysis-a brief tutorial. In: Institute for Signal and information Processing. 1998. p. 1–8.
136. Hilbe JM. Logistic regression models. CRC press; 2009.
137. Jarvis RA, Patrick EA. Clustering using a similarity measure based on shared near neighbors. IEEE Trans Comput. 1973;100(11):1025–1034.
138. Rohan J. Learn with an example : Hierarchical Clustering [Internet]. 2018. Disponible sur: https://medium.com/@rohanjoseph_91119/learn-with-an-example-hierarchical-clustering-873b5b50890c
139. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.
140. Cristianini N, Shawe-Taylor J. An introduction to support vector machines: and other kernel-based learning methods. Cambridge ; New York: Cambridge University Press; 2000. 189 p.
141. Loh W. Classification and regression trees. WIREs Data Min Knowl Discov. 2011;1(1):14-23.
142. Amit Y, Geman D. Shape Quantization and Recognition with Randomized Trees. Neural Comput. 1997;9(7):1545-88.
143. Chakure A. Random Forest Regression [Internet]. 2019. Disponible sur: https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f
144. Kriesel D. A Brief Introduction to Neural Networks [Internet]. 2007. Disponible sur: available at http://www.dkriesel.com
145. Tosco P, Mackey M. Lessons and Successes in the Use of Molecular Fields. In: Comprehensive Medicinal Chemistry III [Internet]. Elsevier; 2017. p. 253-96. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/B9780124095472123534
146. Refaeilzadeh P, Tang L, Liu H. Cross-Validation. Encycl Database Syst. 2009;5:532–538.
147. Tropsha A. Predictive Quantitative Structure–Activity Relationship Modeling. In: Comprehensive Medicinal Chemistry II [Internet]. Elsevier; 2007. p. 149-65. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/B008045044X002480
148. Sushko I. Applicability domain of QSAR models [Internet]. 2011. Disponible sur: https://mediatum.ub.tum.de/doc/1004002/1004002.pdf
149. Gonzalez Amaya JA, Cabrera DZ, Matallana AM, Arevalo KG, Guevara-Pulido J. In-silico design of new enalapril analogs (ACE inhibitors) using QSAR and molecular docking models. Inform Med Unlocked. 2020;19:100336.
150. Kwang LS. In silico high-throughput screening for ADME/Tox properties: PreADMET program. In: Abstr Conf Comb Chem Jpn. 2005. p. 22–28.
151. Göller AH, Kuhnke L, Montanari F, Bonin A, Schneckener S, ter Laak A, et al. Bayer's in silico ADMET platform: a journey of machine learning over the past two decades. Drug Discov Today. 2020;S1359644620302609.
152. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure Appl Chem. 1998;70(5):1129-43.
153. Beckett A, Harper N, Clitherow J. The importance of stereoisomerism in muscarinic activity. J Pharm Pharmacol. 1963;15(1):362–371.

154. Kier L. Molecular Orbital Theory in Drug Research. Academic Press. 1971;
155. Dror O, Shulman-Peleg A, Nussinov R, Wolfson H. Predicting Molecular Interactions in silico: I. A Guide to Pharmacophore Identification and its Applications to Drug Design. Curr Med Chem. 2004;11(1):71-90.
156. Reutlinger M, Koch CP, Reker D, Todoroff N, Schneider P, Rodrigues T, et al. Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. Mol Inform. 2013;32(2):133-8.
157. Rarey M, Dixon JS. Feature trees: A new molecular similarity measure based on tree matching. J Comput Aided Mol Des. 1998;12(5):471-90.
158. ter Laak AM, Venhorst J, Donne-Op den Kelder GM, Timmerman H. The Histamine H1-Receptor Antagonist Binding Site. A Stereoselective Pharmacophoric Model Based upon (Semi-)Rigid H1-Antagonists and Including a Known Interaction Site on the Receptor. J Med Chem. 1995;38(17):3351-60.
159. Barnum D, Greene J, Smellie A, Sprague P. Identification of Common Functional Configurations Among Molecules. J Chem Inf Comput Sci. 1996;36(3):563-71.
160. Jacobsson M, Lidén P, Stjernschantz E, Boström H, Norinder U. Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data. J Med Chem. 2003;46(26):5781-9.
161. Güner OF. Pharmacophore perception, development, and use in drug design. Vol. 2. Internat'l University Line; 2000.
162. Seidel T, Ibis G, Bendix F, Wolber G. Strategies for 3D pharmacophore-based virtual screening. Drug Discov Today Technol. 2010;7(4):e221-8.
163. Markt P, Feldmann C, Rollinger JM, Raduner S, Schuster D, Kirchmair J, et al. Discovery of Novel CB 2 Receptor Ligands by a Pharmacophore-Based Virtual Screening Workflow. J Med Chem. 2009;52(2):369-78.
164. Gao Q, Yang L, Zhu Y. Pharmacophore Based Drug Design Approach as a Practical Process in Drug Discovery. Curr Comput Aided-Drug Des. 2010;6(1):37-49.
165. Clement O, Mehl A. HipHop: pharmacophores based on multiple common-feature alignments. Pharmacophore Percept Dev Use Drug Des. 2000;69–84.

Université de Lille
FACULTE DE PHARMACIE DE LILLE
DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE
Année Universitaire 2019/2020

Nom : BEDART
Prénom : Corentin

Titre de la thèse : Contribution des méthodes *in silico* dans le processus de conception de médicaments

Mots-clés : Conception de médicaments ; Recherche pharmaceutique ; Développement pharmaceutique ; *In silico* ; Chémoinformatique ; Amarrage moléculaire ; Criblage virtuel à haut débit ; Dynamique moléculaire ; QSAR ; QSPR ; Intelligence artificielle ; Pharmacophore

---

Résumé :

La conception de médicaments est un processus complexe qui dure plus d'une décennie et demande des investissements de l'ordre du milliard de dollars. Peu de candidats médicaments développés vont atteindre le marché des médicaments en raison de l'amélioration des standards de soins modernes, des attentes des autorités de santé et de la recherche de traitements pour des pathologies complexes.

Les méthodes *in silico*, notamment au travers de la chémoinformatique, permettent d'optimiser le processus de conception à partir d'une vaste palette de techniques informatiques. Ces techniques, se basant sur la connaissance de ligands de référence ou de structures tridimensionnelles de la cible visée, permettent l'identification et l'optimisation de nouveaux candidats médicaments, ainsi que la prédiction de leur activité biologique, de leur toxicité, ou de leurs propriétés moléculaires essentielles.

Ce mémoire permet de comprendre le processus de conception de médicaments, l'impact des méthodes *in silico* et de la chémoinformatique sur celui-ci, ainsi que la vulgarisation scientifique de certaines méthodes.

<u>Membres du jury</u> :

Président : Pr. CHAVATTE Philippe, Docteur en Pharmacie, Professeur des Universités, Faculté de Pharmacie de Lille

Assesseur : Dr. STANDAERT Annie, Docteur en Pharmacie, Maître de Conférences des Universités, Faculté de Pharmacie de Lille

Membre extérieur : Dr. PLAETEVOET Marina, Docteur en Pharmacie, Pharmacie Plaetevoet à Flines-lez-Râches